Accurate Surveillance of Diabetes Mellitus in Nova Scotia within the General Population and the Five First Nations of Cape Breton

by

Roderick Clark

Submitted in partial fulfillment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
October 2011

DALHOUSIE UNIVERSITY

DEPARTMENT OF COMMUNITY HEALTH AND EPIDEMIOLOGY

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "ACCURATE SURVEILLANCE OF DIABETES MELLITUS IN NOVA SCOTIA WITHIN THE GENERAL POPULATION AND THE FIVE FIRST NATIONS OF CAPE BRETON" by Roderick Clark in partial fulfillment of the requirements for the degree of Master of Science.

Dated: October 3, 2011

Supervisor: _____

Readers: _____

_____

DALHOUSIE UNIVERSITY

DATE: October 3, 2011

AUTHOR: Roderick Clark

TITLE: Accurate Surveillance of Diabetes Mellitus in Nova Scotia within the General Population and the Five First Nations of Cape Breton

DEPARTMEMT OR SCHOOL: Department of Community Health & Epidemiology

DEGREE: MSc Convocation: May Year: 2012

_____

Signature of Author

TABLE OF CONTENTS

List of Tables....vi

List of Figures....vii

Abstract....viii

List of Abbreviations and Symbols Used....ix

Acknowledgements....x

Chapter 1: Introduction....1

Chapter 2: Literature Review....2

Importance of Disease Surveillance....2

Challenges in Current Approaches to Surveillance for Disease Management....5

New Directions/Approaches....12

Compensating for Estimation Error in surveillance methods....13

Models for Estimating Parameters of Interest in the Absence of a Gold-Standard....15

Estimation Methods....23

Chapter 3: Manuscript 1....24

Validation of an Administrative Case Definition for Identifying Cases of Diagnosed Diabetes within an Aboriginal Population at the Sub-Provincial Level....24

Introduction....24

Method....27

Data and Subjects....27

Statistical Analysis....29

Results....32

Discussion....37

Conclusion.....40

References.....40

Chapter 4: Manuscript 2.....43

Data Source Dependence in the Estimation of Diabetes Prevalence within the General Population of Nova Scotia.....43

Introduction.....43

Method.....46

Data and Subjects.....46

Measures.....47

Statistical Analysis.....49

Results.....52

Discussion.....55

Conclusion.....57

References.....57

Appendix A.....60

Chapter 5: Conclusion.....61

References.....63

LIST OF TABLES

Table 1. The effect of sensitivity, specificity, and true disease prevalence on estimates of disease prevalence using various types of surveillance data....10

Table 2-1 Administrative case definitions for non-gold standard data sources....29

Table 2-2. Percent Agreement and Kappa Statistics for Administrative Case Definition and EM Registry....34

Table 2-3. Estimates of Sensitivity for the EMR and Manitoba Administrative Case Definition....34

Table 2-4. Estimates of Specificity for the Manitoba Administrative Case Definition....34

Table 2-5. Bayesian Estimates of Diabetes prevalence Manitoba rule 2009....35

Table 2-6. Diabetes Prevalence Estimates from Cross Tabulation of EMR and Admin Data (Manitoba rule)....36

Table 2-7. Estimates of Positive Predictive Values for Administrative Case Definitions....36

Table 3-1. Data Sources and Description (DCPNS, 2009)....47

Table 3-2. Data Source Case Definitions....49

Table 3-3. Sensitivity of the administrative case definition and dependence estimates between administrative and gold standard case measures for individuals identified as gold standard cases of diabetes within the NSPP-S and DCPNS measures....53

Table 3-4. Sensitivity of Administrative case definition and conditional covariance estimates for individuals identified within the DCPNS measure....55

LIST OF FIGURES

Figure 1. One population with two independent diagnostics.....16

Figure 2. Two populations (different disease prevalence) with two independent diagnostics.....19

Figure 3. One population with three independent diagnostics.....20

Figure 4. Two Populations with two dependence diagnostics.....22

Figure 5. Two populations (with different disease prevalence) with two independent diagnostics.....30

ABSTRACT

Using administrative data and data sources which contained gold standard cases of diabetes, this thesis examined (1) the validity of commonly used administrative case definitions for identifying cases of diagnosed diabetes within an Aboriginal population, and (2) the effect of conditional covariance on parameter estimates of an administrative case definition used to identify cases of diagnoses diabetes within the general population of Nova Scotia. We found significant differences in the sensitivity and specificity of a commonly used administrative case when applied to an Aboriginal population at the sub-provincial level. For the general population of Nova Scotia, we found that including a parameter to estimate conditional covariance between data sources resulted in significant variation in sensitivity, specificity, and prevalence estimates as compared to a study which did not consider this parameter.

List of Abbreviations and Symbols Used

		True Condition +	True Condition -	
Test Outcome	+	True Pos	False Pos	PPV= TP/(TP + FP)
	-	False Neg	True Neg	NPV= TN/(FP + TN)
		Sensitivity= TP/(TP + FN)	Specificity= TN/(FP +TN)	Total

True positive	Individuals who have a condition of interest who are correctly identified
False positive	Individuals who do not have a condition of interest who are incorrectly identified as having the condition
True negative	Individuals who do not have a condition of interest who are correctly identified as not having the condition
False negative	Individuals who have a condition of interest who are incorrectly identified as not having the condition
Prevalence	The number of cases of a given disease or other attributes (e.g. drug use, obesity) that exist in a defined population at a specified time. It is also sometimes referred to as the prevalence number.
Sensitivity	The ability of a screening test to identify correctly those who have the disease. Sensitivity is measured as the percent of those with the disease who test positive for the disease on a screening test .
Specificity	The ability of a screening test to identify correctly those who do not have the disease. Specificity is measures as the percent of those without the disease who test negative for the disease on a screening test.

ACKNOWLEDGEMENTS

The data used in [part of] this research was made available by the Unama'ki Data Access Committee. Any opinions expressed by the authors do not necessarily reflect the opinion of the Unama'ki Data Access Committee.

The data used in [part of] this research was made available by the Diabetes Care Program of Nova Scotia. Any opinions expressed by the authors do not necessarily reflect the opinion of DCPNS.

Thanks to my thesis supervisors Dr. George Kephart and Dr. Pantelis Andreou for their support throughout this process.

Thanks to my numerous past academic mentors, especially Dr. Chris Stewart, Dr. Gayle MacDonald Dr. Sherry Stewart, Dr. Ingrid Sketris, Dr. Judith Fisher, and Dr. George Kephart for your guidance throughout my academic career.

Thanks for financial support throughout my Master's Program, especially the Canadian Agency for Drugs and Technology in Health (CADTH), the Initiative for Medication Management, Policy Analysis, Research & Training (IMPART), Nova Scotia Health Research Foundation (NSHRF), Nova Scotia Gaming Foundation (NSGF), Dr. Gordon Flowerdew, the Department of Community Health and Epidemiology, Dr. George Kephart, and the Tui'kn Partnership.

Most importantly, thanks to my family.

CHAPTER 1: INTRODUCTION

Administrative data is one of the most commonly used data sources for diagnosed diabetes surveillance within Canada. The main advantages of this data source are its population coverage and wide availability for disease surveillance systems. Despite the fact that diabetes coding in administrative data has been found to have high validity in many studies (Hux, 2002., Health Canada, 2003., Johnson, 2009) the magnitude of misclassification errors due to coding error remains a significant concern (Strom, 2001.; Feinstein, 1989; Kephart, 2004). This concern has prompted interest in both estimating the validity of administrative case definitions in the absence of a gold standard data source, and combining administrative data with data from other sources to improve the validity of diabetes surveillance.

Models developed by Hui and Walter (1980) and Bayesian methods to simultaneously adjust for and estimate sensitivity and specificity of data source case definitions can be used to improve the validity of diabetes surveillance systems which utilize administrative data. This thesis project uses these methods to estimate and adjust for error within administrative case definitions, and evaluates the utility of combining diabetes surveillance data from various sources to improve diabetes surveillance within Nova Scotia.

This thesis manuscript provides a review of the literature in Chapter 2. Chapters 3 and 4 present the results from two distinct studies, in the form of research articles, and Chapter 5 provides a conclusion of the full master's thesis project.

CHAPTER 2: LITERATURE REVIEW

Importance of Disease Surveillance

Disease surveillance systems have been in operation since epidemiologists first attempted to understand trends in disease progression. These systems are a central component of any modern, effective public health system. The Center for Disease Control defines epidemiologic surveillance as,

> "… the ongoing systematic collection, analysis, and interpretation of health data essential to the planning, implementation, and evaluation of public health practice, closely integrated with the timely dissemination of these data to those who need to know."(Thacker, 1988)

Effective disease surveillance is essential in planning appropriate response to disease, monitoring population health, and mitigating costs associated with disease. Choi (1998) argues that a modern surveillance system must be systematic, ongoing, and population-based. Such a system could be used to identify emerging health risks and to develop and evaluate evidence-based disease control and prevention programs.

While there are common features to all modern surveillance systems, programs can be conceptualized into several distinct types based on their focus:

> **1) Surveillance for prevention**. Focuses on the risk factors and antecedents of disease. In the prevention of diabetes, such a system would focus on monitoring recognized risk factors in the development of type 2 diabetes such as smoking, obesity, diet and exercise (Chipkin, 1996). Surveillance for prevention allows public health officials to act proactively rather than reactively to prevent the spread of disease. "Upstream" disease prevention has tremendous potential for success, but is more difficult to for policy makers to accept. The ideal outcome from upstream prevention is low incidence of disease, which is less tangible than the outcomes from other health programs (such as primary and acute care).
>
> 2) **Surveillance of disease prevalence, incidence and diagnosis.** Focuses on the incidence and prevalence of disease within a population. The key challenge is in the identification of true cases of disease in the population, many of which may be undiagnosed. With Type II diabetes, many individuals may meet the clinical definition of disease, but are asymptomatic and undiagnosed. These individuals must be identified in

order to prevent complication or slow the progression of disease. For diabetes health policy planning, monitoring the time from development of disease to diagnoses is crucial in order to improve screening and disease management programs.

3) **Surveillance of disease management**. Focuses on identifying diagnosed cases of disease to monitor disease management and outcomes. This type of surveillance is very common, and is the focus of this research project. Disease management systems provide us with progress indicators on how we are doing in addressing specific diseases. In diabetes care, this type of system is crucial to ensure that, (1) patients receive adequate follow-up care once they have been diagnosed with disease, (2) to ensure that patients receive timely care at diabetes centres, and (3) to monitor common diabetic co-morbidities/complications (e.g. cardiovascular disease and diabetic retinopathy). Monitoring these indicators helps ensure effective treatment and resource planning/deployment to prevent or ameliorate the negative consequences of diabetes.

While it is important to delineate types of surveillance, it is equally important to identify where and how surveillance information is used. Surveillance data is used at the national, provincial, regional, and local levels. The Public Health Agency of Canada monitors national trends in disease in order to plan broad country-wide health initiatives and target resources. Since health care delivery is under provincial jurisdiction in Canada, provinces and territories use disease surveillance systems to identify specific disease trends within their populations. At the regional level, district health authorities use surveillance data to monitor disease trends at the sub-provincial level, within place such as Cape Breton. Finally, Communities use surveillance data at the sub-population level, and are often responsible for the implementation of initiatives and programs to address disease within their communities.

Over the past two decades within Canada, there has been a movement towards the devolution of health care planning and decision making to smaller geographical and sub-population levels (Lewis, 2004). Surveillance systems which are able to provide information at the local and sub-population level are needed, as this is where health care is increasingly being planned and delivered. Disease trends at the national and provincial level may not be reflected at

the local level, and with increasing population diversity, it is important to understand disease management and risk factors at the local level for specific groups of individuals.

An excellent example of the need for surveillance data at the local level is within Aboriginal communities. While it is believed that Aboriginal people are at higher risk of developing diabetes (FNIHB, 2010), there is great diversity between Aboriginal communities in Canada, and the risk may not be equal between or even within communities. Moreover, the need for Aboriginal communities to have greater control over their own health care planning and delivery has been widely recognized in major reports on health in Canada (Commission of the Future of Health Care in Canada, 2002., Royal Commission on Aboriginal Peoples, 1996).

Despite the need for surveillance of diabetes in Aboriginal communities, there are considerable gaps in the data needed to support surveillance at this level. Many Canadian Aboriginal communities are not included in national and provincial health surveys. The First Nations Longitudinal Regional Health Survey (RHS) was initiated to address this need for disease surveillance within Aboriginal communities. The RHS is a First Nations governed, longitudinal national health survey (RHS, 2010). Data from the RHS is controlled by the Canadian First Nations due to complex self governance issues within many Aboriginal communities. The main advantage of this self-governance is that Aboriginal communities can claim ownership of their health data under the principals of OCAP (ownership, control, access, possession). This survey provides excellent coverage of First Nations communities within Canada, but due to its national scope, does not have adequate statistical power to produce reliable indicators at the community level. Using national or regional level indicators for planning at the community level is also problematic, as minority populations are often erroneously treated as homogeneous (for example, First Nations, Inuit, and Métis people are

distinct groups, but are often considered to be part of a homogenous group of Aboriginal peoples) (Burrows, 2004). Given the heterogeneity and diversity of Aboriginal peoples within Canada, there is very likely to be considerable variation in disease indicators between communities. Moreover, the accuracy of data sources may also vary across communities (Liao, 2004).

Data from Administrative databases (such as hospital discharge abstracts and physician billing data) is an attractive option for surveillance within Aboriginal communities because of its population coverage. It could be used to attempt to provide the coverage and statistical power needed for Aboriginal health surveillance, but it can be problematic to identify Aboriginal community members using this type of data, as individuals who are not registered under the Indian Act cannot be identified (Pohar, 2007). Geographic proxies for Aboriginal individuals can also be inaccurate due to the wide geographical dispersion of Aboriginal peoples (Assembly of First Nations, 2010). The use of local sources of data from Aboriginal communities, such as electronic medical records, is also being explored. Unfortunately this type of information is not typically organized or accessible for disease surveillance, although they are of great potential for surveillance systems.

Health data issues within Aboriginal communities are reflective of the problems that arise with the devolution of health care planning and implementation to the local level. We will now examine these and other challenges to disease surveillance more generally within Canada.

