ECONSTOR
Make Your Publications Visible.

A Service of
ZBW
Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Kjellsson, Gustav; Clarke, Philip; Gerdtham, Ulf-G.

Working Paper

Forgetting to Remember or Remembering to Forget - A Study of the Recall Period Length in Health Care Survey Questions

Working Paper, No. 2013:1

Provided in Cooperation with:
Department of Economics, School of Economics and Management, Lund University

Suggested Citation: Kjellsson, Gustav; Clarke, Philip; Gerdtham, Ulf-G. (2013) : Forgetting to Remember or Remembering to Forget - A Study of the Recall Period Length in Health Care Survey Questions, Working Paper, No. 2013:1, Lund University, School of Economics and Management, Department of Economics, Lund

This Version is available at:
https://hdl.handle.net/10419/260059

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

Working Paper 2013:1

Department of Economics
School of Economics and Management

Forgetting to Remember or Remembering to Forget - A Study of the Recall Period Length in Health Care Survey Questions

Gustav Kjellsson
Philip Clarke
Ulf-G Gerdtham

January 2013

LUND UNIVERSITY

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*

[*]*Corresponding Author: Contact Info: email: gustav.kjellsson@nek.lu.se, tel: +46 46 2227911,*

Abstract: Self-reported data on utilization of health care is a key input into a range of studies. However, the length of the recall period in self-reported health care questions varies between surveys and this variation may affect the results of the studies. While longer recall periods include more information, shorter recall periods generally imply smaller bias. This article examines the role of the recall period length for the quality of self-reported data by comparing registered hospitalization with self-reported hospitalizations of respondents that are exposed to a varying recall period length of one, three, six, or twelve month. Our findings have conflicting implications for survey design as the preferred length of recall period depends on the objective of analysis. If the objective is an aggregated measure of hospitalization, longer recall periods are preferred whereas shorter recall periods may be considered for a more micro-oriented level analysis since the association between individual characteristics (e.g. education) and recall error increases with the length of the recall period.

Keywords: Survey Methods; Health survey; Hospitalization; Recall error; recall periods

JEL: C42, C83; I10

Acknowledgments: We thank Kjeld Møller Pedersen, Mirja Elisabeth Kløjgaard, Jens Dietrichson, Anton Nilsson and Joanna Thorpe for comments on previous versions of this paper. 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).

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 period range from two weeks to 14 months with a significant proportion of surveys using either one or 12 months (Heijink, 2011). While information on hospitalization tends to be collected over longer recall periods than physician visits, there is a surprising degree of variation between surveys. For example in the case of hospitalizations, 36% of the surveys reviewed use a one month recall period, while 46% use one year.

It has been well established that self-reported behaviors such as health care use is 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 lesser the accuracy of the reported estimates (Stull et. al. 2009 ; Bhandari and Wagner, 2006). However, even though the likelihood of recall error increases for a longer recall period, so does the information provided and so there is a potential trade-off between recall error and information. This implicit requirement for this trade-off when designing health surveys may explain the high degree of variation in recall period that are used for the same type of health care indicated above.

The appropriate length of recall period also depends on the type of health care consumption and intended use of the information. First, events that are more salient call for a longer 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 GP-visit. 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 utilization for an infrequent and unpredictable event will probably yield estimates that are at best weakly related to the actual utilization (e.g., Deaton, 1997). Third, as individual characteristics, such as cognitive ability or socioeconomic variables, potentially also 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 utilization 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). Although a longer recall period in self-reported health care probably implies a larger amount of recall error, few studies examine how the length of the period affects the association between recall error and socioeconomic variables.

The primary aim of this article is to use a large survey experiment to examine the role of the recall period length for recall error of hospitalization. By comparing self-reported data from a public health survey with registered data (treating the latter as the gold standard), the paper explores the nature of recall error. It then examines the implications for two aspects of survey design. First, it extends the framework suggested by Clarke et al. (2008) in order to determine an optimal length for a recall period for an aggregated measure of hospitalization e.g. estimating the mean number of nights. The second part relates to how individual characteristics affects the quality of self-reported data and this examines the degree of association between years of schooling (which is a proxy for cognitive ability) and recall errors over different recall periods. We know of no published comparable experiment to quantify recall error of a type of health care use. All previous validation studies generally use small samples and have no variation in the length of the recall period (for an overview see Bandhari and Wagner, 2006). Therefore, this article contributes to the literature by exploiting variation in the length of the recall period for a large sample.

2. The nature of recall error

While numerous studies have compared reported and actual usage for a range of health care almost all previous studies have examined only one period over which the respondent is asked to recall their prior use. Comparisons of the degree of reporting error of studies with varying

recall lengths are confounded by other potential differences between 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. In this study we employ an experiment to control for these confounders by allocated respondents to different versions of the same question that differ only in the time period over they are being asked to recall past use. Below we describe such an experiment and the insights it provides on the nature of recall error for hospital use.

2.1 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 to answer the question "How many nights were you hospitalized during the last year/X months", where the length of the recall period was decided by the quarter of birth. For respondents born between January and March (Group 1), between April and June (Group 2), between July to September (Group 3), and between October to December (Group 4), the length of the recall period is 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 corresponds to the registered event. In addition to this question, respondents were also asked to state whether they have been admitted to the hospital during the last three month (admission).

The population in the public health survey (Folkhälsoenkät Skåne 2008) consists of all individuals between the age of 18-80 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 (Rosvall et al., 2009). This study is based on a subset of 7500 respondents that answered the questionnaire on the web as the exact date the survey completion was known.[1] The survey data, which also includes information on self-assessed health, living conditions, and background information such as age and country of birth, is linked to registry data on income, education and hospitalization. The link to registry data allows us to compare the self-reported measure of hospitalization with the registered number of nights spend at a hospital. The National Board

[1] For remainder of the sample only the dates of receipt of the survey form when it was mailed back was known and so the exact recall period could not be obtained.

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 includes hospitalizations at public hospitals within Region Skåne as well as in other county councils, but it does not include nights spend at private hospitals. As the registry data does not include private care, we may overestimate the number of individuals that falsely report 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 as there is only a very limited amount of private clinics for in-patient care within Region Skåne.

Out of the 7500 observations, 365 have missing values on either reported or registered hospitalization and an additional 136 have missing values on either years of schooling or income. Thus, the analysis uses the remaining 6999 observations.

2.2 Results of the experiment

In line with the previous literature, the level of agreement between that self-reported and registered data decreases with the length of the recall period (compare Bhandari and Wagner, 2006). The percentage of correctly self-reported hospitalization is 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 (illustrated in Figure 1) when we divide the error into positive (i.e. over reporting due to commission and forward telescoping) and negative components (i.e. under reporting due to omission and backward telescoping). Importantly, short recall periods such as one month are not free from error as there is a large positive error on average (0.43). By contrast, there are very small negative errors for the short recall periods that rise dramatically for the longest period of one year. So while the proportion of the sample making errors is continuously rising, the degree to which this leads to bias in the reporting of the mean is relatively constant over the year (e.g. in the longest period of one year, the overall bias is -0.045 and imparts only a slightly bias in absolute terms to the bias for one month i.e. 0.041).

Some further insights into the nature of recall error can be gained by plotting the difference between actual and reported nights for each individual with some evidence of hospitalization either during the recall period, or up to a year prior to the recall period. Figure 2 illustrates these plots for the four recall periods. The y-axis illustrates the time in days from the recall boundary (i.e. days before this boundary fall within the recall window which are denoted as

negative values in Figure 2 and days prior to the recall period which are positive values). So for example, if we consider the 30 days recall window, 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 window.

To examine the nature of the error relative to actual hospitalizations, Figure 2 plots the degree of error in the reported number of nights (where positive values indicate over reporting) against the day of hospitalization closest to the recall window. This is defined by the day of admission for spells ending after the recall boundary and discharge for spells starting before the recall boundary. The graph illustrates that respondents over reporting nights often have hospitalization in the period of up to 120 days outside the recall window, which is strongly suggestive of forward telescoping. The long recall periods of 183 and especially 365 days also show an increasing propensity to under-report events (i.e. negative values) when there is evidence of hospitalizations within the recall window.

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 exhibits that the respondents in group 4 (w=365) appear to be slightly different in terms education and 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 if they differ by their error proneness. As everyone is exposed to the same recall period, we can examine if there are any systematic differences between the groups in terms of reporting incorrect answers (see Table 1b, column 7). Even though there are statistically significant differences in some observable aspects, the groups do not seem to be more or less error prone. F-test of equal means cannot reject the null of the four groups being equal in terms of being either a false negative or a false positive (see Table 1b, column 7).

