# Can the prevalence of diagnosed diabetes be estimated from linked national health records?

## The validity of a method applied in New Zealand

**Simon Thornley** MPH;[1] **Craig Wright** BSc;[2] **Roger Marshall** PhD;[1] **Gary Jackson** MPH;[3] **Paul L Drury** MA, MB BChir;[4] **Susan Wells** PhD;[1] **James Smith** MPH;[5] **Wing Cheuk Chan** MPH;[6] **Romana Pylypchuk** MPH;[1] **Rod Jackson** PhD[1]

[1] The University of Auckland, Auckland, New Zealand

[2] Ministry of Health, New Zealand

[3] Health Partners Consulting Group, Auckland

[4] Diabetes Services, Auckland

[5] Queensland Health, New Zealand

[6] Counties Manukau District Health Board, New Zealand

## ABSTRACT

**INTRODUCTION:** With projected global increases in the prevalence of Type 2 diabetes, the health sector requires timely assessments of the prevalence of this disease to monitor trends, plan services, and measure the efficacy of prevention programmes.

**AIM:** To assess the validity of a method to estimate the prevalence of diagnosed diabetes from linked national health records.

**METHODS:** We measured the agreement between a diabetes diagnosis (using combined national lists of drug dispensing, outpatient attendance, laboratory tests (HbA1c) and hospital diagnoses) and a primary care diabetes diagnosis in a (PREDICT™) cohort of 53 911 adult New Zealanders. The completeness of the diagnosis of diabetes in the cohort was estimated using capture-recapture methods.

**RESULTS:** The primary care cohort had a high prevalence of recorded diabetes (20.9%, 11 266/53 911), similar to our derived prevalence of 20.1%. Of the participants with a diagnosis of diabetes, 89% (10 182/11 266) had a similar derived diagnosis, indicating that only about one in 10 people with a primary care diagnosis had not been either admitted to hospital, seen at outpatient clinics, prescribed diabetes drugs or undertaken regular HbA1c tests. The capture-recapture prevalence of diagnosed diabetes in this cohort was 23.7% indicating that primary care diagnoses in the cohort were about 90% complete.

**DISCUSSION:** A method for estimating the prevalence of diagnosed diabetes from national health data shows high-level agreement with primary care records. Linked health data can provide an efficient method for estimating the prevalence of diagnosed diabetes in regions where such records are individually linked.

**KEYWORDS:** Diabetes mellitus, Type 2; epidemiology; prevalence; medical records

J PRIM HEALTH CARE 2011;3(4):262–268.

**CORRESPONDENCE TO:**
**Simon Thornley**
Department of Epidemiology and Biostatistics, School of Population Health, The University of Auckland, PB 92019 Auckland, New Zealand
sithor@woosh.co.nz

## Introduction

The prevalence of diabetes has increased dramatically over the last half century, despite public health efforts to restrain it.[1,2] Diabetes is an important risk factor for cardiovascular disease, and the health costs associated with treating the disease and its sequelae are considerable. Surveillance of diabetes prevalence is important to assist resource allocation decisions, assess the efficacy of nutrition regulation and health promotion programmes intended to reduce the incidence of this disorder, monitor services, and inform quality-of-care programmes for diabetes.

Regular large community surveys would be the ideal way to assess trends in diabetes prevalence, but two major problems beset this approach: non-response bias and expense. Non-response bias limits the generalisability of results and may bias estimates. Further, response rates to community surveys have been falling over the past 25 years.

For example, the 2006 New Zealand Health Survey had a non-response rate of 30%.[3]

An alternative to surveys is the use of health-related administrative datasets. In New Zealand the National Health Index (NHI) number is a unique identifier that is assigned to almost all New Zealand residents (98% in 2008). This allows the merging of different records to develop a picture of an individual's health and treatment received. Use of some drugs is largely limited to the treatment of single diseases; thus oral hypoglycaemic agents and insulin are almost exclusively restricted to the treatment of diabetes. From this assumption, dispensing lists may be combined with recorded hospital discharge diagnoses of diabetes to infer that an individual has diabetes.

We have been investigating the use of linked national health datasets for several years to assess diabetes prevalence and these are currently used to plan health services. The accuracy of diabetes prevalence estimates derived from such combined datasets is uncertain. In the current study, we assessed their accuracy in a large primary care–based cohort which had its diabetes status formally documented while undergoing cardiovascular risk assessment. The level of under-count of diagnosed diabetes in the cohort was also estimated using capture-recapture methods.