Challenges in Current Approaches to Surveillance for Disease Management

There are three critical challenges in surveillance of diagnosed cases of disease within the local and regional populations;

1) To obtain data with sufficient statistical power to support surveillance at the local and regional level,

2) To obtain data that is sufficiently accurate to avoid bias that could compromise surveillance, and

3) To obtain data that has a sufficient breadth of information to support a variety of surveillance requirements (i.e. co-morbidities, processes of care and outcomes).

Utilizing specific types of data necessitates tradeoffs between these three challenges. Consideration of the strengths and weaknesses of different surveillance systems using different data sources illustrates these tradeoffs.

Survey data provides a large breadth of information and is reasonably accurate, but typically lacks adequate statistical power for surveillance at the local level. The advantage of survey data is that it is often collected for research purposes, and surveys can be targeted towards addressing priority health concerns. They thus contain a wealth of data which can be tailored to meet information needs. A primary disadvantage of using survey data is that it must rely on self-reported diagnosis of disease from individuals. Without linkages to other data sources, or other types of verification procedures, there may be flaws in the data as individuals may forget to report health conditions or misunderstand specific questions.

Nationally, the Canadian Community Health Survey (CCHS) is the flagship survey for health surveillance. This survey represents one of the largest Canadian efforts to collect health information, and carries significant expense in its administration. The CCHS is a cross sectional survey designed to provide reliable disease estimates at the health region level (Statistics Canada, 2010). This survey targets all Canadians over the age of 12, but excludes individuals living on Reserves, crown lands, institutionalized individuals, residents of certain remote regions, and members of the Canadian Forces. The advantages of this survey are that it has a large sample size, employs sampling methods to avoid bias at the national level, and also collects a broad

breadth of demographic and outcome information. Despite its large sample size and coverage of the Canadian population, the CCHS often has inadequate statistical power to produce reliable health estimates at the health region or local level. This inadequacy is exacerbated by the fact that some information which is used to estimate health indicators at the health region level is collected in only a sub-sample of respondents.

For Aboriginal communities, a primary survey data resource is the First Nations Regional Longitudinal Health Survey (RHS). This survey also employs sampling methods to correct for bias and error in the data, as well as collects information on a range of health indicators. The RHS uses a cluster sampling design, where Aboriginal communities are selected and approached to participate in this survey, and then samples of individuals are randomly selected from these communities to complete the survey. As discussed earlier, this survey has excellent coverage of Aboriginal communities within Canada, but does not have adequate coverage of Aboriginal communities to produce estimators of diabetes at the community level even though large samples were drawn from many communities across Canada. The samples of Aboriginal people on reserves ranged from as high as 53.8% in Newfoundland and Labrador, to 2.1% in Ontario. The Nova Scotia sample was approximately 14.2% of the population of interest (First Nations Centre, 2002), which is not adequate for making inter-community comparisons within this province. Since the RHS is targeted towards Aboriginal communities only, it does have excellent coverage of the Canadian Aboriginal population living on reserves; unfortunately, it does not cover Aboriginal peoples living off reserve.

A popular alternative surveillance data source is administrative health data. This type of data is routinely collected by provincial governments in the process of administering Medicare (e.g. Provincial Health Insurance Programs). Administrative data includes physician billing data,

hospital discharge abstract data (DAD), and Pharmacare data. DAD consists of demographic, administrative and clinical data on all patients discharged from hospitals or who receive day surgery (CIHI, 2010). Physician billing data is typically submitted to provincial health insurance programs by physicians on a fee-for-service pay schedule to receive reimbursement for services rendered. Many provinces also collect prescription claims data as part of provincial drug programs. Pharmacare data has been used to help identify individuals with diabetes within specific populations which may not be extensively covered in other data sources (e.g. frail elderly), but has not been widely used within disease surveillance systems. Insulin and oral antihyperglycemic agents are a reliable proxy for diagnosed diabetes among seniors which also can be used (DCPNS, 2009).

By linking and combining these administrative data sources, provinces and the Public Health Agency of Canada have established surveillance systems for diabetes and a number of other chronic diseases.

The primary advantages of surveillance systems based on administrative data include their wide availability and population coverage. These types of surveillance system have minimal issues with statistical power since they collect information on every person who accesses health services within Canada. Surveillance systems based on administrative data also suffer from some major limitations. Administrative data is typically not collected for surveillance or research purposes, and thus there may be gaps and inaccuracies in the data collected. For example, while Nova Scotia allows for multiple diagnostic codes within hospital DAD and MSI submissions, patients may only be assigned a single diagnostic code even though they present with multiple complaints. A patient with who visits their general practitioner to manage high blood pressure and discuss their diabetes may be identified in administrative data as having high

blood pressure only, and may not be coded as a diabetic. Administrative data also suffers from accuracy problems. Diagnostic disease coding in administrative data is known to have inaccuracies that can result in significant misclassification error when used for disease surveillance (Strom, 2001.; Feinstein, 1989; Kephart 2004). Diagnostic errors can result as coding on physician claims is often entered by physicians or their support staff, and not professional coders. Checking procedures to ensure data integrity may not be followed within a busy clinic, and can result in data inaccuracies.

One type of error which commonly happens within administrative data source is misclassification error. This type of error typically arises in two situations, when individuals with a specific disease are falsely classified as not having that disease (False Negatives), and when individuals who do not have a specific disease are falsely classified as having this disease (False Positives). Misclassification error causes problems as it creates an imbalance between the number of false positives and false negatives within the data source. False cases are a function of the data's sensitivity (the percent of true cases captured) and its specificity (the percent of non-cases captured). Decreases in the sensitivity of the data source over time will result in underestimates of the prevalence of the disease of interest, through the generation of false negative cases, while decreases in the specificity of the data source will result in overestimation of the prevalence of disease, through the generation of false-positive cases.

The true prevalence of disease also changes how the sensitivity and specificity will affect prevalence estimates. At low disease prevalence, specificity has a larger influence on the validity of administrative case definitions, as it will results in a larger numbers of false positive (see table 1). Even if the sensitivity and specificity of a data source did not change, bias could differ as a function of variations in the true prevalence of disease between communities. This

interrelationship is especially pertinent to our analysis of diabetes rates within and between minority populations as it has been shown that both the prevalence of disease, and the sensitivity and specificity of the data, may vary across these populations (Sackett, 2002). Combined, variation in these three factors can result in biased estimates of true disease prevalence, with differential bias between communities.

Table 1. The effect of Sensitivity, Specificity, and True Disease Prevalence on Estimates of Disease Prevalence using various types of Surveillance Data.

Data source	True Prevalence	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value	Estimated Prevalence (error)
Administrative Data	2%	85%	96%	43%	99.7%	5.6% (+3.6)
	2%	85%	99%	63%	99.7%	2.7% (+0.7)
Clinical Registries	2%	50%	100%	100%	99%	1% (-1.0)
Pharmacare Claims	2%	40%	100%	100%	99%	8% (+6.0%)
Administrative Data	10%	85%	96%	70%	98%	12% (+2.0)
	10%	85%	99%	82%	98%	10% (0)
Clinical Registries	10%	50%	100%	100%	95%	5% (-5.0)
Pharmacare Claims	10%	40%	100%	100%	94%	4% (-6.0)

Administrative data typically has high sensitivity and specificity for identifying cases of diabetes (table 1) (Kephart, 2004), although the accuracy of these data sources is highly affected

by the true prevalence rate. We can also see, that when the specificity of the data is altered, even by a small amount, the error in the estimates of disease prevalence increase drastically (table 1).

While there are concerns about the accuracy of administrative data, it is often the only information which is available at the population level to monitor disease. The National Diabetes Surveillance System was a collaborative network of provincial and territorial information systems designed to improve the breadth of diabetes surveillance. Administrative data from provincial health insurance registries were linked to physician billing and hospitalization data in each province and territory, and compiled by the Public Health Agency of Canada (PHAC, 2009). The strength of the NDSS was that it had excellent coverage and statistical power since it is population-based. Several weaknesses of the NDSS have been identified including;

1) The overestimation of diabetes prevalence when it is used over time. False positive cases accumulate as the data is used over time and the ratio of false positive to false negative cases become unbalanced (Health Canada, 2003), and

2) In Nova Scotia, and elsewhere, many physicians are switching to alternate pay structures where they are not required to submit billing information, and thus the validity of the NDSS case definition may have declined.

Clinical registries are another data source sometimes used in disease surveillance. Clinical registries contain information on patients with a particular condition of interest, within which information is collected and can be used for surveillance and research purposes. The Diabetes Care Program of Nova Scotia (DCPNS) maintains a clinical registry (DCPNS registry) of all cases of diabetes and pre-diabetes referred to their Diabetes Centres around the province. The DCPNS Registry is not subject to some of the limitations of the NDSS. The DCPNS Registry contains information collected through the NS Diabetes Centres; as such, all cases appearing in this Registry have clinically diagnosed diabetes or pre-diabetes (i.e., 100% or no false positives). A limitation of this Registry is that individuals with diabetes who never accessed care at a

Diabetes Centre do not appear within the DCPNS Registry. When comparing the DCPNS Registry against the NDSS, over 70% of the estimated cases identified by the NDSS methodology appear in the DCPNS Registry, with this percentage varying by age (DCONS, personal correspondence). The population-based DCPNS Registry is actively used for diabetes surveillance at the provincial District Health Authority, and community level with the recognition that it does not capture all diabetes cases in the province, but rather those who have received care through Diabetes Centres.

We have seen that each data source has particular strengths and weaknesses which must be addressed. With this in mind, we will consider new approaches to surveillance which seek to overcome the limitations inherent in the use of any one particular data source.

New Directions/Approaches

Given the limitations of using individual data sources for surveillance purpose, there has recently been considerable interest in methods to both combine information from multiple data sources and correct for misclassification error. Through the combination of multiple data sources, it is possible to balance the strengths and weaknesses of each data source, as well as increase the power of data sources to perform subpopulation analyses. Administrative data often has the potential to be linked with disease registries and survey data through the use of personal identifiers such as health card number, or from multiple cell sources such as name, date of birth, and sex using probabilistic matching. While our current data resources may have individual weaknesses, a combination of these sources of information would produce a powerful new tool for disease surveillance.

The Nova Scotia Diabetes Repository (NSDR) is an example of such an approach to combining data sources for disease surveillance. The NSDR was a pilot project that combined

the advantages of the population coverage of administrative data sources used in the NDSS (physician billings data and hospital DAD data) with data rich information in the DCPNS Registry, along with supplemental information from Nova Scotia Pharmacare on medication use related to diabetes (i.e. oral antihyperglycemic agents and insulin), and the Nova Scotia Atlee Perinatal Database (NSAPD), which contains information on all pregnancies in Nova Scotia with information on pre-existing diabetes and new cases of gestational diabetes.

To validate NSDR cases, Kephart and Andreou (2009) used a Bayesian statistical method to evaluate and combine multiple databases to increase the accuracy of surveillance data. To estimate diabetes in persons aged over 65, Pharmacare information, data from the DCPNS registry, and administrative data were combined into a model to simultaneously estimate the sensitivity and specificity of each data source, and to estimate the prevalence of diabetes overall. For people under 65 years of age, data from the diabetes care program and administrative data were used. While this approach yielded good estimates of the prevalence of diabetes, prevalence estimates were conservative, and values for sensitivity of the data sources were likely biased as the authors did not account for dependence between data sources. In this thesis, we will apply the methods used by Kephart and Andreou to estimate the accuracy of data sources for identifying diabetes case definitions in Aboriginal populations. In addition, we will extend the models used by Kephart and Andreou to further refine estimates of the accuracy of various data sources used in the NSDR.

Compensating for Estimation Error in surveillance methods

One approach to address misclassification error in data sources is to adjust estimates of prevalence for misclassification error, based on knowledge of sensitivity and specificity. For example, we can adjust estimate of apparent disease prevalence using a likelihood approach.

Using the observed prevalence as an estimate, we can multiply this estimate by the likelihood ratio (which is estimated from sensitivity and specificity), which will result in an estimate of prevalence which more accurately estimates the true prevalence. Maximum likelihood estimates are a set of estimates which are determined to be the most likely to have generated the observed data. This type of estimation is useful in the absence of certainty about the condition of interest since they will give us the most "likely" estimate of the parameter of interest to match the observed data. A similar approach is to use a Rogan-Gladen estimator to obtain prevalence estimates (Greiner, 2003). A Rogan-Gladen estimator provides an approximately unbiased Bayesian estimate of prevalence conditional on the apparent prevalence (the estimates of prevalence are obtained by applying a diagnostic test to the population) and the sensitivity and specificity (Kephart, 2004).

These two approaches to adjustment have been extensively used within the literature but have several important limitations to their use. The primary issue is that both rely on knowing with certainty the sensitivity and specificity of an ideal reference. There are many difficulties with identifying an ideal reference standard with which to evaluate diagnostic test parameters. The second weakness of these approaches is related to the first, in that it does not allow users to incorporate uncertainty about sensitivity, specificity and prevalence into the adjustment process. This issue is especially important for disease surveillance at the local level and within subpopulation groups, as sensitivity and specificity can vary across or within these levels of analysis, and they can only be estimated with error.

The sensitivity, specificity and prevalence of disease cannot be known with complete certainty, unless an ideal (gold) reference standard is available. An ideal reference standard should meet three criteria (Reitsma, 2009): the reference standard provides error-free

classification of all subjects, the same reference standard is used to verify all index test results, and the index test and reference standard can be performed within a short interval to avoid changes in target conditional status. Even with an ideal reference standard, there will always be a margin of error associated with parameter estimates, and it is important to account for this uncertainty within the adjustment procedure (Lawrence, 1995).

Reitsma et al (2009) argues that the criteria for a reference standard are rarely met with any diagnostic as there are many types of error which can impede the construction of an ideal reference standard including:

1) **Misclassification**: if the condition of interest does not manifest typically, or is not detectable at the time of diagnosis, this would cause error in the reference standard if that individual was not classified as having the condition of interest,

2) **Dichotomizing**: Conditions of interest often occur along a spectrum, and not simply as present or absent, and as different standards may use different cut-points, this can result in error in the reference standard,

3) **Failure in the reference standard protocol**: practitioners who are administering the protocol may not adhere to the highest standard and may misclassify individuals who have the condition of interest, and

4) **Interpretation errors by observers**: There may be error in the entering of results of the reference standard into databases, which would result in errors in the reference standard.

