Contents lists available at ScienceDirect

Journal of Health Economics

journal homepage: www.elsevier.com/locate/econbase

Forgetting to remember or remembering to forget: A study of the recall period length in health care survey questions

Gustav Kjellsson[a,b,*], Philip Clarke[c], Ulf-G. Gerdtham[a,b,d]

[a] Department of Economics, Lund University, P.O. Box 7082, SE-220 07 Lund, Sweden
[b] Health Economics & Management, Institute of Economic Research, Lund University, Sweden
[c] Centre for Health Policy, Programs and Economics, School of Population Health, The University of Melbourne, Melbourne, Victoria 3010, Australia
[d] Center for Primary Health Care Research, Lund University/Region Skåne, Sweden

ARTICLE INFO

Article history:
Received 11 February 2013
Received in revised form 19 January 2014
Accepted 21 January 2014
Available online 7 February 2014

JEL classification:
C42
C83
I10

Keywords:
Survey methods
Health survey
Hospitalization
Recall error
Recall periods

ABSTRACT

Self-reported data on health care use is a key input in a range of studies. However, the length of recall period in self-reported health care questions varies between surveys, and this variation may affect the results of the studies. This study uses a large survey experiment to examine the role of the length of recall periods for the quality of self-reported hospitalization data by comparing registered with self-reported hospitalizations of respondents exposed to recall periods of one, three, six, or twelve months. Our findings have conflicting implications for survey design, as the preferred length of recall period depends on the objective of the analysis. For an aggregated measure of hospitalization, longer recall periods are preferred. For analysis oriented more to the micro-level, shorter recall periods may be considered since the association between individual characteristics (e.g., education) and recall error increases with the length of the recall period.

1. Introduction

A large and growing number of health economic studies rely on survey-based self-reported data to obtain information on health care use, out-of-pocket expenses, and health behaviors. The design of these surveys will inevitably affect the result, and possibly, the conclusions of research, which, in turn, may influence our beliefs and future policy. One feature that varies greatly between different surveys is the period over which people are asked to recall prior events. A recent review of almost 90 country-level health surveys reports that the recall periods range from 2 weeks to 14 months with a significant proportion of surveys using either 1 or 12 months (Heijink et al., 2011). While information tends to be collected over longer recall periods for hospitalizations than physician visits, there is still a surprising degree of variation between surveys. For example, in the case of hospitalizations, 36% of the surveys use a one month recall period, while 46% use one year.[1]

It has been well established that self-reported behaviors such as health care use are subject to error. Gaskell et al. (2000) suggest four types of recall error:

> "Respondents may forget details on even entire events. Although less common, respondents may recall events that did not occur. These are referred to as errors of omission and commission, respectively . . . another type of error . . . [is] telescoping. Respondents may recall an event but report that it happened earlier than it actually did (backward telescoping) or report that it happened more recently (forward telescoping)."

It has also been recognized that the longer the recall period, the less accurate the reported estimates (Stull et al., 2009; Bhandari and Wagner, 2006). However, even though the likelihood of recall error

* Corresponding author at: Department of Economics, Lund University, P.O. Box 7082, SE-220 07 Lund, Sweden. Tel.: +46 46 2227911.
E-mail address: gustav.kjellsson@nek.lu.se (G. Kjellsson).

[1] Debate over the appropriate length of the recall period is not confined to health care use. See Arnold et al. (2013) for an examination of trade-offs when collecting information on childhood illness in developing countries.

0167-6296 © 2014 The Authors. Published by Elsevier B.V. Open access under CC BY-NC-ND license.
http://dx.doi.org/10.1016/j.jhealeco.2014.01.007

increases with longer recall periods, so does the amount of information provided, so there is a potential trade-off between recall error and information. The presence of this implicit trade-off when designing health surveys may explain the high degree of variation in recall periods used for the same types of health care.

The appropriate length of recall period also depends on the type of health care consumption and the intended use of the information. First, events that are more salient call for a longer recall period, while events that are more frequent call for a shorter period; the probability of remembering spending a night at the hospital is likely to be higher than the probability of remembering a visit to a GP. Second, while an overall average for a given target period may be well approximated (given no seasonality) by scaling up an estimate from a shorter recall period, the same exercise with the objective of estimating individual health care use for an infrequent and unpredictable event will probably yield estimates that are at best weakly related to the actual use (e.g., Deaton, 1997). Third, because individual characteristics such as cognitive ability or socioeconomic variables also potentially affect the process of recalling information (Bhandari and Wagner, 2006; Bound et al., 2001), the consequences of recall error may be more severe if the data is intended for studying the relationship between consumption of care and socioeconomic variables (e.g., studying demand or consumption using regression analysis). Unless recall error is orthogonal to individual characteristics, it is problematic to recover the relative impact of variables, and the bias induced by the recall error may falsely affect our understanding of the relationships of interest (e.g., Wooldridge, 2010).

While numerous studies have compared reported and actual use for a range of health care variables, almost all previous studies have examined only one period over which the respondent is asked to recall their prior use (for an overview see Bhandari and Wagner, 2006). It is hard to draw general conclusions about the nature of recall error as there are many differences between such studies, including the type of health care use examined, the nature of the survey (e.g., face-to-face interview vs. mail questionnaire), and the characteristics of the respondents. One way to control for these confounders is by allocating respondents to versions of the same question that differ only in the time period over which they are asked to recall past use. Das et al. (2012) performed such an experiment in India finding significant variation in reported doctor visits between those collected using a one-month recall period and those collected using four weekly reports, as well as differences in reporting behavior between rich and poor. However, this experiment could only document differences in patterns of reporting, not differences in patterns of *reporting error*, i.e., the degree to which self-reported use differs from recorded information on actual use.

The primary aim of this study is to use a large survey experiment to examine the role of the length of recall period in recall error about hospitalization. By comparing self-reported data gathered from a public health survey with registered data (treating the latter as the gold standard), we explore the nature of recall error and examine its implications for two aspects of survey design. First, we extend the framework suggested by Clarke et al. (2008) to determine an optimal length for a recall period for an aggregated measure of hospitalization, i.e., estimating the mean number of nights of stay. Second, we report how individual characteristics affect the quality of self-reported data and examine the degree of association between years of schooling (a proxy for cognitive ability) and recall errors over different recall periods. We know of no comparable published experiment to quantify recall error for a type of health care use. Therefore, this study contributes to the literature by exploiting variation in the length of the recall period for a large sample.

2. Description of a household survey experiment

This household survey experiment uses data from two different sources—Swedish registry data and a public health survey from the most southern Swedish county council (i.e., Region Skåne)—to examine how the length of the recall period affects the accuracy of self-reported hospitalization. Respondents in the public health survey were asked

> "How many nights were you hospitalized during the last year/X months?"

Respondents were assigned to one of four groups, each with a different recall period, based on the quarter of their birth. For respondents born in the months January to March (Group 1), April to June (Group 2), July to September (Group 3), and October to December (Group 4), the lengths of the recall period were one month ($w = 30$), three months ($w = 91$), six months ($w = 183$), and twelve months ($w = 365$), respectively. The wording of the question, specifically asking for hospital nights rather than days, was chosen to assure that the respondents' perception of the event corresponded to the registered event. In addition to this question, respondents were asked to state whether they had been admitted to the hospital during the last three months (admission).

2.1. Experimental population

The population in the public health survey, Folkhälsoenkät Skåne 2008 (Rosvall et al., 2009), consists of all individuals from the ages of 18 to 80 living in Region Skåne, one of the 21 county councils of Sweden. A total of 28 198 out of 52 142 respondents answered the survey. This study is based on the subset of 7500 respondents who answered the questionnaire on the web because the exact date of their survey completion was known.[2] The survey data, which also include information on self-assessed health, living conditions, and background information such as age and country of birth, are linked to registry data on income, education, and hospitalization. The link to registry data allows us to compare self-reported hospitalization with registered number of nights spent at a hospital. The National Board of Health and Welfare (2009) has stressed that the quality of registry data is high for the date of admission to and discharge from the hospital. The registry data include hospitalizations at public hospitals within Region Skåne as well as in other county councils, but they do not include nights spent at private hospitals. As the registry data do not include private care, we may overestimate the number of individuals who falsely reported hospital nights. The bias we observe may therefore be due to consumption of private care. However, this is unlikely to have a significant impact on the results since the share of private in-patient care in Region Skåne is less than one percent (in terms of hospital admission). Out of the 7500 observations, 365 have missing values on either reported or registered hospitalizations and an additional 136 have missing values on either years of schooling or income. Therefore, the analysis uses the remaining 6999 observations.

The definitions of the variables are explained in Table 1a. As the length of the recall period the respondent is exposed to is determined by the quarter of birth and not by randomization, it is important to compare the descriptive statistics for the four groups. Table 1b shows that the four groups are equal in terms of sex, non-Nordic origin, and health care consumption (i.e., the proportion being admitted during the last three month, admission, and the

[2] For the remainder of the sample only the dates of receipt of the mailed returned survey form were known, so the exact recall period could not be obtained.

Table 1a
Definitions of variables.

