

Improving Regional Level Estimates from National Surveys Using Census and Administrative Data: A case study using the New Zealand Health Survey

Robert Templeton

Principal Technical Specialist (Statistics), Ministry of Health

This report was commissioned by Official Statistics Research, through Statistics New Zealand. The opinions, findings, recommendations and conclusions expressed in this report are those of the author(s), do not necessarily represent Statistics New Zealand and should not be reported as those of Statistics New Zealand. The department takes no responsibility for any omissions or errors in the information contained here.

Abstract

The project investigated small area methods and their use in improving national survey estimates at a regional level for the New Zealand Health Survey. It had four specific aims:

- ensure methods and results are explained in lay terms as much as possible
- be able to incorporate data from administrative sources
- be able to produce measures of quality of the resulting estimates
- implement the methods in SAS.

In this case we illustrate the use of a small area estimator in the New Zealand Health Survey to produce estimates at the District Health Board (DHB) level. A statistical model (namely a generalised linear mixed model) can be used to produce the DHB estimates and assess their quality. A set of small area diagnostics are also examined to help assess the validity of the results. The modelling and diagnostic tools are those developed by Ray Chambers and other authors, who in various papers (referenced below) illustrate the methods as applied to the production of small area statistics from the UK Labour Force Survey. The procedures used in this case study are implemented in SAS programs.

Keywords

Small area estimation, composite estimation, random effects, generalised linear mixed models.

Reproduction of material

Material in this report may be reproduced and published, provided that it does not purport to be published under government authority and that acknowledgement is made of this source.

Citation

Templeton, R. (2009). Improving regional level estimates from national surveys using Census and administrative data: A case study using the New Zealand Health Survey. *Official Statistics Research Series, 4*. Available from <http://www.statisphere.govt.nz/official-statistics-research/series/default.htm>

Published by

Statistics New Zealand
Tauranga Aotearoa
Wellington, New Zealand

ISSN 1177-5017 (Online)
ISBN 978-0-478-31569-1 (Online)

Acknowledgements

I would like to thank Ray Chambers at the Centre for Statistical and Survey Methodology, University of Wollongong NSW for his very helpful advice. I would also like to thank Craig Wright (NZ Ministry of Health), who helped me greatly by developing and supplying much of the administrative indicator data used in the study and to Sarah Gerritsen (NZ Ministry of Health) the project manager of the 2006/07 Health Survey for her help and guidance on this project. In all cases however I take full responsibility for any errors or misinterpretation of the excellent advice given.

Contents

1 Background: The New Zealand Health Monitor, the New Zealand Health Survey and District Health Boards	6
2 Previous attempts at DHB estimates from the NZHS	7
3 Methods investigated for the 2006/07 NZHS.....	8
3.1 Estimators compared.....	8
3.2 Overview of the estimation process:.....	8
3.3 Generalised linear mixed model framework.....	9
4 Simplified interpretation of the GLMM approach.....	10
5 Implementation of the estimation process	11
6 The auxiliary administrative and Census data	12
7 Results.....	14
7.1 Results: DHB level prevalence of Diabetes	14
7.2 Results: DHB level prevalence of adequate vegetable intake	17
7.3 Results: DHB level prevalence of smoking	20
8 Overall Conclusions	23
References	25
Appendix 1 Summary of the 2006/07 NZ Health Survey sample design.....	26
Appendix 2 Description of ‘clustered synthetic’ methods used for DHB estimates in 2002/03	27
Appendix 3 Assessment of the DHB estimates for a selection of key adult variables	29
Appendix 4 Technical details of the estimation process	31
Appendix 4.1 Creating a dataset with DHB estimates by age/cell cell and effective sample sizes	31
Appendix 4.2 The generalised linear mixed model estimates and MSEs	31
Iterative procedure to find REML (residual maximum likelihood) estimates.....	31
Estimating totals, rates and their MSEs.....	32
Appendix 4.3 Estimating the synthetic estimates and their MSEs.....	32
Appendix 4.4 Small Area Diagnostics.....	33
Goodness of fit diagnostics	33
Coverage diagnostics	33
Bias diagnostics.....	34
Calibration diagnostics.....	34

List of tables

Table 1 DHB population sizes and NZHS sample sizes	7
Table 2 Explanatory variables used in the synthetic estimates.....	13
Table 3 DHB estimates of prevalence of diabetes.....	15
Table 4 Goodness of fit diagnostics – diabetes	15
Table 5 Bias diagnostics - diabetes.....	16
Table 6 Calibration diagnostics – diabetes.....	17
Table 7 DHB estimates of prevalence of adequate vegetable intake.....	18
Table 8 Goodness of fit diagnostics for adequate vegetable intake.....	18
Table 9 Bias diagnostics for prevalence of adequate vegetable intake.....	19
Table 10 Calibration diagnostics for prevalence of adequate vegetable intake.....	19
Table 11 DHB estimates of prevalence of current smoking	21
Table 12 Goodness of fit diagnostics for smoking:	21
Table 13 Bias diagnostics for smoking.....	22
Table 14 Calibration diagnostics for smoking	22
Table 15 Comparing estimates of RMSE	23
Table 16 Testing normality of regional effects	23
Table 17 Clusters of DHBs used in 2002/03 synthetic methods	27
Table 18 Analysis of DHB composite estimates for key health outcome variables	29

1 Background: The New Zealand Health Monitor, the New Zealand Health Survey and District Health Boards

The New Zealand Health Monitor (NZHM) is an integrated programme of population health surveys and record linkage studies managed by Public Health Intelligence (the epidemiology group of the Ministry of Health). NZHM surveys collect health information that cannot be collected during the process of health care. This includes information on the health status of New Zealanders, risk and protective factors associated with health, access to and the use of health services (especially primary health care services), and inequalities in health between population groups (for example, by age, sex, ethnicity and socio-economic position).

The New Zealand Health Survey (NZHS) is a general health survey and forms the foundation of the NZHM. Data are currently collected approximately every four years from a representative sample of New Zealand households. Previous NZHS only sampled adult (15 years old+) New Zealanders; however, children (0-14 years old) have been included in the 2006/07 NZHS and are planned for inclusion in future NZHSs. The 2006/07 NZHS was in the field from September 2006 for one year, with descriptive results due for release in mid-2008.

The target population of the NZHS is the total usually resident civilian population living in permanent private dwellings. The smallest sample needed to achieve sufficient accuracy is chosen in an attempt to minimise respondent load. Previous surveys have had samples of 7,500 to 12,000 people; however the 2006/07 NZHS has sampled approximately 12,000 adults and 5,000 children.

Data are collected face-to-face via trained interviewers in the respondents' homes. The NZHS is primarily a health interview survey, but also includes a small health examination component (measuring height, weight and waist circumference).

A summary of the sample design of the NZHS is included in Appendix 1.

The objectives of the NZHS are to:

- measure the health status of New Zealanders, and the prevalence of selected health conditions
- measure the prevalence of risk and protective factors associated with these health conditions
- measure the use of health services, including barriers to accessing health services
- examine differences between population groups (as defined by age, gender, ethnicity, and socio-economic position)
- examine changes in key NZHS data over time.

The focus of the NZHS is to produce national level data, however regional (District Health Board area) level data is increasingly requested due to the key role DHBs have in the health sector.

District Health Boards (DHBs) are responsible for providing, or funding the provision of, health and disability services in their district. There are 21 DHBs in New Zealand and they have existed since 1 January 2001 when the New Zealand Public Health and Disability Act 2000 came into force.

The statutory objectives of DHBs include:

- improving, promoting and protecting the health of communities
- promoting the integration of health services, especially primary and secondary care services
- promoting effective care or support of those in need of personal health services or disability support.

Other DHB objectives include promoting the inclusion and participation in society and independence of people with disabilities, reducing health disparities by improving health outcomes for Maori and other population groups, and to reduce toward elimination, health outcome disparities between various population groups.

Table 1
DHB population sizes and NZHS sample sizes

	DHB	NZHS survey population	Adult size NZHS sample 2006/07
1	Northland	112,000	712
2	Waitemata	379,000	1213
3	Auckland	323,000	1104
4	Counties Manukau	322,000	1301
5	Waikato	259,000	1417
6	Lakes	73,000	577
7	Bay of Plenty	150,000	847
8	Tairāwhiti	32,000	326
9	Taranaki	80,000	307
10	Hawke's Bay	112,000	586
11	Whanganui	47,000	225
12	MidCentral	122,000	489
13	Hutt	105,000	469
14	Capital & Coast	212,000	678
15	Wairarapa	30,000	117
16	Nelson Marlborough	102,000	291
17	West Coast	24,000	100
18	Canterbury	371,000	1019
19	South Canterbury	43,000	113
20	Otago	142,000	316
21	Southland	83,000	281

2 Previous attempts at DHB estimates from the NZHS

Table 1 shows clearly that there is a good deal of variation in the sizes of the DHBs. This means that the quality of the DHB estimates if produced directly from the survey data would also be quite variable, with a good deal of the key outputs not of very good quality.