Models for Estimating Parameters of Interest in the Absence of a Gold-Standard

A growing body of literature explores methods for estimating sensitivity, specificity, and disease prevalence in the absence of a gold standard (Reitsma, 2009., Rutjes, 2007). To illustrate the basics of the approaches, and the challenges involved, it is useful to start with a simple example. The simplest data which can be used to represent an evaluation of diagnostic parameters is for one population of interest with two independent diagnostics. The independence of the diagnostics refers to the fact that the error in each diagnostic database is not related to the

error in the other. For the remainder of these examples, we will use the word diagnostic to refer to a diagnostic database used to identify cases of diagnosed disease.

The simplest way to represent the findings of two diagnostic tests within one population is to form a 2 X 2 table. In this table we have the results of diagnostics "2" represented in the rows, and for diagnostic "1" in the columns. (Figure 2). Since we have applied two diagnostics to the population of interest, our simple table also shows the degree of agreement or disagreement between the diagnostics, represented as the cells of the tables. For instance, if a person from the population is identified as having a condition of interest by test 1, but as not having the condition by test 2, they would be included in cell X_{12}.

		Diagnostic 2	
		no	yes
Diagnostic 1	no	X_{11}	X_{21}
	yes	X_{12}	X_{22}

$$X_{11}= [P[D^+]* [1-Se_1]*[1-Se_2] + P[D^-]*[Sp_1]* [Sp_2]]*n$$
$$X_{21}= [P[D^+]*[1-Se_1]*[Se_2] + P[D^-]*[Sp_1]* [1-Sp_2]]*n$$
$$X_{12}= [P[D^+]*[Se_1]*[1-Se_2] + P[D^-]*[1-Sp_1]* [Sp_2]]*n$$
$$X_{22}= [P[D^+]*[Se_1]*[Se_2] + P[D^-]*[1-Sp_1]* [1-Sp_2]]*n$$

$P[D^+]$ = Probability of having Diabetes
$P[D^-]$= Probability of not having Diabetes
$[Se_1]$= Probability of being identified as a Diabetic within Diagnostic 1 conditional on true status as a diabetic
$[1-Se_1]$= Probability of being identified as a non-Diabetic within Diagnostic 1 conditional on true status as a diabetic
$[Sp_1]$= Probability of being identified as a non-Diabetic within Diagnostic 1 conditional on true status as a non-Diabetic
$[1-Sp_1]$= Probability of being a identified as a Diabetic within Diagnostic 1 conditional on true status as a non-diabetic
$[Se_2]$= Probability of being identified as a Diabetic within Diagnostic 2 conditional on true status as a diabetic
$[1-Se_2]$= Probability of being identified as a non-Diabetic within Diagnostic 2 conditional on true status as a diabetic
$[Sp_2]$= Probability of being identified as a non-Diabetic within Diagnostic 2 conditional on true status as a non-Diabetic
$[1-Sp_2]$= Probability of being a identified as a Diabetic within Diagnostic 2 conditional on true status as a non-diabetic
n= total population $(X_{11}+ X_{21}+ X_{12}+ X_{22})$

Figure 1. One population with two independent Diagnostic Tests

In the instance of one population with two independent diagnostics, using probability theory, cells can be represented as a function of the probability that an individual has diabetes, and the sensitivity and specificity of diagnostics 1 and 2.

We can then write equations to solve for our parameters of interest, and see from these equations that there are 5 unknown properties, P, Se_1, Se_2, Sp_1, and Sp_2. In a 2 X 2 table (figure 1), if we know the values of three of the four cells we can calculate the value of the unknown cell, and thus the data has three degrees of freedom. With five unknown parameters, and only three degrees of freedom, we cannot specify a unique solution to any of the unknown parameters in the equations (Lawrence, 1995).

The only way to estimate model parameters is to specify values for at least some of the parameters, thereby reducing the number of unknown parameters. For example, if previous research showed that the values of specificity and prevalence were 95% and 7% respectively, we could substitute these values into our equations, and reduce the number of unknown parameters to three with three degrees of freedom. Using this method, we could solve the equations if there was enough research on our diagnostics tests for us to confidently make these assumptions about some of the parameters. As discussed earlier, this information is typically not accessible due to the non-availability, or inadequacy, of diagnostic gold-standard databases.

A more sophisticated, but similar approach to solving this equation is to use Bayesian statistical methods, which employ a more formalized and integrated approach to including past research to estimate model parameters. This approach allows us to incorporate prior information about our parameter estimates, while also incorporating uncertainty about these values. An example would be if we thought that the specificity and prevalence of our diagnostic data had means of 95% and 7% respectively, with associated credibility intervals. A Bayesian methodological approach would allow us to specify prior distributions for model parameters (e.g. a beta-distributions corresponding to a mean and confidence interval from previous research). Bayesian estimation techniques combine the prior distributions of model parameters with the

data to derive "posterior" distributions of the model parameters. Providing the prior distributions and the data contain enough informative information, the posterior distributions can provide reasonably precise estimates of the model parameters, conditional on the prior distributions, the data and the model on which the estimation is based (i.e. the equations in Figure1).

If more detailed data are available, fewer assumptions (or informative prior distributions) are required, and estimates of model parameters can be based more exclusively on the data. Consider, for example the situation where we have data on two populations, with different prevalence of disease, and data from two independent diagnostics on each population (Figure 3). There are now 6 unknown parameters, and 6 degrees of freedom. Knowing this, we can calculate a unique solution for all the parameters of interest (Lawrence, 1995). Bayesian methods can still be used to incorporate prior knowledge (and uncertainty) into our parameter estimates, and can enhance the precision and accuracy of estimates.

Population A

Diagnostic 2

Diagnostic 1	+	-
+	X_{11}	X_{21}
-	X_{12}	X_{22}

$X_{11}= [P[D_A^+]* [1-Se_1]* [1-Se_2] + P[D_A^-]* [Sp_1]* [Sp_2]]*n$
$X_{21}= [P[D_A^+]* [1-Se_1]* [Se_2] + P[D_A^-]* [Sp_1]* [1-Sp_2]]*n$
$X_{12}= [P[D_A^+]* [Se_1]* [1-Se_2] + P[D_A^-]* [1-Sp_1]* [Sp_2]]*n$
$X_{22}= [P[D_A^+]* [Se_1]* [Se_2] + P[D_A^-]* [1-Sp_1]* [1-Sp_2]]*n$
$Y_{11}= [P[D_B^+]* [1-Se_1]* [1-Se_2] + P[D_B^-]* [Sp_1]* [Sp_2]]*n$
$Y_{21}= [P[D_B^+]* [1-Se_1]* [Se_2] + P[D_B^-]* [Sp_1]* [1-Sp_2]]*n$
$Y_{12}= [P[D_B^+]* [Se_1]* [1-Se_2] + P[D_B^-]* [Sp_1]* [Sp_2]]*n$
$Y_{22}= [P[D_B^+]* [Se_1]* [Se_2] + P[D_B^-]* [1-Sp_1]* [1-Sp_2]]*n$

Population B

Diagnostic 1

Diagnostic 2	+	-
+	Y_{11}	Y_{21}
-	Y_{12}	Y_{22}

$P[D^+]$ = Probability of having Diabetes
$P[D^-]$= Probability of not having Diabetes
$[Se_1]$= Probability of being identified as a Diabetic within Diagnostic 1 conditional on true status as a diabetic
$[1-Se_1]$= Probability of being identified as a non-Diabetic within Diagnostic 1 conditional on true status as a diabetic
$[Sp_1]$= Probability of being identified as a non-Diabetic within Diagnostic 1 conditional on true status as a non-Diabetic
$[1-Sp_1]$= Probability of being a identified as a Diabetic within Diagnostic 1, conditional on true status as a non-diabetic
$[Se_2]$= Probability of being identified as a Diabetic within Diagnostic 2 conditional on true status as a diabetic
$[1-Se_2]$= Probability of being identified as a non-Diabetic within Diagnostic 2 conditional on true status as a diabetic
$[Sp_2]$= Probability of being identified as a non-Diabetic within Diagnostic 2 conditional on true status as a non-Diabetic
$[1-Sp_2]$= Probability of being a identified as a Diabetic within Diagnostic 2 conditional on true status as a non-diabetic
n= total population $(X_{11}+ X_{21}+ X_{12}+ X_{22})$

Figure 2. Two Populations (different disease prevalence) with Two Independent Diagnostics

While more data from separate populations can provide enough degrees of freedom to solve our equations, this is not a straightforward relationship. Sometimes there is only information on one population of interest, say if the administrative data does not contain much demographic information, but we have data from three independent diagnostics tests (Figure 4). In this model there are 7 unknown parameters, and 7 degrees of freedom. For this model we now have enough information to find a unique solution to our parameters of interest. For this situation, the Bayesian method is again useful as it will allow us to use prior information to combine with the data, and obtain meaningful estimates of model parameters.

			Diagnostic 3	
		Diagnostic 2	yes	no
Diagnostic 1	yes	yes	X_{111}	X_{112}
		no	X_{121}	X_{122}
	no	yes	X_{211}	X_{212}
		no	X_{221}	X_{222}

$$X_{111}= [P[D^+]* [Se_1]* [Se_2]* [Se_3] + P[D^-]* [1-Sp_1]* [1-Sp_2]* [1-Sp_3]]*n$$
$$X_{112}= [P[D^+]* [Se_1]* [Se_2]* [1-Se_3] + P[D^-]* [1-Sp_1]* [1-Sp_2]* [Sp_3]]*n$$
$$X_{121}= [P[D^+]* [Se_1]* [1-Se_2]* [Se_3] + P[D^-]* [1-Sp_1]* [Sp_2]* [1-Sp_3]]*n$$
$$X_{122}= [P[D^+]* [Se]* [1-Se_2]* [1-Se_3] + P[D^-]* [1-Sp_1]* [Sp_2]* [Sp_3]]*n$$
$$X_{211}= [P[D^+]* [1-Se_1]* [Se_2]* [Se_3] + P[D^-]* [Sp_1]* [1-Sp_2]* [1-Sp_3]]*n$$
$$X_{212}= [P[D^+]* [1-Se_1]* [Se_2]* [1-Se_3] + P[D^-]* [Sp_1]* [1-Sp_2]* [Sp_3]]*n$$
$$X_{221}= [P[D^+]* [1-Se_1]* [1-Se_2]* [Se_3] + P[D^-]* [Sp_1]* [Sp_2]* [1-Sp_3]]*n$$
$$X_{222}= [P[D^+]* [1-Se_1]* [1-Se_2]* [1-Se_3] + P[D^-]* [Sp_1]* [Sp_2]* [Sp_3]]*n$$

$P[D^+]$ = Probability of having Diabetes
$P[D^-]$= Probability of not having Diabetes
$[Se_1]$= Probability of being identified as a Diabetic within Diagnostic 1 conditional on true status as a diabetic
$[1-Se_1]$= Probability of being identified as a non-Diabetic within Diagnostic 1 conditional on true status as a diabetic
$[Sp_1]$= Probability of being identified as a non-Diabetic within Diagnostic 1 conditional on true status as a non-Diabetic
$[1-Sp_1]$= Probability of being a identified as a Diabetic within Diagnostic 1 conditional on true status as a non-diabetic
$[Se_2]$= Probability of being identified as a Diabetic within Diagnostic 2 conditional on true status as a diabetic
$[1-Se_2]$= Probability of being identified as a non-Diabetic within Diagnostic 2 conditional on true status as a diabetic
$[Sp_2]$= Probability of being identified as a non-Diabetic within Diagnostic 2 conditional on true status as a non-Diabetic
$[1-Sp_2]$= Probability of being a identified as a Diabetic within Diagnostic 2 conditional on true status as a non-diabetic
$[Se_3]$= Probability of being identified as a Diabetic within Diagnostic 3 conditional on true status as a diabetic
$[1-Se_3]$= Probability of being identified as a non-Diabetic within Diagnostic 3 conditional on true status as a diabetic
$[Sp_3]$= Probability of being identified as a non-Diabetic within Diagnostic 3 conditional on true status as a non-Diabetic
$[1-Sp_3]$= Probability of being a identified as a Diabetic within Diagnostic 3 conditional on true status as a non-diabetic
n= total population $(X_{11}+ X_{21}+ X_{12}+ X_{22})$

Figure 3. One Population with Three Independent Diagnostics

The previous models all make a strong assumption about the independence of error in diagnostic tests. Especially with different data sources for identifying a disease, errors may not be independent. In many cases, coding of disease from different sources may be related. For example, providers who do not code a disease in one data source may not do so in another, or errors may be geographically clustered because of the way patients access care. Error in different administrative data sources is possible if the same health professional records diagnostic information that gets incorporated into physician billings data and hospital discharge data.

Dependence is a very realistic situation within health data diagnostic databases, but the difficulty with dependence is that we often do not have good information on how the diagnostic

parameters are related. For example, given a positive result on Diagnostic 1, we could be more likely to have a positive result on diagnostic 2, but this does not support the assumption that given a negative result on 2, we are more likely to see a negative result on 1. Within this study, we parameterize dependence in the context of validation of a data source which contains gold standard cases of disease and one which does not. Dependence of data sources is estimated through the inclusion of a covariance term to express the conditional covariance of error between data sources.

While the dependence assumption of a model of interest is likely more representative of the true nature of using diagnostic tests, it is more complicated, and in order to model the relationship, we must employ more complicated models, with new parameters to reflect dependence. Thus, more data and assumptions will be required to estimate model parameters (figure 5).

