Identifying biases arising from combining census and administrative data – the fertility of migrants in the Office for National Statistics Longitudinal Study

James Robards, [1,2] Ann Berrington, [1,2,3] and Andrew Hinde [1,3]

[1]Division of Social Statistics and Demography, School of Social Sciences, University of Southampton

[2]ESRC Centre for Population Change, School of Social Sciences, University of Southampton

[3]Southampton Statistical Sciences Research Institute, University of Southampton

james.robards@soton.ac.uk

Abstract

Demographic research is increasingly making use of longitudinal and life history data, given its strong analytical potential. Such data are frequently produced by linking and matching records from multiple sources. Where this is the case, there is the potential for a person's appearance in one source of data to be conditional on an event in another source of data. This can lead to bias in estimating occurrence/exposure rates concerning the event in question, unless the correct exposure can be identified. Achieving the latter requires understanding the reasons governing entry to the data. The Office for National Statistics (ONS) Longitudinal Study (LS) for England and Wales is a 1% sample of the population, constructed by combining data from the census, vital registrations (births and deaths) and the National Health Service Central Register (NHSCR). This paper examines the difficulties in obtaining the correct exposure for rates in complex data sets by studying the fertility of migrants using the ONS LS. Three tests in relation to the fertility of female migrants to England and Wales illustrate the possible association between exposure to risk and subsequent events. The first identifies the ability of the data set to record new migrants, the second is concerned with the mode of entry to the data set and subsequent fertility, and the third illustrates how the recorded fertility of migrants depends upon the way migration is measured.

Keywords: Office for National Statistics Longitudinal Study, migration, National Health Service Central Register, fertility, longitudinal data.

1. Introduction

In demographic research there has been increasing interest in the exploration of associations and causation using longitudinal data sources, especially since the reporting of life course history and events in survey data can be incomplete (Murphy, 2009; Ní Bhrolcháin, Beaujouan and Murphy, 2011). Within longitudinal and life course research, it is becoming more common to combine data from different sources to produce complex data sets (Ford et al., 2009; Lyons et al., 2009). Indeed there is currently discussion as to whether linked administrative data sources could replace the decennial census for England and Wales (Ralphs and Staples, 2012).

Linked data sets can provide detailed information on dates of events and event sequencing. However, complications can occur as a result of the combination of data sources. Appearance in some sources can be related to the events which the researcher wishes to measure using the combined data set, leading to potential bias in occurrence/exposure rates. In this paper we examine the potential biases that arise as a result of associations between life events and capture within a particular source. Using the Office for National Statistics (ONS) Longitudinal Study (LS), the issue is illustrated by studying female LS members' entry into the LS and the degree to which the timing of entry is related to subsequent fertility.

The ONS LS may be suitable for studying migrant fertility because of its large sample of migrants and accurate recording of births from registration data. Few studies have considered the timing of fertility among migrants to England and Wales (Waller, Berrington and Raymer, 2012); research has instead considered the absolute level of migrant fertility (Tromans, Natamba and Jefferies, 2009; Zumpe, Dormon and Jefferies, 2012). In other countries elevated fertility shortly after migration has been identified (Toulemon, 2004).

The ONS LS is composed of data from the 1971-2011 censuses, the National Health Service Central Register (NHSCR) and the vital registration system (births and deaths) (Adelstein, 1976; Hattersley and Creeser, 1995; Blackwell, Lynch, Smith and Goldblatt, 2003). The sample consists of persons born on one of four dates of the year, representing around 1% of the population of England and Wales. Individuals born in England and Wales on one of the four dates become new LS members at birth. For new migrants, entry to the ONS LS is made either by registering with a National Health Service (NHS) General Practitioner (GP) and

reporting that the previous address was overseas, or by being recorded at the decennial census for the first time.

