How long is long enough? Optimizing the length of recall in health care utilization surveys

WORK IN PROGRESS: NOT FOR CITATION, OR QUOTATION WITHOUT PERMISSION

Philip Clarke[1] and Ulf Gerdtham[2]

1. Health Economic Research Centre, University of Oxford, UK.

2. Department of Community Medicine, Lund University, Sweden.

Corresponding author: Philip Clarke, HERC, Institute of Health Sciences, Old Road Campus, Headington, OX3 7LF; Email: Philip.clarke@dphpc.ox.ac.uk

Abstract: Self-reported health care utilisation data forms the basis of a wide variety of health economic research ranging from evaluation of new interventions to the measurement of inequity of access to care. While this type of data is normally collected through an annual survey there is little consensus on the most appropriate time period over which to collect prior utilisation information. Health surveys currently use questions to ascertain health care usage for periods ranging from two weeks to a year. In determining the recall period there are two potential sources of error that work in opposite directions. On the one hand, shortening the recall period reduces the likelihood of error due to misreporting, but on the other hand it increases the error associated with using this information to predict longer-term health care usage at an individual level. In this study we develop a method for determining the optimal recall period based on predefined statistical criteria. In particular, we estimate functions to represent these two sources of error and then find the optimum recall period that minimises the mean squared error. To illustrate the method we attempt to find the optimal recall period for hospital usage based on data from Statistics Sweden's Survey of Living Conditions (the ULF survey) that has been linked to patient registry data on actual health services in Sweden.

Paper presented to the Health Economists' Study Group Conference

Glasgow, 30th June- July 2nd 2004.

Introduction

Self-reported health care utilisation is used in many facets of health economics. For example, economic evaluations of health care interventions are often based on self-reported health care data that is collected during the course of a study. It is also routinely collected in national health surveys such as the Health Survey for England, or the Medical Expenditure Survey in the US. While these surveys involve interviews conducted over the course of a year, there is considerable variation in the period over which subjects are asked to recall their previous health care use. This has been highlighted by a recent OCED study comparing inequity in access to health care across 21 developed countries based on health surveys in different countries. The period of recall for primary care and aspects of hospital use ranged from 2 weeks in Australian National Health Survey to one year in the European Community Household Panel (Health Equity Research Group, 2004). As these surveys are being used to estimate the same measure of inequality, the question arises as to which recall period more accurately captures inequity in care.

It has been widely recognized that there is an inverse relationship between the length of time over which subjects are asked to recall prior use and the accuracy of the reported estimates. The longer the period of recall the greater the likelihood of error. A recent review of validation studies comparing reported with actual use suggests that under-reporting is more likely than over-reporting, especially in primary care (Evans and Crawford, 1999). This has implications for evaluation, particularly when this results in differing degrees of recall between interventions. As Evans and Crawford (1999) demonstrate there can be a serious distortion to the incremental average costs which can impact on the estimated cost-effectiveness ratio.

While recall error is undesirable, a survey with a short length of recall provides very little information about an individual's normal health care use. As Deaton (1997) has noted in the context of measuring household consumption expenditure:

> "…if the object of the exercise is to estimate average consumption over a year, one extreme is to approach a sample of households on January 1 and ask each to recall expenditures for the last year. The other extreme is to divide the sample over the days of the year, and to ask each to report consumption for the previous day. The first method would yield a good picture of each household's consumption, but runs the risk of measurement error… [t]he second method… will give a good estimate of the mean consumption over all households… [but] will yield estimates of individual expenditure that… are only weakly related to normal expenditures…"(p. 24)

The purpose of this study is to demonstrate how statistical methods can be used to estimate an optimal recall period for health care consumption surveys. To limit the scope, we will focus on the design of an annual survey whose purpose is not only to measure average use in the sample, but also estimate annual individual use of health care services. In determining an optimal recall period it is important to recognize that there are two potential sources of error that work in opposite directions. Shortening the recall period reduces the likelihood of error due to misreporting, however on the other hand it increases the error associated with using this information to predict longer-term health care usage at an individual level. For example, consider a survey of health care use that informs an evaluation alongside a clinical trial in which follow-up is conducted on an annual basis. The researcher may not only be interested in measuring the difference in average use between the intervention and control groups, but may also want to measure individual use in order to determine the correlation between costs and effects which is necessary to accurately account for uncertainty. Estimates of individual health care use are also required in many other applications such as the measurement of inequity in access to care (outlined above) where the focus is on dispersion of care rather than overall mean use.