Population A

Diagnostic 1	Diagnostic 2: no	Diagnostic 2: yes
no	X_{11}	X_{21}
yes	X_{12}	X_{22}

Population B

Diagnostic 1	Diagnostic 2: no	Diagnostic 2: yes
no	Y_{11}	Y_{21}
yes	Y_{12}	Y_{22}

$$X_{11}= [P[D^+]* [1-Se_1]* [1-Se_2] + cov] + P[D^-]* [Sp_1]* [Sp_2]]*n$$
$$X_{21}= [P[D^+]* [1-Se_1]* [Se_2] – cov] + P[D^-]* [Sp_1]* [1-Sp_2]]*n$$
$$X_{12}= [P[D^+]* [Se_1]* [1-Se_2]- cov] + P[D^-]* [1-Sp_1]* [Sp_2]]*n$$
$$X_{22}= [P[D^+]* [Se_1]* [Se_2] + cov] + P[D^-]* [1-Sp_1]* [1-Sp_2]]*n$$

$$Y_{11}= [P[D^+]* [1-Se_1]* [1-Se_2] + cov] + P[D^-]* [Sp_1]* [Sp_2]]*n$$
$$Y_{21}= [P[D^+]* [1-Se_1]* [Se_2] – cov] + P[D^-]* [Sp_1]* [1-Sp_2]]*n$$
$$Y_{12}= [P[D^+]* [Se_1]* [1-Se_2] – cov] + P[D^-]* [1-Sp_1]* [Sp_2]]*n$$
$$Y_{22}= [P[D^+]* [Se_1]* [Se_2] + cov] + P[D^-]* [1-Sp_1]* [1-Sp_2]]*n$$

$P[D^+]$ = Probability of being identified as having Diabetes
$P[D^-]$= Probability of not being identified as having Diabetes
$[Se_1]$= Probability of being identified as a Diabetic in diagnostic 1 conditional on true status as a diabetic
$[1-Se_1]$= Probability of being identified as a non-Diabetic in diagnostic 1 conditional on true status as a diabetic
$[Sp_1]$= Probability of being identified as a non-Diabetic in diagnostic 1 conditional on true status as a non-Diabetic
$[1-Sp_1]$= Probability of being a identified as a Diabetic in diagnostic 1 conditional on true status as a non-diabetic
$[Se_2]$= Probability of being identified as a Diabetic in diagnostic 2 conditional on true status as a diabetic
$[1-Se_2]$= Probability of being identified as a non-Diabetic in diagnostic 2 conditional on true status as a diabetic
$[Sp_2]$= Probability of being identified as a non-Diabetic in diagnostic 2 conditional on true status as a non-Diabetic
$[1-Sp_2]$= Probability of being a identified as a Diabetic within Diagnostic 2 conditional on true status as a non-diabetic
Cov= Conditional covariance of error in sensitivities between data sources
n= total population $(X_{11}+ X_{21}+ X_{12}+ X_{22})$

Figure 5. Two population with two dependent Diagnostics

From figure 4, we can now see that since we must an additional parameter (cov) to estimate the conditional covariance of sensitivities between data sources. We do not estimate dependence in specificity, as research shows that the specificity of administrative case definitions are typically very high (generally greater than .92 (DCPNS, 2009., Hux, 2002)). While this model may have sufficient degrees of freedom to make parameter estimates, it is sometime necessary to include additional informative prior information to converge reasonable

estimates. As the model become more complicated we must make more assumptions based on prior information in order to produce unique solutions for our parameters of interest.

Estimation Methods

A variety of methods exist to estimate parameters of interest once a model has been specified for the data, including Frequentist based on maximum-likelihood estimation (Enøe, 2000), the Expectation-Maximization algorithm (Enøe, 2000), and Bayesian approaches using a Gibbs sampler (Jospeh, 1995). As previously discussed, the advantage of the Bayesian method lays primarily in the fact that, while both approaches must rely on assumptions when models are under-identified (more parameters of interest than degrees of freedom), the Bayesian approach provides a framework for combining prior information and knowledge with data to estimate model parameters.

The Bayesian approach typically employs a Gibbs sampling technique. A Gibbs sampler chooses a random value for each parameter of interest which lies between the distribution estimate which was specified (usually between 0 and 1), and a sample is drawn from each distribution. The Sampler then builds on all further arbitrary selections by incorporating the values of previous samples into further estimations. This is the method which will be used in the following analyses.

In summary, we have seen that there is a need for reliable and accurate health data at the community level in order to meet the needs of policy makers. Current data sources are inadequate to meet this need, and so we must use methods to balance the strengths of multiple data sources which also explicitly acknowledge their weaknesses. In order to estimate the accuracy of data sources for identifying cases of diagnosed Diabetes in Aboriginal versus non-

Aboriginal populations we will use a Bayesian method to estimate parameters (sensitivity, specificity, and disease prevalence) within each population.

CHAPTER 3: Manuscript 1

Validation of an Administrative Case Definition for Identifying Cases of Diagnosed Diabetes within an Aboriginal Population at the Sub-Provincial Level

Introduction

While there has been a plethora of research using administrative data (Physician Claims and Hospital Discharge Data) to estimate the prevalence of diabetes in Aboriginal communities (Dyck, 2010., Johnson, 2009., Martens, 2007., Hemmelgarn, 2007., Green, 2003), it has typically relied on case definitions which have not been validated within these populations. For example, most Canadian research has used the "Manitoba" case definition (2 or more physician claims or 1 or more hospital claims within a two-year window), which has been validated in national and provincial populations (Hux, 2002., Health Canada, 2003., Johnson, 2009), but not for Aboriginal populations at the sub-provincial level. The sensitivity and specificity of administrative case definitions in Aboriginal populations may differ as a result of increased awareness of diabetes as a concern in this population, which could contribute to increased screening for diabetes. Sub-provincial populations are also served by a smaller number of health care providers, where the contribution of even a single coder deviating from average coding practices would have a substantial effect on the validity of administrative data case definitions.

Validation of administrative case definitions to identify cases of diagnosed diabetes at the sub-provincial level in Aboriginal populations is important for two reasons. The first reason is significant variation in the prevalence of diabetes between Aboriginal communities (Delisle, 1993., Yu, 2007., Oster, 2009). Difference in the prevalence of diabetes is likely a reflection of the heterogeneity of Aboriginal populations at the national and provincial level in Canada. The Canadian National Indian Registry broadly recognizes three groups of Aboriginal peoples; First

Nations, Metis, and Inuit each of whom have distinct history, language and cultural practices (Aboriginal Affairs and Northern Development Canada, 2011), and thus may not have the same prevalence or risk factors of diabetes. It is thus important to assess the validity of administrative case definitions of diabetes at the sub-provincial level for specific Aboriginal groups.

The second reason for validation of administrative case definitions within Aboriginal populations at this level is that the prevalence of diabetes is expected to be higher (Dyck, 2010), and the onset of diabetes occurs at a younger age (Oster, 2009), which will directly affect the positive predictive value (PPV) of administrative case definitions (PPV is the proportion of individuals who are identified as have a condition of interest in a surveillance system who are correctly identified). In any disease surveillance system, PPV is the most important parameter to consider when assessing accuracy as it directly illustrates the utility of the surveillance system to correctly identify individuals of interest. The relationship between the prevalence of disease and PPV is straightforward. For example, for any administrative case definition with a sensitivity (the proportion of true diabetics who are correctly classified) less than 100% and a fixed specificity (the proportion of non-diabetics who are correctly classified), the PPV will decrease as the population prevalence decreases. Again, it is thus important to assess the validity of administrative case definitions for specific Aboriginal populations at the sub-provincial level.

Further, through the validation of administrative case definitions within Aboriginal sub-populations, data from local diabetes surveillance systems can be combined with administrative data to improve diabetes prevalence estimation (Tu, 2010). Combining diabetes surveillance data from various sources is becoming increasingly important as more physicians move to alternative physician reimbursement schedules, which could affect the quality and population coverage of administrative data. If data from Electronic Medical Records's (EMR) is to be incorporated into

diabetes surveillance systems for Aboriginal communities, the validity of alternative administrative case definition's of diabetes first needs to be assessed.

The purpose of this study was to estimate the sensitivity and specificity of administrative case definitions for diabetes within five Aboriginal communities in Nova Scotia, Canada. Further, we also examined the utility of an electronic medical record based diabetes registry to improve apparent prevalence estimates of diabetes.

Method

Data and subjects

The subjects for this study were members of five Aboriginal communities in Cape Breton, Nova Scotia (Eskasoni, Membertou, Wagmatcook, Waycobah, and Chapel Island) in the year 2009 (n= 8380). For analysis, data on diagnosed diabetes was pooled across the five communities. Members of the communities were identified using the Unama'ki Client Registry (UCR), which was assembled by the communities, in partnership with the Population Health Research Unit at Dalhousie University, Medavie Blue Cross, and the Nova Scotia Department of Health and Wellness as an attempt to identify their communities for health policy research and planning. The UCR was assembled from EMR information, Indian and Northern Affairs Canada member data, and data from the Nova Scotia Medical Services Insurance Eligibility File. The UCR includes provincial health card numbers, as well as demographic and community information. The inclusion of health card numbers permitted linkage of the registry to provincial administrative data (Medical Services Insurance (MSI) Claims, and Canadian Institutes of Health Information Hospital Discharge Abstract Database (DAD)), as well as EMR disease registries maintained at local community health clinics.

Several case definitions were compared from the Administrative data (Table 1). Cases of Gestational Diabetes were excluded from the administrative data for these analyses. MSI claims data contains diagnostic data for health services rendered by a physician which has been reimbursed through Nova Scotia's MSI program. DAD contains information on all discharge diagnoses of identified patients from Nova Scotia hospitals. Using the UCR to identify members of the five communities, administrative data ensures that access to health services outside of the community specific health centres will be captured.

Cases of diabetes from the EMR diabetes registries were extracted according to coding practices within clinics located in each community. The five communities of interest all use Practimax© EMR software. Communities use the EMR software for clinical records, to submit physician billings reimbursement information, and also to assemble registries for diabetes surveillance. While all communities use their EMR's for the aforementioned purposes, three of the communities collect information on type of diabetes (Type 1, Type 2, and Gestational Diabetes) in their diabetes registries, while two do not. Thus, cases of gestational diabetes were excluded from analyses for three communities, but not for the two which did not collect information on type of diabetes.

The EMR was assumed to contain gold standard cases of diabetes (having perfect specificity) but was not considered to contain gold standard non-cases of diabetes as it does not have full population-coverage. Not all community members receive health care through the clinics. Some community members have primary care givers located in adjacent communities and thus may not be captured in the EMR.

Table 1. Administrative case definitions for non-gold standard data sources

Case Selection	Case Definition
1 in 1	One or more physician or hospital claims with a diagnosis of diabetes within the previous year, excluding identified cases of gestational diabetes.
2 in 1	At least two physician claims or at least one hospital claim with a diagnosis of diabetes within the previous year excluding identified cases of gestational diabetes.
Manitoba Rule	At least two physician claims or at least one hospital claim with a diagnosis of diabetes within the previous two years excluding identified cases of gestational diabetes.

Statistical Analysis

To assess the validity of diabetes case definitions, we began by examining percent agreement and kappa statistic (Cohen, 1960) between the EMR and Administrative data sources by age group and sex.

We adapted a model developed by Hui and Walter (1980) to estimate the sensitivity and specificity of administrative case definitions and the EMR diabetes registries. Hui and Walter showed that the sensitivity and specificity of two diagnostic tests can be estimated, in the absence of having a gold standard, if the tests are assumed to be independent (i.e. their errors are uncorrelated) and they are applied two populations with substantially different disease prevalence. As shown in figure 1, the model specifies the relationship between the observed data (cross-classifications of measured diabetes status according to the two tests in each population) and six model parameters: the sensitivity and specificity for each test, and the disease prevalence in each population.

Population A

EMR Data

Admin Data		EMR +	EMR -
	+	X_{11}	X_{21}
	-	X_{12}	X_{22}

$X_{22}= [P[D^+]* [1-Se_1]* [1-Se_2] + P[D_A^-]* [Sp_1]* [Sp_2]]*n$
$X_{21}= [P[D^+]* [1-Se_1]* [Se_2] + P[D_A^-]* [Sp_1]* [1-Sp_2]]*n$
$X_{12}= [P[D^+]* [Se_1]* [1-Se_2] + P[D_A^-]* [1-Sp_1]* [Sp_2]]*n$
$X_{11}= [P[D^+]* [Se_1]* [Se_2] + P[D_A^-]* [1-Sp_1]* [1-Sp_2]]*n$
$Y_{22}= [P[D^+]* [1-Se_1]* [1-Se_2] + P[D_B^-]* [Sp_1]* [Sp_2]]*n$
$Y_{21}= [P[D^+]* [1-Se_1]* [Se_2] + P[D_B^-]* [Sp_1]* [1-Sp_2]]*n$
$Y_{12}= [P[D^+]* [Se_1]* [1-Se_2] + P[D_B^-]* [Sp_1]* [Sp_2]]*n$
$Y_{11}= [P[D^+]* [Se_1]* [Se_2] + P[D_B^-]* [1-Sp_1]* [1-Sp_2]]*n$

Population B

EMR Data

Admin Data		EMR +	EMR -
	+	Y_{11}	Y_{21}
	-	Y_{12}	Y_{22}

$P[D^+]$ = Probability of having Diabetes
$P[D^-]$= Probability of not having Diabetes
$[Se_1]$= Probability of being identified as a Diabetic within EMR conditional on true status as a diabetic
$[1-Se_1]$= Probability of being identified as a non-Diabetic within EMR conditional on true status as a diabetic
$[Sp_1]$= Probability of being identified as a non-Diabetic within EMR conditional on true status as a non-Diabetic
$[1-Sp_1]$= Probability of being a identified as a Diabetic within EMR, conditional on true status as a non-diabetic
$[Se_2]$= Probability of being identified as a Diabetic within Admin conditional on true status as a diabetic
$[1-Se_2]$= Probability of being identified as a non-Diabetic within Admin conditional on true status as a diabetic
$[Sp_2]$= Probability of being identified as a non-Diabetic within Admin conditional on true status as a non-Diabetic
$[1-Sp_2]$= Probability of being a identified as a Diabetic within Admin conditional on true status as a non-diabetic
n= total population $(X_{11}+ X_{21}+ X_{12}+ X_{22})$

Figure 5. Two populations (with different disease prevalence) with two independent diagnostics

For model estimation, we treated age groups as separate populations, as the prevalence of diabetes rises sharply with age. As per figure 1, we estimated models using the following three pairs of age groups: <45 and 45-54, 55-64 and 65-74, and 65-74 and 75-84 as separate populations. Models including individuals over the age of 85 were not included as this group had an insufficient number of diabetic individuals. In all models, the Manitoba administrative case definition, and the EMR definition were treated as independent measures of diabetes (no conditional dependence). Separate sets of models were estimated for males and females.

Model parameters were estimated by Bayesian methods using WinBUGS software (Lunn, 2000), which utilizes Gibbs sampling, a Markov Chain Monte Carlo technique, to

generate estimates of model parameters. Non-informative priors were specified for the prevalence of diabetes in each population, the specificity of administrative case definitions, and the two sensitivities. Because the identification of diabetes in the EMR data is based on clinical assessment, we assumed the specificity of the EMR data was nearly perfect (1 coding error within 1000 entries), which was entered into the WinBUGS models as an informative prior (beta distribution). Models were estimated based on 50,000 iterations, with a burn in of 5000 iterations. Model convergence was assessed by checking density, trace and history plots.