Variables	Definitions
School	Years of schooling
Bad health	1 if health is reported bad or very bad (=0 if fair, good, or very good)
Income	Total income – individual level in 2007
ln (income)	Logarithm of total income – individual level in 2007
Non-Nordic	1 if born outside of the Nordic countries
Male	1 if male, 0 if female
Age	Age of the respondent
Age 18–30	1 if 17 < age ≤ 30, 0 otherwise
Age 31–45	1 if 30 < age ≤ 45, 0 otherwise
Age 46–60	1 if 45 < age ≤ 60, 0 otherwise
Age > 60	1 if 60 < age ≤ 75, 0 otherwise
Y_i^w	Registered hospitalization (number of nights) during the recall period
Y_i^S	Registered hospitalization (number of nights) during the previous 365 days
X_i^w	Reported number of hospital nights during the recall period
abs_error	1 if the respondent reported any hospitalization without having a registered event during the recall period ($X_i^w > 0, Y_i^w = 0$), 0 otherwise
pos_error	1 if the respondent reported no hospitalization, but had at least one registered night during the period ($X_i^w = 0, Y_i^w > 0$) otherwise
neg_error	1 if either a false positive or a false negative report (*false_pos* = 1 or *false_neg* = 1), 0 otherwise
abs_error	The absolute difference between reported and registered hospital nights, $\left\|X_i^w - Y_i^w\right\|$
pos_error	The degree of positive errors, ($X_i^w - Y_i^w$ if $X_i^w > Y_i^w$, = 0 otherwise)
neg_error	The degree of negative errors, ($X_i^w - Y_i^w X_i^w - Y_i^w$ if $X_i^w < Y_i^w = 0$, otherwise)
Admission	1 if registered admission to hospital during the last 3 month
alt_bin_error	1 if (a) the respondent reported admission and Admission = 0; (b) the respondent reported no admission and Admission = 1

Note: The notation X_i^w, Y_i^w, and Y_i^S is used in the framework presented in Section 3.

Table 1b
Descriptive statistics.

Variables	(1)	(2)	(3)	(4)	(5)	(7)	(8)
	Total	$w = 30$	$w = 91$	$w = 183$	$w = 365$	*F*-test	Prob.
Age	43.89	43.44	42.92	44.96	44.25	6.178	0.000
Male	0.520	0.521	0.532	0.499	0.525	1.398	0.241
Non-Nordic	0.084	0.090	0.079	0.076	0.092	1.386	0.245
School	12.670	12.720	12.744	12.772	12.476	5.670	0.001
Bad health	0.046	0.043	0.042	0.042	0.057	2.147	0.092
Income	263 787	270 041	264 125	262 072	259 396	0.871	0.455
Y_i^S	0.369	0.329	0.411	0.413	0.327	0.380	0.767
Y_i^w	0.153	0.029	0.088	0.147	0.327	12.95	0.000
Admission	0.019	0.015	0.018	0.021	0.021	0.789	0.499
alt_bin_error[a]	0.022	0.021	0.021	0.017	0.026	1.238	0.336
abs_error	0.119	0.045	0.055	0.153	0.213	10.28	0.000
pos_error	0.074	0.043	0.045	0.132	0.078	4.363	0.005
neg_error	−0.045	−0.002	−0.009	−0.021	−0.134	14.73	0.000
false_neg	0.008	0.000	0.001	0.008	0.024	26.79	0.000
false_pos	0.015	0.012	0.009	0.020	0.020	3.823	0.009
bin_error	0.024	0.012	0.009	0.028	0.043	20.02	0.000
#	6999	1704	1722	1662	1911		

Note: Columns 1–5 show the mean values of the variables for the total sample and the four groups that are exposed to different recall periods. Columns 7 and 8 show the *F*-statistics and corresponding *p*-values for testing the hypothesis of equal means for the four groups. The variables *abs_error* to *bin_error* are further discussed in Section 4.2.

[a] For alt.binary the number of observations is reduced to 6840 (1661; 1671; 1635; 1866).

number of hospital nights in the last year, denoted Y_i^S). However, the respondents in Group 4 ($w = 365$) appear to be slightly different in terms of schooling and self-assessed bad health (confirmed using *F*-tests of equal means). We may also use the binary question of admission that uses a common three-month recall period for the four groups to test whether they differ by their proneness to misreport (*alt_bin_error* is coded as 1 if the respondent either [a] reports being admitted to the hospital without having a registered event during the recall period, or [b] reports no admission although has had at least one registered admission during the recall period[3]). As everyone is exposed to the same recall period in this question, we can examine whether there are any systematic differences between the groups in terms of reporting incorrectly. Even though there are statistically significant differences for some observable characteristics, the groups do not differ in their degree of recall error. *F*-test of equal means cannot reject the null hypothesis of the four groups being equal.

2.2. Results of the experiment

In line with previous studies, the level of agreement between self-reported and registered data decreases with the length of the recall period (cf. Bhandari and Wagner, 2006). The percentages of correctly self-reported hospitalization are 98.5, 98.4, 96.0, and 93.6 for $w = 30$, $w = 91$, $w = 183$, and $w = 365$, respectively. However, the pattern of error is asymmetric (as illustrated in Fig. 1b) when we distinguish between positive error (i.e., over-reporting due to commission and forward telescoping) and negative error (i.e., under-reporting due to omission and backward telescoping). Importantly, short recall periods such as one month are not free

[3] Later in the article we will define (a) and (b) as false positive and false negative, respectively. Separate analysis of the proportions of false negatives and false positives yields the same conclusion.

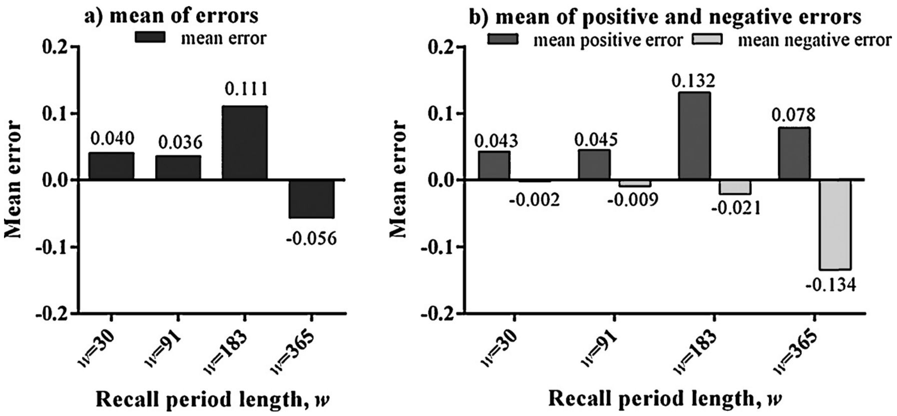

Fig. 1. Mean errors by recall period. *Note*: The graph in a) exhibits the mean number of errors (i.e., bias) for each of the four recall periods and the graph in b) exhibits the mean number of positive errors (over-reporting) and negative errors (under-reporting) or each of the four recall periods. F-tests reject the null of equal means for the four groups (For a) Table 5b, for b) F-test = 7.27, p = 0.000).

from error, and in fact show a large positive error on average (0.43). By contrast, there are very small negative errors for short recall periods, but those errors rise dramatically for the longest (one-year) recall period in the study. In actual numbers, the group exposed to $w = 30$ had 49 registered hospital nights during the recall period. Of these, 45 were correctly reported. The respondents reported 73 additional hospital nights that were not registered. For the longest recall period, $w = 365$, there were 625 registered hospital nights (368 reported and 257 unreported). The respondents also reported an additional 150 hospital nights that were not registered. (Table A1 in the supplementary online appendix reports these numbers for all four recall periods.) So while the proportion of the sample making errors rises continuously at a considerable rate, the degree to which this leads to bias in reporting of the mean is relatively constant over the year (e.g., Fig. 1a illustrates that for the longest period of one year the bias is −0.054 and imparts only a slightly greater bias in absolute terms than the bias for one month [0.040]).

Some further insights into the nature of recall error can be gained by plotting the error (the difference between reported and registered hospital nights) for each individual with some evidence of hospitalization, either during the recall period or up to a year prior to the recall period.[4] Fig. 2 plots the degree of error, where positive values indicate over-reporting, against the day of hospitalization closest to the recall boundary. This is defined by the day of admission for spells ending after the recall boundary, and by the day of discharge for spells beginning before the boundary. The x-axis illustrates the time in days from the recall boundary: days prior to this boundary fall outside the recall period and are denoted as positive values, while days after the boundary and up to the day of the survey fall within the recall period and are denoted as negative values. So for example, if we consider the 30 days recall period, the x-axis starts at −30 (i.e., the day of the survey), day 0 is the recall boundary, and day 10 would be 40 days prior to day of the survey, or 10 days outside the recall period.

Fig. 2 presents these plots for the four recall periods. The graphs illustrate that respondents who over-reported nights often had hospitalizations in the period of up to 120 days outside the recall period, which is strongly suggestive of forward telescoping (i.e., individuals include episodes that happened before the recall boundary). On the other side of the recall boundary (i.e., when the hospitalization episode is within the recall period), the graphs suggest—especially for the longer recall periods of 183 and 365 days—that the propensity to under-report increases and negative errors becomes larger closer to the recall boundary. The data do not, however, allow for discrimination between omission and backward telescoping.

We next explore the implications of the pattern of recall error revealed by the experiment for the choice of recall period for two situations: for estimating an overall summary measure of hospital use and for studying the relationship between hospitalization and individual characteristics.

3. Implications for choosing an optimal recall period for an aggregated mean

3.1. Framework

Clarke et al. (2008) develop a framework for evaluating the optimal recall period and apply it to recall data involving a single recall period (use over the last three months) and therefore have to make assumptions regarding the error over shorter durations. By contrast, the respondents in our data were exposed to recall periods of varying lengths and, thus, we can perform an analysis with fewer restrictive assumptions.