Next, we explore the implications of the pattern of recall error revealed by the experiment for the choice of recall period in two situations: for estimating overall summary measure of hospital use; 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. As the data in their analysis only includes one single recall period, their analysis is based on extrapolation and a set of restrictive assumptions. By contrast, the respondents in our data have been exposed to recall periods with varying length and, thus, we can perform an analysis with less 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 denote this recall period 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 of using X_i^w as an estimate of Y_i^w if individuals had perfect recall during w.[2] It would probably also be possible to choose a recall period short enough to eliminate any recall error. 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 is discussed in section 3, our data observe w= {30, 91,183,365}.

To evaluate the appropriate length of the recall period, Benitez-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 utilization of 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 of individuals' hospitalization increases as the length of the recall period increase, 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

[2] 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 that provoke strategic behavior of the respondents

choose w to minimize root mean square error (RMSE). We exploit the varying recall periods comparing the four recall periods in regard of bias, variance, and RMSE. Further on, we first present the framework in Clarke et al. (2008) and then also considers the different nature of recall errors within their framework.

Formally, Clarke et al. (2008) let

$$X_i^w = Y_i^w + v_i^w \qquad \textit{Eq(1)}$$

where v_i^w represents the measurement error. Our intended objective is to obtain a measure of mean utilization in the target period $E\left(Y_i^S\right)$. 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:

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

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:

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

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

$$\text{Bias}(\bar{X}_w^S) = E(\bar{X}_w^S) - E\left(Y_i^S\right) \qquad \textit{Eq(5)}$$

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

$$\text{RMSE}(\bar{X}_w^S) = \text{sqrt}(\text{Bias}(\bar{X}_w^S))^2 + Var(\bar{X}_w^S)) \qquad \textit{Eq(6)}$$

To get an estimate of the bias, we need an empirical definition of $E\left(Y_i^S\right)$. An obvious candidate is to use 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 $E\left(Y_i^S\right)$. However, as a robustness test, and to assure that seasonality do not introduce further bias, we also calculate the bias as:

$$\text{bias}(\bar{X}_w^S) = E(\bar{X}_w^S) - \left(\frac{S}{w}\right) E(Y_i^w) \qquad \textit{Eq(7)}$$

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 the appendix.

To be able 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 v_i^w is independent of Y_i^w, but allows

$$E(v_i^w) = h(w^*) = h(w)\left(\frac{w}{S}\right)\mu \qquad Eq(8)$$

and

$$Var(v_i^w) = g(w^*) = g(w)\left(\frac{w}{S}\right)\sigma^2 \qquad Eq(9)$$

where $\mu = E(Y_i^S)$ and $\sigma^2 = Var(Y_i^S)$. Thus, $Eq(8)$ and $Eq(9)$ 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 exist 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 causing the errors; the individuals proneness to over reporting (i.e. commission and forward-telescoping) and underreport (i.e. omission and backward telescoping), defined as $h^+(w)$ and $h^-(w)$, respectively. We therefore redefine $Eq(8)$ into

$$E(v_i^w) = h(w)\left(\frac{w}{S}\right)\mu = (h^+(w) - h^-(w))\left(\frac{w}{S}\right)\mu \qquad Eq(10)$$

Splitting the error in this way allows the optimal recall period to be determined for a wide variety of different error structures. Figure 3 shows four stylized graphs of positive and negative error structures.

[Figure 3 about here]

To simplify, we may assume that the error proneness either increases proportionally over the period or increases rapidly in the beginning and then is relatively constant over the remainder of the interval. Case a represents a symmetric error structure where both increases proportionally (with a slightly different speed), Case b represents a symmetric error structure where the error proneness increases rapidly for short recall periods but stays relatively constant over the rest of the interval. For both these cases, $h(w)$ is monotonically increasing, and shorter periods are preferred in terms of bias (although the propensity to over and under report tend to cancel out for all recall periods). By contrast, Case c and d represent asymmetric error structures, where 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 d in Figure 3 that the recall error process may not be monotonic increasing or decreasing functions of the period length.[3] This implies that the bias is not necessarily largest for the longest recall period. To see how the two processes affect the optimal recall period it is illuminating to consider that also g is affected by changes in $h^{+}(w)$ and $h^{-}(w)$ and express the MSE as

$$MSE(\bar{X}_w^S) = \left[1 + g\left(w, h^{+}(w), h^{-}(w)\right)\right]\left(\frac{w}{S}\right)\left(\frac{\sigma^2}{N}\right) + \mu^2\left(h^{+}(w) - h^{-}(w)\right)^2 \qquad Eq(11)$$

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 MSE, through $g(w)$ (i.e. $\frac{\partial g(w)}{\partial h^{+}(w)} > 0$ or $\frac{\partial g\left(w,h^{+}(w),h^{-}(w)\right)}{\partial h^{-}(w)} < 0$, while $\frac{\partial g\left(w,h^{+}(w),h^{-}(w)\right)}{\partial h(w)}$ may equal zero).[4]

When these two processes are equal, i.e. $h(w) = 0$, there will be no recall bias, but non-zero values of $h^{+}(w)$ and $h^{-}(w)$ still implies that recall errors exists and will increase the dispersion. However, the determinant of bias is the relative magnitude of $h^{+}(w)$ and $h^{-}(w)$ which stresses the importance of not using a recall period where either of these two processes strongly dominates the other (compare case c and d).

[3] Note that Clarke et al. (2008) do consider this division of errors in the bivariate case.
[4] The exact relationship between $g(w)$ and the two functions, $h^{+}(w)$ and $h^{-}(w)$, depends on the actual distributions.

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 Figure 4-5 and Table 2a-b is clear.

[Figure 4-5 about here]

[Table 2 about here]

Table 2 shows that while the variance of X_i^w generally increases for longer *w*,[5] $Var(\bar{X}_w^S)$ decreases. The decrease in variance is anticipated since the information of individuals' hospitalization increases as the length of the recall period increase. 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).[6] Thus, as the graph in Figure 4 illustrates, when comparing the four recall periods in regard of bias, variance, and RMSE our experiments 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 appendix.[7]

Recall that Figure 2 shows that although the amount of total errors increases with the length of the recall period (so does the proportion of false positive and false negative), this is not the case for the number of positive errors (or proportion of false positives). Thus, the decrease in bias appears to be driven by the increasing number of under-reporters (negative errors) as the length of the recall period increases. In terms of the optimal recall period framework our observations of $h(w)$, plotted as blue squares in Figure 5, are decreasing in the interval from

[5] The exception is the decrease between *w*=181 and *w*=365 in the continuous case in Table 3a

[6] The two anomalies in the pattern (i.e. the increase in bias and Var(X_i^w) between *w*=91 and *w*=183) disappears if we exclude the two respondents that have more than 100 registered hospital nights within a period of *w*+365 days.

[7] Separate analysis of the proportions of false negative and the proportion of false positive yields the same conclusion. Furthermore, calculating bias, variance, and RMSE of the alternative question as a placebo-analysis also confirm 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 hospitalization > 50 or >30 during the last year. Thus, the results are not driven by a few extreme observations. These results are available upon request.

w=30 to S. The decreasing pattern is caused by the asymmetry of the two recall processes (also plotted in Figure 5); $h^{-}(w)$ is monotonically increasing, while $h^{+}(w)$ (i.e. the process of commission or telescoping) appears to be high already for a short period and decreases or is at least relatively constant over the interval. That is, $h^{+}(w)$ strongly dominates $h^{-}(w)$ for the shorter periods. Thus, even though the amount of recall errors increases over the interval, over- and underreporting becomes more equally distributed. That respondents tend to overstate their hospitalization for shorter recall periods and underreport for longer is also in line with previous research (for an overview see Bhandari and Wagner, 2006).

The graphs in Figure 2 previously presented in Section 2 further support this interpretation as telescoping appears to generate large positive errors for 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, but may very well include an event that occurred before the start of the recall period (e.g. telescoping) (cf. Sudman and Bradburn, 1973).

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 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 purpose therefore examines if the association between years of schooling and recall errors differs with the length of the recall period. The motivation of using years of schooling as an example of socioeconomic variable or individual characteristics relates to the possible link between the 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 the 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 all are related to the cognitive ability of the respondents (Bound et al 2001; Tourangeau 1984; Sudman, Bradburn and Schwarz, 1996; Bhandari and Wagner, 2006). As we cannot directly observe cognitive ability, we may consider years of schooling as a proxy for cognitive ability.

Previous findings for the association between socioeconomic variables and recall error are mixed. For example, Das et al. (2012) recently showed 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 (and even changing the sign of the coefficient). Others (e.g. Wolinsky et al., 2007 and Ritter et al., 2001) find no socioeconomic differences (although, Wolinsky et al., 2007 show that the health of the respondents matters). However, neither of the studies is able to further examine how the length of the period affects the bias. By contrast, we are able to exploit the variation in length of the recall window provided by our experiment.

