NATSEM
at the University of Canberra

Moving beyond comparative validation: Predictive abilities of APPSIM's health module

NATSEM Working Paper 11/10

Sharyn Lymer
Alan Duncan
Laurie Brown

June 2011

MOVING BEYOND COMPARATIVE VALIDATION: PREDICTIVE ABILITIES OF APPSIM'S HEALTH MODULE

Authors:
Sharyn Lymer
PhD Student
National Centre for Social and Economic Modelling (NATSEM),
University of Canberra, ACT, 2601, Australia
Sharyn.lymer@natsem.canberra.edu.au
P: 61 2 6201 2766
F: 61 2 6201 2751

Alan Duncan
Director
National Centre for Social and Economic Modelling (NATSEM),
University of Canberra, ACT, 2601, Australia
alan.duncan@natsem.canberra.edu.au
P: 61 2 6201 2781
F: 61 2 6201 2751

Prof. Laurie Brown
Research Director, Health
National Centre for Social and Economic Modelling (NATSEM),
University of Canberra, ACT, 2601, Australia
laurie.brown@natsem.canberra.edu.au
P: 61 2 6201 2770
F: 61 2 6201 2751

ABSTRACT

In the development of a dynamic microsimulation model, validation lends credibility to the model, making it more likely to be accepted by policy makers. In the validation of multi-module models such as the Australian Population and Policy Simulation model (APPSIM) currently under development at the National Centre for Social and Economic Modelling (NATSEM), one is faced with additional validation challenges. Not only do the specific characteristics of the module under consideration need validating but so too do the data inputs from earlier developed modules. Moreover, validation needs to account for dynamic interactions with the other modules in the model system. This paper considers the validation of the health module of APPSIM, which is itself highly reliant on prerequisite modules such as education and earnings. Commonly used comparative validation methods are used. In addition, as part of the validation of the health module, issues related to variable inputs are disentangled from the health prediction models. To achieve this, a version of the APPSIM health module is developed in SAS and run against the Australian longitudinal HILDA dataset between 2006 and 2008. This provides an alternate data source for the socio-economic variables used in the model, providing health outcomes to assess the equation's predictive quality. Since these data are not used in the development of the prediction equations, the proposed method provides an independent dataset for testing the module's generalisability. Using this method, it is possible to verify both the baseline imputation and the transition equations used in the dynamic simulation model. The validation of the equations' predictive qualities includes 'confusion tables' to consider at a simple level the predictive accuracy of the model, precision measures looking at the ratio of true positive results to all positive results, and ROC curves to consider the true positives against the false positives. As an illustration, a validation of predictions of being overweight or obese from the APPSIM health module is presented.

1 BACKGROUND – WHAT IS VALIDATION, STRATEGIES FOR VALIDATION, HOW IMPORTANT IS VALIDATION

Validation offers a treatment to gain an understanding of strength and accuracy of the microsimulation model being developed (Cassells et al. 2006). Validation is a systematic process that is used to determine any possible threats to the models outputs and possible credibility (Morrison 2008). By determining a level of confidence in the model and its outcomes, policy-makers are more inclined to use its output as a contribution for informing policy decision-making. In the validation of a model, the process should be ongoing and built into the output that is routinely provided for the model (Morrison 2008). More recent work has offered an expanded framework of validation to include not only the aspects of the model's performance but also the process of model development and the quality of the decisions based on the model(Kopec et al. 2010). From the beginning of the model's development, a validation plan needs to be outlined looking at the indicators of model quality that will be required. However, the validation plan is not a rigid document, but one that develops over time with the model and its changes.

An important issue is what measures/analysis/graphics can be considered to provide sufficient confidence in the model. A variety of techniques have been suggested to contribute to the validation process of a microsimulation model. There are three key areas discussed in the validation of microsimulation models: external comparison, sensitivity analysis and determination of confidence intervals. The primary focus for validation is on the comparative analysis that starts to provide evidence about the quality of the model. Techniques related to sensitivity analysis and confidence intervals provide further confidence about the quality of the model but are not necessarily the primary focus of the validation process.

It is important to compare characteristics within the model against credible external data, both macro level totals and distributions (Morrison 2008). Compatibility of model's output with data from multiple sources provides strong confidence in the model's capabilities. Marked deviation of the model from actual data indicates the need for the model to be reconsidered in an attempt to better represent the processes of the system being modelled, and the requirement for calibration or alignment to be implemented. Validation can also include checking that the model aligns with other projections from alternate sources using different methods. Further, for dynamic models not only must there be cross-sectional comparisons but also longitudinal comparisons (Harding et al. forthcoming). Aspects considered in longitudinal comparisons include comparisons of the rates of change between individual states, how frequently transitions are occurring and how many transitions occur in a lifetime.

Sensitivity analysis allows analysis of the impact of input parameters on the outcomes of interest, by varying the parameter inputs, whilst holding other aspects constant. It also provides insights to what parameters may offer the most leverage in policy. As such, it provides outputs that contribute to our understanding of the systems being studied(Zaidi and Rake 2001). Within this group of techniques the running of counterfactuals in the model and looking at their results can be included. There is more

confidence in the model if the counterfactual results are in line with previous modelling in the literature and fit with theoretical models of the system.

The third set of techniques that can contribute to validation is the provision of a measure of the uncertainty surrounding the estimates from the simulation. This gives an indication of the precision of the outputs, which in combination with adequate comparative results helps establish confidence in the model. This is a difficult aspect of validation to be implemented within microsimulation due to the wide variety of elements that can contribute to the uncertainty surrounding the model's estimates. There are few examples of microsimulation models where this aspect of validation has been attempted. In the Future Elderly Model (Goldman et al. 2004) resampling techniques were used to provide a prediction interval. The use of resampling techniques provides a way of estimating the sources of sample variability and provides a method of assessing the size of the error resulting from the monte carlo processes used in the microsimulation model (O'Donoghue 2001). This method does not provide an indication of the variability created due to the estimation of the model parameters and the choice of parameters. A more complex analysis has been used in the MITTS model where a simulation process estimates the sampling distribution of aggregate measures based on the sampling distribution of the estimated parameters within the statistical equation(Creedy et al. 2007). Pudney and Sutherland calculated asymptotic confidence intervals for key summary measures in POLIMOD(1994). These estimates account for the reweighting of the survey data used as a basefile in the microsimulation model and show the levels of sampling error that affect the microsimulation output.

Additional techniques have been used in other areas of modelling for validation. In particular, the focus of this paper is on the methods that have been adapted from the biostatistical literature on validation of prognostic. Data mining literature also offers similar solutions around issues of how to validate models (Han and Kamber 2006). Measures of the predictive quality of the statistical models being used in the microsimulation model, such as presented in confusion tables and ROC curves allow consideration of the quality of the statistical models being introduced into the microsimulation model. These measures are described in some detail in the next section.

1.1 ALTERNATIVE MEASURES OF COMPARISON

Confusion tables (also known as hit and miss tables) summarise information about the actual and predicted classifications for categorical variables. The specification of a confusion table for a binary variable is shown in Table 1. From the comparison of actual data to simulated categories, *a* is the number of simulated individuals that did not have the observed characteristic, *b* is the number of incorrect simulations where the person has the characteristic, *c* is the number of incorrect simulations where the person does not have the observed characteristic and *d* is the number of correct simulation where the individual does have the characteristic.

Table 1 Confusion table layout

		Simulated	
		Negative	Positive
Actual	Negative	a	b
	Positive	c	d

For binary variables, there is a range of summary statistics that can be calculated from this table. These measures that can be calculated to provide qualitative measures of the ability of the simulation to "correctly" allocate a person's characteristics. Selections of these types of measures are presented in Box 1.

Box 1 Measures of Predictive Ability for Binary Variables

Accuracy is the proportion of the simulated outcomes that were correct.