For the 2002/03 NZHS a set of DHB level estimates was produced using a form of synthetic estimation. The approach adopted was to group DHBs into clusters that have similar characteristics with respect to health outcomes. Estimates for a particular DHB were obtained by re-weighting the (larger) cluster sample as though it were drawn only from the specific DHB population. The re-weighting process was a weighting adjustment using generalised regression to ensure the final weighted totals of eligible adult respondents are consistent with independent population estimates for that particular DHB. The survey was benchmarked to the 2001 Census population.

The key variables for which outputs were produced were a subset of the data produced at the national level. These outputs were available disaggregated by gender by Maori/Non-Maori, and for crude and age-standardised rates.

Sampling errors were produced but these did not reflect the likely error arising from the assumptions implied by the methodology (namely that the DHB's within clusters were the same with respect to the key outcome variables).

See Appendix 2 for a description of the methods used to calculate synthetic estimates for the 2002/03 NZHS.

3 Methods investigated for the 2006/07 NZHS

There is an ever growing body of literature on small area estimation. Rao (2003) gives a very comprehensive account of this. The logic behind choosing a generalised linear mixed model (GLMM) approach, but not others, goes something like this:

- We need to rely on statistical models (and hence make some assumptions) but this is inevitable given it is unlikely we can squeeze any more accuracy out of the direct survey data.
- We should use more than just a simple synthetic approach because a composite estimator is quite likely to result in estimates with lower mean squared error.
- Estimates derived using random effects models can be shown (at least for linear models) to be equivalent to composite-type estimators and the models provide a thorough statistical framework within which we can assess the quality of the estimates.
- Different techniques have been discussed to estimate the parameters of the models, including maximum likelihood and Bayesian techniques. Given that we want to produce official statistics, we should probably follow an approach with as few assumptions as necessary, as this will match the philosophy underpinning official statistics as best we can.

Considering these issues, the approach we have taken is to use a GLMM approach using maximum likelihood estimators when we fit the models. This approach lends itself to a description of the process (see section 4) which can side-step some of the more complicated detail involved in the modelling and estimation calculations (which will inevitably be beyond many of those seeking to use these sorts of estimates) but will still give them an understanding of the estimates and their basic properties.

3.1 Estimators compared

Three different estimators are compared in this study:

Direct survey estimates. These are standard survey sub-domain estimates for each of the DHBs. For many of the larger DHBs these will be of reasonable quality (and likely to be at least as good as any modelled estimate we can come up with). However for many of the smaller DHBs these will not be very good.

Synthetic estimates (or GLM estimates). These are estimates derived from a generalised linear regression model. This means we use relationships between the variable of interest and factors such as age, gender and ethnicity to estimate the variable for the DHB. A variant of this method is what was used in the 2002/03 NZHS to produce DHB estimates.

GLMM estimates (or approximate composite estimates). These estimates are derived from a generalised linear mixed model (i.e. modelled estimates where the regression model includes DHB indicators which are modelled as random effects). For simple models these can be shown to be equivalent to composite estimates (i.e. a linear combination of direct estimates and synthetic estimates.) Thus at least one attraction behind choosing the GLMM (or approximate composite estimates) is to make the best of the two sets of rival estimates (direct and synthetic), with the direct estimates being unbiased but more highly variable and the synthetic estimates tending to have low variability but involving significant bias.

Note: we also examined **explicit** composite estimates. Rather than constructing these directly from the GLMM modelling process, we create the synthetic and direct estimates separately and the explicit composite estimator is a weighted sum of the two. In practise these were found to be very close to the GLMM estimates (with correlations over .98 for most key outcome variables) and have not been presented separately in this report.

3.2 Overview of the estimation process:

There are two basic stages:

- i) The synthetic model is fitted and we analyse the results to ensure that *the synthetic estimates are of reasonable value* and that our assessment of their quality is robust. The whole idea here is that we can come up with an alternative set of estimates to ‘rival’ the direct survey estimates. This stage involves basic model fitting diagnostics:
 - demonstrating which factors provide significant discrimination to the model
 - looking at standardised residual plots
 - checking for any extreme values that might affect the model fit
 - looking at overall measure of model fit
 - looking at estimated MSE for the synthetic DHB estimates.
- ii) Looking at the estimated mean square error (MSE) and small area diagnostics to *assess the performance of the final GLMM DHB estimates*:
 - looking at the MSE at the DHB level
 - examining bias and fit of the DHB estimates in relation to the original direct estimates.
 - examining the coverage of the MSE derived confidence intervals
 - examining the amount of calibration needed to ensure consistency between the DHB and national level estimates.

We will calculate their MSE using robust variances estimation methods described in Ambler et al for the synthetic estimators and Saei and Chambers (1) for the GLMM estimators. The small area diagnostics (including measures of bias, goodness of fit, coverage and calibration) are as described in Brown et al.

3.3 Generalised linear mixed model framework

The model based estimators that we examine can be thought of as arising from a generalised linear mixed model framework. The following is a sketch of the modelling background. Fuller details of the models and estimators can be found in Saei and Chambers (2) and in appendix 4 of this report.

Firstly I will simplify this so that we are in fact considering a linear mixed model, which still allows us to explore the different estimators but without some of the details involved in the *generalised* linear model complicating the discussion.

This is what I will term the full model.

$$\hat{Y}_{c,dhb} = \underline{x}_{c,dhb} \underline{\beta} + u_{dhb} + e_{c,dhb}$$

The model is fitted using aggregate data so that $\hat{Y}_{c,dhb}$ is a DHB survey estimate of the total for an age/sex cell (the subscript c will denote age/sex cell. In this example there are three age groups used and hence 6 age/sex cells). The vector $\underline{x}_{c,dhb}$ represents a set of known covariates (usually age, gender, ethnicity, and perhaps some other explanatory variables derived from administrative sources). The parameters $\underline{\beta}$ are referred to as *fixed effects*. Putting all the DHB’s and cells together into a matrix we get the following form of the model in matrix terms:

$$\underline{y} = X\underline{\beta} + Z\underline{u} + \underline{e}$$

Where \underline{y} is now a vector of the survey estimates for each DHB by age/sex cell, X is matrix of the covariate information and the matrix Z represents a matrix of DHB indicators. The vector \underline{u} is a multivariate normal random vector with zero mean and covariance (σ_u^2). The elements of \underline{u} (u_{dhb}) are referred to as the *random effects*. The vector \underline{e} is a multivariate

normal random vector with each element having zero mean and variance ($\sigma_{c,dhb}^2$). (e and u are independent.)

Synthetic estimators can be derived from a simplified model (where we drop the Z matrix and the DHB random effects) and base the DHB estimates on the resulting predicted values from the model.

$$\hat{Y}_{syn,dhb} = \sum_{c=1}^6 X_{c,dhb} \hat{\beta} = \sum_{c=1}^6 \sum_{i=1}^{N_{c,dhb}} x_i \hat{\beta}$$

We can also form, what are termed empirical best linear unbiased predictors (EBLUPs) for the DHB estimates, based on the results of fitting the full mixed model:

$$\hat{Y}_{lmm,dhb} = \sum_c \left\{ \sum_{i=1}^{n_{c,dhb}} y_i + \sum_{i=1}^{N_{c,dhb} - n_{c,dhb}} (x_i \hat{\beta} + u_{dhb}) \right\}$$

In this project all the variables analysed were indicator variables. Hence we actually used a linear **logistic** mixed model we model the logit of the probability underlying the rate in each cell:

$$\logit(R_{c,dhb}) = \ln(R_{c,dhb} / (1 - R_{c,dhb})) = (x_{c,dhb} \beta + u_{dhb})$$

where

$$R_{c,dhb} = \frac{Y_{c,dhb}}{N_{c,dhb}}$$

We use a form of maximum likelihood estimation (REML= residual maximum likelihood) to find estimates of β and u and σ_u^2 . This involves an iterative algorithm the details of which are given in Appendix 4.

In practise we find the DHB estimates based on the mixed model are very close to a simply constructed composite estimate:

$$\hat{Y}_{glmm,dhb} \approx \hat{Y}_{comp,dhb} = \lambda_{dhb} \cdot \hat{Y}_{dir,dhb} + (1 - \lambda_{dhb}) \cdot \hat{Y}_{syn,dhb}$$

where

$$\lambda_{dhb} = \sigma_u^2 / (\sigma_u^2 + \text{var}(\hat{Y}_{dir,dhb}))$$

It is this approximate form that we will concentrate on in terms of trying to explain the properties of the resulting estimates. This approximation will start to breakdown for rare or very common health characteristics. The lowest correlation between the GLMM estimates and the explicit composite estimates was for stroke which has a prevalence less than 2%. However the correlation between the two sets of estimates was still pretty high at .96. For diabetes (with a prevalence of 5%), the correlation was .98, and for many of the more prevalent outcomes the correlation was over .99.