4.2. Framework & Method

Health survey data is often collected for other purposes 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. As an illustrative example, Table 3 presents the coefficients from univariate regressions of hospitalization and years of schooling. Column (1) presents results from the reported hospitalization for each period, Column (2) the registered hospitalization for the corresponding sample and Column (3) the registered hospitalization for the full sample (6999 observation).

[*Table 3*]

Although, Column (1) and (2) show that there are some differences between the coefficients, the general pattern in Table 3 show that the uncertainty surrounding the estimates decreases as recall period or sample size increases. Thus, the table highlights the trade-off between information and accuracy. 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 we know from the previous section that a longer period results in a larger share of individuals misreporting the length their hospital stay.

What also needs to be considered is the relationship between reporting error and individual characteristics. As Wooldridge (2010) notes if recall errors are systematically associated with any observed or unobserved characteristics of the respondents, the coefficients in regression analysis may be seriously biased and may alter researcher's conclusions. While we cannot test

for an association with unobserved characteristics, we may test if the recall error is systematically associated with observed characteristics such as education and income.

To study the association between recall error and socioeconomic variables we use regression analysis for two sets of outcome variables (denoted as $error_i$); first, the number of negative errors (neg_error), positive errors (pos_error), and absolute errors (abs_error); second, the probability of being a false positive (X>0; Y=0, falsepos), a false negative (X=0; Y>0, falseneg), and a binary indicator of being either of the two (binary) (a full set of definitions is given in Table 4).

[Table 4 about here]

We examine how the length of the 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[8] with heteroskedasticity-consistent standard errors for all $error_i$

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

where $\boldsymbol{Z_i}$ is 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 (also see Table 4). 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 the findings for the association between recall error and socioeconomic variables.

4.3. Results

Table 5a and 5b initially presents the results without any interaction, that is assuming $\boldsymbol{\delta} = 0$ in $Eq(11)$, before Table 6a and 6b presents results with interactions between years of schooling and the recall period dummies (i.e. allowing $\boldsymbol{\delta} \neq 0$). The overall pattern[9] indicates a significant association between years of schooling and recall error, when controlling for

[8] On average a Linear Probability Model (LPM) and binary choice models as a Logit or a Probit provide the same results (cf. Wooldridge, 2010 p563).

[9] The exception is falsepos and pos_error which are not statistically significant, but we note the negative sign indicating that the propensity to make these errors decreases by years of education.

demographics, income, and the number of registered nights at the hospital. Note that as the mean 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 decreases the number absolute errors as well as the probability of misreport the binary indicator of hospitalization (i.e. being either a false negative or a false positive). The decrease is driven by the significantly decreased number of negative errors and decreased probability of being a false negative, respectively. For the number of positive errors and the probability of being a false positive, the coefficients are negative but insignificant. However, these are also the dependent variables with the smallest variation.

[Table 5-6 about here]

The results for the control variables are mixed, but some clear tendencies emerge. In line with our beliefs, the health dummy is significantly associated with all indicators except falseneg and neg_error; that is, individuals in bad health do not under-report (given a certain level of hospitalization), but the degree of over-reporting contributes to the increased misreporting (absolute errors and the binary indicator). In contrast to health (and education), 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 falseneg and the binary indicator, where individuals age 30-60 (or 45-75 for falseneg) are significantly less error prone than the reference group (age<30).

When estimating *Eq*(11) allowing for $\boldsymbol{\delta} \neq 0$, the results indicate that the association between schooling and recall is affected by the length of *w*. Table 6a and 6b present the results for the continuous and the binary dependent variables, respectively, with and without the number of registered *hospital nights* among the controls (even columns presents results of the model being estimated conditional on the number of registered *hospital nights*). 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 error proneness increases with the number of registered hospital nights.

The schooling variable follows the same pattern as in Table 5a-b. 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, i.e. *w*=365.[10] 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 exactly observe such a pattern for the binary indicator of being either a false negative or a false positive (Table 6a Column 1-2), the major differences in the association are in general between on the one hand the reference group (w=365) and on the other hand respondents that are exposed to a shorter period (w<365). Without conditioning on *hospital nights*, the interactions are jointly significant in all models (for falsepos and abs_error only at a 10% level) except for the number of positive errors (F-statistics are presented in Table 6a and 6b). The overall pattern remains when conditioning on the number of hospital nights. 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. binary and abs_errors) are driven by differences in under-reporting (falseneg and neg_err). The coefficients may throughout the models seem small in magnitude and are statistical 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 mean of the dependent variables are low), we cannot expect to measure the differences with strong precision.

4.4. Placebo analysis using a common three month recall question

As the descriptive statistics in Section 2.2 indicate some differences in observable characteristics between the groups, there is a risk that group composition rather than the length of the recall period may drive the observed differences in the association between years of schooling and recall error. We are able to examine this issue as all respondents where also asked to state whether they have been admitted to the hospital during the last three month.

If variations in composition of groups impact on analysis we would expect similar differences between the four groups to occur also when respondents answer this three month recall period. However, when estimating *Eq*(11) using the binary indicators of recall error from the alternative recall question as dependent variables no such patterns emerge (See Table 7). The F-statistics testing the joint significance of the interactions presented in in the bottom of

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

Table 7 show that there are no significant differences in the degree of reporting error between groups.

[Table 7 about here]

5. Discussion

The first part of our empirical analysis shows that bias and variance decrease for longer recall periods. Thus, the potential trade-off between information and recall error that the RMSE was intended to capture does not appear in the data and if the intended objective of the survey is an annual summary measure of hospitalization using a one-year recall period is preferable. As the overall level of recall error increases with the length of the period, the composition in terms of under- and over reporting changes. In line with previous research, under-reporting is relatively to over reporting a larger problem for longer periods (cf. Bandhari and Wagner, 2006). While forgetting to report a salient event that recently occurred may be unlikely, telescoping appears to be a problem for shorter periods. Another explanation of the pattern of error relates to anchoring; as individuals may relate to reoccurring events – e.g. birthdays, holidays, and other landmarks events of which individuals may know the date of (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, the errors are more equally distributed between under- and over-reporting.

A key finding of our experiment 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 shorten the recall window one can remove error (cf. Das et al. 2012). Our experiment does not support this view as at one month there is significant over-reporting which 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 the respondents are re-interpreting this question to report any recent hospitalization (i.e. with the last few months). Furthermore, short recall windows for infrequent event such as hospitalization provide vary little information and will be subject to large variations due to chance.

Although we have a relatively large experiment with 1500-1900 respondents in each our four recall periods, we still observe on relatively small proportion of hospitalizations. Unlike a previous smaller study by Reijneveld and Stronks (2001) that did not find association between measures of socio-economic status and reporting error, we find relatively large coefficients for schooling and a pattern of a larger association for the longer period. Our results are

potentially troubling for researchers wanting to examine the relationship between hospitalizations and such measures and there is no obvious solution. Shortening the recall window may reduce the association between error and the degree of education, but comes at the high cost of much less information. It also poses issues for the interpretation of analyses of data sets which have samples that are of a similar size to groups in our survey such as country level results involving the European Community Household Panel (ECHP,or EU-SILC) and Survey of Health, Aging and Retirement in Europe (SHARE) that are commonly undertaken by health economists (e.g. van Doorslaer et al., 2006; Allin et al., 2009).

One avenue for future research is to conduct randomized experiments of alternative wording of health care use questions in an attempt to reduce recall error. For example, reminding respondents that they should not include prior events outside the recall window could potentially reduce the effect of telescoping. Alternatively getting the respondent to think of significant events occurring at or near the recall boundary (such as birthday) may also assist them in remembering the timing of health care use in relation to the survey recall period. Given the worldwide reliance on surveys to collect information on health care use it is surprising that almost no studies of this kind have been conducted.

It is worth highlighting some possible limitations of the study such as the potential implications of the measurements errors in registry data. To what extent the registry data actually is the gold standard depends on (a) the level of private consumption and (b) if registered nights of hospital actually corresponds to spending the night 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 increased number of 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 discharge a patient that is on nightly permission and spends the night at home. Since patients probably spend the first night at the hospital, the binary analysis that is presented in 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 the general conclusions.

The length of the recall period the respondent is exposed to is not randomly decided but instead due to the quarter of birth. Nevertheless, we claim that it is possible to interpret the result as in the context of a randomized trial. Had the effect the quarter of birth may have on

health and cognitive abilities been substantial,[11] we would expected this effect to cause a similar pattern when the four groups are exposed to the same recall period of three months. However, our placebo analysis using the alternative question of a common three months recall period shows the opposite 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 the result to other populations? An important aspect of our experiment is that it uses a sample only consisting 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 have, then the amount of recall error in our sample should be smaller than for the full 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 our results show that less healthy individuals are more prone to misreport (in line with the results in Das et al. 2012). Thus, for example, if the objective of the study is to obtain an estimate of the number of hospital nights for chronically-ill individuals one would have to consider that hospitalization may be perceived as a less salient event for this population. Thus, because the optimal length of the recall period depends on the type of event, 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.