= (a+d)/ (a+b+c+d)

Sensitivity is the proportion of true positives that are correctly identified by the test.

= d/(c+d)

Specificity is the proportion of true negatives that are correctly identified by the test

a/(a+b)

(Altman and Bland 1994a)

Positive predictive value (PPV) is the proportion of individuals who have the characteristic who are correctly allocated in the simulation.

$$= \frac{\text{sensitivity} \times \text{prevalence}}{\text{sensitivity} \times \text{prevalence} + (1 - \text{specificity}) \times (1 - \text{prevalence})}$$

Negative Predictive Value (NPV)

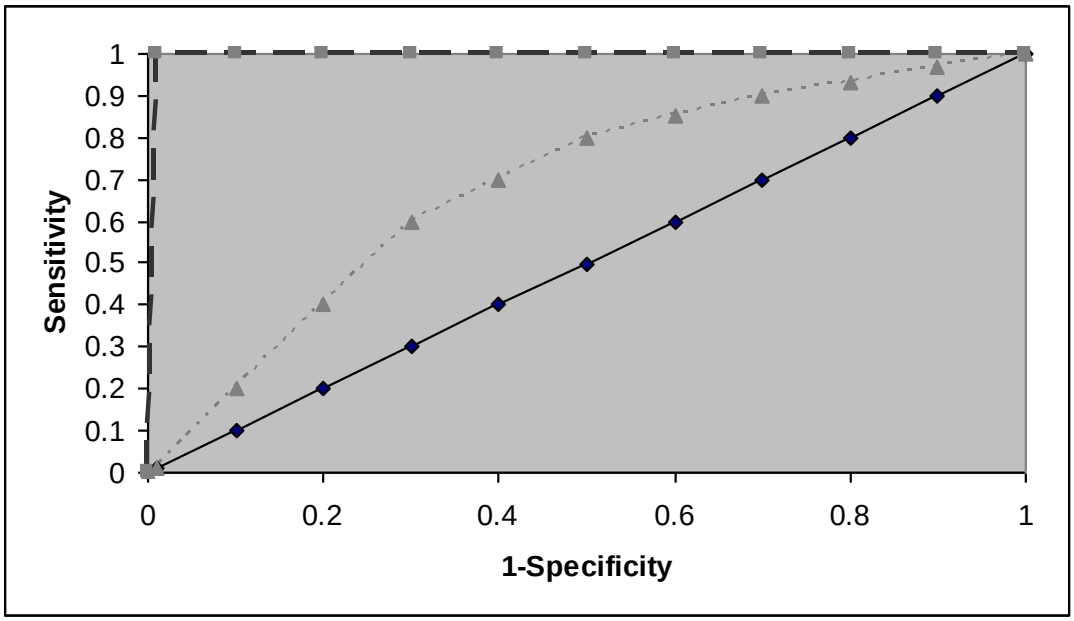
$$= \frac{(1 - \text{sensitivity}) \times \text{prevalence}}{(1 - \text{sensitivity}) \times \text{prevalence} + \text{specificity} \times (1 - \text{prevalence})}$$

(Altman and Bland 1994b)

For binary variables where a person either has an attribute or not and probabilities are used in the determination of having the characteristic, receiver operating characteristic (ROC) plots can provide insights into how useful the statistical equation is in distinguishing the two groups. It is assumed high probabilities are more likely to have the attribute of interest. The ROC plot is developed by calculating the sensitivity and specificity of every individual's data and plotting sensitivity (true positive rate) against 1-specificity (false positive rate)(Altman and Bland 1994c). The calculation of sensitivity and specificity is made for multiple cut-points between 0 and 1 (the probability range). It is expected that there will be an overlap of probabilities between the group who has the attribute and the group that does not. An example of what this graph may look like is presented in Figure 1. If the equation was to discriminate perfectly between the two groups then the curve would follow the left side of the graph and across the top (that is the grey dashed line on Figure 1). If the statistical equation is providing no information to distinguish the two groups the plot would be a straight line from the bottom left corner to the top right corner. This is the black line on Figure 1

where sensitivity and 1-specificity are equal. This is the point where the equation is performing no better than a random allocation. Usually the curve lies between these two extremes (as is the case with the grey dotted line on Figure 1). The area under the curve (that is the area between the grey dotted line and the black line) can provide a measure of the accuracy of the equation. These measures allow the comparison of alternative models with respect to their accuracy in allocating characteristics in the microsimulation model. This type of graphic also allows the comparison of how the test is performing between different sub-groups of the population, for example, genders. The higher the dotted line in the graph is above the black line the better the equation performs at distinguishing the groups with and without the attribute because it has a larger true positive rate to a lower false positive rate.

Figure 1 ROC Curve



2 APPSIM – THE MODELLING FRAMEWORK

The framework of the health module is Australian Population and Policy Simulator Model (APPSIM), a dynamic microsimulation model developed at National Centre for Social and Economic Modelling (NATSEM) (Kelly 2011). It is a population-based, closed, dynamic microsimulation model that operates in discrete time. The model operates by processing through 12 modules (see Figure 1) sequentially. The simulated events include death, immigration and emigration, marriage, divorce, childbirth, ageing, education, labour force participation, earnings, retirement, moving to aged care, changing health status and the use of health services. Through these events, people earn income, receive social security, pay taxes and accumulate assets. The structure of APPSIM, like most other dynamic microsimulation models, is such that it has an initial starting population, a simulation cycle and an output. Within the simulation cycle, the probability of an event occurring is determined by sets of functions or tables of probabilities.

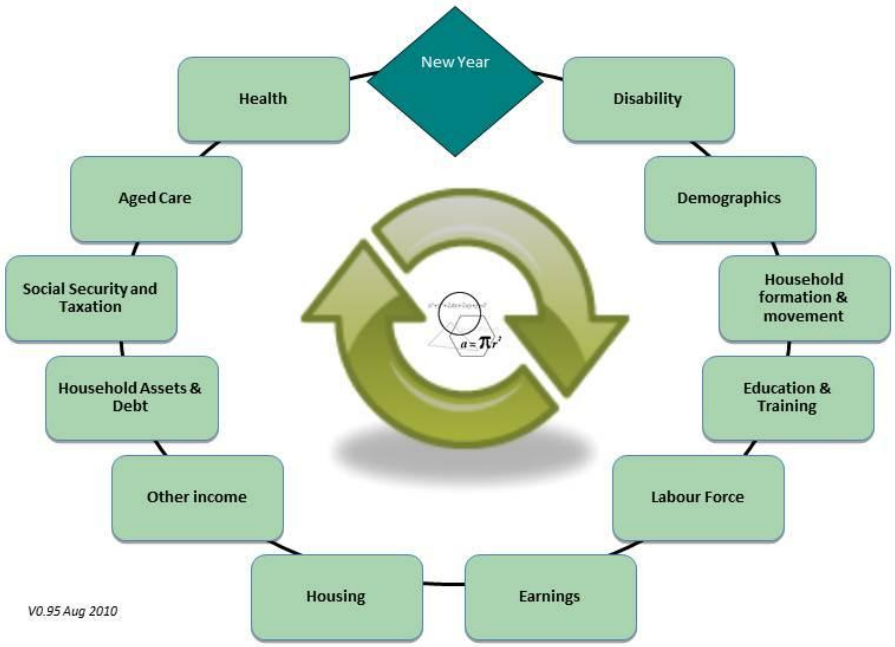
The simulation uses these transition probabilities to determine if an event occurs or not. The probabilities are based on the individual's demographic and socio-economic

characteristics, history and the simulated time. As the simulation moves from year to year, the probability of a person changing from one state to another is considered (for example changing from being classified as "obese" to "not being obese". Most of the equations used in APPSIM are based on the Household, Income and Labour Dynamics in Australia (HILDA) Survey (Watson 2010), which is a longitudinal survey of Australian households that has been running since 2001.

