Diabetes Care Program of Nova Scotia
Suite 548, Bethune Building
1276 Tower Road, Halifax, Nova Scotia, B3H 2Y9

DEVELOPMENT OF A NOVA SCOTIA DIABETES REPOSITORY

Provincial Report

August 2009

Contact Person

Peggy Dunbar,
Coordinator, Diabetes Care Program on Nova Scotia
Phone: 902.473.3219
Fax: 902.437.3911
Email: peggydunbar@diabetescareprogram.ns.c

Prepared by
Diabetes Care Program of Nova Scotia
Suite 548 Bethune Building
1276 South Park Street
Halifax, Nova Scotia
B3H 2Y9

Acknowledgements

The Diabetes Care Program of Nova Scotia (DCPNS) would like to acknowledge and thank those directly involved in the development of the provisional Nova Scotia Diabetes Repository (NSDR). The successful completion of this project would not have been possible without the tireless efforts of the Project Manager, Pam Talbot and the DCPNS Director of Diabetes Surveillance, Zlatko Karlovic as well as the many hours of dedicated work by our volunteer committee members (see below).

We also want to acknowledge the important contribution of Yan Wang (Population Health Research Unit) and Pantelis Andreou (Biostatistical Consulting Unit) to the completion of the validation component of this project. This work was completed under the guidance of George Kephart, Department of Community Health & Epidemiology, Dalhousie University.

Finally, we would like to thank the Public Health Agency of Canada (PHAC) for the funding that made this important project possible. We also acknowledge the in-kind support of the project partners: DCPNS, Nova Scotia Department of Health (DoH) divisions of Acute & Tertiary Care, Information Access & Privacy, Information Analysis & Reporting, Information Technology; Nova Scotia Pharmacare Program (NSPP); Reproductive Care Program of Nova Scotia (RCP), and Dalhousie University.

NSDR Advisory Committee

- Linda Dodds & George Kephart
  Dalhousie University
- Beth Cummings, Peggy Dunbar, Zlatko Karlovic, Jennifer Payne, & Pam Talbot†
  DCPNS
- Lynn Edwards & Katherine Fraser
  NS DoH Acute & Tertiary Care
- Michelle MacDonald
  NS DoH, Information Access & Privacy
- Jill Casey, Barb Harvie, & Cory Russell
  NS DoH, Information Analysis & Reporting
- John MacCallum
  NS DoH, Information Technology
- Dawn Frail, & Steve Graham
  NS DoH, NSPP
- Becky Attenborough
  RCP

NSDR Technical Working Group

- Linda Dodds* & George Kephart
  Dalhousie University
- Beth Cummings*, Igor Grahovic, Zlatko Karlovic, Jennifer Payne, Robin Read, & Pam Talbot†
  DCPNS
- Jill Casey, George Doyle-Bedwell, Mary Ridgewell, & Cory Russell
  NS DoH, Information Analysis & Reporting
- Charmaine Cooke, Dawn Frail, Steve Graham, & Kenneth Huang
  NS DoH, NSPP
- John Fahey
  RCP

NSDR Interim Data Access & Governance Committee

- Peggy Dunbar & Pam Talbot†
  DCPNS
- Lynn Edwards
  NS DoH Acute & Tertiary Care
- Michelle MacDonald
  NS DoH, Information Access & Privacy
- Jill Casey & Barb Harvie
  NS DoH, Information Analysis & Reporting
- Dawn Frail
  NS DoH, NSPP
- Becky Attenborough
  RCP

† Project Manager
* Ad Hoc

Glossary

CCI	Canadian Classification of Health Interventions
CCP	Canadian Classification of Diagnostic, Therapeutic and Surgical Procedures
CIHI-DAD	Canadian Institute for Health Information Discharge Abstract Database
DC	Diabetes Centre
DCPNS	Diabetes Care Program of Nova Scotia
DHA	District Health Authority
DIN	Drug identification number
DM	Diabetes mellitus
DoH	Department of Health
GDM	Gestational diabetes mellitus
HCN	Health card number
ICD-9-CM	International Statistical Classification of Diseases, Injuries and Causes of Death, Ninth Revision, Clinical Modification
ICD-10-CA	International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, Canadian Enhancement
IPR	Insured Patient Registry
MSI	Medical Services Insurance
NDSS	National Diabetes Surveillance System
NS	Nova Scotia
NSAPD	Nova Scotia Atlee Perinatal Database
NSDR	Nova Scotia Diabetes Repository
NSPP	Nova Scotia Pharmacare Program
NSPP-S	Nova Scotia Pharmacare Program for Seniors
OAA	Oral anti-hyperglycaemic agent
OGTT	Oral glucose tolerance test
PG	Plasma glucose
PHAC	Public Health Agency of Canada
PHRU	Population Health Research Unit
PIA	Privacy Impact Assessment
RCP	Reproductive Care Program of Nova Scotia

Table of Contents

Executive Summary

Specific information about the true prevalence and incidence of diabetes (DM) in Nova Scotia (NS), as with other provinces and territories, is difficult to ascertain. At the national level, the National Diabetes Surveillance System (NDSS) was created to address this information gap; however, it has proven to be insufficient due to its sole reliance on administrative health records. In Nova Scotia, the Diabetes Care Program of Nova Scotia (DCPNS) houses a Registry (DCPNS Registry) with over 70,000 clinically confirmed DM/PreDM cases – a resource that is unmatched in Canada. In additional, other provincial programs in NS capture information about confirmed DM cases for limited populations (e.g., pregnant women, seniors).

There is a need for complete and accurate information about DM and other chronic diseases in NS. Although no single source of data is sufficient to accurately describe the burden of DM, combining the many sources into a single repository holds promise. In the 20-month period between May 2007 and December 2008, the DCPNS partnered with the Nova Scotia Pharmacare Program (NSPP), the Reproductive Care Program of Nova Scotia (RCP), various divisions within the NS Department of Health (DoH), and Dalhousie University to identify and address the organizational and technical requirements related to the development and long-term sustainability of a NS Diabetes Repository (NSDR). As part of the NSDR Pilot Project, the DCPNS and its partners constructed a provisional NSDR and tested its utility and feasibility using seven test requests.

The chief goal of NSDR Pilot Project was to enhance DM surveillance in NS through the development of the provisional NSDR and the accompanying data access and data security protocol. The construction and subsequent testing of the provisional NSDR demonstrated that provincial programs can effectively share data while protecting privacy. Currently, the provisional NSDR is comprised of 107,643 unique DM cases derived the DCPNS Registry, RCP's NS Atlee Perinatal Database, NSPP's drug claims database, and the NDSS. One third of these cases appeared in a maximum of one source, highlighting the importance of a collaborative approach to DM surveillance.

While the provisional NSDR was being constructed, additional work was underway to validate an administrative case definition for the NSDR using Bayesian methodology. The final report for this work was completed in July 2009. The validation work highlighted the need for further refinement of the Bayesian case definition before it can be used to identify additional DM cases for the NSDR. In the interim, it was recommended that DM cases can be adequately identified using the Manitoba Rule in a cross-sectional fashion.

Overall the NSDR Pilot Project was a successful experience. One of the keys to success for the project was the broad base of highly engaged partners. Now that the groundwork has been laid, the challenge is for key stakeholders to address outstanding issues pertaining to the sustainability of NSDR. These issues include the management and ongoing maintenance of the NSDR, the formalization of data access guidelines, amendments to the Privacy Impact Assessment to address data access and governance and to facilitate the addition of new partners, and the synthesis of provisional NSDR protocol into a comprehensive NSDR Policy and Procedures Manual.

The potential of the NSDR has been recognized by a wide range of audiences. The DCPNS leveraged the provisional NSDR to secure federal funding for a 15-month project examining factors associated with time to comorbidity and time to death for a cohort of clinically confirmed DM cases. Moreover, presentations about the NSDR Pilot Project have been well received at the local and national level. A further presentation for an international audience is scheduled for October 2009.

Chapter 1 Why Does Nova Scotia Need a Diabetes Repository?

The Diabetes Care Program of Nova Scotia & Its Mandate

The Diabetes Care Program of Nova Scotia (DCPNS) was established in 1991 and is one of nine provincial programs funded by the Nova Scotia (NS) Department of Health (DoH). Working closely with all Diabetes Centres (DCs) in the province, the DCPNS advises the DoH on service delivery models; establishes, promotes, and monitors adherence to diabetes (DM) care guidelines; provides support, services, and resources to DM healthcare providers; and collects, analyses, and distributes DM-related information and statistics for Nova Scotia. As part of this mandate, DCPNS works collaboratively with many partners, including other provincial programs, to enhance *the health of Nova Scotians living with, affected by, or at risk of developing diabetes*. To carry out this important work, the DCPNS requires accurate and complete information about the true burden of DM in the province; however, this information is difficult to ascertain.

The earliest DCs in NS can be traced back to the 1960s; however, it was not until the 1990s that their approach to care and education was standardized – an achievement accomplished with the assistance of the DCPNS. The DCPNS ensures that these centres promote self-care, survey for and monitor the development/progression of DM complications, and follow national and provincial guidelines for optimal care. The DCPNS supports all DCs with activities focused on knowledge transfer, networking in support of best/better practice, and standardization aimed at quality/equitable care. Much of this important work has been facilitated by the creation of the DCPNS Registry.

A Critical Information Gap

The DCPNS Registry, developed and maintained by DCPNS staff, includes records for all new referrals to the province's DCs since April 1st, 1994. Currently, this Registry contains information for more than 70,000 individuals with DM/PreDM. Although the DCPNS Registry strives to capture all cases of DM in Nova Scotia, certain populations are known to be underrepresented – namely non-attendees of the DCs including the frail elderly. Coverage for the paediatric population, however, is near 100%. Recent validation work (2007/08) demonstrated that the DCPNS Registry captures about 70% of the DM cases 20 years of age and older identified through the NS component of the National Diabetes Surveillance System (NDSS)[1] When interpreting this rate of capture, it is important to note that an unknown percentage of cases in the NDSS are incorrectly classified as having DM when in fact DM is not present and this unknown number increases with each additional year of data used. As such, NS DCs may actually provide services for more than 70% of the province's DM cases.

Although the DCPNS Registry does not contain all cases of DM in the province, the cases that are in the Registry are true – or gold standard –cases. Other provincial program databases also contain gold standard DM cases. The Nova Scotia Atlee Perinatal Database (NSAPD) maintained by the Reproductive Care Program of Nova Scotia (RCP) records data about all pregnancies in Nova Scotia including pre-existing and gestational DM (GDM) status. The Nova Scotia Pharmacare Program (NSPP) drug claims database contains records of prescriptions filled for insulin, oral anti-hyperglycaemic agents (OAA), and DM testing supplies for all persons enrolled in the Nova Scotia Pharmacare Program for Seniors (NSPP-S; ≈75% of NS seniors) and the Community Services Pharmacare Programs. While the cases of DM

captured by the NSAPD and NSPP-S database are true cases, they represent only a subset of the general population (i.e., pregnant women or seniors).

The NDSS is another major source of data about DM in Nova Scotia. This mechanism collects nationally comparable data about DM using a case definition based on hospital discharge abstracts and physician billings. An important distinction between DM cases identified by the NDSS as opposed to those identified by the provincial programs is that the NDSS cases are estimates and not gold standard cases of DM. The NDSS is useful for DM surveillance across relatively short periods of time; however, it tends to overestimate DM prevalence when used across longer periods as false positive cases accumulate over time.[2] Moreover, approximately 90% of NDSS cases are identified using physician billings data[3]; thus, the validity of the NDSS case definition is sensitive to physicians billing practices. With NS physicians progressively receiving remuneration through alternative payment structures that do not necessitate the submission of billing claims in order to receive payment, the validity of the NDSS case definition will likely decline.

An Opportunity to Close the Information Gap

The recognition of this critical information gap is not new. In 2004, a group of key stakeholders (including the DCPNS Program Manager, NS DoH Director of Acute and Tertiary Care and Director of Information and Evaluation) discussed the issue of DM surveillance in a province where DM data are held by multiple custodians. This discussion culminated in the DCPNS commission of a report by Jim Pyra exploring the feasibility of data sharing among the various custodians holding DM data. After one year and 22 key informant interviews, Pyra made 5 key recommendations (September 2005):

1. Build on work of DoH/Provincial Program Data Access/Privacy Working Group
2. Pilot test data sharing mechanism between provincial programs
3. Examine output of Recommendation 1 & 2
4. Review cost recovery mechanisms
5. Ensure output of Recommendation 3 is incorporated into requirements for provincial information systems initiatives & provincial efforts toward meeting national initiatives

The creation of the Department of Health Promotion and Protection, in addition to the DoH, demonstrates the province's growing appreciation of the importance of chronic disease prevention, management, and surveillance. This paradigm shift also highlights the need for information that quantifies the true burden of disease at the population level. Although the richness of the DM data housed in DCPNS Registry is unmatched in Canada, case ascertainment is not complete. Other provincial programs also capture information about gold standard DM cases but only for limited populations. The NDSS was created to fill the information gap about DM prevalence and incidence, but unfortunately, it has proven to be insufficient due to its reliance on administrative health records.

Clearly, there is a need for complete and accurate information about DM in Nova Scotia. Although no single source of DM data is sufficient to accurately describe the burden of DM in the province, combining the many sources of DM data into a single repository holds promise. Nova Scotia is uniquely positioned to become a leading sentinel surveillance site for DM within Canada by supplementing information from the DCPNS Registry with that from other provincial program databases and routinely collected

administrative health records. The development and maintenance of such a repository would greatly enhance DM case ascertainment, providing a surveillance model that could be used for other chronic diseases. Moreover, such an endeavour would foster the growth of strong collaborative partnerships between provincial programs.

