

# BMJ Open Impact of socioeconomic and risk factors on cardiovascular disease and type II diabetes in Australia: comparison of results from longitudinal and cross-sectional designs

Jinjing Li,<sup>1</sup> Yohannes Kinfu<sup>2,3</sup>

**To cite:** Li J, Kinfu Y. Impact of socioeconomic and risk factors on cardiovascular disease and type II diabetes in Australia: comparison of results from longitudinal and cross-sectional designs. *BMJ Open* 2016;**6**:e010215. doi:10.1136/bmjopen-2015-010215

► Prepublication history for this paper is available online. To view these files please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2015-010215>).

Received 13 October 2015

Revised 14 February 2016

Accepted 2 March 2016



CrossMark

For numbered affiliations see end of article.

## Correspondence to

Dr Yohannes Kinfu;  
yohannes.kinfu@canberra.edu.au

## ABSTRACT

**Objective:** Existing large-scale studies do not take into account comorbidity or control for selection and endogeneity biases. This study addresses these shortcomings.

**Participants:** We use information on individuals aged between 35 and 70 years from a nationally representative longitudinal survey conducted in Australia between 2001 and 2013. Participants were approached annually, and updates on their characteristics, including health status, were ascertained through self-reporting.

**Method:** We develop three different analytical designs. The first model is a cross-sectional analysis against which our improved models are compared. In the second model, we follow the same approach but control for prior health conditions. The last preferred model additionally adjusts for characteristics and risk profile of respondents prior to onset of conditions. It also allows for comorbidity and controls for selection bias.

**Results:** Once comorbidity and changes over time in the participant's characteristics are controlled for, body mass index (BMI), alcohol consumption and physical activity exhibit a stronger impact than in the models without these controls. A unit increase in BMI increases the risk of developing a cardiovascular disease (CVD) condition within 2 years by 1.3 percentage points ( $\beta=0.11$ , 95% CI 0.05 to 0.16) and regular alcohol intake increases the risk of CVD by 3.0 percentage points ( $\beta=0.24$ , 95% CI 0.09 to 0.39). Both factors lose significance without proper control for endogenous behavioural change. We also note that frequent physical activity reduces the risk of developing diabetes by 0.9 percentage point ( $\beta=-0.40$ , 95% CI  $-0.72$  to  $-0.07$ ).

**Conclusions:** Our result shows a greater importance of certain lifestyle and risk factors than was previously suggested.

## BACKGROUND

Cardiovascular disease (CVD) and type II diabetes are currently the leading causes of

## Strengths and limitations of this study

- The study reports results for diabetes and cardiovascular conditions, which are major causes of illness and premature mortality in Australia and across the world.
- It employs quasi-longitudinal observational data to examine issues of reverse causation and selection biases that were often overlooked in previous studies investigating the effects of socioeconomic status and risk factors on non-communicable conditions.
- Unlike most other previous analyses that focus on each health condition separately, we take into account comorbidity in our models.
- The outcome variables on health status are self-reported data and not those based on biometric indicators or a doctor's examination reports.

disability and premature mortality in the world and are expected to remain major public health challenges in the years ahead.<sup>1 2</sup> Global estimates suggest that approximately 17 million people have died of CVD in the world in 2013, a figure which is about 41% larger than the estimate for 1990 and represents about a fifth of total global deaths estimated for 2013.<sup>3</sup> Although Australia has made substantial progress in cardiovascular health in recent years, the disease continues to cause a heavy burden on the country, with currently one in six Australians suffering from at least one cardiovascular condition.<sup>4</sup> Each year, CVD also claims more than 50 000 lives, which is higher than for any other disease category.<sup>4</sup> A similar trend is observed with respect to type II diabetes, whose rates of prevalence in the country have quadrupled since the 1980s.<sup>5 6</sup>

The huge burden of CVD and type II diabetes and the need to address their impact

on people's health have generated worldwide research interest in identifying the factors associated with the disorders.<sup>7 8</sup> Many of these studies draw on prospective data, often but not always, collected in clinical or controlled settings.<sup>9–11</sup> Follow-up studies monitor disease progression, risk profile and characteristics of patients on an ongoing basis, which provide unique opportunities for understanding the relationship between one's health status and these time-varying indicators. However, such studies are expensive to carry out, as they require substantial resources and a complex research infrastructure that are in short supply. The fact that they are often localised and based on a relatively small sample size means that generalisation beyond the immediate reference population has always been a concern.<sup>12 13</sup> Few among these studies also track progression of multiple disorders, which means that comorbidity can sometimes be ignored despite its importance.<sup>14 15</sup>

Cross-sectional population surveys are therefore sometimes used as the primary data source as they have generally a larger sample size, cover more health disorders and are relatively less expensive to conduct.<sup>16–18</sup> However, a number of limitations are associated with cross-sectional study designs.<sup>19</sup> First, the population that is identified in surveys as having experienced the condition of interest—and used for estimating prevalence rates—represents only the survivors of those who experienced the event before the survey, resulting in what is generally known as selection bias. Second, since some respondents are likely to experience the event after the arbitrary date of the survey, the same population also represents only a subset of those who will eventually succumb to the disease in their lifetime. Third, the absence of information on disease progression and characteristics of respondents before and after diagnosis means that reported health conditions can only be linked with risk profiles and characteristics reported at the time of the survey rather than those that prevailed at the time of experiencing the event itself. This, in turn, requires a strong assumption on how well contemporary variables can approximate the characteristics of the individual at the time of the occurrence of the event of interest. For example, some characteristics of respondents—such as alcohol consumption, smoking, physical activity and labour force status—may even be shaped by a respondent's health condition rather than the other way round, further creating what is known as endogeneity bias in such analyses.