Methods

Statistical background

We define y_i^t as a measure of consumption of some type of health care (e.g. days in hospital over the previous year) for an individual i over a time period of length t. A survey may ask respondents to recall y_i^t, or recall consumption over a shorter period of time w, which we denote as z_i^w, where $w<t$. The consumption of health care within t that is not covered by the survey is denoted as v_i^k, where $k = t - w$.

While a survey will inform the researcher about z_i^w the real interest is in measuring y_i^t. One way of estimating y_i^t is to regard v_i^k as missing data that can be imputed by the researcher. For example a common method for imputing univariate missing data is to regress:

$$z_i^w = \beta X_i + \varepsilon_i \qquad (1)$$

where the vector X_i contains information on the characteristics of individual (that are unlikely to vary over t, or can be predicted in period k, e.g. an individual's age) and dummies to account for seasonal variation in health care use, and β a vector regression coefficients. Equation. (1) can be used to predict the health care use not observed by the survey (Schafer and Olsen1998). We denote the prediction as $\hat{v}_i^k$ and when combined with z_i^w, it provides a point estimate of overall use, i.e.:

$$\hat{y}_i^t = z_i^w + \hat{v}_i^k \qquad (2)$$

Given that $\hat{y}_i^t$ is based in part on v_i^k which is not known to the researcher, multiple imputation could be used to account for the additional uncertainty surrounding the use of simulated rather than actual data of health care use (Briggs et. al., 2003).

In this study we are considering self reported utilization of z_i^w which may be subject to error. Hence a survey will obtain reported use (x_i^w):

$$x_i^w = z_i^w + \mu_i^w \tag{3}$$

where μ_i^w is recall error. Substituting Eq. (3) into Eq. (1) leads to:

$$x_i^w = \beta X_i + \varepsilon_i - \mu_i^w \tag{4}$$

the consequences depend on the properties of μ_i^w. In particular if μ_i^w is independently and identically distributed with zero mean and variance σ^2 it will be absorbed into error term and can be ignored (Greene, 1993). However, evidence on the nature of reporting error (outline above) suggests that under reporting is common, implying $E(\mu_i^w) < 0$, which will result in $E(x_i^w) \neq E(z_i^w)$, and $E(\hat{v}_i^k) \neq E(v_i^k)$ leading to a biased estimate of y_i^t. In these circumstances, we suggest two possible criteria for determining an optimal recall period are:

(i) $E[\hat{y}_i^t] - E[y_i^t] = 0$ i.e. that the predicted average use is an unbiased estimate of actual average use;

and (ii) the minimizing $E[(\hat{y}_i^t - y_i^t)^2]$, i.e. the means squared error of individual level prediction.

The choice between these criteria depend what type of application the survey is being used to inform.

Construction of a simulation model of self-reported health care use

A strategy to determine an optimal recall period in a health use survey is to examine how different durations of recall impact on summary statistics such as the mean and the mean squared error. This requires accurate information on actual use (that could be provided by a patient register, or administrative data) and reported use from surveys employing different recall periods. Ideally, this information could come from a trial where patients are randomised to health care use questions that are identical except for the period of recall. A function relating recall error to duration could then be calculated. Unfortunately, we are unaware of this type of trial being conducted and so have constructed a simulation model to explore these issues.

The key parameters of the model are based on information from Statistics Sweden's Survey of Living Conditions (the ULF survey) that has been linked to the national Patient Register (the National Board of Health and Welfare). Every year, Statistics Sweden conducts systematic surveys of living conditions, in the form of one-hour personal interviews with randomly selected adults aged 16-84 years. Since 1975 around 7,000 individuals have been interviewed each year. The Patient Register includes information about ICD-codes, hospital admissions, and the total length of stay in hospital over the past three months.

Information on the length of stay in hospital (number of night) over the last three months by 11,698 patients from the 1996 and 1997 surveys was compared with registry data to determine the accuracy of reporting over this period. For the purposes of developing this simulation we have assumed that the patient registry is the *gold standard* by which the accuracy of reported use can be judged.