The timing of the Pyra Report and the contemporaneous shift toward an integrated chronic disease strategy for the province coincided with a national movement to improve chronic disease surveillance. The Public Health Agency of Canada (PHAC) issued a national call for proposals under the Enhanced Chronic Disease Surveillance Grants competition; successful projects received funding for up to 12 months to address one or more of the following:

1. Develop new & unique data sources for chronic disease surveillance
2. Document rigorous methodology for combining data sources into a surveillance resource
3. Generate data & knowledge transferrable to other jurisdictions

In January 2007, the DCPNS in partnership with RCP, NSPP, and various other divisions within the NS DoH responded to this call, submitting a grant proposal to PHAC for funding to carry out the second recommendation of the Pyra Report – to pilot test a data sharing mechanism between provincial programs through the development of a provisional Nova Scotia Diabetes Repository (NSDR). The NS submission spoke to all of PHAC's objectives. The provisional NSDR would be a new data source to be used for DM surveillance, the methodology for combining the disparate data sources into the provisional NSDR would be documented, and the framework used to construct the provisional NSDR could be applied to other chronic diseases, possibly yielding a NS Chronic Disease Repository.

Chapter 2: Description of the NSDR Pilot Project

Purpose

In an effort to provide the best possible evidence for responsive and responsible decision making, the DCPNS partnered with the RCP, the NSPP, and various other divisions of the NS DoH to enhance DM surveillance in NS by optimizing the use of existing data sources through the development of the provisional NSDR and the accompanying data access and data security protocol.

Objectives

Objective 1: Identify and address the organizational requirements for incorporating various data sources into a sustainable, ongoing DM surveillance mechanism

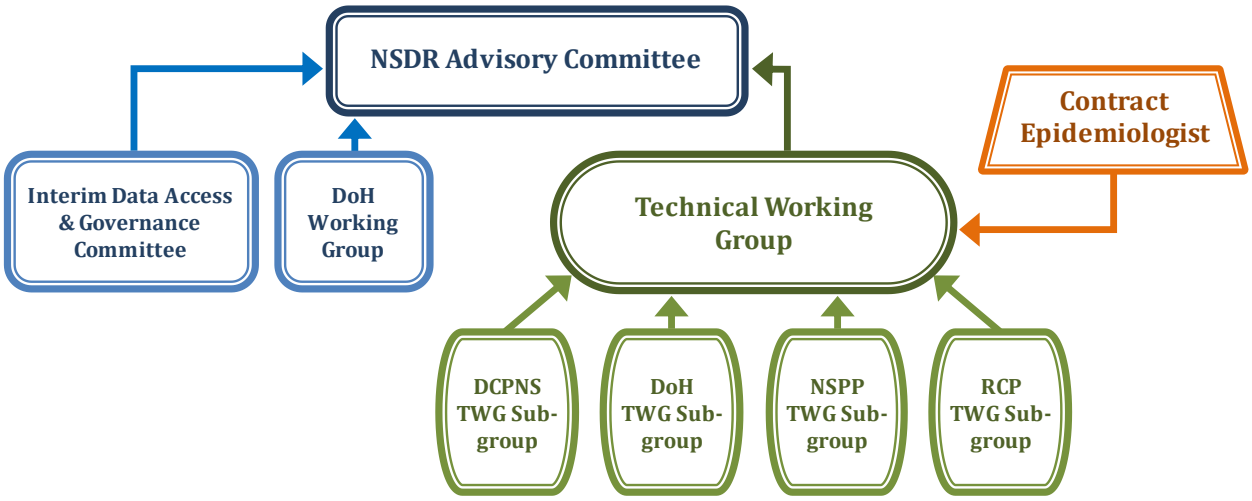
Objective 2: Identify and address the technical requirements for combining DM cases from disparate data sources into a single, sustainable, ongoing DM surveillance mechanism

Objective 3: Validate a Bayesian case definition for the NSDR

Organizational Structure

The NSDR Pilot Project was guided by an Advisory Committee with the assistance of a Technical Working group and various subcommittees with content-specific expertise. Both the Advisory Committee and the Technical Working Group (TWG) were comprised of at least one representative from each project partner as well as other content experts. In addition, an epidemiologist was contracted to validate a new administrative case definition for DM based on Bayesian methodology. The contract epidemiologist reported to the Technical Working Group, which, in turn, reported to the Advisory Committee.

Figure 1: Organization structure for NSDR Pilot Project



Activities

The NSDR Pilot Project was divided into three parallel streams that correspond to the three objectives listed above

Objective 1: Organizational Requirements

The first step in realizing the goal of a NSDR was to identify and address any organizational requirements. The NSDR Advisory committee and various sub-committees carried this work out early in the process. Over the course of several months, numerous meetings were held with key stakeholders to identify issues that could impede the development of the NSDR.

Details about the organizational requirements for the NSDR are presented in Chapter 3 (pp. 8-10)

Objective 2: Technical Requirements

The NSDR Technical Working Group comprised of one or more representatives from each contributing partner examined the technical requirements of the NSDR, in preparation for the day when all the organizational requirements were addressed and the green light was given for the provisional NSDR to be constructed. For practical purposes, smaller sub-groups dealt with program-specific tasks such as the development of program-specific case definitions and program-specific data transfer procedures. The larger group addressed more global issues and reviewed the work of the contract epidemiologist hired to validate a new administrative case definition for the NSDR.

Details about the technical requirements for the NSDR are presented in Chapter 4 (pp. 11-18).

Objective 3: Validation Work

An original partner (and epidemiologist) was contracted to carry out the validation work for the NSDR Pilot Project. This work happened in tandem with the constructions of the provisional NSDR. Due to time constraints of the 12-month funding window, the validation work was completed using an existing data access process through the Population Health Research Unit (PHRU) at Dalhousie University. Although not ideal, linking administrative health records housed at PHRU (including NSPP drug claims database) to data from DCPNS and RCP allowed the validation work to take place while the provisional NSDR was being constructed.

In total, three data access applications (PHRU, RCP, and DCPNS) and two research ethics submissions (Dalhousie, IWK Health Centre) were required for the validation work, a point that highlights the inefficiency of the existing data access process.

Details about the validation work are presented in Chapter 6 (pp. 21-23) and in Appendix A (pp. 29-48)

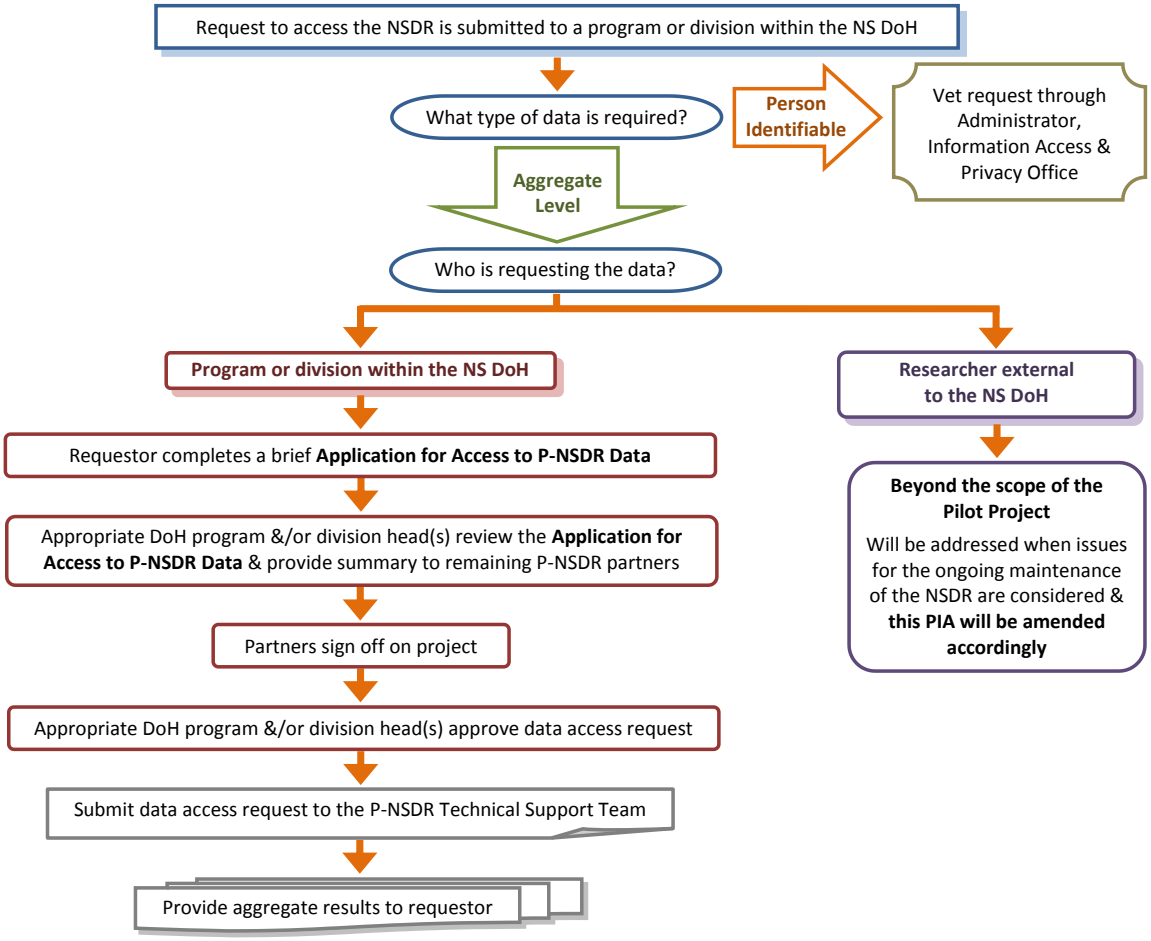
Chapter 3 Organizational Requirements for the NSDR

Privacy & Data Security Issues

Data held by the DCPNS and its partners include personal identifiers thus posing a potential threat to privacy. The NSDR Advisory Committee and the Interim Data Access and Governance Committee (with one or more representatives from each contributing partner) addressed this issue in detail and decided to vet the development of a provisional NSDR through a Privacy Impact Assessment (PIA) with the NS DoH. Key components of the PIA included, but were not limited to, the following: detailed summary of the NSDR Pilot Project; intended scope; data access guidelines; conceptual technical architecture; data flow mechanism; planned data tests; authority to collect, use, and disclose personal information; retention and destruction of personal information; and privacy risks and mitigation. The approved PIA is a living document that can be amended as required (e.g., new partner contributes to the NSDR).

A key challenge to the development of the provisional NSDR was the fact that a separate set of data access guidelines governs access to data held by each of the contributing partners. As such, a potential requestor must complete a separate data access application for each partner. To streamline access to the provisional NSDR, the Interim Data Access and Governance Committee agreed to use a single *Application for Access to Provisional NSDR Data* for data requests from programs and departments within the DoH (see Appendix B, p. 49). Using this process, data requests from within the DoH will be reviewed and summarized by the most applicable program (e.g., DCPNS for DM related requests, RCP for pregnancy related, etc), and the summary will be sent to the remaining partners for sign-off. When all partners sign-off on the request, it will be sent the NSDR Technical Support team to be fulfilled, and aggregate results will be forwarded to the requestor (see Figure 2). Data requests from researchers external to the DoH are beyond the scope of this project.

Figure 2: Provisional NSDR data access process



One of the greatest challenges regarding privacy was the development of protocol to transfer data files containing personal identifiers to the provisional NSDR from contributing partners with data residing outside the NS DoH server. These files must be handled in a manner that minimizes the risk of unintentional disclosure. A suitable resolution to this problem was found after much consultation with the Information Access and Privacy Unit of the DoH. Data transfers between the provisional NSDR and affected partners (RCP & DCPNS) were carried out using the provincial government's secure file transfer system (MOVEit® DMZ). These transfers were initiated only when the receiving party was available to download immediately the incoming file and then delete it from the MOVEit® DMZ server. For more details, see Chapter 4 – Data Handling Protocol (pp.16-17)

Prior to the NSDR Pilot Project, data sharing between the contributing partners was limited to specific projects that received ministerial approval – there was no precedent for an ongoing data sharing agreement. After extensive consultation with the Information Access and Privacy Unit of the NS DoH, it was determined that data sharing for the development and ongoing maintenance of the NSDR is in keeping with provincial legislation governing the collection, use, and disclosure of personal information. The DCPNS and RCP have ministerial authorization under the NS Hospitals Act to allow District Health Authorities (DHAs) to disclose personal information to these organizations. The collection, use, and disclosure of personal information for the NSPP, DCPNS, and RCP is governed by the NS Freedom of

Information and Protection of Privacy Act. Under Section 27 of this act, public bodies including the DoH and all of its programs and services (i.e., DCPNS, RCP, & NSPP) are authorized to disclose personal information for the purpose for which it was obtained or a compatible purpose. Personal information also may be disclosed to a public body for the necessary requirements of government operation.

Jurisdictional Issues

With multiple partners contributing to a single data repository, jurisdictional issues are inevitable. First, it was necessary to clarify who had ownership rights to the provisional NSDR. To date, all contributing partners are programs or divisions of the NS DoH; as such, the ultimate owner of the provisional NSDR is the DoH. It follows that the provisional NSDR resides on a secure server at the DoH. However, access to the provisional NSDR is open to all members of the provisional NSDR Technical Support Team (comprised of one or more representatives from each contributing partner).

