Metodología de Encuestas
Volumen 15, 2013, 5-25
ISSN: 1575-7803

Implications of Retrospective Measurement Error in Event History Analysis

José Pina-Sánchez
Johan Koskinen
Ian Plewis

University of Manchester

Correo electrónico: Jose.pinasanchez@postgrad.manchester.ac.uk

Dirección: CCSR
School of Social Sciences
Humanities Bridgeford St Building
University of Manchester
Manchester
M13 9PL

Las consecuencias de los errores de medida retrospectivos en análisis de la supervivencia.

RESUMEN: Es comúnmente aceptado que el uso de preguntas retrospectivas en las encuestas requiere un mayor esfuerzo cognitivo por parte del encuestado y por lo tanto conduce a mediciones menos precisas. En este trabajo se evalúa el efecto de utilizar los datos derivados de las preguntas retrospectivas como la variable de respuesta en diferentes modelos de análisis del historial de eventos: Weibull, exponencial, Cox y logit. El impacto del error de medición se evalúa mediante la comparación de las estimaciones obtenidas al especificar los modelos usando duraciones de desempleo derivadas de una pregunta retrospectiva frente a aquellas obtenidas usando los datos de validación derivados del registro sueco de desempleo. Los resultados muestran grandes efectos de atenuación en todos los coeficientes de regresión. Además, estos efectos son relativamente similares en todos los modelos estudiados.

PALABRAS CLAVE: Error de medición, Análisis de las biografías, Pregunta retrospectiva, Datos de registro, Desempleo.

Implications of Retrospective Measurement Error in Event History Analysis.

ABSTRACT: It is commonly accepted that the use of retrospective questions in surveys makes interviewees face harder cognitive challenges and therefore leads to less precise measures than questions asking about current states. In this paper we evaluate the effect of using data derived from retrospective questions as the response variable in different event history analysis models: an accelerated life Weibull, an accelerated life exponential, a proportional hazards Cox, and a proportional odds logit. The impact of measurement error is assessed by a comparison of the estimates obtained when the models are specified using durations of unemployment derived from a retrospective question against those obtained using validation data derived from the Swedish register of unemployment. Results show large attenuation effects in all the regression coefficients. Furthermore, these effects are relatively similar across models.

KEYWORDS: Measurement error, Event history analysis, Retrospective question, Register data, Unemployment.

Recibido: 2 de julio de 2013
Revisado: 4 de noviembre de 2013
Aceptado: 14 de noviembre de 2013

1. Introduction

Retrospective questions are a widely used tool in surveys when there is an interest in capturing changes over time. These types of questions ask respondents for information about events from the past. They obtain information about a particular timespan on a single occasion, and thus normally turn out to be cheaper than the alternative approach of repeatedly contacting respondents during that time span as in longitudinal or prospective designs.

Since the interviewee is contacted only once, there is no risk of attrition (that is subjects dropping out of the study) or lack of consistency derived from, for example, changes in the wording of questions over time. Moreover, retrospective questions can capture information on the full history of an event for a particular period of time, whereas repeated questions on current state are only able to provide a series of snapshots[1].

The major problem for retrospective questions stems from their higher propensity to generate measurement error (ME) in the responses. In particular, interviewees answering retrospective questions are faced with a higher cognitive challenge since not only do they need to interpret the question correctly but they also need to recall it. Furthermore, the memory failures that generate ME in retrospective questions are often interrelated with the nature of the topic and with the relative difficulty of reporting it (low saliency, social desirability, etc.), resulting in complex error-generating mechanisms[2].

In this paper we study the implications of using data collected from retrospective questions in statistical models used for longitudinal data. In particular we consider the consequences of using data derived from these questions as the response variable in event history analysis (EHA) models. The impact of ME is assessed by comparing estimates obtained from models that are specified using durations of unemployment derived from a retrospective question against those obtained using validation data derived from a register of unemployment.

In choosing to study the consequences of ME in the response variable of EHA models we address an area which has not been widely researched. In the analysis of ME a majority of studies have focused on settings where the explanatory variables were those which were prone to ME, in what is known as the “errors in variables” problem. This focus on the predictors can be explained from the general opinion that ME affecting the response variable only affects the precision of the model and thus it is a lesser problem. In addition, the study of ME was until recently restricted to analyses using linear models, with the seminal work of Fuller (1987) as the main reference. In the last decade the study of ME has been extended to other non-linear models until recently, especially after the publication of the work of Carroll, Ruppert, Stefanski and Crainiceanu (2006). However, the study of ME in EHA

[1] See Solga 2001 for a comparison of data quality derived from prospective and retrospective questions.

[2] See Pina-Sánchez, Koskinen, and Plewis (2013) for an analysis of the error generating mechanisms affecting retrospective questions on unemployment.

models has been identified by many authors as an area which requires further research. (Augustin (1999: 2), Pyy-Martikainen and Rendtel (2009: 140), Skinner and Humphries (1999: 23), and Jäckle (2008: 2).

The paper is structured as follows: in the next section we present a summary of findings from other studies in the literature, in Section 3 we describe the characteristics of the data that we use in our analysis, in Section 4 we present the results of our analyses, and in Section 5 we conclude with a summary of the results and how these relate to previous research.

2. Literature Review

According to the research design used, we can identify two main groups of studies which have assessed the impact of ME in EHA. These can be either analytical or empirical. The former imply tracing out the impact of ME in EHA models algebraically. However, because of the greater complexity of EHA models the number of settings explored is much more limited than in the case of linear models. In fact, until the 1990s research was concentrated on classical ME affecting covariates in the proportional hazards (PH) Cox model. Some examples are Prentice (1982) and Nakamura (1992) who presented an analytical development of the bias found in the parameter estimates of PH Cox models with classical ME[3] in the covariates. In this context, both authors found attenuation bias in all the regression coefficients.