4 Simplified interpretation of the GLMM approach

For those of us who do not find the linear mixed model based framework particularly illuminating it is possible to describe the resulting estimators without much reference to this. The following is a very simplified step by step description of what happens using these sorts of estimators:

Step 1: Produce DHB estimates directly from the survey data (\hat{Y}_{dir})

Step 2: Produce modelled DHB estimates based on a simple regression using the demographic profile of the DHB, the DHB level administrative data we have and the national level survey data.

Step 3: Calculate the differences between the results from 1 and 2, and calculate an estimate of how much variation between these differences can be attributed to true 'DHB' effects ($= \hat{\sigma}_u^2$) and how much is most likely due to the known level of sampling error.

Step 4: Create a final set of DHB estimates which are a sum of the modelled estimates plus the estimated DHB effects. The estimated DHB effects will reflect the pattern of DHB differences found in step 3, but these DHB differences will be scaled down ('shrunk') so that they reflect the amount of 'true' DHB variation (i.e. by a factor which is in proportion to $\hat{\sigma}_u^2$ as estimated in step 3,)

These final DHB estimates can then be thought of as simple weighted combinations of two main components:

\hat{Y}_{syn} = Modelled estimate based on demographic profile and the administrative/census data for each DHB

\hat{Y}_{dir} = Direct estimate based on observations of people in the survey and who live in the DHB

The weighting is based on the measure of 'true' DHB variation. Thus if there is strong DHB variation exhibited (even after accounting for the demographic and/or administrative rate differences) then the direct estimate will get more weight, but otherwise (if there is not much evidence for DHB variation) more weight will go to the modelled estimates.

5 Implementation of the estimation process

Much of this methodology was developed by Saei and Chambers (2003) using the UK Labour Force Survey as the example dataset. This survey is effectively a simple random sample, which the NZHS survey is not. The NZHS has variation in weights and clustering effects which need to be accounted for. To do this we used effective sample sizes in the model fitting process, rather than actual sample sizes. This is not a perfect solution but should ensure we have largely accounted for the design effects of the NZHS. Examination of covariances between cell (age/sex) estimates of various key variables would suggest the assumptions needed for this approximation are not being violated greatly.

The general process for creating these estimates is as follows: (note this is done for each variable of interest separately). More detail discussion of each step is given in Appendix 4.

Step 1: Create a dataset with the DHB estimates and effective samples sizes.

Step 2: Fit the mixed model and calculate the GLMM estimates and their MSE

- We have written a SAS program (based on PROC IML) to fit the mixed model and calculate the MSE of these composite estimators... This gives us the GLMM estimates and an estimate of $\hat{\sigma}_u^2$ the variance of the random effect vector (this term is then used in the estimates of the MSE of the synthetic estimators.)

Step 3: Fit reduced versions of the full model to get the synthetic estimates

- Use the SAS procedure *Proc Surveylogistic* to fit the fixed effect models to get the synthetic estimates. The parameters and covariance matrices are saved and used in the calculation of the MSE's of this estimator

- Another SAS programme is then used to produce estimates of the MSE of these estimates, implementing methodology detailed in Ambler et al. For the estimates of MSE for the synthetic estimator the $\hat{\sigma}_u^2$ term is taken from when the full model was fitted in step 2.

Step 4: Calibrate the modelled estimates so that they are consistent with the direct estimates for the nine groupings of DHBs as published in Portrait of Health (2008). The grouping represents the lowest level of regional data which we felt was reliable to publish directly from the survey data. A simple raking ratio method is used to accomplish the calibration.

Northland/Lakes/Tairāwhiti/Whanganui/Hawkes Bay
Waitemata
Auckland
Counties-Manakau
Waikato
Bay of Plenty/Taranaki/Mid-Central
Wellington/Hutt/Wairarapa
Canterbury
Other South Island DHBs (Nelson-Marlborough, West Coast, South Canterbury, Otago and Southland)

Step 5: The diagnostics outlined in Brown et al are then examined for synthetic and GLMM estimates.

Note that: as a check on the implementation of the random effects approach, explicit composite estimates (and estimates of their MSE) can be produced using the direct estimators, the synthetic estimators and the SE and MSE of the direct and synthetic estimates (from step 1 and step 3). We found these to be quite close to the estimates produced directly from the random effects models.

6 The auxiliary administrative and Census data

One of the key aspects of this method is the use of DHB-level administrative (and census) data to create indicators for various health outcomes at the DHB level.

Mostly we are using data derived from the New Zealand Health Information Service (NZHIS) National Minimum Dataset (Hospital Events) (NMDS). The NMDS is a national collection of public and private hospital discharge information, including clinical information, for inpatients and day patients. Unit record data is collected and stored. All records must have a valid unique National Health Index (NHI) number. Data has been submitted electronically in an agreed format by public hospitals since 1993. The private hospital discharge information for publicly funded events, eg, birth events and geriatric care, has been collected since 1997. Other data is being added as it becomes available electronically.

Another key source of administrative data is the Pharmaceutical Collection (known as Pharms). Pharms contains claim and payment information from pharmacists for subsidised dispensings that have been processed by the HealthPAC General Transaction Processing System (GTPS). As at October 2002, Pharms holds over 270 million claims. Approximately 3.5 million rows of data are added each month.

NZHIS's Primary Health Organisation Enrolment Collection (PHO datamart) can also be used to capture primary health care events such as a "Get Checked" free annual health

check for patients with diabetes. The PHO Enrolment Collection provides a national collection that holds Primary Healthcare System (PHCS) patient enrolment.

Essentially we have combined data from the NDMS, Pharms and the PHO Datamart using the NHI to try to ensure events are not double counted, to produce approximate prevalence rates of some of the key chronic conditions in the NZHS.

We also use the 2006 New Zealand Census. This provides data on the socio-demographic profile of the population in each DHB and also provides data on smoking rates by DHB within different age, gender and ethnic groups.

Table 2
Explanatory variables used in the synthetic estimates

Socio-demographic variables	Rates	Source
<i>Ethnicity</i>	Maori	Census
	Pacific	
	Asian	
<i>Education</i>	No qualifications	Census
<i>Country of birth</i>	Born in NZ	Census
<i>Household size</i>	3 or more adults in household	Census
<i>Socio-economic status</i>	In high NZDEP ¹ meshblock	Census
	Household income less than \$25,000	
<i>Employment</i>	Employed	Census
<i>Social welfare receipt</i>	Unemployment benefit	Census
	Domestic purposes benefit	
	Sickness benefit	
	Invalids benefit	
<i>Urban/Rural living</i>	Living in urban area	Census
Health indicator		
<i>Smoking</i>	Current regular smoker (from census)	Census
<i>Diabetes</i>	Rates based on diagnoses in public hospitals and prescriptions for medicines related to treatment of these conditions found in NZHS datasets.	NDMS/PHARMS
<i>Asthma</i>		NDMS/PHARMS
<i>Chronic obstructive pulmonary disorder</i>		NDMS/PHARMS
<i>High Cholesterol</i>		NDMS/PHARMS
<i>High Blood pressure</i>		NDMS/PHARMS
<i>Cardiovascular disorders</i>		NDMS/PHARMS

A set of rates by DHB by age group and gender, for each of these 21 indicators was created and available for use as auxiliary variables in the synthetic estimates for each DHB. For any particular outcome variable only explanatory variables which showed significant effects when fitted as part of the logistic modelling were included. This was determined separately for each outcome variable of interest.

¹ NZDEP is an index of socio-economic deprivation based on 2006 Census data

7 Results

7.1 Results: DHB level prevalence of Diabetes

These results are based on data from the 2006/07 survey. The first variable we look at is the prevalence of diabetes. Here we have as explanatory variables indicators for each age/sex cell and DHB diabetes rate for this age/sex cell as measured from NZHIS data. Note that factors such as ethnicity are significant factors but do not seem to add much if we already have the administrative indicator in the model.