Following Clarke et al. (2008), we denote the variable of interest for each individual i in a population of size N as Y_i, which in our application is the registered hospitalization. The survey design problem is to estimate an aggregated measure of mean hospitalization within a target period S. The target period can be divided into sub-periods, and in a survey individuals may be asked to state their hospitalization during a sub-period of the target period. We refer to this sub-period as the recall period and denote it as w (in our application $w \leq S$). We further denote self-reported hospitalization during this period as X_i^w and actual hospitalization within this period as Y_i^w (thus, the index w refers to the length of the recall period).

There would be no problem using X_i^w as an estimate of Y_i^w if individuals had perfect recall during w.[5] It may also be possible to choose a recall period short enough to eliminate any recall error.

[4] Note that for $w = 365$ we only observe hospitalization for 265 days prior to the recall boundary (i.e., 365 + 265 days in total).

[5] This framework does not consider other sources of measurement errors that are not associated with recall problems (strategic behavior and false reporting). We believe that hospitalization is neither a sensitive question nor a question likely to provoke strategic behavior of the respondents.

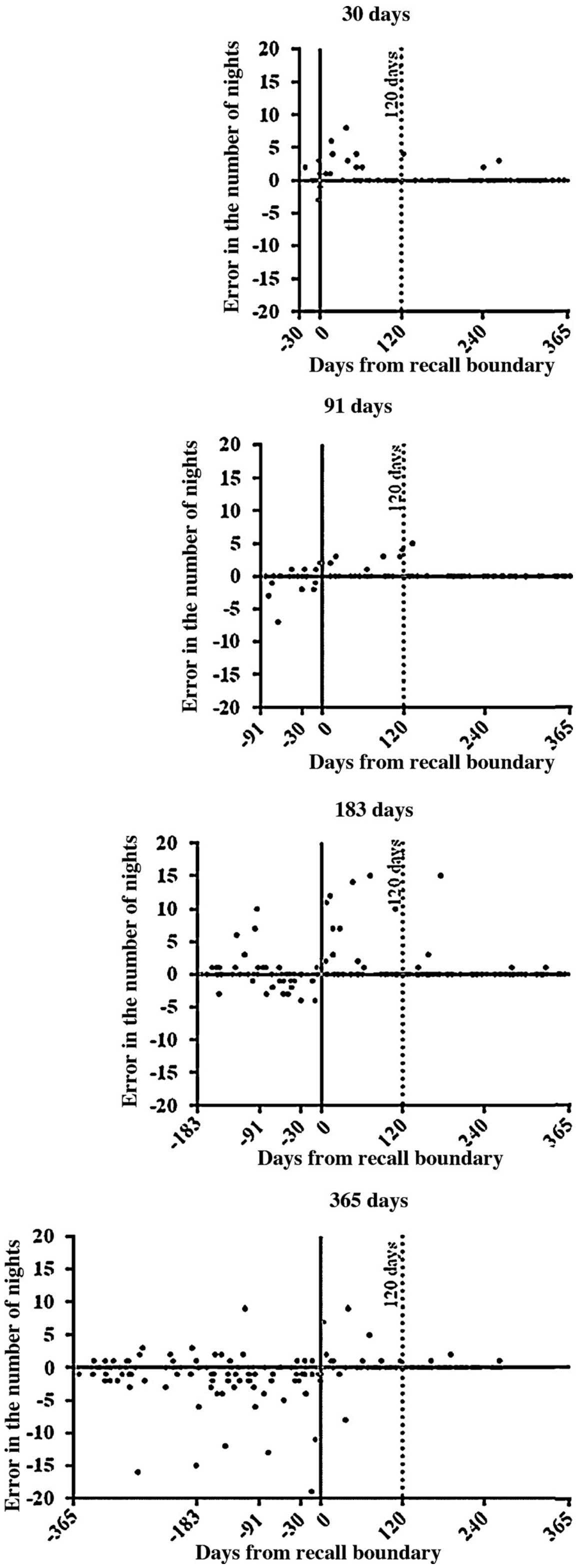

Fig. 2. Individual level error in relation to the recall window. *Note*: The figure plots error (the difference between reported and actual hospital nights) on the y-axis and the distance between the recall boundary and the closest day of hospitalization on the x-axis conditional on some evidence of hospitalization during the recall period plus a span of additional 365 days. (The closest day is defined as the day admission for spells ending after the recall boundary and discharge for spells starting before the recall boundary). Negative values on the y-axis denote days between survey completion and the recall boundary and positive values denote days prior to the recall boundary. Note that for the w = 365 we only observe hospitalization for 365 + 265 days.

However, if our recall period is shorter than the target period, we need to undertake an imputation process to estimate the hospitalization within the target period. If policy makers are interested in the mean of annual hospitalization (i.e., $S = 365$), then the question we need to answer is whether it is better to ask individuals to report hospitalization for a shorter recall period and then undertake the imputation process or to use the target period as the recall period (i.e., $S = w$). As discussed in Section 2, our data observe $w = \{30, 91, 183, 365\}$.

To evaluate the appropriate length of the recall period, Benítez-Silva et al. (2004) and Das et al. (2012) focus only on how w affects the bias of X_i^w as an estimator of mean use during the target period. However, focusing on unbiasedness alone does not consider the mechanisms at play; there is a possible trade-off between more information and bias. Since the information on individuals' hospitalization increases with the length of the recall period, the variance reasonably decreases. Clarke et al. (2008) further suggest combining the variance and the bias in a single measure using quadratic loss so that the survey design problem is to choose w to minimize root mean square error (RMSE). We exploit the varying length of the recall periods by comparing the four recall periods for bias, variance, and RMSE. Next, we first present the framework in Clarke et al. (2008) and then also consider the different nature of recall errors within their framework.

Formally, Clarke et al. (2008) let

$$X_i^w = Y_i^w + \upsilon_i^w \tag{1}$$

where υ_i^w represents the measurement error. Our intended objective is to obtain a measure of mean use in the target period $E(Y_i^S)$. Given a certain recall period, an obvious estimator (given no seasonality) would be to scale up the reported hospital nights within the sub-period to an estimate of the target period as:

$$\bar{X}_w^S = N^{-1}\sum_{i=1}^{N}\left(\frac{S}{w}\right)X_i^w \tag{2}$$

To evaluate the length of the recall period, we consider the two sides of the potential trade-off; recall bias and less information. We first estimate the expected value, the variance, and the bias of $\bar{X}_w^S$ for the four different recall periods as:

$$E(\bar{X}_w^S) = N^{-1}\left(\frac{S}{W}\right)\sum_{i=1}^{N}E(X_i^w) \tag{3}$$

$$\text{Var}(\bar{X}_w^S) = N^{-2}\left(\frac{S}{W}\right)^2\sum_{i=1}^{N}\text{Var}(X_i^w) \tag{4}$$

$$\text{Bias}(\bar{X}_w^S) = E(\bar{X}_w^S) - E(Y_i^S) \tag{5}$$

Following Clarke et al. (2008), we also use RMSE to combine the bias and variance into a single measure:

$$\text{RMSE}(\bar{X}_w^S) = \sqrt{[\text{Bias}(\bar{X}_w^S)]^2 + \text{Var}(\bar{X}_w^S)} \tag{6}$$

To get an estimate of the bias, we need an empirical definition of $E(Y_i^S)$. An obvious candidate is the registered hospitalization during the target period for each of the four groups. This is also in line with the theoretical concept and the actual survey design problem: how good is $E(\bar{X}_w^S)$ as an estimator of $E(\bar{X}_w^S)$? However, as a robustness test, and to ensure that seasonality does not introduce further bias, we also calculate the bias as:

$$\text{Bias}(\bar{X}_w^S) = E(\bar{X}_w^S) - \left(\frac{S}{w}\right)E(Y_i^w) \tag{7}$$

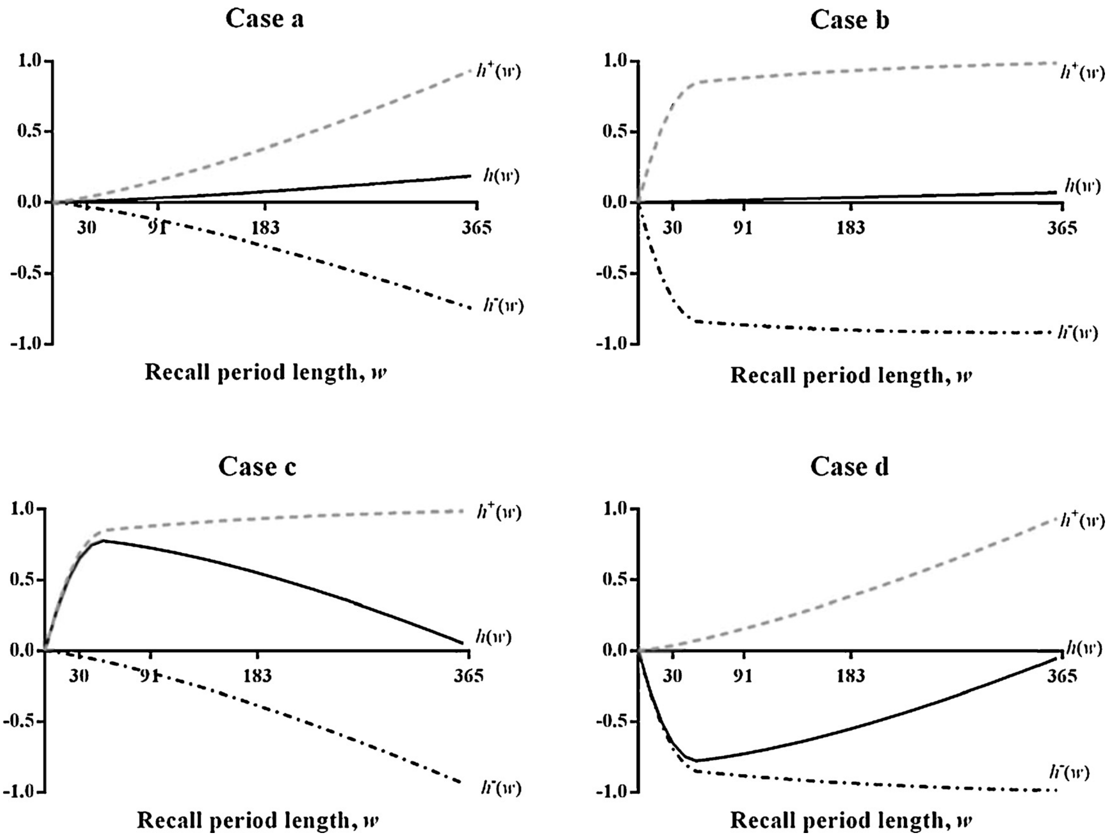

Fig. 3. Stylized graph of hypothetical error structures. *Note*: The graph represents four hypothetical error structures; $h(w)$ is the sum of $h^+(w)$and $h^-(w)$, which respectively represent the individual's proneness to over- or underreport.