## WHAT GAP THIS FILLS

**What we already know:** Extensively linked health data may indicate whether individuals in the New Zealand population have diabetes. This information may be aggregated to derive population estimates of diabetes prevalence; however, the accuracy of this method is unknown.

**What this study adds:** Combined diabetes drug use, laboratory test, hospital diagnosis, and outpatient clinic data show high-level agreement with an independently derived primary care diagnosis of diabetes. Such methods can yield an accurate estimate of prevalence of diagnosed diabetes in New Zealand.

**Surveillance of diabetes prevalence is important to assist resource allocation decisions, assess the efficacy of nutrition regulation and health promotion programmes... monitor services, and inform quality-of-care programmes for diabetes.**

## Methods

### Study population

We used diabetes diagnosis status in a cohort of 53 911 primary care patients who had completed a formal cardiovascular risk assessment as the comparator for assessing the validity of our derived diabetes prevalence estimate. The cohort had been risk assessed using PREDICT™, which is a web-based clinical decision support system, to generate a Framingham-based cardiovascular risk assessment and provide patient-specific management advice. PREDICT™ automatically writes the risk profile to the patient's electronic health record and also anonymously stores a copy on a secure server identified only by an encrypted NHI code. We used this latter database for our analyses.

The PREDICT™ cohort has been described elsewhere.[4] For these analyses, we used a subset of patients who had been risk assessed between 1 January 2007 and 15 December 2008 as part of routine primary care practice. This population consisted of 53 911 patients, mainly from the Auckland and Northland regions of New Zealand, who had attended primary care practices that use PREDICT™. They were expected to have a higher prevalence of diabetes than the general population because, at the time of these analyses, fewer than 20% of the eligible population had been assessed, and initial screening targeted higher-risk patients. Cohort participants who died during this period were excluded from

the dataset. Therefore, all study participants were alive throughout the study period and had the same 'health exposure time' so that they could have been recorded in national datasets.

*Documented primary care diagnosis of diabetes*

As part of the risk assessment, general practitioners or practice nurses must assign a diabetes status to all patients on a template to either 'no-diabetes', 'type-1', 'type-2' or 'type-unknown'. A label of diabetes in the PREDICT™ cohort was assumed to be an accurate 'documented primary care diabetes diagnosis' because, when a patient was assigned a diagnosis of diabetes, a series of additional questions specifically relevant to this diagnosis automatically appeared on the template and had to be answered before the risk assessment could be completed and the data stored. This will be referred to as the 'Predict diabetes diagnosis' from here on.

*Derived diagnosis of diabetes from national health data*

A range of recorded health care activities from national, routinely collected health data was gathered on this Predict cohort, and combined to derive a diagnosis of diabetes to compare with the Predict diabetes diagnosis. Any appearance in one of the following four national lists was used as evidence of diabetes:

1. A hospital discharge diagnosis of diabetes anywhere among the coding (ICD 10: E10-E14, O24.0 to O24.3, ICD 9:250 all excluding ICD 10:O24.4 (diabetes arising in pregnancy)), taken from 1998 to 2008
2. Outpatient visits to specialist diabetes clinics (2004 to 2007)
3. Dispensing lists from community pharmacies for oral hypoglycaemic agents or insulin (2001 to 2008)
4. Five or more plasma HbA1c tests between 1 July 2006 and 30 June 2008 in lists of laboratory test claims (these indicate occurrence of the test only, not the results).

In contrast to the Predict diabetes diagnosis, we refer to this indicator of diabetes as the 'derived diabetes diagnosis'. Five or more HbA1c test claims on the laboratory list in two years were used as the final criterion because preliminary investigation showed this cut-off resulted in the best agreement with the other three lists for patients diagnosed with diabetes.

## Capture-recapture and statistical methods

To test whether the prevalence of diabetes diagnoses (from either method—Predict or derived) was complete, we estimated the overall prevalence of diagnosed diabetes in this cohort by using a statistical technique known as capture-recapture. Such a method links the three national lists most likely to accurately represent diabetes diagnoses (hospital discharge codes, hospital diabetes outpatient clinic visits, dispensing of diabetes drugs, and diabetes from Predict) to the counts found in intersecting combinations of these lists.

Each of these lists was considered a proxy for diagnosed diabetes, but none were expected to be complete; in contrast, capture-recapture methods estimate a 'virtual' complete total. Capture-recapture estimates of diabetes prevalence have been calculated in Italy[5] and the UK[6] using similar datasets to those available in New Zealand.