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 if the association between individual characteristics and the recall error increases with the length of the period. Although the overall level of recall error increases with the length of the period, our study indicates that a recall period of one year is preferable to scaling up a recall period of one, three, or six month 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, we exploit the variation provided by our experiment. Using years of schooling as an example (and as a proxy for cognitive ability) we show that the association may increase with the length of the period.

[11] 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 individuals relative age within the class (e.g. Angrist and Krüger, 1992; Black, Devereux, and Salvanes, 2011)

Consequently, the results of the two parts have conflicting implications for survey design concluding that the appropriate length of the 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 from a longer period.

References

Allin S, Masseria C, Mossialos E. Measuring Socioeconomic Differences in Use of Health Care Services by Wealth Versus by Income. American Journal of Public Health 2009; 99; 1849–1855.

Angrist JD, Krueger AB. Does Compulsory School Attendance Affect Schooling and Earnings? The Quarterly Journal of Economics 1991; 106; 979–1014.

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

Bhandari A. Self-Reported Utilization of Health Care Services: Improving Measurement and Accuracy. Medical Care Research and Review 2006; 63; 217–235.

Black SE, Devereux PJ, Salvanes KG. Too Young to Leave the Nest? The Effects of School Starting Age. Review of Economics and Statistics 2011; 93; 455–467.

Bound J, Brown C, Mathiowetz N. Chapter 59 Measurement error in survey data. Ed. Edward Leamer and James Heckman. Handbook of Econometrics 2001; 5; 3705–3843.

Clarke PM, Fiebig DG, Gerdtham U-G. Optimal recall length in survey design. Journal of Health Economics 2008; 27; 1275–1284.

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

Deaton A. The analysis of household surveys : a microeconometric approach to development policy. Published for the World Bank by Johns Hopkins University Press: Baltimore, MD; 1997.

Gaskell GD, Wright DB, O'Muircheartaigh CA. Telescoping of Landmark Events: Implications for Survey Research. The Public Opinion Quarterly 2000; 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; 2011.

Means B, Nigam A, Zarrow M, Loftus EF, Donaldson M. Autobiographical Memory for Health-Related Events. National Center for Health Statistics. Vital Health Stat 1989; 6.

The National Board of Health and Welfare. Kvalitet och innehåll i patientregistret. Utskrivningar från slutenvården 1964–2007 och besök i specialiserad öppenvård (exklusive primärvårdsbesök) 1997–2007. 2009. http://www.socialstyrelsen.se/Lists/Artikelkatalog/Attachments/8306/2009-125-15_200912515_rev2.pdf.

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

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

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

van Doorslaer E, Masseria C, Koolman X. Inequalities in access to medical care by income in developed countries. Canadian Medical Association Journal 2006; 174; 177–183.

Sudman S, Bradburn N, Schwarz N. Thinking about answers : the application of cognitive processes to survey methodology. Jossey-Bass (Pod): [S.l.]; 2010.

Sudman S, Bradburn N. Effects of Time and Memory Factors on Response in Surveys. Journal of the American Statistical Association 1973; 68; 805–815.Stull et. al. 2009

Tourangeau R. Cognitive sciences and survey methods. In: Jabine TB, Straf ML, Tanur JM, Tourangeau R (Ed). Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines. National Academy Press: Washington DC.

Wooldridge JM. Econometric analysis of cross section and panel data. MIT Press: Cambridge, Mass.; 2010.

Wolinsky FD, Miller TR, An H, Geweke JF, Wallace RB, Wright KB, Chrischilles EA, et al. Hospital Episodes and Physician Visits: The Concordance Between Self-Reports and Medical Claims. Medical Care 2007; 45; 300–307.

Table 1a: Descriptive statistics

VARIABLES	#	(1) Binary	(2) abs_error	(3) pos_err	(4) neg_err	(5) Error range	(6) alt_binary
Total	6,999	0.024	0.119	0.074	-0.045	(-33) – 33	0.021
w=30	1,704	0.012	0.045	0.043	-0.002	(-3) – 12	0.021
w=91	1,722	0.009	0.055	0.045	-0.009	(-7) – 30	0.021
w=183	1,662	0.028	0.153	0.132	-0.021	(-4) – 30	0.017
w=365	1,911	0.043	0.213	0.078	-0.134	(-33) – 9	0.026
F-test		20.02	10.05	4.304	14.09		1.128
prob.		0.000	0.000	0.005	0.000		0.336

Table 1b: Descriptive statistics

VARIABLES	(1) Age	(2) Male	(3) Non-Nordic	(4) School	(5) Bad Health	(6) Income	(7) hnight
Total	43.89	0.520	0.0843	12.67	0.0461	263787	0.153
w=30	43.44	0.521	0.090	12.720	0.043	270041	0.029
w=91	42.92	0.532	0.079	12.744	0.042	264125	0.088
w=183	44.96	0.499	0.076	12.772	0.042	262072	0.147
w=365	44.25	0.525	0.092	12.476	0.057	259396	0.327
F-test	6.178	1.398	1.386	5.670	2.147	0.871	12.95
prob.	0.000	0.241	0.245	0.001	0.092	0.455	0.000

Table2a: Aggregated measure; Continuous case

	$E(X_i^W)$	$E(\bar{X}_w^S)$	$E(Y_i^W)$	$E(Y_i^S)$	$\left(\frac{S}{w}\right)E(Y_i^W)$	**S/w**	$E(X_i^W) - E(Y_i^W)$	**Var(X_i^W)**
w=30	0.069	0.843	0.029	0.329	3.998	12.167	0.040	0.592
w=91	0.124	0.498	0.088	0.411	1.649	4.011	0.036	1.979
w=183	0.258	0.514	0.147	0.413	0.824	1.995	0.111	3.717
w=365	0.271	0.271	0.327	0.327	0.327	1.000	-0.056	3.132

	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	Var($\bar{X}_w^S$)	Bias($\bar{X}_W^S$)	RMSE	Var($\bar{X}_w^S$)	Bias($\bar{X}_W^S$)
w=30	0.562	0.051	0.514	0.294	0.051	0.493
w=91	0.162	0.018	0.087	0.039	0.018	0.144
w=183	0.138	0.009	0.100	0.058	0.009	0.221
w=365	0.069	0.002	-0.056	0.005	0.002	-0.056

Note: Using paired bootstrap (9999 replications), the difference in RMSE between w=31 and the remaining three recall period is statistically significant from zero (10% level for w=91 & w=183, and 5% level for w=365).

Table 3: Regression Coefficients: Hospital nights on years of schooling

	(1)	(2)	(3)
	Reported	Registered	Registered Full sample
w=30	-0.008	-0.010	-0.006*
	(0.009)	(0.009)	(0.004)
w=91	0.001	0.010	-0.009
	(0.022)	(0.018)	(0.008)
w=183	-0.022	-0.001	-0.017*
	(0.015)	(0.007)	(0.009)
w=365	-0.034	-0.055**	-0.031***
	(0.021)	(0.026)	(0.012)

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 4: Definitions of variables

Variable	Definition
Absolute error	= $\lvert X_i^w - Y_i^w \rvert$
Neg_errror	= $X_i^w - Y_i^w$ if $X_i^w < Y_i^w$, = 0 otherwise
Pos_error	= $X_i^w - Y_i^w$ if $X_i^w > Y_i^w$, = 0 otherwise
Binary	= 1 if either $X_i^w = 0$ and $Y_i^w > 0$ or $X_i^w >$ and $Y_i^w = 0$, 0 otherwise
False negative	=1 if $X_i^w = 0$ and $Y_i^w > 0$, 0 otherwise
False positive	= 1 if $X_i^w > 0$ and $Y_i^w = 0$, 0 otherwise
Hnights	= registered hospitalization (number of nights) during Y_i^w
School	= years of school
BadHealth	= 1 if health is reported bad or very bad, (=0 if fair, good, or very good)
ln(income)	=Logarithm of total income – individual level in 2007
Non-Nordic	=1 if born outside of the Nordic countries
Sex	= 1 if male, 0 if female
Age18_30	= 1 if 17<age<30, 0 otherwise
Age31_45	= 1 if 30<age<45, 0 otherwise
Age45_60	= 1 if 45<age<60, 0 otherwise
Age60_	= 1 if 60<age<75 0 otherwise
Admission	= registered admission to hospital

Table 5a:Results without interaction terms (Binary indicators)

