Explanatory Notes

Modelled estimates for small areas based on the 2011-12 Australian Health Survey

Prepared by the Health National Statistics Centre,
Australian Bureau of Statistics

for the
Australian Institute of Health and Welfare

Release: 29 May 2020

LICENCE CONDITIONS

This customised report carries the following licence:

ABS material used 'as supplied'

Provided you have not modified or transformed ABS material in any way, for example by:

- changing the ABS text
- calculating percentage changes
- graphing or charting data
- deriving new statistics from unpublished ABS statistics

Material contained in this customised report may be reused provided one of the following attributions is given:

Source: Australian Bureau of Statistics or
Source: ABS

Derivative material

If you have modified or transformed ABS material, or derived new material from those of the ABS in any way, one of the following attributions must be used:

Based on Australian Bureau of Statistics data or
Based on ABS data

Citing customised reports

If you are required to cite material from this report please be guided by the examples below.

In-text and reference list/bibliography

- In-text:
  (ABS 2020)
- In reference list/bibliography:
  ABS 2020, Customised report.

In-text only

- (ABS, Customised report, 2020)

CONTENTS

1 Introduction

The Australian Institute of Health and Welfare (AIHW) requested the Australian Bureau of Statistics (ABS) to provide modelled estimates of characteristics associated with health at a small area level for the Australian population. To meet this request, and by mutual agreement between ABS and AIHW, the ABS has provided modelled estimates for public release based on the Australian Health Survey, 2011-12 (AHS). These explanatory notes accompany the modelled estimates for small areas, provided as Excel worksheets (Datacube), and describe the methodology used to produce them, as well as how to use them.

2 Methodology used

A modelled estimate can be interpreted as the expected prevalence of a health condition for an area in Australia based on the demographic information available for that area. The process of producing modelled small area estimates for health conditions measured in the AHS consisted of the following components, described in detail in sections 2.1 to 2.9:

1. Identification of the outcome variables

2. Identification of the geographical areas

3. Selection of the predictor variables

4. Scoping the data

5. Creation of binary and proportion variables

6. Aggregating observations and merging datasets

7. Model selection

8. Age standardisation

9. Assessment of the modelled estimates

The process for this AIHW request was simplified to take advantage of what was already produced as part of previous requests for the Public Health Information Development Unit (PHIDU) and AIHW.

2.1 Identification of the outcome variables

From the AHS, modelled small area estimates (counts, proportions, measure of error) have been produced for persons with chronic kidney disease. The outcome variables are:

- chronic kidney disease, by Population Health Area (PHA) by age group
- chronic kidney disease, by Primary Health Network (PHN) by sex by age group

For age groups:

- 18-54 years
- 55-74 years
- 75 years and over

Sex is defined as the following:

- Males
- Females

For more information about chronic kidney disease, including the definition, see the comments in the Datacube.

A range of modelled small area estimates on chronic kidney disease have previously been provided to AIHW based on the AHS data. The modelled small area estimates in this consultancy use the same definition of chronic kidney disease as the previous product, however this consultancy provides data for a more extensive range of age and sex breakdowns than previously provided. Also, in this consultancy direct age-standardised estimates have been produced, whereas indirect age-standardised estimates were provided in the previous consultancy. The direct age-standardised estimates are discussed separately in section 2.8.

The data from this consultancy should be used in preference to the previously supplied data.

2.2 Identification of the geographical areas

The modelled estimates for small areas have been produced at the PHA and PHN levels for each jurisdiction, with the exception of:

- areas classified as very remote
- areas classified as discrete Aboriginal and Torres Strait Islander communities
- areas that had an adjusted 2011-12 ERP of zero residents, or a 2011 Census population of zero residents

PHAs were developed by PHIDU in consultation with state and territory health agencies and are comprised of a combination of whole SA2s (42.8% of PHAs) and aggregates of SA2s with relatively small populations. For further information, refer to the Population Health Areas: Overview in PHIDU's website .

The PHN website states that "PHNs have been established with the key objectives of increasing the efficiency and effectiveness of medical services for patients, particularly those at risk of poor health outcomes, and improving coordination of care to ensure patients receive the right care in the right place at the right time". Modelled estimates for the PHNs of Christmas Island, Cocos (Keeling) Islands, Jervis Bay and Norfolk Island are excluded due to insufficient availability of data.