2. Composition of the ONS LS and implications for analysis

2.1 Data combined to produce life histories in the ONS LS

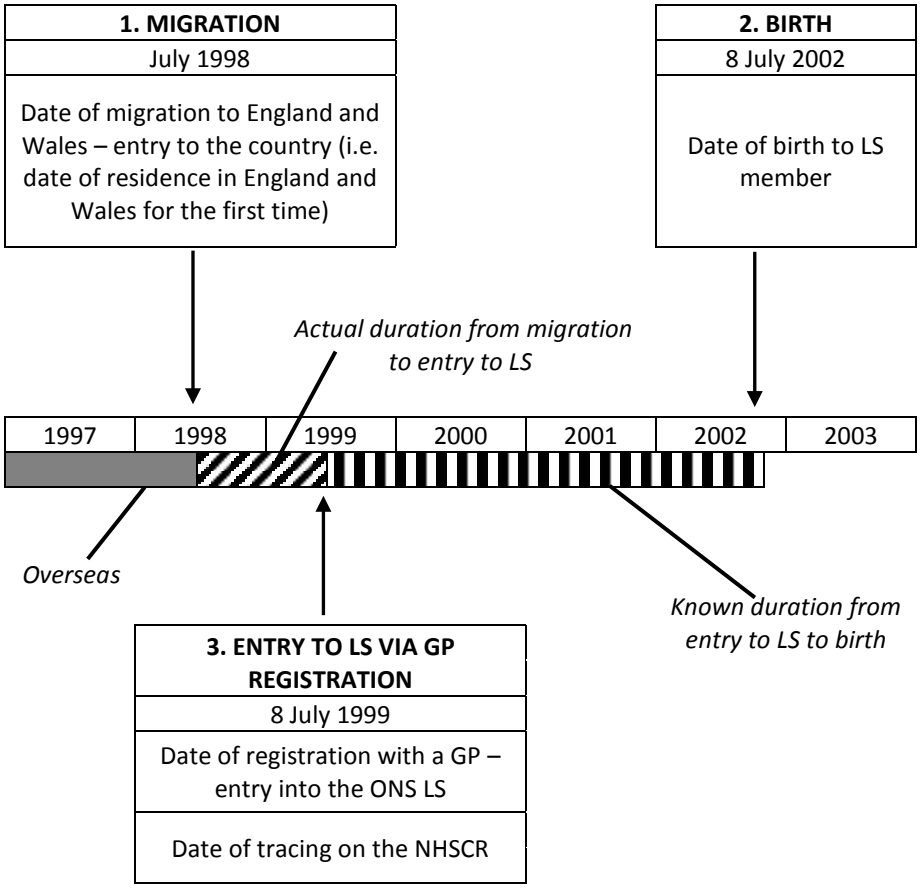
At each census, records for persons born on an LS birth date are selected. From the census, socio-economic and household information is recorded for the LS member. An attempt is then made to trace and match each census record to the NHSCR[1]. This enables the attachment of subsequent events recorded on the NHSCR for these LS members. Births on the four LS dates, and deaths of those born on the four dates, are extracted from vital registration data, as are all births to female LS members. Linkage rates have improved since the early 1990s (Blackwell et al., 2003). Immigrants with one of the four birth dates are captured either through censuses (persons enumerated at each census who were born on one of the four dates and who were not already LS members are added), or by newly registering with a NHS GP. In the latter context, the LS defines 'immigrants' as persons arriving from outside England and Wales, including Scotland, Northern Ireland and the Channel Islands, and '[t]he category of immigrant includes not only those individuals who describe themselves to their general practitioners as such, but also those who, having quoted a previous address abroad, cannot be matched to an existing NHS number' (Hattersley and Creeser, 1995 p.25).

Ideally, new immigrants will register with a GP soon after they arrive in England and Wales and be entered on the NHSCR, so that they will be captured by the LS close to the actual date they arrived. However, as there is no legal requirement for registration with a GP, many new arrivals in fact are not identified until they appear in the census.

Figure 1 shows a hypothetical entry to the data set for a female LS member and the way in which this is captured within the LS. Time point 1 (July 1998) is the date of migration. The date of birth of children born to migrants is known exactly (e.g. this woman gave birth on 8 July 2002 – time point 2). However, there was a time lag of one year before she registered with an NHS GP (on 8 July 1999 - point 3). This time lag is denoted by the period labelled 'Unknown duration from migration to entry to LS'. The duration is 'unknown' because, although the migration was in July 1998, the exact date of migration is not recorded. Upon

registration with a GP, she is recorded in the NHSCR and enters the LS. The recorded exposure to the risk of an event (e.g. death or childbearing) will be the duration from the date of NHSCR registration, and will not include the exposure between the actual date of migration and the date of registration with a GP.