Where possible, we sought to assess assumptions in the Hui-Walter model that might be violated. The model assumes the sensitivity of the EMR and administrative case definitions are constant across age group pairings (e.g. Se constant for individuals age <45 and 45-54). To check this assumption we estimated and compared parameters of interest using alternative pairs of age groups. A second assumption was that the prevalence of diabetes was substantially different between age groups. This assumption was checked using cross-tabulated calculations of apparent prevalence from the Manitoba administrative case definition (Table 6). A third assumption was that the parameters of interest were constant across the five communities. Unfortunately, due to the small population size of most of the communities, we did not have a sufficient number of diagnosed cases of diabetes to support stratified analysis. A final assumption was that error in our administrative case definitions would not be correlated with error in our EMR diabetes registry case definition. This is known as the independence assumption, and while it is very likely that dependence between our data sources exists (as indicated by high percent agreement and kappa values (Table 2)), we do not have sufficient information on our population of interest to specify a model which accounts for data source dependence.

Practically, the degree of misclassification error from diabetes case definitions depends on the positive and negative predictive values. Accordingly, we computed predictive values directly from estimates generated in Bayesian models for sensitivity and prevalence. Only positive predictive values are presented, as the negative predictive values are consistently very high (DCPNS, 2009).

Ethics approval for this study was obtained through Mi'kmaq Ethics Watch Board at Cape Breton University and Human Research Ethics Board at Dalhousie University. Data Access permission was granted though the Unama'ki Data Access Committee, and the Population Health Research Unit Data Access Committee at Dalhousie University.

Results

Table 2 shows the percent agreement and kappa statistics between administrative case definition rules and the EMR registry. Percent agreement values are relatively high among the younger and older age groups, but reach a low point in the 64-75 age group in both sexes across all administrative case definitions. Observed values for the kappa statistic likely reflect that this measure is a poor indicator of agreement for this data as the prevalence of disease and number of individuals within each age group varies significantly. High agreement could result from either accuracy in the case definitions, or highly correlated errors. Most Unama'ki communities use the EMR systems to submit physician billings information and to code clinical cases of diabetes. The overall highest percent agreement and kappa statistics are observed for the Manitoba rule.

Table 2. Percent Agreement and Kappa Statistics for Administrative Case Definition and EMR Registry

	1 and 1		2 and 1		Manitoba rule	
	% Agreement	Kappa	% Agreement	Kappa	% Agreement	Kappa
Female						
<45	95.3	53.6	95.4	49.8	95.8	58.3
45-54	86.2	56.3	85.5	48.6	87.2	60.3
55-64	86.0	64.7	81.4	47.6	86.0	64.4
65-74	73.2	31.1	75.7	30.9	75.2	37.3
75-84	79.6	43.1	79.6	35.2	80.5	44.7
Male						
<45	96.8	49.4	96.9	38.3	97.2	54.6
45-54	82.7	50.9	81.7	42.6	83.4	53.2
55-64	78.0	48.5	79.5	48.2	80.2	54.6
65-74	76.4	48.2	72.7	35.0	75.0	45.0
75-84	80.0	45.1	80.0	40.9	80.0	45.1

Sensitivity of the Administrative and EMR data case definitions were generated using our estimation method (Table 3). These estimates show that the sensitivity of the Manitoba administrative case definition are slightly higher for females than for males and generally increase with age .The 95% credibility intervals (CI) are wide and increase with age, indicated that these estimates are only moderately precise. Sensitivity estimates for the EMR case definition are relatively low within all age groups and are generally higher for females than males with the exception in the 65-84 age group. The 95% CI's are again large, indicating imprecise estimates of these parameters in older age groups.

Table 3. Estimates of Sensitivity for the EMR and Manitoba Administrative Case Definition

	Manitoba Case Definition	95% CI		EMR Case Definition	95% CI	
	Mean	Lower	Upper	Mean	Lower	Upper
Females						
<55	0.639	0.571	0.720	0.658	0.593	0.725
55-74	0.814	0.738	0.881	0.837	0.540	0.995
65-84	0.828	0.707	0.923	0.563	0.332	0.944
Males						
<55	0.628	0.545	0.720	0.608	0.532	0.687
55-74	0.790	0.711	0.860	0.724	0.537	0.976
65-84	0.766	0.636	0.875	0.700	0.456	0.975

Specificity estimates for the Manitoba administrative case definition (Table 4) are high in the youngest age group, but become low within the older age groups, and are quite close between males and females. The 95% CI's once again show that the results are only moderately precise, and generally become wider in older age groups.

Table 4. Estimates of Specificity for the Manitoba Administrative Case Definition

	Manitoba Case Definition	95% Credible Intervals		Manitoba case Definition
	Mean	Lower	Upper	Median
Females				
<55	0.997	0.993	0.999	0.998
55-74	0.859	0.792	0.986	0.844
65-84	0.867	0.742	0.991	0.865
Males				
<55	0.998	0.993	0.999	0.998
55-74	0.884	0.768	0.993	0.883
65-84	0.847	0.722	0.985	0.842

The final output of our Bayesian estimation method is estimates of the apparent prevalence of diabetes (Table 5). The apparent prevalence of diabetes increases with age and is higher in older age groups, but drops in the oldest age group. This result is consistent with other research on diabetes prevalence as this population has higher mortality and a lower life expectancy. These estimates have small CI's in the younger age groups, but the CI's become

large in older age groups. In order to converge estimates for parameters of interest, it was necessary to assume that the sensitivity of case definitions remained constant within age group pairs, but we were able to estimate apparent prevalence for independent age groups.

Table 5. Bayesian Estimates of Diabetes prevalence Manitoba rule 2009

	Mean Prevalence	95% Credible Intervals Lower	Upper
Females			
<45	0.079	0.065	0.092
45-54	0.297	0.250	0.347
55-64	0.277	0.210	0.378
65-74	0.303	0.154	0.456
75-84	0.241	0.108	0.400
Males			
<45	0.048	0.036	0.059
45-54	0.362	0.306	0.421
55-64	0.366	0.253	0.481
65-74	0.408	0.257	0.557
75-84	0.239	0.102	0.431

Estimates of the apparent prevalence of diabetes were also calculated from cross-tabulations of the Manitoba administrative and EMR case definitions (Table 6). The Manitoba administrative case definition yields much higher apparent prevalence estimates of diagnosed diabetes when compared to the EMR case definition across all age groups except in the under 45 age group. The higher apparent prevalence estimates of the administrative data are as expected since this data source captures information on all individual within the UCR, regardless of where they access physician or hospital services, while the EMR registry only capture cases of diagnosed diabetes who are seen within the local community health clinics. Males have a higher apparent prevalence of diabetes than females, as is consistent with diabetes trends in other populations. When prevalence estimates from cross-tabulations of our data and Bayesian estimates of parameters are compared, we see that the Bayesian analysis estimation method yields higher estimates of apparent prevalence for females, except in the under 45 age group. For

males, the Bayesian method had higher estimates of apparent prevalence than the cross-tabulation method, again except in the < 45 age group.

Table 6. Diabetes Prevalence Estimates from Cross Tabulation of EMR and Admin Data (Manitoba rule)

	Manitoba Administrative Case Definition	EMR Diabetes Registry Case Definition
Female		
<45	4.79	5.82
45-54	22.41	17.46
55-64	29.48	24.01
65-74	34.65	15.84
75-84	30.10	12.62
Male		
<45	2.90	3.33
45-54	24.76	20.87
55-64	36.40	26.50
65-74	39.71	27.94
75-84	30.91	14.55

Table 7 shows the PPV's for the three alternative case definitions that were compared within this study. The Manitoba rule was found to have the highest overall PPV's of the administrative case definitions across both sexes.

Table 7. Estimates of Positive Predictive Values for Administrative Case Definitions*

	Administrative Case Definition		
	1 in 1 (%)	2 in 1 (%)	Manitoba Rule (%)
Females			
<45	12.0	7.2	13.1
45-54	39.6	26.7	42.7
55-64	60.5	36.3	62.6
65-74	59.6	44.0	67.7
75-84	52.0	37.3	60.5
Males			
<45	7.2	3.4	7.8
45-54	46.7	30.2	48.9
55-64	63.4	48.5	68.4
65-74	70.5	53.1	68.8
75-84	53.1	32.6	50.6

* PPV calculated from Bayesian estimated parameters using formula:

PPV= (Prevalence)(Senesitivity)/(Prevalence)(Senesitivity) + (1- Prevalence)(1- Senesitivity)

Discussion

To our knowledge, this is the first study to validate administrative case definitions within an Aboriginal population at the sub-provincial level. There have been many studies examining the prevalence of diabetes in Aboriginal communities at the provincial level using administrative data (Dyck, 2010., Johnson, 2009., Martens, 2007., Hemmelgarn, 2007., Green, 2003), and studies examining Aboriginal diabetes prevalence at the community level using diabetes registries, (Horn, 2007) and chart review (Brassard, 1993), but this is the first study to our knowledge to validate administrative case definitions at the local level. Our estimation method allowed us to simultaneously estimate the sensitivity and specificity of our administrative case definition, while estimating the prevalence of diabetes within our population of interest without relying on a gold-standard data source.

Our estimates of the sensitivity and specificity of the Manitoba administrative case definition in these Aboriginal communities were generally lower than those obtained from other studies in Ontario (Hux, 2002) and Nova Scotia (DCPNS, 2009). Hux et al, (2002) found that the sensitivity of the Manitoba administrative case definition was 86% overall, with estimates of specificity at 97%. Our results showed lower sensitivities, except in the over 75 age group for females, and lower specificities except in the females under the age of 55. In validation work which used similar methods to this study, the Diabetes Care Program of Nova Scotia (2009) showed similar estimates of sensitivity of the 1 in 1 administrative case definition in all groups, but estimates of specificity from .95 to .99 across all age groups, as compared to specificity estimates in this Aboriginal population which spanned between .84 to .99.

These results indicate that the validity of administrative case definitions in Aboriginal communities cannot be assumed to mirror estimates of validity for the general population. Many

studies using administrative data to study Aboriginal populations have, essentially, made this assumption (Dyck, 2010., Johnson, 2009., Martens, 2007., Hemmelgarn, 2007., Green, 2003), and as a result may have obtained biased estimates of prevalence. The lower specificity estimates as compared to the published validation data indicate that when administrative data is used to identify cases of diagnoses diabetes, there are likely to be many false-positive cases. This could result in an overestimation of the true prevalence of diabetes within these communities

This concern may not be specific to Aboriginal communities, as it is likely that the validity of administrative data may vary considerably across smaller populations and communities. This is due to the fact that a smaller number of physicians typically serve these populations, and thus deviations in coding practice in even a few of these individuals could result in significant misclassification error in administrative data for these populations. Differences in the salience of diabetes within these small populations could, for example, also result in higher rates of miscoding of pre-diabetes as diabetes within administrative data.

Our estimates of diabetes prevalence were considerably higher than reported for the general population when using administrative data at the national or provincial level (Health Canada, 2003., DCPNS, 2009). Our apparent prevalence for this Aboriginal population was approximately 2.5 times higher in older age groups than was found in a validation study which used similar methods in the general population of Nova Scotia (DCPNS, 2009). However, similarly high estimates have been obtained in several studies examining Aboriginal diabetes which rely solely on administrative data (Dyck, 2010, Oster, 2009, Green, 2003), and in several studies of Aboriginal communities which do not solely rely on administrative data (Oster, 2009, Kaler, 2006). Further, when prevalence estimates were presented to the health directors of the

five Aboriginal communities, they were not surprised, and commented that they were in line with their impressions of diabetes rates within their communities.

Even when combining data from all five communities of interest, estimates of parameters of interest had large 95% credibility intervals, indicating that these results are imprecise. This was likely a result of small numbers of diabetics in the larger age groups within these communities, as a large proportion of this population is made up of individuals in the < 45 age group. This effect could also be due to poor model fit. If the model is an oversimplification of the relationship between our data and parameters of interest, this would result in biased estimates as well as large CI's, regardless of statistical power.

The model assumed conditional independence in the error between data sources, and it is plausible that this assumption is violated. This assumption is not likely to be true as we found that there was a large percent agreement between data sources, and kappa statistics were high across most age groups. This dependence is likely a consequence of the use of EMR to submit physician billings claims. Thus, if a person is missed in the EMR diabetes registry, it is less likely that they will be identified within the administrative data. Within this study it was not possible to estimate the effect of dependence between data sources as there were not sufficient numbers of identified diabetes or diagnostic data sources to support models which could estimate parameters of interest in the presence of dependence between data sources.

Given the similarity of apparent prevalence estimates to the published literature, it is important to note that the PPV's estimated from this study indicate that many diabetics identified from administrative data alone could be false positive cases when using the Manitoba administrative case definition. This is the result of low estimates of the specificity of this administrative case definition. These low PPV values again indicate that researchers and health

policy planners should be cautious when interpreting estimates of diabetes based solely on administrative data, and should be aware that the risk of misclassification error is substantial.

Conclusion

These analyses show that the Manitoba administrative case definition has lower sensitivity and specificity when applied to Aboriginal communities at the sub-provincial level for identifying cases of diagnosed diabetes. This case definition has low Positive Predictive values in older age groups, and thus likely overestimates the prevalence of diagnosed diabetes within Aboriginal communities at this population level. Apparent prevalence estimates from diabetes surveillance systems which rely solely on administrative data should be interpreted with considerable caution, and alternative methods for identifying cases of diagnosed diabetes should be identified. Given these cautions, it is likely that diagnosed diabetes is more prevalent within some Aboriginal communities at the sub-provincial level as compared to the general population.

References

Aboriginal Affairs and Northern Development Canada. (2011). Aboriginal peoples and communities. Accessed: August 16th, 2011 from: http://www.ainc-inac.gc.ca/ap/index-eng.asp

Brassard, P., Robinson, E., Lavallee, C. (1993). Prevalence of diabetes mellitus among the James Bay Cree of northern Quebec. Canadian Medical Association Journal. 149(3), 303-308.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement., 20, 37-46.

Delise, H., Ekoe, J. (1993). Prevalence of non-insulin-dependent diabetes mellitus and impared glucose tolerance in two Algonquin communities in Quebec. Canadian Medical Association Journal., 148(1), 41- 47.

Diabetes Care Program of Nova Scotia. (2009). Development of a Nova Scotia diabetes repository. Diabetes Care Program of Nova Scotia, Halifax, Nova Scotia.