After the calculation of a transition probability (which by definition falls between 0.0 and 1.0), monte carlo simulation is used to determine if the event actually occurs. That is, the estimated probability is compared with a random number drawn from a uniform 0-1 distribution. Based on the result of this comparison, the transition may be flagged to occur. For instance, if the probability of being obese is 0.20, then if a random number of 0.15 is drawn the person is deemed to be obese in this time period. However, if a random number of 0.80 was drawn, the person would be classed as not being obese in this time period.

The APPSIM model is written in C#, with the simulation reading in the basefile - the 2001 Census one per cent sample file - from a CSV file. A series of Microsoft Excel® spreadsheets hold the parameters for the equations and benchmarking/alignment data if that capacity is to be used.

Figure 2 The APPSIM Framework



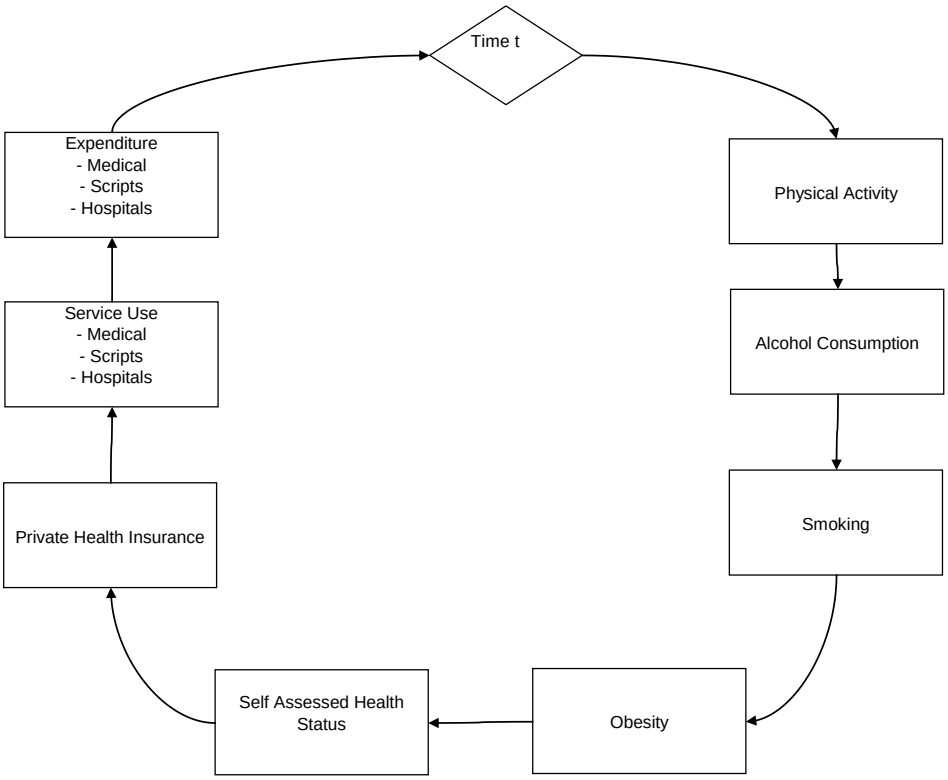
Source: Kelly 2011

3 THE HEALTH MODULE – HOW OBESITY WAS MODELLED.

To facilitate the illustration of some of the validation undertaken in the health module of APPSIM, the obesity outcome will be the focus of this analysis. The health system is quite complex, being made up of 8 sub-modules that are processed sequentially (as illustrated in Figure 2). Obesity was one of the health risk factors being modelled within the health module. It was modelled as a binary variable, characterising an individual as being obese or not. A persons was defined as being obese if their body mass index[1](BMI) was greater than 30 (a recognised cut-point for increased health risk)(WHO 2011). Obesity was initially imputed onto the base file using a logistic regression model to determine the probabilities to inform the monte carlo simulation. The probabilities were based on inputs of age, education, marital status, labour force status, household income, number of adults in the household, number of children in the household, physical activity, smoking status and alcohol consumption. Separate models were developed for men and women. Transition probabilities for obesity were calculated using a pooled dynamic logistic regression model. For the transition equation the input parameters for the calculation of probabilities were simplified to: age, couple, bachelor degree or higher, health status, smoking status, physical activity status, alcohol consumption and obesity in the previous time period. The co-efficients for the input parameters of these equations were estimated from the HILDA survey, wave 6 for the baseline and waves 6 to 8 for the transition equation. HILDA is a nationally representative survey that used complex multistage sampling to choose the households interviewed. Approximately 8000 individuals have responded to the three waves of interest (wave 6 – wave 8) in this analysis. Obesity is determined using self-reported data from the survey, which means that the estimates of obesity are highly likely to be underestimates. Research has shown that individuals tend to over estimate how tall they are and under estimate their weight (Taylor et al. 2006).

[1] Body mass index is equal to weight in kg/(height in metres) squared.

Figure 3 Health sub-module framework



4 VALIDATION PROCESS IN HEALTH MODULE

The validation process for the health module of APPSIM has followed some of the key methods mentioned in the introduction. Sensitivity analysis based on manually manipulating the input parameters was done. The sensitivity analysis presented in this paper focused on the effect the variation of physical activity and education, each considered separately, had on obesity.

Comparative validation methods were undertaken. Firstly, cross-sectional comparisons have been done. The simulation runs from 2001 (the base year) out to 2051. There is actual data for the years 2001 through to 2008 for some variables, with which the models output's can be compared. In the case of obesity, there was only data for HILDA for the years 2006 to 2008. In addition, the ABS National Health Surveys provided estimates of obesity for the years 2001 and 2005 which can be used for comparison. Longitudinal aspects of the model have also been considered, both from the perspective of the trend of the data over time and whether this seems reasonable compared to alternative projections. Further, persistence of obesity between years was calculated and compared between the actual data and the simulation model.

In the process of doing the more standard validation of the health module, it became evident that the comparative approach to validation, which was to have made up most of the validation plan, was not enough to look at the predictive abilities of the statistical

models developed for the health module. In light of the rest of APPSIM being developed at the same time as the development of the health module, there were potential discrepancies in the other APPSIM modules that were effecting the health module outputs. Part of the validation process was the untangling of issues between the prototype APPSIM module output, which provided input parameters such as education, labour force status and income, but were not a reflection of the distributions found in the source data and poor performance of the health equations. The input parameters from the prototype APPSIM needed to be verified that they were providing a reasonable representation of the Australian population in the years between 2001 and 2009. As the prototype APPSIM proved to be a poor reflection of some aspects of the Australian population, with respect to the input parameters, this had untoward effects on the health variables being imputed onto the APPSIM basefile. In turn, due to the high levels of persistence seen in the health variables, any misrepresentation of the health variables in the initial imputation of these variables onto the basefile can cause ongoing problems throughout the projection. If the input parameters, for the transition of health variables, obtained from other modules of APPSIM continue to be a poor representation of the expected experience, this will affect the health variable outcomes in an adverse manner.

Usually, in dynamic microsimulation, one method to overcome the issues of the interactions of the health module with other modules of APPSIM, is to use alignment methods for the other modules. That keeps those modules outcomes at the same level and at a consistent and expected distribution. Alignment is not available on the prototype APPSIM for all the modules and the relevant variables that were contributing inputs to the health module. Consequently, additional methods were used to test the quality of the equations used to calculate probabilities. The equations used in the health module – both the baseline imputation equations and the transition equations - were trialled on the HILDA basefile, which gives consistent inputs on which to trial the equations. The 2006 wave of HILDA was used as the starting population upon which the health baseline imputation equations were applied. This year was chosen because it provided all the socio-economic variables that were input variables for the health module. Also, this was the first year that the all the health risk factors, health status and private health insurance estimators were available on HILDA. Having these variables available allowed review of the quality of the predictions of the statistical models being used in the health module for the health risk factors, health status and private health insurance, without interference of any less than perfect input parameters.