That is, scaling up the bias for the specific recall period to an estimate of the bias within the target period. Clarke et al. (2008) also develop a similar framework for a binary case, where the survey design problem consists of estimating the probability of spending at least one night at the hospital during one year. We consider the binary case to be redundant for this application, but results can be found in a supplementary online appendix.

To extrapolate the RMSE over the interval 1 to S, Clarke et al. (2008) introduces the two functions $h(w)$ and $g(w)$ to relate the mean and the variance for a given w to the moments for the target period S. They further make the classical error-in-variables assumption, i.e., that υ_i^w is independent of Y_i^w but allows

$$E(\upsilon_i^w) = h(w)\left(\frac{w}{S}\right)\mu \tag{8}$$

and

$$\mathrm{Var}(\upsilon_i^w) = g(w)\left(\frac{w}{S}\right)\sigma^2 \tag{9}$$

where $\mu = E(Y_i^S)$ and $\sigma^2 = \mathrm{Var}(Y_i^S)$. Thus, Eqs. (8) and (9) show that the mean and the variance of the recall error depend on the length of the recall period, and increasing values of the two functions h and g imply increasing recall error or dispersion of error (i.e., noisy measurements), respectively. Clarke et al. (2008) further assume that there exists a period short enough to eliminate all recall errors, i.e., $g(1) = h(1) = 0$, and both h and g are monotonic functions over the interval 1 to S.

The results of our experiment presented in the previous section suggest relaxing the second assumption. Considering the types of errors discussed in the introduction, we decompose $h(w)$ into a function of two processes that cause the errors: the individual's proneness to over-reporting (i.e., through commission and forward-telescoping) and under-reporting (i.e., through omission and backward telescoping), defined as $h^+(w)$ and $h^-(w)$, respectively. We therefore redefine Eq. (8) as

$$E(\upsilon_i^w) = h(w)\left(\frac{w}{S}\right)\mu = (h^+(w) - h^-(w))\left(\frac{w}{S}\right)\mu \tag{10}$$

Splitting the error in this way allows the optimal recall period to be determined for a wide variety of different error structures. For illustrative purposes, Fig. 3 shows four stylized graphs of positive and negative error structures, in which the individual's propensity to over-report (and under-report) is limited either to increasing proportionally over the period or increasing rapidly in the beginning and being relatively constant over the remainder of the interval.

Case (a) represents a symmetric error structure in which the propensities to over- and under-report increase proportionally (with slightly different speeds); case (b) represents a symmetric error structure in which the propensities to over- and under-report increase rapidly for short recall periods, but stays relatively constant over the rest of the interval. For both of these cases, $h(w)$ is monotonically increasing, and shorter periods are preferred in terms of bias (although the propensities to over- and under-report tend to cancel out for all recall periods). By contrast, case (c) and case (d) represent asymmetric error structures, in which either of the two processes increases rapidly in the beginning and is then fairly constant over the period, while the other process increases proportionally over the period.

It is clear, as can be seen in case (c) and case (d) in Fig. 3, that the recall error process may not be monotonic increasing or

Table 2
RMSE, variance, bias.

	$\text{Var}(X_i^w)$	Eq. (5): Bias $= E(\bar{X}_w^S) - E(Y_i^S)$			Eq. (7): Bias $= E(\bar{X}_w^S) - (S/w)E(Y_i^w)$		
		RMSE	$\text{Var}(X_w^S)$	$\text{Bias}(\bar{X}_w^S)$	RMSE	$\text{Var}(\bar{X}_w^S)$	$\text{Bias}(\bar{X}_w^S)$
$w = 30$	0.592	0.562	0.051	0.514	0.294	0.051	0.493
$w = 91$	1.979	0.162	0.018	0.087	0.039	0.018	0.144
$w = 183$	3.717	0.138	0.009	0.100	0.058	0.009	0.221
$w = 365$	3.132	0.069	0.002	−0.056	0.005	0.002	−0.056

Note: The table shows $\text{Var}(X_i^w)$ and the RMSE, $\text{Var}(\bar{X}_w^S)$, and $\text{Bias}(\bar{X}_w^S)$ for the two definitions of bias, Eqs. (5) and (7), for each of the four recall periods. Using paired bootstrap (9999 replications), the difference in RMSE between $w = 31$ and the remaining three recall periods is statistically significant (10% for $w = 91$ and $w = 183$, and 5% for $w = 365$).

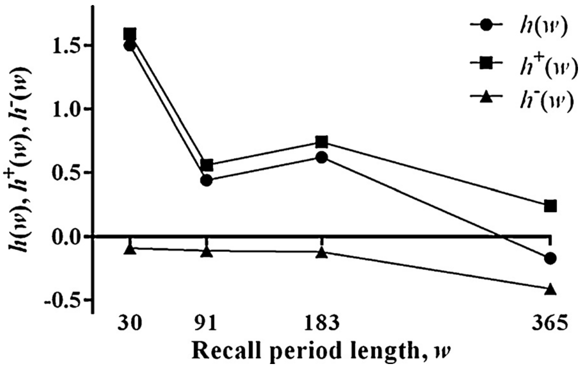

Fig. 4. Observed error structure. *Note*: Fig. 4 plots the four observations of $h(w)$, $h^+(w)$, and $h^-(w)$ as defined in *Eq*(10) for comparisons with the hypothetical error structures in Fig. 3. $h(w)$, $h^+(w)$, and $h^-(w)$ are calculated as $h(w) = \frac{365}{w}\left(\frac{E(v_i^w)}{E(Y_w^S)}\right)$, $h^+(w) = \frac{365}{w}\left(\frac{N^{w+}}{N}\right)\left(\frac{E(v_i^{w+})}{E(Y_w^S)}\right)$, $h^-(w) = \frac{365}{w}\left(\frac{N^{w-}}{N}\right)(\frac{E(v_i^{w-})}{E(Y_w^S)})$. Here, v_i^{w+} (v_i^{w-}) denotes positive (negative) errors and N^{w+} (N^{w-}) the number of respondents with positive (negative) errors.

decreasing functions of the period length.[6] This implies that the bias is not necessarily largest for the longest recall period. Rather, the determinant of bias is the relative magnitude of $h^+(w)$ and $h^-(w)$, which stresses the importance of not using a recall period in which either of these two processes strongly dominates the other (compare case (c) and case (d)). Note however that unlike a change in $h(w)$, which is not necessarily related to $g(w)$, conditional changes of either $h^+(w)$ or $h^-(w)$ will affect the dispersion, and thus the RMSE, through $g(w)$.

3.2. Applying the results of the recall error experiment to determine an optimal recall period

This section presents the results of using experimental data to determine an optimal recall period. The pattern of the results presented in Table 2 and Fig. 4 is clear.

Table 2 reports that while the variance of X_i^w generally increases for longer w,[7] $\text{Var}(\bar{X}_w^S)$ decreases. The decrease in variance is anticipated since information about individuals' hospitalization increases with the length of the recall period. However, the results show that the expected trade-off between variance and bias is not present. With the exception of going from $w = 91$ to $w = 183$, the size of bias decreases as w increases (it is positive for all recall periods such that $w < S$, while for $w = S$ the bias is negative).[8] Thus, as the graph in Fig. 4 illustrates, when comparing the four recall periods for bias, variance, and RMSE our experiment indicates that the one-year period is preferable (i.e., setting $w = S$). In terms of RMSE and bias, the major difference is between $w = 30$ and the three longer periods (i.e., $w > 30$). These differences are also statistically significant using paired bootstrap (at a 5% level for $w = 365$ and at a 10% level for $w = 91$ and $w = 181$). The results are robust to using the alternative definition of bias defined in Eq. (7), and the pattern is even clearer for the binary case presented in the online appendix.[9]

The decrease in bias is driven by the increasing number of under-reporters (negative errors) as the length of the recall period increases. Remember the pattern from Fig. 1; although the amount of total errors increases with the length of the recall period, the bias (without extrapolation) is relatively constant and the degree of positive errors does not increase as much as the degree of negative errors.[10] Thus, the longer recall period is preferred, not because the respondents give a better estimate of their hospitalization, but because the respondents that under-report balance out the respondents that over-report. In terms of the optimal recall period framework, our observations of $h(w)$, plotted as circles in Fig. 4, are decreasing in the interval from $w = 30$ to S. The decreasing pattern is caused by the asymmetry of the two recall processes (also plotted in Fig. 4); $h^-(w)$ is monotonically increasing, while $h^+(w)$ (i.e., the process of commission or telescoping) appears to be high even for the short recall period and to decrease, or at least remain relatively constant over the interval. That is, $h^+(w)$ strongly dominates $h^-(w)$ in the shorter periods, reiterating that even though the degree of recall errors increases over the interval, over- and under-reporting become more equally distributed. That respondents tend to overstate their hospitalization for shorter recall periods and under-report for longer is also in line with previous research (for an overview see Bhandari and Wagner, 2006).

