Dalziel Kim (Orcid ID: 0000-0003-4972-8871)

ACCURACY OF PATIENT RECALL FOR SELF-REPORTED DOCTOR VISITS: IS SHORTER RECALL BETTER?

Kim Dalziel[a], Jinhu Li[b], Anthony Scott[b] and Philip Clarke[a]

[a]*Centre for Health Policy, Melbourne School of Global and Population Health, The University of Melbourne, Victoria 3053*

[b]*Melbourne Institute of Applied Economic and Social Research, Faculty of Business and Economics, The University of Melbourne, Victoria 3053*

Running head: Accuracy of patient recall for doctor visits

Keywords: recall, self-report, doctor visits, health care utilisation, memory, health surveys

Manuscript word count: 4976

Table count: 4

Figure count: 2

Corresponding Author:

Dr. Kim Dalziel

Level 4, 207 Bouverie St, Carlton 3053 VIC, Australia

Tel: +61(0)401591310 Fax:

Email: kim.dalziel@unimelb.edu.au

Funding sources: This work was supported by The University of Melbourne and The Diabetes Care Project.

Ethics: the collection of data for the Diabetes Care Project obtained ethical clearance from the human research ethics committees of the Department of Health and Ageing (Commonwealth Government), Department of Human Services (Commonwealth Government), Australian Institute of Health and Welfare (Commonwealth Government), SA Department of Health (South Australian Government), Queensland Department of Health (Queensland Government), Department of Health Victoria (Victorian Government), and the Aboriginal Health Research Ethics Committee (Aboriginal Health Council of South Australia).

Declarations: The authors declare that this is original unpublished work that is not submitted elsewhere for publication. The authors declare that they have no conflicts of interest. The authors declare that they each meet all three conditions for authorship (1. conception, design, analysis and interpretation of data, 2. Drafting or revising article, and 3. Approval of manuscript).

Keywords: recall, self-report, doctor visits, health care utilisation, memory, health surveys

SUMMARY

In health economics, the use of patient recall of health care utilisation information is common, including in national health surveys. However, the types and magnitude of measurement error that relate to different recall periods are not well understood. This study assessed the accuracy of recalled doctor visits over 2-week, 3-month and 12-month periods by comparing self-report to routine administrative Australian Medicare data. Approximately 5,000 patients enrolled in an Australian study were pseudo-randomised using birth dates to report visits to a doctor over three separate recall periods. When comparing patient recall to visits recorded in administrative information from Medicare Australia, both bias and variance were minimised for the 12-month recall period. This may reflect telescoping that occurs with shorter recall periods (participants pulling in important events that fall outside the period). Using shorter recall periods scaled to represent longer periods is likely to bias results. There were associations between recall error and patient characteristics. The impact of recall error is demonstrated with a cost-effectiveness analysis using costs of doctor visits and a regression example predicting number of doctor visits. The findings have important implications for

surveying health service utilisation for use in economic evaluation, econometric analyses and routine national health surveys.

1. INTRODUCTION

Self-reported health service utilisation data collected via surveys is used in a wide range of applications for health economic, health service and population research. Self-reported methods of recalling health service utilisation are used routinely in population surveys, including national health surveys. Validating self-reported measures against those from administrative sources such as billing records is particularly important, as it is well known that recall of prior health service use is subject to both measurement error and limitations of memory, which can lead to both under- and over-reporting (Bhandari and Wagner 2006).

The use of self-reported health care use data in surveys remains an important means of capturing health service utilisation information. Administrative data are not always available, and there can be significant barriers to access. A review of 90 household surveys reported markedly different practices in regard to recall periods used for health service utilisation. For example, 53% of surveys used 1-month recall periods for physician visits, compared to 9% that used 3-month recall and 27% that used 12-month recall (Heijink et al., 2011). In regard to specific surveys, at one extreme are surveys such as the Australian Health Survey (ABS, 2013) and older versions of the United States National Health Interview Survey (Edwards et al, 1996), which use a two-week recall period for visits to a doctor. The Survey for England (HSCI, 2013) and the National Health Interview Survey United States, after its redesign in 1997 (CDC, 2017), use a 12-month recall. Given that such surveys are often used to answer similar types of research questions, it is unclear why greater efforts are not made to understand the implications of chosen recall periods on health economic research.

There are two main errors that occur with memory; the first is forgetting relevant information (omission), and the second is including events that did not occur (commission). Telescoping refers to the recall of events falling outside of the recall period, leading to over-estimation (Rubin and Baddeley, 1989; Sudman and Bradburn, 1973). Telescoping is particularly problematic for frequently occurring events and shorter recall periods, and it is thought to occur due to confusion about dates and the keenness of participants to please the interviewer or researchers (Rubin and Baddeley, 1989; Sudman and Bradburn, 1973). Longer periods of recall may be more susceptible to errors of omission.

There has been relatively little research assessing different recall lengths for self-reported health care utilisation. Bhandari and Wagner (2006) reviewed 13 studies that assessed the accuracy of a single recall period for doctor visits; the recall periods ranged in length from 1 month to 18 months (Bhandari and Wagner, 2006). They found under-reporting ranging from 5% to 68%, agreement ranging from 12% to 67% and over-reporting ranging from 14% to 40%. In general, they found that under-reporting was greater than over-reporting for longer recall periods (e.g., 12 months), and the opposite was true for shorter recall (e.g., <=3 months). Importantly, none of the studies in this review involved any type of randomised experimental design to compare different recall periods. This makes it hard to draw overall conclusions, as there are many population characteristics that can confound an analysis of the impact of the recall period on reporting error.

Experimental research that directly compares different recall periods and investigates distinct types of error and bias across time frames for recall of health care is limited, and we are aware of only two studies of this topic. One experiment, involving nearly 7000 Swedish survey respondents, directly compared 1-month, 3-month, 6-month and 12-month self-reported hospitalization recall and found that shorter periods appeared prone to telescoping, which imparted an upward bias on reported use (Kjellsson et al., 2014). Recently, a German study compared recall for two periods (3 and 6 months) by randomising 432 people with diabetes and asking them to recall physician visits reported in administrative insurance data. They found that while a 3-month recall period produced more accurate results, there was a higher proportion of respondents who over-reported and a lower proportion of respondents who under-reported when compared to the 6-month reporting period (Icks et.al, 2017).

The aim of this research is to use a pseudo-randomised experiment, conducted as part of a large diabetes study, to directly compare recall periods ranging from two weeks to one year. A further aim was to examine how individual characteristics such as age, ethnicity, income and health are related to recall error for doctor visits.

2. METHODS

The Diabetes Care Project (DCP) was a large Australian multicentre cluster randomised controlled trial conducted in general practice to assess the impact of new coordinated care interventions for patients with diabetes (Leach *et al.*, 2013; Fountaine and Bennett 2016). To assess the accuracy of three different recall periods, the following question was added to the final follow-up questionnaire:

> "How many visits have you made to a doctor (general practitioner GP
> or specialist) in the last X/weeks or months?"

Study participants were pseudo-randomised to a recall period by asking them to answer the question over a 2-week period ($w = 14$) if their date of birth (DOB) fell between the 1st and the 10th of the month, to answer over a 3-month period ($w = 90$) if their DOB fell between the 11th and 20th of the month and to answer over a 12-month period ($w = 365$) if their DOB fell between the 21st and 31st of the month (where w is defined as the recall period in days). The wording of the question presented all three recall periods but asked them to answer only the question corresponding to their day of birth and to indicate “today’s date” or the date they were answering the questionnaire (See Appendix 1 for question wording). This was to enable accurate matching of time period with administrative data. Participants needed to answer the survey to be randomised and included in the analysis.

