Can Existing Cohort Studies with Self-Reported Diagnosis be used in the Study of Rare Cancer?

Evaluating the Feasibility of Rare Cancer Research in Emerging Cohorts in Canada

by

Emily G.D. Maplethorpe

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

Master of Science

in

Epidemiology

School of Public Health

University of Alberta

ABSTRACT

Rare cancers affect few people individually, but collectively account for approximately 22% of new cancer cases in Canada. However, prevention and population-level research on rare cancers have been limited as low case numbers make them difficult to study. The emergence of large, collaborative cohorts may offer the opportunity to study rare cancers with a sufficiently large sample size. The nested case-control design allows all cases to be utilized while maintaining the temporal advantage of a cohort design. This thesis evaluated the feasibility of large cohort studies as a tool for rare cancer research using data from Alberta's Tomorrow Project (ATP). Two questions were addressed: 1) Are self-reported diagnoses of rare cancer valid to use as an outcome in research when cancer registry linkage is unavailable? and 2) Is a nested case-control design a feasible option to study rare cancers in large cohort studies?

The validity of self-reported cancer diagnoses was explored through ATP linkage to the Alberta Cancer Registry (ACR). The first instance of self-reported cancer was compared to the first cancer diagnosis in the ACR after enrollment. Sensitivity and positive predictive value (PPV) were estimated for the reporting of overall cancer status, the reporting of common or rare cancer, and the reporting of site-specific cancer. Logistic regression analysis explored factors associated with false positive, false negative, and incorrect site reporting. Overall, rare cancers had a lower sensitivity and PPV than common cancers. Participants with a rare cancer were more likely to report an incorrect site than those with a common cancer. Rare cancers were also less likely to be captured by active follow-up than common cancers. Therefore, registry linkage is necessary to capture rare cancer diagnoses completely and accurately in large cohort studies.

A pilot etiologic study on pancreatic cancer assessed the feasibility of the nested case-control design for rare cancer research in the ATP cohort. Incidence density sampling was used to match controls to cases on follow-up time and other factors. Conditional logistic regression was used to investigate the association of pancreatic cancer with well-established risk factors and with less established dietary risk factors. The analysis was adequately powered to find estimated effects of established risk factors that were consistent with other literature. However, the dietary risk factor analysis was not adequately powered to detect low to moderate effects. Attrition limited the eligible pool of controls and introduced the possibility of healthy volunteer bias. Using a larger, national-scale cohort that would produce more cases and linking to vital statistics for passive follow-up of controls can mitigate these issues.

This analysis found that rare cancer research may be feasible in large cohort studies at a national scale if linkage to cancer registry and vital statistics is available. Removing barriers that currently prevent the sharing of linked data cross-provincially would allow for these opportunities in rare cancer research to be explored.

PREFACE

This thesis is an original work by Emily Maplethorpe. The research project of which this thesis is a part received research ethics approval from the Health Research Ethics Board of Alberta (Study ID CC-16-0880).

ACKNOWLEDGEMENTS

Firstly, thank you to my supervisors, Dr. Yan Yuan and Dr. Faith Davis, for their mentorship, guidance, patience, and support throughout this process. I am very grateful for the opportunity to have worked with them and for their contribution in making this an incredibly valuable experience. I would also like to thank Dr. Emily Walker for her role in data acquisition, and Trenton Smith for his role in data management and helpful advice.

Secondly, I would like to thank the staff and funders of the Alberta Tomorrow Project for their hard work and commitment to health research. Alberta's Tomorrow Project is only possible due to the commitment of its research participants, its staff and its funders: Alberta Health and the Alberta Cancer Prevention Legacy Fund, Alberta Cancer Foundation, Canadian Partnership Against Cancer and substantial in kind funding from Alberta Health Services. The views expressed herein represent the views of the author(s) and not of Alberta's Tomorrow Project or any of its funders.

I would also like to thank those who funded my graduate endeavours: University of Alberta School of Public Health and Canadian Institutes of Health Research (Frederick Banting and Charles Best Canada Graduate Scholarship).

Thirdly, I am incredibly grateful for the support and friendship of my peers here at the School of Public Health. Finally, my deepest thanks go out to my friends and family for their ongoing support, love, and encouragement, without which I cannot imagine making it to this point.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: Introduction

1.1 Rare Cancer: A Paradox

Rare diseases, including rare cancer, experience a paradox in their occurrence in the population. Individually, rare cancers affect very few individuals relative to common cancers. Collectively, however, rare cancers affect many people. Rare cancers account for 25% of cancer diagnoses in the US, which defines rare cancers as those with an incidence rate of <15/100,000/year[1]. In Europe, which uses finer cancer categories and a stricter incidence rate definition of <6/100,000/year, rare cancers account for 22% of cancer diagnoses[2]. These proportions will likely increase; as cancer biology continually advances, new molecular subtypes result in the separation of previously common cancers into several different types of rare cancer[3]. Though these separations are often molecularly based, resulting cancer types may also have different etiology, prognosis, or treatment implications that are clinically significant. Clinical differences are likely to alter the categorization of cancer for practice and research purposes. As a result of these biology and taxonomy changes, rare cancer discussions at present may underestimate the burden of rare cancer in the future[4].

1.2 The Rare Cancer Burden

Rare cancers contribute disproportionately to cancer-related mortality and morbidity. In Europe, the 5-year relative survival for rare cancers was 47%, compared to 65% for common cancers[2]. On average, survival declined faster over time for rare cancers compared to common cancers, supporting the idea that treatments for rare cancers are less effective than common cancers. Interestingly, for patients aged 0-39, a group in which rare cancers (usually embryonal or haematological types, for which effective treatments are available) are “more common” than common cancers, survival was similar between common and rare cancers. In 75-99 year olds, survival of rare cancers (usually rare epithelial forms, which are poorly understood and have few treatments available) was almost half that of common cancers[2]. Rare cancer survival is also poorer than common cancer survival in the US. Five-year survival for rare cancers was 55% for males and 60% for females, compared to common cancer 5-year survival of 74% for males and 75% for females[5]. Rare cancer survival is worse than common cancers across all age groups in the US, but rare cancer survival worsens compared to common cancers as patients get older, similar to the pattern seen in Europe[5].

Just as rare cancers account for a significant amount of cancer incidence, they also account for a significant proportion of cancer prevalence[2]. Cancers have a significant impact on society due to associated morbidity and high financial burden to the healthcare system[6]. For example, primary malignant brain tumor patients in the US can incur as much as 20 times the health-related costs of cancer free controls during the course of their disease[7]. Though all cancers have significant health care costs associated, rare cancers tend to have more outpatient and ER visits, hospital admissions, laboratory and radiographic procedures, and pharmaceuticals associated with their diagnosis and treatment. A 2004 analysis in the US found that less common cancer types, such as pancreatic cancer, had significantly higher associated costs than more common cancers, such as prostate cancer, though only a few cancer types were looked at individually[8].

A diagnosis of rare cancer is also burdensome to individual patients. Patients often experience a lack of understanding of their diagnosis or lack confidence in clinical decision making. It may take longer to comprehend their situation[6,9]. Though any cancer diagnosis is overwhelming, a rare cancer diagnosis often compounds this feeling. Few resources are available for those seeking information on their diagnosis and assessing treatment options[9,10]. Feelings of isolation are common in patients with rare cancer. A patient may not know of anyone else affected, those offering support are usually unfamiliar with the disease, and social supports and sources of information are often scarce[9]. This feeling of isolation is particularly pronounced in rare cancers that have other stigmas, such as anal cancers and rare gynecological cancers[9,10]. Rare cancer patients often travel greater distances and spend more time away from home in order to get treatment[9]. Conversely, ample resources and social supports are available for common cancers such as breast, prostate, and lung cancer in clinical settings, online, and often, in the patient's community.

1.3 Rare Cancer in Canada

Little information exists on rare cancer distribution, burden, cost, or patient experience in Canada. While the Canadian Cancer Society (CCS) publications offer a breadth of information, it focuses on more common cancers and offers little information on "all other cancers". CCS roughly estimates that rare cancers account for approximately 25% of incident cancer cases in Canada, 20-25% of cancer prevalence, and 20-25% of cancer deaths[11]. Survival information is limited to select cancer types. Recent work has suggested that rare cancers make up approximately 22% of incident cancer diagnoses in Canada under the US definition of <15 cases/100 000/year[12]. Cancer is Canada's seventh most costly illness or injury, and the costliest illness in terms of premature death and loss of productivity[11]. The proportion of these costs relating to rare cancer specifically is unknown. Regardless of the exact proportion, it is apparent that rare cancers account for a significant amount of cancer cases in Canada.

1.4 Challenges in Rare Cancer Research

Despite the apparent burden that rare cancer has on society and on patients, knowledge on the causes, etiology, diagnosis, and treatment options of rare cancers is lacking. Significant challenges prevent progress in the study of rare cancer. In particular, traditional clinical research designs are harder to execute for many reasons. There are a small number of patients with the disease, resulting in long recruitment periods, relaxed inclusion criteria, and difficulty setting up a concurrent control group, all of which can affect the internal and external validity of the study[13-15]. A lack of clinical expertise and specialized centers can result in poor diagnostic precision, delayed diagnosis, and therapeutic mismanagement, which can misclassify patients or affect their eligibility for a trial, not to mention their own prognosis[4,13-15]. Achieving a sample size for sufficient statistical power may require national or international collaboration, however, trials of this size are financially burdensome, logistically challenging, and subject to the complexities of differing regulations, management practices, and health priorities across regions[4,15].

For these reasons, among others, willingness to fund and carry out trials of this complexity is limited. Pharmaceutical companies have little incentive to invest in diseases and treatments that have such a small market[14-16]. Rare cancer trials may be plagued with uncertainty, but pharmaceutical companies are still bound by strict requirements of evidence and safety to continue[4,14]. Though new research methods that reduce study execution costs are a possibility to prevent investors from being discouraged, designs must still be approved and supported by health authorities in order to move forward[17].

The challenges facing rare cancer research, coupled with the low incentive for large funders to tackle them, has resulted in a lack of knowledge surrounding rare cancer. This not only makes clinical decision making difficult[1-3], but also hinders the ability of policy makers to implement appropriate policies related to rare cancer care and prevention[2,14,18]. Though the rare cancer burden is recognized by governments and health authorities, little progress has been made to evolve funding and research mechanisms[16].

1.5 Population Data: An Opportunity to Advance Rare Cancer Research

In an effort to overcome some of these challenges and advance rare cancer knowledge, large observational databases have been recognized as an opportunity to study risk factors and natural history of rare disease with a sufficiently large sample size[15,19]. Though it requires extensive costs and resources to carry out a cohort study for the purposes of studying the etiology of a rare disease, existing cohorts can be utilized for this purpose through the cooperation of researchers and research bodies. Rare cancers have smaller case numbers than diseases studied in traditional cohort designs, but other epidemiological study designs, such as the nested case-control, may overcome this limitation.

Data collected in large observational cohorts are often self-reported, as this is a low-cost option to gather information from a large population. Population databases and registries can be used to gain or confirm diagnosis or outcome information, as well as gather additional exposure and health outcome data. In fact, population-based registries are implicated as a critical source of information for the study of rare cancers[2,14,15,20]. In particular, collaboration and linkage between large cohort studies and population registries offers an opportunity to improve epidemiological surveillance and address questions that traditional research designs are unable to answer[14,15]. However, data privacy and ethics regulations create barriers in the sharing of information across institutions and organizations, introducing a major challenge to rare cancer research[14,15]. In Canada, barriers to cross-provincial information sharing have been acknowledged as a challenge in the sharing and linkage of information between provinces, cohorts and administrative sources[20,21]. Without the additional information or confirmation from registry data, cancer diagnosis information from a cohort must be valid enough to use as an outcome in epidemiological studies.

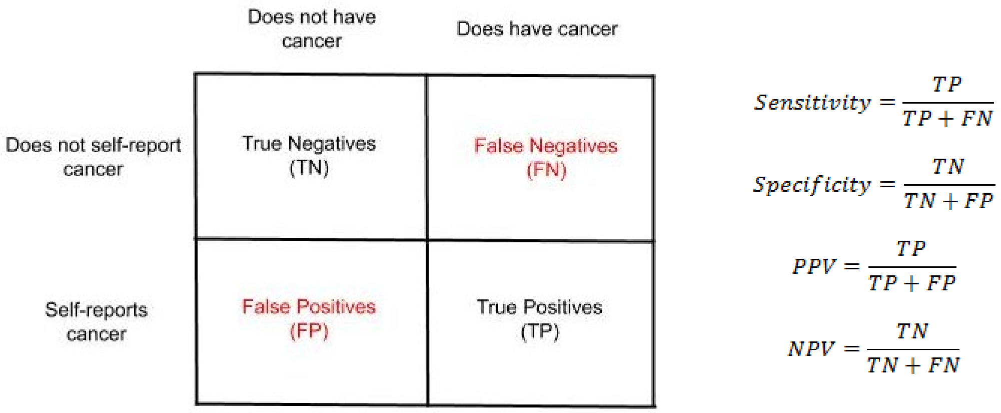
Therefore, two issues arise in the question of whether rare cancer etiology can be practically studied in large observational cohorts in the absence of cancer registry linkage:

1) Is self-reported cancer diagnosis in large observational cohorts a valid outcome for cancer research, in the absence of linkage to cancer registry?
2) Is it feasible to carry out a nested-case control study with sufficient power on rare cancers within large observational cohort studies?

1.6 Accuracy of Self-Reported Cancer Diagnosis

The accuracy of self-reported diagnoses in this context is best evaluated by finding their validity. Adapting a definition of validity by Streiner et al. (2015) to this context, self-report validity is whether or not self-reported diagnosis can draw accurate conclusions about the presence of cancer in an individual[22]. More specifically, criterion validity will be evaluated by comparing self-reports to a ‘gold standard’ measure for the presence of cancer: in this case, the cancer registry. Perhaps most relevant in the evaluation of self-report validity is the sensitivity and positive predictive value (PPV). Sensitivity is the ability of the report to correctly identify those with cancer, or the proportion of those with cancer who report it. The PPV is proportion of people who report cancer that actually have cancer. Low sensitivity implies a high number of people with disease are not reporting (false negatives); using self-reported diagnosis as an outcome will misclassify some cases as disease free. A low PPV implies that disease-free people are reporting that they have disease (false positives) and they will be misclassified as cases

(Figure 1.1). Both a low sensitivity and a low PPV have the potential to bias the results of a study using self-report data due to the misclassification of participants' cancer status. Specificity, or proportion of people without cancer who do not report cancer, and negative predictive value (NPV), or the proportion of those who do not report cancer that truly do not have cancer, are consistently very high in cancer self-report validation studies (both >90%)[23]. Since most people do not get cancer, there are a high number of true negatives compared to false positives and false negatives. Therefore, self-report validity reports commonly focus on sensitivity and PPV (Table S1). Little information exists on the validity of self-reported cancer diagnosis in a Canadian context; research is needed to determine if self-reported diagnosis in Canada is valid and what factors may affect this validity.



	Does not have cancer	Does have cancer
Does not self-report cancer	True Negatives (TN)	False Negatives (FN)
Self-reports cancer	False Positives (FP)	True Positives (TP)

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$PPV = \frac{TP}{TP + FP}$$

$$NPV = \frac{TN}{TN + FN}$$

Figure 1.1. Diagram representing the categorization of self-reports and their notation and equations using these categories for sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV).

1.7 Nested Case-Control: A Potential Design in the Study of Rare Cancer

Provided a valid outcome can be determined, the nested case-control study design may provide an opportunity to study the etiology and risk factors of rare cancers within a large cohort. Ernster (1994) provides a brief summary of the traditional cohort and case-control designs and compares them to the nested case-control design[24]. The nested case-control, as a hybrid design, provides the advantages of case-control design while maintaining the temporal advantage of the cohort design. All participants are disease free at study start and all cases that develop in the cohort can be used in analysis. Exposure information can be obtained at this time before the onset of disease and collected similarly for all participants regardless of disease status. Incidence density sampling, where a control or group of controls is selected from those at risk of the disease at the time of the case, is utilized, allowing for an unbiased estimate of the rate ratio[25]. A control can later become a case, and a control may be selected for more than one case.

To test the feasibility of this design in the study of rare cancer in Canada, pancreatic cancer may be a good candidate. Pancreatic cancer was identified as a rare cancer in a recent, not yet published, analysis by Walker et al.[26], using the US definition of rare cancer: an incidence rate of <15/100,000/year[1]. As pancreatic cancer incidence lies just below this cutoff, there are likely still a sufficient amount of cases to carry out a nested case-control analysis within a smaller population. Like other rare cancers, research on the etiology of pancreatic cancer is lacking[11]. There are however, several well-established risk factors, such as tobacco smoking, a family history of pancreatic cancer, history of diabetes, and body mass index (BMI)[11]. The association between these factors and pancreatic cancer can be estimated in a nested case-control study, provided they are available in the cohort, and the estimates from this study compared to what is expected from previous literature. This provides the opportunity to evaluate the feasibility of this study design and whether it could produce reliable results in the study of other poorly understood rare cancers.

1.8 The Canadian Partnership for Tomorrow Project

In Canada, the Canadian Partnership for Tomorrow Project (CPTP) has the potential to offer the opportunity to explore whether existing large observational cohorts can be used to study the etiology of rare cancers. The CPTP is a collaboration effort between five regional cohorts—from British Columbia, Alberta, Ontario, Quebec, and Atlantic provinces—to create Canada's largest volunteer research participant cohort[27]. A sixth cohort from Manitoba is currently in formation. This cohort aims to address questions about what causes cancer and chronic diseases that cannot be addressed otherwise[27]. The CPTP collected baseline data on health, lifestyle, genetic, and environmental factors from over 300,000 participants since it launched in 2008. Biological samples collected at baseline, such as blood, urine, and saliva, are also available, though not for all participants. Over half of CPTP participants have a venous blood sample, nearly 101,000 have a urine sample, and just over 19,000 have a saliva sample[28]. Toenail clippings are available through the Atlantic regional cohort for their participants[28]. Follow-up data on factors collected through questionnaires is expected to be available in 2019. All data obtained through cohort administered questionnaires, including cancer diagnosis, is self-reported.

Though most participants have consented to linkage with administrative health records and cancer registries, these agreements are made within the original regional cohorts[27]. Researchers cannot currently access the entirety of CPTP linked to the national cancer registry. Legislation specifies that administrative data cannot cross provincial boundaries, and though this challenge has been acknowledged by stakeholders[20], there has yet to be a solution. At the present time, cancer registry linked regional cohort data can be obtained by requesting cohort access from the regional cohorts[28]. In order to get cancer registry linked data for the Canadian cohort, one would need to apply separately to each regional cohort for data access and request linkage to the appropriate provincial registry. Each of these regions have their own application processes, regulations, and fees. Data would be dispensed separately by each cohort according to their timelines, specifications, and standards. This process would be logistically challenging, lengthy, and costly.

1.9 Alberta's Tomorrow Project

Therefore, the ability for CPTP to offer insight on the feasibility of observational cohorts as a tool to study rare cancers is limited due to the barriers that prevent cross-provincial information sharing. Alberta's Tomorrow Project (ATP), a regional cohort in the CPTP, offers a more accessible opportunity to Alberta researchers to evaluate this issue. The ATP is a volunteer cohort study that, like the CPTP, aims to support advances in knowledge of cancer and chronic disease etiology to inform more effective risk reduction strategies[29]. This project explores whether the ATP dataset, as an example of a large observational cohort, can be used to study rare cancers. The two questions stated earlier will be addressed through the following two objectives:

1) Compare self-reported cancer diagnosis in ATP to cancer registry diagnosis to evaluate whether self-reported cancer diagnosis can be used as an outcome in cancer research, and
2) Carry out a pilot nested case-control study on pancreatic cancer to evaluate whether it is feasible to use observational cohort data to explore rare cancer etiology and risk factors.

If it is determined that self-reported cancer diagnoses are valid and rare cancer research is feasible in this cohort, then observational cohorts, including CPTP, may be used to advance rare cancer research, even in the absence of cancer registry linkage. If self-reported data are not valid, but there is still potential to carry out rare cancer research using these cohorts, then cancer registry linkage is a necessary step to unlock the potential CPTP may provide in rare cancer research. Efforts to remove barriers to data linkage and data sharing across provinces may be rationalized to offer valuable opportunities to study rare cancer in Canada.

1.10 References

1. Greenlee RT, Goodman MT, Lynch CF, Platz CE, Havener LA, Howe HL. The occurrence of rare cancers in U.S. adults, 1995-2004. *Public Health Reports*. 2010;125(1):28-43.

2. Gatta G, van der Zwan, JM, Casali PG, et al. Rare cancers are not so rare: The rare cancer burden in Europe. *Eur J Cancer*. 2011;47(17):2493-2511. doi: //dx.doi.org/10.1016/j.ejca.2011.08.008.

3. Boyd N, Dancey JE, Gilks CB, Huntsman DG. Rare cancers: A sea of opportunity. *Lancet Oncology*. 2016;17(2):e61. doi: 10.1016/S1470-2045(15)00386-1.

4. Komatsubara KM, Carvajal RD. The promise and challenges of rare cancer research. *Lancet Oncol*. 2016;17(2):136-138. doi: //dx.doi.org/10.1016/S1470-2045(15)00485-4.

5. DeSantis CE, Kramer JL, Jemal A. The burden of rare cancers in the United States. *CA: A Cancer Journal for Clinicians*. 2017;67(4):261-272. doi: 10.3322/caac.21400.

6. Pillai RK, Jayasree K. Rare cancers: Challenges & issues. *Indian Journal of Medical Research*. 2017;145(1):17-27. doi: 10.4103/ijmr.IJMR_915_14.

7. Kutikova L, Bowman L, Chang S, Long SR, Thornton DE, Crown WH. Utilization and cost of health care services associated with primary malignant brain tumors in the United States. *Journal of Neuro-Oncology*. 2007;81(1):61.

8. Chang S, Long SR, Kutikova L, et al. Estimating the cost of cancer: Results on the basis of claims data analyses for cancer patients diagnosed with seven types of cancer during 1999 to 2000. *Journal of Clinical Oncology*. 2004;22(17):3524-3530. doi: 10.1200/JCO.2004.10.170.

9. Wagland K, Levesque JV, Connors J. Disease isolation: The challenges faced by mothers living with multiple myeloma in rural and regional Australia. *European Journal of Oncology Nursing*. 2015;19(2):148-153. doi: 10.1016/j.ejon.2014.10.003.

10. Longabaugh M. Patient perspective and personal journey of treating a "Rare cancer". *Surg Oncol Clin N Am*. 2017;26(1):1-7. doi: 10.1016/j.soc.2016.07.014.

11. Canadian Cancer Society's Advisory Committee on Cancer Statistics. Canadian Cancer Statistics 2017. cancer.ca/Canadian-CancerStatistics-2017-EN.pdf. Updated 2017. Accessed February, 2018.

12. Maplethorpe E, Walker EV, Davis FG. Occurrence of rare cancer in Canada: The distribution of cancer incidence among the Canadian population from 2009-2013. *Poster session presented at: Canadian Research Data Center Network Conference, Hamilton, ON.* 2018.

13. Kawai A, Goto T, Shibata T, et al. Current state of therapeutic development for rare cancers in Japan, and proposals for improvement. *Cancer Sci*. 2018;109(5):1731-1737. doi: 10.1111/cas.13568.

14. von der Schulenburg JM, Pauer F. Reviews: Rare cancers—Rarity as a cost and value argument. *Journal of Cancer Policy*. 2017;11:54-59. doi: 10.1016/j.jcpo.2016.09.004.

15. Mathoulin-Pélissier S, Pritchard-Jones K. Evidence-based data and rare cancers: The need for a new methodological approach in research and investigation. *European Journal of Surgical Oncology*. 2018;45(1):22-30. doi: 10.1016/j.ejso.2018.02.015.

16. Negrouk A, Lacombe D, Trimble EL, Seymour M. Clinical research for rare cancers: Is it a reality in the global regulatory landscape? The International Rare Cancer Initiative. *Expert Opinion on Orphan Drugs*. 2014;2(5):433-440. doi: 10.1517/21678707.2014.888948.

17. Blay J, Coindre J, Ducimetière F, Ray-Coquard I. The value of research collaborations and consortia in rare cancers. *Lancet Oncology, The*. 2016;17(2):e69. doi: 10.1016/S1470-2045(15)00388-5.