The graphs in Fig. 2 previously presented in Section 2 further support this interpretation, as telescoping appears to generate large positive errors in a short recall period, while the propensity to forget increases with the duration of the recall period. Thus, one is unlikely to forget a recent event or to date it before the start of the recall period (e.g., backward telescoping), but may very well include an event that occurred before the start of the recall period (e.g., forward telescoping). This is also in line with psychological

[6] Note that Clarke et al. (2008) do consider this division of errors in the binary case.

[7] The exception is the decrease between $w = 181$ and $w = 365$ in the continuous case in Table 2.

[8] The two anomalies in the pattern (i.e., the increase in bias and $\text{Var}(X_i^w)$ between $w = 91$ and $w = 183$) disappears if we exclude the two respondents who have more than 100 registered hospital nights within a period of $w + 365$ days.

[9] Furthermore, calculating bias, variance, and RMSE of the alternative question as a placebo-analysis also confirms that there are no general differences between the groups. Further support for the result is that the aggregated analysis is also invariant to the exclusion of individuals with absolute recall errors >15 and >10. The same applies for excluding individuals with hospitalizations >50 or >30 during the last year. Thus, the results are not driven by a few extreme observations. These results are available upon request.

[10] The analysis of the binary case, presented in the appendix, shows that an increase in the proportion of respondents who either report hospitalization without any registered event (false positive) or who fail to report a registered hospitalization (false negative) is driven by a substantial increase in the proportion of false negatives.

survey research in this area (cf. Sudman and Bradburn, 1973; Rubin and Baddeley, 1989; Huttenlocher et al., 1988).

4. Association between socioeconomic variables and recall error

4.1. Background

Even though a longer period is preferred for an aggregated measure of hospitalization, the appropriate length of the recall period may be different if data is intended for further analysis. For example, recall error induces serious bias if it is systematically associated with any observed or unobserved characteristics of the respondents if the objective is to study the relationship between consumption of care and socioeconomic variables. The second part of the analysis therefore examines whether the association between years of schooling and recall errors differs with the length of the recall period. The motivation for using years of schooling as an example of a socioeconomic variable or individual characteristic arises from the possible link between cognitive ability and education. Generally, the quality of self-reported measures depends on the cognitive process of recalling information. Cognitive psychology highlights four parts of this process—comprehension of the question, retrieval of information from memory, assessment of the correspondence between the retrieved information and the requested information, and communication—that are all related to cognitive ability of the respondents (Bound et al., 2001; Tourangeau, 1984; Sudman et al., 1996; Bhandari and Wagner, 2006). As we cannot directly observe cognitive ability, we may consider years of schooling as a proxy for it.

Previous findings for the association between socioeconomic variables and recall error are mixed. For example, Das et al. (2012) show that for GP visits in India, a recall period of one month (compared to the gold standard of four weekly reports) may have huge implications for the association between socioeconomic status (income) and the consumption of care (even changing the sign of the coefficient). Others (e.g., Wolinsky et al., 2007; Ritter et al., 2001) find no socioeconomic differences (although Wolinsky et al. (2007) show that the health of the respondents matters). However, none of these studies are able to examine how the length of the recall period further affects the bias. By contrast, we are able to exploit the variation in the length of the recall period that our experiment provides.

4.2. Method

Health survey data is often collected for purposes other than estimating aggregated measures of use (e.g., the self-reported data may be used as the variable of interest in an inequality index, or as a dependent or independent variable in regression analysis). We know from the previous section that a longer period results in a larger share of individuals misreporting the length of their hospital stay. What also needs to be considered in this context is the relationship between reporting error and individual characteristics, and how this is affected by the length of the recall period. If recall errors are systematically associated with any observed or unobserved characteristics of the respondents, the coefficients in the regression analysis may be seriously biased and alter researchers' conclusions (cf. Wooldridge, 2010). While we cannot test for an association with unobserved characteristics, we can test whether the recall error is systematically associated with observed characteristics such as years of schooling.

To study the association between recall error and socioeconomic variables, we use regression analysis for two sets of outcome variables that may capture different aspects of recall error. The first set consists of three binary variables: *false_pos* is an indicator of a false positive report that equals one if the respondent reports any hospitalization without having a registered event during the recall period (i.e., $X_i^w > 0$ and $Y_i^w = 0$), zero otherwise; *false_neg* is an indicator of a false negative report that equals one if the respondent reports no nights although has had at least one registered night during the period (i.e., $X_i^w = 0$ and $Y_i^w > 0$), zero otherwise; *bin_error* is an indicator of either a false positive or a false negative report that equals one if either $X_i^w = 0$ and $Y_i^w > 0$ or $X_i^w > 0$ and $Y_i^w = 0$, zero otherwise. The second set consists of three continuous indicators of the degree of error in the reported number of hospital nights: *neg_error* denotes the degree of negative errors (equals $X_i^w - Y_i^w$ if $X_i^w < Y_i^w$, 0 otherwise), *pos_error* denotes the degree of positive errors (equals $X_i^w - Y_i^w$ if $X_i^w > Y_i^w$, 0 otherwise), and *abs_error* denotes the absolute value of the difference between reported and registered hospital nights ($|X_i^w - Y_i^w|$). (Descriptive statistics are presented in Table 1b.)

We examine how the length of the recall period affects the association between recall errors and years of schooling, denoted as $school_i$, by including interactions between $school_i$ and a vector of the recall period dummies denoted as $\boldsymbol{w}_i$. Although some of the outcome variables are binary, we estimate the following equation using OLS[11] with heteroskedasticity-consistent standard errors for all dependent variables (denoted as $error_i$)

$$error_i = \alpha + \boldsymbol{w}_i\boldsymbol{\gamma} + \varphi school_i + school_i\boldsymbol{w}_i\boldsymbol{\delta} + \boldsymbol{Z}_i\boldsymbol{\beta} + \varepsilon_i \quad (11)$$

$\boldsymbol{Z}_i$ denotes a vector of controls including demographics (i.e., income, sex, age, and country of birth) and other variables related to the mechanisms that may affect recall (see also Tables 1a and 1b). As chronically ill individuals, or individuals with lower general health, may visit the hospital more frequently, and therefore probably perceive the event as less salient (cf. Das et al., 2012), $\boldsymbol{Z}_i$ also includes a measure of self-assessed health. For the same reason, we also estimate the models with and without conditioning on the registered hospitalization. The next section discusses our findings for the association between recall error and socioeconomic variables.

4.3. Results

Table 3 initially presents the results without any interaction (i.e., assuming $\boldsymbol{\delta} = 0$ in Eq. (11)) and Table 4 then presents results with interactions between years of schooling and the recall period dummies (i.e., allowing $\boldsymbol{\delta} \neq 0$). The overall pattern in Table 3 indicates a significant association between years of schooling and recall error,[12] when controlling for demographics, income, and the number of registered nights at the hospital. Note that because the means of the dependent variables are generally small, the relative differences between the recall periods are large, even though the coefficients are small. More years of schooling significantly decrease the degree of absolute errors (*abs_error*, column 7 and 8) as well as the probability of either a false negative or a false positive report (*bin_error*, columns 1 and 2). These decreases are driven by the significant negative association between years of schooling and the degree of negative errors (*neg_error*, columns 9 and 10), and the probability of a false negative report (*false_neg*, columns 3 and 4), respectively. For the degree of positive errors (*pos_error*, columns

[11] On average a Linear Probability Model (LPM) and binary choice models such as Logit or Probit provide the same results (cf. Wooldridge, 2010, p. 563).

[12] The exceptions are *false_pos* and *pos_error*. These are not statistically significant, but we note the negative sign indicating that the propensity to make these errors decreases by years of schooling.

Table 3
Results without interaction terms.