A unique issue arose with the NSPP data. Part of the data housed by the NSPP is under the jurisdiction of the NS DoH (i.e., recipients of Seniors' Pharmacare) and part is under the jurisdiction of the NS Department of Community Services (i.e., recipients of Community Services Pharmacare). Only data for the NSPP Seniors' Program was used for this pilot project. In the future, Community Services Pharmacare data will be considered for inclusion in the NSDR.

Test Requests

One of the requirements of the PIA was to include a description of any planned tests that would be performed using the NSDR. As such, the Interim Data Access and Governance Committee proposed seven test requests to be included in the PIA. The following questions were used to evaluate the feasibility of using the provisional NSDR data transfer mechanism to respond to various types of information requests – focussing on the time and human resources required to fulfill the request.

1. How many Nova Scotians have DM (any type)/preDM in fiscal year 2005/06?
2. How many Nova Scotians are identified as having DM (any type) by one, two, three, four, or five of case definitions used by the contributing partners?
3. How many Nova Scotians have DM in fiscal years 2001/02, 2002/03, 2003/04, 2004/05 by DM type (T1, T2, GDM, other), age (<20, 20-34, 35-49, 50-64, ≥65), sex, and DHA of residence (1-9)?
4. How many Nova Scotia women had GDM in 2005/06 by age and DHA of residence?
5. How many Nova Scotia women with GDM eventually develop type 2 DM and what is the time to conversion?
6. Of Nova Scotia residents 65 years of age and over, how many have type 2 DM and are taking an OAA in fiscal year 2005/06?
7. As of fiscal year 2005/06, how many Nova Scotians with DM have had a myocardial infarction?

The results from these planned tests are presented in Chapter 5 (pp. 19-20)

Chapter 4 Construction of the Provisional NSDR

Data Sources

The construction of the provisional NSDR was truly a collaborative effort, involving four data custodians and six data sources. Each of the four contributing partners provided a description of the nature and accuracy of their data. A summary of these descriptions follows:

DCPNS Registry

The DCPNS Registry maintains records for all new referrals to 38 of the provinces DCs from April 1st, 1994 onward. Diabetes Centres with an on-site registry also supply the DCPNS Registry with information about follow-up cases that were referred to their clinics prior to 1994. For DCs without an on-site registry, data about follow-up visits is supplied through the DCPNS/Meditech interface. In addition, the Pregnancy and Diabetes Program at the IWK Health Centre contribute data to the DCPNS Registry for new referrals from 1995 forward.

Data supplied to the DCPNS Registry include the following:

- Patient demographics (e.g., date of birth, date of diagnosis, sex, comorbidities at diagnosis, etc)
    - Abstracted from a standardized physician referral sheet,
- DC statistics (e.g., type of DM, type of treatment, type of visit, etc)
- Indicators of care (e.g., weight, height, blood pressure, blood glucose, self-care behaviours, etc)
    - Abstracted from a standardized flow sheet used by all DCs in the province.

There are several checks in place to ensure that the data held in the DCPNS Registry are accurate. The Registry software has a built-in check to prevent the entry of out-of-range or out-of-province health card numbers (HCNs) as well as a Mod 10 check to validate HCNs. Data from all DCs are updated monthly and additional quality checks are run: sex, date of birth, and date of death are checked against the Medical Services Insurance (MSI) Insured Patient Registry file held by Medavie Blue Cross, and frequencies are calculated to determine if there are outliers or unusual data. Reports of any suspected errors are sent to the originating DCs for correction.

Nova Scotia Atlee Perinatal Database (NSAPD)

The RCP is the guardian of the NSAPD, a database that includes information about all live born infants, all foetuses ≥20 weeks gestation or ≥500g, and all co-multiples of the aforementioned that are delivered in NS hospitals or to NS mothers in select out-of-province hospitals from 1998 forward. The NSAPD also captures information about the mothers of the infants or foetuses mentioned above. From 1980 to 1986, data were collected only for deliveries occurring at the IWK Health Centre (formerly the Grace Maternity Hospital). From 1986 to 1987, data were collected for deliveries occurring at the IWK and all regional hospitals within NS. From 1988 onwards, the NSAPD contains information about all births in NS.

Data for the NSAPD are abstracted from hospital records by trained health records coders using a standardized NSAPD coding manual. Starting in 2003, the NSAPD also includes data extracted from files

submitted to the Canadian Institute for Health Information (CIHI) using ICD-10-CA disease codes and CCI procedure codes. Data supplied to the NSAPD include the following:

- Patient demographics (e.g., admission & discharge information, date of birth, sex, etc)
- Behaviour and Lifestyle factors (e.g., prenatal classes, preconceptional folate intake, etc)
- Labour and birth (e.g., method of delivery, APGAR score; infant weight/length; etc)
- Health outcomes – outcome of infant
- Maternal disease (e.g., endocrine disease, heart disease, neoplasms, drug & chemical abuse, etc)
- Maternal procedures (e.g., anaesthesia, transfusion, drug therapies, etc)
- Infant disease (e.g., apnoea, chromosomal abnormalities, neoplasms, respiratory distress, etc)
- Infant procedures (e.g., immunizations, medications, home oxygen, ,etc)

The technical personnel at RCP perform quality checks for the NSAPD. A Mod 10 check is used to identify invalid HCNs, and failures are checked manually. Mother-infant and inter/intra pregnancy linkages are performed, and failures are checked manually. Finally, NSAPD data are audited periodically.

There is an 18-month window from May 1999 to December 2000 for which the dosage used for an Oral Glucose Tolerance Test (OGTT) was changed from 100g to 75g, but the criteria used by health record coders to define GDM was not changed to reflect the new dosage. The criteria for the fasting plasma glucose (PG) and the 1-hour PG tests did not change (≥5.3mmol/L and ≥10.6mmol/L, respectively); however, the 2-hour PG should have been lowered from ≥9.2mmol/L to ≥8.9mmol/L. As a result, some true cases of GDM may have been missed. Cases with a 2-hour PG ≥8.9mmol/L but <9.2mmol/L would have been coded as follows:

- Fasting PG <5.3mmol/L & 1-hour PG <10.6mmol/L
    - Normal, no GDM
- Fasting PG ≥5.3mmol/L or 1-hour PG ≥10.6mmol/L (but not both)
    - Impaired glucose tolerance
- Fasting PG ≥5.3mmol/L and 1-hour PG ≥10.6mmol/L
    - GDM

There is a separate 33-month window from April 1st 2003 to December 31st 2005 for which ICD-10-CA codes were used alone to determine DM status. During this time, hospital coders were instructed that the coding of DM status was optional unless DM impacted length of hospital stay. Starting January 1st 2006, RCP required that DM status be coded irrespective of any effect on hospital stay.

The NSAPD is 95% complete within 6-9 months of fiscal year end (March 31st) and 100% complete within 18 months of fiscal year end (this 5% of the data are primarily from out-of-province hospitals).

Nova Scotia Pharmacare Program Drug Claims Database – Seniors Program

The NSPP maintains records for Nova Scotians enrolled in the provincially funded Pharmacare Program from 1989 forward. Individuals may enrol in the NSPP through the Seniors' Program or through the Community Services' Program. For NSDR Pilot Project, only the NSPP Seniors' Program was used.

Data pertaining to seniors' prescriptions are collected from all NS pharmacies that submit claims to Medavie Blue Cross for reimbursement. Prior to January 1st, 1999, all NS seniors (≥65 years) with a valid HCN could receive benefits through the NSPP-S even if they were enrolled in other programs (e.g., privately insured, veterans, Armed Forces, RCMP, First Nations, and Federal employees). From January 1st, 1999 forward, seniors with drug coverage through a private insurer or other programs could no longer receive benefits through the NSPP.

Data supplied to the NSPP-S include the following.

- Patient demographics (e.g., date of birth, age, sex, etc)
- Providers & pharmacy information (e.g., provider ID and specialty, pharmacy number, etc)
- Pharmaceuticals (e.g., drug identification number, prescription date, amount dispensed, etc)
- Costing (e.g., payee type, amount claimed and approved, prescription cost, dispensing fee, etc)

Data housed by the NSPP-S were originally collected for claims adjudication by Medavie Blue Cross. The adjudication process filters out incorrect HCNs and drug identification numbers (DINs) as well as double entries. Medavie Blue Cross transfers claims data to the NSPP monthly.

Canadian Institute for Health Information Discharge Abstracts Database (CIHI-DAD)

The Nova Scotia DoH houses CIHI-DAD records from April 1st 1996 forward. The CIHI-DAD contains detailed information about all hospital admissions to NS hospitals. Data for the CIHI-DAD are abstracted by trained health record coders using a standardized data abstraction process. Data contained in the CIHI-DAD include the following.

- Administrative information (e.g., admission/separation times, record type, hospital number, etc)
- Patient demographics (e.g., HCN, date of birth, age, sex, etc)
- Patient diagnosis (e.g., ICD-10-CA diagnostic code, diagnostic prefix, diagnosis type, etc)
- Patient procedures (e.g., CCI code, date, doctor, doctor service, location, etc)
- Patient services (e.g., main patient service, subservice, etc)
- Physician information (e.g., most responsible physician, physician specialty, etc)

The Canadian Institute for Health Information uses support personnel, education programs, abstracting software, and data edits to improve the accuracy of their data.[4] Results from a national study examining the quality of DAD data indicated that the level of disagreement for the most responsible diagnosis was 12.8% while the disagreement for co-morbid condition diagnosis was 23.2%.[4] Although these values are quite high, CIHI noted that a substantial proportion of the disagreement stemmed from only a few institutions with highly discrepant data.[4] The CIHI transfers year-to-date DAD data to the NS DoH on a monthly basis.

Medical Services Insurance (MSI) Claims

Medavie Blue Cross maintains the MSI Claims database (i.e., physician billings) on a server accessible by the NS DoH. The MSI Claims database contains claims data for health services rendered by a physician and reimbursed through Nova Scotia's MSI Program from January 1st, 1996 forward. Data for the MSI Claims are abstracted by physicians' clerical staff. Medavie updates the MSI Claims biweekly. Data contained in the MSI Claims database include the following.

- Administrative information (e.g., date of service, treatment location, MSI program, etc)
- Patient demographics (e.g., HCN, date of birth, age, sex, etc)
- Patient diagnosis (e.g., ICD-9-CM diagnostic code, ICD-9-CM injury code, etc)
- Patient procedures (e.g., quantity of treatment, CPP procedure category, etc)
- Physician information (e.g., physician ID, type, main specialty, etc)
- Costing (e.g., unit value of service, amount approved/paid, etc)

Data contained in the MSI Claims were originally collected for claims adjudication by Medavie Blue Cross. This adjudication process filters out incorrect HCNs and double entries. Medavie also conducts random audits and chart reviews to ensure claims data are accurate.

In Nova Scotia, the majority of physicians' services are funded through a fee-for-service structure and thus are captured in the MSI Claims database. However, there are some exceptions. Specifically, 30-45 general practitioners working in rural settings were either salaried or received an income guarantee; rural physicians received lump-sum payments for emergency room and on-call services; and 256 physicians received block funding.[5] Physicians who are remunerated through alternative payment structures are encouraged to submit shadow billing claims; however, there are no regulations in place requiring them to do so. Currently, there is no information available regarding the accuracy of shadow billing claims.

MSI Insured Patient Registry (MSI-IPR)

Medavie Blue Cross maintains the MSI-IPR, a longitudinal database with information about all registered beneficiaries (past and present) of the Nova Scotia MSI Program, on behalf of the NS DoH. Data contained in the MSI-IPR database include the following:

- Demographics information (e.g., HCN, date of birth, sex, most recent postal code, etc)
- Program eligibility (e.g., eligibility status, eligibility start/end date, program, etc)

The MSI-IPR does not capture records for Nova Scotians who have their health care costs covered through other programs such as the Canadian Armed Forces and the RCMP. Medavie updates the MSI-IPR daily with incoming patient information. Data fields pertaining to an individual's address are updated yearly for NS residents who are eligible for programs with a yearly renewal process (e.g., Seniors' Pharmacare, Community Services Pharmacare) and once every four years for all other MSI Program enrolees.

Case Definitions

One major task of the NSDR Technical Working Group was to develop program-specific case definitions to identify DM cases (see Table 1). Three of the partners – DCPNS, RCP, and NSPP – contributed gold standard DM cases from their respective data holdings. The DCPNS also contributed administrative cases from the NDSS. The fourth partner – NS DoH – intended to provide additional administrative cases using a new definition based on Bayesian methodology; however, the validation work for this new case definition showed that it needs further refinement (for more details, see Chapter 6 – pp. 21-23).