Ideally, a study on the effects of respondents' characteristics on health conditions would require a large representative sample of healthy respondents clinically checked at regular intervals for a range of health conditions and a record of their characteristics and risk profile during each of these encounters. In Australia, the AusDiab survey<sup>20</sup> and the 45 Up study<sup>21</sup> provide data that are close to the ideal situation. However, the AusDiab survey is conducted at intervals of 5 years, potentially missing out on important lifestyle events and

characteristic changes that happen in between the follow-up surveys. On the other hand, the 45 UP study is confined to a single jurisdiction, New South Wales, of the eight in the country, which makes it unrepresentative of the nationwide situation. Most other studies in the country, as a result, use data from national health surveys, but such data while being statistically representative remain cross-sectional in nature and hence subject to the biases discussed earlier.

Given the limitations of existing studies, this paper therefore aims to revisit the link between risk factors and the development of CVD and type II diabetes by making use of nationally representative longitudinal data from the Australian Household, Income and Labour Dynamics survey (HILDA). Most importantly, the paper employs a range of statistical designs to these data to address some of the key shortcomings associated with studies based on large-scale cross-sectional surveys or follow-up data collected in controlled settings.

## METHODS

### Data source

This study uses the HILDA, which is a nationally representative longitudinal survey that has been carried out in Australia since the early 2000s. The unique nature of the HILDA survey emanates from the fact that it combines the advantages of both a follow-up study and population-level surveys and contains information on respondents' characteristics, as well as a range of health topics and associated risk profiles collected on an annual basis. For example, just as with population-level surveys, the data are nationally representative, cover a relatively large sample and have information on multiple health conditions. Similarly, just like in clinically based follow-up studies, patients are observed at regular intervals, providing an opportunity to examine a range of time-varying characteristics of respondents, including their health conditions.

Although HILDA occupies a unique position in the Australian data collection landscape, its design resembles other large-scale household panel surveys elsewhere in the developed world, such as the German Socio-Economic Panel (GSOEP) Survey and the British Household Panel Survey (BHPS).<sup>22 23</sup> Traditionally, the survey has been primarily used for economic research, but it is now increasingly also used for demographic and health-related studies.<sup>24–27</sup> The initial sample of households in the HILDA survey was drawn in 2001, using a multistage sampling technique.<sup>23</sup> It included all members of households where at least one person provided an interview in the initial wave. These individuals were then followed annually. In addition, the sample was gradually extended to newly arrived migrants as well as new household members resulting from changes in the composition of the original households.

With respect to content, HILDA covers extensive information such as age and sex of respondents; living

arrangement, family background and relationships; educational attainment and occupation; income and expenditure patterns as well as health status and health service utilisation. Relevant health information obtained by way of self-reporting was included in the HILDA survey in a systematic fashion for the first time in wave 7, and repeated again in wave 9. The data from these waves allow us to identify respondents with type II diabetes and three types of cardiovascular and circulatory-related conditions, namely high-blood pressure, coronary heart disease and other cardiovascular and circulatory conditions. In the analysis, we combine all the responses associated with cardiovascular and heart conditions into a single category known as CVD. We also restrict the analysis to the population aged 35–70 years, given that risks associated with these conditions began to increase rapidly among the Australian population from around age 35 years.<sup>28</sup> We cap the age at 70 years both because of differences in general health conditions in old age and the limited number of observations. The final data set for replicating the cross-sectional study design constitutes all respondents aged 35–70 years, while only those who were free from CVD and type II diabetes in the earlier wave were used in the other models.

### Analytical approach

The availability of extensive information on respondents' health status and other individual characteristics at different time points in HILDA makes it possible to compare results from longitudinal and cross-sectional designs and control for comorbidity and time-varying characteristics of respondents. Suppose  $d_{i,t}^j$  represents the presence of health condition  $j$  in individual  $i$  at time  $t$ , a basic cross-sectional model for the prevalence of disease  $j$  can be expressed as follows:

$$\Pr(d_{i,t}^j = 1) = \Phi(X_{i,t}\beta^j + u_i^j) \quad (1)$$

where  $X$  is a vector that represents respondents' socio-economic status, lifestyle conditions and other key risk factors and  $\beta^j$  is the coefficient vector for disease  $j$ .  $u_i^j$  represents the error term that is assumed to follow a standard normal distribution for all  $J$ .  $d_{i,t}^j$  is a binary outcome variable and  $\Phi(\cdot)$  denotes the standard normal cumulative distribution function.

The model above, however, does not take into account the timing of events (ie, the time a respondent developed the condition) or their characteristics and risk profile before or at the time of being afflicted with the disease, which makes interpretation of observed relationships less straightforward. Given the availability of data at multiple time points in the HILDA survey, we can correct for the bias by restricting the analysis only to individuals who reported being healthy in the early wave of the survey. This design thus allows us to analyse the association between disease progression and risk factors that is not usually possible in conventional cross-sectional population-level surveys. Algebraically, such a

model that corrects for prior health status can be described as,

$$\Pr(d_{i,t}^j = 1 | d_{i,t-1}^j = 0) = \Phi(X_{i,t}\beta^j + u_i^j) \quad (2)$$

While the restriction on prior health condition represents an improvement over the conventional cross-sectional design and allows us to have a clearer interpretation of the role of observed lifestyle and risk profile of respondents on the diseases of interest, such a model is not without problems in that it still suffers from endogeneity bias. One example is when there are respondents who quit smoking, begin regular physical exercise or change their employment status following a diagnosis; this can lead to a finding that less smoking and more exercise are associated with having a condition, which is contrary to both expectations and the actual sequence of events. To address such causation and directional issues, we design a 'lag' model that links prediagnoses risk profile and characteristics of respondents with outcomes at a later period rather than use those characteristics and profiles observed at the end point. The approach is consistent with disease progression<sup>29</sup> and avoids the strict assumptions imposed by some of the alternative approaches such as the one that uses instrumental variables.<sup>30</sup>