2.3 Selection of the predictor variables

In order to predict outcome variables, predictor variables are required on both the AHS dataset and a small area dataset containing population, Census, and administrative data. Predictor variables were created if data were available for small areas for all of urban, rural, and remote Australia and if there was an expectation that they might be good predictors of the outcome variables.

For age and sex predictor variables, data at the small area level were obtained from ABS ERP data from Regional Population Growth, Australia, 2011-12 (Cat. No. 3218.0). This is described below in section 2.4.

For other demographic variables collected in the AHS, data at the small area level were obtained from the 2011 Census of Population and Housing, as this was the most up-to-date comprehensive source of demographic data due to the depth of information at small geographical levels.

Additional variables that were available at the small area but not collected in the AHS were also included in the model. These variables included other demographic variables on the Census, geographic variables, and variables from administrative sources.

Predictor variables that relate to the geographical areas where people reside included:

- remoteness area
- socio-economic indexes for areas (SEIFAs) – population-weighted deciles at the Statistical Area Level 1 (SA1) level
- state and territory
- section of state (major urban/other urban/bounded locality/rural balance)
- Greater Capital City Statistical Area (GCCSA)/balance of state
- design area type (categorises inner city, large and small urban towns, rural towns and remote areas within states and territories for designing the sample of the AHS)

Sources of geographical area data included:

- Australian Statistical Geography Standard (ASGS): Volume 5 - Remoteness Structure, July 2011 (Cat. No. 1270.0.55.005)
- Census of Population and Housing: Socio-Economic Indexes for Areas (SEIFA), Australia, 2011 (Cat. No. 2033.0.55.001)
- Australian Statistical Geography Standard (ASGS): Volume 4 - Significant Urban Areas, Urban Centres and Localities, Section of State, July 2011 (Cat. No. 1270.0.55.004)
- Australian Statistical Geography Standard (ASGS): Volume 1 - Main Structure and Greater Capital City Statistical Areas, July 2011 (Cat. No. 1270.0.55.001)

Predictor variables from administrative data sources included:

- recipients of age pensions from the ABS National Regional Profile (NRP), 2007-2011 (Cat. No. 1379.0.55.001)
- recipients of disability support pensions from the ABS National Regional Profile (NRP), 2007-2011 (Cat. No. 1379.0.55.001)

Within most types of predictor variables (as discussed above), several separate categories or data items were included. The variables considered for inclusion in the model are listed in the *Predictor Variables* tab of the Datacube.

2.4 Scoping the data

The modelled estimates for small areas are applicable to persons who were usual residents of private dwellings to match the scope of the AHS. They exclude:

- non-private dwellings, for example hospitals and aged care facilities
- areas classified as very remote
- areas classified as discrete Aboriginal and Torres Strait Islander communities

The base data source used to compile the modelled small area estimates was the ABS Estimated Resident Population (ERP) data from Regional Population Growth, Australia, 2011-12 (Cat. No. 3218.0). Adjustments were made to the ERP data, by using ratios of private to non-private dwellings, calculated from the 2011 Census to match the scope of the AHS, and then summed to the AHS population state by age by sex estimates. These are the 'population denominator' estimates included in the Datacube. It is important to note that these population estimates are not official estimates and were created solely for analysis of the AHS modelled small area estimates and will not match other population data at the PHA and PHN geography levels.

Adjustments were also made to the Census data, specifically the predictor variables obtained from the Census to match the scope of the AHS. Persons residing in non-private dwellings were easily removed from the small area dataset using persons' dwelling type available on the Census datasets for respondents at home on Census night. However, for persons who were not at home on Census night, information is not collected to determine if the dwelling they usually reside in is a private or non-private dwelling; therefore, their records were deleted from the small area dataset. This data adjustment assumes that the people who were away from home on Census night and live in private dwellings have the same health characteristics as the people who were at home in a private dwelling.

To further match AHS scope, removal of very remote areas and discrete Aboriginal and Torres Strait Islander communities from the ERP and Census datasets was approximately done by deleting persons residing in SA2s that had more than 50% of their population in SA1s classified as very remote or in discrete Aboriginal and Torres Strait Islander communities. SA2s were chosen as the geography to exclude because this was the process undertaken as part of the PHIDU request which was the basis for this AIHW request. Additional exclusions that were applied to the data included:

- small area locations (SA2s) with zero residents in the 2011 Census
- small area locations (SA2s) with an adjusted ERP of zero residents
- residents of Other Territories

- foreign diplomatic personnel and their families were excluded from the modelled estimates because they are not included in Australia's ERP, the Census or the AHS

See the *SA2s Excluded* tab within the Datacube for the full list of SA2s not included in the modelled estimates.