2.1 Study population

Of the 6,853 participants who were enrolled in the trial (Commonwealth Department of Health 2015), 5,305 (77.4%) were randomised to one of the three recall groups by completing the final questionnaire (Figure 1, A). The randomisation question presented all three recall periods but directed participants to only answer for the recall period corresponding to their day of birth (see Appendix 1 for original question). A total of 4,478 (84.4%) patients completed the recall period question correctly according to randomisation by day of birth (Figure 1, B);.1,199 patients (27%) in addition to completing the correct randomised question, also incorrectly answered for one or more of the other recall periods. These additional responses did not correspond to the randomised recall period and have not been used in the analysis to preserve an intention-to-treat approach. A comparison of patient characteristics for those complying with randomisation instructions and those that did not was made to assess the potential bias of excluding these participants from the analysis (Appendix 2). There were differences between the groups which informed the use of the intention to treat analysis as the base case analysis (i.e. all participants included as per randomisation). Four observations with implausible values for self-reported doctor visits were excluded from the analysis, with implausible defined as reporting a visit on more than 75% of days in the recall period (e.g., reporting >10.5 visits within 2 weeks, reporting >67.5 visits within 3 months and reporting >273.75 within 12 months). This resulted in three observations being dropped in the 2-week recall group and one observation being dropped in the 3-month recall group (see Figure 1). The study sample thus consists of a total of 4,399 respondents, including 1,481 in the 2-week recall group, 1,490 in the 3-month recall group and 1,428 in the 12-month recall group (Figure 1, D).

INSERT FIGURE 1

The characteristics of patients randomised to recall groups are described in Table I. The only statistically significant differences among groups (p<0.05) were the proportion of patients who were newly diagnosed with diabetes within the previous 5 years and the proportion of patients who had private health insurance. These characteristics are not expected to differentially influence a patient’s ability to recall doctor visits.

INSERT TABLE I

2.2 Administrative data

Participants' self-reported responses were linked to their date of birth, enabling linkage of self-reported data to administrative registered data that were obtained from Medicare Australia (a complete Government registry of doctor visits according to billing). The specific Medicare item numbers coded as representing "doctor or specialist visits" are reported in Appendix 3. Using the variable "today's date" as recorded on the patient questionnaires, time periods were constructed for 2 weeks, 3 months and 12 months.

2.3 Analysis of recall error

In comparing the Medicare data to self-reported data, a number of continuous error measures were constructed: the number of positive errors (number of reported visits that occurred without a corresponding Medicare visit), number of negative errors (number of Medicare visits without a corresponding reported visit), total errors (number of positive errors plus number of negative errors) and recall error per day (total number of days recalled in error divided by recall period in days). The difference among recall periods in each error measure is compared using an ANOVA test and F statistic to assess statistical independence. Statistical significance is reported at the 0.05 probability level.

To explore the potential impact of telescoping, the self-reported mean results for each time period are first compared to the exact same time period of Medicare data and then compared to the Medicare data plus an additional 2-week window of recall. The proportion of patients making recall errors for their doctor visits was also reported including the condition that they had made at least one visit.

Following Clarke et al. (2008) and Kjellsson et al. (2014), we define Y_s as the actual measure of mean doctor visits in a target period, *s*, averaged over population *i,…,N*. For example, in economic evaluation, *s* is often equal to 12 months. In our context, the gold standard is the measure of doctor visits provided by Medicare data. However, in surveys, Y_s is not observed, and we require an estimate, $E(y_s)$. This estimate may be biased such that $Bias1(Y_s) = E(y_s) - Y_s$. In a survey, the time period by which the estimate is elicited may be shorter or equal to the period *s*, referred to as period *w*. As noted above, situations in which $w<=s$ may be due to researcher beliefs about the higher accuracy of recall in shorter periods. In this shorter period, Y_w is the actual mean, which we do not observe, and $E(y_w)$ is the observed self-reported mean from the survey. $Y_w = E(y_w)$ when there is perfect recall. When *w* is shorter than *s*, researchers need to scale up $E(y_w)$ to provide an estimate of Y_s, so that $E(y_s) =E(y_w)(s/w)$. This process produces two types of errors that cause $E(y_s)$ to be biased and not equal Y_s. The first is recall error when $E(y_w)$ may not equal Y_w because of recall bias. The second error is scaling error. If s/w=2 such that there are 2 time periods in S (e.g., 2x6 month periods if s=12 months) then scaling up assumes $E(y_{w1t}) = Y_{w1} = Y_{w2}$, which may not be the case. The longer the recall period, the lower this error and the lower the overall bias. To account for this, a second measure of bias is $Bias2(Y_s) = E(y_s) - (s/w)Y_w$.

A further issue is that as the recall period increases, *w* approaches *s*, and more information on additional visits is provided, so the variance of our estimate (y_s) decreases according to a quadratic loss function. Clarke et al. (2008) use the root mean square error (RMSE) to combine the bias and variance of the estimate into a single measure:

$$RMSE(y_s) = \sqrt{[Bias(y_s)]^2 + Var(y_s)}$$

The optimal recall period, w, is then chosen to minimise the RMSE. We calculate two sets of RMSE, variance and bias, that differ according to the inclusion of seasonality in the calculation of bias.

3. RESULTS

As shown in Table I, there was an even distribution of patients across the recall period groups. The groups showed a similar level of Medicare doctor visits across a 12-month period, with the average number of doctor visits per year being 13.023 (SD 9.585) and no statistically significant difference among groups. Table II describes the mean registered doctor visits in the Medicare data, the mean reported doctor visits, and the error measures across the three recall periods. Of the whole sample, 95.23% reported a doctor visit during the previous 12 months and 38.46% within the last 2 weeks.

Table II reports mean self-reported versus Medicare visits and an additional value capturing visits over the recall period, plus an additional 2-week window of time to represent telescoping. The period reflecting telescoping produced mean visits more similar to self-reported visits for the 2-week and 3-month periods. These figures show that a shorter recall period (2-week) is more likely to incur error from telescoping.

There were a number of patients with no doctor visits (25.3% overall), and this proportion was predictably greater for shorter recall periods.

INSERT TABLE II

3.1 Proportion of patients reporting errors and type of error

The majority of patients (71.2%) made an error in recalling their visits to the doctor, and this percentage increased when the length of the recall period increased, with the proportion minimized in the 2-week recall group (Table II). This reflects the smaller number of visits to be recalled. Interestingly, the proportion was not improved when those without visits in the corresponding recall periods were dropped, meaning that error is reported conditional on having at least one visit. Similarly, the total number of days reported in error was greatest for the 12-month recall group and was similar for the other two recall periods.

Figure 2 further breaks down the proportion of patients with any error into those with positive and negative errors in each recall group. Overall, the longer recall period was associated with a greater number of both positive and negative errors. The pattern of errors, however, differed with length of recall. Patients recalling shorter recall periods were more likely to make positive errors (over-reporting) compared to negative errors, whereas for patients recalling longer recall periods, there was a greater proportion of negative errors (under-reporting) relative to positive errors.

INSERT FIGURE 2

3.2 Size of errors and relative error

The number of visits reported in error was highest in the 12-month group (Table II). The total average size of error was an additional 1.407 visits (SD 4.192) for positive errors and -1.584

visits (SD 3.817) for negative errors. Longer recall periods increase the average errors for both types. The F-statistics indicate that the differences in all types of errors are statistically significant across the three recall periods.

When relative error was computed as the total visits recalled in error divided by the total number of days recalled, results showed the relative importance of a day recalled in error. The relative impact of errors in the recall of doctor visits is smallest for the 12-month recall period and greatest for the 2-week recall period. This difference was statistically significant across the three recall periods.