Figure 1. Terminology used to describe the entry of migrants to the ONS LS



2.2 Implications of the construction of the ONS LS for fertility research

Because of the way in which immigrants enter the LS through NHSCR registration it is important to consider the possible association between a woman's entering the exposed-to-risk and the risk of a subsequent birth. If migrant women who were pregnant, or intending to become pregnant, were more likely to register with a GP than other women, there would be an association between entry into the LS exposed-to-risk and the chance of a subsequent birth, and fertility rates of migrants computed using the LS would be inflated. Similarly, a birth to an immigrant mother who was born on one of the four LS dates but who was not in the NHSCR would probably trigger NHSCR registration, again leading to an association between the timing of birth and the timing of entry into the LS exposed-to-risk. The presence and magnitude of such a bias has not hitherto been studied.

3. Research questions and method

3.1 Research questions

We answer three related questions which attempt to quantify potential biases in using the LS for migrant fertility research:

1. How complete is the capture of new migrants to England and Wales in non-census years: what proportion is first identified by the census?
2. Is there evidence of an association between registering with a GP and the timing of a subsequent birth?
3. Is there is a relationship between the mode of entry to the ONS LS (between 1991 and 2001) and fertility after the 2001 census?

The questions are related as they study bias which could be arising from entry of migrants to the ONS LS (question 1), the fertility of migrants who register with a GP (question 2) and the fertility of migrants at the 2001 census (question 3).

3.2 Method

To answer Question 1, the number of new LS female migrant members entering the data set in the five years prior to the census is divided by the number entering at the census for the first time. The sample is composed of LS members who entered via an NHSCR registration in the years 1996-2000 and were at the 2001 census, and those LS members who entered at the 2001 census for the first time. New entrants at the 2001 census are defined as those female migrant LS members who did not enter the data set at any point in the past and have not been resident at a past census. The analysis is by single year of age (based on age at the 2001 census) to allow the identification of age-group trends. The years 1996-2000 for new migrants will be used as they provide enough information to answer the question, while keeping to a minimum the risk that persons arriving in the period before the census may have left England and Wales before 2001. The biggest source of incomplete information in the LS for migrants between 1991 and 2001 was 'unrecorded embarkation' or persons leaving England and Wales without leaving a record of their departure.

To answer Question 2, the duration to the first birth after the date of GP registration (date of migration) is recorded in months for each female migrant into the ONS LS who was captured

by the NHSCR. The number of first births per annum by duration from GP registration is calculated for the periods 1991-2000 and 2001-2006. The sample is composed of migrants who were identified by NHSCR registrations between 1991-2000 or 2001-2006 but who may have subsequently left again.

To answer Question 3, we identify four different types of female migrant entering the LS in the 1991-2001 period (Figure 2), and calculate age-specific and hence total fertility rates for each during the calendar years 2001, 2002, 2003 and 2004. The 2001 census asked all those living in England and Wales on the census night, for details of their place of residence one year before the census. Our four groups are defined as in Figure 2. The sample selection is described for each of the four types of migrant and, as with Question 1, we exclude migrants with a date of birth discrepancy[2].

Figure 2. Four types of female ONS LS migrant at the 2001 census

Group and description	1999												2000												2001											
																												C									
	J	F	M	A	M	J	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D
1. Female LS members resident at the 1991 census and 2001 census. No evidence of any time spent living outside England and Wales.																																				
2. Female LS members reporting that they were located at an address overseas one year before the 2001 census. (Entry to LS through NHSCR registration or at the 2001 census).																																				
3. Female LS members entering between the 1991 census and April 2000 who were at the 2001 census. Not located at an address overseas one year before the 2001 census. (Entry to LS through NHSCR registration between 1991 and April 2000).																																				
4. Female LS members entering between April 2000 and the 2001 census and who were at the 2001 census. Not located at an address overseas one year before the 2001 census. (Entry to LS through NHSCR registration between April 2000 and April 2001).																																				