18. Sandrucci S, Naredi P, Bonvalot S. Centers of excellence or excellence networks: The surgical challenge and quality issues in rare cancers. *European Journal of Surgical Oncology*. 2019;45(1):19-21.

19. Armstrong-Wells J, Goldenberg NA. Institution-based prospective inception cohort studies in neonatal rare disease research. *Seminars in Fetal and Neonatal Medicine*. 2011;16(6):355-358. doi: 10.1016/j.siny.2011.07.004.

20. Doiron D, Raina P, Fortier I. Linking Canadian population health data: Maximizing the potential of cohort and administrative data. *Can J Public Health*. 2013;104(3):e261.

21. Brown C. Barriers to accessing data are bad medicine. *CMAJ*. 2014;186(16):1203. doi: 10.1503/cmaj.109-4894.

22. Streiner DL, Norman GR, Cairney J. *Health measurement scales.* 5th ed. Oxford: Oxford Univ. Press; 2015.

23. Navarro C, Chirlaque MD, Tormo MJ, et al. Validity of self reported diagnoses of cancer in a major Spanish prospective cohort study. *Journal Epidemiology Community Health*. 2006;60:593-599.

24. Ernster VL. Nested case-control studies. *Prev Med*. 1994;23(5):587-590. doi: 10.1006/pmed.1994.1093.

25. Pearce N. What does the odds ratio estimate in a case-control study? *International Journal of Epidemiology*. 1993;22(6):1189-1192. doi: 10.1093/ije/22.6.1189.

26. Walker EV, Maplethorpe E, Davis FG. Common and rare cancer incidence rates in the Canadian population: 2009-2013. 2019. In preparation for submission at the time of this thesis.

27. Canadian Partnership for Tomorrow Project. About; https://www.partnershipfortomorrow.ca/about/. Accessed January, 2019.

28. Dummer TJ, Awadalla P, Boileau C, et al. The Canadian Partnership for Tomorrow Project: A pan-Canadian platform for research on chronic disease prevention. *CMAJ*. 2018;190(23):E717. doi: 10.1503/cmaj.170292.

29. Alberta's Tomorrow Project. All about Alberta's Tomorrow Project; https://myatp.ca/about-atp. Updated 2019. Accessed January, 2019.

CHAPTER 2: Comparison of Self-Reported Cancer Diagnosis in a Canadian Cohort to Cancer Registry

2.1 Introduction

Rare cancers account for a significant proportion of cancer cases in Canada and contribute disproportionately to cancer-related morbidity and mortality[1,2]. However, significant challenges prevent progress in the study of rare cancer. Traditional clinical research designs are harder to execute due to small sample size, diagnostic uncertainty, a lack of expertise, and high costs[3-5]. In an effort to overcome some of these challenges, large observational databases are recognized as an opportunity to study the etiology and natural history of rare disease with a sufficiently large sample size[5,6].

However, information in large observational cohort studies is often self-reported, as these methods are an efficient and low-cost option for research in large samples. The accuracy of self-reported cancer diagnosis must be considered if using it as an outcome in research. If self-reported data are not accurate, using it as an outcome measure in etiologic studies can result in biased or erroneous results. In this instance, cohort linkage to population-based registries would be a necessary step to utilize cohort data in the study of rare cancers. Population-based registries, and particularly, their linkage to large cohort studies, are implicated as a critical source of information for the surveillance and study of rare cancers[4,5,7]. However, data privacy and ethics regulations create barriers in the sharing of information across institutions, organizations, and provinces, preventing this linkage of data sources and introducing a major challenge to rare cancer research[4,5,8]. So, while linkage to cancer registry data removes the need to rely on self-reported data, logistic challenges may prevent this. Therefore, evaluation of the accuracy of self-reported cancer diagnoses is required to determine whether they are valid as an outcome. If so, large observational cohorts may be used for etiologic research in the absence of cancer registry linkage.

2.1.1 The Accuracy of Self-Reported Cancer Diagnoses

The accuracy of self-reported diagnoses in this context is best evaluated by finding their validity; that is, whether or not self-reported diagnosis accurately describes the presence of cancer in an individual. Validity, or specifically, criterion validity, is measured by comparing self-reports to a 'gold standard' measure for the presence of cancer[9]: in this case, the cancer registry. Perhaps most relevant in the evaluation of self-report validity is the sensitivity and positive predictive value (PPV); the proportion of those who have cancer that report cancer, and the proportion of people who report cancer that actually have cancer, respectively. Both a low sensitivity (high false negatives) and a low PPV (high false positives) imply a misclassification of disease status and have the potential to bias the results of a study. Specificity, or the proportion of those without cancer who do not report cancer, and negative predictive value (NPV), or the proportion of those who do not report cancer that truly do not have cancer, are consistently very high in cancer self-report validation studies (both >90%)[10]. Since most people do not get cancer, there are many true negatives relative to false positive and false negatives. Therefore, the following discussion of self-report validity will focus on sensitivity and PPV.

Little information exists in the validity of such self-reports in a specifically Canadian context. Self-report validation studies from the US, Australia, and Europe have found overall sensitivities ranging from 57.5%-90.3%[10-16] and overall PPV's ranging from 54.9%-75%[10,11,13-16] (Table S1). PPV was generally lower than the reported sensitivity. Studies from Korea and Japan, however, found lower sensitivities than those from US, Australia, and Europe, but similar PPV's. Sensitivity ranged from 36.0%-53%[17-19], with reported PPV's of 59.7% and 81.9%[17,18].

These differences in sensitivity across countries are largely attributable to cultural differences in disclosing cancer diagnosis to patients and social perceptions of illness[17-19]. As Canada is culturally more similar to the US, Australia, and Europe, we expect self-report sensitivities and PPV's to be similar to those found in these countries.

Self-report sensitivity and PPV has been found to vary greatly by cancer site. Breast cancer is consistently reported with one of the highest sensitivities and PPV's; some report sensitivity as high as 96.4%[12]. Breast cancer PPV ranges from 72-86%[10,11,13-15]. Prostate cancer also consistently has one of the highest sensitivities and PPV's, with sensitivity ranging from 77-94% and PPV ranging from 70-94%[13-16]. Melanoma of skin is often reported to have lower sensitivity and PPV than other commonly reported cancers; studies have reported sensitivity ranging from as low as 37% to upwards of 82%, and PPV of 34% to 60%[12-16]. Most studies looked at the same selection of cancer sites or groups, focusing on more "common" cancers such as breast, prostate, and lung. Exceptions to this trend only included a few additional cancer types and have few other results to compare to. No reports were found that explored whether there was a difference in sensitivity of common cancers and rare cancers collectively. There is not enough data on the validity of rare cancer self-reports to assume that they can be used as an outcome.

Several factors are thought to be associated with correct self-reporting of cancer. Those who are younger, have a shorter time since diagnosis, higher education attainment, family history of cancer, are female, do not smoke, are Caucasian, and have more comorbidities are more likely to correctly self-report cancer[10,12,14,15]. However, the effect of these factors on self-report validity varies from study to study and few studies have looked at their relevance in a Canadian population.

2.1.2 The Canadian Partnership for Tomorrow Project

Provided self-reported cancer validity if sufficient, self-reported diagnoses in large observational cohorts may be useful for advancing research in the etiology of rare cancers. In Canada, the Canadian Partnership for Tomorrow Project (CPTP) may offer this opportunity. The CPTP is a collaboration effort between five regional cohorts—from British Columbia, Alberta, Ontario, Quebec, and Atlantic provinces—to create Canada's largest volunteer research participant cohort[20]. Current efforts are attempting to form a sixth regional cohort in Manitoba. The CPTP collected baseline data on health, lifestyle, genetic, and environmental factors from over 300,000 participants since it launched in 2008. Follow up data on these factors is expected to be available in 2019. The cohort aims to address questions about what causes cancer and chronic diseases that cannot be addressed otherwise[20]. All data in these surveys, including cancer diagnosis, is self-reported. The accuracy of these self-reported diagnoses is not known to researchers accessing the data.

Linkage to cancer registry would eliminate the need to use self-reported diagnosis as an outcome. However, though most participants have consented to linkage with cancer registry, these agreements are made within the original regional cohorts[20] and administrative data cannot cross provincial boundaries[21]. While this challenge has been acknowledged[21], researchers cannot currently obtain the entirety of CPTP data linked to the national cancer registry. Regional cohort data linked to provincial cancer registry can be obtained by requesting cohort access and data linkage from each of the regional cohorts[22]. This would be a lengthy, costly, and logistically challenging process.

2.1.3 Alberta's Tomorrow Project

The ability of the CPTP to offer insight on the accuracy of self-reported diagnoses is limited due to barriers that prevent cross-provincial information sharing. Alberta's Tomorrow Project (ATP), a regional cohort in the CPTP, offers a more accessible opportunity to Alberta researchers to evaluate this issue. The ATP is a volunteer cohort study that, like the CPTP, aims to support advances in knowledge of cancer and chronic disease etiology to inform more effective risk reduction strategies[23]. The ATP may, then, act as an indicator of the accuracy of self-reported diagnoses in the CPTP.

This project compares self-reported primary cancer diagnoses in the ATP to cancer registry diagnoses to evaluate the accuracy, or more specifically, the validity, of self-reported cancer diagnosis in a Canadian context. Factors affecting self-report validity will also be explored. If self-reported diagnoses are valid, observational cohorts may be useful on their own in the study of rare cancers. If self-reports are not valid, efforts to remove some of the barriers to registry linkage can be rationalized to take advantage of the valuable information and opportunities available in these cohorts.

2.2 Methods

2.2.1 Data Source

Alberta's Tomorrow Project

ATP started recruitment in 2000, finished in 2015, and now has 55,000 participants[23]. Albertan residents aged 35-69 with no history of cancer, other than non-melanoma skin cancer, were eligible to enroll. Other enrollment criteria were that participants had to plan to reside in Alberta for at least one year and ability to compete written questionnaires in English[24]. The cohort was recruited in two phases. In Phase 1 (2000-2008), participants were recruited using random digit dialing. Households were selected from each of the 17 regional health authorities across Alberta in 2000 and one or two eligible participants were selected from each household[24]. This resulted in a cohort from across the province representing a wide range of sociodemographic and health-related factors[24]. Phase 2 (2008-2015) began after harmonization with the Canadian Partnership for Tomorrow Project (CPTP). CPTP-ATP recruitment was achieved through volunteer sampling using communication and advocating strategies to reach eligible participants[22].

This project included those participants recruited in Phase 1 (n=31,203). Upon recruitment, these participants completed the baseline Health and Lifestyle Questionnaire (HLQ), and, depending on when they were enrolled, had the opportunity to complete several follow up questionnaires; Survey 2004, Survey 2008, Updated Health and Lifestyle Questionnaire (UHLQ, 2009-2011) and CORE (2011-2015) (Figure S1). These surveys collected information on personal characteristics, lifestyle factors, and health status. These surveys also asked participants if they have ever been diagnosed with cancer, and if so, what type. In 2008, ATP joined the CPTP. The UHLQ and CORE questionnaires have been updated and administered through the CPTP to support this harmonization, though the information collected from these questionnaires is similar to ATP's original follow up surveys. More information on ATP surveys can be found at www.myatpresearch.ca/survey-information. All information in these questionnaires is self-reported.

Alberta Cancer Registry

The Alberta Cancer Registry ACR is a population-based registry that collects information on all new cancer cases and cancer deaths occurring in Alberta[25]. The ACR has achieved a Gold Standard from the North American Association of Central Cancer Registries (NAACCR) for many years[25]. Cancer registries that meet the Gold Standard have achieved the highest NAACCR standard for complete, accurate, and timely data, among other data quality measures[26]. The ACR has consistently achieved completeness, or the extent to which all new cancer cases are accurately captured, of over 95%[25]. Doctors and laboratories in the province are mandated to notify the ACR of new cancer cases[25]. The ACR achieves comparability by applying standard classification and coding practices[25]. The ACR records topography, morphology, and behavior using the International Classification of Diseases for Oncology (ICD-O): ICD-O-2 for cases before 2000 and ICD-O-3 for cases 2001 and onwards. Topography codes are consistent between the two versions of ICD-O, and the few morphology and behavior codes used in this analysis are not affected by the change in versions[27]. The highest level of accuracy is achieved by numerous data edits and additional data quality reviews by the Canadian Cancer Registry (CCR) and NAACCR[25].

As we are interested in extending the results to explore the utility of linkage to the CCR, cancer diagnoses that are not mandated to be reported to the CCR were excluded from the ACR for this analysis. The CCR mandates the reporting of all primary, malignant tumors (behavior 3) and all in situ/intra-epithelial/noninfiltrating/noninvasive tumors (behavior 2), *except* behavior 2 cervix and prostate cancer[28]. Non-melanoma skin cancers of any behavior code are not mandated. Though all borderline malignancies (behavior 1) and some benign tumors (behavior 0) of the brain and central nervous system are mandated to be reported, these were excluded from this analysis as it is unknown to what extent these borderline or benign diagnoses are described as "cancer" to patients. In exploratory analysis, behavior 0 (benign) and 1 (borderline) diagnoses accounted for less than 3% of ACR records, and less than 40% were self-reported.

Therefore, only diagnoses in the ACR of behavior 2 (in situ/noninvasive) or 3 (malignant) were included a diagnosis of cancer, excluding non-melanoma skin cancer and behavior 2 cervix and prostate cancer. Though exploratory analysis indicated that there were a large number of behavior 2 cervix cancer cases in the ACR (n=455), many occurred before baseline and most were not self-reported. ATP did not consider behavior 2 cervix cancer a prior cancer for enrolment purposes.

Ethics Approval

Ethics approval was obtained from the Health Research Ethics Board of Alberta (Study ID CC-16-0880).

Data Linkage

Alberta Cancer Registry data was obtained through linkage with Surveillance & Reporting, C-MORE CancerControl Alberta. ACR and ATP linkage was performed by these agencies prior to dispensing the data. Participants were linked on Alberta Personal Health Care Number, and confirmed on first name, last name, and date of birth[23]. Participants that did not consent to data linkage (n=360) were included in the dataset but given a value corresponding to "no consent" in a variable related to cancer status at baseline.

2.2.2 Determining the Accuracy of Self-Reported Primary Cancer Diagnoses

Data Preparation

There were 360 participants that did not consent to ACR linkage and were excluded from this analysis (Figure 2.1).

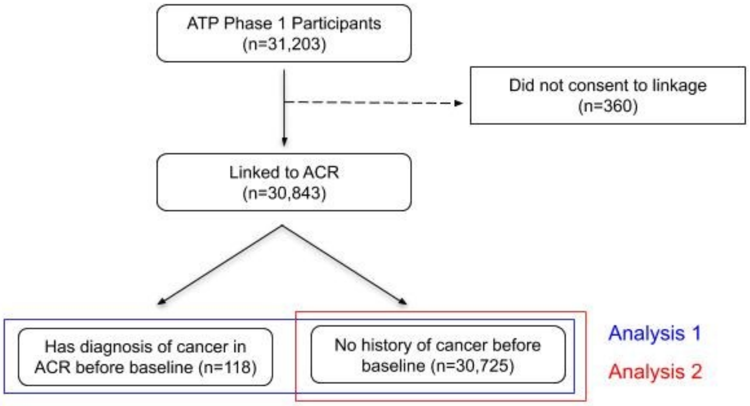

ATP Phase 1 Participants (n=31,203)
Did not consent to linkage (n=360)
Linked to ACR (n=30,843)
Has diagnosis of cancer in ACR before baseline (n=118)
No history of cancer before baseline (n=30,725)
Analysis 1
Analysis 2

Figure 2.1. A flowchart illustrating the steps of participant exclusion and the two study populations for self-report accuracy analyses. Participants with a diagnosis of cancer in the ACR before baseline are included in Analysis 1, but excluded in Analysis 2.
ATP=Alberta Tomorrow Project. ACR=Alberta Cancer Registry.

Though participants are required to have never had cancer at baseline, there were 118 participants that had their age of first cancer diagnosis in the ACR before age at baseline (HLQ). In the HLQ survey, when asked “Has a doctor ever told you that you have cancer (excluding non-melanoma skin cancer)?”, these participants reported that they had not. Due to the ambiguity of what cancer these participants may report in the future, we performed two analyses: one including these participants, where the first cancer in the ACR *after baseline* was compared to the first instance of self-reported cancer (excluding non-melanoma skin cancer), and another analysis excluding these participants, where the first cancer in the ACR and the first instance of self-reported diagnosis (excluding non-melanoma skin cancer) were compared. The first analysis assumes that if a participant does not report a cancer at baseline, a cancer they report in subsequent surveys, when asked the same question, is a new diagnosis that occurred after baseline. Cancer diagnoses in the ACR that occurred after an individual's last survey (ie. age at diagnosis is greater than the age at most recent survey receipt) were not included, as a participant did not have a chance to report. Therefore, only ACR diagnoses within an individual's follow up time were included (Figure 2.2). These diagnoses are termed “self-reportable” diagnoses for the purposes of this thesis.

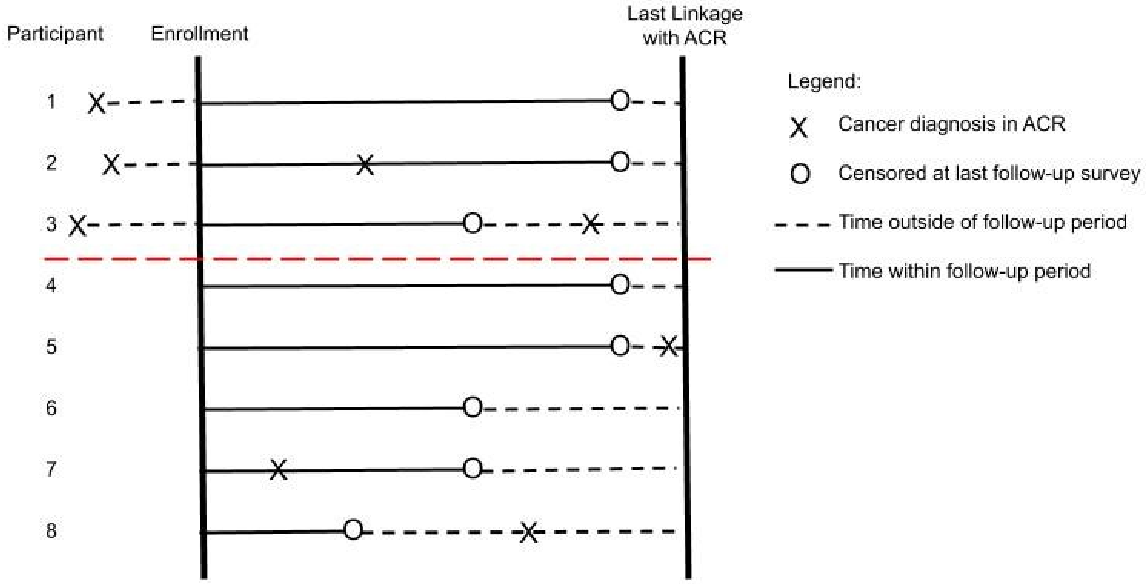

Participant
Enrollment
Last Linkage with ACR
1
2
3
4
5
6
7
8
Legend:
X Cancer diagnosis in ACR
O Censored at last follow-up survey
- - - Time outside of follow-up period
—— Time within follow-up period

Figure 2.2. A schematic demonstrating hypothetical ATP participants 1-8, their cancer diagnoses (X) and censoring time (O) at last follow-up survey. Only cancer diagnoses of behavior 2 or 3 (excluding non-melanoma skin cancer and behavior 2 cervix and prostate cancer) in the Alberta Cancer Registry (ACR) occurring within an individual’s follow up time (solid black line) are included as an ACR diagnosis in the self-report accuracy analyses. Cancer diagnoses occurring outside an individual’s follow-up time (dotted black line) would not have the opportunity to be self-reported and are not included in the analysis. Diagnoses occurring within an individual’s follow-up time had an opportunity to be reported in a later survey and are therefore considered “self-reportable” in this analysis. Participants 1-3 had a diagnosis of cancer in the Alberta Cancer Registry (ACR) before enrollment (baseline). These participants are included in Analysis 1, but their diagnosis before enrollment is not considered. In Analysis 2, participants 1-3 are excluded.

Categorizing ACR Diagnosis Cancer Site

Cancer site was generated from ICD-O-3 topography codes in the ACR, using cancer categories from the Surveillance, Epidemiology, End Results Program (SEER) 2018 classification scheme[29] (Table S2). Several categories were collapsed that were not expected to be differentiated in self-reports: corpus uteri and uterus, NOS (not otherwise specified) were collapsed to a single "Uterus" category and oropharynx, nasopharynx, hypopharynx, and pharynx were collapsed into a single "Throat" category.

Categorizing Self-Reported Diagnosis Cancer Site

Those who self-reported cancer diagnosis in a survey were also asked to record cancer type. The first instance of self-reported cancer type was categorized into an appropriate site category from the revised SEER 2018 scheme. A second observer categorized unique self-report responses with 92% agreement; 92% of unique self-reported type responses were categorized into the same SEER 2018 site category. This resulted in over 99% agreement between observers of the categorization of individual reports; 99% of all self-report responses were categorized into the same SEER 2018 site category. Survey 04, Survey 08, and UHLQ had open text options to record cancer type. The CORE survey had a drop-down menu with 22 cancer types to choose from (which all corresponded to a SEER category) or an "other" option where the participant filled in an open text question. All self-reported diagnoses were considered a cancer diagnosis except skin cancer responses that did not specify "melanoma"; these were assumed to be non-melanoma skin cancer. Though only Survey 04 specified to self-report cancer excluding non-melanoma skin cancer, the open text option in Survey 2004, Survey 2008, and UHLQ allowed melanoma skin cancer to be specified. In the CORE survey, "Skin" was a drop-down option without further specification; participants that chose "Skin" were contacted to determine a skin type. Responses were included in the open text cancer type variable (ATP, personal communication, May 6, 2019). There were 40 participants who self-reported having cancer for which a self-reported site could not be determined (type missing/don't know, type unclear or unspecific, or containing only non-site-specific histological information).

Categorizing Cancer Diagnosis as Common or Rare

Each ACR diagnosis and self-report diagnosis was defined as either common or rare depending on the cancer site. Sites were defined according to a recent analysis[1] that found the age-standardized incidence rate of cancer sites in Canada using the SEER 2018 scheme and defined them as common or rare according to the US definition of a rare cancer: an incidence rate of <15/100,000/year[30].

Sensitivity and PPV Calculations

Cancer categories in the ACR were used as the gold standard. Sensitivity and positive predictive value (PPV) were calculated as follows, with notations defined below:

$$Sensitivity\ = \frac{TP}{TP+FN} \qquad PPV\ = \frac{TP}{TP+FP}$$

$$where\ TP = true\ positive,\ \ FP = false\ positive, and\ FN = false\ negative.$$

Specificity and NPV were also calculated and found to be similarly high to previous studies (>98%), and therefore not presented. Three types of sensitivity and PPV calculations were run including those with cancer before baseline and excluding them: any-cancer overall diagnosis calculation, common or rare cancer site calculations, and site-specific calculations for each cancer site. The terms in these calculations are defined in Table 2.1.

Cancer sites that had at least 10 diagnoses in the ACR and/or 10 self-reported diagnoses were reported in their own in Table 2.2. Due to low sample size for many groups, sites were collapsed, calculated, and reported as anatomically related groups. These groupings can be found in Table S2. Exact 95% confidence intervals were calculated for each proportion.

Table 2.1. Definition of terms in three types of self-report accuracy calculations.

Term	Any-cancer overall diagnosis	Common and rare cancer type diagnosis (Y=common or rare cancer)[a]	Site-specific cancer diagnosis (X=SEER cancer category)[b]
TP	Self-reported cancer in an ATP follow-up survey and had a diagnosis of cancer in the ACR	Self-reported cancer type Y in an ATP follow-up survey and had a diagnosis of cancer Y in the ACR	Self-reported cancer type X in an ATP follow-up survey and had a diagnosis of cancer X in the ACR
TN	Did not report cancer in an ATP follow-up survey and did not have a diagnosis of cancer in the ACR	Did not report cancer type Y in an ATP follow-up survey and did not have a record of cancer Y in the ACR	Did not report cancer type X in an ATP follow-up survey and did not have a record of cancer X in the ACR
FP	Self-reported cancer in an ATP follow-up survey, but did not have a diagnosis of cancer in the ACR	Self-reported cancer type Y in an ATP follow-up survey but did not have a record of cancer Y in the ACR	Self-reported cancer type X in an ATP follow-up survey but did not have a record of cancer X in the ACR
FN	Did not report a cancer diagnosis in ATP follow-up, but had a diagnosis of cancer in the ACR	Did not report cancer type Y in ATP follow-up, but had a record of cancer type Y in the ACR	Did not report cancer type X in ATP follow-up, but had a record of cancer X in the ACR

TP=True Positive, TN=True Negative, FP=False Positive, FN=False Negative

[a] Common or rare cancer types defined as those cancer types with an incidence rate of <15/100,000/year, as per recent analysis by Walker et al. (2019).