3.3 Evaluating optimal recall length: whole sample and subgroups

Table III reports the RMSE, variance and bias for each recall period. Both variance and bias decrease as the length of recall increases. The RMSE, based on both definitions of bias, is more favourable (i.e., minimized) for the 12-month recall period. It is expected that variance decreases with longer recall periods, as the amount of information increases with the length of recall. However, the fact that bias also decreases with longer recall periods shows that the expected trade-off between variance and bias is not present here. As the recall period increases, the respondents who under-report are more likely to balance out the respondents who over-report, resulting in a smaller overall mean bias for the longer recall period (12-month).

INSERT TABLE III

INSERT TABLE IV

Measurement error, variance and bias are presented separately for subgroups of the sample according to patient characteristics deemed to potentially influence error. Subgroups included current age greater or less than 60 years, length of diagnosis greater or less than 5 years, education greater or less than year 12, employment status, income (greater or less than $20,000 per year), and patient risk complexity. Table IV presents the bias, variance and the corresponding RMSE (based on the second definition of bias only) across subgroups of respondents with different individual or socio-economic characteristics. Comparing the RMSE within each recall period, the results show that older, less educated, unemployed, and low-income patients tend to have higher measurement errors in all recall periods. The RMSE is smallest for the 12-month recall period for all subgroups, except for subgroup of patients with fewer than 5 years of diagnosis. Like the RMSE measure, bias also decreases as the length of recall increases, with the measure minimized for the 12-month recall period for all subgroups except for the subgroup of patients with fewer than 5 years of diagnosis. This indicates that a longer recall period is preferred, even considering the potential heterogeneity in the measurement error related to patient characteristics.

3.4 Practical applications

The practical impact of the recall bias identified across the three periods is tested using two example analyses. First, an evaluation of cost inputs for an economic evaluation was performed, where the costs associated with the recall of doctor visits are compared to those generated using Medicare visits. The results showed that the costs of doctor visits were underestimated for the scaled 12-month recall group by 14.0% for the intervention group and by 15.7% for the control group. Costs were overestimated for 2-week and 3-month recall periods for both the intervention (116.9% and 16.6%) and control (93.6% and 19.6%) groups compared to 18 months of registered visits. Further details are presented in Appendix 4.

Second, a regression analysis was conducted, demonstrating the relationship between primary care utilisation (number of doctor visits) and a range of patient and clinical covariates in the diabetic population. The analysis based on recalled doctor visits is compared to an identical analysis using Medicare visits. The results showed that the use of recall data is associated with coefficients that change sign (2 weeks n=5, 3 months n=2 and 12 months n=2) and with statistical significance at p=0.1 (2 weeks n=3, 3 months n=10 and 12 months n=7). There was little consistency across individual covariates in the three analyses. Further details are contained in Appendix 4.

4. DISCUSSION

This study aimed to assess the recall error in self-reported doctor visits with recall periods of varying lengths; it also assessed the optimal recall length for an aggregated mean of doctor visits using evidence from a pseudo-randomised experiment. This involved assigning people, based on birth date, to three different recall periods within the same study, and responses were linked to administrative data reflecting registered doctor visits. This research has important implications for surveying self-reported doctor visits, whether through population household surveys or questionnaires used to inform economic evaluation and other health economics research.

Our research demonstrates that it is important to carefully consider the intended use of the data in order to understand the impact of recall errors and the resulting bias. In terms of accuracy, short recall periods such as 2 weeks are associated with a lower proportion of errors but are not free of error, as there is evidence of telescoping, with respondents reporting twice as many visits than actually occurred (i.e., mean of 1.057 reported vs. 0.549 actual). The relative error likewise showed that the number of visits recalled in error per day recalled was smallest for the longest (12-month) recall period. Such recall bias is also particularly problematic when a short-period recall figure is scaled up to reflect an aggregated mean of doctor visits for longer periods such as one year. Forward telescoping is likely to be an issue for recall of doctor visits over short periods, with evidence indicating that when recalling a 2-week period, patients include an additional 2 weeks of data. Across both measures of bias and variance, the 12-month recall period performed best.

In regard to comparison of our results with other experiments, short recall was associated with over reporting and longer recall was associated with under reporting, which is consistent with a previous experiment involving recall of hospital use (Kjellsson et al., 2014). Our results are not consistent with the reporting of mean visits in the Icks et al. (2017) study, as they show significant underreporting at 3 months (2.3 visits reported vs. 2.8 actual), while our results for the same recall period indicate over-reporting (3.8 vs 3.2), but they do also show that shorter recall periods produce more over-reporting. More generally, we find that the overall proportion of patients who made any error increases as the length of the recall period increases, and the proportion of patients making errors is minimized in the 2-week recall group. This reflects the smaller number of visits to be recalled, which is consistent with other studies examining the difference in absolute accuracy of self-reported health care utilization in single periods (Bhandari and Wagner 2006; Kjellsson *et al.*, 2014). The effect was not significantly modified when we only observed errors conditional on at least one visit, indicating a similar saliency for a lack of visit and a visit.

Consistent with other literature, we found that older patients, less educated patients, unemployed patients, and patients with low incomes tend to have higher measurement errors in all recall groups (Sudman and Bradburn, 1973; Das *et al.*, 2012; Kjellsson *et al.*, 2014). Unemployment and low-income status may be proxies for poorer health status, which is often argued to increase the probability of recall error. There was, however, only one subgroup for which bias was not minimised in the 12-month period: those with diabetes diagnosed fewer than 5 years previously. For this subgroup, the 3-month recall period was optimal. The difference may relate to this group including many newly diagnosed patients who perhaps view their doctor visits as more salient and therefore recall them more accurately over shorter periods. In general, the subgroup analyses provide confidence that the results are likely to translate to heterogeneous patient groups.

Interestingly, in the review of 90 household surveys from around the world, the most common recall period for doctor visits was 1 month (used by 53% of surveys; (Heijink *et al.*, 2011). While our study involves only people with diabetes, it does provide strong evidence to lengthen recall to 1 year if the aim of these surveys is to estimate mean annual use. One such example of a short recall period is The Australian Health Survey (ABS, 2013), which uses a 2-week recall period for doctor visits. It can be recommended that the Australian Bureau of Statistics modify this to be consistent with the Health Survey for England (HSCI, 2013) and the National Health Interview Survey United States (CDC 2016), both of which use 12-month recall. This change would, however, need to be balanced against the need for consistency in reporting across time to allow for within-country analysis of trends. Replacing with 1-year recall would, however, support international comparisons and be associated with less recall error.

This study has important implications for the measurement of patient resources using survey methods. It has been argued that not enough attention has been paid to patient reported cost measurement and that researchers should now afford costing methodologies the same attention as outcome measurement (Thorn *et al.*, 2013). The Database for Resource Use Measurement (DIRUM, 2016) lists 77 instruments for the collection of patient health service utilisation. A recent review of instruments (Ridyard and Hughes, 2015) found that the most widely used is the Client Service Receipt Inventory CSRI (Beecham and Knapp, 1999), which has been used in original or modified form in over 500 studies. The original CSRI asks about self-reported doctor visits over a 3-month period, with periods such as 12 months not generally recommended (Ridyard and Hughes, 2015). Likewise, a review found 85 Health Technology Assessment-funded primary research papers, which included economic evaluation and recording of patient-level resources (Ridyard and Hughes, 2010). Disparity in methods of data collection were found, and a median 4.5 month recall period was found (interquartile range 2 to 6 months). Until now, there has been little methodological evidence to inform the choice of recall period. Validation research has been restricted to single studies of individual time periods comparing recalled data with actual data. It would be valuable to undertake more randomised experiments to provide a sound evidence base for the collection of resource-use data in the future.