One of the key requirements of the proposed lagged design is the availability of a longitudinal data set for estimation. However, longitudinal data sets or follow-up studies are characterised by relatively high levels of non-response and attrition, which at times can be linked with key characteristics of respondents including their health status. Therefore, to cater to such possibilities, our final preferred model explicitly controls for selection bias resulting from missing observations owing to attritions. This is carried out by specifying a probit selection model,<sup>31 32</sup> where the selection equation and the health status equations are estimated jointly and the correlations between the error terms are directly factored in. If we use  $y_i$  to denote whether the individual remains in the sample in the follow-up wave, our approach can be summarised as follows:

$$\begin{aligned} \Pr(d_{i,t}^j = 1 | d_{i,t-1}^j = 0, y_i = 1) &= \Phi(X_{i,t-1}\beta^j + u_i^j) \\ \Pr(y_i = 1 | d_{i,t-1}^j = 0) &= \Phi(z_{i,t-1}\gamma + v_i) \end{aligned} \quad (3)$$

where  $u_i^j \sim N(0, 1)$ ,  $v_i \sim N(0, 1)$ . Both  $d_{i,t}^j$  and  $y_i$  are binary variables and can be estimated using a probit link function. Selection bias can be detected if the correlation between  $u_i^j$  and  $v_i$  is statistically different from zero.  $z_i$  represents variables that are used in the selection equation. Such selection models require the use of exclusion restrictions to obtain reliable estimates. In our case, place of residence and the number of dependent children in each age group were used for this purpose as they are expected to influence attrition but not changes in health status within the 2-year interval used in our

analysis.  $\gamma$  is the coefficient vector for  $z_i$ . The presence of data on multiple health conditions also means that we can model and study comorbidities in the study population. To do so, we use a flexible variance-covariance structure in the error terms of the two disease equations ( $u_i^j$ ).

For comparison purposes, we report the results for our proposed model (equation 3) along with the traditional prevalence-based analyses from the two cross-sectional designs. The first of these (equation 1) is a standard probit model and relates current health status with current personal characteristics of respondents. Equation 2 is also a cross-sectional data-based design but controls for prior health status.

## RESULTS AND DISCUSSION

Table 1 provides basic descriptive statistics for the study population in the final model. In total, the analysis covered 3632 individuals aged 35–70 years in 2007 with valid responses in waves 7 and 9. According to the self-reported data, all of these individuals were free from cardiovascular and type II diabetes diseases in wave 7 (2007), the starting point for analysis. Of these

individuals, 255 respondents reported having developed a CVD in wave 9 (2009), that is, within 2 years of observation and 34 others had developed type II diabetes during the same period (including 10 individuals who developed both conditions). On average, individuals who did not report a condition appeared to be younger, smoked less, had a lower body mass index (BMI) and tended to engage more in physical activity than their counterparts. They were also relatively better educated, more likely to hold managerial and technical positions, earned a higher income and had a partner or were in some form of marital relationship. Similarly, individuals who reported having no CVD or type II diabetes condition tended to report better health status during childhood.

Results shown in table 2 reveal significant changes in the lifestyle and risk profile of respondents over the interview period. For example, after 2 years of follow-up, only 84% of respondents in the original sample maintained their BMI or were within a margin of 2.5 units (half of a typical BMI-derived weight category band) and this fell to below 73% a year later. Similar trends are evident for smoking, alcohol consumption and physical activity. The proportion of respondents who remained in the same relationship and occupational categories also

**Table 1** Data description (age 35–70 years in HILDA wave 7, mean value)

Variable	Disease condition in 2 years		
	No new condition	Developed CVD	Developed type II diabetes
Age	48.34	53.38	55.26
BMI	26.3	28.5	30.53
Regular alcohol consumption	0.67	0.69	0.56
Daily smoker (1=smoke daily, 0 otherwise)	0.18	0.2	0.21
Frequency of physical activity that lasts at least 30 min			
Not at all	0.22	0.29	0.5
1–3 Times/week	0.42	0.38	0.26
3+Times/week	0.36	0.33	0.24
Number of breakfasts taken/ week	5.81	5.63	5.24
Type of jobs			
Manager and technical jobs	0.44	0.34	0.32
Service and sales	0.25	0.22	0.18
Labour/machine jobs	0.11	0.09	0.12
Equivalent income level (log)	10.38	10.24	10.25
Married or having a partner	0.78	0.74	0.76
Education status			
Bachelor's degree or above	0.29	0.16	0.18
Year 12 to Diplomas	0.44	0.42	0.38
Year 11 or below	0.27	0.42	0.44
Female	0.52	0.56	0.38
Poor childhood health	0.14	0.2	0.21
Migrant status			
Non-migrant	0.76	0.74	0.64
English speaking migrants	0.13	0.12	0.12
Non-English speaking migrants	0.11	0.14	0.24
Total number of observations (N)	3632	255	34

Mean values are reported for the continuous variables. For binary variables, yes is coded as 1 while no is coded as 0. The reported values are the mean values of the encoded variables.

BMI, body mass index; CVD, cardiovascular disease; HILDA, Australian Household, Income and Labour Dynamics survey.