[b] SEER=Surveillance, Epidemiology, and End Results. Description of categories in Table S2.

Correct Site and Year Proportions

The number of true positives from the overall calculation, common cancer calculation, and rare cancer calculation that also reported the correct cancer site was determined. Of those who correctly reported cancer site, the proportion that also correctly reported +/- one year from diagnosis year in the ACR and the proportion of those who reported the exact year of diagnosis in the ACR were calculated. All analyses were done with STATA IC v15.

2.2.3. Determining Factors Associated with Incorrect Cancer Reporting

Data Preparation

Participants that did not consent to data linkage (n=360) and that had a diagnosis of cancer before baseline (n=118), as described above, were excluded from logistic regression analyses. Only ACR diagnoses within an individual's follow-up (ACR age of diagnosis less than age at most recent survey receipt) were included. Cancer diagnosis exclusion and categorization were the same as described above.

Outcomes

Three outcomes were analyzed:

(1) Participants who incorrectly reported that they have cancer (FP=1) compared to those who correctly reported that they have cancer (TP=0),

(2) Participants who incorrectly did not report cancer (FN=1) compared to those who correctly reported that they had cancer (TP=0), and

(3) Of those who correctly reported that they have cancer (TP's), participants that reported cancer site incorrectly (1) compared to those who reported site correctly (0).

Covariates

Factors that have been previously associated with self-report accuracy (Table S1) and were available in the ATP were included as covariates. Covariates included age at cancer report (for Outcomes 1 and 3) or age at last follow up (for Outcome 2) , sex (Male (reference) or Female), education level at baseline (High school or less (reference), Some college or university <4 years, University degree ≥4 years) , smoking status (Never (reference), Former, Current), and family history of cancer in a parent or sibling (No (reference) or Yes), and place of birth (In Canada (reference) or Outside of Canada). ACR diagnosis type (Common (reference) or Rare) was also a covariate for outcomes 2 and 3, in which each person in the sample had a true cancer diagnosis in the ACR. Smoking status and family history were recorded at baseline but were changed if updated in a follow-up survey at the time of or before a report of cancer. Place of birth, for which data was not collected at baseline, was investigated in those who took S08, UHLQ, or CORE, and therefore reported this information. There were 39 TP's, 6 FP's, and 1 FN who were missing place of birth information.

Logistic Regression Analysis

Logistic regression analysis estimated odds ratios (OR's) and 95% confidence intervals for covariates associated with each self-report outcome. Covariates in univariate analyses with a likelihood ratio test p<0.2 were included in multivariate analysis. All analyses were done with STATA IC v15.

2.3 Results

2.3.1 Sensitivity and PPV of Self-Reported Cancer Diagnoses

In the 30,843 ATP participants who consented to registry linkage, 810 primary cancer diagnoses, as defined in this report, occurred during participant follow-up time and were "self-reportable", of which 724 were common cancer and 86 were rare cancer. There were 959 self-reports of first, non–melanoma cancer, of which 746 were TP's, for an overall sensitivity of

92.1% (95% CI: 90.0, 93.9) and PPV of 77.8% (95% CI: 75.0, 80.4) (Table 2.2). Reporting a common cancer had a sensitivity and PPV of 89.6% (95% CI: 87.2, 91.8) and 84.5% (95% CI: 81.7, 87.0), respectively. Reporting a rare cancer had a lower sensitivity and PPV of 62.8% (95% CI: 51.7, 73.0) and 35.8% (95% CI: 28.1, 44.0), respectively. Since specification for reporting only non-melanoma skin cancer changed across surveys and skin cancer had a fairly low sensitivity and PPV compared to other common cancers, the sensitivity and PPV of common cancer was also calculated without skin cancer. Sensitivity remained relatively unchanged (90.1% (95% CI: 87.6, 92.3)), but PPV increased to 88.6% (95% CI: 86.0, 90.9). In addition, rare cancers were considered without cervix cancer, as cervix cancer had a very low PPV and behavior 2 cervix cancers were not considered a cancer diagnosis in the ACR in this analysis. While sensitivity remained relatively unchanged, the PPV of the rare cancer group reporting increased to 54.2% (95% CI: 43.7, 64.4) when cervix cancer was excluded.

Of the 746 overall cancer status TP's, 90.5% reported the correct site, 88.1% reported correct site and year of diagnosis within one year, and 68.2% reported the correct site and correct year of diagnosis (Table 2.3). Common cancers were reported more accurately overall, with 97.5% of 649 TP's having the correct site, 95.1% having the correct site and year within one year, and 73.3% having the correct site and year. These percentages remained relatively unchanged when skin cancer was excluded. Rare cancers were reported less accurately. Of 54 TP's, 77.8% had the correct site, 74.1% had the correct site and year within one year, and 61.1% had the correct site and year. Removing cervix cancer had little effect on these percentages.

Table 2.2 includes anatomical group and/or site-specific sensitivity and PPV. Site categories have been collapsed into anatomically related groups (bolded). Sites within these groups that could be reported individually are shown under their respective anatomical group (un-bolded). Male reproductive cancers had the highest sensitivity of all sites and groups reported (96.9% (95% CI: 92.8, 99.0)); most cancers in this category were prostate cancer, which had the highest site-specific sensitivity (96.8% (95% CI: 92.6, 98.9)). Breast cancer had the next highest sensitivity, at 95.6% (95% CI: 91.8, 98.0), followed by digestive/hepatic cancers (89.5% (95% CI: 82.3, 94.4)), and lymphatic cancers (89.3% (95% CI: 71.8, 97.7)). Breast cancer had the highest PPV (93.8% (95% CI: 89.5, 96.6)), followed by male reproductive cancers (91.1% (95% CI: 85.8, 94.9)), which was, again, largely due to prostate cancer (PPV of 90.9% (95% CI: 85.4, 94.8). Melanoma skin cancer had a sensitivity of 79.3% (95% CI: 66.6, 88.8) and a PPV of 50.5% (95% CI: 39.9, 61.2), though since the specification to include only melanoma skin cancer changed across surveys, this may not reflect the true self-reporting accuracy of melanoma skin cancer. Cervix cancer had the lowest PPV at 3.6% (95% CI: 0.4, 12.5). Though there were few (<10) ACR diagnoses of cervix cancer (behavior 3 only), there were many self-reports. There were over 300 women that had cervix cancer (behavior 2 or 3) before baseline; this may be contributing to these low results. Other than cervix cancer, ovarian cancer had the lowest PPV (40.0% (95% CI: 16.3, 67.7)), followed by CNS/Eye cancers (44.4% (95%CI: 18.7, 81.3)). Rectal cancer had the lowest sensitivity (42.1% (95% CI: 20.3, 66.5)).

Table 2.2. Sensitivity and positive predictive value (PPV) of self-reported cancer diagnoses

Cancer Type[a]	*Including* those with cancer before baseline[b]			*Excluding* those with cancer before baseline[b]		
	# ACR[c]	Sensitivity (95% CI)[d]	PPV (95% CI)[e]	# ACR[c]	Sensitivity (95% CI)[d]	PPV (95% CI)[e]
OVERALL	**810**	**92.1 (90.0, 93.9)**	**77.8 (75.0, 80.4)**	**789**	**93.9 (92.0, 95.5)**	**78.9 (76.2, 81.5)**
Common	**724**	**89.6 (87.2, 91.8)**	**84.5 (81.7, 87.0)**	**707**	**91.2 (88.9, 93.2)**	**85.9 (83.2, 88.3)**
Common (no skin)	666	90.1 (87.6, 92.3)	88.6 (86.0, 90.9)	649	91.8 (89.5, 93.8)	89.8 (87.2, 92.0)
Rare	**86**	**62.8 (51.7, 73.0)**	**35.8 (28.1, 44.0)**	**82**	**64.6 (53.3, 74.9)**	**35.8 (28.1, 44.1)**
Rare (no cervix)	84	61.9 (50.7, 72.3)	54.2 (43.7, 64.4)	80	63.8 (52.2, 74.2)	54.8 (44.2, 65.2)
Oral/Respiratory	**41**	**63.4 (46.9, 77.9)**	**60.5 (44.4, 75.0)**	**36**	**69.4 (51.9, 83.7)**	**61.0 (44.5, 75.8)**
Lung & Bronchus	25	76.0 (54.9, 90.6)	76.0 (54.9, 90.6)	23	82.6 (61.2, 95.0)	76.0 (54.9, 90.6)
Digestive/Hepatic	**114**	**89.5 (82.3, 94.4)**	**85.7 (78.1, 91.5)**	**110**	**91.8 (85.0, 96.2)**	**87.1 (79.6, 92.6)**
Large Intestine	68	85.3 (74.6, 92.7)	79.5 (68.4, 88.0)	65	87.7 (77.2, 94.5)	81.4 (70.3, 89.7)
Rectum	19	42.1 (20.3, 66.5)	80.0 (44.4, 97.5)	19	42.1 (20.3, 66.5)	80.0 (44.4, 97.5)
Blood/Hematopoietic	**52**	**75.0 (61.1, 86.0)**	**86.7 (73.2, 94.9)**	**48**	**79.2 (65.0, 89.5)**	**86.4 (72.6, 94.8)**
Skin (melanoma)	**58**	**79.3 (66.6, 88.8)**	**50.5 (39.9, 61.2)**	**58**	**79.3 (66.6, 88.8)**	**52.9 (41.9, 63.7)**
Breast	**204**	**95.6 (91.8, 98.0)**	**93.8 (89.5, 96.6)**	**199**	**97.5 (94.2, 99.2)**	**94.2 (90.0, 97.0)**
Female Reproductive	**67**	**88.1 (77.8, 94.7)**	**46.8 (37.9, 55.9)**	**66**	**89.4 (79.4, 95.6)**	**47.6 (38.5, 56.7)**
Cervix Uteri	-	-	3.6 (0.4, 12.5)	-	-	3.6 (0.4, 12.5)
Uterus[f]	49	85.7 (72.8, 94.1)	79.2 (65.9, 89.2)	48	87.5 (74.8, 95.3)	82.4 (69.1, 91.6)
Ovary	-	-	40.0 (16.3, 67.7)	-	-	40.0 (16.3, 67.7)
Male Reproductive	**159**	**96.9 (92.8, 99.0)**	**91.1 (85.8, 94.9)**	**157**	**97.5 (93.6, 99.3)**	**93.3 (88.3, 96.6)**
Prostate Gland	154	96.8 (92.6, 98.9)	90.9 (85.4, 94.8)	152	97.4 (93.4, 99.3)	93.1 (88.0, 96.5)
Urinary	**51**	**82.4 (69.1, 91.6)**	**87.5 (74.8, 95.3)**	**51**	**82.4 (69.1, 91.6)**	**87.5 (74.8, 95.3)**
Kidney	16	87.5 (61.7, 98.4)	73.7 (48.8, 90.9)	16	87.5 (61.7, 98.4)	73.7 (48.8, 90.9)
Urinary Bladder	31	77.4 (58.9, 90.4)	85.7 (67.3, 96.0)	31	77.4 (58.9, 90.4)	85.7 (67.3, 96.0)
CNS/Eye	**-**	**-**	**44.4 (18.7, 81.3)**	**-**	**-**	**-**
Endocrine	**20**	**70.0 (45.7, 88.1)**	**87.5 (61.7, 98.4)**	**20**	**70.0 (45.7, 88.1)**	**87.5 (61.7, 98.4)**
Thyroid	20	70.0 (45.7, 88.1)	87.5 (61.7, 98.4)	20	70.0 (45.7, 88.1)	87.5 (61.7, 98.4)
Lymphatic	**28**	**89.3 (71.8, 97.7)**	**69.4 (51.9, 83.7)**	**28**	**89.3 (71.8, 97.7)**	**69.4 (51.9, 83.7)**
Other[g]	**-**	**-**	**-**	**-**	**-**	**-**

PPV=Positive predictive value, ACR=Alberta Cancer Registry, CI=Confidence interval, CNS=Central nervous system

[a] Bolded groups generated by combining appropriate SEER 2018 categories. Un-bolded types are specific groups within the bolded group above. See Table S2 for groupings. Only SEER 2018 cancer types with >10 ACR diagnoses and/or >10 self-reported diagnoses were included in the table.

[b] A participant had a diagnosis before baseline if their age of first cancer diagnosis in the ACR was before their age at baseline (n=118).

[c] Number of diagnoses in the ACR. A "-" indicates there was <10. Common and rare diagnoses add up to overall. Bolded cancer site types do not add up to overall as the "Other and "CNS/Eye" groups are not included.

[d] Sensitivities for groups with 10 or more ACR diagnoses are reported. A "-" indicates there was <10.

[e] PPV's for groups with 10 or more self-reported diagnoses are reported. A "-" indicates there was <10.

[f] Combines SEER 2018 categories of corpus uteri and uterus, NOS (not otherwise specified), ie. does not differentiate between the two

[g] Includes SEER 2018 categories of unknown, ill-defined, bones & joints, connective & soft tissue, retroperitoneum & peritoneum.

Table 2.3. Self-reported cancer site and year of diagnosis accuracy among ATP participants who correctly report overall, common, and rare cancer status.

	***Including* those with cancer before baseline**[a]				***Excluding* those with cancer before baseline**[a]			
		% TP's that also have correct:				**% TP's that also have correct:**		
	# TP's[b]	site only	site and year +/-1	site and year	**# TP's**[b]	site only	site and year +/-1	site and year
Common	649	97.5	95.1	73.3	645	97.5	95.0	73.5
Common (no skin)	600	97.8	95.5	73.7	596	97.8	95.5	73.8
Rare	54	77.8	74.1	61.1	53	77.4	73.6	60.4
Rare (no cervix)	52	76.9	73.1	61.5	51	76.5	72.5	60.8
Overall[c]	746	90.5	88.1	68.2	741	90.4	88.0	68.3

ATP=Alberta's Tomorrow Project, TP=True positive

[a] A participant had a diagnosis before baseline if their age of first cancer diagnosis in the ACR was before their age at baseline (n=118).

[b] Those who reported a common, rare, or any cancer (overall) and had a common, rare, or any cancer diagnosis in the ACR within their follow up time.

[c] Overall TP does not equal common TP plus rare TP. An overall TP reported a cancer and had cancer in the ACR, regardless of type. A participant with a common or rare cancer in the ACR had to report a common or rare cancer, respectively, in order to be a common TP or a rare TP.

2.3.2 Sensitivity and PPV Excluding participants with a Diagnosis of Cancer Before Baseline

There were 118 participants who had a cancer in the ACR that occurred before baseline. Upon further investigation of the reporting patterns of these participants, it was clear that while some participants reported an incident cancer in a follow up survey, others reported a cancer that had occurred before baseline. As the first analysis followed the intention of the survey questions and assumed participants were reporting incident cancers, the ambiguity of which cancer they are reporting may affect those results. Therefore, calculations were done excluding these participants to evaluate the accuracy of self-reports in the population that the cohort is intended to include: those with no cancer or history of cancer prior to enrolment.

In the 30,725 ATP participants who consented to registry linkage and had no history of cancer at baseline, there were 789 primary cancer diagnoses in the ACR that occurred within participant follow-up time and 939 self-reports of cancer. The overall sensitivity for self-report of this group was 93.9% (95% CI: 92.0, 95.5) and the PPV was 78.9 (95% CI: 76.2, 81.5), both slightly higher than the first analysis including those with cancer history (Table 2.2). Reporting a common cancer had a sensitivity of 91.2% (95% CI: 88.9, 93.2) and PPV of 85.9% (95% CI: 83.2, 88.3), while reporting a rare cancer had a sensitivity of 64.6% (95% CI: 53.3, 74.9) and PPV of 35.8% (95% CI: 28.1, 44.1).

Of those who correctly reported that they had cancer, 90.4% also reported the correct site, 88.0% reported correct site and within one year of diagnosis, and 68.3% reported correct site and year (Table 2.3). Common cancers were reported more accurately than rare cancers in regards to site and year of diagnosis.

Site-specific sensitivities and PPV's slightly improved or remained unchanged when excluding those with cancer before baseline, with one exception; the PPV of blood/hematopoietic cancers decreased slightly (Table 2.2). The sites with the highest sensitivities were male reproductive cancers (97.5% (95% CI: 93.6, 99.3)) and breast cancer (97.5% (95% CI: 94.2, 99.2)), followed by digestive/hepatic cancers and lymphatic cancer. Breast and male reproductive cancers also had the highest PPV's: 94.2% (95% CI: 90.0, 97.0) and 93.3% (95% CI: 88.3, 96.6), respectively. Cervical cancer and ovarian cancer remained the sites with the lowest PPV. Female reproductive cancers as a group had a PPV of 47.6% (95% CI: 38.5, 56.7). Rectal cancer remained the site with the lowest sensitivity.

2.3.3 Factors Associated with Incorrect Cancer Status or Site Reporting

There were 741 TP's, 198 FP's and 48 FN's for overall self-reporting of cancer in the second analysis excluding those with cancer before baseline. Of the 741 TP's, 71 reported an incorrect cancer site. Predictors of false positive compared to true positive reporting, false negative compared to true positive reporting, and incorrect site compared to correct site reporting are presented in Table 2.4. Older participants were less likely to incorrectly report cancer (false positive) than younger participants (p<0.001), but more likely to incorrectly report cancer site (p<0.001). Participants older than 70 years of age at the time of report had over 4 times the odds of incorrectly reporting cancer site than those <50 years (p=0.017), adjusting for smoking status and common or rare cancer type. Smoking status was associated with incorrectly self-reporting cancer. Former smokers had 59% higher odds of incorrectly report cancer (false positive) than participants who never smoked, adjusted for age (p=0.013). Current smokers had twice the odds of incorrectly reporting cancer (p=0.002). Former smokers had 92% higher odds of neglecting to report cancer compared participants who never smoked (p=0.053). Though participants who were current smokers had almost 50% higher odds of incorrectly reporting cancer site compared to those who never smoked, smoking was not statistically significantly associated with incorrect site reporting (p=0.3111). Finally, participants with a rare cancer had almost 14 times the odds of incorrectly reporting cancer site compared to those with a common cancer, adjusting for age and smoking status (p<0.001). Education, family history, and sex were not significant predictors of false positive, false negative or incorrect site reporting (p>0.05).

In participants that had place of birth information available, those born outside of Canada were somewhat less likely to incorrectly report cancer (false positive) or not report cancer (false negative) compared to those born in Canada (OR [95% CI]: 0.75 [0.46, 1.23] and OR [95% CI]: 0.85 [0.35, 2.06], respectively). Participants born outside of Canada were somewhat more likely to report cancer site incorrectly compared to participants born in Canada (OR [95% CI]: 1.36 [0.70, 2.64]). However, none of the associations between place of birth and incorrect self-reporting were statistically significant.

Table 2.4. Factors associated with false positive, false negative and incorrect site self-reporting.

	Report cancer incorrectly vs correctly (FP vs TP)		**Not report vs report correctly (FN vs TP)**		**Report site incorrectly vs correctly (TP incorrect vs TP correct)**	
	N=939		*N=789*		*N=741*	
Variable	**Adjusted OR[a] (95% CI)**	**p-value**	**OR[b] (95% CI)**	**p-value**	**Adjusted OR[c] (95% CI)**	**p-value**
Age at report or last follow-up[d]						
<50	1				1	
50 to <60	0.54 (0.34, 0.85)	0.008			2.52 (0.80, 7.99)	0.116
60 to <70	0.29 (0.18, 0.45)	<0.001			1.23 (0.38, 3.96)	0.732
>=70	0.35 (0.20, 0.59)	<0.001	NS		4.19 (1.29, 13.6)	0.017
Smoking						
Never	1		1		1	
Former	1.59 (1.10, 2.29)	0.013	1.92 (0.99, 3.72)	0.053	0.80 (0.43, 1.48)	0.478
Current	2.05 (1.29, 3.24)	0.002	1.32 (0.49, 3.51)	0.585	1.48 (0.68, 3.20)	0.325
ACR Diagnosis Type[e]						
Common					1	
Rare	N/A		NS		13.7 (7.60, 24.5)	<0.001
Sex						
Education						
Family History	NS		NS		NS	

FP=False positive, TP=True positive, FN=False negative, OR=Odds Ratio, NS=Not Significant, N/A=Not applicable, ACR=Alberta Cancer Registry

[a] Adjusted for age and smoking status (both significant at $p<0.2$ in univariate likelihood ratio test).

[b] Only smoking status significant at $p<0.2$ in univariate likelihood ratio test.

[c] Adjusted for age, smoking status, and ACR diagnosis type (all significant at $p<0.2$ in univariate likelihood ratio test).

[d] Age at report for FP vs TP and TP incorrect site vs TP correct site outcomes. Age at last follow-up for FN vs TP outcome.

[e] Common and rare cancer types as determined by Walker et al. (2019), using the US definition of rare: an incidence rate of <15 cases/100,000/year. Not applicable to FP vs TP since FP's do not have a true diagnosis in the ACR.

2.4 Discussion

This analysis explored the accuracy of self-reported cancer diagnosis in a Canadian cohort. We evaluated whether self-reported diagnoses are valid as an outcome in etiologic research, particularly for rare cancers. The sensitivity and PPV for reporting overall cancer status, without considering site, was similar to reports from the US and Australia[11,14,16]. PPV was lower than sensitivity; self-report was more likely to misclassify someone as having cancer when they did not than to misclassify someone who had cancer as not having cancer. Those who correctly report cancer status are also likely to correctly report cancer site, however, year of diagnosis is less accurately reported. This was also demonstrated in a US cohort by Bergmann et al., which found that 84% of overall true positives also reported the correct site and correct year of diagnosis within one year[14]. Therefore, self-reported overall cancer status is fairly accurate. However, if the timing of diagnosis is important in analysis, then self-report is less accurate.

Common cancers were reported more accurately overall than rare cancers and, as expected, made up a majority of cancer cases in the cohort. Breast and prostate cancer, the two most common cancers in this cohort, had the highest sensitivity and PPV. These two cancers often have high accuracy across self-report literature[11-14]. Using these self-reports as an outcome would result in relatively few misclassifications and have sufficient case numbers for analysis. Rare cancers, however, had a lower sensitivity than common cancers and were less likely to be captured by self-report. Interestingly, the low sensitivity of rare cancers was not primarily because rare cancer patients were neglecting to report cancer. Rather, they were incorrectly reporting a common site or their response was unclear and could not be categorized. Over 70% of those having but not reporting a rare cancer (false negatives) reported having cancer but reported a common or unclear site. This suggests that rare cancer diagnoses are not well understood by patients. A logistic regression analysis supported this hypothesis; for those that correctly reported overall cancer status, participants who reported cancer site incorrectly were far more likely to have a rare cancer than participants who reported site correctly. A possible explanation for this trend may be that rare cancers often have poor diagnostic precision[4,5]. Previous reports have suggested that ambiguous diagnostic procedures or results are more likely to result in an incorrect or absent self-report[12,13]. Due to the low sensitivity and low PPV of rare cancer sites reported here, it is unlikely that self-reports of rare cancer are valid as an outcome. Misclassification as diseased or disease free is likely.

Cancer registry linkage would not only provide an valid diagnosis, but also serve as a passive follow-up to capture cases more completely. There were 3,187 total cancer diagnoses that developed in this cohort during follow-up in the ACR, but only 810 were “self-reportable” (within active follow-up). In order for rare cancer research in cohort studies to be feasible, all cases that develop must be included to achieve enough sample size. Though participants who developed both common and rare cancers were lost to follow-up, those who developed a rare cancer were more likely to be lost. Rare cancers accounted for approximately 16% of total cases that developed in the cohort, but only 10.6% of “self-reportable” diagnoses. One possible explanation is that participants who develop a rare cancer may be less likely to survive long enough after diagnosis to report. Upon further exploration of the “age at death” information in the ACR, we found that 40% of participants with a rare cancer have died, while 20% of participants with a common cancer have died. For those who died, median time from diagnosis to death was 3.7 years for rare, and over 6 years for common cancers. However, more research and exploration are needed to uncover the reasons behind the differential loss to follow-up for rare cancers.