Table 1: Gold standard and administrative case definitions for the NSDR

Gold Standard Cases	
DCPNS **1992 forward**	Any NS resident with a valid HCN, who is eligible to receive health care services under the MSI program and who has made ≥1 visits to a NS DC – able to distinguish PreDM, type 1 DM, type 2 DM, and GDM
RCP **1980 forward**	Any woman with a White's classification, ICD-10-CA code, or NSAPD code for DM during a pregnancy-related admission – able to distinguish GDM and pre-existing type 1 and type 2 DM
NSPP **1980 forward**	Any senior with one or more claims for insulin, OAA, or test strips[‡] – includes PreDM, type 1 DM, and type 2 DM but cannot distinguish between types
Administrative Cases	
NDSS **1995 forward**	Any individual with one hospitalization or two medical claims (i.e., physician visits) within 730 days that have DM code – includes type 1 and type 2 DM but cannot distinguish between types
DoH	NS Admin Rule: Administrative case definition needs further refinement – see Chapter 6 (pp. 21-23)

[‡] Two definitions were assessed – one with test strips and one without; the definition without test strips is recommended for identifying gold standard DM cases in NSPP-S drug claims database (see Chapter 6, pp)

File Structure

The initial vision of the NSDR was to maintain it as a minimum dataset – using it as a mechanism to construct the project-based datasets required to respond to various research or policy questions. There was vigorous debate among the NSDR Advisory Committee and Technical Working Group members about changing the structure to include some basic demographic variables. Ultimately, it was decided to go with the original intent – agreeing to assess the need to include additional variables at the end of the pilot project.

Although maintaining the provisional NSDR as a minimum dataset will require some or all of the partners to transfer additional data to a project-based dataset each time a new request is made, the partners agreed that it was an acceptable option. One reason for this decision was that it is impossible to anticipate all the

data fields that would be required to answer policy or research questions. As such, including basic demographic variables in the provisional NSDR would by no means preclude the need to go back to the partners for additional data fields. Moreover, by maintaining the provisional NSDR as a minimum dataset, requests are processed using the most up-to-date data available, reflecting the most current case ascertainment and data quality checks.

Each contributing partner used the program-specific case definition to flag DM cases, and then exported the HCNs and the corresponding date when the DM case definition was first met (referred to as DM Date hereafter) for identified cases to the provisional NSDR. The DM Date field was used to create yearly files spanning 1996/97[§] to 2005/06 with one field for HCN and one field for each data source that indicates DM status associated with a given HCN (see Table 2 for a sample file). Files for multiple years can be merged to facilitate longitudinal analyses.

Table2: Sample file for provisional NSDR

HCN	Diabetes Status (yes/no)				
	DCPNS Registry	RCP's NSAPD	NSPP-S Drug Claims	NDSS Case Definition	NS Admin Rule Case Definition
#### ### ##1	Y	Y	N	Y	Pending
#### ### ##2	Y	N	N	N	Pending
#### ### ##3	N	Y	N	Y	Pending
#### ### ##4	N	N	Y	Y	Pending
#### ### ##5	Y	N	Y	N	Pending

[§] The 1996/97 file is comprised of cases derived from all available data, as far back as 1980 for some partners.

Data Handling Protocol

One of the requirements of the PIA was to provide a detailed description of how data are transferred from the contributing partners to the provisional NSDR, from the provisional NSDR to the contributing partners, and from the contributing partners to project-based datasets. This task fell to the Technical Working Group. The exact procedure for transferring data differed depending on the physical location of the contributing partner. The following is a brief description of this process with numeric references to Figure 3.

Populating the Provisional NSDR

The provisional NSDR resides on a secure server within the NS DoH. For contributing partners with data also residing on the DoH server (NDSS – 3a, NSPP – 4a, & DoH – 5a), a simple file transfer was used to move the required data file containing HCN and DM Date from the partner to the provisional NSDR. For data files derived from RCP's NSAPD (1a) and the DCPNS Registry (2a), the data file containing HCN and DM Date was transferred to the provisional NSDR using the NS government's secure file transfer system (MOVEit® DMZ). This system meets Canadian encryption standards for accepting, storing, and transferring electronic files (i.e., conforms to Personal Information Protection and Electronic Documents Act).

Populating Provisional NSDR Project-based datasets

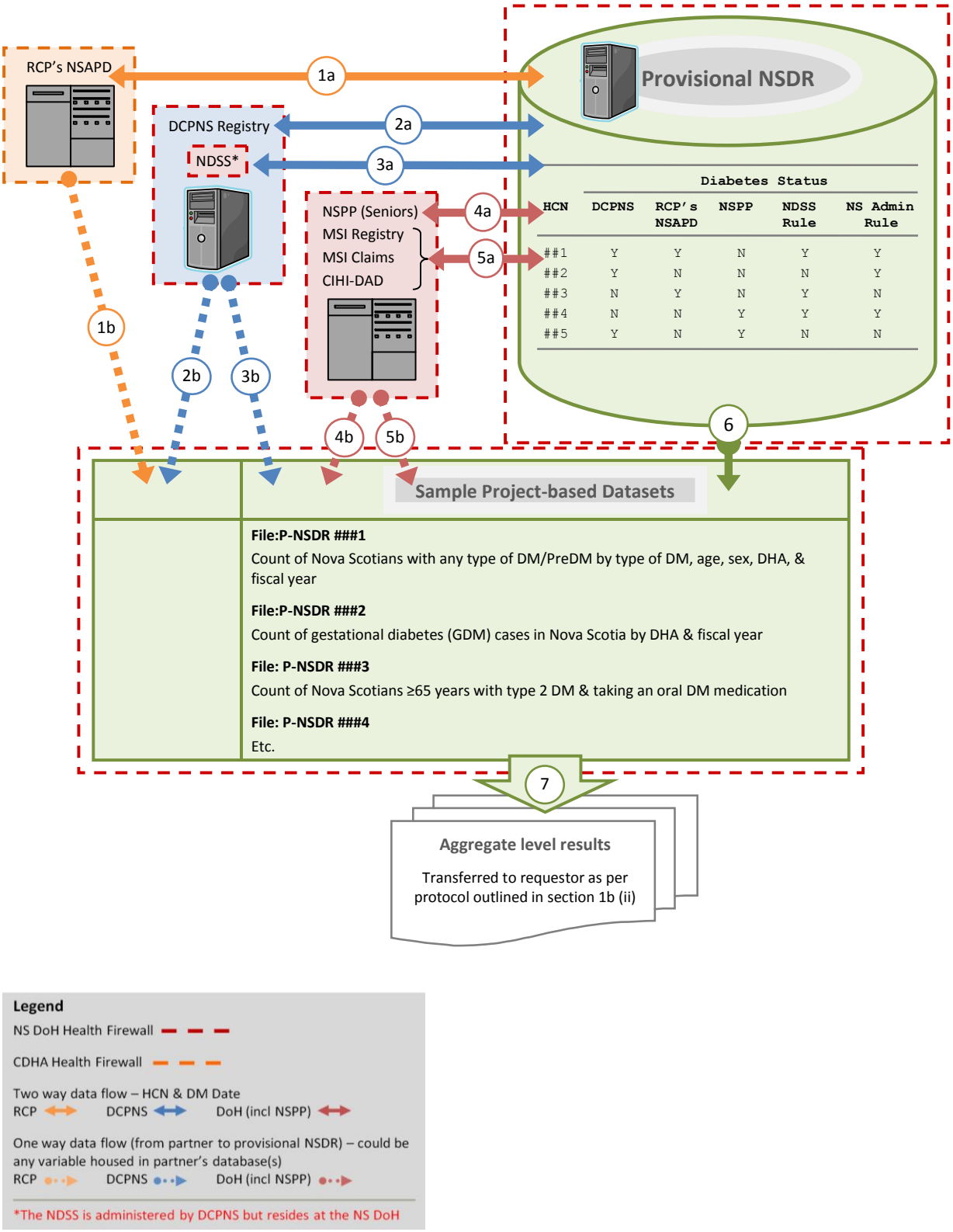
The provisional NSDR Technical Support Team created project-based datasets using the provisional NSDR data flow mechanism in order to respond to the test requests posed in the PIA. This process involved two steps. First, the provisional NSDR Technical Support Team sent a data file containing HCNs from the provisional NSDR to each contributing partner either by transferring the file internally (3a, 4a, & 5a) or by using the NS government's secure file transfer system (1a, 2a). A request for the data fields required for responding to the policy or research question accompanied this file.

The contributing partner added the required data fields to the file of HCNs and returned it to the provisional NSDR Technical Support Team either by transferring the file internally (3b, 4b, & 5b) or by using the NS government's secure file transfer system (1b, 2b). The provisional NSDR Technical Support Team combined these files to form the project-based datasets used to respond to the mock requests. Once assembled, data in the project-based datasets never flow back to the contributing partners. Furthermore, results are presented in aggregate form to protect individual identity and confidentiality.

Data Quality

The data from each partner were checked against the MSI Insured Patient Registry to ensure that the HCNs were valid; all cases with valid HCNs were combined to create the provisional NSDR. Additional data quality checks were run yielding a small number of errors (e.g., pregnant males). These errors were corrected at the source and in the provisional NSDR files. It is noteworthy that all partners benefited from the identification and amendment of erroneous data.

Figure 3: Provisional NSDR conceptual technical architecture



Chapter 5 The Provisional NSDR

Description

Two key objectives of the Public Health Agency of Canada Enhanced Chronic Disease Surveillance Grant were to combine new and unique data sources for the purpose of chronic disease surveillance and to develop rigorous methodological and organizational tools to combine these data into a sustainable surveillance resource. The NSDR Pilot Project met both of these objectives through the development of the provisional NSDR.

Currently, the provisional NSDR is comprised of DM cases from four sources: DCPNS Registry, RCP's NSAPD, NSPP Seniors' Program, and the NDSS. Each partner extracted their prevalent cases using all data at their disposal; as such, the years of coverage vary between partners: RCP and NSPP – 1980 forward, DCPNS – 1992 forward, and NDSS – 1995/96 forward. To date, the provisional NSDR is the most inclusive source of DM data in NS with records for 107,643 unique cases of DM (all types) which includes cases that have since moved out of province or died. This resource will be of benefit to a large variety of stakeholders including the provincial programs, DoH, and DHAs to name a few.

Any project-based datasets assembled in response to policy and research questions will be comprised of the most up-to-date data available from each contributing partner, reflecting the most current case ascertainment and data quality checks.

Results for Test Requests

The NCDR Interim Data Access and Governance Committee developed a series of test questions (see p. 10) with varying degrees of complexity to assess the feasibility and utility of the provisional NSDR data transfer mechanism, procedures, and policies to respond to various types of information requests – focussing on the time and human resources required to fulfill the request. The results of these questions are presented in Box 1.

Time and Human Resources

Overall, the construction of the provisional NSDR and the subsequent project-based datasets was prompt – each being completed within a one month period. One partner noted that the whole process only required an afternoon of programming with some iterative processes over the succeeding two weeks to fix a couple errors. It was noted, however, that a system would need to be devised to prioritize data requests based on the degree of urgency – especially for requests originating within the Department of Health.

Arguably, the data transfer process was facilitated by the fact that each partner knew in advance which data elements would be required for provisional NSDR and the test questions. A more objective "real world" test of the provisional NSDR mechanism will be completed in 2009/10 as part of another DCPNS PHAC funded project titled *Quantifying the Burden of Diabetes: Time to Comorbidity & Time to Death*. Information gathered through this test will help shape a policy about response times to NSDR data requests.

Box 1: Preliminary Data Analysis

As can be seen in Figure 1.1, the number of prevalent DM (all types) and PreDM cases grew by several thousand per year across the last five years (Q3, p. 10). Of the 81,447 DM (all types) and PreDM cases in 2005/06, approximately 80% (n=64,831) are gold standard cases (i.e., from DCPNS, RCP, or NSPP-S) and 20% (n=16,616) are estimated cases based on the NDSS case definition (Q1, p. 10).

Note: A breakdown of cases by DM type, age, sex, and DHA of residence cannot be reported due to small cell counts.

Figure 1.1 Prevalent diabetes (all types) and prediabetes cases by year

One third (n=27,036) of the DM (any type)/PreDM cases in the provisional NSDR were contributed by a maximum of one source (see Table 1.1), highlighting the importance of this collaborative approach to DM surveillance (Q2, p. 10). About half (n=39,329) of the cases are found in the datasets of two contributing partners, and about 18.5% (n=15,082) are present in the datasets of three contributing partners. No single case is present in all four data sources.

As expected, only a small number of cases (n=2) were present in RCP's NSAPD and in the NSPP-S drug claims database. This situation is not impossible as cases from the NSAPD go back to 1980; for example, a 40-year-old woman who gave birth in 1980 would have been eligible to enrol in the NSPP Seniors' Program by 2005.

Table 1.1 DM case contribution by NSDR partners for fiscal year 2005/06

Cases found in one source n=27,036 (33.2%)		Cases found in two sources n=39,329 (48.3%)		Cases found in three sources n=15,082 (18.5%)	
DCPNS	5,914	DCPNS/RCP	1,888	DCPNS/RCP/NDSS	1,175
RCP	2,171	DCPNS/NSPP	569	DCPNS/NSPP/NDSS	13,905
NSPP	2,335	DCPNS/NDSS	28,533	RCP/NSPP/NDSS	2
NDSS	16,616	RCP/NDSS	180		
		NSPP/NDSS	8,159		

In fiscal year 2005/06, there were 294 cases of GDM among NS women. Of these cases, 208 were among women under 35 years of age, and 86 were among women 35 years of age and older (Q4, p. 10).

In NS, seniors (65 years or older) are eligible to receive benefits under the Seniors' Pharmacare Program if they do not receive benefits through a private insurer or other program. In fiscal year 2005/06, a total of 10,786 NS seniors with Type 2 DM were taking OAA (Q6, p. 10).