**Table 2** Percentage of respondents aged 35–70 years who reported identical patterns in the follow-up rounds

Variable	Proportion with identical responses for			
	2 years (%)	3 years (%)	5 years (%)	7 years (%)
Relationship status (6 categories)	95	89	80	73
BMI (fluctuation less than 2.5)	84	73		
Smoking behaviour (5 categories)	90	80	69	61
Alcohol consumption pattern (8 categories)	61	42	26	18
Physical activity pattern (6 categories)	41	21	8	4
Occupation type (8 categories)	77	63	47	37

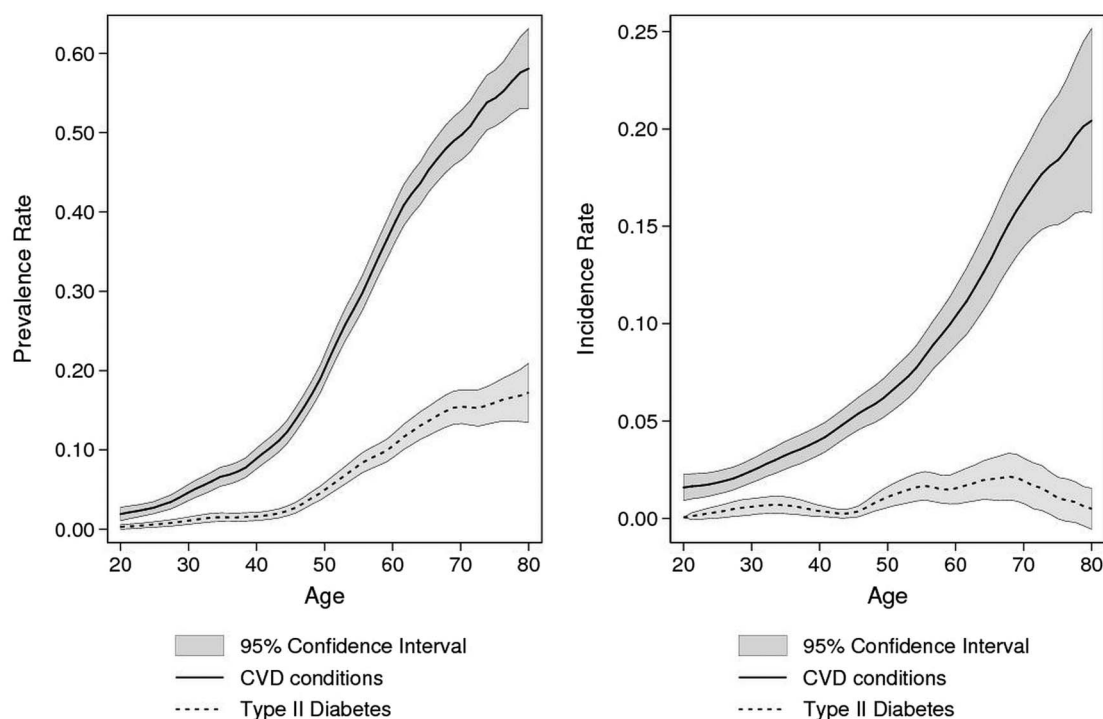
Only individuals who gave a valid response are included. 2009 is the last year used in this table. BMI, body mass index.

declined progressively with time, which suggests that characteristics and risk profiles reported at one point in time may not approximate the characteristics of these individuals in another period, even when the intersurvey interval is reasonably short. Notwithstanding the potential differences in responses to questions over time, resulting from changes in levels of understanding and interpretation of survey questions, the findings from [table 2](#) clearly highlight the problems associated with cross-sectional studies. Lack of information on timing of events and reliance on characteristics reported at the time of survey to predict events that occurred at an unknown period before the survey date can be misleading.

[Figure 1](#) shows age-specific prevalence and incidence rates for CVD and type II diabetes with their associated CIs using local polynomial smoothing. The prevalence

rates for both conditions resemble published results based on the Australian Health Survey (AHS).<sup>33</sup> As expected for both conditions, prevalence rates increased with age of respondents and were higher for CVD than for type II diabetes. By age 50 years, over a third of the population suffered from a lifetime risk of CVD and about 1 in 20 also reported suffering from type II diabetes. Incidence rates also rose with age for CVD, but for type II diabetes they did so only up to the late 60s and declined thereafter.

[Table 3](#) shows the probabilities of transition into the four different health states (healthy; with CVD condition only; with type II diabetes condition only; with both conditions) from a previously healthy state. As can be seen, about 7–8% of the respondents, who reported health status in both waves and were healthy at the beginning of the observation period, became diabetic or developed



**Figure 1** Smoothed prevalence and incidence rate of CVD and type II diabetes in HILDA (wave 9). The rates are smoothed using local polynomial regression estimators with a bandwidth of 2.5. 95% CI is shown in shaded colours. CVD, cardiovascular disease; HILDA, Australian Household, Income and Labour Dynamics survey.

**Table 3** Raw transition probabilities within 2 years for a healthy individual in HILDA between 2007 and 2009

Year 2007	Year 2009			
	Remain healthy (%)	CVD Condition (%)	Type II diabetes (%)	Both CVD and type II diabetes (%)
Male	92.7	6.2	1.0	0.2
Female	92.1	7.1	0.4	0.3

CVD, cardiovascular disease; HILDA, Australian Household, Income and Labour Dynamics survey.

at least one CVD condition 2 years later. The transition probabilities for both genders followed a similar pattern. For both men and women, the probability of developing type II diabetes was lower than the probability of developing CVD conditions, as expected.