Variables	(1) *bin_error*	(2) *bin_error*	(3) *false_neg*	(4) *false_neg*	(5) *false_pos*	(6) *false_pos*	(7) *abs_error*	(8) *abs_error*	(9) *neg_error*	(10) *neg_error*	(11) *pos_error*	(12) *pos_error*
w = 30	−0.030***	−0.027***	−0.023***	−0.020***	−0.007*	−0.008*	−0.157***	−0.063***	0.127***	0.045***	−0.030*	−0.018
	(0.005)	(0.005)	(0.003)	(0.003)	(0.004)	(0.004)	(0.033)	(0.022)	(0.028)	(0.016)	(0.017)	(0.018)
w = 91	−0.033***	−0.030***	−0.023***	−0.020***	−0.010**	−0.010***	−0.146***	−0.071**	0.120***	0.055***	−0.026	−0.016
	(0.005)	(0.005)	(0.003)	(0.003)	(0.004)	(0.004)	(0.036)	(0.029)	(0.028)	(0.019)	(0.023)	(0.024)
w = 183	−0.014**	−0.012**	−0.015***	−0.013***	0.001	0.001	−0.046	0.009	0.107***	0.059***	0.061*	0.068**
	(0.006)	(0.006)	(0.004)	(0.004)	(0.005)	(0.005)	(0.043)	(0.035)	(0.028)	(0.019)	(0.033)	(0.034)
School	−0.002**	−0.001*	−0.001***	−0.001***	−0.000	−0.000	−0.015***	−0.013**	0.007**	0.005*	−0.008	−0.008
	(0.001)	(0.001)	(0.000)	(0.000)	(0.001)	(0.001)	(0.006)	(0.006)	(0.003)	(0.003)	(0.005)	(0.005)
Male	−0.005	−0.005	−0.003	−0.003	−0.002	−0.002	0.009	0.015	−0.002	−0.006	0.008	0.009
	(0.004)	(0.004)	(0.002)	(0.002)	(0.003)	(0.003)	(0.024)	(0.022)	(0.016)	(0.013)	(0.019)	(0.019)
Non-Nordic	0.026***	0.027***	0.004	0.005	0.022***	0.022***	0.085*	0.114**	0.012	−0.013	0.097**	0.101**
	(0.009)	(0.009)	(0.005)	(0.005)	(0.008)	(0.008)	(0.049)	(0.044)	(0.024)	(0.017)	(0.043)	(0.043)
ln (income)	0.000	0.000	0.001***	0.001***	−0.001	−0.001	0.007**	0.006**	−0.004***	−0.004**	0.002	0.002
	(0.001)	(0.001)	(0.000)	(0.000)	(0.001)	(0.001)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)
Age 31–45	−0.009	−0.010*	−0.004	−0.005	−0.005	−0.005	0.008	−0.002	−0.015	−0.005	−0.006	−0.008
	(0.006)	(0.006)	(0.004)	(0.004)	(0.004)	(0.004)	(0.038)	(0.035)	(0.022)	(0.018)	(0.031)	(0.031)
Age 46–60	−0.014**	−0.013**	−0.010***	−0.009***	−0.004	−0.004	−0.040	−0.018	0.025	0.006	−0.015	−0.012
	(0.006)	(0.006)	(0.003)	(0.003)	(0.005)	(0.005)	(0.034)	(0.033)	(0.018)	(0.016)	(0.030)	(0.029)
Age > 60	−0.004	−0.005	−0.008*	−0.009**	0.003	0.003	0.002	−0.023	0.005	0.027	0.007	0.004
	(0.007)	(0.007)	(0.004)	(0.004)	(0.005)	(0.005)	(0.043)	(0.042)	(0.024)	(0.023)	(0.035)	(0.035)
Bad health	0.072***	0.060***	0.018**	0.004	0.054***	0.056***	0.581***	0.212**	−0.302**	0.020	0.279***	0.232***
	(0.016)	(0.016)	(0.009)	(0.009)	(0.014)	(0.014)	(0.157)	(0.091)	(0.138)	(0.058)	(0.080)	(0.074)
Y_i^w		0.011**		0.013***		−0.002***		0.335***		−0.291***		0.043
		(0.005)		(0.005)		(0.001)		(0.070)		(0.072)		(0.027)
Constant	0.063***	0.059***	0.032***	0.027***	0.032***	0.032***	0.289***	0.173**	−0.161***	−0.061	0.128**	0.113**
	(0.012)	(0.012)	(0.007)	(0.007)	(0.011)	(0.011)	(0.072)	(0.070)	(0.045)	(0.047)	(0.056)	(0.057)
Observations	6999	6999	6999	6999	6999	6999	6999	6999	6999	6999	6999	6999
R-squared	0.023	0.035	0.016	0.064	0.014	0.015	0.020	0.245	0.016	0.413	0.009	0.016

Note: Robust standard errors in parentheses. The outcome variables in columns 1–6 are binary indicators of a false negative report (*false_neg* in columns 3 and 4), a false positive report (*false_pos* in columns 5 and 6), and either a false negative or a false positive report (*bin_error* in columns 1 and 2). The outcome variables in columns 7–12 are continuous indicators of the magnitude of recall errors in the number of nights. Columns 7 and 8 present results for degree of absolute error; columns 9 and 10 present results for the degree of negative error; columns present results for degree of positive error.

* $p < 0.1$.
** $p < 0.05$.
*** $p < 0.01$.

Table 4
Results with interactions.

Variables	(1) *bin_error*	(2) *bin_error*	(3) *false_neg*	(4) *false_neg*	(5) *false_pos*	(6) *false_pos*	(7) *abs_error*	(8) *abs_error*	(9) *neg_error*	(10) *neg_error*	(11) *pos_error*	(12) *pos_error*
w = 30	−0.104***	−0.095***	−0.080***	−0.069***	−0.024	−0.026	−0.621***	−0.343**	0.485***	0.243**	−0.136	−0.100
	(0.030)	(0.029)	(0.018)	(0.016)	(0.024)	(0.025)	(0.205)	(0.146)	(0.173)	(0.099)	(0.115)	(0.116)
w = 91	−0.089***	−0.078***	−0.072***	−0.059***	−0.017	−0.019	−0.401	−0.074	0.434**	0.150	0.033	0.076
	(0.030)	(0.029)	(0.019)	(0.017)	(0.023)	(0.023)	(0.255)	(0.209)	(0.172)	(0.097)	(0.191)	(0.192)
w = 183	−0.040	−0.031	−0.080***	−0.069***	0.040	0.038	−0.304	−0.041	0.502***	0.273***	0.198	0.232
	(0.035)	(0.034)	(0.022)	(0.020)	(0.028)	(0.028)	(0.261)	(0.216)	(0.174)	(0.105)	(0.197)	(0.197)
School	−0.005**	−0.004**	−0.005***	−0.004***	−0.000	−0.000	−0.034**	−0.019*	0.028**	0.015**	−0.006	−0.004
	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)	(0.014)	(0.010)	(0.012)	(0.007)	(0.008)	(0.008)
(*w* = 30)*School	0.006**	0.005**	0.005***	0.004***	0.001	0.001	0.037**	0.022**	−0.029**	−0.016**	0.008	0.006
	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)	(0.015)	(0.011)	(0.012)	(0.007)	(0.009)	(0.009)
(*w* = 91)*School	0.004**	0.004*	0.004***	0.003**	0.001	0.001	0.020	0.000	−0.025**	−0.008	−0.005	−0.007
	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)	(0.018)	(0.015)	(0.012)	(0.008)	(0.014)	(0.014)
(*w* = 183)*School	0.002	0.002	0.005***	0.004***	−0.003	−0.003	0.021	0.004	−0.031**	−0.017**	−0.011	−0.013
	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.019)	(0.016)	(0.012)	(0.008)	(0.014)	(0.014)
Constant	0.101***	0.093***	0.073***	0.062***	0.028	0.030	0.529***	0.255*	−0.422***	−0.184**	0.107	0.071
	(0.027)	(0.026)	(0.017)	(0.016)	(0.021)	(0.021)	(0.180)	(0.130)	(0.151)	(0.094)	(0.101)	(0.104)
Observations	6999	6999	6999	6999	6999	6999	6999	6999	6999	6999	6999	6999
R-squared	0.025	0.036	0.019	0.067	0.015	0.016	0.021	0.245	0.018	0.414	0.010	0.016
F-test	2.842	2.572	4.800	4.258	2.328	2.333	2.729	2.220	2.923	2.069	1.159	1.137
Prob.	0.036	0.052	0.002	0.005	0.0723	0.072	0.042	0.084	0.033	0.102	0.324	0.333

Note: Robust standard errors in parentheses. The outcome variables in columns 1–6 are binary indicators of a false negative report (*false_neg* in columns 1 and 2), a false positive report (*false_pos* in columns 3 and 4), and either a false negative or a false positive report (*bin_error* in columns 1 and 2). The outcome variables in columns 7–12 are continuous indicators of the magnitude of recall errors in the number of nights. Columns 7 and 8 present results for degree of absolute error; columns 9 and 10 present results for the degree of negative error; columns 11 and 12 present results for degree of positive error. All models are estimated conditional on male, non-Nordic, ln (income), age, bad health. Models in even columns are also conditioned on the number of hospital nights, Y_i^w. The *F*-statistics test the joint significance of the interactions.

* $p < 0.1$.
** $p < 0.05$.
*** $p < 0.01$.

11 and 12) and the probability of a false positive report (*false_pos*, columns 5 and 6), the coefficients are negative but insignificant. However, these are also the dependent variables with the smallest variation.

All models are estimated with and without the number of registered *hospital nights* during the recall period (i.e., Y_i^w) among the controls (even columns presents results conditional on Y_i^w), as the recall period restricts the possible number of nights one can be hospitalized, which increases with the length of the period. However, increased hospitalization may also be seen as a path for the length of the period to affect recall error. The coefficient for hospital nights confirms that the degree of recall error increases with the number of registered hospital nights. The results for the other control variables are mixed, but some clear tendencies emerge. In line with our beliefs, the health dummy is significantly associated with all indicators except *false_neg* and *neg_error*; that is, individuals in bad health do not under-report (given a certain level of hospitalization), but their higher degree of over-reporting contributes to the increased overall misreporting (i.e., *abs_error* and *bin_error*). In contrast to health (and schooling), the income coefficients are a bit puzzling as individuals with higher income conditional on the other control variables are more error prone. The age-dummies are insignificant in all models except for *false_neg* and *bin_error*, in which individuals aged 30–60 (or 45–75 for *false_neg*) are significantly less likely to misreport than the reference group (18–30).