The only studies that have explored the impact of ME on the response variable in EHA models analytically are Augustin (1999) and Dumangane (2007). They used accelerated life (AL) Weibull models and assumed classical multiplicative errors affecting the recall of durations. In this case, ME in the response was found to produce an attenuation bias in the regression coefficients. However, this particular setting does not account for other types of errors observed in retrospective questions on work histories, such as omission of spells, or misclassification of status. In addition, Augustin (1999) requires the assumption of no right censoring in the data and Dumangane (2007) assumed that the true duration and error distributions are independent. The set of assumptions used in these papers shows both the difficulty of studying the effect of ME in the response variable of EHA models analytically, and how the general expressions developed so far are not really representative of the problems found in retrospective data, which are prone to other types of ME besides mismeasured durations.

Another group of studies assessing the impact of ME in EHA are those which are based on empirical analysis. These studies compare estimates derived from a model that uses prone to ME data against the estimates obtained from replicating the same model but using data free of ME. Korn, Dodd and Freidlin (2010) studied the effects of ME in a PH Weibull model by means of simulating multiplicative log-normal

[3] This is the most commonly used specification of ME, which has its origins in the classical test-theory, “classical test theory […] postulates the existence of a true score, that error scores are uncorrelated with each other and with true scores and that observed, true and error scores are linearly related”. (Novick, p. 1, 1966).

errors in the response. The authors found small downward biases in the hazard rate as long as the ME remains non-differential (i.e. the ME and the error term of the EHA model are independent) and hazard rates relatively high. Considering non-parametric models for discrete data Meier, Richardson and Hughes (2003) assess the bias in the regression coefficients produced by simulating different levels of non-differential false positives and false negatives[4]. The authors conclude that the bias is always toward the null, and that false positives induce greater bias in estimation of the cumulative distribution function and regression coefficients than false negatives when the failure rate is low.

This last group of studies contribute to the understanding of the effect of ME affecting the response variable in duration models, however, just like the studies of Augustin (1999), Dumangane (2007) they consider relatively simple forms of ME. In addition to shortened or extended durations, retrospective ME can also take the form of omitted spells[5] and misclassified status[6].

Turning to studies that use real data, Jäckle (2008) found that ME in the reporting and dating of receipt of unemployment benefits using retrospective questions attenuated both the duration dependence and the regression coefficients from PH cloglog and Weibull models when compared with data from an unemployment register. The recall period used by the retrospective question was only four months, which is perhaps not long enough to see the typical memory failures that characterize retrospective data.

Pyy-Martikainen and Rendtel (2009) used the more common recall frame of one year. Specifically, the authors combined work histories retrospectively reported in five consecutive waves of the European Community Household Panel and matched them against a gold standard obtained from the Finnish register of unemployment. PH Cox and Weibull models for unordered repeated events were specified for the duration of unemployment and both attenuation and augmentation bias were found in the regression coefficients. None of these biases changed the survey estimates by more than 30%, and they were found in the same direction and similar magnitude for both the Cox and Weibull models. Moreover, the comparison of the Cox and Weibull models shows that the baseline hazard was more accurately estimated by the former. The survey baseline hazard from the Weibull model is nearly constant while the register baseline hazard shows positive duration dependence leading to erroneous conclusions about duration dependence. However, the Cox baseline hazards from survey and register both display positive duration dependence.

In summary, it seems that when the response variable of EHA models, regardless of how it is defined (duration logs, hazard rates, or person period cases), is affected

[4] The term false positive refers to values of a binary variable prone to misclassification which indicate that an effect has been observed when none existed. By false negatives the opposite is understood, that is, cases showing no effect when one really existed.

[5] Levine (1993), comparing retrospective questions using a one year recall time with questions asking about the current work status, found that between 35% and 60% of persons failed to report at least one spell of unemployment.

[6] Bound (2001) in a review of the literature concludes that in cross-sectional surveys 11-16% of respondents stated to be unemployed are likely to be misclassified.

by non-differential ME, the regression coefficients of the model are attenuated. On the other hand, when the ME is associated with some of the explanatory variables, the direction of the bias in the coefficients cannot be anticipated. Finally, because of the complexity of tracing the impact of ME in EHA models analytically, more empirical studies using validation datasets are necessary in order to assess both the peculiarities of retrospective ME and the consequences of these types of errors. At present we are only aware of Jäckle (2008), and Pyy-Martikainen and Rendtel (2009).

3. Data

The data we use has been obtained from the "Longitudinal Study of the Unemployed", a research project designed by the Swedish Institute for Social Research (SOFI) at Stockholm University, directed by Sten-Ake Stenberg, and with the collaboration of the register of unemployment (PRESO[7]). This register provided individual-level data on the work status of the participants of three surveys, which were carried out in 1992, 1993 and 2001. The three surveys are relatively similar with respect to the composition of both the sample of participants and the questionnaire. The target sample was composed of subjects registered as unemployed on 28th February 1992 from ages 25 to 55.

In this study we use data derived from a retrospective question on work status from the 1993 survey. This question uses an event-occurrence framework (Lawless, 2003). In particular the question reads as follows:

"Which of the alternative answers on the response card best describes your main activity the first week of 1992? When did this activity start? When did it end?

Which was the subsequent main activity? When did this activity start? When did it end?[8]

In order to simplify the observation scheme we set the beginning of the window of observation at February 28th and only consider subjects who started from a state of unemployment in both the register and the survey. This could be considered the most sensitive approach to follow for researchers who only have access to survey data. That is, in order to reduce the impact of ME, and making use of what is known regarding the sample design, subjects who appeared to have misclassified their work status on 28th February are discarded.