Tables 4 and 5 present results of our regression models for CVD and type II diabetes, respectively. These tables contain three statistical designs, namely results from a cross-sectional approach, a model based on cross-sectional data but estimated on the condition of no prior health condition, and a final model that controls for prior health status and corrects for endogeneity biases. The cross-sectional models were based on wave 9, while for the final model both waves 7 and 9 were used. As stated earlier, in the final model, in addition to estimating the risks of developing CVD and type II diabetes jointly, we also tested the results for the effects of attrition and selection biases.

The analyses suggest notable differences in patterns and the strength of relationships between risk factors and CVD, depending on the strands of the statistical design used for analyses. BMI and a higher education level were consistently and significantly associated with CVD conditions in all three models. However, the strength and significance of the relationships observed for these variables varied among the models. For example, while BMI was significant across all the three models for CVD, its impact was stronger in the model that controlled for endogeneity biases and weakest in the models with a cross-sectional design. Specifically, the point estimates for the final preferred model were more than two SEs higher than those for the cross-sectional estimates. Results for the conditional cross-sectional model were somewhere in between, but closer to the estimates from the conventional cross-sectional model, which suggests that the effects of endogeneity biases are not negligible.

Similar patterns were apparent for alcohol consumption where its effect was found to be insignificant for the standard cross-sectional model. The stronger effect of these factors for the endogeneity-corrected model may suggest that patients diagnosed with CVD conditions were more likely to control their alcohol consumption and physical weight through exercise and diet, a finding that is consistent with a number of earlier clinical trial studies.<sup>34</sup>

On the other hand, once the effects of comorbidity and endogeneity biases were corrected for, occupation and smoking no longer predicted the chance of developing CVD as shown in model 3. Income, education and migration status all showed a stronger impact in the preferred model than the standard cross-sectional (model 1) or the cross-sectional model estimated on condition of no prior health condition (model 2). Additionally, in two of the three models, respondents who reported childhood health problems were more likely to suffer from CVD. In all the three models for CVD, females tended to have a higher risk of developing the condition than their male counterparts, but the difference between them was statistically significant only in the final model, which corrects for endogeneity and selectivity biases.

Interestingly, physical activity did not seem to be an important predictor of risk of developing a CVD condition, although the sign of the coefficient for the variable exhibited the expected adverse impact. However, once the comorbidity and endogeneity biases were corrected for, physical inactivity was significantly associated with the development of type II diabetes. This suggests that the effect of physical activity, as is the case with BMI and alcohol consumption for CVD, may have been underestimated in cross-sectional studies owing to the potential dose–response behaviour. This is consistent with previous medical research on the mechanism and pathways of influence of increased physical activity on risk of diabetes.<sup>35 36</sup>

Certain socioeconomic factors such as employment status and occupation that are usually found to be associated with CVD or type II diabetes in cross-sectional studies showed a similar impact in our cross-sectional designs. However, the findings that these same variables did not have any statistically significant coefficient in the preferred final model seems to suggest that the observed correlation might as well be a result of endogeneity-related issues,<sup>37 38</sup> rather than of any true effect of employment on CVD. This is in line with some of the clinical studies,<sup>39 40</sup> where education, instead of occupation, was found to be a more significant factor. Additionally, the finding suggests that immigrants with a non-English speaking background have a higher risk of developing both CVD and type II diabetes than Australian-born respondents, a finding consistent with some of the earlier research on Australian immigrants.<sup>41 42</sup>

Moreover, the analysis suggests a high and significant correlation between CVD status and the incidence of type 2 diabetes in the population, even after controlling for socioeconomic status and lifestyle factors. The correlation coefficient between the error terms of the two disease equations is positive and statistically significant (coefficient 0.27, p value 0.01), which means that those who developed either of the conditions are more likely to develop the other condition as well. In the preferred model, we also tested for selection-associated or

**Table 4** CVD estimations

Variables	Cross-sectional	Conditional cross-sectional	Preferred model
Social demographic factors			
Female	-0.03 (0.05)	0.10 (0.07)	0.14** (0.07)
Age	0.08*** (0.03)	0.05 (0.05)	0.06 (0.04)
Age squared	0.00* (0.00)	0.00 (0.00)	0.00 (0.00)
Foreign born (English speaking)	-0.02 (0.06)	-0.12 (0.10)	-0.07 (0.10)
Foreign born (Non-English speaking)	0.05 (0.07)	0.18* (0.11)	0.21** (0.11)
Equivalent income	0.04 (0.04)	0.01 (0.05)	-0.11* (0.06)
Married or having a partner	-0.01 (0.05)	-0.06 (0.08)	-0.04 (0.08)
Bachelor's degree or above	-0.16** (0.06)	-0.34*** (0.10)	-0.34*** (0.10)
Year 12 to Diplomas (Baseline: year 11 or below)	-0.08* (0.05)	-0.17** (0.08)	-0.14* (0.08)
Manager and technical jobs (Baseline: not working)	-0.21*** (0.06)	-0.18* (0.10)	0.03 (0.10)
Service and sales	-0.17*** (0.07)	-0.16 (0.11)	-0.03 (0.10)
Labour/machine operators	-0.17** (0.08)	-0.28** (0.14)	-0.09 (0.13)
Mid to poor childhood health	0.09 (0.06)	0.19** (0.09)	0.16* (0.09)
Lifestyle factors			
BMI	0.04* (0.02)	0.06** (0.03)	0.11*** (0.03)
Age BMI interaction	0.00 (0.00)	0.00 (0.00)	0.00** (0.00)
Regular alcohol consumption	0.00 (0.05)	0.26*** (0.08)	0.24*** (0.08)
Daily smoker	-0.11* (0.06)	-0.03 (0.10)	0.08 (0.09)
1-3 times reported physical activity per week (Baseline: <1/week)	0.05 (0.05)	0.02 (0.09)	-0.11 (0.08)
3+ times reported physical activity per week	-0.04 (0.06)	-0.08 (0.09)	-0.10 (0.09)
Number of breakfasts per week	-0.01 (0.01)	-0.03* (0.02)	-0.02 (0.02)
Constant	(1.04)	-4.42*** (1.52)	-4.63*** (1.52)
Observations	5261	3464	3632

SEs in parentheses. All SEs are Huber-White sandwich estimators, which are estimated using the 'robust' option in Stata. \*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

BMI, body mass index; CVD, cardiovascular disease.

attrition-associated biases but found no supporting evidence. The  $p$  value for the correlation between the error terms in the selection equation and the disease equations (see equation 3) were well above the 0.10 threshold.