VARIABLES	(1) Binary	(2) Binary	(3) Falseneg	(4) Falseneg	(5) falsepos	(6) Falsepos
w=30	-0.030***	-0.027***	-0.023***	-0.020***	-0.007*	-0.008*
	(0.005)	(0.005)	(0.003)	(0.003)	(0.004)	(0.004)
w=91	-0.033***	-0.030***	-0.023***	-0.020***	-0.010**	-0.010***
	(0.005)	(0.005)	(0.003)	(0.003)	(0.004)	(0.004)
w=183	-0.014**	-0.012**	-0.015***	-0.013***	0.001	0.001
	(0.006)	(0.006)	(0.004)	(0.004)	(0.005)	(0.005)
School	-0.002**	-0.001*	-0.001***	-0.001***	-0.000	-0.000
	(0.001)	(0.001)	(0.000)	(0.000)	(0.001)	(0.001)
Male	-0.005	-0.005	-0.003	-0.003	-0.002	-0.002
	(0.004)	(0.004)	(0.002)	(0.002)	(0.003)	(0.003)
Non-Nordic	0.026***	0.027***	0.004	0.005	0.022***	0.022***
	(0.009)	(0.009)	(0.005)	(0.005)	(0.008)	(0.008)
ln(income)	0.000	0.000	0.001***	0.001***	-0.001	-0.001
	(0.001)	(0.001)	(0.000)	(0.000)	(0.001)	(0.001)
Age31_45	-0.009	-0.010*	-0.004	-0.005	-0.005	-0.005
	(0.006)	(0.006)	(0.004)	(0.004)	(0.004)	(0.004)
Age45-60	-0.014**	-0.013**	-0.010***	-0.009***	-0.004	-0.004
	(0.006)	(0.006)	(0.003)	(0.003)	(0.005)	(0.005)
Age60_	-0.004	-0.005	-0.008*	-0.009**	0.003	0.003
	(0.007)	(0.007)	(0.004)	(0.004)	(0.005)	(0.005)
BadHealth	0.072***	0.060***	0.018**	0.004	0.054***	0.056***
	(0.016)	(0.016)	(0.009)	(0.009)	(0.014)	(0.014)
Hnight		0.011**		0.013***		-0.002***
		(0.005)		(0.005)		(0.001)
Constant	0.063***	0.059***	0.032***	0.027***	0.032***	0.032***
	(0.012)	(0.012)	(0.007)	(0.007)	(0.011)	(0.011)
Observations	6,999	6,999	6,999	6,999	6,999	6,999
R-squared	0.023	0.035	0.016	0.064	0.014	0.015

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 5b: Results without interaction terms (continuous indicators)

VARIABLES	(1) abs_error	(2) abs_error	(3) neg_error	(4) neg_error	(5) pos_error	(6) pos_error
w=30	-0.157***	-0.063***	0.127***	0.045***	-0.030*	-0.018
	(0.033)	(0.022)	(0.028)	(0.016)	(0.017)	(0.018)
w=91	-0.146***	-0.071**	0.120***	0.055***	-0.026	-0.016
	(0.036)	(0.029)	(0.028)	(0.019)	(0.023)	(0.024)
w=183	-0.046	0.009	0.107***	0.059***	0.061*	0.068**
	(0.043)	(0.035)	(0.028)	(0.019)	(0.033)	(0.034)
School	-0.015***	-0.013**	0.007**	0.005*	-0.008	-0.008
	(0.006)	(0.006)	(0.003)	(0.003)	(0.005)	(0.005)
Male	0.009	0.015	-0.002	-0.006	0.008	0.009
	(0.024)	(0.022)	(0.016)	(0.013)	(0.019)	(0.019)
Non-Nordic	0.085*	0.114**	0.012	-0.013	0.097**	0.101**
	(0.049)	(0.044)	(0.024)	(0.017)	(0.043)	(0.043)
ln(income)	0.007**	0.006**	-0.004***	-0.004**	0.002	0.002
	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)
Age31_45	0.008	-0.002	-0.015	-0.005	-0.006	-0.008
	(0.038)	(0.035)	(0.022)	(0.018)	(0.031)	(0.031)
Age45-60	-0.040	-0.018	0.025	0.006	-0.015	-0.012
	(0.034)	(0.033)	(0.018)	(0.016)	(0.030)	(0.029)
Age60_	0.002	-0.023	0.005	0.027	0.007	0.004
	(0.043)	(0.042)	(0.024)	(0.023)	(0.035)	(0.035)
BadHealth	0.581***	0.212**	-0.302**	0.020	0.279***	0.232***
	(0.157)	(0.091)	(0.138)	(0.058)	(0.080)	(0.074)
Hnight		0.335***		-0.291***		0.043
		(0.070)		(0.072)		(0.027)
Constant	0.289***	0.173**	-0.161***	-0.061	0.128**	0.113**
	(0.072)	(0.070)	(0.045)	(0.047)	(0.056)	(0.057)
Observations	6,999	6,999	6,999	6,999	6,999	6,999
R-squared	0.019	0.212	0.013	0.309	0.010	0.024
Sample	Full	Full	Full	Full	Full	Full

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 6a: Results with interactions (Binary Indicators)

VARIABLES	(1) Binary	(2) Binary	(3) Falseneg	(4) Falseneg	(5) falsepos	(6) Falsepos
w=30	-0.1040***	-0.0949***	-0.0797***	-0.0687***	-0.0243	-0.0262
	(0.030)	(0.029)	(0.018)	(0.016)	(0.024)	(0.025)
w=91	-0.0887***	-0.0781***	-0.0721***	-0.0592***	-0.0166	-0.0189
	(0.030)	(0.029)	(0.019)	(0.017)	(0.023)	(0.023)
w=183	-0.0397	-0.0311	-0.0796***	-0.0692***	0.0399	0.0381
	(0.035)	(0.034)	(0.022)	(0.020)	(0.028)	(0.028)
School	-0.0046**	-0.0041**	-0.0045***	-0.0039***	-0.0000	-0.0002
	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)
(*w*=30)*School	0.0058**	0.0054**	0.0045***	0.0039***	0.0013	0.0014
	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)
(*w*=91)*School	0.0045**	0.0038*	0.0040***	0.0032**	0.0005	0.0007
	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)
(*w*=183)*School	0.0021	0.0015	0.0051***	0.0045***	-0.0030	-0.0029
	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)
Constant	0.1014***	0.0925***	0.0732***	0.0624***	0.0282	0.0301
	(0.027)	(0.026)	(0.017)	(0.016)	(0.021)	(0.021)
Observations	6,999	6,999	6,999	6,999	6,999	6,999
R-squared	0.025	0.036	0.019	0.067	0.015	0.016
F-test	2.842	2.572	4.800	4.258	2.328	2.333
prob.	0.0364	0.0524	0.00242	0.00518	0.0725	0.0720

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Note: All models are estimated conditional on male, Non-Nordic, ln(income), age, BadHealth, and admission (For the binary indicators the number of hospital nights has been replaced with an indicator of whether the respondents have been admitted to the hospital or not.

Table 6b: Results withinteractions (Continuous Indicators)

VARIABLES	(1) abs_error	(2) abs_error	(3) neg_error	(4) neg_error	(5) pos_error	(6) pos_error
w=30	-0.6214***	-0.3428**	0.4850***	0.2427**	-0.1364	-0.1001
	(0.205)	(0.146)	(0.173)	(0.099)	(0.115)	(0.116)
w=91	-0.4014	-0.0745	0.4345**	0.1501	0.0331	0.0757
	(0.255)	(0.209)	(0.172)	(0.097)	(0.191)	(0.192)
w=183	-0.3041	-0.0405	0.5019***	0.2727***	0.1978	0.2321
	(0.261)	(0.216)	(0.174)	(0.105)	(0.197)	(0.197)
School	-0.0342**	-0.0194*	0.0280**	0.0152**	-0.0062	-0.0042
	(0.014)	(0.010)	(0.012)	(0.007)	(0.008)	(0.008)
(*w*=30)*School	0.0369**	0.0221**	-0.0285**	-0.0157**	0.0083	0.0064
	(0.015)	(0.011)	(0.012)	(0.007)	(0.009)	(0.009)
(*w*=91)*School	0.0205	0.0004	-0.0251**	-0.0077	-0.0047	-0.0073
	(0.018)	(0.015)	(0.012)	(0.008)	(0.014)	(0.014)
(*w*=183)*School	0.0206	0.0040	-0.0314**	-0.0169**	-0.0108	-0.0129
	(0.019)	(0.016)	(0.012)	(0.008)	(0.014)	(0.014)
Constant	0.5294***	0.2554*	-0.4222***	-0.1840**	0.1071	0.0714
	(0.180)	(0.130)	(0.151)	(0.094)	(0.101)	(0.104)
Observations	6,999	6,999	6,999	6,999	6,999	6,999
R-squared	0.021	0.245	0.018	0.414	0.010	0.016
F-test	2.729	2.220	2.923	2.069	1.159	1.137
prob.	0.0424	0.0836	0.0326	0.102	0.324	0.333

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Note: All models are estimated conditional on male, Non-Nordic, ln(income), age, BadHealth, and the number of hospital nights.