We used log-linear models to adjust for between-list dependence, using the Rcapture utility[7] of the R-project.[8] Numbers of people with diagnosed diabetes, in varying combinations of lists, were modelled as dependent variables; with independent variables comprising dummy indicators of the included lists. Interaction terms, which accounted for between-list dependence, were included in the models. The model with the least number of interaction terms that also demonstrated evidence of good model fit, was selected to estimate prevalence. Model fit was estimated by comparing Akaike's Information Criterion (AIC), chi-square statistics, and plots of Pearson residuals with predicted values, for competing models. All other calculations were carried out using the R-project (Epicalc[9] utility) or Microsoft Excel™. Scaled rectangle diagrams (similar to Venn diagrams) to display overlap in the datasets used for the combined list and capture-recapture methods were drawn using SPAN software.[10]

*Table 1. Cohort characteristics, by Predict diabetes diagnosis (compared to 2006 Census, aged ≥ 15 years)*

| | Primary care diabetes diagnosis? | | Predict cohort total | Census* |
|---|---|---|---|---|
| | Yes n (row %) | No n (row %) | n (col. %) | (col. %) |
| **TOTALS** | 11 266 | 42 645 | 53 911 | |
| **Gender** | | | | |
| **Male** | 5966 (20.2) | 23 617 (79.8) | 29 583 (54.9) | (48.8) |
| **Female** | 5300 (21.8) | 19 028 (78.2) | 24 328 (45.1) | (51.2) |
| **Age category** | | | | |
| **15–24** | 35 (29.7) | 83 (70.3) | 118 (0.2) | (21.8) |
| **25–34** | 186 (18.3) | 830 (81.7) | 1016 (1.9) | (19.9) |
| **35–44** | 1096 (15.6) | 5943 (84.4) | 7039 (13.1) | (23.5) |
| **45–54** | 2661 (17.4) | 12 614 (82.6) | 15 275 (28.3) | (15.8) |
| **55–64** | 3487 (21.5) | 12 713 (78.5) | 16 200 (30.0) | (10.2) |
| **65–74** | 2548 (25.9) | 7303 (74.1) | 9851 (18.3) | (8.8) |
| **Over 75** | 1253 (28.4) | 3159 (71.6) | 4412 (8.2) | (21.8) |
| **Ethnic group** | | | | |
| **Other†** | 4946 (14.9) | 28 144 (85.1) | 33 090 (61.4) | (73.7) |
| **Pacific** | 3459 (34.9) | 6451 (65.1) | 9910 (18.4) | (5.3) |
| **Maori** | 1988 (23.1) | 6627 (76.9) | 8615 (16) | (12.4) |
| **South Asian** | 873 (38.0) | 1423 (62.0) | 2296 (4.3) | (8.6)* |
| **Deprivation index** | | | | |
| **1 and 2 (least deprived)** | 954 (14.2) | 5783 (85.8) | 6737 (12.5) | (18.6) |
| **3 and 4** | 1378 (16.7) | 6874 (83.3) | 8252 (15.3) | (18.7) |
| **5 and 6** | 1921 (18.4) | 8510 (81.6) | 10 431 (19.4) | (20.5) |
| **7 and 8** | 2775 (21.9) | 9893 (78.1) | 12 668 (23.6) | (21.9) |
| **9 and 10 (most deprived)** | 4217 (26.9) | 11 474 (73.1) | 15 691 (29.2) | (20.3) |

* Census estimate includes all Asian (that is Chinese, South Asian and South East Asian).

† Includes New Zealand European, New Zealand's largest ethnic group.

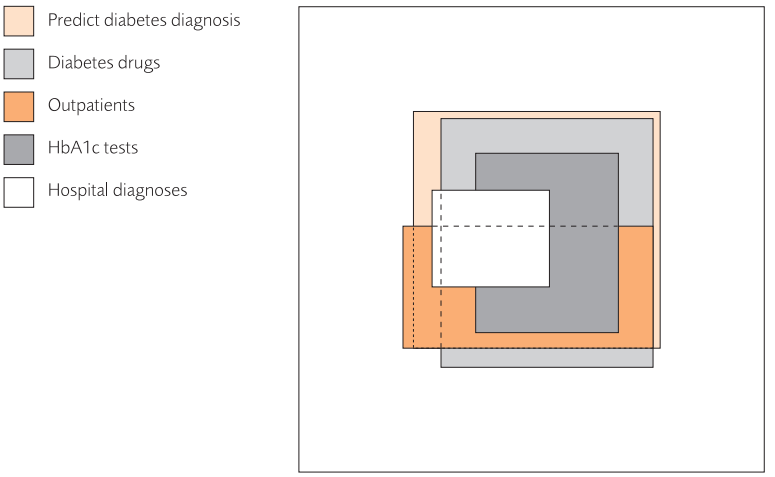
The Predict cohort study was approved by the national Multi-Region Ethics Committee in 2007 (MEC/07/19/EXP).