## CONCLUSIONS

CVD and type II diabetes are two of the major public health challenges of our time, and both conditions are

expected to assume even greater significance in future, given the ongoing demographic transformation in Australia and around the world. In this population-level study, we followed a panel of nationally representative individuals aged 35 years or over with no prior type II diabetes or CVD condition in the initial wave under consideration and re-examined the socioeconomic and risk factors associated with developing these conditions; we also compared the results from cross-sectional study

**Table 5** Type II diabetes estimation

Variables	Cross-sectional	Conditional cross-sectional	Preferred model
Social demographic factors			
Female	-0.40*** (0.07)	-0.37** (0.15)	-0.30** (0.15)
Age	0.16*** (0.05)	0.16 (0.13)	0.12 (0.12)
Age squared	0.00*** (0.00)	0.00 (0.00)	0.00 (0.00)
Foreign born (English speaking)	0.08 (0.10)	-0.06 (0.21)	-0.01 (0.20)
Foreign born (Non-English speaking)	0.22** (0.09)	0.32* (0.18)	0.36** (0.16)
Equivalent income	0.00 (0.06)	0.02 (0.10)	0.01 (0.13)
Married or having a partner	-0.20*** (0.07)	-0.08 (0.16)	0.06 (0.16)
Bachelor's degree or above (Baseline: year 11 or below)	-0.04 (0.10)	-0.04 (0.16)	-0.18 (0.20)
Year 12 to Diplomas	-0.04 (0.08)	-0.24 (0.16)	-0.14 (0.15)
Manager and technical jobs (Baseline: not working)	-0.38*** (0.09)	-0.23 (0.21)	-0.07 (0.19)
Service and sales	-0.27*** (0.10)	-0.21 (0.22)	-0.13 (0.22)
Labour/machine operators	-0.21* (0.12)	-0.11 (0.23)	-0.11 (0.24)
Mid to poor childhood health	-0.06 (0.09)	0.15 (0.18)	0.10 (0.18)
Lifestyle factors			
BMI	0.04 (0.03)	0.02 (0.05)	0.00 (0.06)
Age BMI interaction	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Regular alcohol consumption	-0.46*** (0.07)	-0.24* (0.14)	-0.08 (0.14)
Daily smoker	0.06 (0.10)	-0.39 (0.31)	0.03 (0.18)
1-3 times reported physical activity per week (Baseline: <1/week)	-0.15* (0.08)	-0.24 (0.18)	-0.41** (0.16)
3+ times reported physical activity per week	-0.11 (0.08)	-0.04 (0.17)	-0.40** (0.17)
Number of breakfasts per week	0.05*** (0.02)	0.00 (0.04)	-0.03 (0.03)
Constant	(1.79)	-6.13 (1.79)	-7.67* (3.94)
Observations	5261	3464	3632

SEs in parentheses. All SEs are Huber-White sandwich estimators, which are estimated using the 'robust' option in Stata. \*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

BMI, body mass index.

designs. Specifically, we controlled for selection-related and attrition-related biases and related the outcome variables to prediagnosis behaviour and characteristics rather than to those that prevailed at the time of enquiry. We then compared the results with standard cross-sectional based models to gauge the level of endogeneity-related biases resulting from adopting the latter approach.

Our findings reaffirm some of the findings of earlier studies while at the same time revealing some

differences with results from cross-sectional studies. Our estimations suggest that BMI and alcohol consumption have a much stronger association with risk of developing a CVD condition than cross-sectional estimates would suggest. Physical activity was found to have a significant effect on developing type II diabetes once we controlled for endogeneity biases.

Our analyses suggest that while physical activity and lifestyle factors appear more important in predicting the



probability of developing a CVD in the endogeneity-corrected model, the impact observed with regard to socioeconomic characteristics for this model was rather lower as compared with what is generally reported in cross-sectional studies. For example, employment status and occupation types that are usually found to be associated with CVD or type II diabetes in cross-sectional studies appear to have no impact once the model is corrected for endogeneity biases.

A number of results relevant to policy emerge from our analysis. Given the importance of BMI for CVD prevention and the high levels of population-wide BMI in Australia, it is high time that health intervention programmes pay greater attention to the obesity epidemic in the country. The total annual direct cost of overweight and obesity in Australia in 2005 was \$21 billion,<sup>43</sup> and is likely to be even greater for recent years, given the current trend of obesity in the country. The stronger link between BMI and CVD in this paper therefore suggests that surgical options in the case of excessive obesity<sup>44</sup> appear to be more warranted than was previously thought. Along with medical interventions, efforts to promote healthy lifestyles (such as those related to diet, smoking and physical exercise) also need strengthening in the country. The findings on the importance of physical activity for type II diabetes and smoking for CVD are particularly important and need to be considered in designing future intervention strategies. Further, a targeted screening process for type II diabetes, especially among the country's foreign-born population, could help practitioners identify early-stage patients, given the positive and significant link between migration status and newly developed type II diabetes.