Finally, relying on self-reported diagnosis of cancer at baseline assumes that participants will correctly state that they have no history of cancer at study start. If participants are not cancer-free at baseline but still included in an etiologic study, results from this study may be biased. Controls that are thought to be cancer-free may actually have cancer, biasing towards the null. It is unclear why some participants did not report cancer history at baseline. Though there are likely diverse reasons, participants were aware that being cancer free at baseline was an eligibility requirement of enrolment[24]. Some may have been more inclined to not disclose their previous cancer so that they could enroll in the study.

This analysis contributes to the limited information on the accuracy of self-reported cancer diagnosis in Canada. While most reports focus on common cancer sites, this analysis looked at many different cancer sites or groups and compared the reporting of common and rare cancers. Using the ACR as a gold standard strengthens this analysis due to demonstrated completeness and accuracy in reporting[25]. There are, however, several limitations in this analysis. Firstly, a lack of "self-reportable" diagnoses did not allow for the separate reporting of many individual sites. General anatomical sites were still reported but provide less information. Secondly, some participants may be misclassified as a FN if they experience a lag in learning about their diagnosis. They may have been diagnosed at the time of the survey, but not yet informed of their specific diagnosis. Thirdly, including only behavior 2 (in situ/noninvasive) and 3 (malignant) diagnoses (excluding non-melanoma skin cancer and behavior 2 cervix and prostate cancer) may have impacted the sensitivity and/or PPV of several sites. In particular, skin and cervical cancer may be affected. Self-reports of skin cancer were only included if "melanoma" was specified. This likely underestimated true self-reported melanoma cases. Cervical cancer had a very low PPV; since most cervical cancers in the ACR were behavior 2 and excluded, they were over self-reported. However, somewhat surprisingly, including behavior 2 cervix ACR diagnoses still resulted in a very low PPV (and sensitivity), as a majority of behavior 2 cervical cancers were not reported (exploratory results not shown). Finally, participants who enroll and are followed-up in this cohort are likely healthier or more health conscious than the general population[24]. This could introduce a "healthy volunteer effect"[31]. The accuracy of self-reports in this cohort may overestimate the true accuracy of the general Canadian population. However, these results are likely to be generalizable to other large observational cohorts with similar aims in health research.

In conclusion, while self-reported diagnosis is valid for some common cancer types, other cancer types, particularly rare cancers, require registry linkage to be captured completely and accurately. In order to minimize bias and loss of follow-up in the use of cohort data, such as ATP and CPTP, for rare cancer research, linkage to cancer registry is necessary. Removing barriers that prevent cross-provincial data sharing in Canada would allow researchers to make use of the valuable information on rare cancers that national cohorts and registries may offer.

2.5 References

1. Walker EV, Maplethorpe E, Davis FG. Common and rare cancer incidence rates in the Canadian population: 2009-2013; 2019. In preparation for submission at the time of this thesis.

2. Canadian Cancer Society's Advisory Committee on Cancer Statistics. Canadian Cancer Statistics 2017. cancer.ca/Canadian-CancerStatistics-2017-EN.pdf. Updated 2017. Accessed February, 2018.

3. Boyd N, Dancey JE, Gilks CB, Huntsman DG. Rare cancers: A sea of opportunity. *Lancet Oncology*. 2016;17(2):61. doi: 10.1016/S1470-2045(15)00386-1.

4. von der Schulenburg JM, Pauer F. Reviews: Rare cancers—Rarity as a cost and value argument. *Journal of Cancer Policy*. 2017;11:54-59. doi: 10.1016/j.jcpo.2016.09.004.

5. Mathoulin-Pélissier S, Pritchard-Jones K. Evidence-based data and rare cancers: The need for a new methodological approach in research and investigation. *European Journal of Surgical Oncology*. 2018;45(1):22-30. doi: 10.1016/j.ejso.2018.02.015.

6. Armstrong-Wells J, Goldenberg NA. Institution-based prospective inception cohort studies in neonatal rare disease research. *Seminars in Fetal and Neonatal Medicine*. 2011;16(6):355-358. doi: 10.1016/j.siny.2011.07.004.

7. Gatta G, van der Zwan, Jan Maarten, Casali PG, et al. Rare cancers are not so rare: The rare cancer burden in Europe. *Eur J Cancer*. 2011;47(17):2493-2511. doi: //dx.doi.org/10.1016/j.ejca.2011.08.008.

8. Brown C. Barriers to accessing data are bad medicine. *CMAJ*. 2014;186(16):1203. doi: 10.1503/cmaj.109-4894.

9. Streiner DL, Norman GR, Cairney J. *Health measurement scales.* 5th ed. Oxford: Oxford Univ. Press; 2015.

10. Navarro C, Chirlaque MD, Tormo MJ, et al. Validity of self reported diagnoses of cancer in a major Spanish prospective cohort study. *Journal Epidemiology Community Health*. 2006;60:593-599.

11. Stavrou E, Vajdic CM, Pearson S, Loxton D. The validity of self-reported cancer diagnoses and factors associated with accurate reporting in a cohort of older Australian women. *Cancer Epidemiology*. 2011;35(6):80. doi: 10.1016/j.canep.2011.02.005.

12. Parikh-Patel A, Allen M, Wright WE. Validation of self-reported cancers in the California Teachers Study. *Am J Epidemiol*. 2003;157(6):539-545.

13. Loh V, Harding J, Koshkina V, Barr E, Shaw J, Magliano D. The validity of self-reported cancer in an Australian population study. *Australia NZ J Public Health*. 2014;38:35-38. doi: 10.1111/1753-6405.12164.

14. Bergmann MM, Calle EE, Mervis CA, Miracle-McMahill HL, Thun MJ, Heath CW. Validity of self-reported cancers in a prospective cohort study in comparison with data from state cancer registries. *Am J Epidemiol*. 1998;147(6):556-562.

15. Li J, Cone JE, Alt AK, et al. Performance of self-report to establish cancer diagnoses in disaster responders and survivors, World Trade Center Health Registry, New York, 2001-2007. *Public Health Rep*. 2016;131(3):420-429.

16. Zeig-Owens R, Kablanian A, Webber MP, et al. Agreement between self-reported and confirmed cancer diagnoses in New York City firefighters and EMS workers, 2001-2011. *Public Health Rep*. 2016;131(1):153-159.

17. Cho LY, Kim C, Li L, et al. Validation of self-reported cancer incidence at follow-up in a prospective cohort study. *Ann Epidemiol*. 2009;19(9):644-646. doi: //dx.doi.org/10.1016/j.annepidem.2009.04.011.

18. Inoue M, Sawada N, Shimazu T, et al. Validity of self-reported cancer among a Japanese population: Recent results from a population-based prospective study in Japan (JPHC study). *Cancer Epidemiology*. 2011;35(3):250-253. doi: 10.1016/j.canep.2010.12.002.

19. Yoshinaga A, Sasaki S, Tsugane S. Sensitivity of self-reports of cancer in a population-based prospective study: JPHC study cohort I. *Journal of Clinical Epidemiology*. 2001;54(7):741-746.

20. Canadian Partnership for Tomorrow Project. About. https://www.partnershipfortomorrow.ca/about/. Accessed January, 2019.

21. Doiron D, Raina P, Fortier I. Linking Canadian population health data: Maximizing the potential of cohort and administrative data. *Can J Public Health*. 2013;104(3):261.

22. Dummer TJ, Awadalla P, Boileau C, et al. The Canadian Partnership for Tomorrow Project: A pan-Canadian platform for research on chronic disease prevention. *CMAJ*. 2018;190(23):717. doi: 10.1503/cmaj.170292.

23. Alberta's Tomorrow Project. All about Alberta's Tomorrow Project; https://myatp.ca/about-atp. Updated 2019. Accessed January, 2019.

24. Robson PJ, Solbak NM, Haig TR, et al. Design, methods and demographics from phase I of Alberta's Tomorrow Project cohort: A prospective cohort profile. *Canadian Medical Association Journal Open*. 2016;4(3):527. https://www.clinicalkey.es/playcontent/1-s2.0-S2291002616301345. doi: 10.9778/cmajo.20160005.

25. Cancer Control Alberta. Surveillance and reporting: The 2019 report on cancer statistics in Alberta. *Alberta Health Services*. 2019.

26. North American Association of Central Cancer Registries. Certification criteria. https://www.naaccr.org/certification-criteria/. Updated 2018. Accessed July, 2019.

27. Cancer Statistics Branch, Division of Cancer Control and Population Sciences Surveillance, Epidemiology and End Results Program. Conversion of neoplasms by topography and morphology from the International Classification of Diseases for Oncology, second edition to International Classification of Diseases for Oncology, third edition; 2001.

28. Statistics Canada. Canadian Cancer Registry (CCR). http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3207. Updated 2019. Accessed April, 2019.

29. National Cancer Institute. ICD-O-3 SEER site/histology validation. https://seer.cancer.gov/icd-o-3/sitetype.icdo3.d20150918.pdf. Updated 2015.

30. Greenlee RT, Goodman MT, Lynch CF, Platz CE, Havener LA, Howe HL. The occurrence of rare cancers in U.S. adults, 1995-2004. *Public Health Reports*. 2010;125(1):28-43.

31. Lindsted KD, Fraser GE, Steinkohl M, Beeson WL. Healthy volunteer effect in a cohort study: Temporal resolution in the Adventist Health Study. *Journal of Clinical Epidemiology*. 1996;49(7):783-790. https://www.sciencedirect.com/science/article/pii/0895435696000091. doi: 10.1016/0895-4356(96)00009-1.

CHAPTER 3: Study of Rare Cancer in Canada using Emerging Cohorts: A Pilot Etiologic Study of Pancreatic Cancer in the Alberta Tomorrow Project

3.1. Introduction

Though affecting few people individually, rare cancers account for nearly 22% of cancer cases in Canada collectively[1] and contribute disproportionately to cancer-related morbidity and mortality[2]. Despite the apparent burden that rare cancer has on society and on patients, knowledge on the causes, etiology, diagnosis, and treatment options of rare cancers is lacking. However, significant challenges prevent progress in the study of rare cancer, as traditional clinical research designs are harder to execute[3-5]. In an effort to overcome some of these challenges, large observational databases are implicated as an opportunity to study risk factors and natural history of rare disease with a sufficiently large sample size[4,6].

Traditional cohort designs are inefficient and costly to execute in the study of rare disease. Recruiting enough participants to provide a sufficient amount of cases to develop is rarely feasible. Often, traditional case-control designs are used to mitigate these issues. However, case-control designs cannot distinguish the temporality of exposure-disease association due to the ascertainment of exposure status after disease status and susceptibility to bias. Using existing cohorts for health-related research, however, may be a feasible option to study rare cancers, and less susceptible to systemic biases. The nested case-control design provides the advantages of a case-control design in the study of rare diseases while maintaining the temporal advantage of the cohort design. All participants are disease free at study start and all cases that develop in the cohort can be utilized. Incidence density sampling, where a control or group of controls is selected from those at risk of the disease at the time of the case, is used, allowing for an unbiased estimate of the rate ratio[7]. Therefore, a nested case-control design may be a viable option in the study of rare cancer within an existing cohort.

3.1.1 Alberta's Tomorrow Project

Alberta's Tomorrow Project (ATP) offers the opportunity to test whether this rare cancer research in emerging cohort studies is feasible. The ATP is a cohort study that aims to support advances in knowledge of cancer and chronic disease etiology to inform more effective risk reduction strategies[8]. In order to evaluate whether this dataset and its linkage to the Alberta Cancer Registry (ACR) is a potential tool for etiologic research on rare cancers, we conducted a pilot etiologic study on pancreatic cancer. A recent analysis by Walker et al. (2019)[9], not yet published, has implicated pancreatic cancer as a "rare" cancer in Canada, defined by the US definition of a rare cancer: an incidence rate of <15/100 000/year[10]. As pancreatic cancer incidence lies just below this cutoff, it may have sufficient case numbers to study in a smaller population.

3.1.2 Pancreatic Cancer in Canada

Though pancreatic cancer only accounts for 2.7% of new cancer cases in males and 2.6% of new cancer cases in females, it is the fourth leading cause of cancer death in both sexes[2]. The Canadian Cancer Society expects pancreatic cancer to surpass breast cancer as the third leading cause of cancer death in Canada in the near future[2]. Pancreatic cancer has the poorest 5-year survival of all reported cancers in Canada at 7%, with only 50% of patients surviving beyond 4 months[2,11]. Though several histological subtypes of pancreatic cancer exist, an overwhelming majority of cases in Canada are adenocarcinomas[9,11]. Little change has occurred in the overall age-standardized incidence rate of pancreatic cancer over time, and though modest improvements have occurred in mortality rates, mortality remains high[2,11].

Largely contributing to the lack of progress in improving pancreatic cancer prognosis and survival over time is its tendency to be difficult to detect, exhibit few early symptoms, and have a late stage diagnosis. Over 60% of cases are diagnosed at Stage III or IV and commonly metastasize to the liver, lungs, or brain[2]. The pancreas lies deep in the abdomen, limiting the effectiveness of curative treatments such as surgery, and tumors are relatively unresponsive to chemotherapy or radiation. As screening, detection, and treatment options are limited, primary prevention may play an important role in reducing the burden of pancreatic cancer[2,12]. Since the incidence of pancreatic cancer is generally higher in developed countries, exploring associations with environmental and modifiable lifestyle factors is an opportunity to better understand the etiology of pancreatic cancer and improve its prevention[11,12]. Information for several established risk factors is available in the ATP, which makes pancreatic cancer a good candidate for a pilot nested case-control study. Results relating to these factors can be compared to previous literature to evaluate the feasibility of the ATP in the study of rare cancer etiology.

Like other rare cancers, research on the etiology of pancreatic cancer is lacking[2]. With the exception of a few established risk factors, other lifestyle, genetic, and dietary factors that are studied lack strong evidence and consistency across studies[13,14]. For the purposes of this feasibility study, several well-established risk factors and other dietary factors frequent in the literature were explored based on the data that was available in the ATP. A search of the literature that pertains primarily to these factors of interest is summarized in Supplementary Table 3.

3.1.3 Risk Factors of Pancreatic Cancer: What is Established?

As with most cancers, pancreatic cancer risk increases with age; in Canada, most cases are diagnosed in those 60 years and older[2,11]. Males are more likely to get pancreatic cancer than females[2,11]. This is thought to, at least in part, be attributed to different exposure to risk factors[2]. For example, males are more likely to be smokers than females[15].

Tobacco smoking is considered a definitive risk factor for pancreatic cancer[13,14,16,17]. A meta-analysis by Iodice et al. (2008) including 82 case-control and cohort studies found that current smoking increased the risk of pancreatic cancer by 74%, while former smokers had a 20% increased risk[17]. Maisonneuve et al. (2015) classified smoking as a "moderate" risk factor of pancreatic cancer (RR (relative risk) of 1.5-1.9) after review of over 117 meta-analyses or pooled reports[16]. Other case-control, cohort, or pooled studies generally found OR's (odds ratio) or RR's within this range[18-23]. Anderson et al. (2009) found that current smokers had over 3 times the odds of pancreatic cancer compared to those who had never smoked[24]. The risk of smokeless tobacco products is less consistent[16].

A family history of pancreatic cancer is also an established risk factor of pancreatic cancer[13,16]. Like smoking, family history of pancreatic cancer was considered a "moderate" risk factor (RR of 1.5-1.9) by Maisonnueve et al.[16]. Both a meta-analysis with 9 case-control and cohort studies by Permuth-Wey et al. (2009) and a pooled analysis of 11 case-control and cohort studies by Jacobs et al. (2010) found that a family history of cancer increased pancreatic cancer risk by around 80%[25,26]. Individual case-control studies, however, have found that a family history of pancreatic cancer can increase odds of the disease anywhere from 2 to 4 times compared to those without family history[21,24,27].

Diabetes is also consistently found to moderately increase the risk of pancreatic cancer[13,16]. A 2011 meta-analysis of 35 cohort studies by Ben et al. found that those with diabetes mellitus had 90% greater risk of pancreatic cancer than those without diabetes[28]. A case-control analysis by Rahman et al. found participants with diabetes have 70% higher odds of pancreatic cancer[21]. Other case-control studies have found anywhere from no significant association[24] to an odds ratio of up to 2[23,29].

Finally, body mass index (BMI) is often found to be associated with pancreatic cancer risk[30], though less definitively[13]. Maisonneuve et al. (2015) found BMI to be a "low" risk factor (RR of 1.1-1.4) of pancreatic cancer[16]. A meta-analysis by Larsson et al. (2007) that included 21 prospective studies found that for every 5kg/m^2 increase in BMI, pancreatic cancer risk increased 12%[31]. A pooled analysis of 14 cohort studies found that obese individuals (≥30kg/m^2) had 50% higher risk than those with a BMI of 21-22.9kg/m$^{2(32)}$. Some case-control studies have found those who are obese have 2-3 times the odds of pancreatic cancer compared to normal BMI individuals[20,24]. Other studies, however, have found that BMI was not significant risk factor of pancreatic cancer[19,21,23].

3.1.4 Other Possible Risk Factors of Pancreatic Cancer

Dietary risk factors of pancreatic cancer, including alcohol and coffee, are much inconsistent and not well understood[13,14]. Though alcohol has been found to increase pancreatic cancer risk, significant associations are often only found in heavy drinkers[16,33]. Maisonneuve et al. categorized heavy alcohol consumption as a "low" risk factor (RR of 1.1-1.4) of pancreatic cancer[16]. Lucenteforte, in a pooled anaysis of 10 case-control studies, found a 60% increase in pancreatic cancer risk for heavy drinkers (≥9 drinks/day) compared to those who never or occasionally drink[33]. Hanley et al. and Anderson et al., both case-control analyses, found that light or moderate alcohol consumption, respectively, actually decreased pancreatic cancer risk compared too little to no alcohol[20,24]. Others, however, have found that the alcohol consumption, of any amount, does not exhibit a significant association with pancreatic cancer[18,19,21,23,34]. Coffee or caffeine consumption similarly presents conflicting results. While most studies have found no significant association between coffee drinking and pancreatic cancer[16,18,23,35], a meta-analysis of 14 cohort studies by Dong et al. (2011) found that coffee decreased pancreatic cancer risk in men[36]. Anderson et al. found that coffee increases pancreatic cancer risk[24]. Increased consumption of fruit and vegetables has been implicated as a protective factor[16,23,24,37], while red and processed meat consumption may increase pancreatic cancer risk[16,23,37,38].

There are several other factors that may be associated with pancreatic cancer but will not be explored in this analysis. Chronic pancreatitis is a strong (RR of ≥2) risk factor of pancreatic cancer[16,21,29]. Gallbladder conditions[13,29,39], idiopathic thrombosis[16], thyroid conditions[39], increased sugar[13,16] and fat[13] intake, and *Helicobactor pylori* infection[13,16] have been associated with increased pancreatic cancer risk. A history of allergies is considered a moderate protective factor (RR of 0.5-0.9) of pancreatic cancer[16,40]. Increased folate consumption[13,16,34,41], higher education[24], physical activity[16,20], and increased parity in females[13,42] have also been associated with lower risk of pancreatic cancer.

3.1.5 Pilot Etiologic Study: Nested Case-Control

A nested case-control study on the risk factors of pancreatic cancer was conducted to evaluate whether the ATP dataset is feasible to conduct etiologic research on rare cancer types. Results of more definitive risk factors are compared to what we expect from the literature, and less understood dietary factors are explored in a Canadian context. This will provide insight into the utility of observational cohorts in the study of rare cancers in Canada.

3.2. Methods

3.2.1 Data source

Alberta's Tomorrow Project

ATP started recruitment in 2000, finished in 2015, and now has 55,000 participants[8]. Albertan residents aged 35-69 with no history of cancer, other than non-melanoma skin cancer, were eligible to enroll. Other enrollment criteria were that participants had to plan to reside in Alberta for at least one year and ability to compete written questionnaires in English[43]. The cohort was recruited in two phases. In Phase 1 (2000-2008), participants were recruited using random digit dialing. Households were selected from each of the 17 regional health authorities across Alberta in 2000 and one or two eligible participants were selected from each household[43]. This resulted in a cohort from across the province representing a wide range of sociodemographic and health-related factors[43]. Phase 2 (2008-2015) began after harmonization with the Canadian Partnership for Tomorrow Project (CPTP). CPTP-ATP recruitment was achieved through volunteer sampling using communication and advocating strategies to reach eligible participants[44].

This project included those participants recruited in Phase 1 (n=31,203). Upon recruitment, these participants completed the baseline Health and Lifestyle Questionnaire (HLQ), and, depending on when they were enrolled, had the opportunity to complete several follow-up questionnaires; Survey 2004, Survey 2008, Updated Health and Lifestyle Questionnaire (UHLQ, 2009-2011) and CORE (2011-2015) (Figure S1). These surveys collected information on personal characteristics, lifestyle factors, and health status. Shortly after completing HLQ, participants were also given the Canadian Diet History Questionnaire-I (CDHQ-I). This food frequency questionnaire has been adapted for use in this cohort[43]. In 2008, ATP joined the CPTP.

The UHLQ and CORE questionnaires have been updated and administered through the CPTP to support this harmonization, though the information collected from these questionnaires is similar to ATP's original follow up surveys. More information on ATP surveys can be found at www.myatpresearch.ca/survey-information. All information in these questionnaires is self-reported.

Alberta Cancer Registry

The Alberta Cancer Registry ACR is a population-based registry that collects information on all new cancer cases and cancer deaths occurring in Alberta[45]. The ACR has achieved a Gold Standard from the North American Association of Central Cancer Registries (NAACCR) for many years[45]. Cancer registries that meet the Gold Standard have achieved the highest NAACCR standard for complete, accurate, and timely data, among other data quality measures[46]. The ACR has consistently achieved completeness, or the extent to which all new cancer cases are accurately captured, of over 95%[45]. Doctors and laboratories in the province are mandated to notify the ACR of new cancer cases[45]. The ACR achieves comparability by applying standard classification and coding practices[45]. The ACR records topography, morphology, and behavior using the International Classification of Diseases for Oncology (ICD-O): ICD-O-2 for cases before 2000 and ICD-O-3 for cases 2001 and onwards. Topography codes are consistent between the two versions of ICD-O, and the few morphology and behavior codes used in this analysis are not affected by the change in versions[47]. The highest level of accuracy is achieved by numerous data edits and additional data quality reviews by the Canadian Cancer Registry (CCR) and NAACCR[45].

Diagnoses of cancer in the ACR that were behavior 2 (in situ/noninvasive) or 3 (malignant), *except* non-melanoma skin cancers and behavior 2 cervix and prostate cancer, were considered a cancer diagnosis for case ascertainment, exclusion due to previous cancer, and censoring purposes. This reflected the exclusions diagnoses that would be recognized in a broader Canadian population linked to the CCR. The CCR mandates the reporting of all primary, malignant tumors (behavior 3) and all in situ/intraepithlial/noninfiltrating/noninvasive tumors (behavior 2), *except* behavior 2 cervix and prostate cancer[48]. Non-melanoma skin cancers of any behavior code are not mandated.

Ethics Approval

Ethics approval was obtained from the Health Research Ethics Board of Alberta (Study ID CC-16-0880).

Data Linkage

Alberta Cancer Registry data was obtained through linkage with Surveillance & Reporting, C-MORE CancerControl Alberta. ACR and ATP linkage was performed by these agencies prior to dispensing the data. Participants were linked on Alberta Personal Health Care Number, and confirmed on first name, last name, and date of birth[8]. Participants that did not consent to data linkage (n=360) were included in the dataset but given a value corresponding to "no consent" in a variable related to cancer status at baseline.

3.2.2 Data Preparation for Nested Case-Control Analyses

Exclusions

All data preparation and analysis was done with STATA IC v15. Exclusion criteria applied to the cohort for analysis is shown in Figure 3.1. There were 360 participants that did not consent to linkage and were excluded. Though participants must report that they have never had cancer at baseline to be included in the cohort, there were 118 participants that had age of first cancer diagnosis in the ACR before age at baseline (HLQ); these participants were excluded so the study population included only those that had no history of cancer at baseline.