Figure 1 illustrates a system for classifying the accuracy of patient recall. Responses to a question can be classified into four states (S1- S4), the last of which can be further classified into 3 sub-states:

S1: patient registry and survey indicates zero length of stay (true negative, i.e. no recall error);
S2: patient registry indicates a stay of one or more nights, but no reported use (false negative recall);
S3: patient registry indicates no actual use, but survey indicates stay of one or more nights (false positive recall);
S4: patient registry and survey both indicate positive use (i.e. true positive) which may be classified as either:

(a) survey indicates fewer number of day in hospital than patient registry (under reporting);
(b) survey indicates same number of days as patient registry (exact match);
(c) survey indicates a greater number of days in hospital than patient registry (over reporting);

Based a comparison of responses to the ULF survey and the patient registry we have estimated the following as probabilities for each state for a three month recall period:

(i) $P(S1 \mid z_i^w = 0) = 0.99$;

(ii) $P(S2 \mid z_i^w > 0) = 0.54$;

(iii) $P(S3 \mid z_i^w = 0) = 1 - P(S1 \mid z_i^w = 0) = 0.01$;

(iv) $P(S4 \mid z_i^w > 0) = 1 - P(S2 \mid z_i^w > 0) = 0.46$;

(v) $P(S4(a) \mid S4) = 0.32$;

(vi) $P(S4(b) \mid S4) = 0.36$;

(vii) $P(S4(c) \mid S4) = 0.32$;

From these probabilities we conclude that patients with no actual use during the period have a low probability (1% of patients) of false reporting. While only 46% of patients recall an actual hospital stay over the last three months, of those only 36% accurately recall the number of nights in hospital, and the remainder are evenly divided between over and under reporting.

For those patients in states S3, S4(a) and S4(c) the magnitude of the recall error must be defined. In the case of S3, the mean reported length of day from the ULF survey was 7.3 days (SD 12.3); for S4(a) mean reported length of stay was 35% less than actual use and for S4(c) the mean over reporting was 98% above actual use.

All of this data relates only to a survey asking about use over the previous three months, and so in order to undertake the simulation we must make assumptions about the nature of the error if the survey had adopted other durations of recall. It would seem reasonable to assume that there would be no recall error over very short periods (e.g. the previous day). The degree of error over other periods is less easy to ascertain, with the limited evidence suggesting accuracy declining linearly over a 12 month period (Evans and Crawford 1999).

In this study we make three assumptions regarding the nature of the recall error over time. We will initially assume the error increases at a linear rate for all probabilities outline above. Our alternative assumptions are that errors increase at a logarithmic or exponential rate. In the case of probabilities, the linear and exponential error functions can lead to probabilities less than one. When this occurs we have censored the probabilities at 0.05 (i.e. 95% error).

The algorithm for the simulations is as follows. six recall periods were examined: one month, two months, three months, four months, six months and twelve months (all being factors of twelve so that the interval not covered by the survey is a multiple of these periods). For each of these recall periods, a

simulated response was constructed. This involved comparing a random number drawn from a uniform distribution ranging from zero to one with each of the probabilities (i) to (vii) to determine each individual's response to the survey (ranging from S1-S4). For those in S1 and S4 (b) the actual number of days in hospital from the patient registry was used. In all other states the actual length of stay was modified due to recall error. For example, if the individual is in S2, the reported use is zero even though actual use is positive. For individual's in states S3, S4(a) and S4(c) the amount of error must also be simulated. For simplicity we have not varied the degree of over and under reporting by recall period, but the simulation model could easily be adapted to explore this issue in future work.

Using the simulation model to determine an optimal recall period

The main purpose of the simulation is to provide likely estimates of x_i^w that can then be used to estimate and compare the sample mean and mean squared error for different recall periods. For example, consider the shortest recall of one month. The survey response denoted as x_i^1 for a randomly chosen month is first calculated using the simulation model. In order to predict $\hat{y}_i^t$ a regression equation for x_i^1 (Eq. 4) is then estimated based on explanatory variables such as age and self-reported health obtained from the ULF survey. The estimated equation is then used to predict $\hat{v}_i^k$ for the remaining 11 months not covered by the survey. The predicted use over these 11 months is added to x_i^1 to obtain $\hat{y}_i^t$ and this is used to calculate $E[\hat{y}_i^t]$ and $E[(\hat{y}_i^t - y_i^t)^2]$.

Four scenarios that make different assumptions about the nature of recall error are reported in the results. These are:

A) No recall error;
B) All probabilities of recall error rise/decline at a linear rate;

C) All probabilities of recall error rise/decline at logarithmic rate;
D) All probabilities of recall error rise/decline at an exponential rate;

Table 1 lists the minimum, six monthly and maximum values of all parameters used in the model for each scenario.

We calculate $E[\hat{y}_i^t]$ and $E[(\hat{y}_i^t - y_i^t)^2]$ for all recall periods ranging from one month to the full twelve months to compare the four scenarios.