Of the 81,447 prevalent DM (any type)/PreDM cases in 2005/06, 6,661 had a myocardial infarction at some point in the last 10 years. (Q7, p. 10)

Chapter 6 Validation of the NSDR

Introduction

The validation component of the NSDR Pilot Project was completed by George Kephart (NSDR partner and contract epidemiologist) and his colleague Pantelis Andreou, both from the Department of Community Health And Epidemiology at Dalhousie University. The final report for this work was completed in July 2009. A brief description of this work and the resulting recommendations follow. The complete report is appended (Appendix A, pp. 29-48).

Background

The provisional NSDR combines gold standard DM cases from the DCPNS Registry, RCP's NSAPD, and NSPP-S drug claims database; however, these gold standard sources do not capture all DM cases in the province. In NS, administrative health records (i.e. physician billing claims and hospital discharge abstract) can be linked to program databases, providing an alternative option for identifying additional DM cases. However, a suitably accurate administrative case definition has yet to be developed. As such, validating administrative case definitions to identify DM cases which are not captured by the gold standard sources was the overarching goal of the validation component of the NSDR project.

Unlike the gold standard data sources, case definitions based on administrative data are subject to false-positive as well as false-negative errors. While there has been a considerable amount of validation work pertaining to administrative case definitions for DM, it has focussed primarily on the general population and may not be applicable to a subset of the Nova Scotia population which is not captured by gold standard sources. The accuracy of case definitions derived from administrative data is highly dependent on the prevalence of DM in the population. When a large percentage of DM cases are captured by the gold standard sources, the prevalence of DM in the remaining population declines as does the validity of administrative cases definitions. Accordingly, more stringent case definitions are required.

Objectives

1. To establish case definitions to be used from each of the gold standard data sources for inclusion as gold standard cases.
2. To estimate the accuracy of case ascertainment from each of the data sources providing evidence of diabetes.
3. To evaluate which case definitions based on administrative data should be used to identify additional cases for inclusion in the NSDR.

Methods

A person-year data file for all NS residents with a valid HCN was created for the 10-year period between 1996/97 and 2005/06. A person-year file combines person and time at risk in the denominator so that a person with a valid HCN for the 10-year study period will contribute 10 person-years of data while a

person who moved away after 3.5 years will contribute 3.5 person-years of data. In each year, it was noted whether individuals appeared in one or more of the gold standard data sources (i.e., the gold standard DM cases) or had a DM code in any administrative health record. This information was used to generate cross-tabulations for statistical analyses.

Estimates of prevalence, sensitivity, and specificity were calculated for each data source using Bayesian methodology. These estimates were then used to compute prevalence of DM among the non-gold standard DM population (i.e., people left after the gold standard cases are removed) and then to estimate the positive predictive value for various administrative case definitions. The positive predictive value refers to the percentage individuals identified as having DM by a case definition that truly have DM.

A more detailed methodology is presented in the full validation report (see Appendix A, pp. 29-48)

Results

Diabetes supplies (i.e., test strips and lancets) have been covered in the NSPP-S since the 2004/05 fiscal year. Between 2004-05 and 2006-07, just over 5,000 new DM cases were identified based on a claim for DM supplies alone. A large share of these cases only filled claims for supplies sporadically.

Relatively precise estimates of DM prevalence were generated using Bayesian estimation methods. With the exception of 55-74 year old males, these estimates behaved as one would expect – prevalence was higher for males than for females, and it increased with age.

Overall, the Manitoba Rule has the best positive predictive values. While it may miss some cases, it will yield few false-positive cases. In contrast, the NDSS case definition performs very poorly. Depending on the age and sex group in question, 20-80% of DM cases identified by the NDSS would be false-positives.

More detailed results are presented in the full validation report (see Appendix A, pp. 29-48)

Recommendations

The following recommendations, listed in order of priority, ensued from the validation work.

1. When using the NSPP-S data to define gold standard cases of DM to be incorporated into the NSDR, users of diabetes supplies only (e.g., test strips and lancets) should not be included as cases. Cases should be defined as users of insulin and/or oral medications.
2. The NDSS case definition should not be used to identify additional DM cases for the NSDR as the number of false positive cases is unacceptably high.
3. The Manitoba Rule (≥ 1hospitalization &/or ≥2 physician visits in 2 years for which DM is coded) should be used to identify additional cases of DM which are not captured by the gold standard sources.

   Using the NDSS case definition is likely to result in an unacceptably large number of false-positive cases of DM. The risk of this error is low for the Manitoba Rule because it is applied cross-sectionally rather than longitudinally.
4. Preferably, the prevalence of DM should not be estimated directly from cases in the NSDR. Instead, the Bayesian estimation approach we have used in the validation analysis should be further refined

and used as standard practice for estimating and reporting prevalence. In the interim, cases identified using the Manitoba Rule can be combined with gold standard NSDR cases to determine a reasonably accurate estimate of DM prevalence.

The Bayesian methods we employed estimate the prevalence of DM while simultaneously adjusting for deviations from perfect sensitivity and specificity of the data sources. We found that the prevalence estimates were very robust to alternative model specifications. However, the estimates of prevalence for males age 65-74 were suspect, being lower than expected. Further refinements of the estimates should be undertaken before the method is used on a routine basis to estimate DM prevalence for reporting purposes.

Using the NSDR to directly estimate prevalence assumes that there are no ascertainment errors and that ascertainment errors do not vary by age and sex. These assumptions are not substantiated by our analyses. However, we believe that using gold standard cases in the RCP, DCPNS, and NSPP-S data, supplemented by additional cases based on the Manitoba Rule, can be used as an interim approach for estimating DM prevalence.

5. Additional work should be conducted to further refine the Bayesian estimates of prevalence, sensitivity and specificity of the data sources in the NSDR.

   The Bayesian methods we employed in this study make a number of simplifying assumptions to facilitate estimation. These may result in some bias in our estimates. We believe that further analyses, which relax model assumptions while using more data, show considerable promise for refining estimates and improving surveillance capacity.

6. Additional work should be conducted to refine the Kephart Rule (an administrative case definition based on Bayesian methodology)[6] for use in the NSDR.

   There is an important conceptual distinction between the NDSS case definition and the Kephart Rule on one hand, and the remaining rules on the other. The Kephart Rule (and to some degree the NDSS case definition) are designed to identify cases than can be followed over time. Once a person is identified as a DM case, they remain in that state in subsequent years. This situation is not necessarily true for case definitions like the Manitoba Rule that use different claims data in each year to infer DM status. Accordingly, a person's DM status may be classified differently from year to year.

   The Manitoba Rule will work well when NSDR applications only need to identify DM status of person at specific points in time, which likely pertains to the majority of expected applications. However, for follow-up studies of the patients with DM, this application will be problematic. For such studies, the Kephart Rule should be adapted for use in the NSDR. This endeavour is feasible but will require establishing new probability cut-offs for identifying DM cases with the Kephart Rule.

Chapter 7 Lessons Learned

Partnership is Key

Overall, the Nova Scotia Diabetes Database Pilot Project was a very positive experience. The successful development of the provisional NSDR is a testament to the importance of collaboration with a broad base of highly engaged partners. Most visible, is the importance of the contributing partners, without whom there would be no provisional NSDR. However, no less important are the numerous unseen partners who worked behind the scenes to ensure that the provisional NSDR has a permanent home and that data transfers are, and continue to be, secure.

The Need for Continuity

Although the provisional NSDR is truly a collaborative project, it does require a dedicated individual/partner to oversee its maintenance. If multiple individuals/partners are involved in the maintenance, it is difficult to keep track of what has and has not been done. A dedicated guardian is better able to monitor ongoing activities such as regular updates and scheduled destruction of files. Of course, the individual/partner in charge of these duties can rotate on a regular basis (e.g., yearly) as long as the outgoing guardian maintains sufficient logs for the incoming guardian.

There Is Always More to Learn

Responding to the seven test questions used to assess the feasibility of the provisional NSDR generated some learnings beyond the actual answers to the questions. All partners benefited from the identification and subsequent correction of erroneous data in their datasets. Other anomalies were also noted and are awaiting further investigation, for example; the provisional NSDR data contained 4,523 cases of DM for which there were no valid HCNs.

Plan for Delays

As is often the case, certain activities of the NSDR Pilot Project took longer to complete than first anticipated. For example, the validation work took nearly two years to complete and required five separate submissions to various ethics boards and data access committees to assemble de-identified versions of the data used to construct the provisional NSDR. The redundancies in this process highlight the need to streamline data access for the provisional NSDR. Similarly, the PIA process required numerous revisions across a five-month period. Conducting this type of project is not feasible within a 12-month grant period.

Chapter 8 Next steps

A Real World Test

A number of plans for the provisional NSDR are currently under way. First, the DCPNS has leveraged the provisional NSDR to secure federal funding for a project examining the factors associated with time to comorbidity and time to death for a cohort of clinically confirmed (i.e., gold standard) DM cases. This project will provide a real world test of the entire NSDR data transfer process – from the Application for Access to Provisional NSDR Data to the provision of aggregate level results. The outcome of this test will help shape policies and procedures for an ongoing NSDR.

In addition to the real world test, the seven test questions (see p. 10) will be refined and the analyses rerun so that more meaningful results are produced. At this time, additional policy-relevant questions will be added as deemed necessary.

From Provisional to Ongoing NSDR

The construction of the provisional NSDR has demonstrated that programs can successfully share data while protecting privacy. In essence, an ongoing NSDR is a feasible endeavour. Now that the DCPNS and its partners have laid the groundwork, the NS DoH needs to determine if this type of repository fits within the department's chronic disease strategy. Clearly, the NSDR is a valuable surveillance tool as many partners have bought in to and continue to support the endeavour. Now, key stakeholders need to address outstanding issues about the fate of the NSDR before the momentum is lost:

- How will the NSDR be updated?
    - How often?
- Who will maintain the NSDR?
    - Do we need an administrator to oversee its day-to-day management?
    - How will be ensure continuity if a key partner ceases to be involved
- Are there new partners that should be considered?
    - If so, who?
    - How should they be added?
    - Which data fields are of interest
- How should we move forward?
    - Should we consider a Provincial Program Repository/Chronic Disease Repository?

If the provisional NSDR is to become a sustainable ongoing DM surveillance tool, a number of refinements must be made to the provisional NSDR protocol. The Data Access Guidelines need to be formalized in a cohesive document that addresses data requests from departments and programs within the DoH and requests from departments, organizations, and researchers outside the DoH (e.g., other government departments, university researchers, students, etc). This document should include a plan for

prioritizing data requests as there will be the need to respond expeditiously to urgent requests from within the Ministry

Minor revisions to the PIA also will be necessary if the NSDR becomes permanent. These refinements include, but are not limited to, sections addressing data access and governance, the addition of new partners, and a schedule of contributors. A schedule of recurring questions might also be considered for inclusion in the PIA.

Finally, a manual of NSDR Policies and Procedures needs to be developed. The importance of this type of document cannot be overstated, especially given the shared ownership status of the NSDR. This manual would include all the details about the day-to-day management of the NSDR, guidelines for dealing with recurring questions, and details about the nature and nuances of the data contained in the NSDR.

Future Research Possibilities

Additional funding will be sought to explore the cases contributed solely through the NDSS using a chart audit approach. It is important to understand whether or not these administrative cases are true positive cases of DM. If they are, DCPNS and others need to understand why these individuals are not accessing the province's Diabetes Centres for education and support.

Chapter 9 Dissemination

Provincial Audiences

Dissemination activities have been ongoing since the start of the NSDR Pilot Project. Regular progress updates have been (and continue to be) published in the quarterly DCPNS Newsletter – this newsletter is circulated broadly within the province and to key stakeholders outside the province. In fact, an interested party in Ontario recently contacted the DCPNS to inquire about the provisional NSDR data sharing mechanism after reading about the project in the DCPNS Newsletter.

The DCPNS Project Manager delivered a brief presentation about the NSDR Pilot Project at a meeting of the DCPNS Advisory Council. The DCPNS Coordinator and Director of Diabetes Surveillance also delivered a presentation about the project at a meeting of the NB Diabetes Task Force. Additional presentations are being planned with the NS DoH. We also welcome the opportunity to present our work to those interested in the NDSS and enhanced diabetes surveillance.

National and International Audiences

In February 2009, the DCPNS Project Manager assigned to the NSDR Pilot Project delivered a presentation about the construction of the provisional NSDR at a national workshop sponsored by PHAC: *Working with Provincial/Territorial Administrative Data for Surveillance and Research: Methodological Challenges*. Also an abstract has been accepted for presentation at the International Diabetes Federation's 20th World Diabetes Congress in October 2009.

References

1. Diabetes Care Program of Nova Scotia. Internal validation work comparing DCPNS Registry to NDSS (Software v208). Diabetes Care Program on Nova Scotia, Halifax, Canada, 2008.
2. Health Canada. Responding to the challenge of diabetes in Canada: first report of the National Diabetes Surveillance System (NDSS) 2003.Health Canada Report No. H39-4/21-2003E, Ottawa, 2003.
3. Diabetes Care Program of Nova Scotia. Nova Scotia NDSS submission files v207, December 2007.
4. Canadian Institute for Health Information. Data quality documentation: discharge abstract database 2001-2002. Ottawa, Canada: Canadian Institute for Health Information 2003.
5. Canadian Institute for Health Information. Alternative Payments and the National Physician Database: the status of alternative payment programs for physicians in Canada 2001-2002 and preliminary information for 2002-2003. Ottawa, Canada: CIHI 2004.
6. Kephart G, Casey J, Ranger R, Dunbar P, Karlovic Z. The development and validation of an alternative case definition for the National Diabetes Surveillance System. Final Report to the Data Validation Working Group, National Diabetes Surveillance System. Ottawa, Health Canada, 2004.