Dyck, R., Osgood, N., Lon, T., Gao, A., Stang, M. (2010). Epidemiology of diabetes mellitus among First Nations and non-First Nations Adults. CMAJ., 182(3), 249-256.

Green, C., Blacnchard, J., Young, T., Griffith, J. (2003). The epidemiology of diabetes in the Manitoba-registered First Nation population. Diabetes Care., 26(7), 1993-1998.

Health Canada. (2003). Repsonding to the challenge of diabetes in Canada: first report of the national diabetes surveillance system (NDSS) 2003. Health Canada. Ottawa, Ontario.

Horn, O., Bruegl, A., Jocobs-Whyte, H., Paradis, G., Ing, A., Macaulay, A. (2007). Incidence and prevalence of type 2 diabetes in the First Nation community of Kahnawa:ke, Quebed, Canada, 198-2003. Revue Canadienne De Sante Publique., 98(6), 438-444.

Hui, S., Walter, S. (1980). Estimating the error rates of diagnostic tests. Biometrics., 36, 167-171.

Hux, J., Ivis, F., Flintoft, V., Bica, A. (2002). Determination of prevalence and incidence using a validated administrative data algorithm. Diabetes Care., 25(3), 512-516.

Johnson, J., Vermeulen, S., Toth, E., Hemmelgarn, B,m Raph-Campbell., K., Hugel, G., King, M., Crowshoe, L. (2009). Canadian Journal of Public Health., 100(3), 231-236.

Kaler, S., Ralph-Campbell, K., Pohar, S., King, M., Laboucan, C., Toth, E. (2006). High rates of metabolic syndrome in a First Nations community in western Canada: prevalence and determinants in adults and children. International Journal of Circumpolar Health., 65(5), 389-402.

Lunn, D., Thomas, A., Best, N., Spiegelhalter, D. (2000). WinBUGS- a Bayesian modeling framework: concepts, structure, and extensibility. Statistics and Computing., 10, 325-337.

Martens, P., Martin, B., O'Neil, J., MacKinnon, M. (2007). Diabetes and adverse outcomes in a first nations populations: association with healthcare access, and socioeconomic and geographical factors. Canadian Journal of Diabetes., 37(3): 223-232.

Oster, R., Toth, E. (2009). Differences in the prevalence of diabetes risk factors among First Nation, Metis and non-Aboriginal adults screening clinics in rural Alberta, Canada. Rural and Remote Health. 9. 1170-1178.

Hemmelgarn, B., Toth, E., King, M., Crowshoe, L., Ralph-Campbell, K. (2009) Chapter 10. diabetes and the status Aboriginal population in Alberta in Johnson, J. (ed). Alberta diabetes atlas 2009. Edomonton: Institute of Health Economics.

Tu, K., Manuel, D., Lam, K., Kavanagh, D., Mitiku, T., Guo, H. (2011). Diabetes can be identified in an electronic medical record using laboratory tests and prescriptions. Journal of Clinical Epidemiology., 64. 431-435.

Yu, C., Zinman, B. (2007). Type 2 diabetes and impaired glucose tolerance in Aboriginal populations: a global perspective. Diabetes Research and Clinical Practice., 78, 159-170.

CHAPTER 4: Manuscript 2

Data Source Dependence in the Estimation of Diabetes Prevalence within the General Population of Nova Scotia

Introduction

Administrative data is often relied upon to provide population-based, ongoing data for diabetes surveillance systems. Despite the fact that diabetes coding in administrative data has been found to have high validity in many studies (Hux, 2002., Health Canada, 2003., Johnson, 2009) the magnitude of misclassification errors due to coding error remains a significant concern (Strom, 2001.; Feinstein, 1989; Kephart, 2004).

Assessment of the validity of administrative case definitions for diabetes has typically relied on a gold standard data source (assumed to have perfect sensitivity and specificity) for estimates of test parameters (Valenstein, 1990., Greiner, 2000, Saydah, 2004). When a gold standard data source is available, it is assumed that the disease status of each individual within this source is known with certainty, and the validity of another data source is determined in comparison with the gold standard. Choices of gold standard data sources within diabetes surveillance have included hospital chart abstraction (Wilson, 2001, Guttmann, 2010), health care provider report (Gambassi, 1998) and patient self-report (Hebert, 1999). Unfortunately, the criteria for a gold reference standard are rarely met, as there are many types of error which can impede the construction of an ideal reference standard (Lawrence, 1995., Reitsma, 2009). While validation studies have typically identified data sources with gold standard cases of disease (individuals we are certain have a condition of interest), they often do not contain good gold standard non-cases of diabetes (individuals we are certain do not have a condition of interest).

To address this issue, there has been considerable interest in methods for estimating the sensitivity and specificity of administrative case definitions in the absence of a gold standard (Rutjes, 2006., Enøe, 2000.,Valenstein, 1989). These methods have the advantage of simultaneously estimating and adjusting for the sensitivity and specificity of case definitions while producing prevalence estimates (Joseph, 1995), and where Bayesian estimation is used, can incorporate prior knowledge and information.

While these methods relax the assumption of gold standard accuracy in comparator data, many studies that use these methods also assume that errors between alternative case definitions or tests are conditionally independent (DCPNS, 2009., Joseph, 1995). Two measures are conditionally independent if the sensitivity and specificity of one measure does not depend on whether the subject tests positive or negative to the second test. Conditional dependence implies that the sensitivity of one test varies depending on the result of the other test. Incorrectly assuming that error is conditionally independent between data sources can result in biased estimates of parameters of interest in validation studies (Gardner, 2000., Brenner, 1996., Torrance-Rynard, 1997., Vacek, 1985). For example, when Dendukuri & Joseph (2001) re-examined data from a previous study which assumed that the result from multiple diagnostic tests were independent (Joseph, 1995), they found that including parameters which account for test dependence resulted in substantial changes in estimates of test parameters and disease prevalence.

Conditional covariance of error is a critical issue for the validation of diabetes administrative case measures, and the application of these measures to health research. To the degree conditional covariance in error exists, estimates of sensitivity and specificity obtained from validation of case definitions against other measures, such as clinical registries, will not be

generalizable to the general population. For example, some validation studies have relied on clinical registries for gold standard cases of disease (DCPNS, 2009). Patients attending these clinics are generally a subset of the population with diabetes, and since they are referred to these centres by physicians, they may be more likely to be coded as having diabetes in administrative claims data. Referring physician who perceive diabetes as salient and a priority, will thus be more likely to code diabetes in physician claims and charts. Conversely, if diagnosed diabetes is not as salient (e.g. in an elderly patient with many other more pressing health problems, or lower severity of diabetes), then coding diabetes as the most responsible diagnosis within a reimbursement claim and referring the patient to a local diabetes care centre may both be less likely. Based on this example, the estimated sensitivity of administrative data in a subset of persons identified as gold-standard cases of disease may be higher than in the total diseased population.

Methods for relaxing the conditional independence assumption between data sources have also received considerable attention (Dendukuri, 2001., Gardner, 2000., Torrance-Rynard, 1997., Vacek, 1985). Generally, these methods assess and control for dependent error through the inclusion of covariance terms in the estimation models. However, in the context of the validation of administrative case definitions, a simple and plausible example of dependence of error is that the sensitivity and specificity of an administrative definition differs depending on whether or not it is identified in a gold standard data source, such as a disease registry. For example, independence of error would imply that the sensitivity of an administrative case definition for those in a disease registry is the same as the sensitivity for cases that are not. If this did not hold, it would be an example of dependent error.

The purpose of this study was to assess the generalizability of estimates of the validity administrative case definitions that are obtained from comparing administrative data to two alternative gold standard data sources: clinical registry data and drug claims data. We estimated conditional dependence in the sensitivity between an administrative case definition of diabetes and two different gold standard measures of persons with diabetes, and examined the potential effect of erroneously assuming conditional independence on estimates of sensitivity and prevalence.

Methods

Data and subjects

This study used population-based data previously assembled for the validation component of a pilot project, conducted in 2009 to develop a Nova Scotia Diabetes Repository in Nova Scotia (NSDR) (DCPNS, 2009). The purpose of the NSDR was to assemble information from a number of different data sources to facilitate diabetes surveillance and research, while ensuring security and confidentiality.

The validation component of the NSDR assembled linked data on diagnoses of diabetes, from multiple data sources, for all persons in Nova Scotia who were registered with the provincial health insurance program between 2005-2006 (n= 1,006,687). With the exception of excluding a small percentage of Nova Scotians who have their health care costs covered through other programs, such as the Canadian Armed Forces and the RCMP, the data covers the whole population. Data on diagnoses of diabetes status from multiple data sources, shown in Table 1, were linked using encrypted health card numbers. All records were linked at the Population Health Research Unit at Dalhousie University, which maintains and linked administrative data for research purposes.

Table 1. Data Sources and Description (DCPNS, 2009).

Data Source	Description
Diabetes Care Program of Nova Scotia (DCPNS) Diabetes Registry	The DCPNS maintains records for all new referrals to 38 of the provinces Diabetes Care Centres from April 1st 1994 onwards. This clinical registry provided information on Diabetes Status of individuals identified in the MSI person registry who were found in the DCPNS registry during the study period.
Nova Scotia Pharmacare Program (NSPP) Drug Claims Database	The NSPP maintains records for Nova Scotians enrolled in the provincially funded Pharmacare Program from 1989 forwards. For this project, only data from the NSPP seniors program were used. Data pertaining to seniors' (over age 65 with a valid HCN and who do not have another benefit program since 1999) prescriptions are collected from all NS pharmacies that submit claims for reimbursement. This database provided information on prescription claims of individuals identified in the MSI person registry during the study period.
Nova Scotia Atlee Perinatal Database	A database which contains information about all live born infants and foetuses born of less than 20 weeks gestation and their mothers, in Nova Scotia Hospitals or select out of province hospitals to Nova Scotian mothers from 1998 forward. The database also collects information about the mothers of the infants mentioned above. This database was used to exclude cases of gestational diabetes from the study population.
Canadian Institute for Health Information Discharge Abstracts Database (CIHI-DAD)	This database contains detailed information about all discharges from NS hospitals from 1996 forward. This database provided information on discharge diagnosis of individuals identified in the MSI person registry during the study period.
Medical Services Insurance (MSI) Claims	This database contains claims data for health services rendered by a physician and reimbursed through Nova Scotia's MSI program from January 1st 1996 forward. This database provided information on diabetes diagnosis of individuals identified in the MSI person registry during the study period.

Measures

Using the data sources in Table 1, alternative case definitions of diagnosed diabetes were constructed (Table 2). Two of the data sources, the CIHI-DAD and the MSI claims data, were used to construct an administrative case definition for diabetes. While a two-year case definition is most commonly used in Canadian research (2 or more physician claims or 1 or more hospital claims in 2 years), we employed a simple one-year case definition for this study (1 or more physician or hospital claims in a year), as we wished to explore dependence within a single year

of data, and two-year case definitions were not available in our data sources. Previous work shows that a 1-in1rule has slightly higher sensitivity and slightly lower specificity than the commonly used two-year rule; however, these differences are small and should have minimal impact on the results of our study (DCPNS, 2009). Any physician or hospital claims for diabetes which corresponded to gestational diabetes cases in the Atlee Perinatial Data were excluded from the administrative case definitions.

The NSPP-S and DCPNS Registry were used to create two other measures of diagnosed diabetes status. Both measures identify gold standard cases of diabetes (i.e. we expect near perfect specificity and no false positive cases), but only include a subset of all diabetic cases in the population (i.e. low to moderate sensitivity). From the NSPP-S data, beneficiaries with one or more claims in 2006 for an oral anti-hyperglycaemic or insulin medication were classified as having diabetes. Persons with claims for diabetic supplies, such as test-strips, were not classified as having diabetes, as previous validation work found that this results in many false-positive cases (DCPNS, 2009). The NSPP-S measure only captures diabetes cases which are drug treated and covered under the drug insurance program. The program covers Nova Scotia residents over the age of 65. Excluded are seniors who opted out of the program or who were covered by other drug plans.

DCPNS data includes records for all NS residents with one of more visits to one of 38 Diabetes Centres in Nova Scotia, who were clinically diagnosed as individuals with diabetes (DCPNS, 2009). The percent of adults with diabetes who visit one of these centres is estimated to be in the range of 60-80% on average, with lower capture rates among the elderly (DCPNS, 2009). Because diagnostic data is based on clinical assessment, the number of false positives should be negligible.

We cross-classified all persons in the population by the three measures of diabetes status as presented in Table 2. However, for the population under age 65, no cases are identified by the NSPP-S, and thus only two measures of diabetes status were available for this population. In addition, the data included information on each persons' sex and age (<45, 45-54, 55-64, 65-74, 75-84, 85+). Data was provided to the researchers as an aggregated cross-classification, and small cells (1-4 persons) were assigned a random number between 1 and 4 with a probability corresponding to the frequency distribution of such cells.

This project received ethics approval from the Dalhousie University Research Ethics Board, and Data Access approval from the Diabetes Care Program of Nova Scotia Data Access Committee. All data linkage was conducted through the Population health Research Unit at Dalhousie University

Table 2. Data Source Case Definitions

Data Source	Case Definition
DCPNS 1992 forward	Any NS resident with a valid HCN, who is eligible to receive health care services under the MSI program, who has made one or more visits to a Nova Scotia Diabetes Care Centre, and who is coded as having the following types of diabetes: Type 1 Diabetes, Type 2 Diabetes.
NSPP-S 1989 forward	Any senior with one or more claims for insulin or an oral antihyperglycemic agent during the study period
Administrative Data (CIHI-DAD & MSI claims)	One or more physician or hospital claims with a diagnosis of diabetes (ICD-9 250) within the previous year

Statistical Analysis

We estimated conditional dependence in sensitivity between the administrative case definition and the DCPNS and NSPP-S measures. We did not estimate conditional covariance in

specificity, as research shows that the specificity of administrative case definitions are very high (generally greater than .92 (DCPNS, 2009., Hux, 2002)) and the specificity of the DCPNS and NSPP-S measures are, essentially, perfect. If one measure has perfect specificity, then the measures are by definition conditionally independent (Gardner, 2000).