Identifying Cases

Cases were identified as those ATP participants who had a diagnosis of pancreatic cancer (ICD-O-3 topography codes C25.0-25.4, C25.7-25.9) of behavior 2 (in situ/noninvasive) or 3 (malignant) in the ACR. Only cases that were the participant's first cancer, as defined in this analysis, were included, as the etiology of pancreatic cancer may be different in someone who has had a previous malignancy. A case's follow-up time was between their age at baseline and age at diagnosis. There were 72 incident cases of primary pancreatic cancer in this cohort; all except one were behavior 3 (malignant).

Identifying Eligible Controls

Follow-up time for a control was from their age at baseline to their age at the most recent follow-up survey they completed. There were some participants who only took the baseline survey and did not complete a follow-up survey (n=6,695); these participants had no follow-up time and were excluded (Figure 3.1). The distribution of covariates in those who did not take

HLQ, including cases, was compared to the group who took at least one follow-up survey to evaluate if there were any differences between them (Table 3.1). Those with a record of cancer other than pancreatic cancer in the ACR, within their follow-up time, were censored at their age of diagnosis.

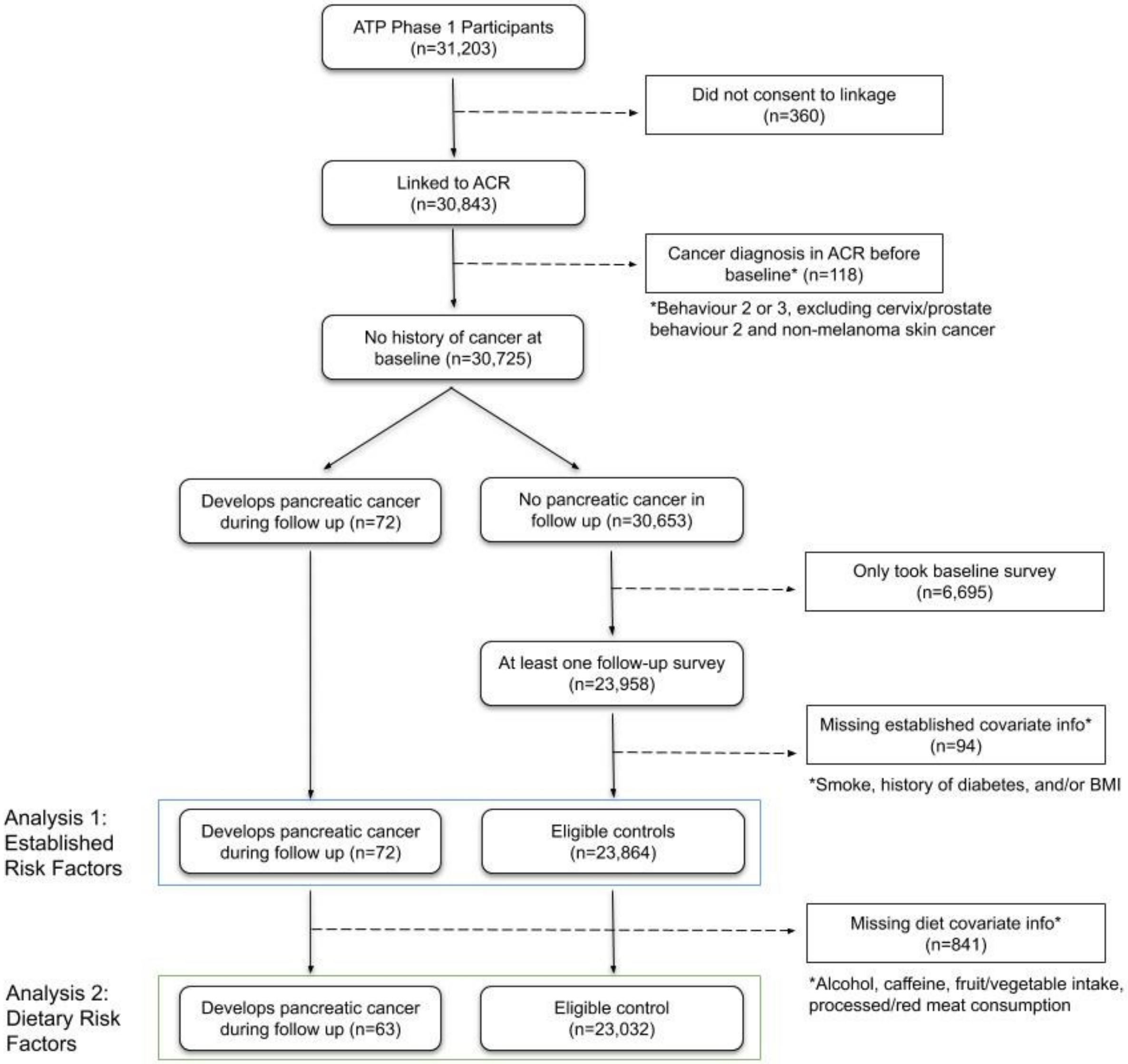

ATP Phase 1 Participants (n=31,203)
Did not consent to linkage (n=360)
Linked to ACR (n=30,843)
Cancer diagnosis in ACR before baseline* (n=118)
*Behaviour 2 or 3, excluding cervix/prostate behaviour 2 and non-melanoma skin cancer
No history of cancer at baseline (n=30,725)
Develops pancreatic cancer during follow up (n=72)
No pancreatic cancer in follow up (n=30,653)
Only took baseline survey (n=6,695)
At least one follow-up survey (n=23,958)
Missing established covariate info* (n=94)
*Smoke, history of diabetes, and/or BMI
Analysis 1: Established Risk Factors
Develops pancreatic cancer during follow up (n=72)
Eligible controls (n=23,864)
Missing diet covariate info* (n=841)
*Alcohol, caffeine, fruit/vegetable intake, processed/red meat consumption
Analysis 2: Dietary Risk Factors
Develops pancreatic cancer during follow up (n=63)
Eligible control (n=23,032)

Figure 3.1. A flowchart illustrating the steps of participant exclusion and the cases and pool of controls used for the first and second analysis.
ATP=Alberta Tomorrow Project. ACR=Alberta Cancer Registry.

3.2.3 First Nested Case-Control Analysis: Established Risk Factors of Pancreatic Cancer

Covariates

Covariates included in the first analysis were smoking status, family history of cancer, history of diabetes, and BMI. All of these covariates were categorical and generated from self-reported data from the baseline survey (HLQ). Participants were grouped into 3 categories for smoking: Never (reference), Former, and Current. Self-reported family history of cancer in a parent or sibling was categorized into three categories: None or other cancer (reference), digestive-related cancer, and pancreatic cancer. Participants with no family history of cancer and family history of other cancer had similar risk of pancreatic cancer and were combined to create the reference category. Digestive-related cancers included any responses that were related to the digestive system, such as digestive, stomach, gastric/gastrointestinal, liver/biliary/hepatic, bowel, colorectal, intestinal, esophagus, appendix, or gall bladder. History of diabetes was a binary variable: No history of diabetes (reference) and History of diabetes. HLQ did not differentiate between type 1, type 2, and gestational diabetes in this question. Self-reported BMI at baseline was grouped into three categories, according to the World Health Organization (WHO) guidelines[49]: Normal weight (<25kg/m^2, reference), Overweight (≥25 and <30kg/m^2), and Obese (≥30kg/m^2). There were some participants with a BMI <18.5, which is classified as underweight, but since most associations with BMI and their biologic mechanisms are for those who are overweight or obese, these participants were included in the reference category[16,30]. All 72 cases had information on these covariates. There were 94 possible controls that were missing at least one of smoking, history of diabetes, and/or BMI, and were excluded from the pool of eligible controls (Figure 3.1).

Matching

Incidence density sampling was used to select controls from those who were at risk at the time of each case when they were diagnosed. As controls are implicitly matched on follow up time in the analysis, and a control can later become a case, the estimated OR is an unbiased estimate of the rate ratio (RR). Controls were randomly selected from those at risk of pancreatic cancer at the time the case is diagnosed, matching on sex (male of female) and age (caliper matched +/- 2 years). Where possible, up to ten controls were selected for each case. All but three cases were matched to ten controls and all cases were matched to at least one control. Sampling was done with replacement; a control can later become a case and an individual can serve as a control for more than one case. This sampling method produced a sample size of 765 for the first analysis.

Conditional Logistic Regression

Conditional logistic regression was conducted for univariate and multivariate analyses. Univariate OR's with 95% confidence intervals, conditional on follow-up time, age (+/- 2 years), and sex, were estimated for smoking status, family history of cancer, history of diabetes, and BMI. Multivariate OR's were estimated for all four covariates from a model containing the other three covariates, as well as for only those covariates with $p<0.2$ in the likelihood ratio test in univariate analyses (smoking status, history of diabetes, and family history of cancer). OR's estimate the rate ratio.

3.2.4 Second Nested Case-Control Analysis: Dietary Risk Factors of Pancreatic Cancer

Covariates

A second nested case-control analysis looked at associations between dietary risk factors of pancreatic cancer, adjusted for covariates that were significant in the first analysis (smoking, family history of cancer, history of diabetes). Covariates in this analysis were alcohol consumption, caffeine consumption, fruit and vegetables consumption, and red/processed meat consumption. All were categorical and generated from self-reported data from the CDHQ-I survey. There were 841 participants, including 9 cases, which did not take the CDHQ-I survey and were therefore missing data on these covariates. They were excluded (Figure 3.1).

Participants who took CDHQ-I were grouped into four categories for alcohol consumption from their reported number of drinks per day: Never (0 drinks/day, reference), Light (≤1 drink/day for women, ≤1.5 drinks per day for men), Moderate (>1 and ≤2 drinks/day for women, >1.5 and ≤3 drinks/day for men), and Heavy (>2 drinks/day for women, >3 drinks/day for men). Heavy drinkers were defined based on Canada's low risk drinking guidelines[50], and the light and moderate categories were based on the distribution of the number of drinks per day in the sample. Heath Canada recommends no more than 400mg of caffeine per day[51], but many participants consumed beyond this. The caffeine reference category was based on this recommendation and two higher consumption categories were based on the distribution of consumption in the cohort: ≤400mg/day, >400 to 700mg/day, and >700mg/day.

World Cancer Research Fund (WCRF) cancer prevention recommendations[52] were used to determine desired consumption of fruit and vegetables (including fruit juices and vegetable juices), and low risk consumption of red and processed meat (including meat from beef, pork, veal, lamb, game, franks, sausages, and luncheon meats). WCRF recommends at least 5 servings/week of fruit and non-starchy vegetables. Participants were grouped into two categories for fruit and vegetable consumption: <5 servings/day (reference), and ≥5 servings/day. The variable in servings/day available from CDHQ-I to generate this covariate did, however, include non-starchy vegetables. WCRF recommends no more than 3 servings of red meat per week (12-18 ounces/week, cooked) and very little, if any, processed meat. There were three categories of red and processed meat consumption: <12 ounces/week (reference), 12-18 ounces/week, and >18 ounces/week.

Matching

Incidence density sampling with replacement, similar to the first analysis, was used to select controls from those who were at risk at the time of each case when they were diagnosed. Controls were randomly selected from those at risk of pancreatic cancer at the time of the case, matched on sex (male or female), age (caliper matched +/- 5 years), smoking status (never, former, or current), and history of diabetes (yes or no). A five-year age range was used for age-matching. This was less restrictive than the two-year caliper in the first analysis since there were other matching factors in this analysis that restricted the number of eligible controls for each case. Though family history of cancer was significant at p<0.2 in the first analysis, matching on this factor further limited control selection. Family history was adjusted for in multivariate analysis rather than matched on to avoid losing power. Where possible, up to ten controls were selected for each case. Two cases were not matched to any controls and were excluded from the analysis. This sampling method produced a sample size of 621 for the second analysis.

Conditional Logistic Regression

Conditional logistic regression analyses were conducted for univariate and multivariate analyses. Univariate OR's with 95% confidence intervals, conditional on follow up time, age (+/-5 years), sex, smoking status, and history of diabetes and adjusted for family history, were calculated for alcohol, caffeine, fruit/vegetable, and red/processed meat consumption. Multivariate OR's were calculated for all four covariates from a model containing the other three covariates and family history of cancer using conditional logistic regression. The OR's in this analysis estimate the rate ratio.

3.3. Results

3.3.1 Comparison of Participants with and without Follow-Up

Table 3.1 displays the distribution of study variables in those that took at least one follow up survey and those that only took the baseline survey in order to evaluate if there were any differences in these two groups. Distribution of the variables was, for the most part, similar between the two groups. There were more males in the group that had no follow up (42.3% vs 37.9% in group that had at least one follow up). Those without follow up were more likely to be current smokers than those with follow up (25.7% vs. 16.5%), and more likely to be obese (31.1% vs 25.9%). Over half of those with no follow up also did not take CDHQ-I, while most of those who had follow up data took CDHQ-I. Of those who took CDHQ-I, the two groups were similar in their consumption of alcohol and fruit and vegetables. Those without follow up consumed slightly more caffeine (30.6% vs 27.4% consuming >700mg/day) and red or processed meat (37.7% vs 33.8% consuming >18 ounces/week).

Table 3.1. Comparison of baseline characteristics of ATP participants who only took the HLQ baseline survey and participants who took at least one follow up survey.

		Only HLQ[a]	**At least one follow up[b]**
		N=6,710	*N=24,015*
Continuous Variables[c]		**Mean (SD)**	**Mean (SD)**
	Age at baseline	48.9 (8.9)	51 (9.2)
	BMI (continuous)	28.2 (5.8)	27.5 (5.4)
Categorical Variables[c]		**n(%)**	**n(%)**
Developed cancer[d]	No	6095 (90.8)	21464 (89.4)
	Other	600 (8.9)	2494 (10.4)
	Pancreatic	15 (0.2)	57 (0.2)
Sex	Male	2840 (42.3)	9102 (37.9)
	Female	3870 (57.7)	14913 (62.1)
Smoke	Never	2622 (39.1)	10953 (45.6)
	Former	2356 (35.1)	9082 (37.8)
	Current	1722 (25.7)	3958 (16.5)
	(missing)	10 (0.2)	22 (0.1)
Family History of cancer	No	3433 (51.2)	11349 (47.3)
	Other	2399 (35.8)	9249 (38.5)
	Digestive	750 (11.2)	2940 (12.2)
	Pancreas	128 (1.9)	477 (2.0)
Diabetes	No	6306 (94.0)	22928 (95.5)
	Yes	404 (6.0)	1087 (4.5)
BMI	Normal	2103 (31.3)	8233 (34.3)
	Overweight	2450 (36.5)	9501 (39.6)
	Obese	2087 (31.1)	6219 (25.9)
	(missing)	70 (1.0)	62 (0.3)
CDHQ Variables[e]		*N=3,561*	*N=23,166*
Alcohol	Never	606 (17.0)	3491 (15.1)
	Light	2426 (68.1)	16150 (69.7)
	Moderate	322 (9.0)	2283 (9.9)
	Heavy	207 (5.8)	1242 (5.4)
Caffeine (mg/day)	≤400	1708 (48.0)	11460 (49.5)
	>400 to 700	762 (21.4)	5363 (23.2)
	>700	1091 (30.6)	6343 (27.4)
Fruit/Vegetables (servings/day)	<5	1805 (50.7)	11629 (50.2)
	≥5	1756 (49.3)	11537 (49.8)
Meat (ounces/week)	<12	1437 (40.4)	9987 (43.1)
	12 to 18	780 (21.9)	5354 (23.1)
	>18	1344 (37.7)	7825 (33.8)

[a] Those who only took the baseline (HLQ) survey.

[b] Participants that took at least one of the follow-up surveys (S04, S08, UHLQ, or CORE).

[c] From baseline (HLQ) survey.

[d] Behavior 2 (in situ/noninvasive) or 3 (malignant), excluding non-melanoma skin cancer and cervix/prostate behavior 2

[e] Participants that took the Canadian Diet and Health Questionnaire (CDHQ-I), taken at the time of or shortly after HLQ.

3.3.2 Established Risk Factors of Pancreatic Cancer

Established risk factors of pancreatic cancer were included in a first analysis to compare their association, if any, in this cohort to previous findings. There were 72 cases and 693 controls in this analysis (Table 3.2). In univariate analysis, cases were twice as likely to be current smokers compared to controls (OR [95% CI]: 2.07 [1.10, 3.89], p=0.023). Former smokers exhibited a slightly decreased risk of pancreatic cancer, though this was not significant (p=0.548). Cases were over 3.6 times as likely as controls to have a family history of pancreatic cancer (OR [95% CI]: 3.61 [1.28, 10.2], p=0.015). Cases were 2.6 times as likely to have a history of diabetes compared to controls (OR [95% CI]: 2.59 [1.16, 5.75], p=0.020). Being overweight or obese was not statistically significantly associated with pancreatic cancer risk.

When smoking, family history, history of diabetes, and BMI were considered in a multivariate analysis, current smoking, family history of cancer, and a history of diabetes continued to be associated with pancreatic cancer. Current smokers had an OR of 2.10 (p=0.024), family history of pancreatic cancer had an OR of 4.27 (p=0.010) and history of diabetes had an OR of 2.32 (p=0.045) (Table 3.2). These associations were maintained and exhibited little change when BMI was removed from the analysis.

Table 3.2. Distribution of pancreatic cases and controls and univariate and multivariate OR's from conditional logistic regression for risk factors of pancreatic cancer.

Variable	Cases (n=72) # (%)	Controls (n=693) # (%)	Univariate OR[a] (95% CI)	Multivariate OR[b] (95% CI)	Multivariate OR[c] (95%CI)
Smoking Status					
Never	25 (34.7)	269 (38.8)	Ref.	Ref.	Ref.
Former	24 (33.3)	312 (45.0)	0.84 (0.47, 1.50)	0.77 (0.43, 1.40)	0.77 (0.42, 1.40)
Current	20 (27.8)	112 (16.2)	2.07 (1.10, 3.89)	2.10 (1.10, 3.98)	2.09 (1.10, 3.95)
Family History of Cancer					
No/Other cancer	—	—	Ref.	Ref.	Ref.
Digestive	—	—	1.30 (0.67, 2.53)	1.30 (0.66, 2.54)	1.32 (0.68, 2.57)
Pancreatic	—	—	3.61 (1.28, 10.2)	4.27 (1.42, 12.9)	4.30 (1.43, 12.9)
History of Diabetes					
No	62 (86.1)	650 (93.8)	Ref.	Ref.	
Yes	10 (13.9)	43 (6.2)	2.59 (1.16, 5.75)	2.32 (1.02, 5.27)	2.34 (1.03, 5.30)
BMI					
Normal	20 (27.8)	184 (26.6)	Ref.	Ref.	
Overweight	28 (38.9)	291 (42.0)	0.85 (0.46, 1.56)	0.89 (0.47, 1.68)	
Obese	24 (33.3)	218 (31.5)	0.97 (0.51, 1.86)	0.98 (0.50, 1.91)	NS

A "—"indicates that at least one cell count of the variable violated data disclosure policy (<10).
OR=Odds Ratio, CI=Confidence Interval, Ref.=reference category, NS=Not significant
[a] Adjusted for age (± 2 years), sex, and length of follow-up time
[b] Adjusted for age (± 2 years), sex, length of follow-up time, and other variables in table
[c] Adjusted for age (± 2 years), sex, length of follow-up time, and variables in table significant with a likelihood ratio test p<0.2 in univariate analysis

As smoking and history of diabetes are associated with pancreatic cancer risk, they were included as matching factors in the second analysis exploring dietary risk factors of pancreatic cancer to prevent confounding from these factors. Family history of cancer was adjusted for as a covariate, as matching on this factor limited control selection and reduced power. An analysis was also done with BMI included as a covariate (not shown) to see if it was a confounder of dietary associations. However, results from the dietary analysis including BMI were not sufficiently different from results without adjusting for BMI, and so BMI was not included in further analysis.

3.3.3 Dietary Risk Factors of Pancreatic Cancer

There were 61 cases and 560 controls in this analysis. None of the dietary factors were significantly associated with pancreatic cancer risk in univariate or multivariate analyses and most estimates hade wide confidence intervals, likely due to low sample size. However, some OR's may still be informative. In univariate analysis, light alcohol consumers had a higher risk of pancreatic cancer (OR [95% CI]: 1.90 [0.84, 4.30]) than those who never drank. Heavy drinkers had 2.9 times the risk (OR [95% CI]: 2.89 [0.81, 10.3]), though this estimate had a very wide confidence interval. These associations persisted in multivariate analysis. Heavy caffeine consumption may have a slight protective effect (multivariate OR [95% CI]: 0.81 [0.42, 1.56], respectively). Higher fruit and vegetables consumption increased pancreatic cancer risk by 20%, adjusted for the other factors, but this association was not statistically significant and the estimate was imprecise (95% CI: 0.68, 2.10). Meat consumption was not associated with pancreatic cancer risk.

Table 3.3. Distribution of pancreatic cases and controls, univariate OR's, and multivariate OR's for dietary risk factors of pancreatic cancer.

Variable	Cases (n=61) # (%)	Controls (n=560) # (%)	Univariate OR[a] (95% CI)	Multivariate OR[b] (95% CI)
Alcohol Consumption				
Never	—	—	Ref.	Ref.
Light	—	—	1.90 (0.84, 4.30)	1.87 (0.83, 4.25)
Moderate	—	—	1.10 (0.31, 3.95)	1.11 (0.31, 4.03)
Heavy	—	—	2.89 (0.81, 10.3)	2.89 (0.80, 10.4)
Caffeine Consumption				
≤400mg/day	27 (44.3)	226 (40.3)	Ref.	Ref.
>400 to 700mg/day	14 (22.9)	123 (22.0)	0.97 (0.48, 1.98)	0.97 (0.47, 2.01)
>700mg/day	20 (32.8)	211 (37.7)	0.80 (0.42, 1.52)	0.81 (0.42, 1.56)
Fruit/Vegetables Consumption				
<5 servings/day	29 (47.5)	293 (52.3)	Ref.	Ref.
≥5 servings/day	32 (52.5)	267 (47.7)	1.24 (0.72, 2.14)	1.19 (0.68, 2.10)
Processed/Red Meat Consumption				
<12 ounces/week	27 (44.3)	242 (43.2)	Ref.	Ref.
12 to 18 ounces/week	13 (21.3)	134 (23.9)	0.95 (0.47, 1.92)	0.96 (0.47, 1.96)
>18 ounces/week	21 (34.4)	184 (32.9)	1.04 (0.53, 2.03)	1.00 (0.50, 2.02)

A "—"indicates that at least one cell count of the variable violated data disclosure policy (<10).
OR=Odds Ratio, CI=Confidence Interval, Ref.=reference category
[a] Adjusted for age (± 5 years), sex, smoking status, history of diabetes, family history of cancer, and length of follow-up time
[b] Adjusted for age (± 5 years), sex, smoking status, history of diabetes, family history of cancer, length of follow-up time, and other variables in table

3.4. Discussion

This project carried out a pilot nested case-control study on pancreatic cancer in the ATP cohort to evaluate whether research on rare cancers may be feasible in large population based cohorts. This analysis is the first project from the ATP, to our knowledge, which performs a nested-case control analysis on a rare outcome[53]. Other ATP projects have proposed or carried out a nested case-control or case-cohort design on more common outcomes, such as breast cancer, lung cancer, and diabetes. Many risk factor estimates for pancreatic cancer are derived from traditional case-control designs (Supplementary Table 3). The nested case-control is a stronger design that is less susceptible to systematic biases such as recall and information bias. Exposure information in this study was collected before disease onset and collected similarly between cases and controls. If sample size is sufficient in this cohort to make this design feasible, the ATP, and other large cohorts, could provide valuable evidence in the etiology of rare cancers.

Several well-established risk factors were explored in this cohort. Overall, the association between pancreatic cancer and smoking, family history of cancer, and diabetes were consistent with other literature (Supplementary Table 3). The estimated effect of diabetes (OR=2.34) was stronger than the effect found from several meta-analyses (1.5-1.9)[16,28], but was consistent with some case-control studies that have estimated an increase in odds of pancreatic cancer of over 2 times[23,29]. Family history of cancer is often considered to have a moderate effect (1.5-1.9)[16,25,26], but this analysis found family history to be a strong risk factor (OR=4.3). Anderson et al (2009) found similar results for family history of pancreatic cancer (OR=4.2)[24]. BMI was not a significant predictor of pancreatic cancer in this analysis. Higher BMI is often considered to increase pancreatic cancer risk[16,30-32], though other studies have found no association with BMI[19,21,23].