## Results

The overall prevalence of a documented diagnosis of diabetes in the Predict cohort was 20.9% (11 266/53 911) (Table 1). The Predict cohort also had higher proportions of males, Pacific and Maori ethnic groups and people of low socioeconomic status compared with proportions derived from census estimates (the final column in the Table 1). The highest prevalence of a documented diagnosis of diabetes was amongst South Asian people (38%; 873/2296), followed by Pacific peoples (35%; 3459/9910) and Maori (23%; 1988/8615). The remainder, who were mainly European, had a diagnosed diabetes prevalence of 15% (4946/33 090). People with diabetes were slightly older and a higher proportion were in low socioeconomic groups compared to those without such a diabetes diagnosis.

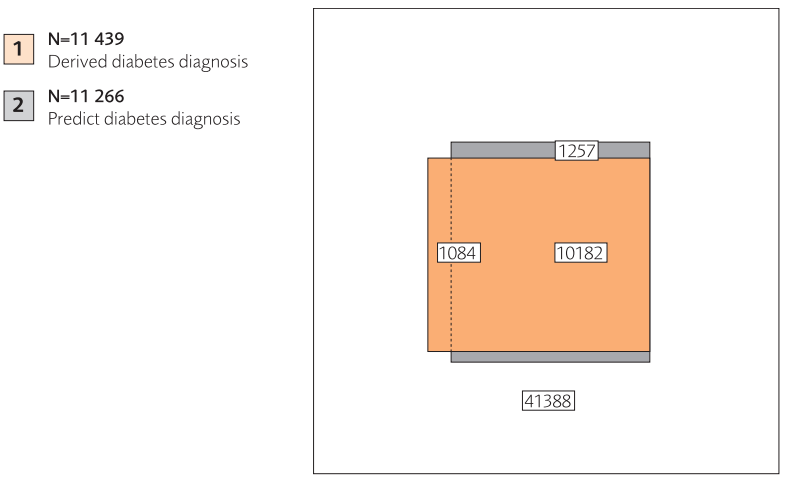
The agreement between groups with diabetes-related activity identified from the four national datasets and those with a Predict diabetes diagnosis in the cohort was high, shown by the

*Figure 1. Scaled rectangle diagram of overlap between elements of the 'algorithm'*




Diabetes drugs: Diabetes medication dispensing
Outpatients: Diabetes outpatient clinic attendance
HbA1c tests: ≥5 or more HbA1c test claims in two years
Hospital diagnoses: Hospital discharge diabetes diagnosis

*Figure 2. Scaled rectangle diagram of agreement between combined list algorithm-based (CLE) label of diabetes and the primary care (PREDICT™) diagnosis of diabetes*




scaled rectangle diagram (Figure 1). The outer rectangle represents the entire Predict cohort of 53 911 participants; the inner rectangles are the five overlapping groups with records of diabetes-related activity in national datasets (elements of the derived diabetes diagnosis) or a Predict diabetes diagnosis. The rectangles are scaled according to size of the groups and the degree of overlap. The diagram shows that nearly all of the patients with diabetes-related activities from the national outpatient (n=6009), inpatient (n=2176) and laboratory datasets (n=5069) are captured by the Predict diabetes diagnosis (n=11 266). While most people on the national drug dispensing list (n=10 157) also possess a Predict diabetes diagnosis, a substantial minority (9% [903/10 157]) did not have a Predict diagnosis of diabetes. Conversely, about one in 10 people (1084/11 266) who had a Predict diabetes diagnosis did not have a corresponding derived diagnosis.

The overlap between the derived and Predict diagnoses is illustrated in Figure 2. The derived diagnosis identified 89% (10 182/11 439) of people who had a Predict diabetes diagnosis, and 97% (41 388/42 472) of people who did not have a Predict diabetes diagnosis had the equivalent derived diagnosis. Of people identified with a derived diabetes diagnosis, 90% (10 182/11 266) had a corresponding Predict diagnosis, while among those without a derived diagnosis, 97% (41 388/42 472) had the same status in diabetes diagnosis in Predict. A non-conditional measure of agreement (Cohen's kappa coefficient) was 0.87, indicative of a high level of concordance between the two indicators.