While out of scope for the AHS, members of non-Australian defence forces (and their dependents) stationed in Australia were unable to be removed from the modelled estimates because they could not be identified in Australia's 2011-12 ERP.

2.5 Creation of binary and proportion variables

On the AHS dataset outcome variables were created as binary variables to make them suitable for the type of modelling undertaken (logistic regression). On both the AHS and the small area datasets, predictor variables that were categorical were also created as binary variables. An observation took the value of 1 if an individual had a characteristic of interest and 0 otherwise. For example:

- in the case of chronic kidney disease, the outcome variable for chronic kidney disease took the value of 1 if an individual had chronic kidney disease and 0 if the individual did not have chronic kidney disease,
- in the case of labour force status, the predictor variable for employed took the value of 1 if an individual was employed and 0 if the individual was unemployed, not in the labour force or aged 0-14 years.

Variables on the small area dataset sourced from administrative data were converted to proportions of their areas' population with the characteristic of interest. For example, a person can live in an area with a proportion of its population receiving a disability support pension.

2.6 Aggregating observations and merging datasets

All data sources were aggregated to a fixed structure (cross classification cell groups) including several levels of geography, five year age group and sex. This decreases the size of the datasets (especially the Census dataset) to increase the efficiency of the modelling process.

The Census, adjusted ERP and administrative datasets were then merged into one small area dataset.

2.7 Model selection

A statistical model was created for the chronic kidney disease outcome variable. However within this outcome variable the same model is used for each output classification, for example geography, age group, and sex.

The model selection method uses the prepared dataset to measure the relationship between the outcome variable and possible predictor variables to determine one set of significant predictor variables. This method assumes that the relationships observed in the survey data at State and National levels also hold at the small area level. The significant predictor variables for the model are listed in the *Predictor Variables* tab of the Datacube.

A random effects logistic regression models is used for the outcome variable. As part of any model selection process an appropriate significance level must be chosen for determining

which predictor variables to include in the models. The 0.05 (95%) level is most commonly used; however, due to AHS' relatively large sample sizes, the Bayesian Information Criterion (BIC) was used to reduce the risk of over-fitting.

To verify that the model adequately predicted the outcome variable, the model is applied to small area data, summed to create National level modelled estimates and compared with reliable direct survey weighted estimates. This property is known as model additivity. Where model additivity was not similar, additional predictor variables were included in the model until suitable model additivity was achieved.

Using the selected model, a mixed estimate comprised of modelled and survey data is then produced for each small area output classification (PHA by age group and PHN by sex by age group). A mixed/composite estimate reflects the best trade-off between the accuracy of the direct survey weighted estimate and the error associated with the modelled estimate. For a small area that happens to have a low sampling error (because of a large sample size within that small area, for example), more weight will be given to the direct estimate when calculating a modelled estimate for that small area. On the other hand, for a small area with high sampling error, more weight will be given to the model-based prediction as this will be more reliable in calculating the modelled estimate for that small area. This takes advantage of what is known about the small area location from the survey to improve the modelled estimates.

The modelled estimates are then adjusted so that they sum to published state direct survey estimates. For Tasmania, NT and ACT, where one PHN encompasses the whole state or territory, the model estimates were adjusted to equal the state/territory direct survey estimates. The associated errors resulting from the modelling process, which improve on direct survey estimates' errors, were not adjusted; and are used for reporting the model error for Tasmania, NT and ACT.

This is a different adjustment than was applied to the modelled small area estimates previously provided to AIHW, where modelled estimates for each age and sex group had different adjustments. The adjustment process undertaken in this consultancy ensures that small area estimates across all tables of the Datacube within the same state will always add up to the state level direct survey estimate.

In the case of PHN by sex and PHN by age group estimates, this means that the modelled small area estimates in this consultancy will differ slightly to those data previously provided. The differences are generally very minor and are solely due to the different adjustment applied. The data from this consultancy should be used in preference to the previously supplied data in these circumstances.

The PHN for Murray includes SA2s in New South Wales (NSW) and Victoria, so the modelled estimates for PHNs for both of these states were summed to totals of NSW plus Victoria.