Results

Table 2 shows the parameters for the OLS regression with a three explanatory variables that were correlated with reported length of stay in hospital: age (mean 45.5; SD 18.3) and reported health status "bad" (mean 0.04) and health status "very bad" (mean 0.01). The coefficients and significance of these variables varied considerably across the different recall periods. Only in the longer periods (six months and over) did these variables appear to be significantly related to the length of stay. Further the degree of fit as measured by the R^2 was generally very poor. Other explanatory variables (e.g. sex) and more complex specifications such as two-part models were explored, but these approaches did not improve the overall fit of the equations. While recognizing that these equations are likely to be poor predictors of annual care we proceeded in order to illustrate how they can be used to evaluate different recall periods.

Figure 3 plots the mean annual length of stay that comprises survey response and prediction for the missing data from the above OLS equations. The actual mean for the 11,698 individuals in the sample was 0.9 days per year. Not surprisingly when there was no recall error (scenario one) there was little difference between the predicted and actual mean over most recall periods. In all the other scenarios there were differences between the predicted and actual mean with the largest variation arising in scenarios C and D.

Figure 4 plots the mean squared error of the difference between actual and reported length of stay at an individual level. Again, not surprisingly under Scenario A the MSE declined continuously as the period of recall increases. Supporting the astounding conclusion that in a world without error the recall period should be as long as possible! The general trend of a declining MSE as the recall period increases was also true of Scenario B and C. This indicates that when the errors increase at a linear, or logarithmic rate the 12 month recall period has the lowest MSE i.e. reported annual data with the highest degree of error has a lower MSE than predictions based on more accurate data from short periods of recall. However, if errors increase at an exponential rate then six month recall has the lowest MSE.

It should be emphasized that these results depend not only on the assumed rate of increase in error, but also the probabilities that have been used to define the 3 month recall error. Other types of health care (or other surveys) may have different patterns of error, so the results should not be generalized at this stage.

Discussion

Health economists often use self-reported health care consumption data in analyses. An important issue that arises, when collecting this type of data, is the length of time over which a respondent be asked to recall previous health care use. In this study we have attempted to explore this systematically by developing a simulation model to examine the performance of survey questions with different recall lengths that are used to estimate annual health care consumption using predefined statistical criteria.

This study is at a relatively early stage and should be regarded more as an illustration of the method than providing results that can be used to inform

future surveys. However two tentative conclusions regarding measurement of hospital use can be drawn:

(i) Recall error clearly impacts on the mean of reported use and this may result in bias across a wide range of survey periods;

(ii) In regard to prediction of individual length of stay a longer recall period would appear preferable, unless recall error rises at exponential rate.

The above findings are based on prediction from regression equations that have a relatively poor degree of fit. It would useful to explore the degree to which the recall period can be reduced when better predictors of health care use are available. We intend to examine this issue in future work.

While this study has only begun to use statistical methods to develop an optimal design for health care consumption surveys, we hope that it has demonstrated that this approach provides an alternative to the arbitrary methods often employed.

References

Briggs AH, Clark T, Wolstenholem J, Clarke P. "Missing... presumed at random: cost-analysis of incomplete data", *Health Economics*, 12, pp. 377-392. 2003

Evans C. and Crawford B. "Patient self-report in pharmacoeconomic studies: Their use and impact on study validity", *Pharmacoeconomics*; 15(3): pp. 241-256. 1999.

Deaton A. *The Analysis of Household surveys: A microeconometric approach to development policy*, John Hopkins University Press, 1997.

Greene WH. *Econometric Methods*, Second Edition, Macmillian 1993.

Health Equity Research Group, "Income-related inequality in the use of medical care in 21 OECD countries", Towards High Performing Health Systems: Policy Studies from the OECD Health Project, OECD, Paris (in press).

Schafer JL and Olsen MK "Multiple imputation for multivariate missing-data problems: a data analyst's perspective", The Pennsylvania State University, Working Paper. 1998.

Table 1: Assumed probabilities by scenario

	Linear rate			logarithmic rate			Exponential rate		
	1 Month	6 months	12 months	1 Month	6 months	12 months	1 Month	6 months	12 months
$P(S1 \mid z_i^w = 0)$	1	.98	0.95	.99	.99	.99	.99	.97	0.76
$P(S2 \mid z_i^w > 0)$	0.85	0.07	0.05	0.65	0.47	0.40	0.76	0.05	0.05
$P(S4(a) \mid S4)$	0.79	0.05	0.05	0.52	0.26	0.16	0.67	0.05	0.05
$P(S4(b) \mid S4)$	0.1	0.48	0.48	0.24	0.37	0.42	0.17	0.48	0.48
$P(S4(c) \mid S4)$	0.1	0.48	0.48	0.24	0.37	0.42	0.17	0.48	0.48

Table 2: Estimated equations of reported length of stay for different recall periods