The four lists (Predict diabetes diagnosis, hospital diabetes diagnosis, diabetes drug dispensing and outpatient diabetes clinic attendance) considered to be most specific for diabetes were combined to estimate the 'total' or 'virtual' prevalence of diagnosed diabetes in the cohort using capture-recapture methods. Plots of Pearson residuals indicated acceptable model fit for the chosen log-linear model, which also minimised Akaike's Information Criterion. The final number of people with a 'total' diagnosis of diabetes in this cohort was estimated at 23.7% (12 778/53 911; 95% CI 22.5%–25.9%), which is slightly larger than that estimated by combining the Predict and derived diagnoses (23.2%; 12 523/53 911). The derived estimate of people with diabetes was about 90% of this estimated capture-recapture total (11 439/12 778), similar to the Predict diabetes diagnosis estimate, which was 90% of the 'virtual' estimate (11 266/12 778).

## Discussion

Our study showed high levels of agreement between a derived diabetes diagnosis, based on combining lists of national, routinely collected health data, and a diagnosis from a primary care database, in people undergoing CVD risk assessment. This suggests that the combined list method may be a useful surrogate for diagnosed diabetes to monitor trends in diabetes prevalence. Secondly, we showed that an electronic clinical decision support system used in routine primary care practice captures up to 90% of all diabetes diagnoses among patients to which it is applied.

The main strength of this study is that we had access to a large primary care cohort in which a documented diagnosis of diabetes is likely to be accurate. Further, the capture-recapture analyses indicated that use of the Predict system results in over 90% of all people with diabetes in the cohort being labelled appropriately.

Conversely, a limitation of the study was the nature of the sampling which led to entry into the Predict cohort. The Predict population had a higher proportion of people with diabetes (21%) compared to the general population, estimated at 4.3% (based on self-report from the last national health survey[3]). This is likely to be due to many of the cohort participants being enrolled by their primary health care providers for a formal cardiovascular risk assessment, if they were believed to be at increased risk of developing CVD, by virtue of advanced age, or risk factors for the disease, or a combination of both.

The capture-recapture estimates of 'total' prevalence need to be interpreted with caution as they are limited by the assumptions that underlie their use. We used four lists, thought to be relatively specific for a diagnosis of diabetes; however, false positives may be present (e.g. metformin is sometimes used to treat polycystic ovarian syndrome) and may inflate prevalence estimates. Also, the number in intersecting lists is assumed to have a Poisson distribution (so that appearance on individual lists is independent of other subjects). Clustering of individuals on list counts by characteristics such as health service use, ethnic group, deprivation and age are likely to occur, and so the precision of the estimates may be overinflated. Another weakness is that patients who are treated with diet or lifestyle measures alone who have not been admitted or been to a diabetes clinic, would not be captured unless they had five or more HbA1c tests within two years—more than recommended by national guidelines.[11]

The use of routinely collected national data has limitations. Because the datasets are not collected primarily for research purposes, significant heterogeneity in record collection in different administrative regions may occur. The completeness of linkage of health data by national health identifier has improved markedly over the last three to four years in New Zealand, but wider variability in the linkage of older health records used may affect the accuracy of the derived diagnosis method. These data quality issues should, however, improve further with time.

Internationally, authors have reported the use of derived diabetes prevalence from health records and capture-recapture estimates of diabetes prevalence, although not in the same study. The combined list estimate method has been used previously in Denmark[12] to describe time trends in diabetes prevalence between 1995 and 2006. The researchers used a similar technique to ours; however, five or more blood glucose measurements in a year, rather than aggregate numbers of HbA1c, was used as laboratory evidence of diabetes. A comparable study from Ontario used a rule that included a diagnosis in hospital discharge or outpatient records in the last two years as evidence that an individual has diabetes.[1] The sensitivity of the diabetes diagnosis associated with our derived diagnosis (89%) is similar to those quoted in Danish (85%)[12] and Canadian (86%) studies.[13] In contrast, a British study showed that the sensitivity of multiple record linkage, similar to the sources used in our study, was 91%, and more sensitive than general practice records.[14]

Local attempts to define the prevalence of diabetes have included combining primary health care registers using diagnosis codes, diabetes medications, laboratory tests and screening programme registers. This method has shown high levels of concordance between the lists, similar to our results; however, the method is limited by

organisational and geographic boundaries.[15] Our derived indicator of diabetes allows national prevalence of diabetes to be calculated.