The Completeness and Accuracy of Data Sources Used in the Provisional NSDR: Results from the Validation Study

July 30, 2009

George Kephart, PhD
Pantelis Andreou, PhD

Department of Community Health and Epidemiology
Faculty of Medicine
Dalhousie University

Executive Summary

The plan for the provisional NSDR was to first combine cases of diabetes (DM) identified through the Diabetes Care Program of Nova Scotia (DCPNS) Registry, the Reproductive Care Program of Nova Scotia's (RCP) Nova Scotia Atlee Perinatal Database, and the drug claims database for the Nova Scotia Pharmacare Program for Seniors (NSPP-S). These sources are considered to be gold standard sources because cases are clinically confirmed and accurately coded. In other words, it was assumed that these sources would not yield any false positive cases. However, these gold standard sources do not capture all cases. The NSPP-S only capture cases for person's age 65 and over who are enrolled in the program (approximately 75%) and fill prescriptions for diabetes treatment. The DCPNS only captures patients who are referred to the province's 39 Diabetes Centres. Together, it was believed that these sources would capture approximately 50-70% of cases. Administrative data (i.e. physician claims and hospital discharge abstract) are thus needed to identify additional cases of DM, but to do so, acceptably accurate case definitions must be established.

Validating administrative case definitions to identify DM cases which are not captured by the gold standard sources was the overarching goal of the validation component of the NSDR project. We linked data from all the sources used in the NSDR and used a variety of statistical models to estimate the prevalence of **diagnosed diabetes** (DDM) along with estimates of the accuracy and completeness of each data source.

Based on our analyses, we make the following recommendations for the NSDR, listed in order of priority.

1. **When using the Nova Scotia Senior's Pharmacare data to define gold standard cases of DDM to be incorporated into the NSDR, users of diabetic supplies only (e.g., test strips and lancets) should not be included as cases. Cases should be defined as users of insulin and/or oral medications.**
2. **National Diabetes Surveillance System (NDSS) case definition should <u>not</u> be used to identify additional DDM cases for the NSDR as the number of false positive cases is unacceptably high.**
3. **The Manitoba Rule (≥ 1hospitalization &/or ≥2 physician visits in 2 years for which DM is coded) should be used to identify additional cases of DDM which are not captured by the gold standard sources.**
4. **Preferably, the prevalence of diagnosed diabetes should not be estimated directly from cases in the NSDR. Instead, the Bayesian estimation approach we have used in the validation analysis should be further refined and used as standard practice for estimating and reporting prevalence. In the interim, cases identified using the Manitoba Rule can be combined with gold standard NSDR cases to determine a reasonably accurate estimate of DM prevalence.**
5. **Additional work should be conducted to further refine the Bayesian estimates of prevalence, sensitivity, and specificity of the data sources in the NSDR.**
6. **Additional work should be conducted to refine the Kephart Rule (an administrative case definition based on Bayesian methodology) for use in the NSDR (Kephart, Casey, Ranger, Dunbar, & Karlovic, 2004).**

Introduction

The plan for the provisional NSDR was to first combine cases of diabetes (DM) identified through the Diabetes Care Program of Nova Scotia (DCPNS) Registry, the Reproductive Care Program of Nova Scotia's (RCP) Nova Scotia Atlee Perinatal Database, and the drug claims database for the Nova Scotia Pharmacare Program for Seniors (NSPP-S). These sources are considered to be gold standard sources because cases are clinically confirmed and accurately coded. In other words, it was assumed that these sources would not yield any false positive cases. However, these gold standard sources do not capture all cases. The NSPP-S only capture cases for person's age 65 and over who are enrolled in the program (approximately 75%) and fill prescriptions for diabetes treatment. The DCPNS only captures patients who are referred to the province's 39 Diabetes Centres. Together, it was believed that these sources would capture approximately 50-70% of cases. Administrative data (i.e. physician claims and hospital discharge abstract) are thus needed to identify additional cases of DM, but to do so, acceptably accurate case definitions must be established.

Validating administrative case definitions to identify DM cases which are not captured by the gold standard sources was the overarching goal of the validation component of the NSDR project. Unlike the gold standard data sources, case definitions based on administrative data are subject to false-positive as well as false-negative errors. While there has been a considerable amount of previous validation work on administrative case definitions of **diagnosed diabetes mellitus** (DDM), they have been developed for the general population and may not be applicable to the subset of the Nova Scotia population which is not captured by gold standard sources. For example, the case definitions currently employed by the National Diabetes Surveillance System (NDSS) may not be appropriate for the NSDR. The reason is that the accuracy of case definitions derived from administrative data is highly dependent on the prevalence of DM in the population to which they are applied. The higher the capture rate of the gold standard sources, the lower will be the prevalence of DM in the remaining population. For any given case definition, the percent of all cases which are false-positives will increase as the prevalence decreases. Accordingly, as prevalence decreases, more stringent case definitions are required.

Objectives

The goals of the validation analysis were as follows:

(1) To establish case definitions to be used from each of the gold standard data sources for inclusion as gold standard cases.
(2) To estimate the accuracy of case ascertainment from each of the data sources providing evidence of diabetes.
(3) To evaluate which case definitions based on administrative data should be used to identify additional cases for inclusion in the NSDR.

Methods

The validation component of the study included the following data and analytic steps:

1. Determine the criteria for selecting cases from each gold standard source.

In consultation with each of the gold standard data holders, we defined the criteria for identifying cases of diagnosed DM from each of the gold standard data sources. The criteria for each gold standard source are shown in Appendix A.

It was believed at the outset of the project that claims for DM supplies in the NSPP-S database could be used to identify gold standard cases, but analyses were required to determine if this assumption was valid. Accordingly, we examined the pattern of claims, within persons, for DM supplies; we examined if persons who filled claims for supplies were also likely to be filling claims for DM medications and whether persons who filled claims for supplies were likely to be captured as DDM by other gold standard sources. Based on these analyses, it was determined that persons with NSPP-S claims for DM supplies only should not be treated as gold standard DM cases. Summary results from these analyses are provided in the Results section (see p.35)

2. Construct a master validation database.

Based on the criteria determined for step 1, ten years of data (up to and including the 2006-07 fiscal year) from gold standard and administrative data sources were linked to MSI person registry data, which includes records for each person in the population and the period of time for which they were covered by MSI. The resulting file provided detailed information on the Nova Scotia population over a ten-year period, consisting of a record for each year in which each person was registered with MSI. Each record included variables indicating whether a person, in that year, was classified as being diagnosed with DM according to each gold standard data source. In addition, each record included variables indicating the number of physician billing claims and the number of hospital discharge abstracts with a diagnosis of DM.

Given the sensitive nature of the data, linkage and analyses were subject to high security measures to protect confidentiality. Linkage and analyses were conducted through the Population Health Research Unit (PHRU) at Dalhousie University. PHRU maintains denominized versions of the MSI patient registry, Seniors' Pharmacare drug claims data, physician claims data and hospital discharge abstract data. Records are identified only by encrypted health card numbers (eHCNs) as the unique personal identifiers. All extracted data for the project from RCP and DCPNS was denominized (i.e., had name, address, and phone number removed) and HCNs were encrypted. The encryption of HCNs was carried out by a responsible third party, Medavie Blue Cross. The encryption algorithm is unknown to the programs, the Population Health Research Unit (PHRU), and the researchers. The linkage was conducted based on year and encrypted health card number (eHCN).

The project was also subject to extensive oversight. The linkage and analysis was approved by the PHRU Data Access Committee, as well as by data access committees for the Reproductive Care Program and the Diabetes Care Program of Nova Scotia. The project was also approved by the Dalhousie University Research Ethics Board. Further oversight was provided by the project Steering Committee and Technical Advisory Committee.

3. Estimate the sensitivity and specificity of each data source, and the population prevalence of diagnosed DM.

For each of the data sources, we estimated the sensitivity and specificity of each data source. Sensitivity is the percent of all cases of DDM in the population identified by a given source. In other words, it is the percent of all DDM cases captured by a data source. Because the sensitivity of the RCP data was very small (less than 1% of all gold standard cases), we did not attempt to estimate sensitivity for this data source. However, the sensitivity and specificity of the DCPNS and the NSPP-S data are important to determine.

Specificity is the percent of all non-cases that are identified correctly by a given data source. Because the DCPNS data and the NSPP-S data (based on use of diabetes drugs only) were assumed to have no false-positive cases, their specificity was assumed to be equal to 100% . Thus, we only estimated specificity for the administrative data.

In all of the previous validation work on the coding of diabetes in Nova Scotia administrative data, estimates of sensitivity and specificity of case definitions were obtained by comparing administrative data to "gold standard" cases and non-cases of diabetes. Gold standard cases were obtained from the DCPNS Registry, and gold standard non-cases were obtained from survey data. Clearly, the latter is tarnished and subject to reporting error, which is a major limitation of previous work (Kephart et al., 2004). The DCPNS cases are "gold standard" cases, as are the NSPP-S cases (if based only on DM drugs, and not supplies – see below), but these cases may not be representative of all cases in the population.

There is now considerable interest in the research literature in using newer, Bayesian methods that estimate sensitivity and specificity of administrative case definitions in the absence of a gold standard (Branscum, Gardner, & Johnson, 2005; Enoe, Georgiadis, & Johnson, 2000; Joseph, Gyorkos, & Coupal, 1995; Rutjes, Reitsma, Coomarasamy, Khan, & Bossuyt, 2007). These newer methods were used for the validation analysis we conducted as part of the NSDR pilot project. They do not require us to assume that any of the data sources have perfect sensitivity and specificity, but allow us to make such assumptions where appropriate (for example, we can assume that the specificity of the DCPNS data is perfect). Another advantage of the Bayesian methods we used is that, in addition to estimating the sensitivity and specificity of multiple data sources, they simultaneously estimate the population prevalence of DM. As noted above, knowledge of the population prevalence of DM is critical for estimating the accuracy of case definitions based on administrative data.

A more detailed description of the Bayesian methods we used, along with some examples, can be found in Appendix B. We stratified all analyses by sex and age group (<45, 45-54, 55-64, 65-74, 75-84, and >84). As noted in the appendix, we estimated a range of models which employed different assumptions.

4. Estimate the positive predictive value (PPV) for alternative administrative case definitions in the population of persons who are not in either of the gold standard data sources.

The positive predictive value (PPV) is the percent of all cases identified by a given data source that are, in fact, true cases. For example, a PPV of 80% would mean that 80% of cases identified are true cases, while the remainder (20%) are false-positives. The PPV is dependent on population prevalence. For a case definition with any given specificity which is less than 100%, and a sensitivity of any given value, the PPV will decrease as the population prevalence declines. For example, suppose one had a case

definition with 99% specificity and 100% sensitivity. If it was applied to a population with 10% disease prevalence, the PPV would be 92%, which is excellent. However, the same case definition administered to a population with .1% prevalence would yield a PPV of only 9.1%, which would be unacceptable. For the actual application of any surveillance system such as the NSDR, PPV is the most important attribute of accuracy to consider. The negative predictive value (NPV) is also of interest, but is always high for case definitions of DDM that have been examined.

This issue is directly relevant to the NSDR pilot project. In the NSDR, administrative case definitions will be used to identify additional cases of DDM which are not captured by the gold standard data sources. This situation is in contrast to how case definitions are used by the NDSS, a system designed for use in the general population for which the overall prevalence of DDM is 5-10%. If, as expected, the gold standard data sources used in the NSDR capture over 50% of all DDM cases, then the prevalence of DDM in the remaining population, to which administrative case definitions will be applied, will be less than half that of the general population. As a result, NDSS case definitions will yield lower, and quite possibly, unacceptable PPVs. We thus estimated PPVs for the NDSS case definition, as well as alternative administrative case definitions, to determine which is optimal for use in the NSDR.

To estimate PPVs, we first combined age group and sex specific estimates of the population prevalence and sensitivity of the DCPNS and NSPP-S to estimate the prevalence of DDM in the population not identified as cases by these data sources (i.e. the NGP: Non-gold-standard population). In doing so, we assumed independence in the sensitivity of the DCPNS and NSPP-S sources. It seems likely that cases captured in one of these data sources are more likely to be captured in the other, which would violate this assumption. If so, then we may underestimate the prevalence of diagnosed DM in the NGP. However, based on the data we do not expect this bias to be large.

Using the estimates of DDM prevalence in the NGP, we were then combined them with estimates of sensitivity and specificity for administrative case definitions to compute PPV estimates. However, we had to decide was estimates of sensitivity and specificity to use. The estimates of specificity of administrative case definitions obtained from our Bayesian estimates were somewhat unstable, particularly for case definitions with high values of specificity, and all the estimates of specificity were higher than we expected based on previous work (Kephart et al., 2004). We thus decided to use estimates of sensitivity and specificity of administrative case definitions that we obtained previously using traditional methods (Kephart et al., 2004). We think that the estimates of specificity that we used may be biased downwards slightly (i.e. they may be lower than the true values), as a result of the "gold standard" comparisons that were used to compute them. However, given the extensive data cleaning that was done to refine the gold standard comparisons that were used, and we believe they are the best estimates of sensitivity and specificity we have available.

Results

Can Claims for Diabetic Supplies in the NSPP-S Data be Used to Identify Gold Standard Cases?