Two practical applications were shown to demonstrate the use of recall data in an economic evaluation and in a regression analysis using socio-economic and clinical variables to predict use of doctor services (Appendix 4). Both examples illustrated that the error inherent in the recalled doctor visit data have the potential to alter health care decisions and to modify the significance and conclusions of empirical analyses. The use of 2-week self-reported doctor visits increased the associated cost by 116.9% for the intervention group and by 93.6% for the control group. The potential impact of this difference in an economic evaluation could be large enough, under certain conditions, to alter health care decision making. The regression analyses based on recalled rather than Medicare doctor visits led to changes in not only the sign of coefficients (such as the income coefficient) but also the statistical significance of a number of covariates. The error in the recall data therefore has the potential to alter conclusions regarding relationships among socio-economic variables, clinical variables and health care utilisation. Given the worldwide reliance on surveys for health economic research, further work exploring how to avoid problems through improved data collection, or how to address these problems through econometric techniques, is warranted. In regard to the former, a fruitful avenue for research would be to see if the accuracy of the reporting of past use could be improved by changing the wording of the questionnaires. Again, randomised experiments are likely to provide the best path forward.

There are some limitations to the current analysis. Whilst the Medicare data are an accurate record of doctor visits billed through Australia's universal health system, there is a lag until billing data are uploaded, which may have resulted in a small number of very recent consults being missing. There are also a small number of patients who may have accessed private doctor consultations, which, in the Australian health system, would only apply to specialist visits. This is, however, unlikely to be a large concern as a purpose of the DCP study was to facilitate access to public services. The other limitation is in the coding of visits as ‘doctor visits’. All professional attendance items available for billing in Australia were thoroughly assessed to determine whether they constituted a visit to a doctor, but there may have been errors in general practice bill coding. For ‘team care planning’ and ‘coordinated care’ types of consultations, we assumed that a patient was only permitted one visit per day to prevent multiple providers for the same visit being coded as multiple visits. It is, however, possible that more than one visit was made per day, and this may have been missed. These limitations relate to underestimation in the registry data, which would serve to further strengthen the research findings. The results are likely robust to these limitations. A further limitation is that whilst data support the notion of telescoping for short recall periods, it is impossible to understand patient thought processes without qualitative data.

The limitations involved when scaling self-reported data are only partially addressed by this research. We highlight that the errors associated with different periods of recall will have a differential effect on bias when scaled. There are, however, at least two different components of scaling error, one of which is the impact of a uniform and predictable error that is magnified through scaling, and the second is the impact of a non-uniform pattern in the data when a subset of data are scaled. This non-uniform pattern may occur for reasons such as a seasonal effect on health service use. This is an important note, but its impact is not directly addressed by this experiment. Further work is required to test for the impact of bias arising when there are non-uniform properties of data that are scaled.

Scaling is only one limitation when self-reported health utilisation data are employed in economic evaluation. Further limitations that were not directly assessed by this experiment include the lack of congruence between the relevant period for co-payment in an economic model and the reporting period in a survey as well as the various ways in which these two periods may differ (Farbmacher and Winter 2013). The current research highlights one set of biases that may arise from a mismatch between periods, and it points to the use of longer recall periods rather than shorter to minimise bias in these instances; still, this research does not resolve the issues related to the need for more rigorous scaling techniques.

This was a large survey with data on nearly 5000 patients with diabetes. While this study is based on a sub-set of the Australian population, the advantage of randomisation using birth date means the impact of any selection on recall applies to all groups, so the relative observed effects should remain. The survey included patients with a common chronic health condition (diabetes), and the subgroup analyses indicated that the 12-month recall period was preferred in all but one scenario, indicating that the results are likely to be generalizable. However, one caveat may be patients with cognitive impairment or acute conditions. Further studies are warranted, as patterns of bias may differ. Acute patients may differ in their perception of the salience of events and their ability to anchor the time periods recalled.

There have been some suggestions about the use of interventions to improve recall. Some research points to the importance of anchoring (Hufford and Shiffman, 2002). This is the ability to reference the bounds of the recall period with a salient event. It is thought that perhaps a 12-month period makes it easier for patients to anchor, and the current results support this reasoning (Means, 1989). Future directions for research include the use of interventions to improve the accuracy of recall for health service utilisation.

The clear empirical result that longer recall periods for doctor visits are associated with less overall error arises from a robust randomised study with a large sample size. These findings are important to inform current efforts to produce more standardised survey instruments for self-reported data as inputs to economic evaluation (such as the CSRI and DIRUM databases, Beecham and Knapp 1999; Ridyard and Hughes 2015). Likewise, these results will be important for those designing and using population health surveys and health services research more generally.

ACKNOWLEDGEMENTS

We would like to acknowledge the Australian Department of Health and the Diabetes Care Project team for their work collecting the data used in this analysis.

REFERENCES

Australian Bureau of Statistics (ABS). 2013. *Australian Health Survey 2012-13*. (Available from: http://www.abs.gov.au/australianhealthsurvey [2 November 2016].)

Australian Government. 2016. *Medicare Benefits Schedule July 2016*. Department of Health. (Available from: http://www.mbsonline.gov.au/ [2 November 2016].)

Beecham J, Knapp M. 1999. *Costing Psychiatric Interventions*. Discussion Paper, 1536.

Bhandari A, Wagner T. 2006. Self-reported utilization of health care services: improving measurement and accuracy. *Medical Care Research and Review* **63**(2) : 217-35.

Bradburn NM, Rips LJ, Shevell SK. 1987. Answering autobiographical questions: The impact of memory and inference on surveys. *Science* **236**(4798) : 157-161.

Branch ER. 1994. The Consumer Expenditure Survey: A Comparative Analysis. *Monthly Labour Review* **117**(12) **:** 47-55.

Celebrezre AJ, Terry LL. 1965. *Health Interview Responses Compared with Medical Records*. National Centre for Health Statistics, Series 2 Number 7, Washington D.C.

Centres for Disease Control and Prevention (CDC). 2016. *National Health Interview Survey*. National Centre for Health Statistics. (Available from: https://www.cdc.gov/nchs/nhis [2 November 2016].)

Centres for Disease Control and Prevention (CDC). 2016. *National Health and Nutrition Examination Survey (NHANES)*. National Centre for Health Statistics. (Available from: http://www.cdc.gov/nchs/nhanes/ [2 November 2016].)

Centres for Disease Control and Prevention (CDC). 2017. Health Measures in the 1997 Redesigned National Health Interview Survey (NHIS). National Center for Health Statistics. (Available from: https://www.cdc.gov/nchs/nhis/nhis_redesign.htm [2 March 2017].)

Clarke PM, Fiebig DG, Gerdtham U-G. 2008. Optimal recall length in survey design. *Journal of Health Economics* **27**(5) : 1275-1284.

Department of Health [website]. *Evaluation report of the Diabetes Care Project.* Canberra: Australian Government; 2015. https://www.health.gov.au/internet/main/publishing.nsf/Content/302DF0372F537A43CA257E35000138E8/$File/DCP%20Evaluation%20Report.pdf (accessed Jan 2018).

Das J, Hammer J, Sánchez-Paramo C. 2012. The impact of recall periods on reported morbidity and health seeking behavior. *Journal of Development Economics* **98**(1) : 76-88.

DIRUM. 2016. Database of Instruments for Resource Use Measurement (Available from: http://www.dirum.org/ [2 November 2016].)

Edwards WS, Winn DM, Collins JG. 1996. *Evaluation of 2-week doctor visit reporting in the National Health Interview Survey.* National Center for Health Statistics. Vrtal Health Stat 2(122).

Farbmacher H, Winder J. 2013. Per-period co-payments and the demand for health care: evidence from a survey and claims data. *Health Economics* **22**(9) : 1111-23.