In terms of the overall fit of the model there does not seem to be a consensus on a standard goodness of fit measure for this situation. The Hosmer and Lemeshow measure partitions the dataset into 10 groups ranked on the model's estimated predicted probability of having diabetes. The difference between the expected rate in the first partition compared to the expected rate in the tenth partition also gives a measure of the ability of the model to discriminate between the diabetes rate in the different cells. In this case the first partition has an expected rate of about 1% and the tenth partition has a rate of nearly 15%, so that the tenth group has nearly 15 times the chance of having diabetes as the first partition. The difference between the observed and predicted numbers in each of these partitions gives us a measure of fit of the model and in this case it is 6.6 (a χ^2 statistic with $df=8$) which suggests a reasonable fit to the basic model. However the validity of this measure as a χ^2 statistic is doubtful. Another simple measure is to look at the amount of variation in the diabetes rates per cell, and measure the proportion of this cell variation that is explained by the model, in other words, a simple 'R square' like goodness of fit measure.

$$cell_level_R^2 = 1 - \frac{\sum_{d,h,b,c} (X_{dir,d,h,b,c} - X_{syn,d,h,b,c})^2}{\sum_{d,h,b,c} (X_{dir,d,h,b,c} - \bar{X}_{dir})^2}$$

In this case this $cell_level_R^2$ measure is .60. So the fit, at the age/sex by DHB cell level is moderate.

So it seems reasonable based on these model-fitting diagnostics that the model will be able to provide, via the synthetic estimates, plausible estimates of the rate of diabetes in each DHB. What is critical to the DHB estimates is the amount of regional variation that remains once age, sex, and the administrative data has been accounted for. When we add DHB random effects to the model and we get DHB estimates as shown in the following table. The extra variation in these estimates compared to the variation in the synthetic estimates is very small (less than 5%) and this suggests the residual regional effects are not very important.

The table of DHB estimates from the different methods and their RMSEs (root mean square errors) is as follows (next page):

Table 3
DHB estimates of prevalence of diabetes

District health Board	Direct	Synthetic	GLMM†
Northland	5.3(1.2)	6.0(0.8)	5.3(0.7)
Waitemata	4.0(0.6)	4.6(0.6)	4.0(0.6)
Auckland	4.9(0.7)	5.2(0.7)	4.9(0.8)
Counties-Manakau	8.2(0.9)	7.9(1.1)	8.2(0.9)
Waikato	5.6(0.7)	4.5(0.6)	5.6(0.7)
Lakes	3.9(0.8)	4.5(0.5)	3.9(0.6)
Bay of Plenty	5.6(1.0)	5.1(0.7)	5.2(0.7)
Tairāwhiti	6.1(1.6)	6.6(0.9)	5.8(0.8)
Taranaki	3.9(1.4)	5.6(0.8)	5.3(0.7)
Hawkes Bay	3.2(0.6)	4.5(0.6)	3.8(0.5)
Whanganui	5.6(2.0)*	4.9(0.7)	4.5(0.7)
MidCentral	4.5(1.0)	3.9(0.6)	4.1(0.6)
Hutt	8.5(1.8)*	5.5(0.8)	6.4(0.8)
Capital and Coast	3.9(0.8)	4.3(0.6)	4.5(0.5)
Wairarapa	2.0(1.1)	4.6(0.8)	4.7(0.6)
Nelson-Marlborough	3.4(1.3)	3.6(0.7)	3.8(0.6)
West Coast	2.7(1.6)*	4.4(0.8)	4.5(0.6)
Canterbury	4.4(0.8)	4.5(0.6)	4.4(0.9)
South Canterbury	5.3(1.9)*	5.8(0.9)	5.9(0.8)
Otago	5.8(1.6)*	4.6(0.7)	4.9(0.6)
Southland	3.3(1.0)	3.5(0.6)	3.6(0.5)

* indicates the RMSE (square root of the mean square error) is greater than the standard error that a direct estimator from a simple random sample of size 200 would have produced.

† GLMM estimates in this table are those which have been through the final calibration stage as outlined in section 5 (step 4).

The reduction in MSE (averaged across the DHBs) is very similar for the synthetic and GLMM estimators.

Goodness of fit diagnostics: These test how close the modelled estimates are to the direct estimates, with the difference weighted by how accurate each estimate is. Essentially where the direct estimates are accurate – we should get modelled estimates which are close to them, and where they are less accurate we will not be as concerned by bigger differences. This test is thus a sensible measure of the overall fit of each model.

Table 4
Goodness of fit diagnostics – diabetes

Estimator	W	$P(\chi^2 > W)$	df
GLMM	9.72	0.98	21
Synthetic	14.40	0.85	21

These values are quite small and not significant (we are looking to see if $(P(\chi^2 > W))$ is less than .05) However one needs to be aware that the W statistic relies on the idea that we are comparing two independent estimates. This is not really the case for the GLMM estimators in some cases and is why we commonly get low values for W for the GLMM estimators. The test does suggest that the synthetic estimates have a good fit to the direct estimates.

Bias diagnostics: The bias diagnostics are based on the idea that the direct estimates are unbiased. So by comparing these to our modelled estimates we can test the extent of bias in them. We do this by doing a simple linear regression between the modelled estimates and direct estimates and test whether the intercept=0 and slope=1.

Table 5
Bias diagnostics - diabetes

tests for intercept=1

Estimator	Slope= β	SE	Prob($\beta = 1$)
GLMM	-1.15	1.05	0.29
Synthetic	-0.51	1.32	0.71

tests for slope=0

Estimator	Intercept= α	SE	Prob ($\alpha = 0$)
GLMM	1.20	0.21	0.34
Synthetic	1.07	0.26	0.81

tests for intercept=0 and slope=0

Estimator	F value	Prob > F	
GLMM	.71	.50	
Synthetic	.28	.76	

Here although the terms for the slope and intercept are quite different from 1 and 0 the tests do not reveal any significant differences. With only 21 estimates we don't have particularly great power with these bias diagnostics in this application.

Coverage diagnostics: These assess how well our measures of RMSE are working, in terms of their ability to help us construct confidence intervals that have the desired coverage properties. We look at the proportion of the confidence intervals (constructed for the direct and for the modelled estimates) which overlap (note these confidence intervals are adjusted so that the expected number of overlapping intervals is .05. In this case we find one non-overlapping interval for the synthetic estimates and one non-overlapping interval for the GLMM estimator. This suggests the MSE calculations are working appropriately. Again we have to acknowledge the fairly limited power we have here with only 21 different intervals to look at.

Calibration diagnostics: These look at how well the modelled estimates add up across the age sex categories to the national level. We are assuming that the direct estimates are

fairly accurate at that level and hence that there should be good correspondence between the aggregated model estimates and direct estimates at the national age/sex level.

Table 6
Calibration diagnostics – diabetes

	Synthetic	GLMM
Males 15-44	-0.05	-0.08
45-64	-0.06	-0.05
64+	-0.09	0.44
Females 15-44	-0.04	-0.11
45-64	-0.06	0.09
64+	-0.08	-0.47

What we see is that the absolute differences in the estimated rates are all quite small.

Conclusion: The MSE's of the synthetic and GLMM estimates of the rates of diabetes are considerably lower than the direct survey estimates. This is mainly because the estimate of the residual regional variation is very small. There is nothing from the analysis of the diagnostics to suggest that either of the sets of estimates are not sensible ones to use.

7.2 Results: DHB level prevalence of adequate vegetable intake

One of the key health measures we look at as a proxy for good nutrition in the NZHS is whether people have an adequate vegetable intake (defined to be 3 or more servings per day). In this case we have no particular administrative data regarding this variable so it is a good example to see how the modelling process might work in this situation.

In this case the age and gender composition of the region and the proportion of the region born in New Zealand are the most significant indicators of different regional rates of adequate vegetable intake. In terms of the overall fit of the model: The first Hosmer and Lemeshow partition has an expected rate of about 41% and the tenth partition has a rate of nearly 81%, so that the tenth group has about 2 times the chance of having adequate vegetable intake as someone in the first partition. So the model has some discriminating power but not as powerful as the diabetes model. The Hosmer and Lemeshow goodness of fit measure is 10.5 (a χ^2 statistic with $df=8$) which suggests an adequate fit to the basic model. The cell_level_ R^2 is .50 only. The residual regional variation is quite large and this means the composite estimator still puts most of its weight on the direct estimates. This means that we get less improvement in the MSE compared to the diabetes example. The table of DHB estimates from the different methods and their RMSEs (root mean square errors) is as follows (next page):