The modelled estimates in the Datacube are in the form of counts (number of persons) and their relative error for each small area location. Prevalence proportions (percentage of population at risk in each small area) have also been calculated. The denominators used in the calculation of proportions at risk were the unofficial population estimates for each PHA and PHN (based on adjusted ERP) described above in section 2.4. Small area estimates in this

data consultancy are not published for small areas (PHAs) that do not meet ABS confidentiality requirements. These cells are indicated by an 'np' (not publishable) in the Excel spreadsheets.

2.8 Age standardisation

In addition to the output defined for crude rates described in Section 2.1, age standardised rates of chronic kidney disease are also provided for the following output categories:

- PHA
- PHN
- PHN, by sex

The national age standardised rates, including by sex, are also provided in the spreadsheet 'National age standardised rates'.

A direct method of age standardisation was used for this AIHW request. The direct Age Standardised Rate (ASR) is:

$$ASR = \sum_i \frac{r_i N_i}{N}$$

where:

r_i = modelled rate for each PHA or PHN (as discussed above) for the outcome variable in age group i
N_i = Standard population in age group i
N = Standard population

The standard population is from the Australian Estimated Resident Population at 30 June 2001 .

Age groups i are as follows:

18-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75+

In addition, a Standard Rate Ratio (SRR) was calculated, which presents the ASR for an area relative to the national age standardised rate:

$$SRR_{area} = \frac{ASR_{area}}{ASR_{national}} \times 100$$

The ASR and SRR, and their relative errors, are provided for each PHA and PHN subject to data for that area meeting the following criteria:

1. The errors for modelled estimates in each age group i are acceptably low across all age groups; and

2. The trend of modelled estimates across each age group i is similar to the national level trend across each age group i. This criterion is only applied to PHAs and PHNs with a substantially different age structure.

For each outcome variable, there are a small number of PHAs which did not meet these criteria, therefore the ASR and SSR are marked as 'np' (not publishable) in the Excel spreadsheets. Data for all PHNs meet these criteria for all outcome variables.

2.9 Assessment of the modelled estimates

Various measures were taken to quality assure the modelled estimates. Modelled estimates were compared with direct survey estimates from the AHS for areas that were sampled. For the survey estimates, 95% Confidence Intervals (CIs) were calculated. These were plotted against the modelled estimates to see if the majority of modelled rates fell within the CIs of the NHS estimates.

Comparisons among the small area estimates and choropleth maps were produced to assess whether the modelled estimates aligned with expectations.

Please see section 5 for further quality assurance practices performed on the small area estimates, including a quality summary for the modelled small area estimates.

3 Accuracy of results

The process undertaken in producing modelled estimates overcomes much of the volatility at the PHA and PHN levels caused by sampling error. However, it should be remembered that the modelled estimates produced are still subject to errors.

The errors associated with the modelled small area estimates fall into four categories, as follows:

1. sampling error
2. non-sampling error
3. modelling error
4. prediction error

These errors are combined into an overall measure of accuracy, the relative root mean squared error (RRMSE), described in section 3.5.

3.1 Sampling Error

Sampling error is introduced into estimates because the AHS data were collected from only a sample of dwellings. Therefore, they are subject to sampling variability; that is, modelled estimates may differ from those that would have been produced if all dwellings had been included in AHS. Furthermore, the smaller the sample obtained within a small area, the greater the sampling error associated with that small area's modelled estimates will be.

3.2 Non-Sampling Error

The imprecision due to sampling error should not be confused with inaccuracies that may occur because of imperfections in reporting by respondents and recording by interviewers, and errors made in coding and processing data. Inaccuracies of this kind are referred to as

non-sampling error, and they occur in any enumeration, whether it be a full count (Census) or a sample. Unlike the other sources of error, non-sampling error is not measurable and therefore isn't accounted for in the measured error (direct or modelled) that accompanies ABS estimates. Every effort is made to reduce non-sampling error to a minimum by careful design of questionnaires, intensive training and supervision of interviewers, and rigorous procedures.

3.3 Modelling Error

Modelling error is introduced by model misspecification. This can occur when the choice of model is incorrect, a key predictor variable is left out or an inappropriate predictor variable is included. Therefore, the variables chosen in the models may result in incorrect modelled estimates for certain small areas, particularly those unusually small areas that do not follow the typical associations between the available predictor variables and the health conditions being modelled. The models that have been chosen have been tested against a range of possible alternative models; however, they are only the most preferred models subject to available data at the time.

3.4 Prediction Error

A strong model does not guarantee statistically accurate modelled estimates. Prediction error is a measure of the statistical accuracy of the predictions made to produce the modelled small area estimates.