The first risk factor analysis was adequately powered to detect moderate to strong risk effects (OR≈2) and produced results consistent with the literature. Dietary risk factor analysis, however, produced more imprecise and uncertain results. This may be partially due to attrition, though the use of self-reported dietary data may also be a factor. Food frequency questionnaires, while a cost-effective and practical option to collect long-term dietary data on a population this size, are subject to recall and social desirability bias[54]. None of the associations were statistically significant, likely due to a lack of power rather than lack of association. Some effect directions were unexpected. Most notably, increased fruit and vegetable intake slightly increased pancreatic cancer risk, contrary to other findings[16,23,37].

Several major considerations arose while evaluating the feasibility of rare cancer etiology research in these cohorts. Firstly, there was a lack of statistical power, particularly in the second analysis, due to low sample size. A small number of cases may result in some exposure categories with a very small sample size that can lead to imprecise estimates. This was demonstrated in the estimates for family history of pancreatic cancer and heavy alcohol drinking. Effect estimates for risk factors are often low to moderate, so small sample size lowers the ability to detect these differences. A power analysis (at 80% power) demonstrated that an analysis with 72 cases (as in the first analysis) could not detect significant effects with an odds ratio (OR) of less than approximately 2.2 (Table 3.4). It must be noted, however, that this is a provincial cohort. Larger cohorts are available that would offer more cases. The ATP is part of the Canadian Partnership for Tomorrow Project (CPTP), a partnership between 5 regional cohorts across Canada[44]. CPTP contains over 300,000 participants, compared to the initial 30,000 ATP participants in this analysis. The CPTP would offer more cases throughout its follow-up period and allow smaller OR's to be detected (Table 3.4). However, linkage to the Canadian Cancer Registry (CCR) would be necessary, as pancreatic cancer cases need to be ascertained by registry data. Very few pancreatic cancer cases were self-reported in the ATP. It is currently not possible for researchers to obtain CPTP data linked directly to the CCR. Participants make data linkage consent agreements within their regional cohorts and current legislation prevents this information from crossing provincial borders[44,55].

Table 3.4. Minimum case number required at 80% power and a 10:1 ratio of controls to cases for different minimum detectable odds ratio's (OR's) and probabilities of exposure in the baseline (control) population[a].

P(exposure|control)[b]	**Minimum Detectable Odds Ration (OR)**						
	1.2	**1.4**	**1.6**	**1.8**	**2.0**	**2.2**	**2.4**
0.05	4,904	1,313	621	370	250	183	141
0.1	2,622	711	340	205	140	104	81
0.2	1,513	420	206	127	88	66	53
0.3	1,182	336	168	105	75	57	46
0.4	1,060	308	157	100	72	56	45
0.5	1,043	309	160	104	76	59	49

[a] Case numbers in dark grey were achieved by the first analysis of this project, which had 72 cases. Cells in light grey may be possible if this analysis was conducted in the Canadian Partnership for Tomorrow Project (CPTP), which has approximately 10 times the enrollment of the Alberta Tomorrow Project (ATP).
[b] The probability of exposure in the population of eligible controls.

A second consideration relates to the loss of follow up in cohort studies and impact on the internal validity of results. Loss to follow up, or attrition, is a problem that has been acknowledged by stakeholders in Canadian cohort studies[56]. In this study, cases were followed passively by ACR linkage. However, controls required active follow-up, through completion of a follow-up survey, to be included. As addressed in the results, there were a large number of ATP participants that did not complete a follow up survey. Table 3.1 demonstrated that there were some differences between those with and without a follow-up survey, introducing a possibility of bias. Those with follow-up were healthier, overall. Further, those who are healthier or more health conscious may take more surveys and have longer follow-up periods. Characteristics of the pool of eligible controls may change as the study progresses—cases with longer follow-up times may be matched to healthier individuals overall than cases with shorter follow-up times. Linkage to vital statistics may help to alleviate this issue, as controls could also be passively followed up. This would also provide a larger pool of eligible controls since even those with no follow-up survey would have a known vital status.

A lack of follow-up for cases and controls also introduces the possibility of bias or misclassification of exposure status over time. Some exposures, such as smoking status, BMI, and dietary intake, may vary over time. However, since most cases lacked follow-up survey information on these variables, baseline data was used for all participants. Particularly for participants with long follow-up times, this could misclassify their exposure status at case ascertainment and control selection. This issue is somewhat mitigated by the fact that many lifestyle risk factors act in a long-term fashion; changes throughout follow-up may not increase or decrease an individual's risk from baseline.

External validity must also be considered in this design. As this is a voluntary participation cohort, a "healthy volunteer effect" may be present in those that decide to enroll in the study[57]. The cohort may not be representative of the general population in terms of exposure status distribution and outcome probability[43]. If there are differences in the etiology of disease between those who are healthier (enroll) and those who are unhealthier (don't enroll), the generalizability of the study is affected.

The results of this pilot study demonstrate that a nested case-control design may be possible to study rare cancer in a large cohort. However, rare cancer research is still limited by the number of cases that develop, missing data, and the availability of longitudinal follow-up data. Using larger cohorts, such as the Canadian Partnership for Tomorrow project, may be more feasible provided that follow-up procedures are adequate for eligible controls and linkage to cancer registry is available to identify cases. If these issues are mitigated, the opportunity these cohorts may hold for rare cancer research can be further explored.

3.5. References

1. Maplethorpe E, Walker EV, Davis FG. Occurrence of rare cancer in Canada: The distribution of cancer incidence among the Canadian population from 2009-2013. *Poster session presented at: Canadian Research Data Center Network Conference, Hamilton, ON.* 2018.

2. Canadian Cancer Society's Advisory Committee on Cancer Statistics. Canadian Cancer Statistics 2017. cancer.ca/Canadian-CancerStatistics-2017-EN.pdf. Updated 2017. Accessed February, 2018.

3. von der Schulenburg JM, Pauer F. Reviews: Rare cancers—Rarity as a cost and value argument. *Journal of Cancer Policy*. 2017;11:54-59. doi: 10.1016/j.jcpo.2016.09.004.

4. Mathoulin-Pélissier S, Pritchard-Jones K. Evidence-based data and rare cancers: The need for a new methodological approach in research and investigation. *European Journal of Surgical Oncology*. 2018;45(1):22-30. doi: 10.1016/j.ejso.2018.02.015.

5. Boyd N, Dancey JE, Gilks CB, Huntsman DG. Rare cancers: A sea of opportunity. *Lancet Oncology*. 2016;17(2):e61. doi: 10.1016/S1470-2045(15)00386-1.

6. Armstrong-Wells J, Goldenberg NA. Institution-based prospective inception cohort studies in neonatal rare disease research. *Seminars in Fetal and Neonatal Medicine*. 2011;16(6):355-358. doi: 10.1016/j.siny.2011.07.004.

7. Pearce N. What does the odds ratio estimate in a case-control study? *International Journal of Epidemiology*. 1993;22(6):1189-1192. doi: 10.1093/ije/22.6.1189.

8. Alberta's Tomorrow Project. All about Alberta's Tomorrow Project. https://myatp.ca/about-atp. Updated 2019. Accessed January, 2019.

9. Walker EV, Maplethorpe E, Davis FG. Common and rare cancer incidence rates in the Canadian population: 2009-2013; 2019. In preparation for submission at the time of this thesis.

10. Greenlee RT, Goodman MT, Lynch CF, Platz CE, Havener LA, Howe HL. The occurrence of rare cancers in U.S. adults, 1995-2004. *Public Health Reports*. 2010;125(1):28-43.

11. Flook R, van Zanten SV. Pancreatic cancer in Canada: Incidence and mortality trends from 1992 to 2005. *Can J Gastroenterol*. 2009;23(8):546-550.

12. Ilic M, Ilic I. Epidemiology of pancreatic cancer. *World Journal of Gastroenterology*. 2016;22(44):9694-9705. doi: 10.3748/wjg.v22.i44.9694.

13. Hart AR, Kennedy H, Harvey I. Pancreatic cancer: A review of the evidence on causation. *Clinical Gastroenterology and Hepatology*. 2008;6(3):275-282. https://www.clinicalkey.es/playcontent/1-s2.0-S154235650701244X. doi: 10.1016/j.cgh.2007.12.041.

14. Lowenfels AB, Maisonneuve P. Epidemiology and risk factors for pancreatic cancer. *Best Practice & Research Clinical Gastroenterology*. 2006;20(2):197-209. https://www.sciencedirect.com/science/article/pii/S152169180500154X. doi: 10.1016/j.bpg.2005.10.001.

15. Statistics Canada. Current smoking trends. Health at a glance; 2015. https://www150.statcan.gc.ca/n1/pub/82-624-x/2012001/article/11676-eng.htm.Accessed June 2019.

16. Maisonneuve P, Lowenfels AB. Risk factors for pancreatic cancer: A summary review of meta-analytical studies. *International Journal of Epidemiology*. 2015;44(1):186-198. https://www.ncbi.nlm.nih.gov/pubmed/25502106. doi: 10.1093/ije/dyu240.

17. Iodice S, Gandini S, Maisonneuve P, Lowenfels AB. Tobacco and the risk of pancreatic cancer: A review and meta-analysis. *Langenbecks Arch Surg*. 2008;393(4):535-545. doi: 10.1007/s00423-007-0266-2.

18. Villeneuve P, Johnson K, Hanley A, Mao Y, Canadian Cancer Registries Epidemiology Research Group. Alcohol, tobacco and coffee consumption and the risk of pancreatic cancer:

Results from the Canadian Enhanced Surveillance System Case-Control Project. *Eur J Cancer Prev*. 2000;9(1):49-58.

19. Kuzmickiene I, Everatt R, Virviciute D, et al. Smoking and other risk factors for pancreatic cancer: A cohort study in men in Lithuania. *Cancer Epidemiology*. 2013;37(2):133-139. doi: 10.1016/j.canep.2012.10.001.

20. Hanley AJG, Johnson KC, Villeneuve PJ, Mao Y, Canadian Cancer Registries Epidemiology Research Group. Physical activity, anthropometric factors and risk of pancreatic cancer : Results from the Canadian Enhanced Cancer Surveillance System. *International Journal of Cancer*. 2001(1):140.

21. Rahman F, Cotterchio M, Cleary SP, Gallinger S. Association between alcohol consumption and pancreatic cancer risk: A case-control study. *PLoS ONE*. 2015;10(4):e0124489. doi: //dx.doi.org/10.1371/journal.pone.0124489.

22. Bosetti C, Lucenteforte E, Silverman DT, et al. Cigarette smoking and pancreatic cancer: An analysis from the International Pancreatic Cancer Case-Control Consortium (Panc4). *Ann Oncol*. 2012;23(7):1880-1888. doi: 10.1093/annonc/mdr541.

23. Liu SZ, Chen WQ, Wang N, Yin MM, Sun XB, He YT. Dietary factors and risk of pancreatic cancer: A multi-centre case-control study in China. *Asian Pacific Journal of Cancer Prevention*. 2014;15(18):7947-7950.

24. Anderson LN, Cotterchio M, Gallinger S. Lifestyle, dietary, and medical history factors associated with pancreatic cancer risk in Ontario, Canada. *Cancer Causes Control*. 2009;20(6):825-834. doi: //dx.doi.org/10.1007/s10552-009-9303-5.

25. Permuth-Wey J, Egan KM. Family history is a significant risk factor for pancreatic cancer: Results from a systematic review and meta-analysis. *Familial Cancer*. 2009;8(2):109-117. doi: 10.1007/s10689-008-9214-8.

26. Jacobs EJ, Chanock SJ, Fuchs CS, et al. Family history of cancer and risk of pancreatic cancer: A pooled analysis from the Pancreatic Cancer Cohort Consortium (PanScan). *International Journal of Cancer*. 2010;127(6):1421-1428. doi: 10.1002/ijc.25148.

27. Austin MA, Kuo E, Van Den Eeden, S. K., et al. Family history of diabetes and pancreatic cancer as risk factors for pancreatic cancer: The PACIFIC study. *Cancer Epidemiology Biomarkers and Prevention*. 2013;22(10):1913-1917. doi: 10.1158/1055-9965.EPI-13-0518.

28. Ben Q, Xu M, Ning X, et al. Diabetes mellitus and risk of pancreatic cancer: A meta-analysis of cohort studies. *Eur J Cancer*. 2011;47(13):1928-1937. doi: 10.1016/j.ejca.2011.03.003.

29. Maisonneuve P, Lowenfels AB, Bueno-de-Mesquita HB, et al. Past medical history and pancreatic cancer risk: Results from a multicenter case-control study. *Ann Epidemiol*. 2010;20(2):92-98. doi: //dx.doi.org/10.1016/j.annepidem.2009.11.010.

30. Bracci PM. Obesity and pancreatic cancer: Overview of epidemiologic evidence and biologic mechanisms. *Molecular Carcinogenesis*. 2012;51(1):53-63. https://onlinelibrary.wiley.com/doi/abs/10.1002/mc.20778. doi: 10.1002/mc.20778.

31. Larsson SC, Orsini N, Wolk A. Body mass index and pancreatic cancer risk: A meta-analysis of prospective studies. *International Journal of Cancer*. 2007;120(9):1993-1998. https://onlinelibrary.wiley.com/doi/abs/10.1002/ijc.22535. doi: 10.1002/ijc.22535.

32. Genkinger JM, Spiegelman D, Giles GG, et al. A pooled analysis of 14 cohort studies of anthropometric factors and pancreatic cancer risk. *International Journal of Cancer (Print)*. 2011(7):1708.

33. Lucenteforte E, La Vecchia C, Silverman D, et al. Alcohol consumption and pancreatic cancer: A pooled analysis in the International Pancreatic Cancer Case-Control Consortium (PanC4). *Annals of Oncology*. 2012;23(2):374-382. https://www.ncbi.nlm.nih.gov/pubmed/21536662. doi: 10.1093/annonc/mdr120.

34. Yallew W, Bamlet WR, Oberg AL, et al. Association between alcohol consumption, folate intake, and risk of pancreatic cancer: A case-control study. *Nutrients*. 2017;9(5):448. doi: 10.3390/nu9050448.

35. Guertin KA, Freedman ND, Loftfield E, Stolzenberg-solomon RZ, Graubard BI, Sinha R. A prospective study of coffee intake and pancreatic cancer: Results from the NIH-AARP diet and health study. *The British Journal of Cancer*. 2015;113(7):1081-1085. https://www.ncbi.nlm.nih.gov/pubmed/26402414. doi: 10.1038/bjc.2015.235.

36. Dong J, Zou J, Yu XF. Coffee drinking and pancreatic cancer risk: A meta-analysis of cohort studies. *World J Gastroenterol*. 2011;17(9):1204-1210. doi: 10.3748/wjg.v17.i9.1204.

37. Ghadirian P, Nkondjock A. Consumption of food groups and the risk of pancreatic cancer: A case-control study. *Journal of Gastrointestinal Cancer*. 2010;41(2):121-129. doi: //dx.doi.org/10.1007/s12029-009-9127-2.

38. Larsson SC, Wolk A. Red and processed meat consumption and risk of pancreatic cancer: Meta-analysis of prospective studies. *British Journal of Cancer*. 2012;106(3):603-607.

39. Ko AH, Wang F, Holly EA. Pancreatic cancer and medical history in a population-based case-control study in the San Francisco Bay Area, California. *Cancer causes & control*. 2007;18(8):809-819. doi: 10.1007/s10552-007-9024-6.

40. Olson SH, Hsu M, Satagopan JM, et al. Allergies and risk of pancreatic cancer: A pooled analysis from the Pancreatic Cancer Case-Control Consortium. *Am J Epidemiol*. 2013;178(5):691-700. doi: 10.1093/aje/kwt052.

41. Marley AR, Fan H, Hoyt ML, Anderson KE, Zhang J. Intake of methyl-related nutrients and risk of pancreatic cancer in a population-based case-control study in Minnesota. *European Journal of Clinical Nutrition*. 2018;72(8):1128-1135. doi: 10.1038/s41430-018-0228-5.

42. Kreiger N, Lacroix J, Sloan M. Hormonal factors and pancreatic cancer in women. *Ann Epidemiol*. 2001;11(8):563-567.

43. Robson PJ, Solbak NM, Haig TR, et al. Design, methods and demographics from phase I of Alberta's Tomorrow Project cohort: A prospective cohort profile. *Canadian Medical Association Journal Open*. 2016;4(3):E527. https://www.clinicalkey.es/playcontent/1-s2.0-S2291002616301345. doi: 10.9778/cmajo.20160005.

44. Dummer TJ, Awadalla P, Boileau C, et al. The Canadian Partnership for Tomorrow Project: A pan-Canadian platform for research on chronic disease prevention. *CMAJ*. 2018;190(23):E717. doi: 10.1503/cmaj.170292.

45. Cancer Control Alberta. Surveillance and reporting: The 2019 report on cancer statistics in Alberta. *Alberta Health Services.* 2019.

46. North American Association of Central Cancer Registries. Certification criteria. https://www.naaccr.org/certification-criteria/. Updated 2018. Accessed July, 2019.

47. Cancer Statistics Branch, Division of Cancer Control and Population Sciences Surveillance, Epidemiology and End Results Program. Conversion of neoplasms by topography and morphology from the International Classification of Diseases for Oncology, second edition to International Classification of Diseases for Oncology, third edition; 2001.

48. Statistics Canada. Canadian Cancer Registry (CCR). http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3207. Updated 2019. Accessed April, 2019.

49. World Health Organization. Body mass index - BMI. http://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi. Accessed February, 2019.

50. Canadian Center on Substance Abuse and Addiction. Drinking guidelines. http://www.ccdus.ca/Eng/topics/alcohol/drinking-guidelines/Pages/default.aspx. Updated 2018. Accessed February, 2019.

51. Government of Canada, Health Canada. Recalls and safety alerts: Health Canada reminds Canadians to manage their caffeine consumption. http://healthycanadians.gc.ca/recall-alert-rappel-avis/hc-sc/2013/34021a-eng.php. Updated 2013. Accessed February, 2019.

52. World Cancer Research Fund. Recommendations and public health and policy implications. 2018.

53. Canadian Partnership for Tomorrow Project. About. https://www.partnershipfortomorrow.ca/about/. Accessed January, 2019.

54. Shim J, Oh K, Kim HC. Dietary assessment methods in epidemiologic studies. *Epidemiology and health*. 2014;36:e2014009. doi: 10.4178/epih/e2014009.

55. Doiron D, Raina P, Fortier I. Linking Canadian population health data: Maximizing the potential of cohort and administrative data. *Can J Public Health*. 2013;104(3):e261.

56. Boffetta P, Colditz GA, Potter JD, et al. Cohorts and consortia conference: A summary report (Banff, Canada, June 17-19, 2009). *Cancer Causes Control*. 2011;22(3):463-468. https://www.jstor.org/stable/41485113. doi: 10.1007/s10552-010-9717-0.

57. Lindsted KD, Fraser GE, Steinkohl M, Beeson WL. Healthy volunteer effect in a cohort study: Temporal resolution in the Adventist Health Study. *Journal of Clinical Epidemiology*. 1996;49(7):783-790. https://www.sciencedirect.com/science/article/pii/0895435696000091. doi: 10.1016/0895-4356(96)00009-1.

CHAPTER 4: Conclusion

The recent emergence of large, collaborative observational cohorts has been implicated as an opportunity to study rare cancer etiology with a sufficiently large sample size[1,2]. In Canada, the Canadian Partnership for Tomorrow Project (CPTP) was recognized to possibly provide this opportunity. Though, ideally, the CPTP would be directly explored for its potential in rare cancer research, CPTP data linked to the Canadian Cancer Registry (CCR) is not available to researchers. Legislation currently limits the sharing of health data cross-provincially, introducing a major barrier to health research in Canada[3,4]. Researchers can, however, access regional cohort data linked to provincial cancer registry[5]. Therefore, Alberta's Tomorrow Project (ATP), the Albertan regional cohort of the CPTP, was used in this thesis. This cohort, as a partner of the CPTP, offers data that is comparable to the CPTP and an accessible option for cancer registry linkage. The ATP was used as an indicator of the opportunities the CPTP may offer for rare cancer research.

This thesis evaluated whether large observational cohorts are feasible to study rare cancers, particularly in the absence of cancer registry linkage. This question was addressed through two objectives. The first objective evaluated whether self-reported cancer diagnosis could be used as an outcome in rare cancer research. In the second objective, a pilot etiologic study on pancreatic cancer was conducted to evaluate the feasibility of rare cancer research in cohorts using a nested case-control design.

4.1 Summary of Main Findings

4.1.1 Validation of Self-Reported Cancer Diagnoses

Self-reported cancer diagnoses in the ATP were compared to Alberta Cancer Registry (ACR) diagnoses to evaluate whether self-reported cancer diagnosis is a valid outcome for etiologic research. Overall, cancer diagnosis was well captured by self-report; 92.1% of those who had cancer reported that they had cancer. Of those who reported that they had cancer, 77.8% actually had cancer. Rare cancers were reported much less accurately than common cancers. Rare cancers had a lower sensitivity than common cancers (62.8% vs. 89.6%) and a much lower PPV than common cancers (35.8% vs. 84.5%). Participants who correctly reported that they had a rare cancer were less likely to get the specific site correct than participants who correctly reported that they had a common cancer. Some common cancer sites, such as breast and prostate cancer, had high sensitivities and PPV's. These self-reported diagnoses could be used as an outcome with little risk of bias due to misclassification of disease status. Self-reported rare cancer diagnoses, however, should not be used as an outcome as the risk of misclassification is high.

Unfortunately, the sensitivity and PPV of some rare cancer sites could not be estimated due to small sample size. For many of these sites, enough participants in the cohort developed the cancer. But, too few were followed-up after their diagnosis to actually have the opportunity to report. This highlighted another important issue in relying on self-reported diagnosis as an outcome: too few rare cancer diagnoses are actually captured in active follow-up. Not only do self-reported rare cancer diagnoses lack validity to use as an outcome, but they are also not feasible. Therefore, this validation study concluded that cancer registry linkage is required in order to provide both valid *and* complete cancer diagnosis outcome data for rare cancer research in large observational cohorts.

4.1.2 Pilot Etiologic Study on Pancreatic Cancer

A pilot etiologic study was conducted to evaluate the feasibility of large cohorts in the study of rare cancer. ATP data linked to the ACR was used as it was determined in the first study that this is a necessary step to research rare cancer. In fact, the pilot study would not have been possible without registry linkage. Very few (<10) pancreatic cancer cases were self-reported, and fewer were actually true primary pancreatic cancer cases.

Two nested case-control analyses estimated the effects of 1) established risk factors of pancreatic cancer (smoking, family history of pancreatic cancer, diabetes, and body mass index (BMI)), and 2) less-established dietary risk factors of pancreatic cancer (alcohol, coffee, fruit/vegetables consumption, and red/processed meat consumption). The first nested-case control was adequately powered to detect moderate-strong effects and produced estimates for the established effects that were relatively consistent with other literature. The second dietary analysis, however, had fewer cases and eligible controls due to loss of follow-up. This reduced power and produced imprecise estimates of effects.

4.2 Limitations and Considerations of Large Cohorts in Rare Cancer Research

Several limitations and considerations were identified regarding the use of large cohorts for rare cancer research. Firstly, cancer registry linkage is a necessary step to obtain a valid cancer diagnosis to study rare cancers in these cohorts. The importance of cancer registries in rare cancer research has been previously acknowledged[1,3,6]. Nonetheless, barriers to cross-provincial data sharing prevent researchers from accessing CPTP data linked to the Canadian

Cancer Registry[3,5]. The CPTP, which would offer more cases than the ATP, still requires registry linkage to utilize these cases as active follow-up neither adequately nor accurately captures the cases that occur.

Attrition was identified as another major limitation of using cohort data in the study of rare cancers. Attrition has been acknowledged as an issue in Canadian cohort studies that threatens internal validity[7]. Indeed, the pilot study in this thesis demonstrated that those who took at least one follow-up survey were somewhat healthier than those who did not take a follow-up survey, introducing selection bias. Attrition could, however, be mitigated by linkage to vital statistics. This would allow for passive follow-up of controls. Currently, ATP data can be linked to vital statistics from Alberta Health if a researcher applies for access to this linked data[8]. However, it is likely that the same cross provincial barriers preventing national cancer registry linkage to the CPTP would also prevent vital statistics linkage. Vital statistics, like cancer registry data, are provincially compiled and then collected by the national database[9].