Finally, a few caveats around the research are in order. One of the major limitations of our analysis relates to the fact that the health status variables used in the study were not based on biomarkers but were obtained by way of self-reporting. Although it is possible that part of this limitation could be offset by the potential correlations between biomarkers and lifestyle factors, of which our data have an extensive collection, self-reported data may contain measurement errors that warrant further investigation as and when such data become available. In view of this and the increasing burden of non-communicable conditions, there appears to be an opportunity for stakeholders of the HILDA survey to introduce biomarkers for selected health conditions, possibly to a limited subsample of the survey. However, given that HILDA is not meant to be a health survey and collects only a limited range of information, future health research in Australia should also look into ways of introducing large-scale nationally representative panel surveys focusing on major health issues in the country. This can be achieved either by extending the 45 UP study to other jurisdictions (beyond New South Wales) or by widening the scope of the AusDiab survey to include other major non-communicable conditions as well as by shortening the interval between rounds for the AusDiab survey from 5

to 3 years. Meanwhile, while analysing existing longitudinal clinical data sets, further research into the socioeconomic and risk factor effects on CVD and diabetes in Australia and elsewhere may consider applying our preferred model to correct for endogeneity and selectivity biases.

#### Author affiliations

<sup>1</sup>National Economic and Social Modelling Centre, Institute of Governance and Policy Analysis, University of Canberra, Canberra, Australian Capital Territory, Australia

<sup>2</sup>Centre for Research and Action in Public Health, Health Research Institute, University of Canberra, Canberra, Australian Capital Territory, Australia

<sup>3</sup>School of Demography, The Australian National University, Canberra, Australian Capital Territory, Australia

**Acknowledgements** This paper uses unit record data from the Household, Income and Labour Dynamics in Australia (HILDA) Survey. The HILDA Project was initiated and is funded by the Australian Government Department of Social Services (DSS) and is managed by the Melbourne Institute of Applied Economic and Social Research (Melbourne Institute). The findings and views reported in this paper, however, are those of the authors and should not be attributed to either DSS or the Melbourne Institute. The authors are grateful for the valuable comments and feedback from Tom Cochrane.

**Contributors** YK conceptualised the study, the design and approach. JL prepared the data and undertook the analysis. Both authors were equally responsible for drafting the manuscript.

**Funding** This work was supported by The National Heart Foundation (NHF) Special Grant to the Centre for Research and Action in Public Health, University of Canberra. However, the findings and views reported in this paper remain those of the authors and do not necessarily reflect the views of either the NHF or its staff.

**Competing interests** None declared.

**Ethics approval** This work reports the secondary analysis of a non-identifiable data set. Ethical approval for the collection and analysis of HILDA has been provided by the Human Ethics Advisory Committee at the University of Melbourne, Australia.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data sharing statement** No additional data are available.

**Open Access** This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>

#### REFERENCES

1. Fox CS, Coady S, Sorlie PD, *et al*. Increasing cardiovascular disease burden due to diabetes mellitus: the Framingham Heart Study. *Circulation* 2007;115:1544–50.
2. Lim SS, Vos T, Flaxman AD, *et al*. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* 2012;380:2224–60.
3. Roth GA, Forouzanfar MH, Moran AE, *et al*. Demographic and Epidemiologic Drivers of Global Cardiovascular Mortality. *N Engl J Med* 2015;372:1333–41.
4. AIHW. *Cardiovascular disease: Australian facts 2011. Cardiovascular disease series. Cat. no. CVD 53*. Canberra: Australian Institute of Health and Welfare, 2011.
5. Shaw JE, Sicree RA, Zimmet PZ. Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes Res Clin Pract* 2010;87:4–14.
6. Magliano DJ, Shaw JE, Shortreed SM, *et al*. Lifetime risk and projected population prevalence of diabetes. *Diabetologia* 2008;51:2179–86.