Capture-recapture study examples include the Casale Monferrato study that monitored the prevalence of diabetes in this area of northwest Italy between 1988 and 2000.[5] Data from diabetes clinics, hospital discharge records, prescribing and sales of reagents and strips were combined to calculate population estimates. Comparing individual to capture-recapture methods, they found a diagnosis rate of about 80%, which is slightly lower than our estimate. In the United Kingdom similar capture-recapture studies of diabetes have been reported.[6,16]

We have previously carried out a more limited validation study of our derived diabetes prevalence method using a hospital-based diabetes register from a disease management programme as a gold standard.[17] We were aware that the register only included about half of the people with a diagnosis of diabetes in the population served by this hospital. Therefore, this earlier study was only able to assess the 'sensitivity' of the derived diagnosis. With these caveats, the proportion with a derived diagnosis, among those on the diabetes register (96%) from the earlier study, was higher than in the present analysis.

## Conclusion

A derived indicator of diabetes diagnoses, based on linking routine national data, shows substantial agreement with a documented primary care diagnosis of diabetes. Routinely collected health data can provide a rapid and efficient way of monitoring the prevalence of diagnosed diabetes, and its change over time, with reasonable accuracy.

**ACKNOWLEDGEMENTS**
We thank Dean Papa for assistance with data management.

**FUNDING**
A small grant was received from the New Zealand Ministry of Health to assist this analysis.

**COMPETING INTERESTS**
None declared.

### References

1. Lipscombe LL, Hux JE. Trends in diabetes prevalence, incidence, and mortality in Ontario, Canada 1995–2005: a population-based study.[See comment]. Lancet. 2007 Mar 3;369(9563):750–6.
2. Ministry of Health. Diabetes surveillance: population-based estimates and projections for New Zealand, 2001–2011: Public Health Intelligence Occasional Bulletin No. 46. Wellington: Ministry of Health, 2007.
3. Ministry of Health. A Portrait of Health: Key Results of the 2006/07 New Zealand Health Survey. Wellington: Ministry of Health, 2008.
4. Wells L. Getting evidence to and from general practice consultations for cardiovascular risk management using computerised decision support. Auckland: University of Auckland; 2009.
5. Bruno G, Merletti F, Bargero G, et al. Changes over time in the prevalence and quality of care of Type 2 diabetes in Italy: the Casale Monferrato surveys, 1988 and 2000. Nutr Metab Cardiovasc Dis. 2008 Jan;18(1):39–45.
6. Gill GV, Ismail AA, Beeching NJ. The use of capture-recapture techniques in determining the prevalence of Type 2 diabetes. QJM. 2001 Jul;94(7):341–6.
7. Baillargeon S, Rivest L-P. Rcapture: loglinear models for capture-recapture in R. Journal of Statistical Software. 2007;19(5):1–31.
8. R Development Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2007.
9. Chongsuvivatwong V. Analysis of epidemiological data using R and Epicalc. Prince of Songkla University; 2008.
10. Marshall RJ. Scaled rectangle diagrams can be used to visualize clinical and epidemiological data. J Clin Epidemiol. 2005 Oct;58(10):974–81.
11. New Zealand Guidelines Group. Management of Type 2 diabetes. Wellington: New Zealand Guidelines Group; 2003.
12. Carstensen B, Kristensen JK, Ottosen P, Borch-Johnsen K, Steering Group of the National Diabetes R. The Danish National Diabetes Register: trends in incidence, prevalence and mortality. Diabetologia. 2008 Dec;51(12):2187–96.
13. Hux JE, Ivis F, Flintoft V, Bica A. Diabetes in Ontario: determination of prevalence and incidence using a validated administrative data algorithm. Diabetes Care. 2002 March 1, 2002;25(3):512-6.
14. Morris AD, Boyle DIR, MacAlpine R, et al. The diabetes audit and research in Tayside Scotland (darts) study: electronic record linkage to create a diabetes register. BMJ. 1997 August 30, 1997;315(7107):524–8.
15. Joshy G. Linking existing databases to monitor and improve diabetes care. Hamilton: University of Auckland; 2010.
16. Ismail AA, Beeching NJ, Gill GV, Bellis MA. How many data sources are needed to determine diabetes prevalence by capture-recapture? Int J Epidemiol. 2000 Jun;29(3):536–41.
17. Thornley S, Marshall R, Jackson G, et al. Estimating diabetes prevalence in South Auckland: how accurate is a method that combines lists of linked health datasets? N Z Med J. 2010;123(1327):76–86.