Table 7: Placebo analysis

VARIABLES	(1) alt_discrete_error2	(2) falseneg_alt	(3) falsepos_alt
w=30	-.0267	-.0027	-.0240
	(0.027)	(0.007)	(0.026)
w=91	.0068	.0057	.0012
	(0.028)	(0.010)	(0.026)
w=183	-.0228	-.00352	-.0193
	(0.028)	(0.007)	(0.028)
School	-.0008	.0002	-.0010
	(0.002)	(0.000)	(0.001)
(*w*=30)*School	.00176	0.0000	.0018
	(0.002)	(0.001)	(0.002)
(*w*=91)*School	-.0009	-.0006	-.0003
	(0.002)	(0.001)	(0.002)
(*w*=183)*School	.0011	.0001	.0010
	(0.002)	(0.001)	(0.002)
Constant	.0508**	-.0026	.0534***
	(0.021)	(0.006)	(0.020)
Observations	6,840	6,840	6,840
R-squared	0.019	0.141	0.008
F-test	0.621	0.432	0.434
prob.	0.601	0.730	0.729

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Note: All models includes the following control variables; male, Non-Nordic, ln(income), age, BadHealth, admission. The number of hospital nights has been replaced with an indicator of whether the respondents have been admitted to the hospital or not (admission). The F-statistic tests the joint hypothesis of all interaction being equal zero.

Figure 1: Magnitude of positive and negative errors by recall period

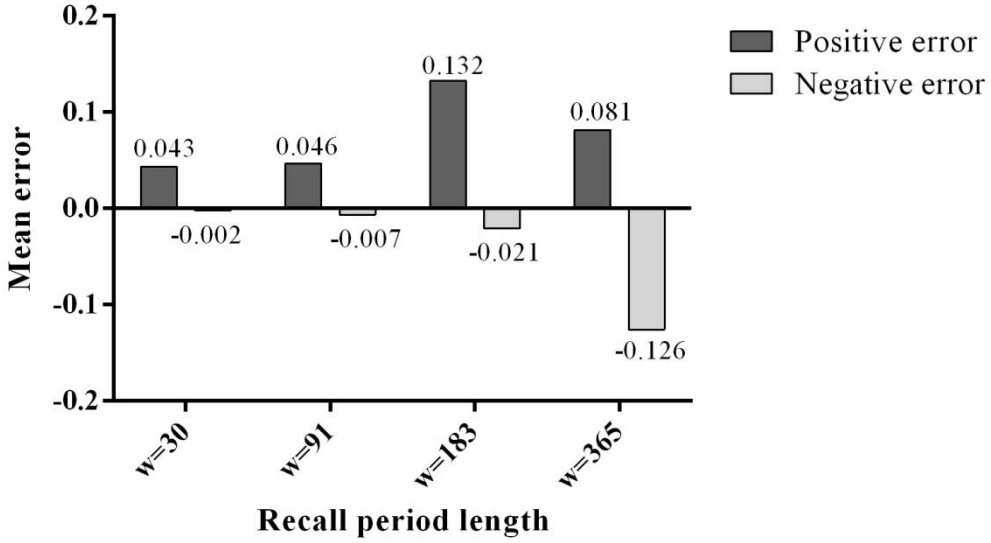


Figure 2: Individual level error in relation to the recall window

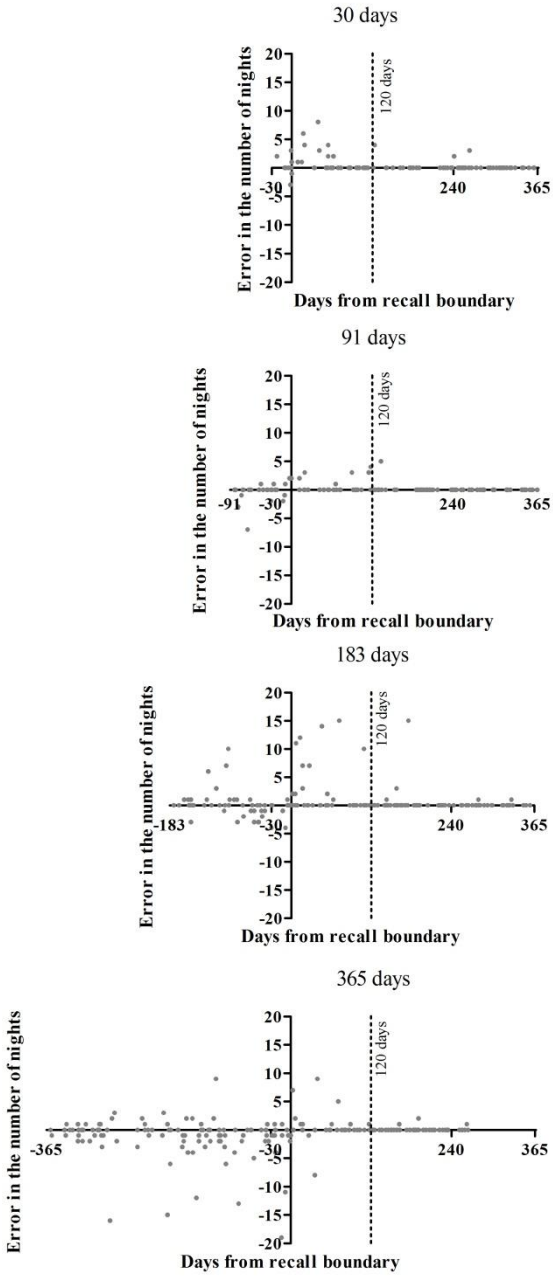
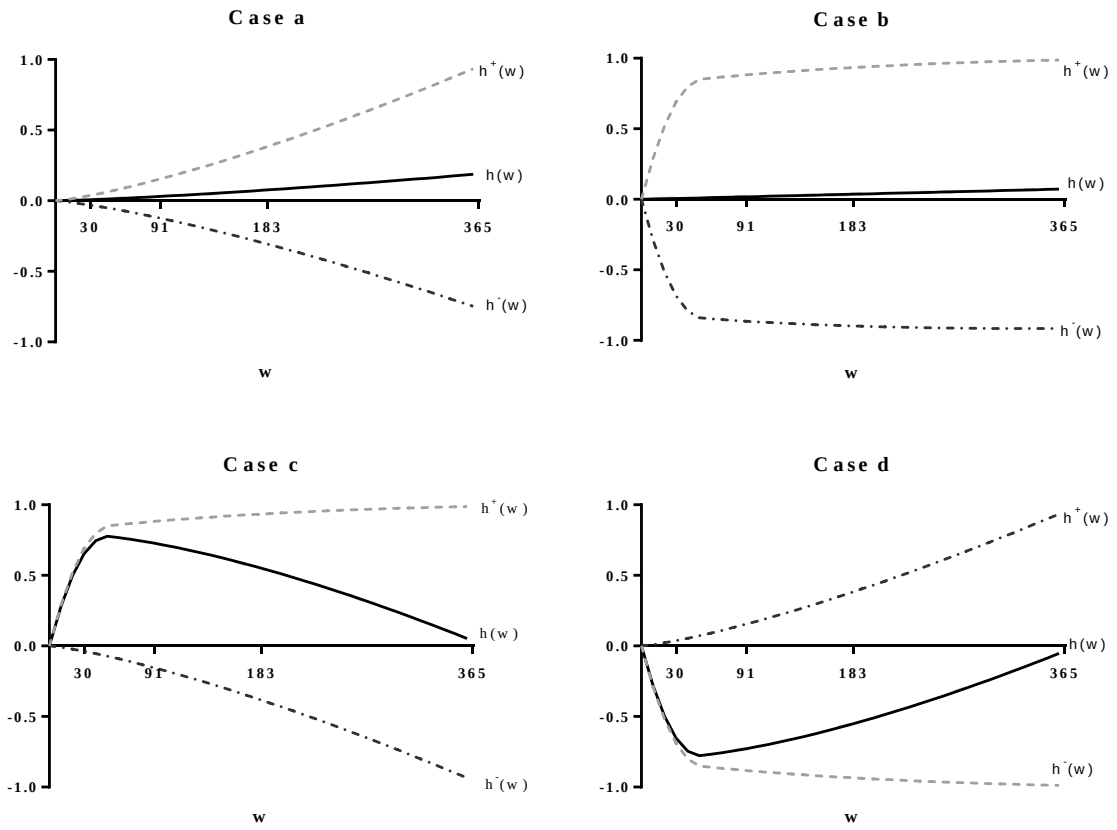


Figure 3: Hypothetical Error Structures



Note: *w* denotes the length of the recall period.