As is conventional in the literature, we measured conditional dependence using the conditional covariance (Dendukuri, 2001., Gardner, 2000., Brenner, 1996). If two tests are administered to a population of diseased individuals, and *p* is the probability of testing positive to both tests, then the conditional covariance for sensitivity (λ_{Se}) is *p* minus the product of the two test sensitivities ($\lambda_{Se} = p - Se1*Se2$). The magnitude of λ_{Se} depends on the magnitude of the sensitivities, and thus is not directly comparable between different models. To enable comparison, we expressed λ_{Se} as a percent of its maximum value, which is readily computed using the estimates of the two sensitivities (Gardner, 2000).

We employed two different approaches to estimating the λ_{Se}. The first approach took advantage of the availability of two over-lapping gold standard measures of having diabetes in the population over age 64. This permitted us to identify a diseased population using one measure (e.g. DCPNS), and directly estimate λ_{Se} between the other two measures (e.g. between the administrative measure and NSPP-S). We estimated λ_{Se} between the administrative measure and NSPP-S for the diabetes population identified using DCPNS, and λ_{Se} between DCPNS and the administrative measure using the diabetes population identified by NSPP-S. These two sets of estimates of λ_{Se} cover only a subset of all diabetes cases, and may not be generalizable to the full population of diabetics. However, this method has the benefit of direct calculation of λ_{Se}. The two sets of λ_{Se} estimates, one for diabetic cases in the NSPP-S and one for diabetic cases identified in the DCPNS, were computed by sex and the three age groups over age 65.

The second method for estimating λ_{Se} employed Bayesian estimation methods commonly used to estimate the sensitivity and specificity of diagnostic tests in the absence of a gold standard. We adapted a model developed by Hui and Walter (1980) to estimate the sensitivity and specificity of administrative case definitions and our data sources containing gold standard cases of disease. The original model that Hui and Walter conceptualized assumed that the test were conditionally independent. We modified the model to included terms for conditional covariance in the sensitivities. As shown in Appendix A, the model specifies the relationship between the observed data (cross-classifications of measured diabetes status according to the two tests in each population) and seven model parameters: the sensitivity and specificity of each test, the disease prevalence in each population, and the conditional covariance in the sensitivities between the two tests. The model treated age groups as separate populations, as the prevalence of disease rises sharply with age. Using DCPNS data, we estimated models using the following three sets of age groups: <45 and 45-54; 55-64 and 65-74; 75-84 and 85+ as separate populations. Separate sets of models were estimated for males and females.

Model parameters were estimated by Bayesian methods using WinBUGS software (Lunn, 2000), which utilizes Gibbs sampling, a Markov Chain Monte Carlo technique, to generate estimates of model parameters. It was necessary to include informative prior information on the prevalence of diabetes within each population in order to converge estimates of parameters of interest. Prevalence estimates were obtained from earlier work with this dataset (DCPNS, 2009) which assumed independence of data sources. One prevalence estimate from this previous work was identified as inconsistent (for males age 65-74), thus an alternative estimate was used for this group. To assess the effect of entering information on prevalence, models were re-run with +5, +10 and +20% prevalence estimates. Because the DCPNS data source was

considered to contain only gold standard cases of disease, we assumed the specificity of this data sources was nearly perfect (1 coding error within 1000 entries entered as a beta distribution), which was entered into the WinBUGS models as an informative prior. Models were estimated based on 50,000 iterations, with a burn in of 5000 iterations. Model convergence was assessed by checking kernel density, trace and history plots.

Results

Direct estimation of conditional covariance λ_{Se} between our administrative data and our gold standard measures shows high levels of dependence between the administrative case definition and the other two measures. Table 4 shows that estimates of conditional covariance ranged between about 20% and 50% of the maximum, suggesting that estimates of sensitivity obtained by comparing administrative case definitions to a gold standard source may not be generalizable to the general population. For example, among female diabetics between the ages of 65 and 74, identified through the NSPP-S data, the conditional covariance in sensitivity between the DCPNS and administrative case definition was 51.5% of the maximum. This translates into substantial differences in the sensitivity of the administrative case definition between those who are also cases in the DCPNS (.917) and those who are not (.808). The former, which corresponds to the estimate that would obtained from classical methods for estimating the sensitivity of administrative case definitions, overestimates the average sensitivity of the administrative case definition by 10.6% (.917 versus .829. Conditional dependence was higher between the NSPP-S measure and the 1in1 measure than it was between the DCPNS measure and the 1in1 measure. For the DCPNS measure, values of λ_{Se} increase in older age groups and are higher for males than females. For the NSPP-S case definition, values of λ_{Se} stay relatively constant across age groups, and are generally higher for males than females, except in

the 75-84 age group. These results indicate that assuming independence between this administrative and our gold standard case definitions would result in biased parameter estimates.

Table 3. Sensitivity of the administrative case definition and dependence estimates between administrative and gold standard case measures for individuals identified as gold standard cases of diabetes within the NSPP-S and DCPNS measures

	NSPP-S data				DCPNS data			
	Se (avg)	Se (in DCPNS)	Se (not in DCPNS)	Cov % of Max	Se (avg)	Se (in Pharm)	Se (not in Pharm)	Cov % of Max
Female								
65-74	0.829	0.917	0.808	51.5	0.895	0.917	0.846	21.0
75-84	0.743	0.861	0.710	46.0	0.810	0.861	0.744	26.9
85+	0.651	0.8	0.606	42.6	0.607	0.8	0.496	49.0
Male								
65-74	0.841	0.918	0.824	48.4	0.884	0.918	0.815	29.0
75-84	0.763	0.885	0.734	51.4	0.811	0.885	0.706	39.0
85+	0.620	0.805	0.587	48.6	0.616	0.805	0.491	49.1

Results from our second approach to estimating λ_{Se}, using Bayesian estimation methods, are shown in Table 5. Results are highly variable and sensitive to prior specification of prevalence, However, credibility intervals are generally narrow across all models, indicating that estimates are precise. Estimates of λ_{Se} using prior prevalence estimates from previous validation, which used the same data but assumed conditional independence, are shown in the top row of each panel in Table 5. They are highly variable and differ considerably from the estimates using the first method (Table 4). Many are negative, which is implausible, and the magnitude of some are close to zero. Accordingly, subsequent models were run specifying different prior prevalences (increasing the prevalence by 5%, 10%, and 20%). Doing so had a large impact on the estimates of λ_{Se}. For example, for females aged between 0-54, an increase of 20% in the informative prior prevalence estimate results is an increase in the covariance percent max from 2.3% to 50.7%. This demonstrates high interdependence between λ_{Se} and prevalence, and that the estimation of one depends upon specifying prior information on the other.

Making the strong assumption that the estimates of conditional covariance observed in Table 4 extend to the full population of diabetics provides a rough basis for selecting which of the estimates in Table 5 are plausible. It is also possible to observe how much higher the prior prevalence would have to be to obtain plausible estimates of λ_{Se} . Plausible estimates should result in λ_{Se} that are positive, and in the range of 20% to 50% of the maximum. For example, for females age 55-74, analysis in table 4 shows that the covariance percent of maximum should be in the range of 21%. To yield model estimates of covariance at approximately these levels, prevalence estimates needed to be increased to 20%. Using this line of reasoning suggests that previous work, which used similar models but assumed conditional independence, may have underestimated prevalence by as much as 15-20%.

Table 5 also shows that estimates of the sensitivity of the case definitions are affected by dependence in errors. As we increase the informative prior prevalence to levels in which the models yield conditional covariance percentage of maximum within an expected range, sensitivity estimates decrease. For example, for females age 55-74, when the prevalence estimates are increased by 20%, we see that estimates of the sensitivity of administrative data decrease from 0.904 to 0.781, a substantial drop. Decreased sensitivity of the administrative case definition will result in increased number of false negative cases of diabetes when using this case definition. Estimates of the specificity of administrative data remain high across age groups, sex, and prevalence adjustment, although specificity dependence between case definitions was not assessed.

Table 4. Sensitivity of Administrative case definition and conditional covariance estimates for individuals identified within the DCPNS measure.

Models	Prevalence Estimate	Se_{admin}	95%CI	Se_{dcpns}	95% CI	Sp_{admin}	Cov / Cov_{max}
Female							
0-54	*	0.855	0.832: 0.880	0.570	0.554: 0.586	0.998	2.3%
	+5%	0.815		0.543		0.998	23.7%
	+10%	0.778		0.570		0.998	36.6%
	+20%	0.716		0.457		0.998	50.7%
55-74	*	0.904	0.890: 0.916	0.614	0.601: 0.627	0.995	-55.1%
	+5%	0.892		0.594		0.991	-39.6%
	+10%	0.825		0.567		0.999	-1.2%
	+20%	0.781		0.52		0.999	30.7%
75-85+	*	0.812	0.748: 0.849	0.54	0.521: 0.557	0.993	-44.5%
	+5%	0.774		0.514		0.993	-19.8%
	+10%	0.700		0.460		0.995	8.7%
	+20%	0.641		0.422		0.994	23.8%
Males							
0-54	*	0.834	0.815: 0.851	0.529	0.515: 0.544	0.999	9.1%
	+5%	0.795		0.505		0.999	26.2%
	+10%	0.758		0.482		0.999	37.9%
	+20%	0.697		0.441		0.999	50.7%
55-74	*	0.903	0.884: 0.915	0.592	0.580: 0.603	0.997	-57.1%
	+5%	0.872		0.568		0.998	-19.4%
	+10%	0.831		0.542		0.998	8.7%
	+20%	0.762		0.497		0.998	35.5%
75-85+	*	0.826	0.759: 0.859	0.547	0.527: 0.566	0.992	-46.3%
	+5%	0.781		0.521		0.990	-16.6%
	+10%	0.745		0.498		0.990	-0.7%
	+20%	0.687		0.456		0.991	17.6%

*Estimates of prevalence were derived from previous validation work which used similar methods, but assumed independence between data sources

Discussion

This study quantitatively demonstrates that there is significant conditional dependence between administrative and the gold standard case definitions used for diabetes surveillance within Nova Scotia. Both direct computation of conditional dependence, as well as differences in the sensitivity of administrative data depending on whether individuals were identified in one or both gold standard databases support this finding. Previous work shows that when conditional

dependence is present, but not accounted for in estimation models, it can result in underestimation of test error rates (Brenner, 1996., Torrance-Rynard, 1997), overestimation of specificity and especially sensitivity of test parameters(Torrance-Rynard, 1997) , and poor estimation of the true prevalence of disease (Georgiadis, 2003., Vacek, 1985).

This study also suggests that previous work using this data source (DCPNS, 2009) underestimated the prevalence of diabetes within this study population. This previous work assumed independence between data sources, which was likely the main contributing factor to bias in these parameter estimates. Our work shows that the prevalence of disease may be up to 20% higher than previously estimated in age groups over 65, especially within the oldest age groups. While we could not directly estimate the magnitude of dependence in age groups under 65 (as there was no second data source which contained gold standard cases of disease), it is likely that there is significant conditional dependence between the DCPNS data source and administrative data within this population, and thus that the prevalence of diabetes has been underestimated. Sensitivity analysis of our models regarding our prevalence estimates highlights a potential weakness of our study. Our models are very sensitive to the prevalence prior information that we enter. Small variation in specific prevalence estimates causes models to behave erratically, as was observed before amending the prevalence estimate for males from validation work which assumed independence (DCPNS, 2009).

While previous work has shown that conditional dependence between data sources contributes to bias in estimation of sensitivity and specificity (Torrance-Rynard, 1997), estimates of these parameters were similar to previous validation work with this data set (DCPNS, 2009). This effect could be partially due to the fact that within this analysis, only models which used the DCPNS as containing gold standard cases and administrative data were used to estimate these

parameters. Models which examined conditional dependence and yielded estimates of test parameters for individuals over the age of 65 using the NSPP-S data source were estimated but yielded inconsistent results and unrealistic parameter estimates. Given the strong theoretical and practical work which has demonstrated that conditional dependence results in overestimation of sensitivity and specificity, (Torrence-Rynard, 2003., Dendukuri, 2001., Georgiadis, 2003., Vacek, 1985) it is possible that we have overestimated these parameters within our analysis. With this possibility, we also were forced to make an assumption that the sensitivity of administrative and DCPNS case definition remained constant across our pairs of age groups. If this assumption is incorrect, it could contribute to error in our estimation of sensitivity and specificity parameters.

Conclusion

These analyses show that there is considerable conditional dependence between an administrative case definition, and case definitions from a clinical diabetes registry (DCPNS), and a drug claims data source (NSPP-S) for individuals over the age of 65. It is likely that conditional dependence is also present between these data source for individuals under the age of 65 as well. This dependence is likely to have contributed to an underestimation of the true prevalence of disease, and biased estimates of sensitivity and specificity of an administrative case definition in previous work. It is important for disease surveillance systems and authors to carefully assess for potential dependence between diagnostic data sources when attempting to validate case definitions and estimate the prevalence of disease.

References

Brenner, H. (1996) How independent are multiple 'independent' diagnostic classifications? Statistics in Medicine. 15, 1377-1386.

Dendukuri, N., Joseph, L. (2001). Bayesian approaches to modeling the conditional

dependence between multiple diagnostic tests. Biometrics. 57, 158-167.

Diabetes Care Program of Nova Scotia (DCPNS). (2009) Development of a Nova Scotia Diabetes Repository. Provincial Report. Halifax, Canada.

Enøe, C., Georgiadis, M., Johnson, W. (2000). Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. Preventative Veterinary Medicine. 45(1-2) 61-81.

Feinstein, A. (1989). Para analysis, faute de mieux, and the perils of riding on a large data barge. Journal of Clincal Epidemiology. 42, 929-935.

Gambassi, G., Landi, F., Peng, L., et al. (1998). Validity of diagnostic and drug data in Standardized nursing home resident assessment potential for geriatric pharmacoepidemiology. Medical Care. 36(2), 167-179.

Gardner, I., Stryhn, H., Lind, P., Collins, M. (2000). Conditional dependence between tests affects the diagnosis and surveillance of animal diseases. Preventative Veterinary Medicine. 45, 107-122.

Georgiadis, M., Johnson, W., Gardner, I., Singh, R. (2003). Correlation-adjusted estimation of sensitivity and specificity of two diagnostic tests. Applied Statistics. 52(1), 63-76.

Greiner, M., Gardner, I. (2000).Epidemiologic issues in the validation of veterinary diagnostic tests. Preventive Veterinary Medicine. 45. 3-22.

Guttmann, A., Nakhla, M., Henderson, M. et al (2010) Validation of a health administrative data algorithm for assessing the epidemiology of diabetes in Canadian children. Pediatric Diabetes. 11. 122-128.