A final limitation is the likely presence of a "healthy volunteer effect" that affects external validity[10]. Those who participate in population-based cohorts are healthier and more health conscious than the general population. Participants in the CPTP are more educated, more affluent, less ethnically diverse, less likely to smoke than the general Canadian population[5]. There is still, however, heterogeneity in many sociodemographic variables and a similar prevalence to the Canadian population of important risk factors for disease[5]. This supports the generalizability of some results and highlights the important questions this cohort can address in terms of disease etiology.

4.3 Opportunities for Further Research

Several research questions directly follow this project. Firstly, this validation study focused on the validity of self-reported diagnoses: whether or not someone correctly identified their cancer site. Due to the repeated follow-ups available in the ATP, the reliability of self-reported diagnoses, or how consistently a participant reports a cancer diagnosis, may also be explored. Secondly, it was suggested in the second study that linkage to vital statistics can improve passive follow-up of controls. Replicating the pilot study after linkage of ATP to vital statistics can verify that passive follow-up may strengthen the study by mitigating the impact of attrition on eligible controls. This would provide further evidence for the necessity of data sharing in rare cancer research and support the removal of barriers to cross-provincial data sharing.

4.4 Implications for Stakeholders and Policy

Researchers should be aware that emerging cohorts can be used at a regional level to study common cancers. Common cancers, in particular breast and prostate cancers, are well-reported and achieve high enough case numbers to be studied in the ATP. However, researchers are advised to take additional steps in order to study rare cancers in cohort studies. Cross-cohort collaboration is required to allow for adequate power to study rare outcomes. The CPTP may offer this opportunity, though linkage to cancer registry and vital statistics is required to provide passive follow-up and accurate diagnosis data. Researchers should continue to advocate for improvements to data sharing and provide evidence for the positive impact it can have.

Cohort leadership should ensure that all data opportunities available within and beyond the cohort, including linkage to other data sources, are promoted to researchers. User friendly protocols to access this data should be prioritized. Stakeholders have previously recognized the importance of detailed data access policies to facilitate transparency in data sharing[7]. Though active follow-up procedures can always be improved, the difficulty of active follow-up on this scale is recognized. The importance of passive follow-up through administrative database linkage, then, should be communicated to stakeholders to move towards facilitating data sharing beyond provincial boundaries.

In order for cross-provincial information sharing to be facilitated, legislative barriers that currently prevent this must be addressed[3]. Policy makers, together with researchers, are encouraged to develop new methods to facilitate data-sharing while maintaining appropriate privacy and ethics protections[3]. All stakeholders must recognize and advocate for the importance of cross-provincial data sharing. The facilitation, creation, and maintenance of linked data sources will require multi-level cooperation between researchers, cohorts, and government agencies[7].

4.5 Conclusion

This thesis explored the accuracy of self-reported cancer diagnosis and carried out a pilot etiologic study on pancreatic cancer within a Canadian cohort study. Large cohorts have the potential to provide enough cases to study rare cancer in Canada. However, this thesis demonstrated that cohort linkage to cancer registry is a necessary step to study rare cancers feasibly. Passive follow-up procedures must also be made more accessible through linkage to other administrative databases. Stakeholders must work together to overcome the barriers that prevent cross-provincial information sharing. Removing these barriers will allow the potential large cohorts offer rare cancer research in Canada to be further explored.

4.6 References

1. Mathoulin-Pélissier S, Pritchard-Jones K. Evidence-based data and rare cancers: The need for a new methodological approach in research and investigation. *European Journal of Surgical Oncology*. 2018;45(1):22-30. doi: 10.1016/j.ejso.2018.02.015.

2. Armstrong-Wells J, Goldenberg NA. Institution-based prospective inception cohort studies in neonatal rare disease research. *Seminars in Fetal and Neonatal Medicine*. 2011;16(6):355-358. doi: 10.1016/j.siny.2011.07.004.

3. Doiron D, Raina P, Fortier I. Linking Canadian population health data: Maximizing the potential of cohort and administrative data. *Can J Public Health*. 2013;104(3):e261.

4. Brown C. Barriers to accessing data are bad medicine. *CMAJ*. 2014;186(16):1203. doi: 10.1503/cmaj.109-4894.

5. Dummer TJ, Awadalla P, Boileau C, et al. The Canadian Partnership for Tomorrow Project: A pan-Canadian platform for research on chronic disease prevention. *CMAJ*. 2018;190(23):E717. doi: 10.1503/cmaj.170292.

6. von der Schulenburg JM, Pauer F. Reviews: Rare cancers—Rarity as a cost and value argument. *Journal of Cancer Policy*. 2017;11:54-59. doi: 10.1016/j.jcpo.2016.09.004.

7. Boffetta P, Colditz GA, Potter JD, et al. Cohorts and consortia conference: A summary report (Banff, Canada, June 17-19, 2009). *Cancer Causes Control*. 2011;22(3):463-468. https://www.jstor.org/stable/41485113. doi: 10.1007/s10552-010-9717-0.

8. Alberta's Tomorrow Project. All about Alberta's Tomorrow Project; https://myatp.ca/about-atp. Updated 2019. Accessed January, 2019.

9. Statistics Canada. Vital statistics - death database (CVSD). http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3233. Updated 2018.

10. Lindsted KD, Fraser GE, Steinkohl M, Beeson WL. Healthy volunteer effect in a cohort study: Temporal resolution in the Adventist Health Study. *Journal of Clinical Epidemiology*. 1996;49(7):783-790. https://www.sciencedirect.com/science/article/pii/0895435696000091. doi: 10.1016/0895-4356(96)00009-1.

REFERENCES

Alberta's Tomorrow Project. All about Alberta's Tomorrow Project; https://myatp.ca/about-atp. Updated 2019. Accessed January, 2019.

Anderson LN, Cotterchio M, Gallinger S. Lifestyle, dietary, and medical history factors associated with pancreatic cancer risk in Ontario, Canada. *Cancer Causes Control*. 2009;20(6):825-834. doi: //dx.doi.org/10.1007/s10552-009-9303-5.

Armstrong-Wells J, Goldenberg NA. Institution-based prospective inception cohort studies in neonatal rare disease research. *Seminars in Fetal and Neonatal Medicine*. 2011;16(6):355-358. doi: 10.1016/j.siny.2011.07.004.

Austin MA, Kuo E, Van Den Eeden, S. K., et al. Family history of diabetes and pancreatic cancer as risk factors for pancreatic cancer: The PACIFIC study. *Cancer Epidemiology Biomarkers and Prevention.* 2013;22(10):1913-1917. doi: 10.1158/1055-9965.EPI-13-0518.

Ben Q, Xu M, Ning X, et al. Diabetes mellitus and risk of pancreatic cancer: A meta-analysis of cohort studies. *Eur J Cancer*. 2011;47(13):1928-1937. doi: 10.1016/j.ejca.2011.03.003.

Bergmann MM, Calle EE, Mervis CA, Miracle-McMahill HL, Thun MJ, Heath CW. Validity of self-reported cancers in a prospective cohort study in comparison with data from state cancer registries. *Am J Epidemiol*. 1998;147(6):556-562.

Blay J, Coindre J, Ducimetière F, Ray-Coquard I. The value of research collaborations and consortia in rare cancers. *Lancet Oncology, The*. 2016;17(2):69. doi: 10.1016/S1470-2045(15)00388-5.

Boffetta P, Colditz GA, Potter JD, et al. Cohorts and consortia conference: A summary report (Banff, Canada, June 17-19, 2009). *Cancer Causes Control*. 2011;22(3):463-468. https://www.jstor.org/stable/41485113. doi: 10.1007/s10552-010-9717-0.

Bosetti C, Lucenteforte E, Silverman DT, et al. Cigarette smoking and pancreatic cancer: An analysis from the International Pancreatic Cancer Case-Control Consortium (Panc4). *Ann Oncol.* 2012;23(7):1880-1888. doi: 10.1093/annonc/mdr541.

Boyd N, Dancey JE, Gilks CB, Huntsman DG. Rare cancers: A sea of opportunity. *Lancet Oncology*. 2016;17(2):61. doi: 10.1016/S1470-2045(15)00386-1.

Bracci PM. Obesity and pancreatic cancer: Overview of epidemiologic evidence and biologic mechanisms. *Molecular Carcinogenesis*. 2012;51(1):53-63. https://onlinelibrary.wiley.com/doi/abs/10.1002/mc.20778. doi: 10.1002/mc.20778.

Brown C. Barriers to accessing data are bad medicine. *CMAJ*. 2014;186(16):1203. doi: 10.1503/cmaj.109-4894.

Canadian Cancer Society's Advisory Committee on Cancer Statistics. Canadian Cancer Statistics 2017. cancer.ca/Canadian-CancerStatistics-2017-EN.pdf. Updated 2017. Accessed February, 2018.

Canadian Center on Substance Abuse and Addiction. Drinking guidelines. http://www.ccdus.ca/Eng/topics/alcohol/drinking-guidelines/Pages/default.aspx. Updated 2018. Accessed February, 2019.

Canadian Partnership for Tomorrow Project. About. https://www.partnershipfortomorrow.ca/about/. Accessed January, 2019.

Cancer Control Alberta. Surveillance and reporting: The 2019 report on cancer statistics in Alberta*.* *Alberta Health Services*. 2019.

Cancer Statistics Branch, Division of Cancer Control and Population Sciences Surveillance, Epidemiology and End Results Program. Conversion of neoplasms by topography and morphology from the International Classification of Diseases for Oncology, second edition to International Classification of Diseases for Oncology, third edition; 2001.

Chang S, Long SR, Kutikova L, et al. Estimating the cost of cancer: Results on the basis of claims data analyses for cancer patients diagnosed with seven types of cancer during 1999 to 2000. *Journal of Clinical Oncology*. 2004;22(17):3524-3530. doi: 10.1200/JCO.2004.10.170.

Cho LY, Kim C, Li L, et al. Validation of self-reported cancer incidence at follow-up in a prospective cohort study. *Ann Epidemiol*. 2009;19(9):644-646. doi: //dx.doi.org/10.1016/j.annepidem.2009.04.011.

DeSantis CE, Kramer JL, Jemal A. The burden of rare cancers in the United States. CA: *A Cancer Journal for Clinicians.* 2017;67(4):261-272. doi: 10.3322/caac.21400.

Doiron D, Raina P, Fortier I. Linking Canadian population health data: Maximizing the potential of cohort and administrative data*. Can J Public Health*. 2013;104(3):261.

Dong J, Zou J, Yu XF. Coffee drinking and pancreatic cancer risk: A meta-analysis of cohort studies. *World J Gastroenterol*. 2011;17(9):1204-1210. doi: 10.3748/wjg.v17.i9.1204.

Dummer TJ, Awadalla P, Boileau C, et al. The Canadian Partnership for Tomorrow Project: A pan-Canadian platform for research on chronic disease prevention. *CMAJ*. 2018;190(23):717. doi: 10.1503/cmaj.170292.

Ernster VL. Nested case-control studies. *Prev Med*. 1994;23(5):587-590. doi: 10.1006/pmed.1994.1093.

Flook R, van Zanten SV. Pancreatic cancer in Canada: Incidence and mortality trends from 1992 to 2005. *Can J Gastroenterol*. 2009;23(8):546-550.

Gatta G, van der Zwan, Jan Maarten, Casali PG, et al. Rare cancers are not so rare: The rare cancer burden in Europe. *Eur J Cancer.* 2011;47(17):2493-2511. doi: //dx.doi.org/10.1016/j.ejca.2011.08.008.

Genkinger JM, Spiegelman D, Giles GG, et al. A pooled analysis of 14 cohort studies of anthropometric factors and pancreatic cancer risk. *International Journal of Cancer (Print).* 2011(7):1708.

Ghadirian P, Nkondjock A. Consumption of food groups and the risk of pancreatic cancer: A case-control study. *Journal of Gastrointestinal Cancer*. 2010;41(2):121-129. doi: //dx.doi.org/10.1007/s12029-009-9127-2.

Government of Canada, Health Canada. Recalls and safety alerts: Health Canada reminds Canadians to manage their caffeine consumption. http://healthycanadians.gc.ca/recall-alert-rappel-avis/hc-sc/2013/34021a-eng.php. Updated 2013. Accessed February, 2019.

Greenlee RT, Goodman MT, Lynch CF, Platz CE, Havener LA, Howe HL. The occurrence of rare cancers in U.S. adults, 1995-2004. *Public Health Reports.* 2010;125(1):28-43.

Guertin KA, Freedman ND, Loftfield E, Stolzenberg-solomon RZ, Graubard BI, Sinha R. A prospective study of coffee intake and pancreatic cancer: Results from the NIH-AARP diet and health study. *The British Journal of Cancer*. 2015;113(7):1081-1085. https://www.ncbi.nlm.nih.gov/pubmed/26402414. doi: 10.1038/bjc.2015.235.

Hanley AJG, Johnson KC, Villeneuve PJ, Mao Y, Canadian Cancer Registries Epidemiology Research Group. Physical activity, anthropometric factors and risk of pancreatic cancer: Results from the Canadian Enhanced Cancer Surveillance System. *International Journal of Cancer*. 2001(1):140.

Hart AR, Kennedy H, Harvey I. Pancreatic cancer: A review of the evidence on causation. *Clinical Gastroenterology and Hepatology*. 2008;6(3):275-282. https://www.clinicalkey.es/playcontent/1-s2.0-S154235650701244X. doi: 10.1016/j.cgh.2007.12.041.

Heinen MM, Verhage BAJ, Goldbohm RA, van den Brandt PA. Meat and fat intake and pancreatic cancer risk in the Netherlands Cohort Study. *International Journal of Cancer*. 2009;125(5):1118-1126. doi: https://doi.org/10.1002/ijc.24387

Ilic M, Ilic I. Epidemiology of pancreatic cancer. *World Journal of Gastroenterology*. 2016;22(44):9694-9705. doi: 10.3748/wjg.v22.i44.9694.

Inoue M, Sawada N, Shimazu T, et al. Validity of self-reported cancer among a Japanese population: Recent results from a population-based prospective study in Japan (JPHC study). *Cancer Epidemiology*. 2011;35(3):250-253. doi: 10.1016/j.canep.2010.12.002.

Iodice S, Gandini S, Maisonneuve P, Lowenfels AB. Tobacco and the risk of pancreatic cancer: A review and meta-analysis. *Langenbecks Arch Surg*. 2008;393(4):535-545. doi: 10.1007/s00423-007-0266-2.

Jacobs EJ, Chanock SJ, Fuchs CS, et al. Family history of cancer and risk of pancreatic cancer: A pooled analysis from the Pancreatic Cancer Cohort Consortium (PanScan). *International Journal of Cancer*. 2010;127(6):1421-1428. doi: 10.1002/ijc.25148.

Kawai A, Goto T, Shibata T, et al. Current state of therapeutic development for rare cancers in Japan, and proposals for improvement. *Cancer Sci.* 2018;109(5):1731-1737. doi: 10.1111/cas.13568.

Ko AH, Wang F, Holly EA. Pancreatic cancer and medical history in a population-based case-control study in the San Francisco Bay Area, California. *Cancer Causes & Control*. 2007;18(8):809-819. doi: 10.1007/s10552-007-9024-6.

Komatsubara KM, Carvajal RD. The promise and challenges of rare cancer research. *Lancet Oncol*. 2016;17(2):136-138. doi: //dx.doi.org/10.1016/S1470-2045(15)00485-4.

Kreiger N, Lacroix J, Sloan M. Hormonal factors and pancreatic cancer in women. *Ann Epidemiol*. 2001;11(8):563-567.

Kutikova L, Bowman L, Chang S, Long SR, Thornton DE, Crown WH. Utilization and cost of health care services associated with primary malignant brain tumors in the United States. *Journal of Neuro-Oncology.* 2007;81(1):61.

Kuzmickiene I, Everatt R, Virviciute D, et al. Smoking and other risk factors for pancreatic cancer: A cohort study in men in Lithuania. *Cancer Epidemiology*. 2013;37(2):133-139. doi: 10.1016/j.canep.2012.10.001.

Larsson SC, Orsini N, Wolk A. Body mass index and pancreatic cancer risk: A meta-analysis of prospective studies. *International Journal of Cancer*. 2007;120(9):1993-1998. https://onlinelibrary.wiley.com/doi/abs/10.1002/ijc.22535. doi: 10.1002/ijc.22535.

Larsson SC, Wolk A. Red and processed meat consumption and risk of pancreatic cancer: Meta-analysis of prospective studies. *British Journal of Cancer*. 2012;106(3):603-607.

Li J, Cone JE, Alt AK, et al. Performance of self-report to establish cancer diagnoses in disaster responders and survivors, World Trade Center Health Registry, New York, 2001-2007. *Public Health Rep.* 2016;131(3):420-429.

Lindsted KD, Fraser GE, Steinkohl M, Beeson WL. Healthy volunteer effect in a cohort study: Temporal resolution in the Adventist Health Study. *Journal of Clinical Epidemiology.* 1996;49(7):783-790. https://www.sciencedirect.com/science/article/pii/0895435696000091. doi: 10.1016/0895-4356(96)00009-1.

Liu SZ, Chen WQ, Wang N, Yin MM, Sun XB, He YT. Dietary factors and risk of pancreatic cancer: A multi-centre case-control study in China. *Asian Pacific Journal of Cancer Prevention*. 2014;15(18):7947-7950.

Loh V, Harding J, Koshkina V, Barr E, Shaw J, Magliano D. The validity of self-reported cancer in an Australian population study. *Australia NZ J Public Health*. 2014;38:35-38. doi: 10.1111/1753-6405.12164.

Longabaugh M. Patient perspective and personal journey of treating a "Rare cancer". *Surg Oncol Clin N Am*. 2017;26(1):1-7. doi: 10.1016/j.soc.2016.07.014.

Lowenfels AB, Maisonneuve P. Epidemiology and risk factors for pancreatic cancer. *Best Practice & Research Clinical Gastroenterology*. 2006;20(2):197-209. https://www.sciencedirect.com/science/article/pii/S152169180500154X. doi: 10.1016/j.bpg.2005.10.001.

Lucenteforte E, La Vecchia C, Silverman D, et al. Alcohol consumption and pancreatic cancer: A pooled analysis in the International Pancreatic Cancer Case-Control Consortium (PanC4). *Annals of Oncology*. 2012;23(2):374-382. https://www.ncbi.nlm.nih.gov/pubmed/21536662. doi: 10.1093/annonc/mdr120.

Maisonneuve P, Lowenfels AB, Bueno-de-Mesquita HB, et al. Past medical history and pancreatic cancer risk: Results from a multicenter case-control study. *Ann Epidemiol*. 2010;20(2):92-98. doi: //dx.doi.org/10.1016/j.annepidem.2009.11.010.

Maisonneuve P, Lowenfels AB. Risk factors for pancreatic cancer: A summary review of meta-analytical studies. *International Journal of Epidemiology*. 2015;44(1):186-198. https://www.ncbi.nlm.nih.gov/pubmed/25502106. doi: 10.1093/ije/dyu240.

Maplethorpe E, Walker EV, Davis FG. Occurrence of rare cancer in Canada: The distribution of cancer incidence among the Canadian population from 2009-2013. *Poster session presented at: Canadian Research Data Center Network Conference, Hamilton, ON*. 2018.

Marley AR, Fan H, Hoyt ML, Anderson KE, Zhang J. Intake of methyl-related nutrients and risk of pancreatic cancer in a population-based case-control study in Minnesota. *European Journal of Clinical Nutrition*. 2018;72(8):1128-1135. doi: 10.1038/s41430-018-0228-5.

Mathoulin-Pélissier S, Pritchard-Jones K. Evidence-based data and rare cancers: The need for a new methodological approach in research and investigation. *European Journal of Surgical Oncology*. 2018;45(1):22-30. doi: 10.1016/j.ejso.2018.02.015.

National Cancer Institute. ICD-O-3 SEER site/histology validation. https://seer.cancer.gov/icd-o-3/sitetype.icdo3.d20150918.pdf. Updated 2015.

Navarro C, Chirlaque MD, Tormo MJ, et al. Validity of self reported diagnoses of cancer in a major Spanish prospective cohort study. *Journal Epidemiology Community Health*. 2006;60:593-599.

Negrouk A, Lacombe D, Trimble EL, Seymour M. Clinical research for rare cancers: Is it a reality in the global regulatory landscape? The International Rare Cancer Initiative. *Expert Opinion on Orphan Drugs*. 2014;2(5):433-440. doi: 10.1517/21678707.2014.888948.

North American Association of Central Cancer Registries. Certification criteria. https://www.naaccr.org/certification-criteria/. Updated 2018. Accessed July, 2019.

Olson SH, Hsu M, Satagopan JM, et al. Allergies and risk of pancreatic cancer: A pooled analysis from the Pancreatic Cancer Case-Control Consortium. *Am J Epidemiol.* 2013;178(5):691-700. doi: 10.1093/aje/kwt052.

Parikh-Patel A, Allen M, Wright WE. Validation of self-reported cancers in the California Teachers Study. *Am J Epidemiol*. 2003;157(6):539-545.

Pearce N. What does the odds ratio estimate in a case-control study? *International Journal of Epidemiology.* 1993;22(6):1189-1192. doi: 10.1093/ije/22.6.1189.

Permuth-Wey J, Egan KM. Family history is a significant risk factor for pancreatic cancer: Results from a systematic review and meta-analysis. *Familial Cancer*. 2009;8(2):109-117. doi: 10.1007/s10689-008-9214-8.

Pillai RK, Jayasree K. Rare cancers: Challenges & issues. *Indian Journal of Medical Research*. 2017;145(1):17-27. doi: 10.4103/ijmr.IJMR_915_14.

Rahman F, Cotterchio M, Cleary SP, Gallinger S. Association between alcohol consumption and pancreatic cancer risk: A case-control study. *PLoS ONE*. 2015;10(4):e0124489. doi: //dx.doi.org/10.1371/journal.pone.0124489.

Robson PJ, Solbak NM, Haig TR, et al. Design, methods and demographics from phase I of Alberta's Tomorrow Project cohort: A prospective cohort profile. *Canadian Medical Association Journal Open*. 2016;4(3):527. https://www.clinicalkey.es/playcontent/1-s2.0-S2291002616301345. doi: 10.9778/cmajo.20160005.

Sandrucci S, Naredi P, Bonvalot S. Centers of excellence or excellence networks: The surgical challenge and quality issues in rare cancers. *European Journal of Surgical Oncology*. 2019;45(1):19-21.

Shim J, Oh K, Kim HC. Dietary assessment methods in epidemiologic studies. *Epidemiology and health*. 2014;36:e2014009. doi: 10.4178/epih/e2014009.

Statistics Canada. Canadian Cancer Registry (CCR). http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3207. Updated 2019. Accessed April, 2019.

Statistics Canada. Current smoking trends. Health at a glance; 2015. https://www150.statcan.gc.ca/n1/pub/82-624-x/2012001/article/11676-eng.htm.Accessed June 2019.

Statistics Canada. Vital statistics - death database (CVSD). http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3233. Updated 2018.

Stavrou E, Vajdic CM, Pearson S, Loxton D. The validity of self-reported cancer diagnoses and factors associated with accurate reporting in a cohort of older Australian women. *Cancer Epidemiology*. 2011;35(6):80. doi: 10.1016/j.canep.2011.02.005.

Streiner DL, Norman GR, Cairney J. *Health measurement scales*. 5th ed. Oxford: Oxford Univ. Press; 2015.

Villeneuve P, Johnson K, Hanley A, Mao Y, Canadian Cancer Registries Epidemiology Research Group. Alcohol, tobacco and coffee consumption and the risk of pancreatic cancer:

Results from the Canadian Enhanced Surveillance System Case-Control Project. *Eur J Cancer Prev.* 2000;9(1):49-58.

von der Schulenburg JM, Pauer F. Reviews: Rare cancers—Rarity as a cost and value argument. *Journal of Cancer Policy*. 2017;11:54-59. doi: 10.1016/j.jcpo.2016.09.004.

Wagland K, Levesque JV, Connors J. Disease isolation: The challenges faced by mothers living with multiple myeloma in rural and regional Australia. *European Journal of Oncology Nursing*. 2015;19(2):148-153. doi: 10.1016/j.ejon.2014.10.003.

Walker EV, Maplethorpe E, Davis FG. Common and rare cancer incidence rates in the Canadian population: 2009-2013; 2019. In preparation for submission at the time of this thesis.

World Cancer Research Fund. Recommendations and public health and policy implications. 2018.