Figure 4: RMSE, Variance, Bias

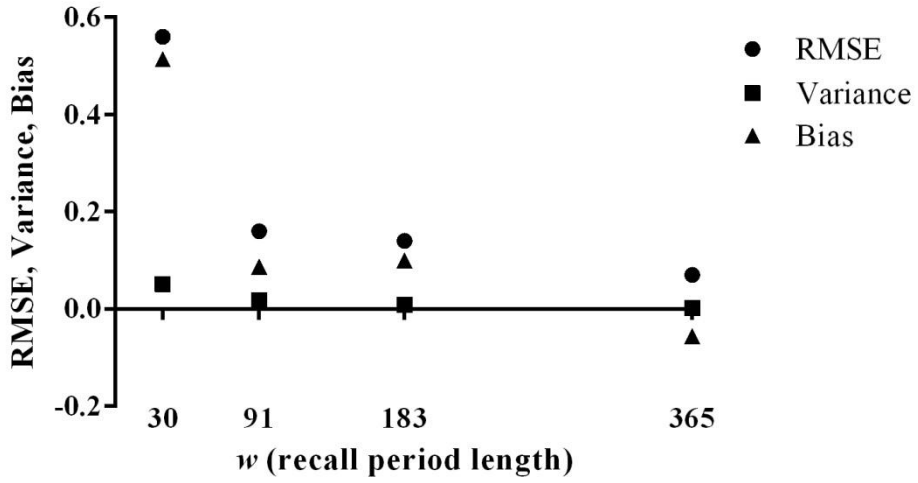
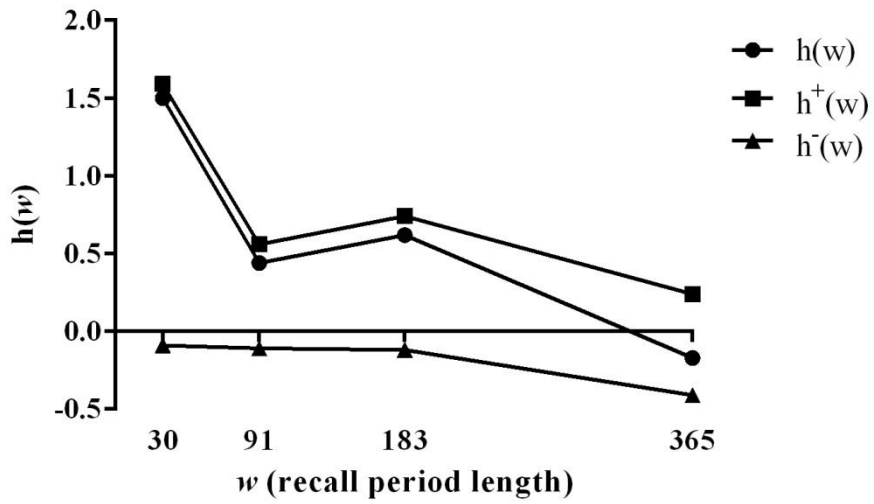


Figure 5: Observed Error Structure



Appendix

A1. Framework for an optimal recall period in a binary case

For the binary case the survey design problem consists of estimating the probability of spending at least one night at the hospital during one year. The measurement errors are now the probability of being a false positive $\Pr(X_i^w > 0 | Y_i^w = 0)$ and the probability of being a false negative $\Pr(X_i^w = 0 | Y_i^w > 0)$. The major difference from the continuous case is the imputation process when scaling the probability of consuming care during *w* to the probability of consuming care during S. For simplicity, Clarke et al. (2008) again use *Eq(2)* and the same imputation process as in the continuous case, i.e., $\Pr\left(X_i^S > 0\right) = \left(\frac{S}{w}\right) \Pr(X_i^w > 0)$. For the binary case, this relationship only holds if the individuals that have been hospitalized during the recall period are (very) unlikely to also have been hospitalized during the remaining part of the target period S (i.e. the part of S that do not coincide with *w*). Because this assumption is unlikely to hold, such a scaling procedure would induce further bias by overestimating $\Pr\left(X_i^S > 0\right)$. We therefore complement $\left(\frac{S}{w}\right)$ by a scaling factor less likely to overestimate the probability of being hospitalized during S; the ratio between the observed probability of consuming care during S and the observed probability of consuming care during *w*, i.e. $\left(\frac{\Pr(\widehat{Y_i^S} > 0)}{\Pr(\widehat{Y_i^w} > 0)}\right)$. [12] For this imputation process, the two definitions of the bias *Eq*(5) and *Eq*(7) are equal. The results are presented in Table A1-A2 below and are in line the results from the continuous case. RMSE, variance, and bias all decrease as the recall window increase indicating that the one year period is preferable.

Table A1a: Descriptive statistics

Reported data		**Registry data**		
		0	**1**	**Total**
w=30	**0**	1,675	0	1,675
	1	20	9	29
	Total	1,695	9	1,704
				Total
w=91	**0**	1,679	1	1,680
	1	15	27	42
	Total	1,694	28	1,722

[12] That is, we assume the imputation process to be the same for the self-reported as for the registered information. We calculate the ratio individually for each group. Observe, however, that we treat the scaling function as known (i.e. a constant with zero variance) to avoid exaggeration of the variance for the shorter periods.

				Total
w=183	**0**	1,568	13	1,581
	1	33	48	81
	Total	1,601	61	1,662
				Total
w=365	**0**	1,751	45	1,796
	1	38	77	115
	Total	1,789	122	1,911

Note: reported/registered events are coded as 1.

Table A1b: Proportions of false negative and false positive

	False negative		**False positive**	
w=30	(0/9)=	0.000	(20/1695)=	0.012
w=91	(1/28)=	0.036	(15/1694)=	0.009
w=183	(13/61=	0.213	(33/1601)=	0.021
w=365	(45/122)=	0.369	(38/1789)=	0.021

TableA2: Aggregated measure; Binary case

	$E(X_i^W)$	$f(.)E(X_i^W)$	(S/w) $E(X_i^W)$	$E(Y_i^W)$	$E(Y_i^S)$	$\frac{\Pr(\widehat{Y_i^S} > 0)}{\Pr(\widehat{Y_i^w} > 0)}$	$E(X_i^W) - E(Y_i^W)$	**Var(** X_i^W)
w=30	0.017	0.174	0.207	0.005	0.054	10.222	0.012	0.017
w=91	0.024	0.080	0.098	0.016	0.053	3.286	0.008	0.024
w=183	0.049	0.093	0.097	0.037	0.070	1.918	0.012	0.046
w=365	0.060	0.060	0.060	0.064	0.064	1	-0.004	0.057

Scaling procedure	$\Pr(Y_i^S > 0)/\Pr(Y_i^w > 0)$			S/w		
	RMSE	**Var**($\bar{X}_W^S$)	Bias($\bar{X}_W^S$)	RMSE	**Var**($\bar{X}_W^S$)	Bias($\bar{X}_W^S$)
w=30	0.124	0.00103	0.120	0.158	0.00145	0.153
w=91	0.029	0.00015	0.027	0.047	0.00022	0.044
w=183	0.025	0.00010	0.023	0.029	0.00011	0.027
w=365	0.007	0.00003	-0.004	0.007	0.00003	-0.004