With that proviso our sample shares the structure seen in state-based samples (Holt, McDonald and Skinner, 1991), where the sample frame is created out of individuals who are known to be in a particular state. Our final sample size captures 381 individuals and the window of observation encompasses spells from 28/2/92 to 30/03/93, where the ending date represents the earliest day interviews were taken. Right censoring is present in both datasets. Extensions to account for unobserved heterogeneity[9] were not implemented.

[7] PRESO is a register from the Swedish employment office (Arbetsmarknadsstyrelsen).

[8] This and the following quote are translations from the original in Swedish.

[9] Frailty models are not considered because the interest lies in ascertaining the impact of ME in the regression coefficients, not in finding causal associations between the response and the explanatory

The explanatory variables in the models considered here are age, experience, and their interaction term. Although durations of unemployment would be more appropriately specified using additional variables such as education or gender, we decided to include a short list of explanatory variables to facilitate comparisons of the effect of ME in the regression coefficients between different EHA models.

Experience captures self-reported levels of experience in the type of work that the subject applied for on a scale with three levels (low, medium, and high). Both variables are drawn from the register; the value for age is taken in January 1993, while for experience the mean of the monthly reported levels in 1992 is used. Given that age is an important variable in the register the probability that it is prone to ME is very low. This is different for experience since it is a self-reported value. However, in our analysis we assume that both of them are free of ME. In our sample the mean age is 37 and the standard deviation 8.8, while for experience these are 2.59 and .60 respectively. Finally, regarding these two variables the ME can be considered non-differential since the Spearman correlation coefficients for the misclassification of person-day observations with age and experience were .01 and .03, respectively[10].

4. Analysis

In order to assess the impact of retrospective ME affecting the response variable in EHA models we use a design similar to Jäckle (2008), and Pyy-Martikainen and Rendtel (2009). We specify EHA models using duration of spells of unemployment derived from the retrospective question presented in the previous section and compare their estimates to the ones that are obtained by specifying the same type of models, for the same subjects, time-frame, and explanatory variables, but using durations derived from PRESO, the Swedish register of unemployment. This register is assumed to be a gold standard; consequently differences in the estimates of the models using survey data with respect to those obtained using register data are understood as evidence of the impact of ME. For the sake of completeness we analyse the effect of ME on four different models. These are: an AL Weibull and an AL exponential representing parametric models, a PH Cox from the semi-parametric models, and a proportional odds (PO) logit representing non-parametric models.

We use four measures to assess the differences found in the regression coefficients when the models are specified using the survey and the register data. The simplest of the four is the bias, calculated as the difference between the regression coefficient obtained from the model using survey data and the same obtained using register data,

variables.

[10] This result is also corroborated when ME is operationalised as the difference between the survey and register durations. In this case the Pearson correlation coefficient of age and experience with the ME is -.01 and .07, respectively.

$$BIAS = \hat{\beta}_s - \hat{\beta}_r \tag{1}$$

where s stands for survey and r for register. A second measure particularly useful for making comparisons between models and between explanatory variables that use different scales for their regression coefficients is the relative bias,

$$R.BIAS = \frac{\left|\left(\hat{\beta}_s - \hat{\beta}_r\right)100\right|}{\left|\hat{\beta}_r\right|} \tag{2}$$

In order to take into account impacts on the precision of the estimates we also use the root mean squared error, which is the square root of the addition of the squared bias and the variance of the regression coefficient obtained from the survey,

$$RMSE\left(\hat{\beta}_s\right) = \sqrt{MSE\left(\hat{\beta}_s\right)} = \sqrt{E\left[\hat{\beta}_s - \hat{\beta}_r\right]} = \sqrt{Var\left(\hat{\beta}_s\right) + (BIAS)^2} \tag{3}$$

Finally, in order to facilitate comparisons between models in terms of the RMSE, we also use the relative root mean squared error,

$$R.RMSE = \frac{\left(RMSE\left(\hat{\beta}_s\right) - RMSE\left(\hat{\beta}_r\right)\right)100}{RMSE\left(\hat{\beta}_r\right)} \tag{4}$$

We start the study of the impact of ME in EHA using an exploratory analysis. Figure 1 below shows the Kaplan-Meier estimate of the survivor functions for the registered and reported time in unemployment. The two datasets show a similar path for the first 30 days; from that point until about day 100 the two measures diverge due to an accelerated failure rate in the survey; from then on the two survivor functions behave roughly similarly and the gap between them is maintained. At the end of the window of observation 35% (n = 133) of the spells of unemployment in the register were right-censored, whereas in the survey this was only 6% (n = 21).

Measures of central tendency for the registered and reported durations also show substantial differences. These are included in Table 1 together with their standard deviations.

The mean duration in the register is 241, while the median is 253 days. In the survey these figures were 136 and 92, respectively. The higher median than the mean in the register indicates that the probability density function of durations is skewed to the left, whereas the opposite can be deduced for the distribution of durations from the survey. On the other hand, measures of dispersion are similar. These features can be seen graphically in Figure 2, where the probability density functions for spells of unemployment in the register and survey data are plotted.