World Health Organization. Body mass index - BMI. http://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi. Accessed February, 2019.

Yallew W, Bamlet WR, Oberg AL, et al. Association between alcohol consumption, folate intake, and risk of pancreatic cancer: A case-control study. *Nutrients*. 2017;9(5):448. doi: 10.3390/nu9050448.

Yoshinaga A, Sasaki S, Tsugane S. Sensitivity of self-reports of cancer in a population-based prospective study: JPHC study cohort I. *Journal of Clinical Epidemiology*. 2001;54(7):741-746.

Zeig-Owens R, Kablanian A, Webber MP, et al. Agreement between self-reported and confirmed cancer diagnoses in New York City firefighters and EMS workers, 2001-2011. *Public Health Rep*. 2016;131(1):153-159.

APPENDICES

Appendix 1: Supplementary Tables

Table S1. Summary of literature on self-reported cancer diagnosis sensitivity and positive predictive value (PPV) using population-based cancer registries as the gold standard.

Author Year	Country	Population	Overall Sensitivity (%)	Overall PPV (%)	Site-Specific Sensitivities (%)	Site-Specific PPV's (%)
Navarro et al. 2006	Spain	Prospective cohort study (EPIC, adults aged 29-69)	57.5%	70.8%	Breast (84.5), thyroid (61.9), lung (50.0), stomach (50.0), oral cavity (23.5), bladder (21.7), colorectal (17.4), corpus uteri (15.0), cervix uteri (13.2)	Stomach (100), bladder (100), thyroid (86.7), breast (82.6), colo-rectal (80.0), lung (57.1), cervix uteri (50.0), uterus (41.8), oral cavity (40.0), corpus uteri (37.5)
Stavrou et al. 2011	Australia	Prospective cohort study (ALSWH, women aged 70 and older)	89.2% (excluding melanoma), 88.3% (incident cases)	66.5% (excluding melanoma) 80.2% (incident cases)	Prevalent cases: lung (100), breast (93.1), colorectal (90.0), cervical (50) Incident cases: Breast (82.6), colorectal (76.2)	Prevalent cases: colorectal (64.3), breast (59.4) Incident cases: breast (77.5), colorectal (72.7)
Parikh-Patel et al. 2003	US	Prospective cohort study (California Teachers Study, adult females)	Not reported	Not reported	Breast (96.4), thyroid (92.9), ovary (85.9), hodgkins lymphoma (84.0), colon/ rectum (82.5), leukemia (82.0), lung (80.0), melanoma (73.2), cervix (44.3), endometrial (69.1), other skin (53.6)	Not reported
Loh et al. 2014	Australia	Prospective cohort study (AusDiab, adults aged 25 and older)	71.1%	65.7%	Breast (90.7), bowel (77.8), prostate (77.1), melanoma (36.9)	Breast (72.1), bowel (70.0), prostate (70.0), melanoma (60.5)
Bergmann et al. 1998	US	Prospective cohort study (CPS-II Nutrition Survey, adults)	93% (any cancer), 79% (exact type & +/- 1 year diagnosis)	75%	Breast (91), prostate (90), lung (90), colon (85), uterus (71), bladder (67), non-hodgkins lymphoma (64), leukemia (61), melanoma (53), rectal (16)	Breast (85), prostate (80), uterus (79), lung (72), bladder (72), rectum (71), non-hodgkins lymphoma (69), colon (54), leukemia (41), melanoma (34)

Li et al. 2016	US	Prospective cohort study (WTCHR, World Trade Center disaster responders and survivors aged 18 and older)	83.9%	54.9%	Pancreas (90.9), multiple myeloma (84.6), testis (82.4), lymphoma/leukemia (77.1), prostate (76.1), breast (75.1), thyroid (73.3), lung (66.7), bladder (63.8), corpus uterus (63.6), colorectal (59.7), kidney (58.7), melanoma (51.5), ovary (50.0), stomach (46.2), brain/NS (31.8), oral cavity/pharynx (22.2)	Multiple myeloma (100), prostate (93.5), testis (93.3), breast (86.2), pancreas (83.3), bladder (81.1), kidney (77.1), lung (76.5), lymphoma/ leukemia (71.8), thyroid (69.8), corpus uterus (65.6), colorectal (62.2), brain/NS (58.3), oral cavity/pharynx (54.6), stomach (50.0), ovary (42.1), melanoma (40.5)
Zeig-Owens et al. 2016	US	New York Fire Department and EMS Workers	90.3%	69.1%	Esophageal/gastric (100), lung (95.8), prostate (94.6), thyroid (92.5), testicular (92.0), oral/ nasal/throat (89.7), bladder/ kidney (87.9), melanoma (82.1), colorectal (80.0), lymphoma (79.3), ovarian/ uterine/cervical (66.7) hematologic (64.9), bone/ sarcoma (61.5), brain/CNS (33.3)	Prostate (94.2), bladder/kidney (85.0), thyroid (82.2), testicular (82.1), lymphoma (73.0), esophageal/gastric (63.2), colorectal (58.1), lung (57.5), oral/nasal/throat (56.5), hematologic (51.1), ovarian/ uterine/cervical (28.6), brain/ CNS (28.6), melanoma (26.3), bone/sarcoma (16.3)
Inoue et al. 2011	Japan	Prospective cohort study (JPHC, adults aged 40-69)	52.6%	59.7%	Breast (82.4), stomach (61.9), uterus (59.1), lung (56.5), colorectal (38.3), liver (33.6)	Breast (58.4), stomach (51.6), colorectal (47.1), lung (45.5), liver (30.7), uterus (21.7)
Yoshinaga et al. 2001	Japan	Prospective cohort study (JPHC cohort I, adults aged 40-65)	36%	Not reported	Breast (81), uterus (42), stomach (41), lung (26), colorectal (14), liver (8)	Not reported
Cho et al. 2017	Korea	Prospective cohort study (HEXA, adults aged 35-79)	72.0%	81.9%	Breast (81.2), stomach (78.0), thyroid (69.3), prostate (67.0), lung (65.6), colon/rectum (57.9), bladder (56.0), liver (53.7), cervix uteri (52.1)	Thyroid (96.1), prostate (96.1), colon/rectum (94.3), stomach (93.0), lung (89.7), liver (84.0), breast (80.8), bladder (70.4), cervix uteri (43.7)

Table S2. SEER 2018 cancer categories based on ICD-O-3 topography, corresponding self-report accuracy type, and combined groupings for reporting purposes.

ICD_0-3 (Cxxx) Topography	SEER 2018 Cancer Type(ref)	Self-report Accuracy Type	Combined Groups
C000-C006, C008-C009	Lip	Lip	Oral/ Respiratory
C019-C024, C028-C029	Tongue	Tongue	
C030-C031, C039-C041,C048-C052,C058-C062, C068-C069	Gum, Floor of Mouth, & Other Mouth	Gum, Floor of Mouth, & Other Mouth	
C079-C081,C088-C089	Salivary Gland	Salivary Gland	
C090-C091,C098-C104,C108-C109	Oropharynx	Throat	
C110-C113,C118-C119	Nasopharynx		
C129-C132,C138-C139	Hypopharynx		
C140,C142,C148	Pharynx		
C150-C155,C158-C159	Esophagus	Esophagus	Digestive/ Hepatic
C160-C166,C168-C169	Stomach	Stomach	
C170-C173,C178-C179	Small Intestine	Small Intestine	
C180, C182-C189, C199	Large Intestine	Large Intestine	
C181	Appendix	Appendix	
C209	Rectum	Rectum	
C210-C212,C218	Anal Canal & Anus	Anal Canal & Anus	
C260,C268-C269	Unspecified Digest. Organs	Unspecified Digest. Organs	
C220	Liver	Liver	
C221	Intrahepatic Bile Duct	Intrahepatic Bile Duct	
C239-C241,C248-C249	Gallbladder & Extrahepatic Bile Duct	Gallbladder & Extrahepatic Bile Duct	
C250-C254,C257-C259	Pancreas	Pancreas	
C300	Nasal Cavity (Including Nasal Cartilage)	Nasal Cavity (Including Nasal Cartilage)	Oral/ Respiratory
C301, C310-C313,C318-C319	Accessory, Sinuses, Middle & Inner Ear	Accessory, Sinuses, Middle & Inner Ear	
C320-C323,C328-C329	Larynx	Larynx	
C339	Trachea	Trachea	
C340-C343,C348-C349	Lung & Bronchus	Lung & Bronchus	
C379	Thymus	Thymus	
C380	Heart	Heart	
C381-C383, C388	Mediastinum	Mediastinum	
C384	Pleura	Pleura	
C390,C398-C399	Respiratory, NOS	Respiratory, NOS	
C400-C403,C408-C414,C418-C419	Bones & Joints	Bones & Joints	Other
C420, C421, C424	Blood, Bone Marrow & Hematopoietic Sys	Blood, Bone Marrow & Hematopoietic Sys	Blood/ Hemato-poietic
C422	Spleen	Spleen	
C423	Reticulo-Endothelial	Reticulo-Endothelial	
C440-C449	Skin-US SEER Definition	Skin-US SEER Definition	Skin
C470-C476,C478-C479,C490-C496,C498-C499	Connective & Soft Tissue	Connective & Soft Tissue	Other
C480-C482,C488	Retroperitoneum & Peritoneum	Retroperitoneum & Peritoneum	

C500-C506,C508-C509	Breast	Breast	Breast
C510-C512,C518, C529	Vagina & Labia	Vagina & Labia	Female Reproductive
C519	Vulva, NOS	Vulva, NOS	
C530-C531,C538-C539	Cervix Uteri	Cervix Uteri	
C540-C543,C548-C549	Corpus Uteri	Uterus/Endometrial	
C559	Uterus, NOS		
C569	Ovary	Ovary	
C570-C574,C577-C579	Other Female Genital	Other Female Genital	
C589	Placenta	Placenta	
C600-C602,C608-C609, C632	Penis & Scrotum	Penis & Scrotum	Male Reproductive
C619	Prostate Gland	Prostate Gland	
C620-C621,C629	Testis	Testis	
C630, C631, C637-C639	Epididymis, Spermatic Cord, Male Genital, NOS	Epididymis, Spermatic Cord, Male Genital, NOS	
C649	Kidney	Kidney	Urinary
C659, C669	Renal Pelvis, Ureter	Renal Pelvis, Ureter	
C670-C679	Urinary Bladder	Urinary Bladder	
C680-C681,C688-C689	Other Urinary Organs	Other Urinary Organs	
C690-C691, C693, C695-C698	Orbit & Lacrimal Gland (Excl. Retina, Eye, NOS)	Orbit & Lacrimal Gland (Excl. Retina, Eye, NOS)	CNS/Eye
C692	Retina	Retina	
C694	Eyeball	Eyeball	
C699	Eye, NOS	Eye, NOS	
C700-C701,C709	Meninges (Cerebral, Spinal)	Meninges (Cerebral, Spinal)	
C710-C714, C717-C719, C720-C725	Brain, Cranial Nerves, & Spinal Cord (Excl. Ventricle,	Brain, Cranial Nerves, & Spinal Cord (Excl. Ventricle,	
C715	Ventricle	Ventricle	
C716	Cerebellum	Cerebellum	
C728-C729	Nervous	Nervous	
C739	Thyroid	Thyroid	Endocrine
C740-C741,C749	Adrenal Glands	Adrenal Glands	
C750	Parathyroid	Parathyroid	
C751	Pituitary Gland	Pituitary Gland	
C753	Pineal Gland	Pineal Gland	
C754-C755,C758-C759	Other Endocrine Glands	Other Endocrine Glands	
C760-C768	Ill-Defined	Ill-Defined	Other
C770-C775,C778-C779	Lymph Nodes	Lymph Nodes	Lymphatic
C809	Unknown	Unknown	Other
--	--	Histology Only, site unclear[a]	
--	--	Site Unclear, histology not specified[a]	
--	--	Type Missing[a,b]	

SEER= Surveillance, Epidemiology and End Results Program, ICD-O-3=International Classification of Diseases for Oncology, Third Edition

[a] Sites generated to mark self-reported cancer types that did not fit into a site category, either because they specified histology only, were unclear or too broad, or type was missing.

[b] For self-reported type that was missing if the participant had self-reported they were diagnosed with cancer.

Table S3. Summary of literature on risk factors of pancreatic cancer (PC).

Author	Study Design[a]	Population	Results[b]		Summary	Comments
Year			Factor	Association		
Maisonneuve et al. 2015	Review of 117 meta-analyses and pooled reports (case-control and cohort studies)	Adults	Smoking	Moderate risk (RR 1.5-1.9)	↑ risk of PC: -smoking, diabetes, increased BMI, heavy alcohol intake, red/processed meat consumption, pancreatitis, idiopathic thrombosis, ↓ risk of PC: -Fruit/vegetables consumption, allergies	Categorized by RR based on average of associations found in literature. Other factors explored but not included in this table (not lifestyle related or not explored in this analysis) Some articles included multiple cancer sites.
			Family history of PC	Moderate risk (RR 1.5-1.9)		
			Diabetes/metabolic syndrome/ use of antidiabetic drugs (except metformin)	Moderate risk (RR 1.5-1.9)		
			BMI	Low risk (RR 1.1-1.4)		
			Alcohol (heavy intake)	Low risk (RR 1.1-1.4)		
			Fruit/Vegetables	Protective (RR 0.5-0.9)		
			Red/Processed Meat	Low risk (RR 1.1-1.4)		
			Allergies	Protective (RR 0.5-0.9)		
			Chronic pancreatitis	High risk (RR≥2)		
			Idiopathic thrombosis	High risk (RR≥2)		
			Coffee, tea	NS		
Iodice et al. 2008	Meta-analysis of 82 case-controls and cohorts	Adults	Smoking	Current: RR=1.74 Former: RR=1.20 (compared to never smokers)	↑ risk of PC: -smoking	
Bosetti et al. 2012	Pooled analysis of 12 case-controls	Adults	Smoking	Current: OR=2.20 Former: OR=1.17 (compared to never)	↑ odds of PC: -smoking	Dose-response relationship for cigarettes/day.
Permuth-Wey et al. 2009	Meta-analysis of 9 case-controls and cohorts	Adults	Family history of PC	Family history: RR=1.80 (compared to no family history)	↑ risk of PC: -family history	Did not have to be first degree relative for overall estimate
Jacobs et al. 2010	Pooled analysis of 11 case-control and cohort studies	Adults	Family history of PC	Family history in first degree relative: OR=1.76	↑ odds of PC: -family history	Family history of prostate cancer also increased odds of PC
Ben et al. 2011	Meta-analysis of 35 cohort studies	Adults	Diabetes	Diabetes mellitus: RR=1.94		
Larsson et al. 2007	Meta-analysis of 21 prospective studies	Adults	BMI	Per 5kg/m^2 increase: RR=1.12	↑ risk of PC: -increasing BMI	

Genkinger et al. 2011	Pooled analysis of 14 cohort studies	Adults	BMI	Obese (≥30kg/m^2): RR=1.47 (compared to 21-22.9 kg/m^2)	↑ risk of PC: -increased BMI	BMI at baseline, also looked at BMI in early adulthood.
Lucenteforte et al. 2012	Pooled analysis of 10 case-controls	Adults	Alcohol	Heavy (≥9drinks/day): OR=1.60 (compared to none or occasional)	↑ odds of PC: -heavy alcohol drinking	
Dong et al. 2011	Meta-analysis of 14 cohort studies	Adults	Coffee	Regular drinker: RR=0.82 (pooled) Low/Moderate drinker: RR=0.86 High drinker: RR=0.68 (compared too little to no coffee)	↓ risk of PC: -Coffee drinking	In subgroup analyses, association significant in men but not in women.
Larsson et al. 2012	Meta-analysis of 11 prospective studies	Adults	Red meat	For 120g/day increase: RR=1.29 in males, NS in females.	↑ risk of PC: -processed meat, red meat in males	
			Processed meat	For 50g/day increase: RR=1.19		
Kuzmickiene et al. 2013	Cohort, prospective	Adult men in Lithuania	Smoking	Current smoking: HR=1.79 (compared to neve smoking)	↑ risk of PC: -smoking	Other measures of smoking also reported.
			Alcohol, BMI,	NS		
Guertin et al. 2015	Cohort, prospective	Adults aged 50-71 years in USA	Coffee	NS		
Maisonneuve et al. 2010	Case-control, population-based	Adults in Australia, Canada, Netherlands, Poland	Diabetes	History of diabetes: OR=2.16 (compared to no history)	↑ odds of PC: -diabetes, pancreatitis, gallbladder condition, smoking, alcohol drinking, increased BMI ↓ odds of PC: -Allergies, increased education	Smoking, alcohol education, and BMI associated with PC but only looked at as confounders or EM's. EM of pancreatitis by smoking, alcohol drinking. EM of diabetes by BMI.
			Allergies	History of allergies: OR=0.64 Particularly eczema and asthma. (compared to no history)		
			Pancreatitis	History of pancreatitis: OR=4.68 (compared to no history)		
			Gallbladder Condition	History of condition: OR=1.42 (compared to no history)		
Anderson et al. 2009	Case-control, population-based	Adults <75 years of age in Ontario, Canada	Smoking	Current smoking: OR=3.24 Former smoking: NS (compared to never smoking)	↑ odds of PC: -current smokers, family history of PC, higher BMI, caffeine consumption, ↓ odds of PC: -Moderate alcohol consumption, fruit	EM by smoking status for caffeine, family history of PC, BMI, and fruit.
			Family history of PC	OR=4.16 (compared to no family history)		
			BMI	Overweight (25-29.9 kg/m^2): OR=1.77 Obese (>30): OR=3.51		

				(compared to normal [<25])	consumption, allergies, some university/college education	
			Alcohol	1-6 drinks/week: OR=0.50 ≥7 drinks/week: NS (compared to <1 drink/week)		
			Caffeine	1-2 drinks/day: OR=2.37 ≥3/day: OR=2.29 (compared to <1/day)		
			Fruit/Vegetables	8-14 fruit servings/week: OR=0.59 >14 fruit servings/week OR=0.54 (compared to ≤7/week)		
			Allergies	Allergies/hay fever: OR=0.40 (compared to no allergies)		
			Education	Some college/university: OR=0.56 College/university grad: NS (compared to high school only)		
			Diabetes, Red meat (servings/week), vegetable consumption	NS		
Rahman et al. 2015	Case-control, population-based	Adults <89 years of age in Ontario, Canada	Smoking	Current smoking: OR=1.9 (compared to never smokers)	↑ odds of PC: -smoking, family history of PC, diabetes, pancreatitis ↓ odds of PC: -Drinking several alcohol types	OR's univariate (age-adjusted), but confounding was tested.
			Family history of PC	Family history: OR=2.4 (compared to no family history)		
			Alcohol	Drinking several alcohol types: OR=0.67 (compared to never drinking any alcohol) Amount of alcohol: NS		
			Diabetes	Diabetes: OR=1.7 (compared to no diabetes)		
			Pancreatitis	Pancreatitis: OR=2.4 (compared to no pancreatitis)		
			BMI	NS		
Austin et al. 2013	Case-control	Adults in Seattle and California, USA	Family history of PC	Any first degree relative: OR=2.79 Parent or sibling: OR=2.63 (compared to no family history)	↑ odds of PC: -family history of PC, family history of diabetes	Proportion with personal history of diabetes higher in cases than controls, but association not explored.
			(Family history of) Diabetes	Any first degree relative: OR=1.37 Parent of sibling: OR=1.34 Offspring: OR=1.95		

				(compared to no family history)		
Hanley et al. 2001	Case-control, population-based	Adult men and women in Canada	Smoking	15 to <25 cigarette pack-years: OR 1.95 in females ≥25 pack-years: OR=2.38 in females (compared to <5 pack-years) NS in men.	↑ odds of PC: -Heavy smoking in females, high BMI in males. ↓ odds of PC: - light alcohol consumption in females, % change in weight in males, high physical activity in males.	Models fit for males and females separately because of reproductive and hormonal factors.
			BMI	BMI ≥28.3: OR=1.90 in males (compared to <23.7) NS in females.		
			Alcohol	Light (>0 to <3 drinks/week): OR=0.52 in females (compared to no drinks) NS in males.		
			% change in weight	2.9 to <5.7%: OR=0.35 in males ≥5.7%: OR=0.45 in males (compared to <2.9% change) NS in females.		
			Physical activity	High moderate/strenuous: OR=0.42/OR=0.53 in males (compared to very low levels) NS in females		
Liu et al. 2014	Case-control	Adults in China	Smoking	Smoker: OR=1.50 (compared to non-smoker)	↑ odds of PC: -smoking, diabetes, higher meat consumption ↓ odds of PC: -Higher fruit and vegetable consumption, tea, peanuts	
			Diabetes	History of diabetes: OR=2.69 (compared to no history)		
			Fruit	1-2 times/week: OR=1.73 (compared to ≥3 times/week)		
			Vegetables	1-2 times/week: OR=2.29 (compared to ≥3 times/week)		
			Meat	1-2 times/week: OR=0.59 (compared to ≥3 times/week)		
			Tea	Drinking tea: OR=0.49 (compared to not drinking tea)		
			Peanuts	1-2 times/week: OR=0.56 (compared to <1 time/week)		
			BMI, alcohol, coffee	NS		

Villeneuve et al. 2000	Case-control, population-based	Adults aged 30-76 in Canada	Smoking	≥35 cigarette pack-years: OR=1.46 (males), OR=1.84 (females) (compared to 0 pack-years)	↑ odds of PC: -smoking, some alcohol types (particularly in non-smokers)	Results presented separately for males and females
			Alcohol	>1 liquor drink/day: OR=1.83 in males (compared to 0-3 times/month)		
			Coffee, total alcohol	NS		
Yallew et al. 2017	Case-control	Adults in USA	Alcohol	NS (amount and type)	↓ odds of PC: -Folate	
			Folate (natural)	≥267.66 mcg/day: OR=0.41 (compared to <188.14 mcg/day)		
Heinen et al. 2009	Case-cohort	Adults aged 55-69 years in Netherlands	Red/Processed Meat	NS		Looked at other types of meat as well, all NS.
Ghadirian et al. 2010	Case-control	Adults aged 35-79 years in Montreal, Canada	Vegetables	Increased consumption decreases odds (highest quartile: OR=0.47)	↑ odds of PC: -meat ↓ odds of PC: -vegetables	Sausages/luncheon meats, beef NS
			Meat (lamb, veal, and game)	Increased consumption increase odds (highest quartile: OR=2.24)		

PC=Pancreatic Cancer, NS=Not significant, OR=Odds ratio, EM=Effect Modification, BMI=Body mass index

[a] Design as stated by the author.

[b] Results summarized for the factors looked at in each study. OR's are multivariate OR's, adjusting for other variables or confounding factors authors deemed relevant, unless otherwise stated in comments column. Only OR's that are statistically significant (in multivariate analysis) are included in table, but non-significant factors explored in this analysis are noted (NS). Factors that not explored in this thesis are noted only if statistically significant.

Appendix 2: Supplementary Figures

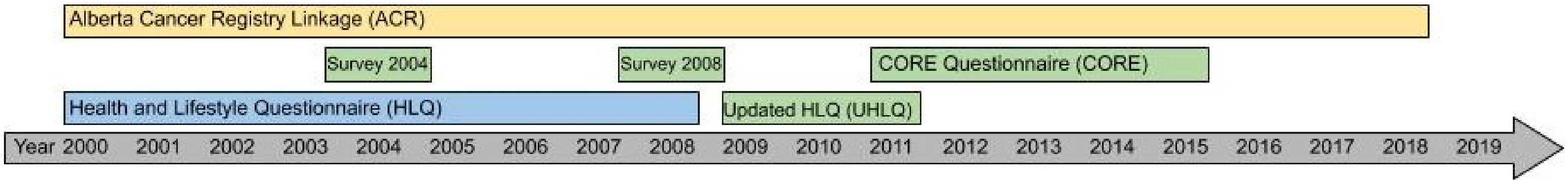


Figure S1. A timeline illustrating the years that Alberta's Tomorrow Project (ATP) surveys were administered. The baseline questionnaire (blue), the Health and Lifestyle Questionnaire (HLQ), was administered from 2000-2008; this was the enrollment phase for the study population in this project. Depending when they completed HLQ, participants had the opportunity to complete 4 follow-up questionnaires (green). Passive follow-up was completed by routine linkage to the Alberta Cancer Registry (ACR) (yellow), with the last linkage for this study population occurring in 2018.