In validation protocols it is best to test equations against an alternative dataset (hold-out sample) rather than using the data on which the statistically models were estimated. This helps in the confirmation of the generalisability of the model that has been developed. In the case of obesity, the luxury of using an alternative dataset was not possible as this information was only available for three waves of HILDA data. Consequently, only a small sample was available to consider transitions between obesity states. The implication of this is an over optimistic view of the quality of the predictions for obesity status. For the testing of the transition equations developed for the health module, the HILDA dataset from years 2007 and 2008 were used.

To date calculation of confidence intervals or estimation of monte carlo error have not been completed. These are areas be scheduled for further research.

5 EXAMPLE VALIDATION RESULTS FOR OBESITY MODELLING

5.1 OBESITY PROBABILITIES

The first aspect of validation considered is comparison of some of the key health outcomes and their probability distributions. The obesity modelling probability distributions were graphed (see Figures 4 and 5). For males, the probabilities ranged between 0.00 and 0.68, whilst for females the range was between 0.03 and 0.77. In both cases the probabilities are skewed with many individuals having low probabilities of being obese but there being a tail into the higher probabilities.

Figure 4 Estimated probability of being obese, Males 2006

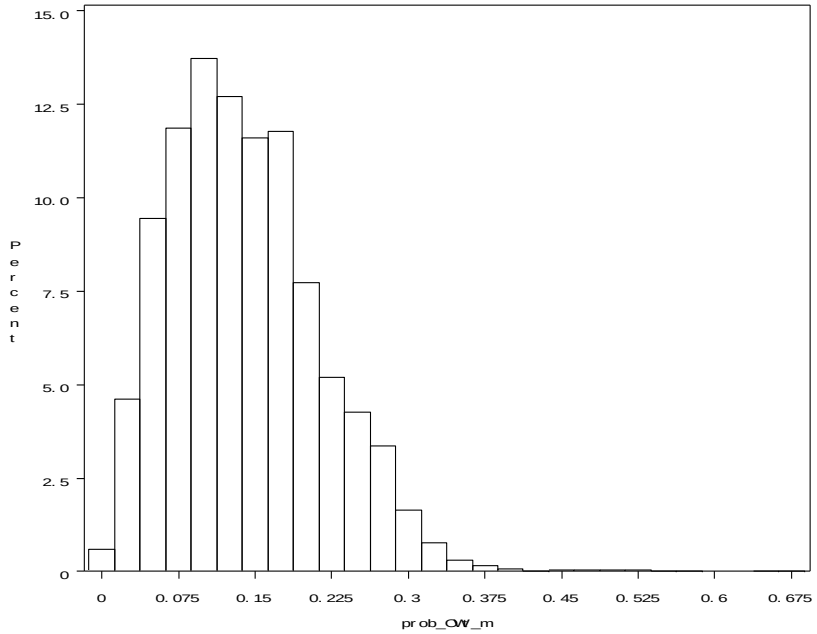
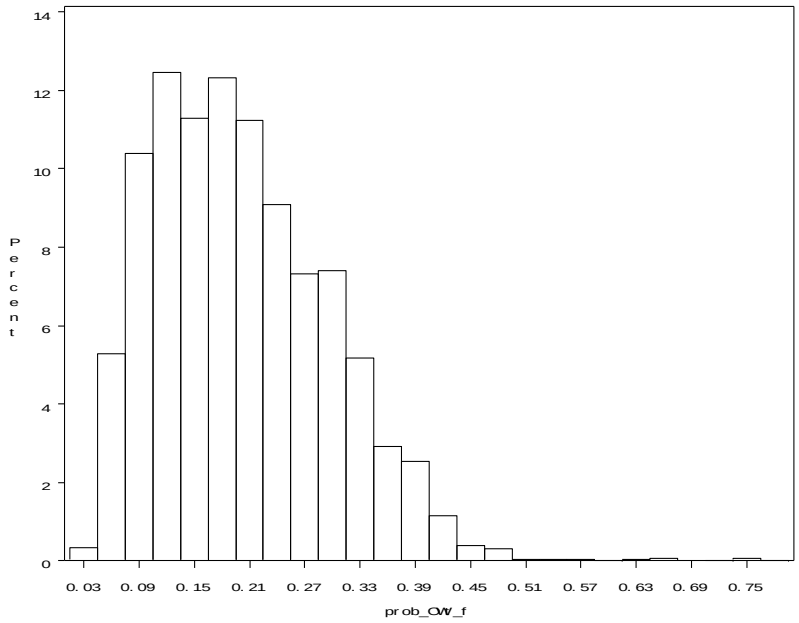


Figure 5 Estimated probability of being obese, Females 2006



The distribution of obesity transition probabilities are shown in Figures 6 and 7. A bimodal distribution is obvious. The probabilities are being split based on the individuals who were obese in the previous time period and those who were not obese. For males the transition probabilities for being obese ranged from 0.00 to 0.89. Female transition probabilities ranged from 0.00 to 0.92 chance of being obese in the time period.

Figure 6 Estimated transition probability of being obese, Males 2007

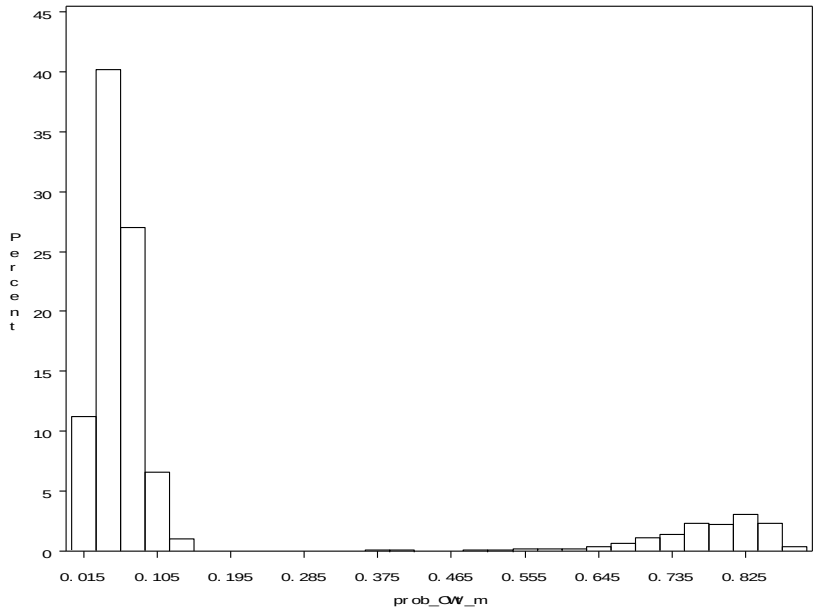
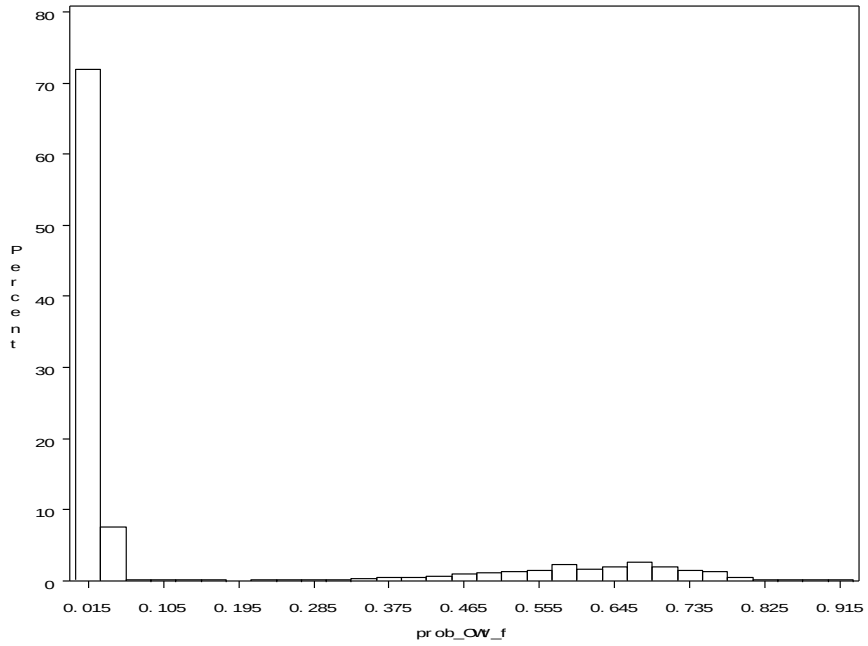