Figure 1.
Survivor function for the register and survey data

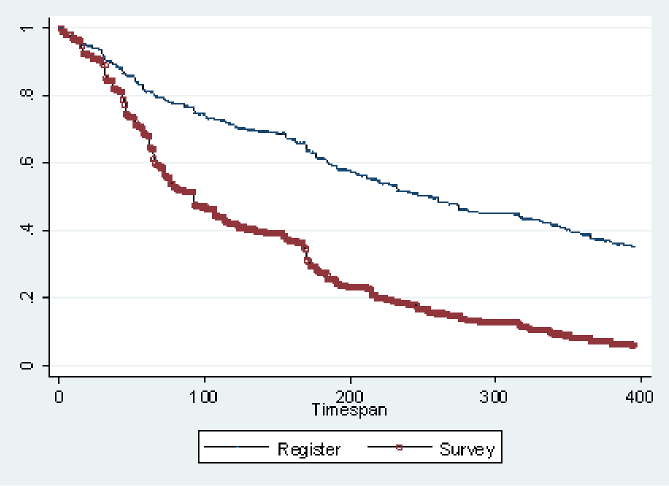


Table 1.
Descriptive Statistics of the Unemployment Durations

	Mean	Median	Std Dev
Register	241	253	145
Survey	136	92	113

The impact of using this prone to error data in EHA models is analysed next. A separate description is included for each family of EHA models, and at the end their relative performance in the presence of ME is assessed.

In Tables 2 and 3 below we show the results obtained when comparing the AL Weibull models using register and survey data. In the model using the register data the main effects for both age and experience are negative and statistically significant, while their interaction effect is also significant but positive. When the main effects of age and experience are taken into account, we found that the older and more experienced the subjects are, the longer it will take them to obtain employment. However, this claim is attenuated by the positive interaction term, which indicates that subjects who are both old and highly experienced make that transition more rapidly.

Figure 2.

Probability density function of single spells

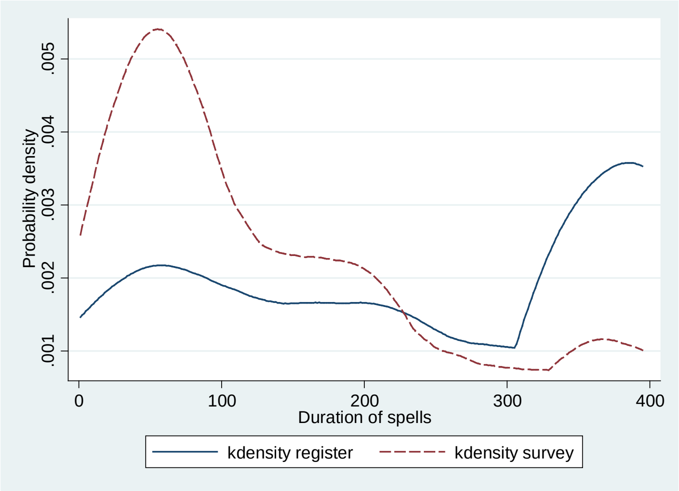


Table 2.

AL Weibull model using register and survey data,***

	Register		Survey	
Age	-0.087	(0.039)	0.001	(0.029)
Experience	-1.38	(0.51)	-0.09	(0.38)
Age*Exp	0.038	(0.014)	0.001	(0.010)
Constant	9.08	(1.40)	5.07	(1.02)
α	0.98	(0.05)	1.11	(0.05)
LR Chi^2 (3)	11.34		0.98	

*From here on estimates standard errors are represented between brackets to the right of the regression coefficients, which when in bold indicate that they were significantly different from zero at the 5% level.

**α represents the shape parameter of the baseline hazard function.

In considering the impact of ME, the first result to note is the attenuation of all the regression coefficients -age, experience etc- their interaction, and the constant, as a consequence of using survey data. It can be argued that attenuation bias represents the least worst type of bias since it only buffers the estimated effect size, therefore leading to type II errors (Korn et al. 2010). However, the substantial size of the biases found here makes them non-negligible. Standard errors (SE) of the regression coefficients have also been underestimated, although this was to be expected given the attenuation of the regression coefficients, which now represent smaller effects.

Table 3 shows the four measures set out at the beginning of this section to assess the impact of ME. The relative biases in the coefficients of the explanatory variables age and experience are very large, 101.1% and 93.1%, respectively. These results indicate that the size of the bias is roughly the size of the true estimate. The interaction effect suffers from a similar effect, with a R.BIAS of 97.4%; being more moderate only in the case of the constant term, 44.2%.

Table 3.
*Bias in the AL Weibull model**

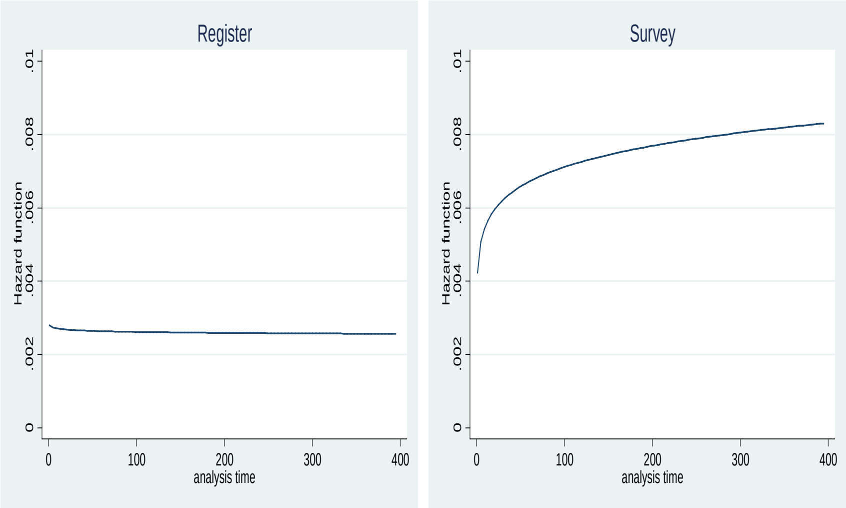
	BIAS	R.BIAS	RMSE	R.RMSE
Age	0.088	101.1%	0.093	137.6%
Experience	1.290	93.1%	1.340	161.9%
Age*Exp	-0.037	97.4%	0.038	173.8%
Constant	-4.010	44.2%	4.140	196.6%
α	0.130	13.3%	0.139	178.6%