Fountaine T. and Bennett CC. Health care homes: lessons from the Diabetes Care Project *Med J Aust* 2016; **205** (9): 389-391

Gordon LG, Patrao T, Hawkes AL. 2012. Can colorectal cancer survivors recall their medications and doctor visits reliably? *BMC Health Services Research* **12** : 440.

Health and Social Care Information Centre (HSCI). 2013. *Health Survey for England 2013*. (Available from: https://data.gov.uk/dataset/health_survey_for_england? [2 November 2016].)

Heijink R, Xu K, Saksena P, Evans D. 2011. *Validity and comparability of out-of-pocket health expenditure from household surveys: a review of the literature and current survey instruments.* World Health Organisation, Discussion Paper 1. Department "Health Systems Financing" (HSF) Cluster "Health Systems and Services" (HSS).

Hufford MR, Shiffman S. (2002). Methodological issues affecting the value of patient-reported outcomes data. *Expert Review of Pharmacoeconomics and Outcomes Research* **2**(2) : 119-128 .

Icks A, Dittrich A, Brüne M, Kuss O, Hoyer A, Haastert B, Begun A, Andrich S, Hoffmann J, Kaltheuner M, Chernyak N. Agreement found between self-reported and health insurance data on physician visits comparing different recall lengths. *J Clin Epidemiol*. 2017 Feb;82:167-172.

Kjellsson G, Clarke P, Gerdtham UG. 2014. Forgetting to remember or remembering to forget: a study of the recall period length in health care survey questions. *Journal of Health Economics* **35** : 34-46.

Leach MJ, Segal L, Esterman A, Armour C, McDermott R, Fountaine T. 2013. The Diabetes Care Project: an Australian multicentre, cluster randomised controlled trial [study protocol]. *BMC Public Health* **13**(1) **:** 1212.

Means B. 1989. *Autobiographical memory for health-related events*: US Department of Health and Human Services, Public Health Service, Centers for Disease Control, National Center for Health Statistics.

Norquist JM., Girman C, Fehnel S, DeMuro-Mercon C, Santanello N. 2012. Choice of recall period for patient-reported outcome (PRO) measures: criteria for consideration. *Quality of Life Research* **21**(6) : 1013-20.

Petrou S, Murray L, Cooper P, Davidson LL. 2002. The accuracy of self-reported healthcare resource utilization in health economic studies. *International Journal of Technology Assessment in Health Care* 18(03) : 705-710.

Ridyard CH, Hughes DA. 2010. Methods for the Collection of Resource Use Data within Clinical Trials: A Systematic Review of Studies Funded by the UK Health Technology Assessment Program. *Value in Health* **13**(8) : 867–872

Ridyard CH, Hughes DA. 2015. *Review of resource-use measures in UK economic evaluations*. Personal Social Services Research Unit (PSSRU), Unit Costs of Health and Social Care report. (Available from: http://www.dirum.org/newsitems/details/28? [2 November 2016].)

Ritter PL, Stewart AL, Kaymaz H, Sobel DS, Block DA, Lorig KR. 2001. Self-reports of health care utilization compared to provider records. *Journal of Clinical Epidemiology* **54**(2) : 136-141.

Rubin DC, Baddeley AD. 1989. Telescoping is not time compression: A model. *Memory & Cognition* **17**(6) : 653-661.

Sudman S, Bradburn NM. 1973. Effects of time and memory factors on response in surveys. *Journal of the American Statistical Association* **68**(344) : 805-815.

Thorn JC, Coast J, Cohen D, Hollingworth W, Knapp M, Noble SM, Ridyard C, Wordsworth S, Hughes D. 2013. Resource-Use Measurement Based on Patient Recall: Issues and Challenges for Economic Evaluation .*Applied Health Economics and Health Policy* **11** :155–161

Table I. Description of participants

Variables	Total completing survey	2-week recall group	3-month recall group	12-month recall group	F-test† (p-value)
Total number of patients	5,305	1,771	1,796	1,738	-
	Total number of patients with characteristic (presented as a proportion of column total in parentheses %)				
Female	2,372 (44.71)	803 (45.34)	792 (44.10)	777 (44.71)	0.01 (0.9906)
Anglo/European ethnicity	3,962 (74.68)	1,302 (73.52)	1,341 (74.67)	1,319 (75.89)	2.19 (0.1394)
Aboriginal or Torres Strait Islander	49 (0.92)	14 (0.79)	23 (1.28)	12 (0.69)	1.20 (0.2740)
Completed year 12 or above education	1,891 (35.65)	626 (35.35)	652 (36.30)	613 (35.27)	0.12 (0.7324)
Employed	1,154 (21.75)	372 (21.01)	393 (21.88)	389 (22.38)	0.69 (0.4066)
Retired	3,116 (58.74)	1,026 (57.93)	1,060 (59.02)	1,030 (59.26)	0.36 (0.5474)
Income less than $20,000 per year	2,097 (39.53)	686 (38.74)	718 (39.98)	693 (39.89)	0.18 (0.6699)
Risk: complex	2,667 (50.27)	878 (49.58)	894 (49.78)	895 (51.50)	1.56 (0.2117)
Risk: out of range	3,092 (58.28)	992 (56.01)	1,062 (59.15)	1,038 (59.72)	2.72 (0.0989)
Risk: newly diagnosed	785 (14.80)	263 (14.75)	297 (16.54)	225 (12.95)	6.30 (0.0121)*
Concession health care card	2,940 (55.42)	993 (56.07)	977 (54.40)	970 (55.81)	0.09 (0.7588)

Private health insurance	2,472 (46.60)	810 (45.74)	817 (45.49)	845 (48.62)	4.16 (0.0415)*
Type 1 diabetes	352 (6.64)	106 (5.99)	125 (6.96)	121 (6.96)	0.60 (0.4403)
Type 2 diabetes	4,426 (83.32)	1,477 (83.40)	1,493 (83.13)	1,450 (83.43)	0.02 (0.8952)
Type 1 and type 2 diabetes	213 (4.02)	75 (4.23)	71 (3.95)	67 (3.86)	0.20 (0.6513)
			Mean (SD)		
Medicare visits during the last year**	13.023 (9.585)	13.408 (10.893)	12.813 (8.664)	12.842 (8.026)	1.81 (0.1645)
Length of diagnosis in years	11.41 (9.10)	11.10 (8.93)	11.64 (9.14)	11.46 (9.21)	0.75 (0.9211)
Age in years	68.31 (11.15)	68.40 (11.38)	68.24 (10.85)	68.29 (11.22)	0.86 (0.8130)
SEIFA disadvantage score	979 (66.51)	978 (67.26)	981 (65.16)	979 (67.12)	0.78 (0.9912)

*statistically significant difference between groups at p=0.05 level. **Last year is defined as 1 year from the date of survey administration, defined according to the date the survey was completed †to assess statistical independence

Table II Description of error variables by recall period

Variable	Total	2 weeks	3 months	12 months	F test (p-value)
	N=4399	N=1481	N=1490	N=1428	
	Mean (SD)				
Medicare visits during recall period	5.452 (7.531)	0.549 (0.855)	3.242 (2.749)	12.842 (9.026)	2065.91 (<0.00001)
Medicare visits during recall period plus 2 week telescoping window	5.734 (7.421)	1.107 (1.394)	3.721 (3.020)	13.321 (9.373)	1855.48 (<0.00001)
Proportion of patients with no Medicare visit during recall period	0.253 (0.435)	0.609 (0.488)	0.101 (0.302)	0.041 (0.199)	1153.68 (<0.00001)
Recalled visits during recall period	5.275 (7.605)	1.057 (1.254)	3.838 (3.928)	11.147 (10.267)	983.85 (<0.00001)
Proportion of patients recalling no visit	0.125 (0.331)	0.341 (0.474)	0.029 (0.167)	0.003 (0.053)	604.93 (<0.00001)
Proportion of patients with reported visits not equal to Medicare- absolute error	0.712 (0.453)	0.463 (0.499)	0.758 (0.428)	0.924 (0.266)	472.15 (<0.00001)
Proportion of patients with recalled visits not equal to Medicare- absolute error conditional on at least 1 visit	0.767 (0.423)	0.446 (0.497)	0.748 (0.434)	0.920 (0.271)	306.45 (<0.00001)
Number of total days reported in error- total error (reported minus Medicare)	-0.177 (6.051)	0.508 (1.176)	0.596 (3.610)	-1.695 (9.715)	68.63 (<0.00001)
Number of additional days reported- positive error (reported minus Medicare)	1.407 (4.192)	0.597 (1.067)	1.307 (3.002)	2.353 (6.480)	66.34 (<0.00001)