Figure 7 Estimated transition probability of being obese, Females 2007



5.2 COMPARATIVE VALIDATION

For the obesity output to have face validity it needs to show an increasing trend over time as has been the case over the previous 20 to 30 years. The initial model output (shown in Figure 8) did not show the trends of increasing obesity levels over time that would be expected. Revision of the model to include a time component as well as the socio-economic and the individual's obesity history resulted in a more intuitive outcome (see Figure 9). That is, over the time period there is an increase in the prevalence of obesity from 21 per cent in 2011 to approximately 30 per cent in 2041. This is still a very conservative outlook compared to other projections regarding obesity in Australian adults over the next 20 years. Included on Figure 9 is a projection point (the black square) of obesity prevalence for 2025 of 34 per cent of adults obese put forward by Wall et al(2011) based on dynamic life tables. This provides just one alternative view point (not a benchmark of where APPSIM output should be hitting) of where obesity may progress in the near future. There projections used information from the AUSDIAB data (Barr et al. 2006). AUSDIAB was a survey run in Australia in 2000 and 2005 that followed up the same individuals to look at the progression of diabetes in the community and its associated risk factors. As part of the survey they measured height and weight allowing consideration of measured BMI over a 5-year time period. The dots on Figure 9 are a representation of the HILDA estimates of obesity prevalence which are slowly increasing over the three years from 20 to 21 per cent. Further, the APPSIM obesity simulation has an issue with the run of the first four years of the simulation from which it stabilises and then shows a trend of increasing obesity rates over time. It is believed that this may be a result of the socio-economic inputs from other modules of APPSIM that are currently being revised.

Figure 8 Simulated per cent obese (no time element), 2001-2041

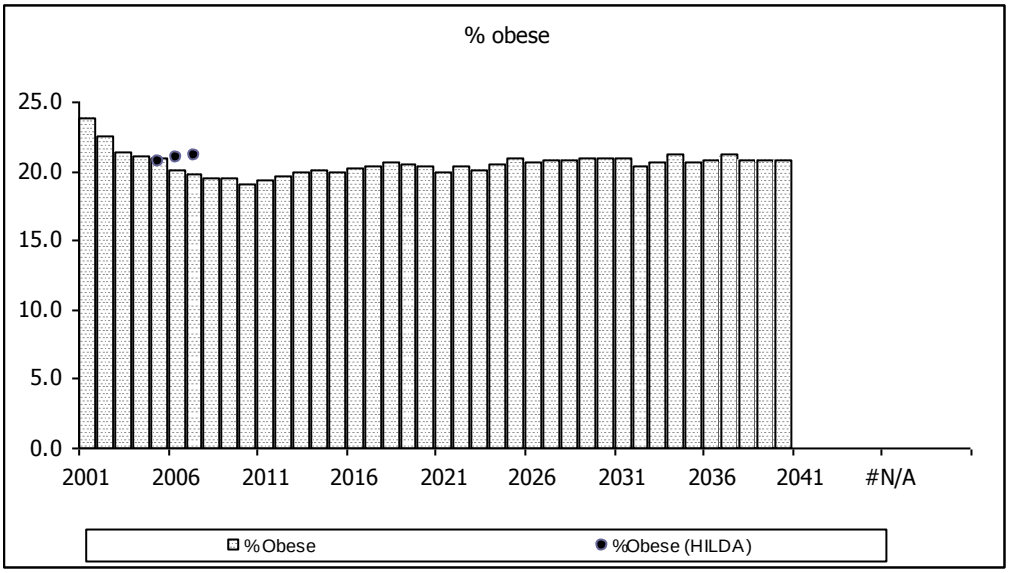
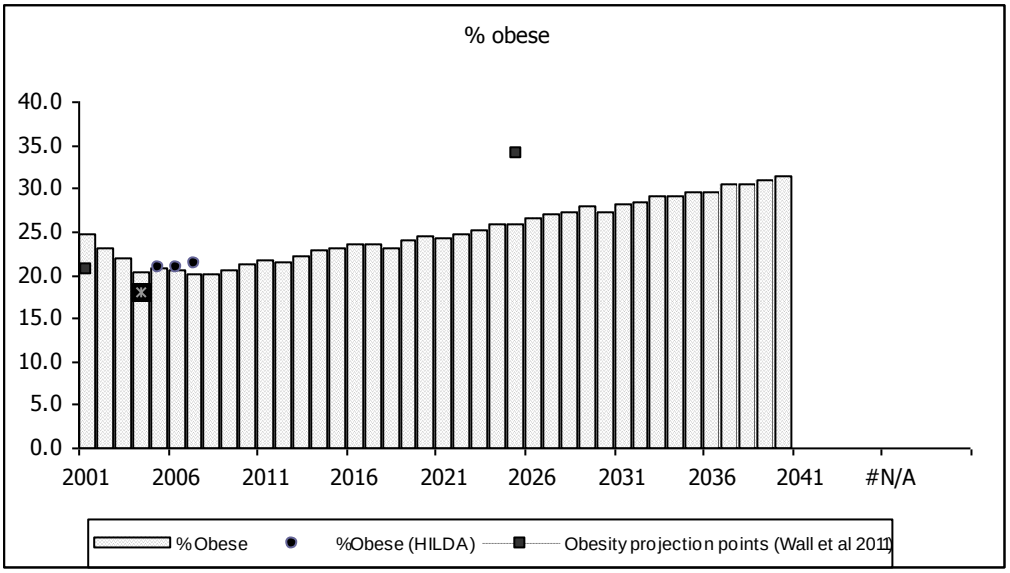


Figure 9 Simulated per cent obese (time element), 2001-2041

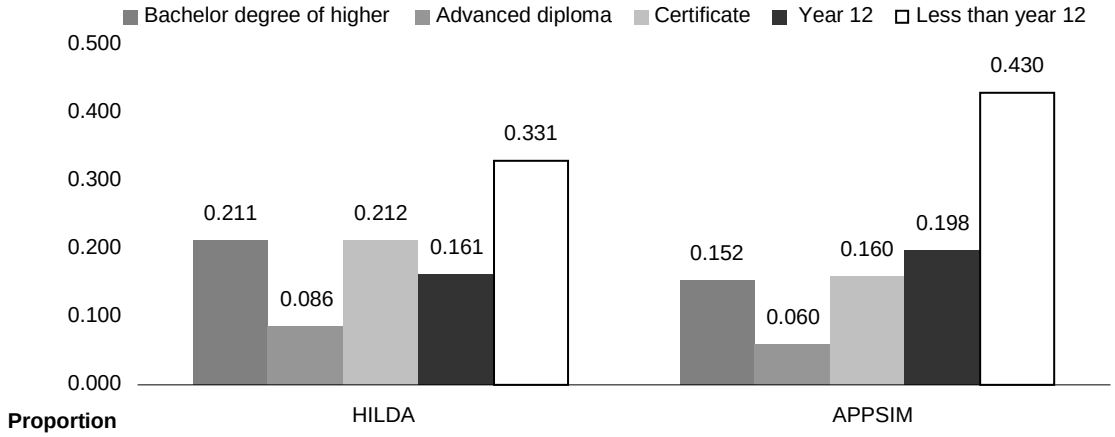


5.3 INPUTS TO THE HEALTH MODULE – COMPARISON WITH BENCHMARK DATA

To understand the processes underlying the obesity outcomes from APPSIM, the distribution of the input parameters that were external to the health module where plotted. Poor performance in these APPSIM outputs has the potential to impact on the health module through poor simulation results even if the transition equations within the health module can be shown to be reasonable by other means. In the series of Figure 10 to 13, the HILDA distributions of some key input parameters for the obesity transition