The impact of ME on α, the parameter used in the Weibull model to estimate the shape of the baseline hazard function, might seem relatively unimportant compared to what has been seen in the other coefficients since the true estimate is 0.98 and the one found using survey data is 1.11. However, as Skinner and Humphreys (1999) point out, in some settings, there is interest not only in the size of this estimate but also in the distinction between $\alpha<1$, $\alpha=1$, and $\alpha>1$, or equivalently between a decreasing, constant or increasing hazard function respectively. For example, Chesher, Dumangane and Smith (2002) anticipate that in the analysis of unemployment durations it is well known that uncontrolled across-individual heterogeneity in hazard functions can lead to the appearance of negative duration dependence. In our case we observe the opposite effect when the model is specified using survey data, while the model using register data shows no effect in either direction. Here the impact of ME differs from what we have seen for the rest of coefficients, indicating a positive effect where there is none, which represents a type I error.

Figure 3 shows the shapes of the baseline hazard functions for the register and the survey data. In spite of the different signs of the slopes, it is worth noting that the shape of the baseline hazard function from the survey data mimics quite well the one from the register data. However, this result was to be expected. Due to the constraints of the Weibull model, where only one shape parameter is used, baseline hazard functions are bound to be either monotonically increasing or decreasing.

Another characteristic to be noted in Figure 3 is the flatness of both hazard functions, which are almost constant across the window of observation. This feature suggests the possibility of using a simpler model to parameterize the baseline hazard function. In particular, the AL exponential appears to be a good alternative because it assumes a constant baseline hazard function.

Figure 3.
Weibull baseline hazard function for the register and survey data



A likelihood ratio test between the two models using register data (taking the exponential model to be nested in the Weibull) corroborates this intuition. The test shows that the difference in deviances (0.13) for 1 degree of freedom is not statistically significant (p > 0.7). The better specification of the exponential model can also be concluded from the lower SEs for age, experience and the constant term. The results are shown in Tables 4 and 5.

Table 4.
AL Exponential model using register and survey data

	Register		Survey	
Age	**-0.087**	(0.038)	-0.003	(0.032)
Experience	**-1.370**	(0.500)	-0.110	(0.420)
Age*Exp	**0.037**	(0.014)	0.002	(0.012)
Constant	**9.040**	(1.370)	**5.080**	(1.140)
LR Chi^2 (3)	**11.47**		0.81	

In addition, the exponential model seems to perform marginally better at buffering the effects of ME; at least in terms of R.BIAS which is now lower for all the coefficients. It is possible that parametric EHA models are more sensitive to ME in the response when the baseline hazard function is misspecified.

Table 5.
Bias in the AL Exponential mode

	BIAS	R.BIAS	RMSE	R.RMSE
Age	0.084	96.6%	0.090	136.5%
Experience	1.250	91.7%	1.320	163.9%
Age*Exp	-0.035	94.6%	0.037	164.3%
Constant	-3.950	43.7%	4.120	200.9%

Tables 6 and 7 below show the results of the PH Cox. Estimates from the PH Cox model are often presented on a hazard rate scale. However, here we show the untransformed coefficients to facilitate comparisons between models[11].

Table 6.
*PH Cox model using register and survey data**

	Register		Survey	
Age	**0.086**	(0.038)	0.002	(0.032)
Experience	**1.350**	(0.500)	0.130	(0.420)
Age*Exp	**-0.037**	(0.014)	-0.002	(0.012)
LR Chi^2 (3)	**11.03**		0.77	

*Compared to the AL models signs of regression coefficients are now reversed since an increase in the hazards corresponds to a decrease in the expected (log-) durations, and vice-versa.

Results regarding the impact of ME on the PH Cox model show a very similar picture to what was found in the previous models. The regression coefficients are again heavily attenuated. Interestingly, the PH Cox model performs slightly better in terms of RMSE. This is surprising given the higher precision obtained from parametric models when they are correctly specified. This result suggests that, in spite of using an optimal parametric form for the true durations (from the register data), the baseline hazard function will probably change when there is ME, and less restrictive models such as the PH Cox are then a better choice.

The Cox baseline functions for the survey and register data are displayed in Figure 4. From the comparison of the two functions it can be seen that the former is overestimated, as it was in the Weibull model. Furthermore, now that the baseline function is freely estimated, it can be seen that the survey data exaggerated two bumps around days 220 and 330. These shocks were not captured by the exponential nor the Weibull baseline functions for survey data because of their parametric restriction. However, since the Cox baseline function using survey data remains

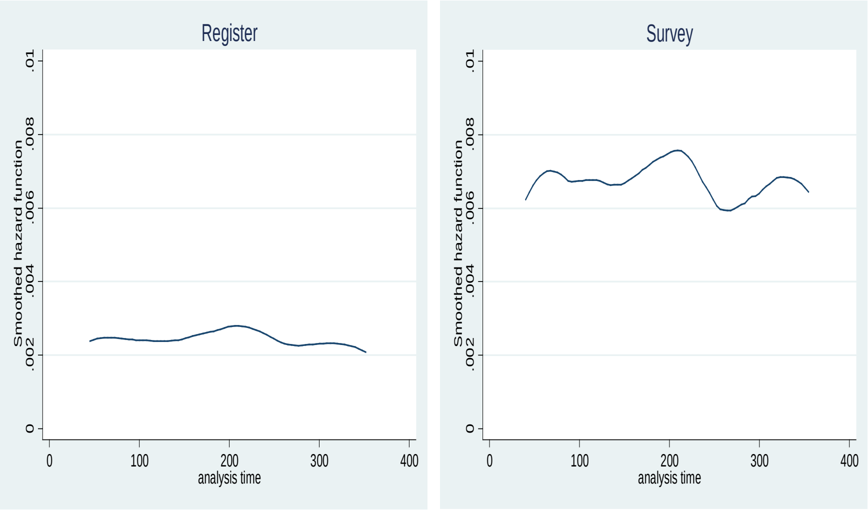
[11] That is we report $\hat{\beta}_i$ instead of $\exp(\hat{\beta}_i)$

roughly constant while that of Weibull shows a positive slope, the Weibull one shows a positive slope, we might say that the former reflects the true function more faithfully.