Number of missing days from report- negative error (reported minus Medicare)	-1.584 (3.817)	-0.088 (0.371)	-0.711 (1.469)	-4.048 (5.773)	565.62 (<0.00001)
Relative error: total number of days reported in error per day recalled	-	0.036 (0.084)	0.007 (0.040)	0.005 (0.027)	208.01 (<0.00001)

Notes: The results are based on our study sample of 4,399 respondents (group of respondents shown in Figure 1, D).

Table III RMSE, variance and bias

		Bias1(Y_s)= E(y_s)-Y_s			Bias2(Y_s)= E(y_s)- (s/w)E(y_w)		
	Var (y_w) Variance for recall visits during each period	RMSE	Var(y_s)	Bias(y_s)	RMSE	Var(y_s)	Bias(y_s)
2 weeks	1.572	14.185	0.720	14.159	13.312	0.720	13.285
3 months	15.426	2.602	0.166	2.570	2.420	0.166	2.386
12 months	105.421	**1.716**	0.074	-1.695	**1.716**	0.074	-1.695

Notes: The results are based on our study sample of 4,399 respondents (group of respondents shown in Figure 1, D).
S=total time period (12 months), w=recall period
Bold= lowest RMSE (least biased) for each group/subgroup

Table IV RMSE, variance and bias (Equation Bias 2) by socio-demographic and clinical variables

	RMSE	Var (y_s)	Bias2(Y_s)	RMSE	Var (y_s)	Bias2(Y_s)	RMSE	Var (y_s)	Bias2(y_s)
		2 weeks			3 months			12 months	
Main study sample	13.312	0.720	13.285	2.420	0.166	2.386	**1.716**	0.074	-1.695
Age<60 years	9.504	2.583	9.367	2.532	0.685	2.393	**0.711**	0.443	0.250
Age>60 years	14.460	0.971	14.427	2.427	0.217	2.382	**2.258**	0.086	-2.238
Length of diagnosis >5 years	14.242	1.010	14.207	2.644	0.212	2.604	**1.505**	0.106	-1.470
Length of diagnosis <5 years	10.056	2.082	9.952	**1.831**	0.766	1.609	2.467	0.202	-2.426
Less than year 12 education	15.196	1.267	15.154	2.951	0.364	2.889	**1.893**	0.140	-1.855
Year 12 or greater education	11.021	1.958	10.932	1.942	0.312	1.860	**1.351**	0.184	-1.281
Unemployed	15.133	1.049	15.098	2.617	0.251	2.569	**2.066**	0.115	-2.038
Employed	7.080	1.860	6.947	1.728	0.352	1.623	**0.931**	0.141	-0.852
Income>$20k	9.562	1.135	9.502	1.723	0.209	1.661	**1.412**	0.120	-1.369
Income<$20k per year	17.547	2.229	17.483	3.529	0.606	3.442	**2.221**	0.234	-2.168

Risk: Complex (No)	11.253	1.011	11.208	2.102	0.330	2.022	**2.036**	0.100	-2.012
Risk: Complex (Yes)	15.411	1.861	15.351	2.822	0.324	2.764	**1.462**	0.182	-1.398
Risk: Out of range (No)	14.385	2.058	14.313	2.250	0.506	2.135	**1.068**	0.207	-0.966
Risk: Out of range (Yes)	12.486	1.038	12.445	2.611	0.236	2.565	**2.193**	0.113	-2.167

Notes: The results are based on our study sample of 4,399 respondents (group of respondents shown as Consort D in Figure 1).
Bold= lowest RMSE for each group/subgroup

APPENDIX 1: DCP GP recall question

We are interested in the way people remember their doctor visits as time passes. To help us understand this better please answer 1 question only from the following.

Please answer only the question that corresponds to the day of your birth.

If you were born between the 1st & 10th of the month (eg 3rd March, 7th Oct etc)	If you were born between the 11th & 20th of the month (eg 14th May, 19th Jan)	If you were born between the 21st & 31st of the month (eg 25th Aug, 31st Feb)
⬇	⬇	⬇
How many visits have you made to a doctor (general practitioner GP or specialist) in the last **2 WEEKS**? Answer: □□	How many visits have you made to a doctor (general practitioner GP or specialist) in the last **3 MONTHS**? Answer: □□	How many visits have you made to a doctor (general practitioner GP or specialist) in the last **12 MONTHS**? Answer: □□

Please also indicate today's date (use format DD/MM/YY) : / /

APPENDIX 2 Comparison of patient characteristics for those who answered only randomised period compared to those who incorrectly also answered additional periods not according to randomisation instructions

Variables	Patients who answered only as randomised (n=4102)	Patients who also answered additional recall periods (1199)	Difference between groups % (t test p value)
	Total n (%)	Total n (%)	
Female	1,847 (44.98)	525 (43.79)	1.19 (0.4636)
Anglo/European ethnicity	3,137 (76.40)	825 (68.81)	7.59 (<0.00001)*
Aboriginal or Torres Strait Islander	35 (0.85)	14 (1.17)	-0.3 (0.3155)
Completed year 12 or above education	1,554 (37.85)	337 (28.11)	9.7 (<0.00001)*
Employed	973 (23.70)	181 (15.10)	8.6 (<0.00001)*
Retired	2,385 (58.09)	731 (60.97)	-2.9 (0.0746)
Income less than $20,000 per year	1,545 (37.63)	552 (46.04)	-8.4 (<0.00001)*
Risk: complex	2,057 (50.10)	610 (50.88)	-0.7 (0.6354)
Risk: out of range	2,414 (58.79)	678 (56.55)	2.2 (0.1655)
Risk: newly diagnosed	594 (14.47)	191	-1.5 (0.2094)

		(15.93)	
Concession health care card (2)	2,196 (53.48)	744 (62.05)	-8.7 (<0.00001)*
Private health insurance (2)	2,034 (49.54)	438 (36.53)	13.0 (<0.00001)*
Type 1 diabetes	280 (6.82)	72 (6.01)	0.8 (0.3190)
Type 2 diabetes	3,447 (83.95)	973 (81.15)	2.8 (0.0222)*
Type 1 and type 2 diabetes	149 (3.63)	64 (5.34)	-1.7 (0.0080)*
	Mean (SD)	**Mean (SD)**	**Difference between groups % (t test p value)**
Length of diagnosis in years**	11.42 (9.16)	11.36 (8.81)	0.6 (0.8691)
Age in years	67.81 (11.18)	70.05 (10.88)	2.3 (<0.00001)*
SEIFA disadvantage score	982 (66.00)	970 (67.44)	11.8 (<0.00001)*

*statistical significant difference between groups p <0.05 level. **Note methodology of calculation is different, todays date is used from survey respondents, for non-respondents their official last day in trial is used.