When estimating Eq. (11) allowing for $\delta \neq 0$, the results indicate that the association between schooling and recall is affected by the length of *w*. The schooling variable in Table 4 follows the same pattern as in Table 3. As we have interacted years of schooling with the recall period dummies, we interpret the coefficient of schooling as the association for the reference group ($w = 365$).[13] We further interpret the coefficients of the interactions as the difference in the association between the specific recall period and $w = 365$. Thus, to observe an increasing association between years of schooling and recall error, the coefficient of schooling should be significant and the interactions should increase in absolute terms and be of the opposite sign to schooling.

Although we observe exactly such a pattern for the binary indicator of either a false negative or a false positive report (*bin_error*, Table 3, column 1 and 2), the major differences in the association are generally among the reference group ($w = 365$) on the one hand, and respondents who are exposed to a shorter period ($w < 365$) on the other. Without conditioning on hospital nights (odd columns), the interactions are jointly significant in all models (for *false_pos* and *abs_error* only at a 10% level) except for the number of positive errors (*F*-statistics are presented in Table 4). The overall pattern remains when conditioning on the number of hospital nights (even columns). Nevertheless, the results are not as strong, primarily for the continuous outcomes variables. In general, the differences in the association between total recall errors (i.e., *bin_error* and *abs_errors*) are driven by differences in under-reporting (i.e., *false_neg* and *neg_error*). The coefficients may seem small in magnitude throughout the models and are statistically significant, in some cases, only at a 10% level. However, considering the small amount of observed hospitalization (and thus possible errors) for each recall period (i.e., the means of the dependent variables are low), we cannot expect to measure the differences with strong precision.

4.4. Applications and scope for correction

As an illustrative example, and to calibrate the extent to which the increased association between schooling and recall error affect summary statistics of education-related inequality in health care use, Table 5 presents the coefficients from bivariate regressions of hospitalization and years of schooling in columns 1–3 and the concentration index suggested by Erreygers (2009) in columns 4–7. Erreygers' index is a measure of absolute inequality adapted to bounded variables. Columns 1 and 4 present results for the reported hospitalization for each period, columns 2 and 5 present the results for the registered hospitalization for the corresponding sample, and columns 3 and 6 present the results for the registered hospitalization of the full sample (6999 observation) for each of the recall periods. Column 7 is discussed below.

Although not statistically significant, the differences between reported and registered hospitalizations in columns 1 and 2—and 4 and 5—are smallest for the shortest recall period and exhibit a tendency to increase (in absolute terms) with the length of recall period. This may be interpreted as, at least suggestive, evidence for the increasing association between schooling and recall error in the previous section to also be translated into an increasing bias. The exception is the difference of the inequality indicators for the longest recall period of one year, which is either equal (columns 1 and 2) or smaller than the medium length periods (columns 4 and 5). The general pattern in Table 5 also shows that the uncertainty surrounding the estimates decreases as the recall period or sample size increases. Thus, the table highlights that the trade-off between information and accuracy is present when the objective of the data is a less aggregated analysis (in contrast to the first part of the analysis). A short recall period for a salient event such as hospitalization implies a small number of actual events, which reduces the power of the test (i.e., a short recall period requires a larger *N*). On the other hand, a longer period results in a larger share of individuals misreporting the length of their hospital stay, and to some extent, a stronger association between reporting error and schooling.

One way to correct for recall error would be to use functions of the nature of the error to adjust reported data. For comparison with the reported and registered hospitalization, column 7 presents Erreygers' concentration index for reported hospitalization corrected for measurement errors.[14] As seen in the table, the correction is generally closer to registered hospitalization, suggesting a scope for using correction equations to correct for measurement error.

5. Discussion

The purpose of this study is to use a survey experiment to improve the understanding of the nature of recall error in self-reports of nights spent in hospital. A unique aspect of the experiment is that respondents were assigned to a wide range of recall periods ranging from one month to one year in the same health survey. This variation provides the opportunity to study how the length of the recall period affects both estimates of an overall measure of hospitalization and the association between hospitalization and a measure of socioeconomic status. Both of these represent applications in which health economists often rely on survey data.

[13] As $w = 30$ has few observed hospitalization events, we use $w = 365$ as our reference group.

[14] The correction is based on predictions from a regression of recall error on the same set of variables as in the previous analysis. The predictions are subtracted from the reported hospital nights. Presenting a similar analysis for the regressions coefficients is redundant as such a correction yields the same coefficients as in column 2.

Table 5
Summary statistics of education-related inequality.

	Regression coefficients			Erreygers' concentration index			
Recall period	(1) Reported	(2) Registered	(3) Registered (full sample)	(4) Reported	(5) Registered	(6) Registered (full sample)	(7) Reported (corrected)
$w = 30$	−0.0085	−0.0097	−0.0062*	−0.00180	−0.00168	−0.00122*	−0.00200
	(0.009)	(0.009)	(0.004)	(0.0016)	(0.0016)	(0.0007)	(0.0016)
$w = 91$	0.0013	0.0103	−0.0085	−0.00001	0.00052	−0.00055	0.0003
	(0.022)	(0.018)	(0.008)	(0.0013)	(0.0005)	(0.0005)	(0.0011)
$w = 183$	−0.0217	−0.0011	−0.0171*	−0.00060	0.00001	−0.00053*	0.00004
	(0.015)	(0.007)	(0.009)	(0.0005)	(0.0002)	(0.0003)	(0.0002)
$w = 365$	−0.0341	−0.0552**	−0.0308***	−0.00056*	−0.00083**	−0.00045**	−0.00088**
	(0.015)	(0.026)	(0.012)	(0.0003)	(0.0004)	(0.0002)	(0.0004)

Robust standard errors, columns 1–3, and bootstrapped standard errors, column 4–7, in parentheses.
Note: The table reports regression coefficients of hospital nights on education in columns 1–3 and the Erreygers' concentration index (cf. Erreygers, 2009) of hospital nights using years of schooling as the ranking variable in columns 4–7. Columns 1 and 4 use reported data during w; columns 2 and 5 use registered data during w for the group exposed to w; columns 3 and 6 use registered data during w for the full sample independent of recall period exposure. Column 7 presents results for reported hospitalization corrected for measurement errors. Corrections are based on a regression of error on schooling, male, non-Nordic, ln (income), age, and bad health.

* $p < 0.1$.
** $p < 0.05$.
*** $p < 0.01$.

The first part of our empirical analysis, which focuses on overall hospital use, has important implications for the optimal recall period for hospital-use questions in future surveys. Using a framework previously developed by Clarke et al. (2008) and assuming the purpose is to estimate average annual use, the results of our experiment show that bias and variance decrease over longer recall periods. Because of the tendency for some respondents to over-report previous hospital use in shorter recall periods, there is no trade-off between increasing the length of the recall period to include more information and getting a precise estimate. Hence under these circumstances our results indicate that a survey involving a one-year recall period is preferable.

Our experiment demonstrates that while the overall level of recall error increases with the length of the recall period, the composition in terms of under- versus over-reporting changes. In line with previous research, under-reporting is a relatively larger problem than over-reporting for longer recall periods (cf. Bhandari and Wagner, 2006). While forgetting to report a salient event that occurred recently may be unlikely, forward telescoping appears to be a problem for shorter periods. Another explanation of the pattern of error relates to anchoring; as individuals may relate to recurring events—e.g., birthdays, holidays, and other landmarks events that individuals may know by date (cf. Means et al., 1989)—a year may be a more natural unit to use as a reference point for the perception of time rather than a certain number of months. Therefore, although the total amount of recall error is larger for a one-year period, the errors are more equally distributed between over- and under-reporting.

A key insight from our results is that none of the recall periods is short enough to eliminate all bias. A common presumption of many survey designers appears to be that by shortening the recall period, one can remove error (cf. Das et al., 2012). Our experiment does not support this view, as even at one month there is significant over-reporting that appears to be due to forward telescoping. It is unclear why this behavior would be lessened if the recall period was shortened further, particularly if respondents re-interpret this question to report any recent hospitalization (i.e., with the last few months). Furthermore, short recall periods for infrequent events such as hospitalizations provide very little information and will be subject to large variations due to chance. Even for the relatively large sample of our experiment, with 1500–1900 respondents in each of our four recall periods, there is still a relatively small proportion of hospitalizations.

Together with the results of the second part of the empirical analysis, this lack of information highlights the potential trade-off between information and bias that must be considered when deciding upon the best length of recall period for analyzing the relationship between hospitalization and socioeconomic variables. Unlike a previous smaller study by Reijneveld and Stronks (2001) that found no association between measures of socioeconomic status and reporting error, we find relatively large coefficients for schooling and a pattern of a larger association between socioeconomic status and reporting error for the longer period. Our results are potentially troubling for researchers wanting to examine the relationship between hospitalizations and socioeconomic measures, and there is no obvious solution. Shortening the recall period may reduce the association between error and years of schooling, but will come at the high cost of much less information.

One avenue for future research is to conduct additional randomized experiments to further understand and find ways of reducing recall error. It will be important to try to understand the cause of telescoping, as there are two alternative explanations in the literature. In the model of Sudman and Bradburn (1973), [forward] telescoping is explained as time compression: individuals extend the effective recall period to include events outside the actual recall period, and this extended period increases with the length of the recall period. This effect is counteracted by decay in recall over time and respondents are, consequently, less prone to report events in the distant past. An alternative explanation is in the variance models suggested by Rubin and Baddeley (1989) and Huttenlocher et al. (1988). These models still assume that recall decay over time, but claims that over-reporting of the number of events during a recall period may be observed without any systematic error in dating events. Rather, the over-reporting is due to an increase of the variance in dating a recalled event as time passes.