Table 7
DHB estimates of prevalence of adequate vegetable intake

District Health Board	Direct	Synthetic	GLMM†
Northland	67.8(2.4)	69.6(6.2)*	68.8(2.2)
Waitemata	55.7(2.3)	57.5(7.2)*	55.7(2.3)
Auckland	56.1(2.2)	56.6(7.0)*	56.1(2.3)
Counties-Manakau	51.8(2.2)	55.3(7.2)*	51.8(2.2)
Waikato	70.8(1.9)	65.9(6.4)*	70.8(1.8)
Lakes	74.1(2.5)	63.7(6.6)*	72.6(2.4)
Bay of Plenty	68.7(2.1)	68.0(6.2)*	68.3(2.0)
Tairāwhiti	73.2(4.5)*	70.0(6.1)*	73.5(3.8)*
Taranaki	64.6(3.7)*	68.4(6.2)*	64.9(3.1)
Hawkes Bay	67.4(2.8)	69.5(6.0)*	68.1(2.5)
Whanganui	77.9(3.4)*	69.0(6.1)*	75.9(3.0)
MidCentral	62.3(2.7)	67.3(6.4)*	62.5(2.5)
Hutt	57.9(3.4)	66.4(6.4)*	58.6(3.0)
Capital and Coast	64.8(2.7)	65.1(6.6)*	64.4(2.5)
Wairarapa	69.4(4.5)*	76.7(4.8)*	70.2(3.6)*
Nelson-Marlborough	70.7(2.7)	69.5(6.0)*	71.7(2.5)
West Coast	59.0(7.0)*	75.4(5.3)*	64.8(5.1)*
Canterbury	69.4(2.2)	68.0(6.2)*	69.4(2.2)
South Canterbury	83.6(4.4)*	71.4(5.9)*	78.9(3.9)*
Otago	72.6(2.6)	67.2(6.4)*	72.0(2.5)
Southland	72.4(3.4)*	71.1(5.7)*	73.0(3.1)

* indicates the RMSE (square root of the mean square error) is greater than the standard error that a direct estimator from a simple random sample of size 200 would have produced.

† GLMM estimates in this table are those which have been through the final calibration stage as outlined in section 5 (step 4).

The reduction in MSE (averaged across the DHBs) is 11% for the GLMM estimates. So there is only a minor reduction in MSE by going to the random effects models for the vegetable intake variable. This is mainly because the synthetic estimates have larger MSEs than the direct estimates so do not end up contributing heavily to the GLMM estimates.

Table 8
Goodness of fit diagnostics for adequate vegetable intake

Estimator	W	$P(\chi^2 > W)$	df
GLMM	1.77	1.00	21
Synthetic	15.30	0.81	21

As for diabetes all these values are quite small and not significant (we are looking to see if $P(\chi^2 > W)$ is less than .05)

Table 9
Bias diagnostics for prevalence of adequate vegetable intake

tests for slope=1

estimator	Slope= β	SE	Prob($\beta =1$)
GLMM	1.09	0.06	0.13
Synthetic	0.83	0.27	0.53

tests for intercept=0

estimator	Intercept= α	SE	Prob ($\alpha =0$)
GLMM	-6.03	3.75	0.12
Synthetic	11.65	18.54	0.54

tests for intercept=0 and slope=1

Estimator	F value	Prob > F	
GLMM	1.29	.30	
Synthetic	.20	.82	

Again the slopes are quite different from 1 but do not reveal any significant biases.

Coverage diagnostics for vegetable intake: These assess how well our measures of MSE are working, in terms of their ability to help us construct confidence intervals that have the desired coverage properties. In this case we find no non-overlapping interval for the synthetic estimates nor for the GLMM estimator. This suggests the MSE calculations are working appropriately. We construct that the intervals so that they should overlap 95% of the time and so with 21 DHB estimates, so we expect to see 0,1 or 2 non-overlapping intervals).

Table 10
Calibration diagnostics for prevalence of adequate vegetable intake

	Synthetic	GLMM
Males 15-44	0.00	0.11
45-64	0.02	0.62
64+	0.05	1.27
Females 15-44	0.02	-0.97
45-64	0.06	0.09
64+	0.06	0.23

What we see is that the differences in the estimated rates are fairly small with only one greater than one percentage point.

Conclusion: The GLMM estimates of the rates of adequate vegetable intake are very close to the direct survey estimates. This is mainly because while the logistic regression model has some indicators of adequate vegetable intake, there remains considerable unexplained regional variation. A larger number of the DHB estimates remain with an estimated standard error which we would regard as large.

7.3 Results: DHB level prevalence of smoking

One of the key health behaviour measures we look at in the NZHS is whether people are current smokers. In this case we have recent data from the Census for the proportion of the adult population who are current smokers, so we have very closely related auxiliary data. (Towards the end of the section we look at how we can also use the census data to help us assess the validity of our general approach to DHB estimates).

Firstly though we examine the strength of the logistic model for smoking. In this case many of our potential explanatory variables are significant indicators of different rates of smoking; however the census indicator is the most powerful term and is in fact all we need to get a very good fitting model. In terms of the overall fit of the model the Hosmer and Lemeshow is 1.1 (a χ^2 statistic with $df=8$) which suggests a very good fit to the basic model. The first HL partition has an expected rate of about 6% and the tenth partition has a rate of nearly 36%, so that the tenth group has nearly 6 times the chance of being a smoker as someone in the first partition. The cell_level_ R^2 is .64.

In summary, it seems reasonable based on these model-fitting diagnostics that the model will be able to provide, via the synthetic estimates, quite good estimates of the rate of smoking in each DHB. The residual DHB variation is very small ($\hat{\sigma}_u^2 = .009$) and the lambda term in the GLMM estimates averages to be about .3 which means most weight is given to the synthetic estimates

The table of DHB estimates from the different methods and their RMSEs (root mean square errors) is as follows (next page):

Table 11
DHB estimates of prevalence of current smoking

District Health Board	Direct	Synthetic	GLMM†
Northland	22.8(2.2)	24.8(1.5)	24.5(1.7)
Waitemata	15.3(1.5)	16.2(1.1)	15.3(1.6)
Auckland	17.2(1.6)	15.5(1.1)	17.2(1.6)
Counties-Manakau	21.0(1.6)	20.8(1.1)	21.0(1.6)
Waikato	23.6(1.6)	22.1(1.2)	23.6(1.6)
Lakes	29.5(3.2)*	27.9(1.7)	28.3(1.9)
Bay of Plenty	22.7(2.1)	21.5(1.2)	21.0(1.5)
Tairāwhiti	26.7(3.6)*	29.0(2.0)	29.1(2.2)
Taranaki	19.4(2.8)	20.8(1.2)	20.0(1.4)
Hawkes Bay	24.1(2.3)	25.0(1.4)	25.1(1.7)
Whanganui	33.4(3.8)*	26.0(1.6)	27.3(2.0)
MidCentral	19.3(2.3)	22.2(1.2)	20.9(1.5)
Hutt	17.5(1.9)	21.8(1.2)	18.8(1.4)
Capital and Coast	13.7(1.8)	16.7(1.1)	14.5(1.3)
Wairarapa	32.4(5.9)*	23.8(1.5)	22.1(1.9)
Nelson-Marlborough	18.7(2.7)	18.3(1.2)	19.1(1.4)
West Coast	27.2(5.8)*	24.7(1.7)	25.6(2.0)
Canterbury	18.3(1.5)	17.7(1.1)	18.3(1.5)
South Canterbury	18.9(4.1)*	19.6(1.3)	20.2(1.5)
Otago	20.8(3.8)*	18.5(1.1)	19.1(1.4)
Southland	22.0(3.5)*	23.5(1.3)	24.3(1.7)

* indicates the RMSE (square root of the mean square error) is greater than the standard error that a direct estimator from a simple random sample of size 200 would have produced.

† GLMM estimates in this table are those which have been through the final calibration stage as outlined in section 5 (step 4).

The reduction in RMSE (averaged across the DHBs) is 42% for the synthetic and GLMM estimates. So there is quite good improvement in RMSE by going to the synthetic or random effects models for the smoking variable.

Table 12
Goodness of fit diagnostics for smoking:

Estimator	W	$P(\chi^2 > W)$	df
GLMM	7.73	1.00	21
Synthetic	16.18	0.76	21

As for diabetes all these values are quite small and not significant (we are looking to see if $P(\chi^2 > W)$ is less than .05)

Table 13
Bias diagnostics for smoking

:
tests for slope=1

estimator	Slope= β	SE	Prob($\beta =1$)
GLMM	1.09	0.17	0.58
synthetic	1.14	0.19	0.48

tests for intercept=0

estimator	Intercept= α	SE	Prob ($\alpha =0$)
GLMM	-1.55	3.84	0.67
synthetic	-2.56	-4.16	0.55

Test for slope=1 and intercept=0

estimator	F value	Prob > F
GLMM	.38	.69
synthetic	.42	.66

Here the tests do not reveal any significant biases.

Coverage diagnostics for smoking: These assess how well our measures of RMSE are working, in terms of their ability to help us construct confidence intervals that have the desired coverage properties. In this case we find one non-overlapping intervals for the synthetic estimates and no non-overlapping interval for the GLMM estimator. This suggests the MSE calculations are working appropriately.