Health Canada. (2003). Responding to the challenge of diabetes in Canada: first report of the national diabetes surveillance system (NDSS) 2003. Health Canada. Ottawa, Ontario.

Hui, S., Walter, S. (1980). Estimating the error rates of diagnostic tests. Biometrics. 36, 167-171.

Hux, J., Ivis, F., Flintoft, V., Bica, A. (2002). Determination of prevalence and incidence using a validated administrative data algorithm. Diabetes Care., 25(3), 512-516.

Johnson, J., Vermeulen, S., Toth, E., Hemmelgarn, B,m Raph-Campbell., K., Hugel, G., King, M., Crowshoe, L. (2009). Increasing incidence and prevalence of diabetes among the status Aboriginal population in urban and rural Alberta, 1995-2006. Canadian Journal of Public Health., 100(3), 231-236.

Joseph, L., Gyorkos, T., Coupal, L. (1995). Bayesian estimation of disease prevalence

and the parameters of diagnostic tests in the absence of a gold standard. American Journal of Epidemiology. 141(3), 263-272.

Kephart, G., Casey, J., Ranger, R., Dunbar, P., Karlovic, K. (2004). The development and validation of an alternative case definition for the national diabetes surveillance system. Population Health Research Unit. Halifax, Nova Scotia.

Lunn, D., Thomas, A., Best, N., Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing. 10, 325-337.

Reitsma, J., Rutjes, A., Kahn, K., Coomarasamy, A., Bossuyt, P. (2009). A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. Journal of Clinical Epidemiology. 62(8), 797-806.

Rutjes, A., Reitsma, J., Di Nisio, B., Smidt, N., van Rijn, J., Bossuyt, P. (2006). Evidence of bias in diagnostic accuracy studies. CMAJ. 174(4), 469-476.

Saydah, S., Geiss, L., Tierney, E. et al (2004). Review of the performance of methods to identify diabetes cases among vital statistics, administrative, and survey data. Annals of Epidemiology. 14. 507-516.

Strom, B., (2001). Data validity using claims data. Pharmacoepidemiology and Drug Safety. 10, 389-392.

Torrance-Rynard, V., Walter, S. (1997). Effects of dependent errors in the assessment of diagnostic test performance. Statistics in Medicine. 16, 2157-2175.

Vacek, P. (1985). The effect of conditional dependence on the evaluation of diagnostic tests. Biometrics. 41(4), 959-968.

Valenstein, P. (1990). Evaluating diagnostic tests with imperfect standards. American Journal of Clinical Pathology. 93(2), 252-258.

Wilson, C., Susan, L., Lynch, A., Saria, R., Peterson, D. (2001). Patients with diagnosed diabetes mellitus can be accurately identified in an Indian health service patient registration database. Association of Schools of Public Health. 116(1) 45-50.

Appendix A: Dependence Model

Population A

Diagnostic 1	Diagnostic 2: no	Diagnostic 2: yes
no	X_{11}	X_{21}
yes	X_{12}	X_{22}

Population B

Diagnostic 1	Diagnostic 2: no	Diagnostic 2: yes
no	Y_{11}	Y_{21}
yes	Y_{12}	Y_{22}

$$X_{11}= [P[D^+]* [1-Se_1]* [1-Se_2] + cov] + P[D^-]* [Sp_1]* [Sp_2]]*n$$
$$X_{21}= [P[D^+]* [1-Se_1]* [Se_2] – cov] + P[D^-]* [Sp_1]* [1-Sp_2]]*n$$
$$X_{12}= [P[D^+]* [Se_1]* [1-Se_2]- cov] + P[D^-]* [1-Sp_1]* [Sp_2]]*n$$
$$X_{22}= [P[D^+]* [Se_1]* [Se_2] + cov] + P[D^-]* [1-Sp_1]* [1-Sp_2]]*n$$

$$Y_{11}= [P[D^+]* [1-Se_1]* [1-Se_2] + cov] + P[D^-]* [Sp_1]* [Sp_2]]*n$$
$$Y_{21}= [P[D^+]* [1-Se_1]* [Se_2] – cov] + P[D^-]* [Sp_1]* [1-Sp_2]]*n$$
$$Y_{12}= [P[D^+]* [Se_1]* [1-Se_2] – cov] + P[D^-]* [1-Sp_1]* [Sp_2]]*n$$
$$Y_{22}= [P[D^+]* [Se_1]* [Se_2] + cov] + P[D^-]* [1-Sp_1]* [1-Sp_2]]*n$$

$P[D^+]$ = Probability of being identified as having Diabetes
$P[D^-]$= Probability of not being identified as having Diabetes
$[Se_1]$= Probability of being identified as a Diabetic conditional on true status as a diabetic
$[1-Se_1]$= Probability of being identified as a non-Diabetic conditional on true status as a diabetic
$[Sp_1]$= Probability of being identified as a non-Diabetic conditional on true status as a non-Diabetic
$[1-Sp_1]$= Probability of being a identified as a Diabetic conditional on true status as a non-diabetic
$[Se_2]$= Probability of being identified as a Diabetic conditional on true status as a diabetic
$[1-Se_2]$= Probability of being identified as a non-Diabetic conditional on true status as a diabetic
$[Sp_2]$= Probability of being identified as a non-Diabetic conditional on true status as a non-Diabetic
$[1-Sp_2]$= Probability of being a identified as a Diabetic within Diagnostic 2 conditional on true status as a non-diabetic
cov= conditional covariance in sensitivity between tests
n= total population $(X_{11}+ X_{21}+ X_{12}+ X_{22})$

Chapter 5: Conclusion

Population level surveillance of diabetes helps to inform health service planning and delivery. Administrative data is commonly used for diabetes surveillance systems, but there are concerns about the validity of common administrative case definitions within Aboriginal populations. Within majority populations, there is also concern regarding the validity of administrative case definitions when combining surveillance data from administrative and other data sources due to data measure conditional dependence. This thesis project addresses these validation issues using data from two Nova Scotian populations.

In our first manuscript, we validated a commonly used administrative case definition for the identification of cases of diagnosed diabetes within an Aboriginal community. We found that the case definitions examined had lower values of sensitivity and specificity than were identified when validated within majority populations. This finding illustrates that caution should be taken when interpreting prevalence estimates of diabetes within Aboriginal communities using the Manitoba administrative case definition, as it could be overestimating the prevalence of diabetes. Given these concerns, it is still likely that Aboriginal populations do experience a higher prevalence of diabetes as compared to majority populations. This work has important implications for other minority populations who may have different diabetes prevalence rates than the general population, as administrative case definitions may contain bias as a consequence of the particular risk factors unique to that population.

Future work should focus on chart review within Aboriginal communities to validate diabetes diagnoses within diabetes surveillance databases. Since diabetes is widely acknowledged as an issue within Aboriginal communities, increased screening and coding of

pre-diabetic individuals as diabetic could be a concern. Clinical validation of diabetes cases would help to improve and validate the findings presented in this study.

The incorporation of conditional covariance estimation into models used to validate administrative case definitions of diabetes is another potential area for future work. Given the high levels of agreement between Aboriginal data sources used in this study, it is likely that conditional covariance is indeed present, which could contribute to bias in parameter estimation. In order to support estimation of conditional covariance, high quality prior information on sensitivity, specificity, or prevalence will need to be identified to support model convergence.

Our second manuscript focused on evaluating the effect of incorporating conditional covariance estimates in models to estimate parameters of interest in the general population of Nova Scotia. We found that high levels of covariance were present, which likely contributed to an underestimation of the prevalence of disease, and an overestimation of the sensitivity of our administrative case definition. This study highlighted that diabetes surveillance systems which incorporate diabetes data from several sources must address conditional covariance between these measures if they are to yield accurate estimates of the prevalence of disease.

Future work within this area should focus on improving the quality of informative prior information to support more precise estimation of conditional covariance terms. Dependence between data sources is complex, but through the identification of high quality additional sources of diabetes surveillance data, informative prior information to converge models which include conditional covariance estimates can improve parameter estimates.

REFERENCES

Assembly of First Nations. (2010) Assembly of first nations- the story. Accessed on: September 15th 2010 at: http://www.afn.ca/article.asp?id=59.

Burrows, N.R., Lojo, J., Engelgau, M.M., Geiss, M.A. (2004). Using survey data for diabetes surveillance among minority populations: a report of the centers for disease control and prevention's expert panel meeting. Preventing Chronid Disease: Public Health Research, Practice, and Policy. 1(2): A03.

Chipkin, S., Gottlieh, P., Bogorad, D., Parker, F. (1996) Endocrine, diabetes, and metabolism in Primary care medicine J. Noble (ed). Mosby-Year Book Inc. United States of America.

Choi, B. (1998). Perspectives on epidemiologic surveillance in the 21st century. Chronic Disease in Canada. 19(4); 152-156.

Canadian Institute for Health Information (CIHI). Discharge Abstract Database. Accessed: July 20th 2010 from: http://www.cihi.ca/cihiweb/dispPage.jsp?cw_page=services_dad_e

Commission of the Future of Health Care in Canada (2002). Building on values: The future of healthcare in Canada. Accessed: July 27th 2010. at: http://dsp-psd.pwgsc.gc.ca/Collection/CP32-85-2002E.pdf

Diabetes Care Program of Nova Scotia. (2007). Nova Scotia NDSS submission v207. December.

Diabetes Care Program of Nova Scotia. (2008) Internal Validation work Comparing Diabetes Care Program Nova Scotia Registry to National Diabetes Surveillance System (Software v 2.08). Diabetes Care Program of Nova Scotia. Halifax, Canada

Diabetes Care Program of Nova Scotia (DCPNS). (2009) Development of a Nova Scotia Diabetes Repository. Provincial Report. Halifax, Canada.

EnØe, C., Georgiadis, M., Johnson, W. (2000). Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. Prevetative Veterinary Medicine. 45; 61-81.

Feinstein, A.R. (1989). Para-analysis, faute de mieux, and the perils of riding on a data barge. Journal of Clinical Epidemiology, 42; 929-935.

First Nations Centre. (2002). First nations longitudinal health survey: Report on process and methods. Accessed on: August 8th, 2010 at: http://www.rhs-ers.ca/english/pdf/rhs2002-03reports/RHS_002-03-Report_on_Process_and_Methods-CONDENSED_VERSION.pdf

First Nations, Inuit and Aboriginal Health Branch (FNIHB) (2010). Diabetes. Accessed: January 25th 2011 at: http://www.hc-sc.gc.ca/fniah-spnia/diseases-maladies/diabete/index-eng.php#a9

Greiner, M. (2003) A decision criterion for the application of the Rogan-Gladen estimator in prevalence studies International Symposia on Veterinary Epidemiology and Economics (ISVEE) proceedings, ISVEE 10: Vina del Mar, Chile, Statistical methods session, p 240, Nov 2003

Health Canada. (2003). Responding to the challenge of diabetes in Canada: first report of the national diabetes surveillance system. Health Canada report No. H39-4/21-2003E, Ottawa.

Hui, S., Walter, S. (1980). Estimating the error rates of diagnostic tests. Biometrics. 36, 167-171.

Hux, J., Ivis, F., Flintoft, V., Bica, A. (2002). Determination of prevalence and incidence using a validated administrative data algorithm. Diabetes Care., 25(3), 512-516.

Johnson, J., Vermeulen, S., Toth, E., Hemmelgarn, B,m Raph-Campbell., K., Hugel, G., King, M., Crowshoe, L. (2009). Increasing incidence and prevalence of diabetes among the status Aboriginal population in urban and rural Alberta, 1995-2006. Canadian Journal of Public Health., 100(3), 231-236.

Joseph, L., Gyorkos, T., Coupal, L. (1995). Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. American Journal of Epidemiology. 141(3), 263-272.

Kephart, G., Casey, J., Ranger, R., Dunbar, P., Karlovic, Z. (2004). The development and validation of an alternate case definition for the national diabetes surveillance system. Health Canada. Canada.

Lawrence, J., Goyorkos, T., Coupal, L. (1995). Bayesian estimation of disease prevakence and the parameters of diagnostic tests in the absence of a gold standard. American Journal of Epidemiology. 141(3) 263-273

Lewis, S., Kouri, D. (2004). Regionalization: Making sense of the Canadian experience. Healthcare Papers. 5(1); 12-31.

Liao, Y., Tucker, P., Okoro, C., Giles, W., Mokdad, A., Harris, V. (2004). REACH 2010 surveillance for health status in minority communites- United States, 2001-2002. Centers for Disease Control and Prevention: Surveillance Summaries. 53(SS-6) 1-36.

Public Health Agency of Canada (PHAC). Report from the national diabetes surveillance system: diabetes in Canada 2009. Accessed on: July 26th, 2010, at: http://www.phac-aspc.gc.ca/publicat/2009/ndssdic-snsddac-09/pdf/report-2009-eng.pdf

Pohar, S., Johnson, J. (2007). Health care utilization and costs in Saskatchewan's registered Indian population with diabetes. BMC Health Services Research. 7(126).

Reitsma, J., Rutjes, A., Khan, K., Coomarasamy, A., Bossuyt, P. (2009). A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. Journal of Clinical Epidemiology, 62; 797-806.

Regional Health Survey (RHS). First Nations Regional Longitudinal Health Survey. Accessed: July 19th 2010 from: http://www.rhs-ers.ca/english/

Royal Commission on Aboriginal Peoples. (1996). Report of the royal commission on Aboriginal peoples. Accessed: July 27th, 2010. at: http://www.collectionscanada.gc.ca/webarchives/20071115053257/http://www.ainc-inac.gc.ca/ch/rcap/sg/sgmm_e.html

Rutjes, A., Reitsma, J., Coomarasamy, A., Bossuyt, P. (2007). Evaluation of diagnostic tests when there is no gold standard: a review of methods. Health Technology Assessment 11(50): ix-51

Sackett, D.L., Haynes, R.B. (2002). The architecture of diagnostic research. BMJ. 324(7336); 539-541.

Statistics Canada. (2010) The Canadian community health survey. Accessed: July 5th 2010 from: http://www.statcan.gc.ca/cgi-bin/imdb/p2SV.pl?Function=getSurvey&SDDS=3226&lang=en&db=imdb&adm=8&dis=2#b3

Strom B.L. (2001). Data validity using claims data. Pharmacoepidemiology of Drug Safety, 10; 389-392.

Thacker, S., Berkelman, R. (1988). Public health surveillance in the United States. Epidemiological Review. 10: 164-190.