DM supplies have only been covered in the NSPP-S since the 2004-2005 fiscal year. Between 2004-05 and 2006-07, and using the NSPP-S data alone, 44,514 apparent new (i.e. incident) cases of DDM were identified in the NSPP-S data based on claims for supplies or DM medications. Of these, most were based on a first claim for oral medications (74%) and insulin (14.6%). New cases based on a claims for DM supplies alone accounted for 11.3% of all cases (N=5031).

More detailed examination of NSPP-S supply cases calls into question the use of supplies to identify cases of DDM. A large share of the 5031 supply cases only filled claims for supplies sporadically. Moreover, only 1.6% had subsequent claims for insulin and only 26% had a subsequent claim for an oral DM medication in the same or future years. Moreover, the evidence for supply cases being diagnosed with DM in other data sources was low. Only 12.4% of the 5,031 supply cases were identified as DDM cases in the DCPNS registry. While the majority of supply cases (63.2%) had a physician claim indicating a diagnosis of DM, the 40% that did not is concerning. Only 5.8% ever had a hospital abstract with a diagnosis of diabetes. Thus, we concluded that claims for DM supplies should not be used to identify gold standard cases of DDM. Further work should explore the utility of using claims for supplies as evidence in administrative case definitions.

Estimates of the population prevalence of DDM, and the sensitivity and specificity of data sources

Table 1 shows the Bayesian estimates of DDM prevalence that were used in all subsequent steps. These estimates were based on Bayesian estimation methods described in Appendix A, and employed an administrative case definition of 1 or more claims for DDM in either physician claims or hospital data within a 1 year (the 1-in-1 rule). Using other administrative case definitions produced very similar results for prevalence as well as for the sensitivity of the DCPNS and NSPP-S case definitions.

The estimates of prevalence in Table 1 are precise, as indicated by the narrow 95% credibility intervals, plausible, and follow expected patterns by age and sex. Males have higher prevalence than females, and prevalence generally increases with age. The decrease in prevalence in the oldest age group is as expected, and is the result of higher mortality and lower life expectancy of persons with DM. However, the decline in estimated prevalence for males between the 55-64 and 65-74 age groups is puzzling, and modifications to the estimation procedure were not successful in removing this anomaly. We suspect that we have underestimated prevalence in the 65-74 age group. Further work is required to determine what factors account for this result.

Estimates of the sensitivity of the 1-in-1 administrative case definition, the DCPNS data and the NSPP-S data are shown in Tables 2-4. The estimates are precise, as indicated by the narrow 95% credibility intervals, and are largely consistent with expectations. The exception is results for males in the 65-74 age group which, like the prevalence estimate for this group, are lower than for adjacent age groups. Further work is needed to refine this estimate. The estimates of sensitivity for the 1-in-1 rule are consistent with the results of previous validation work. They decline with age, and are slightly higher for females than for males.

The sensitivity estimates for the DCPNS and NSPP-S data indicate the coverage rates for these sources, which to this point have not been available. With the exception of the odd results for males aged 65-74, the sensitivity estimates for the DCPNS data indicate that between 50-60% of all 2006 cases of DDM

under the age of 85 were seen at DCPNS clinics within a ten-year window. However, fewer older DDM patients have visits to DCPNS clinics. It is known that coverage for paediatric cases is nearly 100%, so the in the youngest age group, the missed cases would be concentrated in the older end of the age range. The NSPP-S data captures about 20% percent of DDM cases for the population over age 65. There are several potential explanations for this low capture rate. First, only about 75% of the population age 65 and over are enrolled in the Seniors' Pharmacare Program. Second, not all DM medications are covered under the Seniors' Pharmacare Program – namely long-acting insulins, TZDs, and newer agents pending approval). Finally, not all patients with DDM are on pharmacologic therapy, and thus will not be captured using this data source.

Table 1. Bayesian Estimates of DDM Prevalence, 2006

	Mean Prev (%)	95% Credible Intervals Lower	Upper
Males			
< 45	1.56%	(1.50% ,	1.61%)
45 - 54	9.15%	(8.93% ,	9.37%)
55 - 64	18.37%	(18.04% ,	18.71%)
65 - 74	15.85%	(14.34% ,	17.26%)
75 - 84	22.07%	(20.81% ,	23.43%)
> 84	20.50%	(18.04% ,	22.64%)
Females			
< 45	1.49%	(1.42% ,	1.55%)
45 - 54	7.15%	(6.95% ,	7.34%)
55 - 64	14.16%	(13.86% ,	14.55%)
65 - 74	16.81%	(16.15% ,	17.51%)
75 - 84	19.76%	(18.77% ,	20.82%)
> 84	15.85%	(14.34% ,	17.26%)

Table 2. Estimates of Sensitivity for the 1 in 1 Administrative Case Definition

	1 in 1 rule Mean	(Lower ,	Upper)
Males			
< 65	83.4%	(82.8% ,	86.0%)
65 - 74	60.3%	(57.5% ,	63.1%)
75 - 84	75.7%	(74.1% ,	77.3%)
> 84	59.0%	(55.1% ,	62.8%)
Females			
< 65	85.6%	(84.9% ,	86.2%)
65 - 74	83.1%	(82.0% ,	84.1%)
75 - 84	74.2%	(72.8% ,	75.7%)
> 84	60.3%	(57.5% ,	63.1%)

Table 3. Estimates of Sensitivity for the DCPNS Registry

	DCPNS Registry Mean	(Lower , Upper)
Males		
< 65	55.1%	(54.3% , 55.9%)
65 - 74	36.7%	(33.2% , 40.5%)
75 - 84	58.9%	(55.4% , 62.3%)
> 84	42.3%	(37.9% , 48.2%)
Females		
< 65	57.7%	(56.6% , 58.7%)
65 - 74	69.1%	(66.4% , 71.6%)
75 - 84	56.5%	(53.6% , 59.3%)
> 84	36.4%	(33.2% , 40.5%)

Table 4. Estimates of Sensitivity for Pharmacare Cases

	Pharmacare Drugs Mean	(Lower , Upper)
Males		
< 65	n/a	
65 - 74	23.4%	(20.9% , 26.3%)
75 - 84	19.0%	(17.4% , 20.6%)
> 84	16.0%	(13.6% , 18.9%)
Females		
< 65	n/a	
65 - 74	19.1%	(17.9% , 20.2%)
75 - 84	21.4%	(19.9% , 22.9%)
> 84	23.4%	(20.9% , 26.3%)

Table 5. Estimates of Specificity for the 1 in 1 Administrative Case Definition.

	1 in 1 rule Mean	95% Credible Interval (Lower , Upper)
Males		
< 65	99.97%	(99.93% , 100.00%)
65 - 74	99.04%	(98.00% , 99.91%)
75 - 84	95.42%	(94.34% , 96.64%)
> 84	98.86%	(97.19% , 99.94%)
Females		
< 65	99.88%	(99.83% , 99.93%)
65 - 74	95.80%	(95.24% , 96.42%)
75 - 84	97.50%	(96.68% , 98.41%)
> 84	99.04%	(98.00% , 99.91%)

Note: The specificity of DCPNS and Pharmacare were assumed to be 1.0 since they are gold standard cases.

Table 5 shows the estimated specificity for the 1-in-1 administrative case definition. The estimates are quite variables between age-sex groups, and most are very high, approaching 100%. Overall, these estimates are higher than have been obtained in previous work (Kephart et al., 2004). This fact coupled with the lack of a clear pattern by age call in to question the trustworthiness of these specificity estimates.

Estimates of the positive predictive value (PPV) for alternative administrative case definitions in the population of persons who are not in either of the gold standard data sources.

We next estimated the prevalence of DDM in the non-gold-standard case population, and then combined the NGP prevalence estimates with estimates of sensitivity and specificity for alternative case definitions to compute their positive predictive values. As noted, we used estimates of sensitivity and specificity from previous work for the calculations (Kephart et al., 2004). Table 6 shows estimates of prevalence in the total population (from Table 1) and the NGS population. The prevalence in the NGS population is less than half that of the general population. Table 7 describes the alternative administrative case definitions for which PPV were calculated.

Table 6. Estimates of Prevalence for the Total and NGS Population, 2006

		Population Prevalence	NGS Prevalence
Males			
	< 45	1.6%	0.7%
	45 - 54	9.1%	4.1%
	55 - 64	18.4%	8.2%
	65 - 74	15.9%	7.7%
	75 - 84	22.1%	7.4%
	> 84	20.5%	9.9%
Females			
	< 45	1.5%	0.6%
	45 - 54	7.1%	3.0%
	55 - 64	14.2%	6.0%
	65 - 74	16.8%	4.2%
	75 - 84	19.8%	6.8%
	> 84	15.9%	7.7%

Table 7. Definitions of Alternative Administrative Case Definitions

1 in 1	One or more physician or hospital claims with a diagnosis of diabetes (ICD-9 250) within the previous year.
2 in 1	At least two physician claims or at least one hospital claim with a diagnosis of diabetes within the previous year.
Manitoba rule	At least two physician claims or at least one hospital claim with a diagnosis of diabetes within the previous two years (i.e. the Manitoba Rule).
NDSS	Met the Manitoba rule anytime within the last five years (The NDSS has now reduced this to three years)
Kephart Rule	A Bayesian rule, developed in our previous validation work, which uses all of the information on all the years of administrative data to estimate the probabilities of having DDM. Cases are defined based on probability cutoffs.

Table 8 shows the computed PPVs for alternative administrative case definitions. Overall, the Manitoba Rule has the best positive predictive values. While it may miss some cases, it will yield few false-positive cases. In contrast, the NDSS case definition performs very poorly. Depending on the age and sex group used, 20-80% of cases identified by the NDSS would be false-positives. In the population under age 45, nearly 80% of cases identified by the NDSS rule would be false positives. It should be noted, however, that the version of the NDSS rule used in Table 8 is based on 5-year run in of data. The NDSS methodology has since been modified to only use a three -year run in of data. As a result, better PPVs would be expected, but they will still be considerably lower than the PPVs for the Manitoba rule.

Table 8. Estimates of Positive Predictive Values (PPVs) for Alternative Administrative Case Definitions

		NDSS (se=.926, sp=.977)	Kephart (se=.853, sp=.988)	Manitoba (se=.771, sp=.999)	2 in 1 (se=.613, sp=.999)	1 in 1 (se=.779, sp=.996)
Males						
	< 45	22.1%	33.3%	84.4%	81.2%	57.8%
	45 - 54	63.3%	75.3%	97.1%	96.3%	89.3%
	55 - 64	78.4%	86.5%	98.6%	98.2%	94.6%
	65 - 74	77.0%	85.6%	98.5%	98.1%	94.2%
	75 - 84	76.2%	84.9%	98.4%	98.0%	93.9%
	> 84	81.6%	88.7%	98.8%	98.5%	95.6%
Females						
	< 45	20.3%	31.1%	83.0%	79.5%	55.3%
	45 - 54	55.7%	68.9%	96.0%	95.0%	85.9%
	55 - 64	72.0%	81.9%	98.0%	97.5%	92.5%
	65 - 74	63.9%	75.8%	97.1%	96.4%	89.5%
	75 - 84	74.5%	83.7%	98.2%	97.8%	93.4%
	> 84	77.1%	85.6%	98.5%	98.1%	94.2%

Key Recommendations for the NSDR

Based on our analyses, we make the following recommendations, listed in order of priority.

1. **When using the Nova Scotia Senior's Pharmacare data to define gold standard cases of DDM to be incorporated into the NSDR, users of diabetic supplies only (e.g., test strips and lancets) should not be included as cases. Cases should be defined as users of insulin and/or oral medications.**

2. **National Diabetes Surveillance System (NDSS) case definition should not be used to identify additional DDM cases for the NSDR as the number of false positive cases is unacceptably high.**

3. **The Manitoba Rule (≥ 1hospitalization &/or ≥2 physician visits in 2 years for which DM is coded) should be used to identify additional cases of DDM which are not captured by the gold standard sources.**

   Rational: Using the NDSS case definition is likely to result in an unacceptably large number of false-positive cases of DDM. The risk of this error is low for the Manitoba case definition because it is applied cross-sectionally rather than longitudinally.

4. **Preferably, the prevalence of diagnosed diabetes should not be estimated directly from cases in the NSDR. Instead, the Bayesian estimation approach we have used in the validation analysis should be further refined and used as standard practice for estimating and reporting prevalence. In the interim, cases identified using the Manitoba Rule can be combined with gold standard NSDR cases to determine a reasonably accurate estimate of DM prevalence.**

   Rational: The Bayesian methods we employed estimate the prevalence of DDM while simultaneously adjusting for deviations from perfect sensitivity and specificity of the data sources. We found that the prevalence estimates were very robust to alternative model specifications. However, the estimates of prevalence for males age 65-74 were suspect, being lower than expected. Further refinements of the estimates should be undertaken, and the method used on a routine basis to estimate DDM prevalence for reporting purposes.

   Using the NSDR to directly estimate prevalence assumes that there are no ascertainment errors, and that ascertainment errors do not vary by age and sex. These assumptions are not substantiated by our analyses. However, we believe that using gold standard cases in the RCP, DCPNS and NSPP-S data, supplemented by additional cases based on the Manitoba Rule, can be used as an interim approach for estimating DDM prevalence.

5. **Additional work should be conducted to further refine the Bayesian estimates of prevalence, sensitivity and specificity of the data sources in the NSDR.**

   Rational: The Bayesian methods we employed in this study make a number of simplifying assumptions to facilitate estimation. These may result in some bias in our estimates. We believe that further analyses, which relaxes model assumptions while using more data, shows considerable promise for refining estimates and improving surveillance capacity.