equation are compared with the APPSIM[2] outputs. The figures presented show the areas where APPSIM had performed less than adequately. For the education parameter, which was concerned with the highest qualification achieved, APPSIM was allocating substantially greater proportion of persons to have less than year 12 education, whilst not allocating enough to the higher education categories. With respect to the percentage married, APPSIM progressively performed worse moving forward from the starting year, such that by 2008 there were 51 per cent allocated to be married on APPSIM, but on HILDA there were actually 61 per cent who were married. APPSIM has a substantial over-estimation of the proportion of persons in a lone adult household and under estimation of persons in households with three or more adults. In 2008, APPSIM allocated 35 per cent of the population to be in single adult households whilst on HILDA there were estimated to be just over 10 per cent of the population in single adult households. Finally, the income distribution is markedly different between HILDA and APPSIM. APPSIM has a greater proportion of persons in the lower weekly income levels and substantially less in the high income levels compared to HILDA.

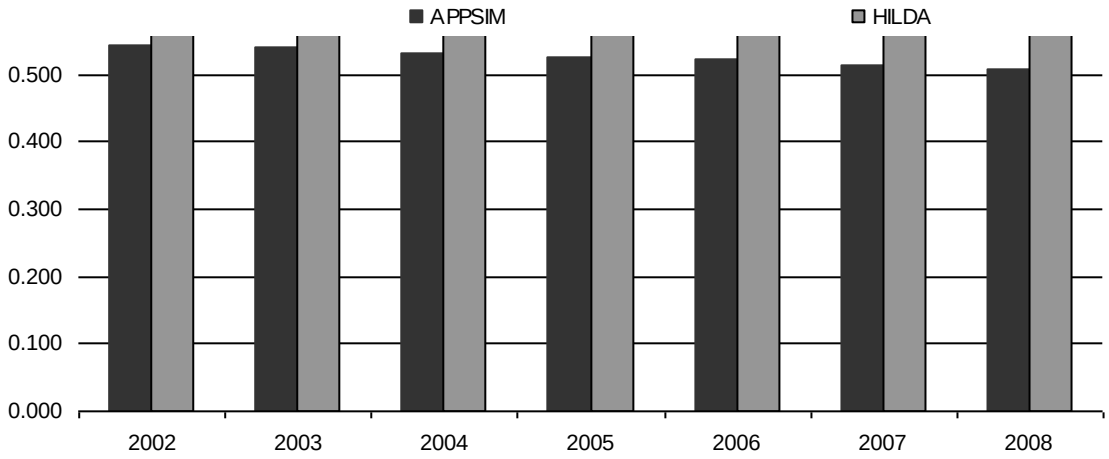
Figure 10 Highest Qualification Distribution, 2008



Source: HILDA version 8, APPSIM March 2011

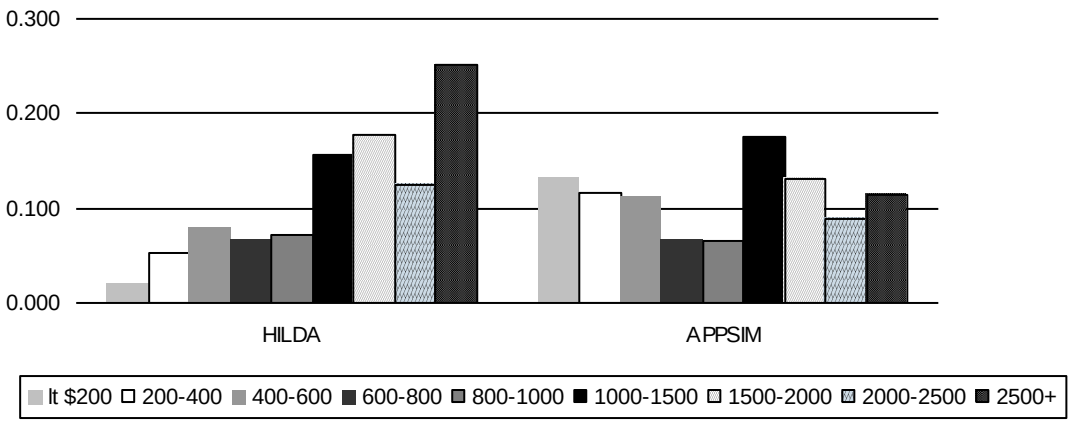
[2] This analysis is based on the March 2011 version of APPSIM. Subsequently there has been a new release of the APPSIM model, which may perform differently with respect to these parameters.

Figure 11 Proportion Married on Not, 2002-2008



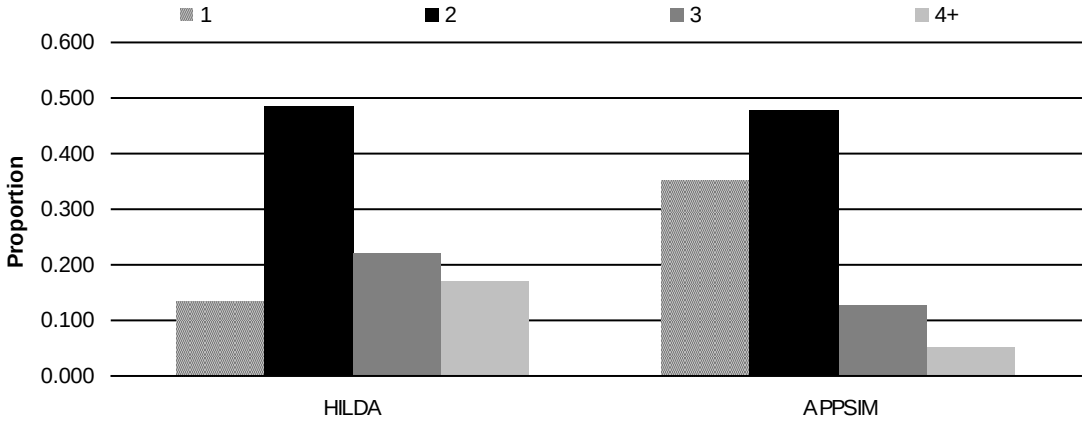
Source: HILDA version 8, APPSIM March 2011