Table 14
Calibration diagnostics for smoking

	Synthetic	GLMM
Males 15-44	-0.04	-0.36
45-64	-0.07	-0.03
64+	-0.13	0.22
Females 15-44	-0.03	0.17
45-64	-0.05	0.10
64+	-0.10	0.01

The differences in the estimated rates are fairly small with only two marginally greater than one percentage point.

Conclusion: Given that the auxiliary data that we have from the census which is so close to our variable of interest, it is not surprising that models produce estimates with low MSE and are a considerable improvement on the direct survey estimates.

As a slight diversion: with the variable ‘current smoker’ we are in the situation where we have fairly recent census data which matches the survey question quite closely. This

means that we can hold back the census indicator, assume it is the true rate of smoking in each DHB and hence use it instead to examine the success or otherwise of a (reduced) version of the modelling process and measures of quality of the modelled estimates.

If we do not use the census smoking data in the process, and assume that the census data tells us the true smoking rates, observed without sample error, then we can look at the following:

- Create a measure of the observed squared differences between the modelled and/or direct survey estimates compared to the census rates and check if these are close to the measures of RMSE that we have constructed.

Table 15
Comparing estimates of RMSE

Estimator	Estimated RMSE (averaged across DHBs)	$\sqrt{(\text{Estimate}-\text{Census})^2}$ (averaged across DHBs)
Direct	2.8	3.2
Synthetic	2.5	2.1
GLMM	1.9	1.6

So we can see that our RMSE measures give us reasonably close estimates of these independent estimates of RMSE (assuming the smoking data is the true value for each DHB). Again this suggests the RMSE calculations we have are working reasonably well (at least in the average sense across DHBs).

- Check that the measure of residual DHB variation derived from the census is close to that derived from the survey data. By fitting a logistic mixed model to the census data we get $\hat{\sigma}_u^2 = .039$ (SE=.012) and with the survey data we get $\hat{\sigma}_u^2 = .024$ (SE=.015) which are different but unlikely to be significantly different given the size of their respective SEs.
- Check that the residual regional effects fitted using the census data look like they could be random effects from a normal distribution. All the standard tests for normality that Proc Univariate in SAS undertakes indicate no reason to reject this hypothesis.

Table 16
Testing normality of regional effects

Test	Statistic	p Value
Shapiro-Wilk	W 0.948441	Pr < W 0.3182
Kolmogorov-Smirnov	D 0.126124	Pr > D >0.1500
Cramer-von Mises	W-Sq 0.084914	Pr > W-Sq 0.1739
Anderson-Darling	A-Sq 0.500675	Pr > A-Sq 0.1938

All these help re-assure us that (for this variable at least) the synthetic and composite estimates and our measures of their respective MSEs have some validity.

8 Overall Conclusions

The variables diabetes and adequate vegetable intake illustrate the two most common situations we have when trying to produce DHB-level estimates from the NZHS. One where we have quite good synthetic estimates (and hence even better composite estimates) and one where, despite a reasonable logistic model, the synthetic estimates have quite high MSE and where the direct survey estimates need to be relied upon heavily for the DHB estimates. A summary which shows the performance of the models and different estimators across a fuller range of key variables from the study are shown in appendix 4.

In terms of the different estimators examined, the synthetic estimators have been shown to perform worse than the direct estimates for some variables and hence it is not wise to use them in place of the direct survey estimates all the time. In each case we see that the GLMM estimates perform better than either the synthetic or the direct survey estimates alone and hence we recommend that this approach be taken to produce DHB estimates for the 2006/07 NZHS.

As with any modelling exercise, some assumptions need to be made. In this case, the assumption of the randomness of the regional effects is one that is difficult to test given the small number of regions that we have in this study. The smoking data has allowed us to verify (under the assumption that the census data provides estimates of the true smoking rates without any sampling error) that our methods are valid and produce sensible MSE estimates for this variable and that it is valid to assume the regional smoking effects can be treated as random effects.

For more assessments of DHB estimates produced from the 2006/07 NZHS, refer to Appendix 3.

There remain a number of issues for further discussion, investigation and future research:

- The small area diagnostics examined were not particularly powerful given such a small number of actual small areas examined (i.e. only 21 DHBs) and so had limited usefulness in assessing these estimates. It would be good to do some further work to see whether some alternatives could be found that worked better for this situation.
- The methods used are complex and not currently available via standard SAS procedures. The programmes had to be hand crafted using SAS Proc IML. Given the complexity of the underlying calculations there will always be a preference to use more developed 'off the shelf' software products.
- Some assumptions had to be made to cope with the complex survey design of the New Zealand Health survey. The validity of these assumptions could be more thoroughly examined using STATA or MLWIN.
- There is the possibility of linking NZHS health data with some of the administrative data at the individual level. Hence it would be good to further investigate the gain from having this available to help fit the models.
- The standard errors of the estimates of the residual regional variation ($\hat{\sigma}_u^2$) are quite large. This parameter is really quite critical to the whole process so it would be good to do some further work to see whether these could be made more robust, perhaps by pooling data across repeated surveys.

References

- Ambler, R., Caplan, D., Chambers, R., Kovacevic, M. & Wang, S. (2001). Combining unemployment benefits data and LFS data to estimate ILO unemployment for small areas: An application of a modified Fay-Herriot method, *Proceedings of the International Association of Survey Statisticians*, Meeting of the International Statistical Institute, Seoul August 2001.
- Brown, G., Chambers, R. Heady, P. & Heasman, D. (2001). Evaluation of small area estimation methods – an application to unemployment estimates from the UK LFS, *Proceedings of Statistics Canada Symposium 2001*.
- Clark, R and Gerritsen (2006). Sampling the Māori Population in the 2006/2007 New Zealand Health Survey, *Proceedings of Statistics Canada Symposium 2006* (available from Ministry of Health).
- Ministry of Health (2004). *2002/03 New Zealand Health Survey- DHB snapshot datacube help documentation*, New Zealand Ministry of Health
- Ministry of Health (2008). *2006/07 Portrait of Health, descriptive results from the 2006/07 New Zealand Health Survey*, New Zealand Ministry of Health
- Ministry of Health website: www.moh.govt.nz
- Office of National Statistics (2006). *Model-Based Estimates of ILO Unemployment for LAD/UAs in Great Britain – Guide for Users*
- Rao, J.N.K. (2003). *Small Area Estimation*. Wiley
- Saei, A. and Chambers, R (2003). *Small Area Estimation under linear and generalized linear mixed models with time and area effects*, Southampton Statistical Sciences Research Institute, University of Southampton, S³RI Methodology Working Paper M03/15
- Saei, A. and Chambers, R (2003). *Small Area Estimation: A Review of Methods Based on the Application of Mixed Models*, Southampton Statistical Sciences Research Institute, University of Southampton, S³RI Methodology Working Paper M03/16

Appendix 1 Summary of the 2006/07 NZ Health Survey sample design

The sample design and methodology for the 2006/07 New Zealand Health Survey was developed by Dr. Robert Clark of the Centre for Statistical & Survey Methodology, University of Wollongong, New South Wales, Australia.

The mode of data collection is a face-to-face computer-assisted (CAPI) survey, which includes an interview and a short health measurements section for each respondent. The NZHS collects information about the New Zealand civilian population of all ages living in permanent private dwellings.

An area-based frame of Statistics New Zealand's meshblocks was used, based on NZ 2001 Census meshblocks and a 3-step selection process was used to achieve the NZHS sample:

1. Selection of meshblocks

Meshblocks vary considerably in size and were therefore selected by probability proportional to size, i.e., larger meshblocks have an increased chance of selection in the design. Those DHBs with higher concentrations of Maori had a slightly increased chance of meshblock selection. Approximately 1400 meshblocks were selected throughout the country for inclusion in the 2006/07 New Zealand Health Survey.

2. Selection of households within meshblocks

Within each meshblock, some households were selected to form the *core* sample, and some households were selected to form the *screened* or "booster" sample. A minimum of 100 households from each DHB were included in the sample. Households in the core sample were selected by a systematic procedure of beginning at a random point in the meshblock. Households in the screened sample were selected by a similar systematic process and only included if the householder reported that the household contained a person who identifies as Maori, Pacific or Asian ethnicity. Approximately 12,000 households were approached for the core sample and approximately 23,000 households were approached for the screened sample.

3. Selection of respondents within households

One adult (aged 15 years or over), and one child if there were any in the household, of each selected household were randomly chosen to participate in the New Zealand Health Survey. Approximately 17,000 adults and 7000 children were selected to participate in the survey, resulting in approximately 12,500 completed adult interviews and 5,000 completed child interviews (given a 70% response rate).