7. Kanjilal S, Gregg EW, Cheng YJ, *et al.* Socioeconomic status and trends in disparities in 4 major risk factors for cardiovascular disease among US adults, 1971-2002. *Arch Intern Med* 2006;166:2348-55.
8. Powell KE, Thompson PD, Caspersen CJ, *et al.* Physical activity and the incidence of coronary heart disease. *Annu Rev Public Health* 1987;8:253-87.
9. Sjöström L, Lindroos AK, Peltonen M, *et al.* Lifestyle, diabetes, and cardiovascular risk factors 10 years after bariatric surgery. *N Engl J Med* 2004;351:2683-93.
10. Wing RR, Look AHEAD Research Group. Long term effects of a lifestyle intervention on weight and cardiovascular risk factors in individuals with type 2 diabetes: four year results of the Look AHEAD trial. *Arch Intern Med* 2010;170:1566.
11. Pi-Sunyer X, Blackburn G, Brancati FL, Look AHEAD Research Group. Reduction in weight and cardiovascular disease risk factors in individuals with type 2 diabetes: one-year results of the look AHEAD trial. *Diabetes Care* 2007;30:1374-83.
12. Aviña-Zubieta JA, Choi HK, Sadatsafavi M, *et al.* Risk of cardiovascular mortality in patients with rheumatoid arthritis: a meta-analysis of observational studies. *Arthritis Rheum* 2008;59:1690-7.
13. Ruige JB, Assendelft WJ, Dekker JM, *et al.* Insulin and risk of cardiovascular disease: a meta-analysis. *Circulation* 1998;97:996-1001.
14. Tong B, Stevenson C. *Comorbidity of cardiovascular disease, diabetes and chronic kidney disease in Australia.* Australian Institute of Health and Welfare, 2007.
15. Sowers JR, Epstein M, Frohlich ED. Diabetes, hypertension, and cardiovascular disease an update. *Hypertension* 2001;37:1053-9.
16. Johnson JV, Hall EM. Job strain, work place social support, and cardiovascular disease: a cross-sectional study of a random sample of the Swedish working population. *Am J Public Health* 1988;78:1336-42.
17. Bhopal R, Unwin N, White M, *et al.* Heterogeneity of coronary heart disease risk factors in Indian, Pakistani, Bangladeshi, and European origin populations: cross sectional study. *BMJ* 1999;319:215-20.
18. Imai K, Nakachi K. Cross sectional study of effects of drinking green tea on cardiovascular and liver diseases. *BMJ* 1995;310:693-6.
19. Carlson MDA, Morrison RS. Study design, precision, and validity in observational studies. *J Palliat Med* 2009;12:77-82.
20. Cameron AJ, Welborn TA, Zimmet PZ, *et al.* Overweight and obesity in Australia: the 1999-2000 Australian diabetes, obesity and lifestyle study (AusDiab). *Med J Aust* 2003;178:427-32.
21. Banks E, Redman S, Jorm L, 45 and Up Study Collaborators. Cohort profile: the 45 and up study. *Int J Epidemiol* 2008;37:941.
22. Frick JR, Jenkins SP, Lillard DR, *et al.* The Cross-National Equivalent File (CNEF) and its member country household panel studies. *Schmollers Jahrbuch: J Applied Social Science Studies* 2007;127:627-54.
23. Watson N, Wooden MP. The HILDA Survey: a case study in the design and development of a successful household panel survey. *Longitudinal Life Course Stud* 2012;3:369-81.
24. Butterworth P, Crosier T. The validity of the SF-36 in an Australian National Household Survey: demonstrating the applicability of the Household Income and Labour Dynamics in Australia (HILDA) Survey to examination of health inequalities. *BMC Public Health* 2004;4:44.
25. Berry HL, Welsh JA. Social capital and health in Australia: an overview from the household, income and labour dynamics in Australia survey. *Soc Sci Med* 2010;70:588-96.
26. Butterworth P, Leach LS, Rodgers B, *et al.* Psychosocial job adversity and health in Australia: analysis of data from the HILDA Survey. *Aust N Z J Public Health* 2011;35:564-71.
27. Crosier T, Butterworth P, Rodgers B. Mental health problems among single and partnered mothers. The role of financial hardship and social support. *Soc Psychiatry Psychiatr Epidemiol* 2007;42:6-13.
28. National Vascular Disease Prevention Alliance. *Absolute cardiovascular disease management.* Quick reference guide for health professionals, 2012.
29. Rosin BL. The progression of cardiovascular risk to cardiovascular disease. *Rev Cardiovasc Med* 2007;8(Suppl 4):S3-8.
30. Didelez V, Meng S, Sheehan NA. Assumptions of IV methods for observational epidemiology. *Statistical Sci* 2010;22-40.
31. Van de Ven WP, Van Praag B. The demand for deductibles in private health insurance: a probit model with sample selection. *J Econometrics* 1981;17:229-52.
32. Heckman JJ. Sample selection bias as a specification error. *Econometrica* 1979:153-61.
33. AIHW. *Cardiovascular disease, diabetes and chronic kidney disease: Australian facts: prevalence and incidence.* Canberra: Australian Institute of Health and Welfare, 2014.
34. Stevens VJ, Obarzanek E, Cook NR, *et al.* Long-term weight loss and changes in blood pressure: results of the Trials of Hypertension Prevention, phase II. *Ann Intern Med* 2001;134:1-11.
35. Knowler WC, Barrett-Connor E, Fowler SE, Diabetes Prevention Program Research Group. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N Engl J Med* 2002;346:393.
36. Lindström J, Ilanne-Parikka P, Peltonen M, *et al.* Sustained reduction in the incidence of type 2 diabetes by lifestyle intervention: follow-up of the Finnish Diabetes Prevention Study. *Lancet* 2006;368:1673-9.
37. Church TS, Thomas DM, Tudor-Locke C, *et al.* Trends over 5 decades in US occupation-related physical activity and their associations with obesity. *PLoS ONE* 2011;6:e19657.
38. Söderberg M, Rosengren A, Hillström J, *et al.* A cross-sectional study of the relationship between job demand-control, effort-reward imbalance and cardiovascular heart disease risk factors. *BMC Public Health* 2012;12:1102.
39. Maty SC, Everson-Rose SA, Haan MN, *et al.* Education, income, occupation, and the 34-year incidence (1965-99) of type 2 diabetes in the Alameda County Study. *Int J Epidemiol* 2005;34:1274-81.
40. Winkleby MA, Jatulis DE, Frank E, *et al.* Socioeconomic status and health: how education, income, and occupation contribute to risk factors for cardiovascular disease. *Am J Public Health* 1992;82:816-20.
41. Dassanayake J, Gurrin L, Payne WR, *et al.* Is country of birth a risk factor for acute hospitalization for cardiovascular disease in Victoria, Australia? *Asia Pac J Public Health* 2011;23:280-7.
42. Dassanayake J, Dharmage SC, Gurrin L, *et al.* Are Australian immigrants at a risk of being physically inactive. *Int J Behav Nutr Phys Act* 2011;8:53.
43. Colagiuri S, Lee CM, Colagiuri R, *et al.* The cost of overweight and obesity in Australia. *Med J Aust* 2010;192:260-4.
44. AIHW. *Weight loss surgery in Australia. Cat. no. HSE 91.* Canberra: Australian Institute of Health and Welfare, 2010.