Figure 12 Weekly Household Income Distribution ($Au), 2008



Source: HILDA version 8, APPSIM March 2011

Figure 13 Distribution of Number of Adults in the Household, 2008

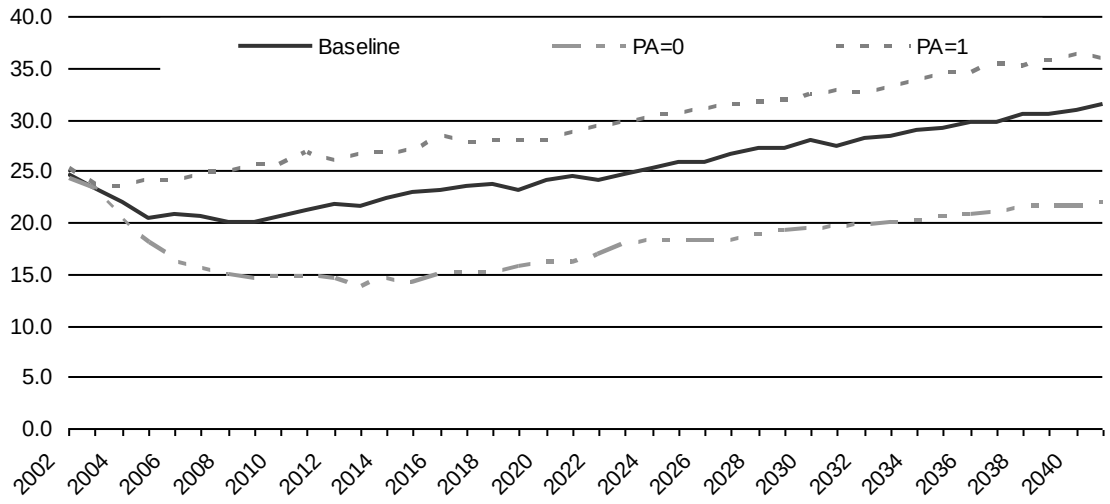


Source: HILDA version 8, APPSIM March 2011

5.4 SENSITIVITY ANALYSIS

The next aspect of validation was some sensitivity analysis around the impact of changed physical activity levels on obesity. Two simulations were run at the extremes of physical activity inputs for the obesity module. Firstly, physical activity levels were set to having all the population not having adequate activity and the second simulation set all the population to having adequate activity. The impacts of these changes on obesity are shown in Figure 14. The change in obesity with the changing physical activity levels was consistent with expectation. An increase in physical activity levels within the population decreased obesity. Conversely, a decrease in physical activity levels resulted in an increase in obesity levels.

Figure 14 Simulated per cent obese – Physical Activity scenarios, 2001-2041



5.5 PREDICTIVE ABILITY

Finally, the review of the obesity equations using confusion tables, their summary measures and ROC curves are presented. The baseline and transition simulations achieve approximately 70 per cent accuracy, which is better than a pure random selection but not as good as allocating the most likely outcome which would provide approximately 80 per cent accuracy. Due to the higher probability of not being obese (approximately 80 per cent) the equations provide better distinction for that outcome with a specificity of almost 0.80 and a negative predictive value of 0.96 (see tables 2-4 for the details of the predictive outcomes). The results were similar for the baseline imputation equation and the transition equations. Having these quantitative allows for comparison of alternative equations that may be implemented in the microsimulation model.

Table 2 Baseline Obesity Equation Confusion Table (2006)

		Simulated (APPSIM Equations)		
		Negative	Positive	Total
Actual (HILDA)	Negative	0.66	0.17	0.83
	Positive	0.13	0.04	0.17
	Total	0.79	0.21	1.00

Note: n=13,683,935

Source: APPSIM March 2011; HILDA version 8

Table 3 Transition Obesity Equation Confusion Table (2006-7)

		Simulated (APPSIM Equation)		
		Negative	Positive	Total
Actual (HILDA)	Negative	0.68	0.18	0.86
	Positive	0.11	0.03	0.14
	Total	0.79	0.21	1.00

Note: n=12,834,853

Source: APPSIM March 2011; HILDA version 8

Table 4 Summary Statistics of Obesity Equations Predictive Ability

Qualitative Measure	Baseline Imputation	Transition
Accuracy	0.7028	0.7127
Sensitivity	0.2370	0.2366
Specificity	0.7990	0.7905
Positive Predictive Value[a]	0.2056	0.1987
Negative Predictive Value[a]	0.9627	0.9717

a Prevalence of obesity assumed to be 0.18

Source: APPSIM March 2011; HILDA version 8

The visual summary of these quantitative measures is presented in the ROC curve (see Figure 15 and Figure 16). The ROC plot in Figure 15 indicates that the current baseline modelling of obesity performs better than random chance, since the curve is above the 45 degree line. However, the transition equation does not perform as well as the baseline transition with its ROC curve only being just above the 45 degree line. The transition equation is still performing better than random chance, but only marginally (see Figure 16). This outcome may in part be due to the low probabilities of a person's obesity status changing – only 3 per cent of the population change from normal weight to obese from year to year.

Figure 15 ROC Curve baseline Obesity Imputation

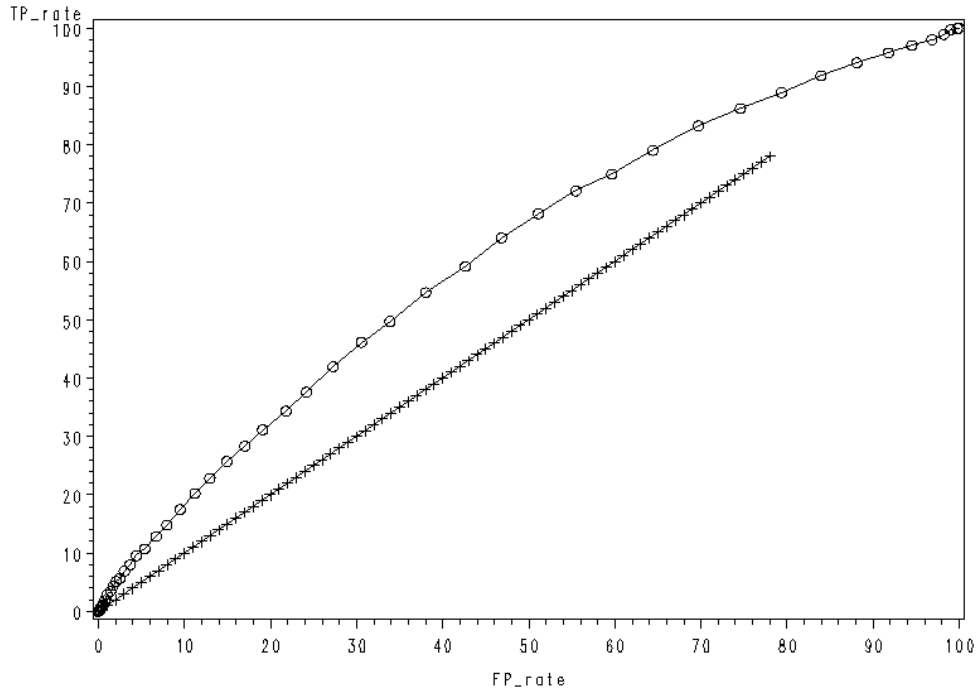
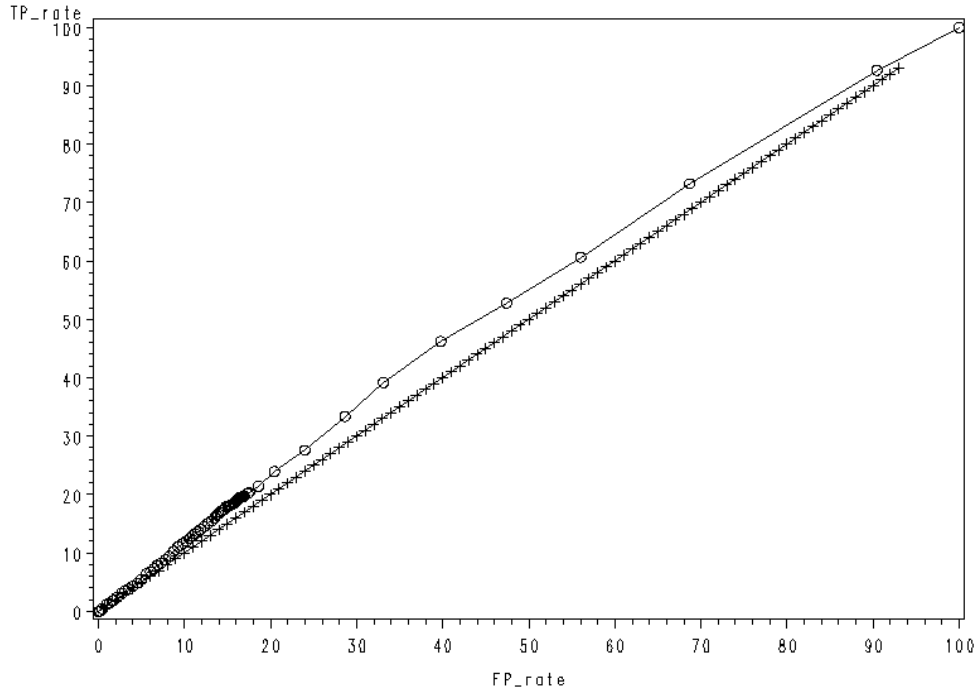