	One Month		Two months		Three months		Four months		Six months		12 months	
	B	t value	β	t value	β	t value	β	t value	β	t value	β	t value
Age	0.000	1.940	0.001	0.930	0.002	1.270	0.001	0.520	0.009	5.240	0.036	10.110
Bad health	0.013	0.780	0.086	1.160	0.220	1.730	0.151	1.050	0.565	3.370	5.475	8.730
Very Bad health	0.074	2.290	0.140	1.010	0.307	1.280	0.397	1.460	1.880	5.930	2.219	6.710
Const	-0.005	-0.590	0.088	2.280	0.108	1.610	0.292	3.880	0.023	0.270	-0.509	-2.930
R^2	0.00		0.00		0.00		0.00		0.01		0.02	

Figure 1: Classification of self-reported use of health services

Actual use \ Reported use	$x_i^w = 0$	$x_i^w > 0$		
$z_i^w = 0$	**S1** True negative No recall error	**S2** False positive		
$z_i^w > 0$	**S3** False negative	**S4** True positive		
		(a) under-reporting	(b) exact match	(c) over-reporting

Figure 2 : Functions representing recall error over time

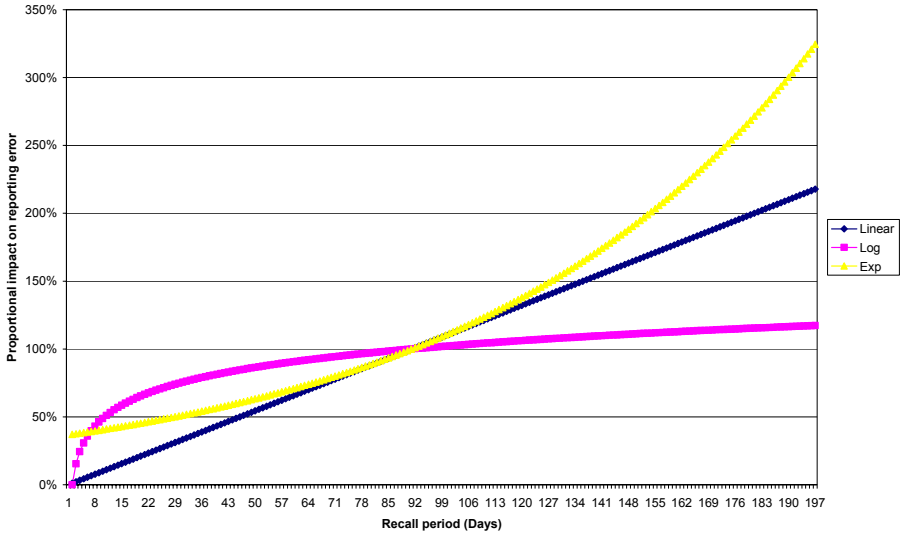

Figure 3 : Sample mean by recall period

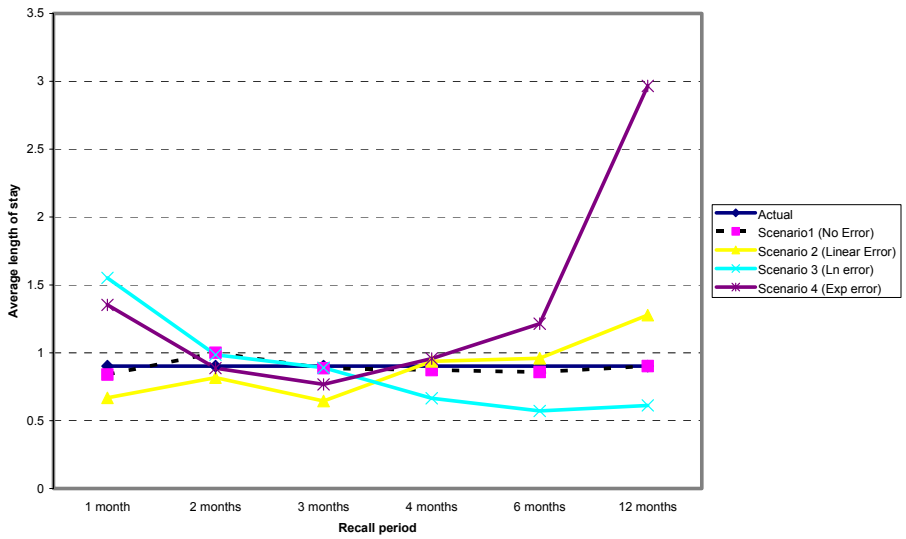

Figure 4: MSE by recall period