Table 7.
Bias in the PH Cox model

	BIAS	R.BIAS	RMSE	R.RMSE
Age	-0.084	97.7%	0.090	136.5%
Experience	-1.220	90.5%	1.290	157.2%
Age*Experience	0.035	94.6%	0.037	164.3%

Figure 4.
Cox baseline hazard function for the register and survey data



So, when considering which model to use in the presence of ME in the response variable, we agree with Pyy-Martikainen and Rendtel (2009) in asserting that the flexibility of the Cox model makes it a better choice than a parametric approach. The only exception to this would be where the true baseline hazard function can be properly approximated by a parametric form, as was shown for the case of the AL exponential model. In those cases the restrictive form of a parametric function could be beneficial. However, knowing the true baseline function conditional on a set of explanatory variables represents a major challenge, and it becomes even harder in the presence of ME.

Finally we review the effect of retrospective ME in the response variable on a model from the non-parametric family, a PO logit model. Here, a series of temporal dummies are included in the model in order to specify the baseline logit-hazard function. Each of the dummy variables represents a period of the time-frame, in what is called a piecewise-constant hazards model. This is a reasonable solution when coarse time units relative to the window of observation are used. However, this creates some problems for our case. First, the degrees of freedom are drastically reduced from the inclusion of 395 dummy variables, one for each day. Second, some of the days capture the same number of failures, which produces a problem of perfect multicollinearity in the model. In order to prevent these two problems we used temporal dummies that aggregated failures by weeks.

The results for the PO logit model are shown in Tables 8 and 9. The dummy variables representing the 56 weeks considered in the window of observation are not included in the tables for reasons of space, but they are shown in Figure 5 below as the dots composing the baseline hazard functions. In addition, the sample size is now 89,842 person-day cases in the register, and 50,366 in the survey[12].

Table 8.
PO logit model using register and survey data

	Register		Survey	
Age	**0.086**	(0.038)	0.002	(0.032)
Experience	**1.360**	(0.500)	0.140	(0.420)
Age*Experience	**-0.037**	(0.014)	-0.003	(0.012)
Constant	**-9.410**	(1.450)	**-5.760**	(1.190)
LR Chi^2 (61)	**77.29**		**161.33**	

The outcomes of the two models are again very similar to what was found in the previous EHA specifications. The expected lower precision due to the 56 additional parameters that needed to be estimated to reproduce the baseline hazard function was not as problematic as first thought. In fact, the same SEs as in the AL exponential and the PH Cox were obtained for age, mobility, and the interaction effect when the survey data is used.

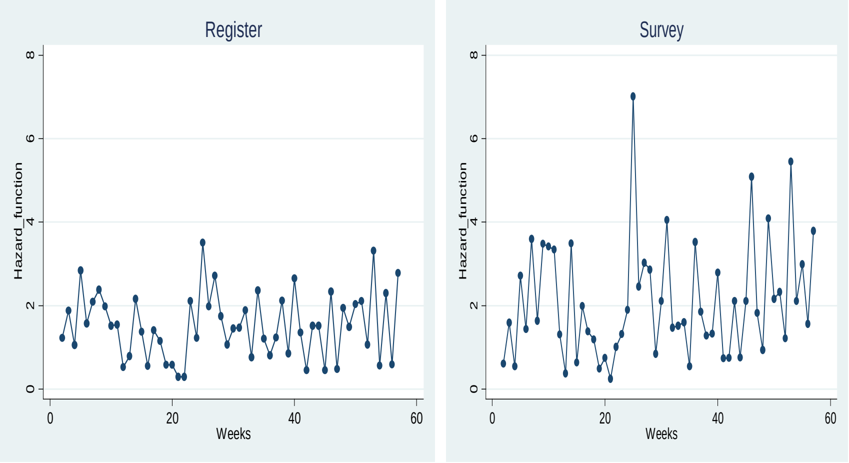
Duo to the effect of aggregating days into weeks, both baseline hazard functions differ from the PH Cox ones, in particular the function for the register data is no longer smooth. Also, unlike in the previous cases where the effect of ME was expressed as a higher baseline function, here what we see is more volatility between time-periods (weeks), resulting in a more jagged baseline function.

[12] The two datasets differ in their sample size because of the transformations required in the specification of EHA models for discrete data; from a dataset capturing one case for each subject to another capturing person-week cases.

Table 9.
Bias in the PO logit model model

	BIAS	R.BIAS	RMSE	R.RMSE
Age	-0.084	97.7%	0.090	136.5%
Experience	-1.220	89.7%	1.290	157.0%
Age*Experience	0.034	91.9%	0.036	157.5%
Constant	3.650	38.8%	3.840	165.4%

Figure 5.
PO baseline hazard function for the register and survey data,***



*The hazard is measured in odds ratios.
**Values for the first week were omitted to prevent multicollinearity in the model.

To sum up, we have seen that the consequences of using retrospective data in EHA models are not negligible and are very similar across different models. Strong attenuation effects were found in all the regression coefficients. In Table 10 we summarize these results by taking the R.BIAS and R.RMSE average in age and mobility[13] for each of the four models studied.