Figure 16 ROC Curve Obesity Transition



The final consideration was the rate of change to gain an understanding of how the model transitioned from year to year. The rate of change between years 2006 and 2007 was calculated. From the simulation, 6 per cent of the population changed to having normal weight from being obese, 3 per cent become obese, 11 per cent stayed obese and 80 per cent remain with normal weight. On HILDA, between 2006 and 2007 there was 4 per cent who

became obese, 4 per cent who became normal weight, 75 per cent who remained in the normal weight range and 17 per cent who stayed in the obese category. The simulation is moving slightly too many into normal weight and not enough to being obese, in comparison to HILDA estimates. Of note, the 3 per cent transition from normal weight to obesity was slightly lower than the results from AUSDIAB (Walls et al. 2011)

6 DISCUSSION/CONCLUSION

In a complex model such as APPSIM the validation process is quite onerous but vital in understanding the working of the model and providing user confidence in the outputs from the simulation model. Each sub-module within each module requires testing. The importance of automating the process when dealing with many modules which have multiple outcomes interest becomes clear.

There are many aspects of a module that need to be considered in validation including face validity of the simulation outputs, comparisons of the simulation outputs with external data, both cross-sectional and longitudinal indicators. Further, to distinguish issues of input parameters' quality from inadequate probability equations is important when several interacting modules are being developed simultaneously. For the health module in APPSIM, predictive measures borrowed from biostatistical (and data mining) literature on model development such as specificity, sensitivity, positive predictive value and negative predictive value, have been used to try to quantify the quality of the proposed probability models. These measures provide indicators to compare whether new health probability models perform better than currently implemented ones as the APPSIM health module is further developed.

Further, the interactions between modules and sub-modules in the model need to be considered. It is difficult to validate across modules as the data required for comparisons is often not available. But processes such as sensitivity analysis can provide some indication whether at a face level movement of one outcome results in the intuitively correct movement of the other module as was illustrated with the effects of changing physical activity levels on obesity prevalence.

Longitudinal validation is also problematic, as there is only limited data available to consider the transition rates of persons from one health state to another (or continued undertaking of a particular health risk behaviour). Currently the main measure available is looking at whether the simulation model returns similar transitions between years as the original data source.

With more waves of HILDA data becoming available, it will be possible to look at testing the health probability models developed for APPSIM against a sample that has not been used in the development of the actual probability models. This has the potential to lend more strength to our confidence in the probability models that have been developed and their generalisability. Other future work in the validation of the health module of APPSIM will involve consideration of how to most effectively determine predictive intervals around the point estimates of the simulation.

This paper has shown the importance of considering alternative ways (from other areas of research) to validate and thus quantify the "accuracy" and strength of the model. At a

broader level, there still is much room for research regarding how much is enough validation to provide confidence about the module and the whole model.

REFERENCES

Altman, D.G. and Bland, J.M. 1994a, Diagnostic tests 1: sensitivity and specificity, *BMJ*, vol.308, pp. 1552.

Altman, D.G. and Bland, J.M. 1994b, Diagnostic tests 2: predictive values, *BMJ*, vol.309, pp. 102.

Altman, D.G. and Bland, J.M. 1994c, Diagnostic tests 3: receiver operating characteristic plots, *BMJ*, vol.309, pp. 188.

Barr, E.L.M., Magliano, D.J., Zimmet, P.Z., Polkinghorne, K.R., Atkins, R.C., Dunstan, D.W., Murray, S.G. and Shaw, J.E. 2006, *AusDiab 2005 The Australian Diabetes, Obesity and Lifestyle Study. Tracking the Accelerating Epidemic: Its Causes and Outcomes*, Melbourne, International Diabetes Institute, Melbourne, Australia.

Cassells, R., Harding, A. and Kelly, S. 2006, Problems and Prospects for Dynamic Microsimulation: A Review and Lessons for APPSIM, *Discussion Paper no. 63*, Canberra,NATSEM, University of Canberra

Creedy, J., Kalb, G. and Kew, H. 2007, Confidence Intervals for Policy Reforms in Behavioural Tax Microsimulation Modelling, *Bulletin of Economic Research*, vol.59, no. 1, pp. 0307-3378.

Goldman, D.P., Shekelle, P.G., Bhattacharya, J., Hurd, M., Joyce, G.F., Lakdawalla, D.N., Matsui, D.H., Newberry, S.J., Panis, C.W.A. and Shang, B. 2004, *Health Status and Medical Treatment of the Future Elderly* TR 169, Santa Monica, RAND Health for the Centers for Medicare and Medicaid Services.

Han, J. and Kamber, M. 2006, *Data mining: concepts and techniques*, 2nd edn, San Francisco, Elsevier.

Harding, A., Keegan, M. and Kelly, S. forthcoming, Validating a dynamic population microsimulation model: Recent experience in Australia, *International Journal of Microsimulation*, vol.

Kelly, S. 2011, *A Detailed Guide to APPSIM: A description of the APPSIM model, its functions and the data produced*, Canberra, NATSEM, Univeristy of Canberra.

Kopec, J.A., Fines, P., Manuel, D.G., Buckeridge, D.L., Flanagan, W.M., Oderkirk, J., Abrahamowicz, M., Harper, S., Sharif, B., Okhmatovskaia, A., Sayre, E.C., Rahman, M.M. and Wolfson, M.C. 2010, Validation of population-based disease simulation models: a review of concepts and methods, *BMC Public Health*, vol.10, no. 710.

Morrison, R. 2008, *Validation of Longitudinal Microsimulation Models: DYNACAN Practices and Plans*, Working Paper No. 8 Online working paper no. 8, Canberra, National Centre for Social and Economic Modelling, University of Canberra, March.

O'Donoghue, C. 2001, Dynamic Microsimulation: A Methodological Survey, *Brazilian Electronic Journal of Economics*, vol.4, no. 2.

Pudney, S. and Sutherland, H. 1994, How reliable are microsimulation results? : An analysis of the role of sampling error in a U.K. tax-benefit model, *Journal of Public Economics*, vol.53, no. 3, pp. 327.

Taylor, A.W., Grande, E.D., Gill, T.K., Chittleborough, C.R., Wilson, D.H., Adams, R.J., Grant, J.F., Phillips, P., Appleton, S. and Ruffin, R.E. 2006, How valid are self-reported height and weight? A comparison between CATI self-report and clinic measurements usinga large cohort study, *Australian and New Zealand Journal of Public Health*, vol.30, no. 3, pp. 238-246.

Walls, H.L., Magliano, D.J., Stevenson, C.E., Backholer, K., Mannan, H.R., Shaw, J.E. and Peeters, A. 2011 Projected Progression of the Prevalence of Obesity in Australia, *Obesity* vol (13 January 2011) DOI: doi:10.1038/oby.2010.338.,

Watson, N., (ed.). 2010, *HILDA User Manual - Release 8*, Melbourne, Melbourne Institute of Applied Economic and Social Research, University of Melbourne.

WHO 2011, Fact Sheet No. 311: Obesity and overweight, http://www.who.int/mediacentre/factsheets/fs311/en/, Accessed 5/05 2011.

Zaidi, A. and Rake, K. 2001, Dynamic Microsimulation Models: a review and some lessons for SAGE, *SAGE Discussion Paper no. 2*, London,ESRC SAGE Research Group, London School of Economics