**note only correct responses as randomised are used in the main analysis due to the difference in characteristics and in order to preserve intention to treat

APPENDIX 3

Medicare Benefits Schedule (MBS) Items numbers coded as doctor or specialist visits listed below.
For a full list of MBS Item Numbers refer to http://www.mbsonline.gov.au/ [Accessed 28th August, 2015]

Items: 3-51, 132, 133, 141-147, 160, 164, 193, 195, 197, 199, 597-600, 701, 703, 705, 707, 715, 721, 723, 729, 731, 732, 901, 902, 2497-2559, 2620-2635, 2664-2677, 2700, 2701, 2713,
2715, 2717, 2712, 2801, 2806, 2814, 2824, 2832, 2840, 3005, 3010, 3014, 3018, 3023, 3028, 5000-5067, 5200, 5203, 5207, 5208, 6007-6015,

APPENDIX 4: PRACTICAL APPLICATIONS

Economic evaluation costs

The costs associated with doctor visits were calculated using self-reported doctor data to test the impact of self-reported recall error on cost inputs to an economic evaluation. The DCP data for the intervention and control group were used to provide the number of self-reported and actual doctor visits over an 18-month period. The cost of each visit was assumed equal to a Level B professional attendance in consulting rooms with a fee of AU$37.05 (Australian Government, 2016). The self-reported visits (2 weeks, 3 months and 12 months recall) were scaled to represent 18 months to reflect the DCP study period. These were then compared to the 18 months of Medicare visits for the DCP control group (gold standard). The differences between recall groups were compared to assess impact for economic evaluation.

The number of doctor visits and associated cost of visits are reported in the table below. The 2-week scaled recall group resulted in the largest variation from the gold standard. The 12-month recall data led to the smallest absolute variation from gold standard albeit underestimated. This can potentially change the conclusions regarding i) the actual magnitude of GP visit costs, which are approximately doubled when using 2 week recall, and ii) the difference in costs between the intervention and control group as using 3 month recall leads to the intervention group appearing to be less costly than the control group.

Table: Cost of DCP intervention and control group general practitioner (GP) visits

Average per patient	2 weeks scaled data	3 months scaled data	12 months scaled da
Intervention Group*			

Number of patients	1084	1079	1027
Number of GP visits	42.56	22.99	16.86
Cost of GP visits	$1,577	$848	$625
Difference from gold standard	+$850 (116.9%)	+$121 (16.6%)	-$102 (14.0%)
Control Group			
Number of patients	397	411	401
Number of GP visits	37.62	23.24	16.37
Cost of GP visits	$1,394	$861	$607
Difference from gold standard	+$674 (93.6%)	+$141 (19.6%)	-$113 (15.7%)

Notes: The results are based on our study sample of 4,399 respondents (group of respondents shown as Consort D in Figure 1).

Bold=gold standard cost per person *composed of DCP intervention 1 and intervention 2 groups combined

Regression analyses: predicting number of doctor visits

Among a wide range of count data models, a negative binomial model was selected based on Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to demonstrate the relationship between primary care utilisation (number of doctor visits) and a range of patient (income, education, gender, ethnicity, concessional status, pension status, private insurance) and clinical (newly diagnosed, type 1 and/or 2 diabetes diagnosis, length of diagnosis in years, existence of complexity, whether hba1c measure is out of range- hba1c is a glycated haemoglobin test indicating how well diabetes is being controlled) covariates in the diabetes population. The aim of the analyses was to determine the impact of substituting the recalled doctor visits with the Medicare visits (gold standard) across recall periods. Impact was determined through a change in coefficient sign or statistical significance (p< or > 0.1) of covariates. The percentage differences in coefficients and standard errors with the use of registered compared to recall data were calculated.

Results of the negative binomial regression modelling are shown in figure and table below with direct comparison of recall data and Medicare data for doctor visits. There was little consistency across individual covariates in the three analyses. For example, when using recall data on doctor visits the predictor variable "income less than AU$20,000" changed from a negative to a positive coefficient in the 2 week analysis, became statistically significant in the 2 week and 3month analyses and lost statistical significance in the 12 month analyses.

The figure below shows the percentage change in coefficients and standard errors across the three recall analyses. Overall the largest percentage changes occurred for the analysis using 2-week recall data compared to Medicare data. The standard errors were always lower in the 2-week recall period, and higher for the 3-month recall period. There was little consistency in either the direction or size of impact across the analyses.

Table: Comparison of regressions results based on Medicare and recall data by recall period

	2 weeks recall						3 months recall						12 months recall					
VARIABLES	Medicare data		Recall data		flip in sign	flip in p-value	Medicare data		Recall data		flip in sign	flip in p-value	Medicare data		Recall data		flip in sign	flip in p-value
	coef	pvalue	coef	pvalue			coef	pvalue	coef	pvalue			coef	pvalue	coef	pvalue		
year12 or more	-0.1794*	0.077	-0.1679**	0.031			-0.0921*	0.063	-0.1341**	0.013			-0.0607	0.121	-0.0903*	0.056		Y
income <20k	-0.0137	0.901	0.1775**	0.035	Y	Y	0.0406	0.452	0.1397**	0.018		Y	0.1161***	0.008	0.0582	0.270		Y
employed	-0.3395**	0.041	-0.3129**	0.018			-0.3136***	0.000	-0.3202***	0.001			-0.3951***	0.000	-0.5094***	0.000		
pension	-0.1435	0.297	-0.0297	0.782			-0.1844**	0.014	-0.0884	0.289		Y	-0.1432**	0.019	-0.1447**	0.048		
current age	0.0049	0.345	0.0028	0.489			0.0118***	0.000	0.0026	0.426		Y	0.0066***	0.003	0.0025	0.336		Y
length diagnosis	-0.0010	0.865	0.0039	0.369	Y		0.0018	0.510	0.0098***	0.001		Y	0.0047**	0.034	0.0060**	0.031		
type1	0.2018	0.611	0.0806	0.774			-0.4576***	0.004	-0.1088	0.553		Y	0.0530	0.691	0.2475	0.128		
type2	0.0760	0.831	-0.0904	0.716	Y		-0.2628**	0.045	0.0566	0.720	Y	Y	0.0076	0.947	0.0876	0.534		
type1 & 2	-0.2003	0.656	-0.3190	0.318			-0.2296	0.208	0.5683***	0.005	Y	Y	0.1359	0.352	0.1905	0.287		
female	0.0400	0.677	-0.0693	0.348	Y		0.0711	0.134	0.1174**	0.023		Y	0.0617	0.103	0.0995**	0.031		Y
anglo european	-0.0036	0.979	-0.0700	0.503			-0.0570	0.434	-0.0093	0.910			0.0140	0.796	0.1487**	0.025		Y
indigenous	0.5469	0.236	0.2947	0.430			0.0996	0.607	0.0555	0.796			-0.3532	0.179	-0.3745	0.242		
health card	0.0724	0.524	0.1468*	0.096		Y	0.0816	0.125	0.1343**	0.020		Y	0.0754*	0.080	0.0597	0.250		Y
private insurance	-0.2395**	0.022	-0.2187***	0.006			-0.0853*	0.087	-0.0456	0.407		Y	-0.1028***	0.010	-0.1216**	0.013		

complex	0.0895	0.366	0.1850**	0.015		Y	0.1899***	0.000	0.1985***	0.000			0.1264***	0.001	0.1810***	0.000		
out of range	-0.0034	0.973	-0.0903	0.238			0.0289	0.563	0.0033	0.952			0.1090***	0.005	-0.0491	0.297	Y	Y
newly diagnosed	-0.0405	0.776	0.1212	0.248	Y		0.0925	0.142	0.0397	0.567			-0.0014	0.981	0.0949	0.179	Y	
Number of observations	911		911				975		975				955		955			