We believe it would be possible to distinguish between these two explanations if either more information on the timing of events were reported or the same individuals were asked to report usage in multiple recall periods (i.e., usage "within the last month" and "between two and three months ago"). However, if respondents questioned about the more recent period were aware of their opportunity to answer a question about the more distant recall period and disclose a more distant but salient event, this might influence their response. Indeed such an approach has been suggested as a way to reduce reporting error (c.f. Sudman et al., 1984).[15]

It is worth highlighting some possible limitations of this study, such as the potential implications of the measurements errors in the registry data. To the extent the registry data truly is the gold standard depends on (a) the level of private (unregistered) health care consumption and (b) whether registered nights of hospital stay correspond to actual nights spent at the hospital. However, we do not believe that these are major issues. For the first problem (a), there are only a limited number of private clinics in the county council and we do not observe more errors in the geographical areas where these are located. The second problem (b) is related to the economic incentives for the hospital (although it is publically run) or the nurse (e.g., lower administrative burden) to not formally discharge a patient on nightly permission who actually spends the night at home. Since patients probably spend the first night at the hospital, the binary analysis presented in the supplementary online appendix may be less exposed to such bias. As the results are in line with the continuous case, we claim that this issue does not affect our general conclusions.

Although the length of the recall period for each respondent was not randomly decided, but instead assigned by quarter of birth, we claim that the results may be interpreted as valid as those in a randomized trial. Had the effect the quarter of birth may have on health and cognitive abilities been substantial,[16] we would have expected this effect to cause a similar pattern when the four groups were exposed to the same recall period of three months. However, a placebo analysis estimating Eq. (11) using the alternative question in which all respondents were also asked to state whether they had been admitted to the hospital during the last three months shows the opposite (see results in the supplementary online appendix), and thus supports the internal validity of the findings. Instead, the main limitation of the study is, as for a randomized trial, external validity. Is it possible to generalize these results to other populations?

Our sample consists only of individuals who chose to answer the questionnaire online. If we believe that individuals answering a survey online have on average higher cognitive abilities than individuals choosing to fill out a paper form, then the amount of recall error in our sample should be smaller than that in the general population. If this is the case, the association between recall error and the individual characteristics in our sample is probably an underestimation of the association in the total population. We also note that the optimal length of the recall period probably depends on the type of event, and thus we cannot directly extrapolate the results to other types of health care consumption without considering the saliency and frequency of the care we have in mind.

6. Summary and conclusion

In this article, we have used experimental data to study how the length of the recall period affects recall error. The twofold purpose was (a) to examine the optimal length of the recall period for an aggregated measure and (b) to examine whether the association between individual characteristics and recall error increases with the length of the recall period. Although the overall level of recall error increases with the length of the period, our study indicates that using a recall period of one year is preferable to scaling up a recall period of one, three, or six months to a target period of one year. In our analysis of how the length of the recall period affects the association between recall error and individual characteristics (using years of schooling as an example), we show that the association may increase with the length of the recall period. Consequently, the results of the two parts have conflicting implications for survey design concluding that the appropriate length of the recall period depends on the intended objectives of the survey data. For an aggregate measure of hospitalization, a longer recall period is preferable. However, if the objective of the survey is to study the relation between hospitalization and individual characteristics (e.g., for inequality indices or regression analysis), the researcher needs to seriously consider the trade-off between the lower bias of a shorter recall period and the larger amount of information available from a longer period.

Acknowledgments

We thank Kjeld Møller Pedersen, Mirja Elisabeth Kløjgaard, Jens Dietrichson, Anton Nilsson, Joanna Thorpe and two anonymous referees for comments on previous versions of this article. We are also grateful to seminar participants at the 2011 NHESG in Odense, 2012 HESG in Oxford and Lund University. Financial support from the Swedish Council for Working Life and Social Research (FAS) (dnr 2012-0419) is gratefully acknowledged. The Health Economics Program (HEP) at Lund University also receives core funding from FAS (dnr. 2006-1660), the Government Grant for Clinical Research ("ALF"), and Region Skåne (Gerdtham).

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jhealeco.2014.01.007.

References

Arnold, B.F., Galiani, S., Ram, P.K., Hubbard, A.E., Briceno, B., Gertler, P.J., Colford, J.M., 2013. Optimal recall period for caregiver-reported illness in risk factor and intervention studies: a multicountry study. American Journal of Epidemiology 177, 361–370.

Angrist, J.D., Krueger, A.B., 1991. Does compulsory school attendance affect schooling and earnings? The Quarterly Journal of Economics 106, 979–1014.

Benítez-Silva, H., Buchinsky, M., Man Chan, H., Cheidvasser, S., Rust, J., 2004. How large is the bias in self-reported disability? Journal of Applied Econometrics 19, 649–670.

Bhandari, A., Wagner, T., 2006. Self-reported utilization of health care services: improving measurement and accuracy. Medical Care Research and Review 63, 217–235.

Black, S.E., Devereux, P.J., Salvanes, K.G., 2011. Too young to leave the nest? The effects of school starting age. Review of Economics and Statistics 93, 455–467.

Bound, J., Brown, C., Mathiowetz, N., 2001. Measurement error in survey data. In: Leamer, E., Heckman, J. (Eds.), Handbook of Econometrics, vol. 5, pp. 3705–3843 (Chapter 59).

Clarke, P.M., Fiebig, D.G., Gerdtham, U.-G., 2008. Optimal recall length in survey design. Journal of Health Economics 27, 1275–1284.

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, 76–88.

Deaton, A., 1997. The analysis of Household Surveys: A Microeconometric Approach to Development Policy. Johns Hopkins University Press, Baltimore, MD.

Erreygers, G., 2009. Correcting the concentration index. Journal of Health Economics 28, 504–515.

Gaskell, G.D., Wright, D.B., O'Muircheartaigh, C.A., 2000. Telescoping of landmark events: implications for survey research. The Public Opinion Quarterly 64, 77–89.

Heijink, R., Xe, 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 Organization, Geneva, pp. 2011.

[15] There are other potential ways to reduce recall error, such as reminding respondents not to include prior events outside the recall period, or encouraging the respondent to think of significant events occurring at or near the recall boundary (such as a birthday).

[16] The accumulation of human capital such as cognitive and non-cognitive skills may be affected by quarter of birth through both the absolute age of school-start and the individual's relative age within the class (e.g. Angrist and Krueger, 1991; Black et al., 2011).

Huttenlocher, J., Hedges, L., Prohaska, V., 1988. Hierarchical organization in ordered domains: estimating the dates of events. Psychological Review 95, 471–484.

Means, B., Nigam, A., Zarrow, M., Loftus, E.F., Donaldson, M., 1989. Autobiographical memory for health-related events. National Center for Health Statistics. Vital and Health Statistics 6 (2).

National Board of Health and Welfare, 2009. Kvalitet och innehåll i patientregistret. Utskrivningar från slutenvården 1964–2007 och besök i specialiserad öppenvård (exklusive primärvårdsbesök) 1997–2007 (Quality and content in the patient register: discharges from inpatient care 1964–2007and visits to specialist outpatient care (excluding primary care visits) 1997–2007). http://www.socialstyrelsen.se/Lists/Artikelkatalog/Attachments/8306/2009-125-15_200912515_rev2.pdf

Reijneveld, S.A., Stronks, K., 2001. The validity of self-reported use of health care across socioeconomic strata: a comparison of survey and registration data. International Journal of Epidemiology 30, 1407–1414.

Ritter, P.L., Stewart, A.L., Kaymaz, H., Sobel, D.S., Block, D.A., Lorig, K.R., 2001. Self-reports of health care utilization compared to provider records. Journal of Clinical Epidemiology 54, 136–141.

Rosvall, M., Grahn, M., Modén, B., Merlo, J., 2009. Hälsoförhållanden i Skåne-Folkhälsoenkät Skåne 2008. Region Skåne http://www.skane.se/upload/Webbplatser/UMAS/VERKSAMHETER%20UMAS/Socialmedicin/09-dokument/fh-08-rapport-v3.pdf

Rubin, D., Baddeley, A., 1989. Telescoping is not time compression: a model. Memory & Cognition 17, 653–661.

Stull, D.E., Leidy, N.K., Parasuraman, B., Chassany, O., 2009. Optimal recall periods for patient-reported outcomes: challenges and potential solutions. Current Medical Research and Opinion 25, 929–942.

Sudman, S., Bradburn, N., 1973. Effects of time and memory factors on response in surveys. Journal of the American Statistical Association 68, 805–815.

Sudman, S., Bradburn, N., Schwarz, N., 1996. Thinking About Answers: The Application of Cognitive Processes to Survey Methodology. Jossey-Bass, San Francisco.

Sudman, S., Finn, A., Lannom, L., 1984. The use of bounded recall procedures in single interviews. Public Opinion Quarterly 48, 520.

Tourangeau, R., 1984. Cognitive sciences survey methods. In: Jabine, T.B., Straf, M.L., Tanur, J.M., Tourangeau, R. (Eds.), Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines. National Academy Press, Washington, DC.

Wooldridge, J.M., 2010. Econometric Analysis of Cross Section and Panel Data. MIT Press, Cambridge, MA.

Wolinsky, F.D., Miller, T.R., An, H., Geweke, J.F., Wallace, R.B., Wright, K.B., Chrischilles, E.A., et al., 2007. Hospital episodes and physician visits: the concordance between self-reports and medical claims. Medical Care 45, 300–307.