Figure 1: Magnitude of positive and negative errors by recall period

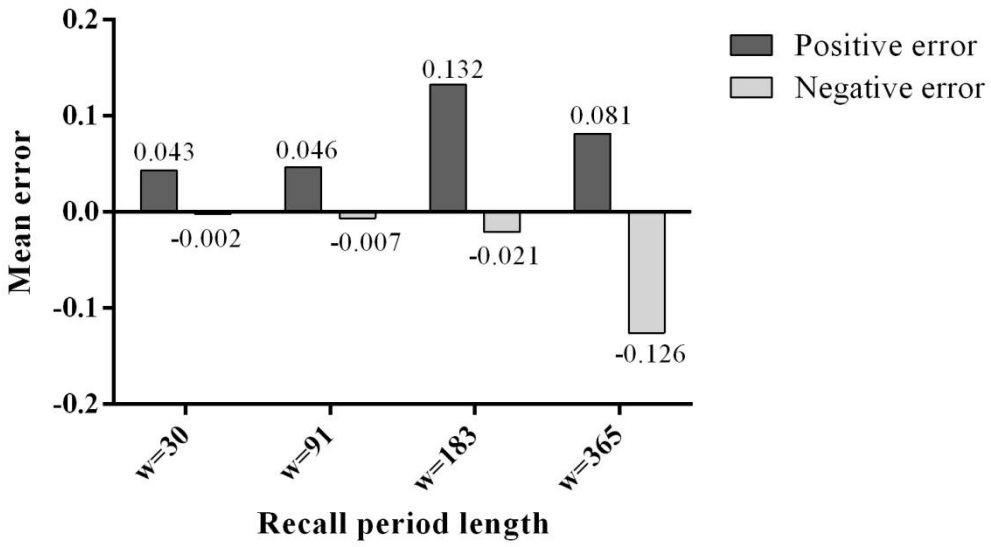


Figure 2: Individual level error in relation to the recall window

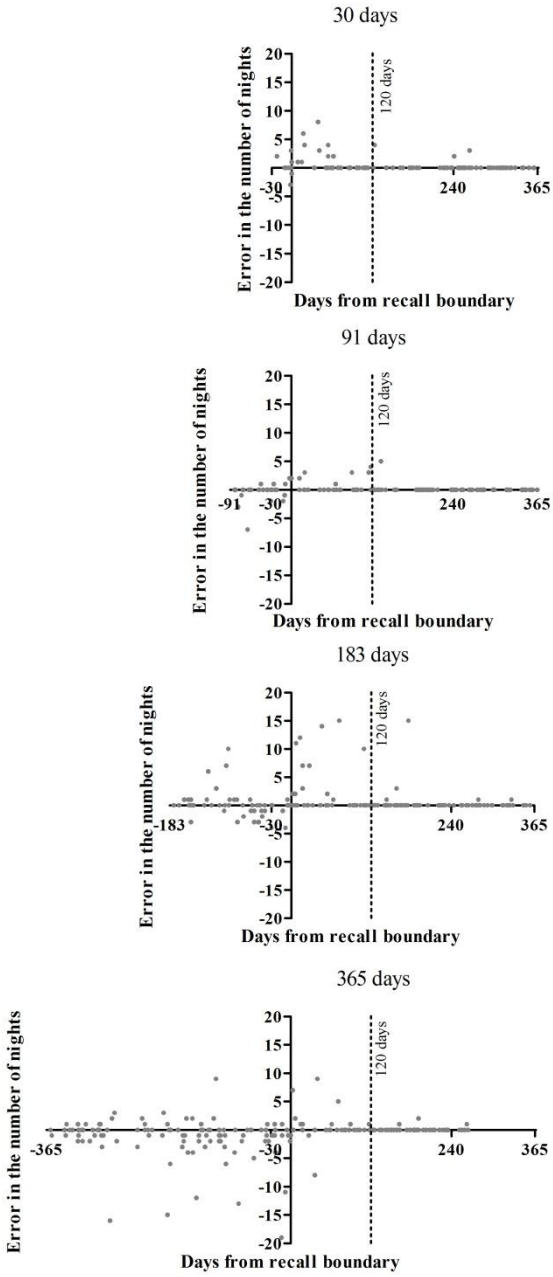
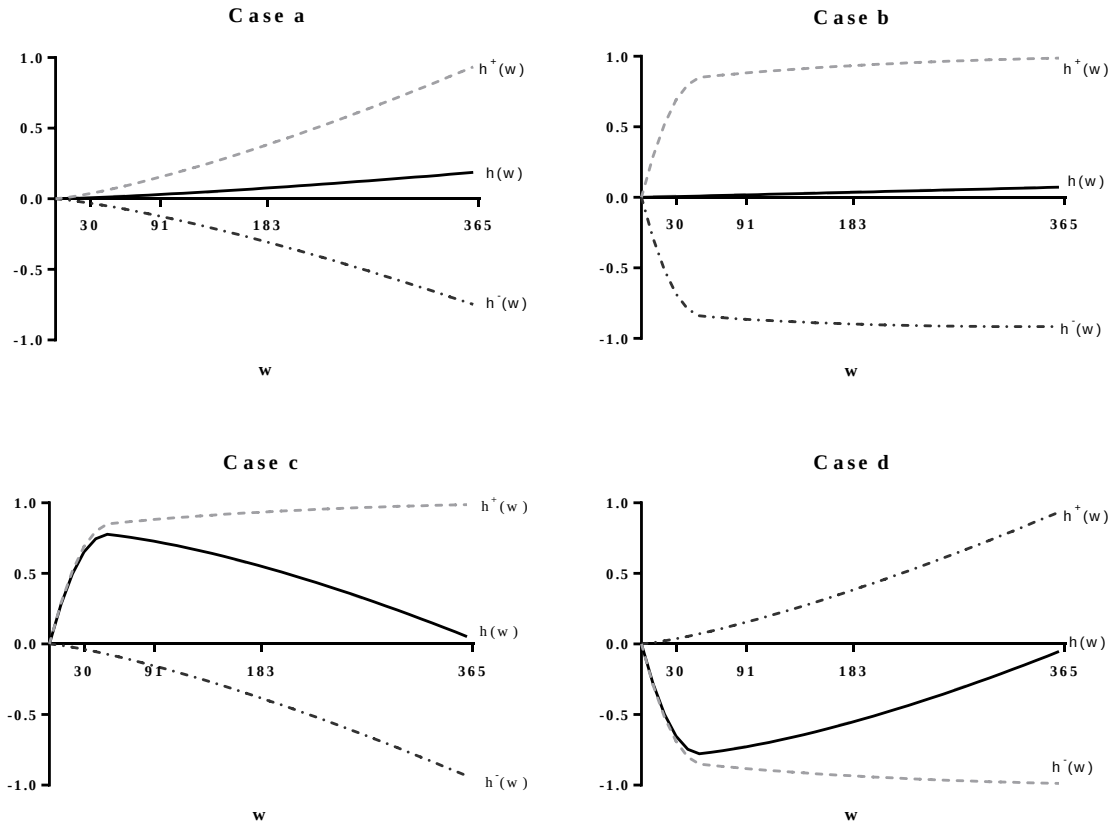


Figure 3: Hypothetical Error Structures



Note: *w* denotes the length of the recall period.

Figure 4: RMSE, Variance, Bias

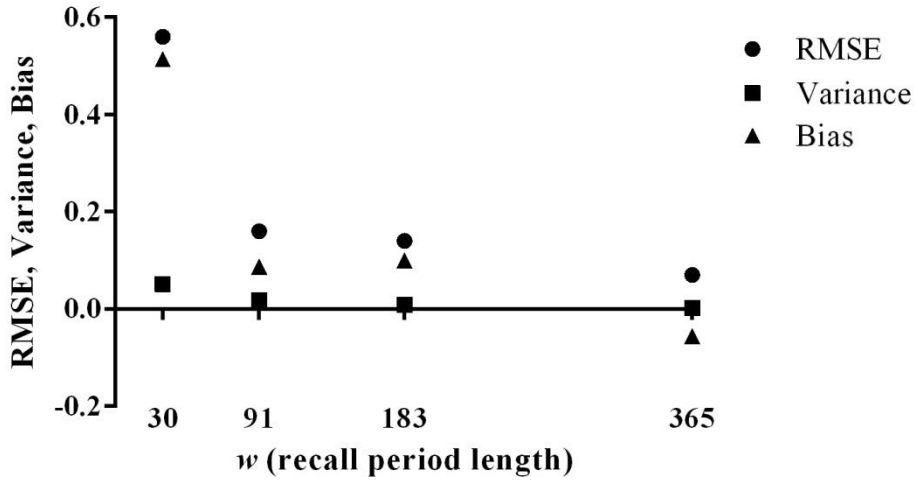
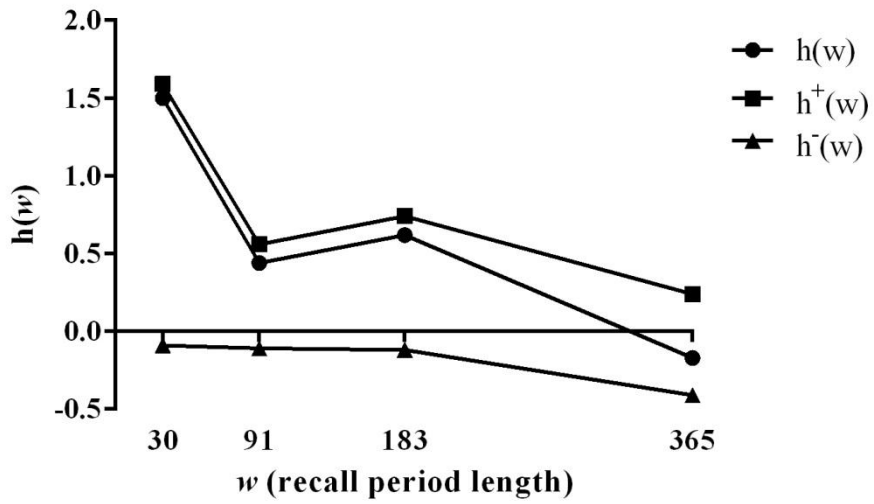



Figure 5: Observed Error Structure