6. **Additional work should be conducted to refine the Kephart Rule (an administrative case definition based on Bayesian methodology) for use in the NSDR (Kephart et al., 2004).**

   Rational: There is an important conceptual distinction between the NDSS and Kephart rules on one hand, and the remaining rules on the other. The Kephart rule (and to some degree the NDSS rule) are designed to identify cases than can be followed over time. Once a person is identified as a DDM case, they remain in that state in subsequent years. This is not necessarily true for case definitions like the Manitoba rule, which use different claims data in each year to infer DDM status. Accordingly, a person's DDM status may be classified differently from year to year.

   The Manitoba rule will work well when NSDR applications only need to identify DDM status of person as specific points in time. This will apply to the majority of expected applications. However, for follow-up studies of the patients with DDM, this will be problematic. For such studies, the Kephart rule should be adapted for use in the NSDR. This is feasible, but will require establishing new probability cutoffs for identifying DDM cases.

Appendix A: Gold Standard Case Definitions for Diagnosed Diabetes Mellitus

DCPNS **1992 forward**	Any NS resident with a valid HCN, who is eligible to receive health care services under the MSI program and who has made ≥1 visits to a NS DC – able to distinguish PreDM, type 1 DM, type 2 DM, and GDM
RCP **1980 forward**	Any woman with a White's classification, ICD-10-CA code, or NSAPD code for DM during a pregnancy-related admission – able to distinguish GDM and pre-existing type 1 and type 2 DM
NSPP **1980 forward**	Any senior with one or more claims for insulin, OAA, or test strips[‡] – includes PreDM, type 1 DM, and type 2 DM but cannot distinguish between types

Appendix B: Detailed Description of Bayesian Estimation of the Accuracy of Case Definitions for the NSDR Pilot Project

In all of the previous validation work on the coding of diabetes in Nova Scotia administrative data, estimates of sensitivity and specificity of case definitions were obtained by comparing administrative data to "gold standard" cases and non-cases of diabetes. Gold standard cases were obtained from the DCPNS Registry, and gold standard non-cases were obtained from survey data. Clearly, the latter is tarnished and subject to reporting error, which is a major limitation of previous work (Kephart et al., 2004).

There is now considerable interest in the literature in using newer, Bayesian methods that estimate sensitivity and specificity of administrative case definitions in the absence of a gold standard (Branscum et al., 2005; Enoe et al., 2000; Joseph et al., 1995; Rutjes et al., 2007). These newer methods were used for most of the validation analysis we conducted as part of the NSDR pilot project. The purpose of this document is to provide an accessible introduction and explanation of the methods used.

A Simplified Example

To describe the approach and the estimation challenges, we use a simple example. The simplest form of data we used was extracted from our validation data set. We constructed a frequency table like the following, which classifies all persons in the population according to two tests (e.g., an administrative case rule and DCPNS). For example, $\mathbf{n_{11}}$ is the number of persons who were classified as having diabetes (i.e. Test +) in both the administrative data and in the DCPNS, and $\mathbf{n_{22}}$ is the number of persons who were not classified as having diabetes (i.e. Test -) in either data source. Note that the table includes 4 pieces of data ($\mathbf{n_{11}, \ldots, n_{22}}$), but only 3 pieces of information (because any one cell frequency can be computed if we know the total and the other three). Thus, there are only 3 "degrees of freedom" in this data.

	DCPNS (test 2)	
Admin Rule (test 1)	**Test +**	**Test -**
Test +	n_{11}	n_{12}
Test -	n_{21}	n_{22}

We want to estimate the following parameters**:**

π = disease prevalence

Sp1 = specificity of admin rule (i.e. the % of true non-cases captured)

Sp2 = specificity of DCPNS

Se1 = sensitivity of admin rule (i.e. the % of true cases captured)

Se2 = Sensitivity of DCPNS

A "model "consists of the relationships between the cell frequencies in the data table and the parameters that we want to estimate. If we assume that the two tests are independent (i.e., the accuracy of one test

does not depend on the result of the other test), then a model consisting of the following set of equations applies:

$$n_{11} = \pi(Se1)(Se2) + (1-\pi)(1-Sp1)(1-Sp2)$$
$$n_{12} = \pi(Se1)(1-Se2) + (1-\pi)(1-Sp1)(Sp2)$$
$$n_{21} = \pi(1-Se1)(Se2) + (1-\pi)(Sp1)(1-Sp2)$$
$$n_{22} = \pi(1-Se1)(1-Se2) + (1-\pi)(Sp1)(Sp2)$$

Basically, the equations express the cell frequencies as a function of the probability of having the disease and the conditional probabilities of testing positive or negative with each test. In other words, each cell frequency is a function of whether or not the person has the disease (given by the prevalence), and whether they are correctly classified as such by each tests (which depends on the sensitivity and specificity of each test). For example, expressed in common language, the first equation says that the n_{11} cell frequency will consist of those who have the disease and correctly test positive to both tests (the first half of the equation), and those who don't have the disease and who falsely test positive to both tests (the second half of the equation).

This simple "model" describing the relationship between the data and the parameters may be wrong because the tests may not be independent. It is likely, for example, that patients seen in Diabetes Centres (DCs) are more likely to be coded as having diabetes in the administrative data patients are referred to the DCs by their physicians. If we want to allow for dependence between the tests, then an alternative model with more complex equations is required. The more complex model includes up to two additional parameters describing the dependence between the tests ("covariance" terms) which need to be included in the equations and estimated.

For the 2-test, independence model above, we attempted to use a Bayesian estimation approach to estimate the parameters using "Gibbs sampling" and Winbugs software. The software draws a sequence of samples from a "prior" distribution of each parameter, in turn, conditional on values of all the other parameters. This sequence is repeated thousands of times to estimate the joint distribution of all the parameters (we used a minimum of 10,000 and usually 100,000 sequences per analysis). The data and information on prior distributions of each parameter are provided as inputs into the software. The prior distribution of each parameter can take many forms. For example, it can be fixed at a given value (e.g. Sp2=1.0), or left as a uniform distribution that includes all possible values. For example, all of the parameters we want to estimate take values in the range of 0-1, so we *usually* used uniform prior distributions with the range 0-1. Because the uniform distribution assumes that all valid values are equally likely, we call it a "non-informative prior". We can also use information from other studies to define an "informative" prior distribution for a parameter. For example, based on previous validation work we could specify a prior distribution for the sensitivity of an administrative data rule. It would be a bell-shaped distribution with a mean and variance corresponding to the previous estimate (we used a beta distribution).

The end result of the Bayesian estimation process is a "posterior" distribution for each parameter. The output includes graphs showing the distributions and detailed tables describing the mean, median and percentiles of the distributions of each parameter. The 2.5^{th} and 97.5^{th} percentiles of the posterior distribution of each parameter provide a type of confidence interval (usually called a 95% credibility interval).

Unfortunately, estimating the parameters for the data and model in this example is difficult. If uniform prior distributions with values from 0-1 are assumed (i.e. "non-informative" priors), this model and data will not yield consistent posterior distributions because we have 5 parameters but only 3 degrees of freedom. The posterior distributions have wide credibility intervals with odd shapes, or back up against one end of the prior distribution (e.g. Sp1=1.0). This situation happens because there is no unique solution to the equations, and many possible combinations of the parameters could produce the observed data. We call such a model "under-identified". To understand why the model cannot be estimated, the problem can also be viewed as an algebra problem. In algebra terms, we could say that we have 6 unknowns and only 4 equations. If one were to do the algebra to try and solve for the unknown parameters, it would not work. There is no unique solution.

We can "identify" (make estimable) the model by making assumptions about the values of some of the parameters. In an identified model, there is a unique solution. To identify the model, we will need to have at least as many degrees of freedom as the number of parameters we want to estimate. For example, we believe that the specificity of the DCPNS data is perfect (there are no false positive cases), so we can reasonably assume that Sp2=1.0. We made this assumption in all of our analyses. However, the model will still not be identified (4 parameters and 3 degrees of freedom). Further assumptions are difficult to make, but if we are willing to do so, we can identify the model and obtain consistent estimates of the model parameters. For example, we have conducted previous research to estimate the sensitivity of the administrative data, and we can choose to set this parameter to previous estimates, or to specify an "informative" prior distribution that corresponds to the estimated sensitivity from earlier work. Using such prior knowledge has been used in some of the published literature on estimating the accuracy of diagnostic tests but has been criticized because it arrives at apparently meaningful and consistent estimates by making lots of assumptions.

How We Approached the Bayesian Analyses

To estimate the parameters of interest, we had to use combinations of data and models that were identified. For two reasons, we chose not to identify the models by using estimates from previous work to define prior distributions. First, the previous work is somewhat dated and may not apply to the current environment. For example, alternative payment and block alternate funding plans to academic clinicians have likely altered the accuracy of the administrative data. In addition, we observed in our data that using different gold standard data sources resulted in different estimates of the accuracy of administrative data. For example, estimates of the sensitivity of administrative data obtained by comparing it to Pharmacare cases were much higher than estimates obtained by comparison to DCPNS cases.

In our Bayesian analysis, we used two different strategies to obtain combinations of data and models that were identified and that could produce consistent estimates of model parameters. The different strategies were used for the population over and under age 65.

The Population Age 65 and Over

For the population age 65 and over, we used data from three tests. Using 2006 data, we constructed tables classifying all MSI eligible persons according to whether or not they were classified as having DDM according to each of three tests:

- Pharmacare (had a claim in 2006 for insulin or an antihyperglycaemic drug)
- DCPNS (had ever appeared as a patient in a Diabetes Centre in the DCPNS data).
- An administrative case definition (we usually used the 1-in-1 rule: ≥1 physician or hospital claim in one year).

This resulted in a 2×2×2 table with 7 degrees of freedom. We assumed that the specificity of both the Pharmacare and DCPNS tests was equal to 1.0 (i.e., no false-positive cases), leaving 5 parameters to be estimated: the sensitivity of each of the three tests, the specificity of the administrative case definition, and the disease prevalence. The model was thus identified. The analyses were stratified by age group (65-74, 75-84, and 85+) and sex, thus yielding six sets of parameter estimates.

The key limitation of this approach was the assumption of independence between tests. It is very likely that the tests are not independent. In particular, we expect that a positive test in Pharmacare case definition is associated with a positive test in the administrative data since both are generated by the same physician. However, accounting for test dependence is challenging. There are six possible covariance terms that could be included and inadequate degrees of freedom to estimate them. We have not estimated models with test dependence in this population. We expect positive covariance between the tests. From simulations we conducted, we have shown that assuming independence when there is positive covariation will result in overestimation of the sensitivity and underestimation of the prevalence.

The Population Under Age 65

For the population under age 65, a different strategy was required. While there is a third gold standard test available (the RCP data), it has very low sensitivity (i.e., captured few cases) and could not be used. Thus, we were restricted to approaches that used two tests.

We adopted an approach which has been widely used in the literature. It is based on data for which two independent tests are administered to two or more independent populations that have different disease prevalence. It is assumed that the sensitivity and specificity of the tests are the same when applied to each population, and independence of the two tests is assumed. The data for the model consists of a 2×2 table for each population). As applied to our data, we used three different age groups as the independent populations (0-44, 45-54, or 55-64). The two tests consisted of the DCPNS and an administrative case definition. To relax the assumption of equal sensitivity and specificity across all three age groups, we also ran the analyses for pairs of age groups (0-44 and 45-54, or 45-54 and 55-64). We also ran analyses that included dependence between the tests.

Using all three age groups, we had three 2×2 tables, each with 3 degrees of freedom (for a total of 9). The model included the following seven parameters.

π_1= disease prevalence in age group 1

π_2= disease prevalence in age group 2

π_3= disease prevalence in age group 3

Sp1 = specificity of the admin rule

Sp2 = specificity of DCPNS

Se1 = sensitivity of admin rule

Se2 = Sensitivity of DCPNS

The specificity of the DCPNS data was assumed to be 1.0, leaving six parameters to be estimated with 9 degrees of freedom. For analyses where just two age groups were used, we had 5 parameters and 6 degrees of freedom.

References

Branscum, A. J., Gardner, I. A., & Johnson, W. O. (2005). Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling. *Prev Vet Med*, 68(2-4), 145-163.

Enoe, C., Georgiadis, M. P., & Johnson, W. O. (2000). Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. *Prev Vet Med*, 45(1-2), 61-81.

Joseph, L., Gyorkos, T. W., & Coupal, L. (1995). Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *Am J Epidemiol*, 141(3), 263-272.

Kephart, G., Casey, J., Ranger, R., Dunbar, P., & Karlovic, Z. (2004). The development and validation of an alternative case definition for the National Diabetes Surveillance System. Final Report to the Data Validation Working Group: National Diabetes Surveillance System. Ottawa, Health Canada.

Rutjes, A. W., Reitsma, J. B., Coomarasamy, A., Khan, K. S., & Bossuyt, P. M. (2007). Evaluation of diagnostic tests when there is no gold standard. A review of methods. *Health Technol Assess*, 11(50), iii, ix-51.

APPLICATION FOR ACCESS TO PROVISIONAL NSDR DATA

Requestor Internal to Department of Health/Provincial Programs

Section A: Administrative Information

Date	
Name	
Position	
Division/Program	
Address	
Phone	
Fax	
Email	

Section B: Data Requirements

Describe the reason for this data request (e.g., question to be answered).

Describe the expected audience.

Describe the format in which the data are required with details about file format or table format.