4. Weighting

Probability weights, reflecting the different stages of selection were calculated and then calibrated using population counts from the 2006 Census broken down by age, gender, District Health Board (DHB) area and ethnic group. Age, gender and ethnic group were included in the calibration weighting because these variables are related to many health conditions, they are related to non-response, and they are a key output classification for the survey. DHB area was included because this is the main geographic classification used in analysing the data. DHB area is also expected to be related to non-response and the variables of interest, although not as strongly as age, gender and ethnic group. Only the population counts from permanent private dwellings were used as the 2006/07 New Zealand Health Survey did not include institutions.

Appendix 2 Description of 'clustered synthetic' methods used for DHB estimates in 2002/03

The following is the description of methods used in 2002/03.

The approach adopted was to cluster DHBs into groups that have similar characteristics (that are strong predictors of health outcomes), on the basis that the individuals with these characteristics would have similar health outcomes.

DHBs were clustered on the basis of the following variables:

- average NZDep2001 decile²
- proportion of population of Māori ethnicity
- proportion of population aged over 65 years
- population density
- prevalence of smoking.

Three distinct clusters were formed on the basis of the clustering variables chosen. These clusters are summarised in Table 17.

Table 17
Clusters of DHBs used in 2002/03 synthetic methods

Cluster	DHBs	Characteristics
1	Waitemata Auckland Capital & Coast Nelson-Malborough Canterbury	Average NZDep2001 decile 4.9. Mostly urban. Low smoking prevalence (compared to the national prevalence). Note that this cluster is comprised of two more-or-less geographically contiguous subgroups: Waitemata and Auckland; Capital & Coast, Nelson-Marlborough and Canterbury.
2	Waikato Taranaki Midcentral Wairarapa Hutt West Coast South Canterbury Otago Southland	Average NZDep2001 decile 5.5. Mostly rural. Note that this cluster comprises two more-or-less geographically contiguous subgroups: Waikato, Taranaki, Midcentral, Wairarapa and Hutt; West Coast, Otago, South Canterbury and Southland.
3	Northland Counties Manakau Bay of Plenty Tairāwhiti Lakes Hawke's Bay Whanganui	Average NZDep2001 decile 6.5. High proportion of Māori. Mixture of urban and rural. Note that this cluster is comprised of two geographically contiguous subgroups: Northland and Counties Manakau; Bay of Plenty, Tairāwhiti, Lakes, Hawke's Bay and Whanganui.

Once the clusters had been established, DHB estimates were obtained by re-weighting each cluster sample as though it were drawn from the cluster member of interest. The re-weighting process was a weighting adjustment using generalised regression to ensure the final weighted totals of eligible adult respondents are consistent with independent population estimates for each DHB. The survey was benchmarked to the 2001 Census population.

For example, to obtain a boosted sample for Hutt DHB, the survey data from cluster 2 would be re-weighted to match the demographic characteristics of the Hutt DHB area.

This method will produce synthetic type estimates (as outlined in section 4) but where the model is fitted separately within each cluster. One advantage of this 'clustered' method (over other synthetic methods) is that members of any one cluster have similar

² NZDep2001 decile 1 is least deprived and decile 10 is most deprived.

characteristics in terms of ethnic and age compositions. Consequently, the resulting weights will not be as variable as it would have been had we tried to re-weight the entire health survey. A disadvantage is that it is very hard to get a stable estimate of the MSE of the resulting estimates, because of the very small number of DHBs in each cluster.

In the analysis of this report we have compared the composite estimates with (unclustered) synthetic estimates. The grouping into 3 clusters is ignored, so that strictly speaking the discussion of the performance of the synthetic estimates does not exactly pertain to the 2002/03 method, but to a simplified version of it.

Appendix 3 Assessment of the DHB estimates for a selection of key adult variables

Table 18
Analysis of DHB composite estimates for key health outcome variables

Variable	1 Unexplained DHB variation= $\hat{\sigma}_u^2$ (SE)	2 Composite Parameter =mean(λ)	3 Mean RMSE: (across DHBs) direct	4 Synthetic (GLM)	5 GLMM	6 Percent Gain in RMSE	7 Largest RMSE: direct	8 GLMM	9 # DHBs where effective <200: direct	10 GLMM
Unmet need for GP	0.254(0.102)	0.8	1.3	2.8	1.2	9	3	2.3	6	4
Adequate vegetable consumption	0.087(0.036)	0.8	3.1	6.2	2.8	11	7	5.1	7	4
Physically active	0.047(0.020)	0.7	3.3	5.4	2.9	12	6.8	4.6	8	5
Hazardous drinking	0.055(0.027)	0.6	2.5	3	2.1	15	5	3.1	5	5
High Cholesterol	0.056(0.028)	0.7	1.5	2.4	1.3	17	3.9	2.4	3	1
Obese	0.024(0.014)	0.5	3	3.3	2.3	24	6.6	3.6	7	2
Seen a dentist	0.018(0.010)	0.5	3.3	3.5	2.5	25	8.9	4	8	1
High blood pressure	0.009(0.010)	0.3	2.1	1.5	1.3	38	4.4	2	6	0
Heart Disease	0.024(0.021)	0.3	1.4	1.1	0.9	40	4.2	1.5	4	0
Diabetes	0.013(0.019)	0.2	1.2	0.7	0.7	42	2	0.9	5	0
Current smoker	0.004(0.007)	0.1	2.8	1.3	1.6	42	5.9	2.2	8	0
Exposed to second hand smoke	0.023(0.024)	0.3	2.1	1.2	1.2	42	4.7	1.5	10	0
Asthma	0.009(0.011)	0.3	2	1.1	1.1	43	4.7	1.4	7	0
Psychological distress	0.011(0.014)	0.2	1.4	0.7	0.8	44	2.8	1	5	0
Stroke	0.028(0.036)	0.2	0.8	0.5	0.4	45	2	0.6	4	0
Adequate fruit consumption	0.000(0.005)	0	2.9	0.9	1.6	46	4.5	2.1	7	0

Notes:

- The first column measures the unexplained DHB variation= $\hat{\sigma}_u^2$, the variance of the random effects term in the mixed model.
- The second column presents the mean composite estimator Parameter = λ . The mean is taken across the DHBs. Where lambda is near 1 a lot of weight is placed on the direct estimates and when it is small more weight is placed on the synthetic estimates.
- The columns (3 to 5) illustrate the sizes of the RMSE as averaged across the DHBs for the direct, synthetic and composite estimators.
- Column 6 gives how much the direct estimators average RMSE has been reduced by moving to the composite estimator
- The columns (7 & 8) illustrate the size of the largest DHB RMSE direct and composite estimators.
- Column 9 shows how many of the direct DHB estimates have an effective sample size of less than 200. hence how many of the DHB estimates are what we have called 'reasonable quality'
- Column 10 gives the same but for the composite DHB estimates.

Appendix 4 Technical details of the estimation process

As we note there are 4 main steps in the process. Some details are given for each step here.

Appendix 4.1 Creating a dataset with DHB estimates by age/cell cell and effective sample sizes

We create a set of DHB by age and sex (direct) estimates and calculate standard errors which fully reflect the sample design and weighting processes used in their production. In the case of the New Zealand health survey this involves using a set of jackknife weights to produce the standard errors. From this we calculate the design effects (DEFF) and hence the effective sample sizes $n_{eff} = n / DEFF$ in each DHB by age by sex cell and the effective sample count

$y_{eff} = round(\hat{R} \cdot n_{eff})$ in terms of the variable of interest (\hat{R} is the weighted estimate of the rate of the variable of interest.)

We then create a dataset where we find y_{eff}^* , the smallest integer in $[y_{eff}, y_{eff} + \partial]$ (in other words close to but larger than the effective sample count) such that $y_{eff}^* \in (\hat{R} \pm \alpha)$. If we can not achieve this we set $y_{eff}^* = y_{eff} + \partial$ and live with a small variation between \hat{R} and

$$r_{eff}^* = y_{eff}^* / n_{eff}.$$

The rationale here is that we want to input into the model fitting process a dataset that reflects original estimates from the survey but also the effective sample sizes from the surveys as close as we can. The original survey has varying design effects and varying weights which means this process is a bit more fiddly than one might have expected. We set $\alpha = .0005$ and $\partial = \max(3, .01 \cdot n_{eff})$ and found in practise there are only very small discrepancies left between \hat{R} and r_{eff}^* .

Appendix 4.2 The generalised linear mixed model estimates and MSEs

Iterative procedure to find REML (residual maximum likelihood) estimates

Assign initial values to β and u and σ_u . In this case we assign β and u vectors with elements .000001 and $\sigma_u = 1$.

The subscript s refers to sample values of the respective quantities and old/new refers to the quantity as they progress through each stage of the iterative fitting process.