In addition to the strong attenuation effects, illustrated by measures of R.BIAS not lower than 93%, the similarity of the effects across models is striking. None of the models seems to buffer the effects of ME better than the others. In fact, the between model variability in terms of R.BIAS and R.RMSE is 1.8% and 3.1% respectively, while the average effect within models is 94.7% for the former and 153.9% for the latter.

[13] In order to make comparisons possible we excluded the constant term from this analysis since the PH Cox model does not estimate it.

Table 10.
EHA models' performance in the presence of retrospective ME

	R.BIAS	R.RMSE
AL Weibull	97.2%	157.8%
AL exponential	94.3%	154.9%
PH Cox	94.3%	152.7%
PO logit	93.1%	150.4%

The analysis presented so far has focused on assessing the impact of ME in each model separately. However, it could be argued that some EHA models are superior to others given the type of data that we are using here. For example, non-parametric models like the PO logit are recommended when there are fewer time units, whereas here we have seen that the exponential model is a better specification than the Weibull model. In order to assess which model performs better in the presence of ME we need to compare them against a common benchmark. That is, the use of a common benchmark allows us to analyze not only comparisons between the same models using error free and prone to error data, but also comparisons between different models when prone to error data is used.

Here, we use results from the PH Cox model based on register data as that benchmark. There are both empirical and theoretical reasons for this choice. First, the PH Cox model has, along with the AL exponential, the lowest SEs in their regression coefficients (see Tables 4 and 6). Second, since the baseline hazard function is freely estimated it cannot be misspecified. Finally, tied events, a flaw affecting models for continuous time such as the Cox model, are not a major issue here. The window of observation covers 395 days, which makes the time-unit approximately continuous, and rarely do two spells or more end on the same day.

This process to assess the relative impact of ME on the different EHA models implies the assumption that the Cox model using register data produces the true estimates. The comparisons can be formally defined by equations 2 and 4, where $\hat{\beta}_r$ is now substituted by $\hat{\beta}_{r,Cox}$.

Table 11.
EHA models' performance compared to the PH Cox

	R.BIAS	R.RMSE
AL Weibull	97.1%	161.6%
AL exponential	94.2%	154.7%
PO logit	93.0%	150.5%

Results are shown in Table 11 above, where it can be seen that the PO logit performs marginally better than the rest. It is also interesting to note that the AL

Weibull offers the worst performance. These results reinforce the idea advanced when discussing the effect of ME in the baseline function: EHA models that do not make use of a restrictive parametric form seem to do better at buffering the effect of ME in the response variable. This seems to be especially true when the parametric form used is not the most appropriate, as it is shown by the worse performance by the Weibull model than the exponential model.

5. Conclusions

In this paper we have explored the implications of using EHA models where the response variable is affected by ME derived from a retrospective question. Evidence of large attenuation biases in the regression coefficients is found across different EHA models. These findings challenge the common belief that ME in the response variable only affects the SEs of the model's estimates, and also contrast with the existing literature.

Our findings are not in accordance with those of Pyy-Martikainen and Rendtel (2009), which is the most similar study available in the literature since they compare retrospectively reported spells of unemployment with data from a register. The authors found both attenuation and augmentation biases affecting the regression coefficients. We argue that the mix of biases found by Pyy-Martikainen and Rendtel (2009) may be related to the ME being associated with some of the explanatory variables. The ME analysed in our study was non-differential with respect to the two regressors that were used (age and mobility), and the direction of the biases was always towards the null. Moreover, these results are consistent with all the other studies that we are aware of that have assessed the impact of non-differential ME in the response in EHA: Augustin (1999), Dumangane (2007), Korn et al. (2010), and Meier et al. (2003)

Another substantive difference between our results and those from Pyy-Martikainen and Rendtel (2009) is the larger size of the biases found in our study. In Table 10 we showed that the average R.BIAS in the regressors of the Weibull model was 97.2%, whereas the biggest bias found in Pyy-Martikainen and Rendtel (2009) for the same model was 30% of the true estimate. These differences may be due to the use of months as time-units in Pyy-Martikainen and Rendtel (2009), and to the much bigger sample size both in terms of both individuals and window of observations.

One original feature of our study is the assessment of different families of EHA models. We found very similar results for the four types of EHA models that were studied (AL Weibull, AL exponential, PH Cox and PO logit), which implies that the way the response variable capturing life course events is defined (duration data, hazard rates, or person-period cases) is not related to the effect of ME on the model estimates. In fact, when true data is used, when estimates of true data are compared against the ones obtained using survey data, and when the PH Cox model is used as a benchmark, all the models performed similarly for the different comparisons carried out. It was perhaps the PO logit model that showed the biggest differences.

This may be due to the inclusion of temporal dummies which were discretized to capture weeks instead of days.

Using the PH Cox model as a benchmark we found that the exponential model is slightly less affected by ME than the Weibull both in terms of R.BIAS and R.RMSE. We argued that this could be due to the fact that the exponential function matched better the true baseline hazard function than the Weibull did, which led us to think about the possibility that parametric models correctly specified might buffer the effect of ME better than when they are misspecified. Moreover, semi and non-parametric models perform similarly to the AL exponential model, even in terms of R.RMSE. This is an interesting result since parametric models, when correctly specified, are expected to obtain more precise estimates. Moreover, ascertaining the shape of the baseline hazard function is complicated, and in general, it could be expected that parametric forms would perform worse than the results we observed. Hence, when the shape of the baseline hazard function cannot be identified, as it is the case in settings that use durations measured with errors, the use of semi- and non-parametric forms are recommended.