Note: “flip in p-value” indicates when comparing across survey data and MBS register data whether the regressor changes from being significant (<0.1) to non-significant (⩾0.1) or vice versa. All regressions use negative binomial specification as it is the preferred model based on AIC and BIC measures. Statistical significance *p<0. 1, **p<0.05, ***p<0.01

Figure 1 Consort Flow Diagram: randomised recall experiment

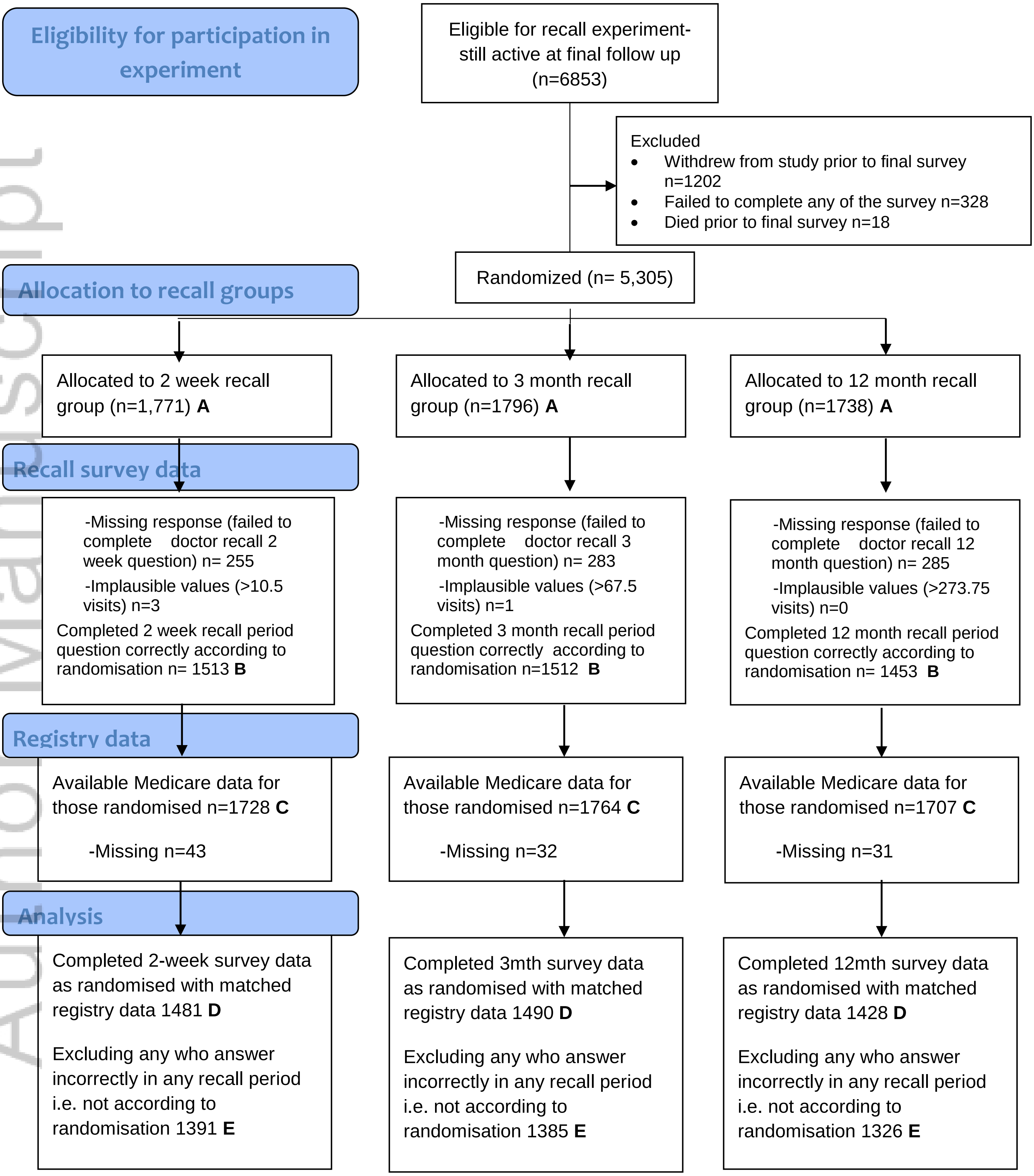
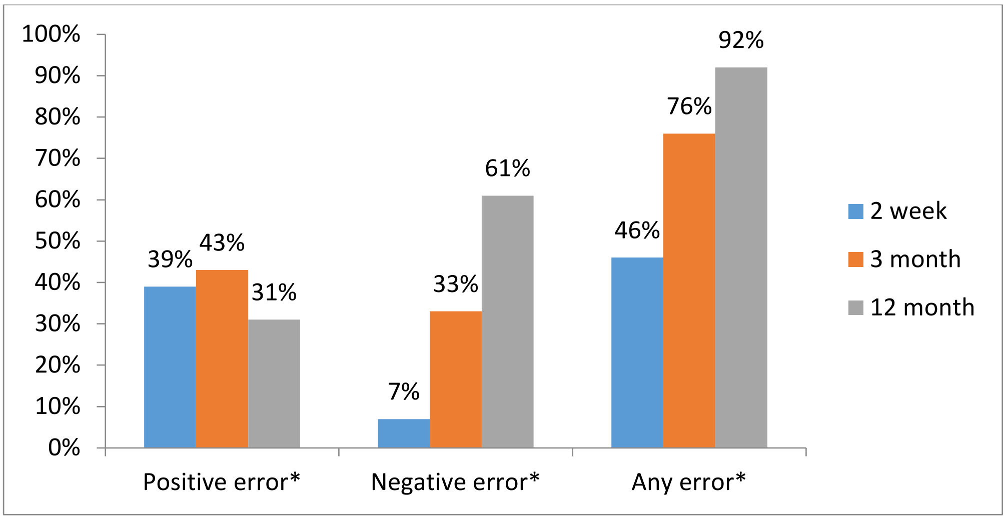

Eligibility for participation in experiment
Eligible for recall experiment- still active at final follow up (n=6853)
Excluded
• Withdrew from study prior to final survey n=1202
• Failed to complete any of the survey n=328
• Died prior to final survey n=18
Allocation to recall groups
Randomized (n= 5,305)
Allocated to 2 week recall group (n=1,771) A
Allocated to 3 month recall group (n=1796) A
Allocated to 12 month recall group (n=1738) A
Recall survey data
-Missing response (failed to complete doctor recall 2 week question) n= 255
-Implausible values (>10.5 visits) n=3
Completed 2 week recall period question correctly according to randomisation n= 1513 B
-Missing response (failed to complete doctor recall 3 month question) n= 283
-Implausible values (>67.5 visits) n=1
Completed 3 month recall period question correctly according to randomisation n=1512 B
-Missing response (failed to complete doctor recall 12 month question) n= 285
-Implausible values (>273.75 visits) n=0
Completed 12 month recall period question correctly according to randomisation n= 1453 B
Registry data
Available Medicare data for those randomised n=1728 C
-Missing n=43
Available Medicare data for those randomised n=1764 C
-Missing n=32
Available Medicare data for those randomised n=1707 C
-Missing n=31
Analysis
Completed 2-week survey data as randomised with matched registry data 1481 D
Excluding any who answer incorrectly in any recall period i.e. not according to randomisation 1391 E
Completed 3mth survey data as randomised with matched registry data 1490 D
Excluding any who answer incorrectly in any recall period i.e. not according to randomisation 1385 E
Completed 12mth survey data as randomised with matched registry data 1428 D
Excluding any who answer incorrectly in any recall period i.e. not according to randomisation 1326 E

Figure 2 Proportion of patients with negative or positive errors in each recall group



*Statistical significance positive error F=22.55, p<0.0001; negative error F=612.2, p<0.0001; Any error F= 472.15, p<0.0001