1. Calculate $\eta_s = X\beta + Zu$ and $p_s = (1 + \exp(-\eta_s))^{-1}$
2. Update β and u via:

$$\begin{bmatrix} \beta_{new} \\ u_{new} \end{bmatrix} = \begin{bmatrix} \beta_{old} \\ u_{old} \end{bmatrix} + V_s^{-1} \begin{bmatrix} X'_s \\ Z'_s \end{bmatrix} (\partial l_1 / \partial \eta_s |_{\beta_{old}, u_{old}}) - V_s^{-1} \begin{bmatrix} 0 \\ \sigma_{u,old}^{-2} I_{21} \end{bmatrix}$$

where

$$V_s = \begin{bmatrix} X'_s \\ Z'_s \end{bmatrix} (\partial^2 l_1 / \partial \eta_s \partial \eta'_s) + \begin{bmatrix} 0 & 0 \\ 0 & \sigma_{u,old}^{-2} I_{21} \end{bmatrix}$$

and $\partial l_1 / \partial \eta_s$, $\partial^2 l_1 / \partial \eta_s \partial \eta'_s$ are first and second order derivatives of the likelihood l_1 with respect to the linear predictor vector η_s . In the case of binomial response data:

$$l_1 = \text{constant} + \sum_{dhb=1}^{21} \sum_{c=1}^6 [y_{s,dhb,c} n_{dhb,c} - n_{dhb,c} \ln(1 + \exp(\eta_{dhb,c}))]$$

so that

$$\partial l_1 / \partial \eta_s = \underline{y} - \underline{n} \cdot \underline{p}_s \quad \text{and}$$

$$\partial^2 l_1 / \partial \eta_s \partial \eta_s' = \text{diag}[-\underline{n} \cdot \underline{p}_s \cdot (1 - \underline{p}_s)]$$

As discussed in section 5 and appendix 4.1 we replace \underline{n} with \underline{n}_{eff}^* and \underline{y} with \underline{y}_{eff}^* , the vectors of adjusted effective sample sizes and counts.

- Recalculate $\sigma_{u,new}^2 = \frac{1}{N_{dhb}} [tr(T_{22} - u'_{new} \cdot u_{new})]$

- where T_{22} comes from $V_s^{-1} = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix}$

- Repeat steps 1 to 3 until σ_u has converged.

Estimating totals, rates and their MSEs

$$\hat{Y}_{dhb,glmm} = \sum_c^6 \{ (N_{dhb,c} - n_{dhb,c}) \hat{p}_{dhb,c} + y_{dhb,c} \}$$

$$\hat{R}_{dhb,glmm} = \frac{\hat{Y}_{dhb,glmm}}{N_{dhb}}$$

$$MSE(\hat{Y}_{dhb,glmm}) = G_1 + G_2 + 2G_3 + G_5$$

$$MSE(\hat{R}_{dhb,glmm}) = MSE(\hat{Y}_{dhb,glmm}) / N_{dhb}^2$$

where the G terms are as outlined in Saei and Chambers (1), on pages 25 to 27.

The G terms arise from deriving the $MSE(Y_{dhb,glmm})$ in a series of steps where we condition on the model parameters separately, so that

G_1 = MSE of the $Y_{dhb,glmm}$ where we consider β , $\sigma_{c,dhb}^2$ and σ_u to be known.

G_2 = the extra contribution to the MSE of the $Y_{dhb,glmm}$ arising from having to estimate β .

G_3 = the extra contribution to the MSE of the $Y_{dhb,glmm}$ arising from having to estimate the variance terms ($\sigma_{c,dhb}^2$ and σ_u).

G_5 = the extra contribution to the MSE of the $Y_{dhb,glmm}$ arising from calibrating these estimates to the direct estimates for the grouped DHBs

Appendix 4.3 Estimating the synthetic estimates and their MSEs

We derive estimates of $MSE(\hat{Y}_{syn,dhb})$ as outlined in Ambler et al.:

$$MSE(\hat{Y}_{syn,dhb}) = n_{dhb}^2 \sum_{c=1}^6 \sum_{d=1}^6 [w_{dhb,c} \{ \hat{\phi}_{dhb,c} C(\tilde{R}_{dhb,c}, \tilde{R}_{dhb,d}) \hat{\phi}_{dhb,d} \} w_{dhb,d}]$$

where:

$$C(\tilde{R}_{d_{hb,c}}, \tilde{R}_{d_{hb,d}}) = \tilde{R}_{d_{hb,c}}(1 - \tilde{R}_{d_{hb,c}})[\hat{\sigma}_u^2 + x'_{d_{hb,c}}V(\beta)x_{d_{hb,d}}]\tilde{R}_{d_{hb,d}}(1 - \tilde{R}_{d_{hb,d}})$$

and

$$\hat{\phi}_{d_{hb,c}} = n_{d_{hb,c}} / n_{d_{hb}}$$

$$w_{d_{hb,c}} = N_{d_{hb,c}} / n_{d_{hb,c}}$$

$$\tilde{R}_{d_{hb,c}} = \hat{R}_{d_{hb,c}} \left(1 - \frac{1}{2}(1 - \hat{R}_{d_{hb,c}})(1 - 2\hat{R}_{d_{hb,c}})[x'_{d_{hb,c}}V(\beta)x_{d_{hb,c}}] \right)$$

$$\hat{R}_{d_{hb,c}} = (1 + \exp(-x'_{d_{hb,c}}\beta))^{-1}$$

$\hat{\sigma}_u$ comes directly from the above process for the REML GLMM estimates.

PROC SURVEYLOGISTIC is used to find β and $v(\beta)$, using the strata (=DHB) and weight options to take into account that (even after making the design effect adjustments as outlined in section 4d) we still have disproportionate sampling by DHB.

Notes that in the expressions for the MSE of the synthetic estimates and the REML GLMM MSE estimates, a term that accounts for variability in estimated population sizes in each cell is ignored. These terms were calculated and found to be very small so have been left out here to help keep things simpler. In any case, in the NZHS we have available auxiliary estimates of these quantities (derived from census data and other sources) and so they will not be affected by the levels of sampling error that these terms assume.

Appendix 4.4 Small Area Diagnostics

Goodness of fit diagnostics

For any particular set of proposed DHB estimates (R), we compare them against the direct estimates

and construct a χ^2 statistic using the estimates of MSE(R) and the variance of the direct estimates.

$$W(\hat{R}_{d_{hb}}) = \sum_{d_{hb}=1}^{21} \left(\frac{(\hat{R}_{d_{hb}} - \hat{R}_{d_{hb,dir}})^2}{MSE(\hat{R}_{d_{hb}}) + Var(\hat{R}_{d_{hb,dir}})} \right) \sim \chi^2_{21}$$

for $\hat{R}_{d_{hb}} = \hat{R}_{glmm,d_{hb}}$ and $\hat{R}_{syn,d_{hb}}$

Coverage diagnostics

Here we see if the confidence intervals constructed using the estimates of the MSE(R) and the variance of the direct estimates overlap. If we compare 95% confidence intervals of two independent variables the expected level of coverage is too high. Hence we need to adjust the intervals so that instead of $z(\alpha=0.05)$ to construct the confidence intervals we use:

$$z'(\alpha) = z(\alpha) \left(1 + \frac{RMSE(\hat{R}_{d_{hb}})}{SE(\hat{R}_{d_{hb,dir}})} \right)^{-1} \sqrt{1 + \frac{MSE(\hat{R}_{d_{hb}})}{var(\hat{R}_{d_{hb,dir}})}}$$

Hence we see how often these two intervals overlap across the DHBs:

$$CI(\hat{R}_{dir,dhb}) = \hat{R}_{dir,dhb} \pm z'(\alpha) \cdot SE(\hat{R}_{dir,dhb}),$$

$$CI(\hat{R}_{dhb}) = \hat{R}_{dhb} \pm z'(\alpha) \cdot RMSE(\hat{R}_{dhb})$$

for $\hat{R}_{dhb} = \hat{R}_{glmm,dhb}$ and $\hat{R}_{syn,dhb}$

Bias diagnostics

Here we fit a simple linear regression:

$$\hat{R}_{dhb} = \alpha_0 + \alpha_1 \hat{R}_{dhb,dir}$$

and test whether $\alpha_0 = 0$ and test whether $\alpha_1 = 1$.

Calibration diagnostics

Here we examine the difference between:

$$R_{c,glmm} = \frac{\sum_{dhb=1}^{21} \hat{Y}_{dhb,glmm,c}}{\sum_{dhb=1}^{21} N_{dhb,c}} \quad \text{and} \quad R_{c,dir} = \frac{\sum_{dhb=1}^{21} \hat{Y}_{dhb,dir,c}}{\sum_{dhb=1}^{21} N_{dhb,c}}$$

for each age/sex cell=c.