Similarly, we have found that inferences about the time-dependency of the event derived from the PH Cox or the PO logit model are less misleading than those obtained from the AL Weibull model, which wrongly indicated that the probability of making a transition out of unemployment increased with time. This result corroborates Pyy-Martikainen and Rendtel (2009), where the authors posited that freely estimated baseline functions offer better results than those which imposed a parametric form. An exception to this precept might be cases where the parametric form perfectly maps the form of the baseline function. This is what we observed here for the case of the AL exponential. However, in most cases, previous knowledge about the shape of the baseline function conditional on a set of regressors is not available, let alone when the durations are affected by ME. Hence, for the estimation of time-dependencies in the event of duration data prone to ME, we recommend using semi-or non-parametric models.

In this study we have used data derived from a retrospective question on work histories for a period of 395 days, yet our findings may be generalized to other cases where retrospective data is used to derive different life course events. In particular, this would be the this would be the case for events that, because of their relatively low saliency, can be subject to recall errors in the form of mismeasuring, miscounting, and misclassification of spells in the same way as spells of unemployment are. On the other hand, we predict less damaging effects when simpler retrospective questions are used. For example, when the interviewee is asked to report one specific event, such as age at menarche, year of retirement, or time spent since leaving the parental home. Similarly, we could expect better results from reports of events less prone to a social desirability bias than unemployment.

However, more research is necessary since some of the findings presented here need to be tested. "Despite the recognition of the existence of measurement errors in survey-based data on event histories, little is known about their effects on an event history analysis" Pyy-Martikainen and Rendtel (2009, p.140). In particular we would like to extend our study to cases where the ME affecting the response variable in EHA is associated with the explanatory variables. Non-differential ME can generate biases that are not necessarily towards the null, but it is not clear what are the levels of association that could cause a change in the direction of an attenuation bias. Another setting of interest would be the extension of the models seen here to the case of competing risks. This would allow contemplating the impact of retrospective data in EHA in greater detail, in particular the influence of misclassified cases.

References

Augustin, T. (1999). Correcting for Measurement Error in Parametric Duration Models by Quasi-likelihood. Munchen Institut fur Statistik, Working Paper, available at: http://epub.ub.uni-muenchen.de/1546/1/paper_157.pdf [accessed 24 January 2011].

Bound, J.; Brown, Ch., and Mathiowetz, N. (2001). Measurement error in survey data. In: J.J. Heckman and E. Leamer, ed. *Handbook of Econometrics*, 5 (59): 3705-3833. Amsterdan: Elsevier.

Carroll, R.; Ruppert, D.; Stefanski, L., and Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models; a Modern Perspective*. Boca Raton: Chapman and Hall.

Chesher, A.; Dumangane, M., and Smith, R. (2002). Duration Response Measurement Error. *Journal of Econometrics*, 111 169-194.

Cox, D. R. (1975). Partial Likelihood. *Biometrika*, 62 (2): 269-276.

Dumangane, M. (2006). Measurement Error Bias Reduction in Unemployment Durations, Centre for Microdata Methods and Practice, Working Paper 3, available at: http://www.cemmap.ac.uk/wps/cwp0603.pdf [accessed 9 September 2011].

Fuller, W. (1987). *Measurement Error Models*. New York: John Wiley and Sons.

Holt, D.; McDonald, J.W., and Skinner, C.J. (1991). The Effect of Measurement Error on Event History Analysis. In: P. Biemer, ed. *Measurement Error in Surveys*, 32: 665-686. New York: John Wiley.

Hughes, M. (1993). Regression Dilution in the Proportional Hazards Model, *Biometrics*, 40: 1056-1066.

Jäckle, A. (2008). Measurement Error and Data Collection Methods: Effects on Estimates from Event History Data, Institute for Social and Economic Research (ISER), Working Paper 2008-13, available at: https://www.iser.essex.ac.uk/publications/working-papers/iser/2008-13 [accessed 15 May 2013].

Korn, E.; Dodd, L., and Freidlin, B. (2010). Measurement Error in the Timing of Events: Effect on Survival Analyses in Randomized Clinical Trials. *Clinical Trials*, 7(6): 626-633.

Lawless, J. (2003). *Statistical Models and Methods for Lifetime Data*. Hoboken: John Wiley and Sons.

Levine, P. (1993). CPS Contemporaneous and Retrospective Unemployment Compared. *Monthly Labor Review*, 116: 33-39.

Meier, A.; Richardson, B., and Hughes, J. (2003). Discrete Proportional Hazards Models for Mismeasured Outcomes. *Biometrics*, 59 (4): 947-954.

Nakamura, T. (1992). Proportional Hazards Model with Covariate Subject to Measurement Error. *Biometrics*, 48 (3): 829-838.

Novick, M. R. (1966). The Axioms and Principal Results of Classical Test Theory. *Journal of Mathematical Psychology*, 3 (1): 1-18.

Pina-Sánchez, J.; Koskinen, J., and Plewis, I. (2012). Measurement Error in Retrospective Reports of Unemployment, CCSR. Working Paper, available at: http://www.ccsr.ac.uk/publications/working/ [accessed 18 June 2012].

Prentice, R. (1982). Covariate Measurement Errors and Parameter Estimation in a Failure-time Regression Model. *Biometrika*, 69 (2): 331-42.

Pyy-Martikainen, M., and Rendtel, U. (2009). Measurement Errors in Retrospective Reports of Event Histories: A Validation Study with Finnish Register Data. *Survey Research Methods*, 3 (3): 139-155.

Skinner, C., and Humphreys, K. (1999). Weibull Regression for Lifetimes Measured with Error. *Lifetime Data Analysis*, 5: 23-37.

Solga, H. (2001). Longitudinal Survey and the Study of Occupational Mobility: Panel and Retrospective Design in Comparison. *Quality and Quantity*, 35: 291-309.