3.5 Relative Root Mean Squared Error (RRMSE)

A measure of the quality of the modelled estimates is the RRMSE. The RRMSE is primarily a measure of prediction error but in its calculation it also inherits some aspects of modelling and sampling error. The RRMSE generally decreases as the population size increases, and is used to assess the reliability of modelled estimates.

As a general rule of thumb, estimates with RRMSEs less than 25% are considered reliable for most purposes, estimates with RRMSEs between 25% and 50% should be used with caution and estimates with RRMSEs greater than 50% are considered too unreliable for general use.

4 Using modelled estimates

The small area modelled estimates can be interpreted as the expected prevalence of a health condition for a typical area in Australia with the same characteristics. For some small area location (PHAs and PHNs), there will be differences between the modelled estimates and the actual number of people with the characteristic of interest. One explanation for this is that significant local information about particular small areas exists but has not been collected for all areas and cannot be incorporated into the models. This sort of information is usually not measurable, and relies on local or expert knowledge.

Small area modelled estimates should be viewed as a tool that when used in conjunction with local area knowledge as well as the consideration of the modelled estimates reliability, can provide useful information that can assist in making decisions for small geographic areas. Care needs to be taken to ensure decisions are not based on inaccurate estimates.

The modelled small area estimates crude rates can be aggregated to larger regions (such as regional planning regions) to help improve decision making, using an approximation formula

outlined in section 6. Aggregation of small areas should be done taking into account local knowledge about these areas.

Crude rates should be used for analysis of the prevalence of chronic kidney disease within and between small areas at the time of the survey, and for comparing rates within a small area over time. Direct age standardised rates on the other hand should be used for analysis of the relative rates between areas, under the assumption that each small area has the same standard age structure. Direct age standardised rates can also be used to compare relative rates over time, assuming the age structure does not change over time. The age standardised rate alone for an individual area is of limited use – instead the Standardised Rate Ratio is the most appropriate statistic to use, as this compares the age standardised rate for an area to the national age standardised rate.

5 Quality summary for modelled estimates

The quality of the modelled estimates were assessed according to the following criteria:

1. median RRMSE, as a measure of prediction accuracy
2. consistency with national direct survey estimates. For example, whether modelled estimates for chronic kidney disease increased proportionally with age
3. the number, range, and applicability of predictor variables included in the models

These culminated in an overall reliability assessment, which has three categories:

- reliable, meaning the modelled estimates are suitable for general use
- less reliable, meaning the modelled estimates should be used with caution
- unreliable, meaning the modelled estimates are unsuitable for general use

Reliability assessment table: PHA estimates

Outcome variable	Median RRMSE	Consistency with National data	Number and range of predictor variables	Overall reliability estimate
Chronic kidney disease, 18 years and over, by age group	6.9%	Reliable	Reliable	Reliable
Chronic kidney disease, 18 years and over	4.2%	Reliable	Reliable	Reliable

Reliability assessment table: PHN estimates

Outcome variable	Median RRMSE	Consistency with National data	Number and range of predictor variables	Overall reliability estimate
Chronic kidney disease, 18 years and over, by age group by sex	5.1%	Reliable	Reliable	Reliable
Chronic kidney disease, 18 years and over, by sex	3.2%	Reliable	Reliable	Reliable
Chronic kidney disease, 18 years and over, by age group	4.5%	Reliable	Reliable	Reliable
Chronic kidney disease, 18 years and over	3.0%	Reliable	Reliable	Reliable

6 Estimating aggregated areas

The following formulas describe the estimation of aggregated areas. This may be done for one of two reasons:

1. Estimates are required for a bespoke small area of interest
2. Where the error (RRMSE) for an area is unacceptably high aggregating areas can decrease the error

Note that the error formula is an approximation only, and that these should only be used where alternative modelled estimates are not available. Aggregation of the modelled small area estimates to large geographies such as capital city or state/territory level is not recommended. If you require capital city or state/territory level data for the characteristics of health conditions provided here at small area level, then use of AHS published data (or use of the TableBuilder product) is recommended.

The following formula is used to estimate the count for an aggregated area.

$$Count_{aggregated\ area} = \sum_{SA2} Count_{SA2}$$

The following formula may be used to approximate the RRMSE for an aggregated area.

$$RRMSE_{aggregated\ area} = \frac{\sqrt{\sum_{SA2}(Count_{SA2}{}^2 \times RRMSE_{SA2}{}^2)}}{Count_{aggregated\ area}}$$