Key

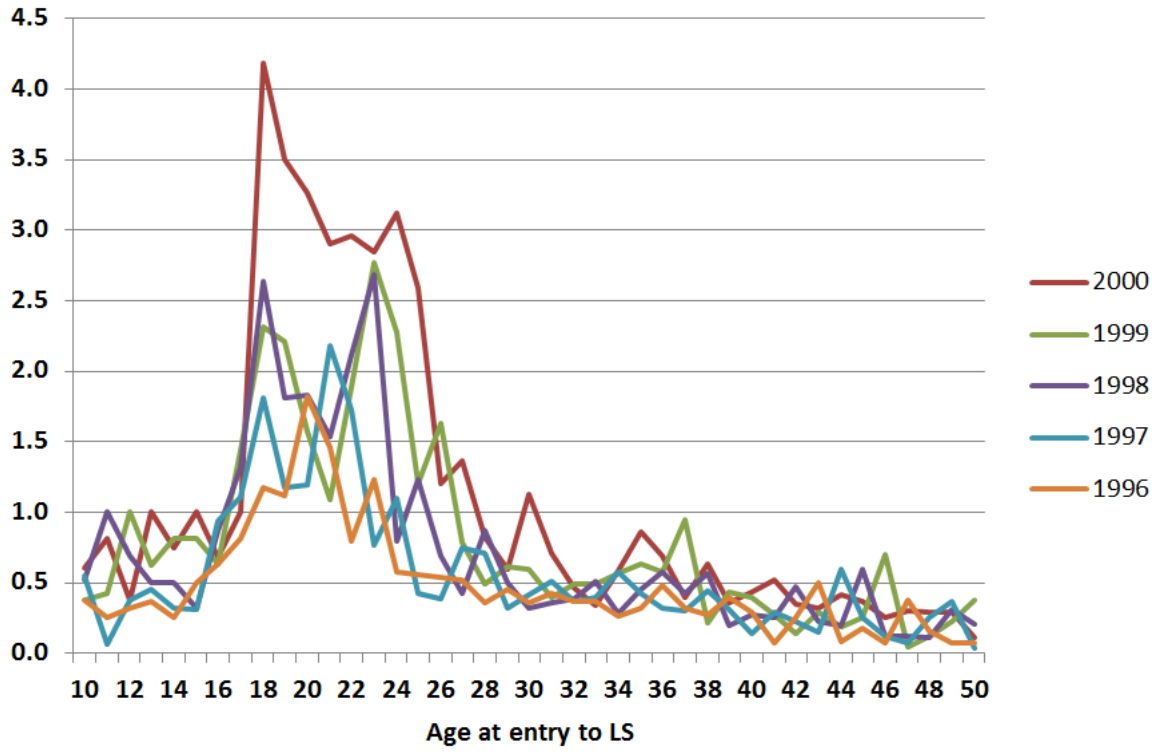
	Continually resident 1991-2001
	Entry to England and Wales and ONS LS within this time period.
	Overseas or not yet entered the ONS LS
C	Denotes Census on 29 April 2001

4. Results

4.1 How complete is the capture of new migrants to England and Wales in non-census years: what proportion is first identified by the census?

We compare the number of female LS members entering through an NHSCR GP registration in the five years before the 2001 census, with the number of female LS members entering at the 2001 census (Figure 3). A ratio below 1.0 indicates that more immigrants were first entered into the LS as a result of being present in the 2001 census, than registering with a GP in one of the years before the census. Among women aged 18-28 years in 2001, more entered the LS through registration with a GP in each of the years preceding the census, than did through being present at the census with an LS date of birth. For other ages, the ratios of entries are below 1.0, and LS members aged over 38 years at 2001 had the lowest ratios, typically below 0.5. We see high proportions of inter-censal capture among LS members entering around age 18 years. The increase in the ratio from age 17 years to age 18 / 19 years coincides with possible demand for reproductive health services. Women in the key reproductive age groups are more likely to register with a GP and enter the LS than women in older and younger age groups.

Figure 3. Ratio of ONS LS joiners in each of the years 1996-2000 to joiners at 2001 census in the same cohort by age at joining

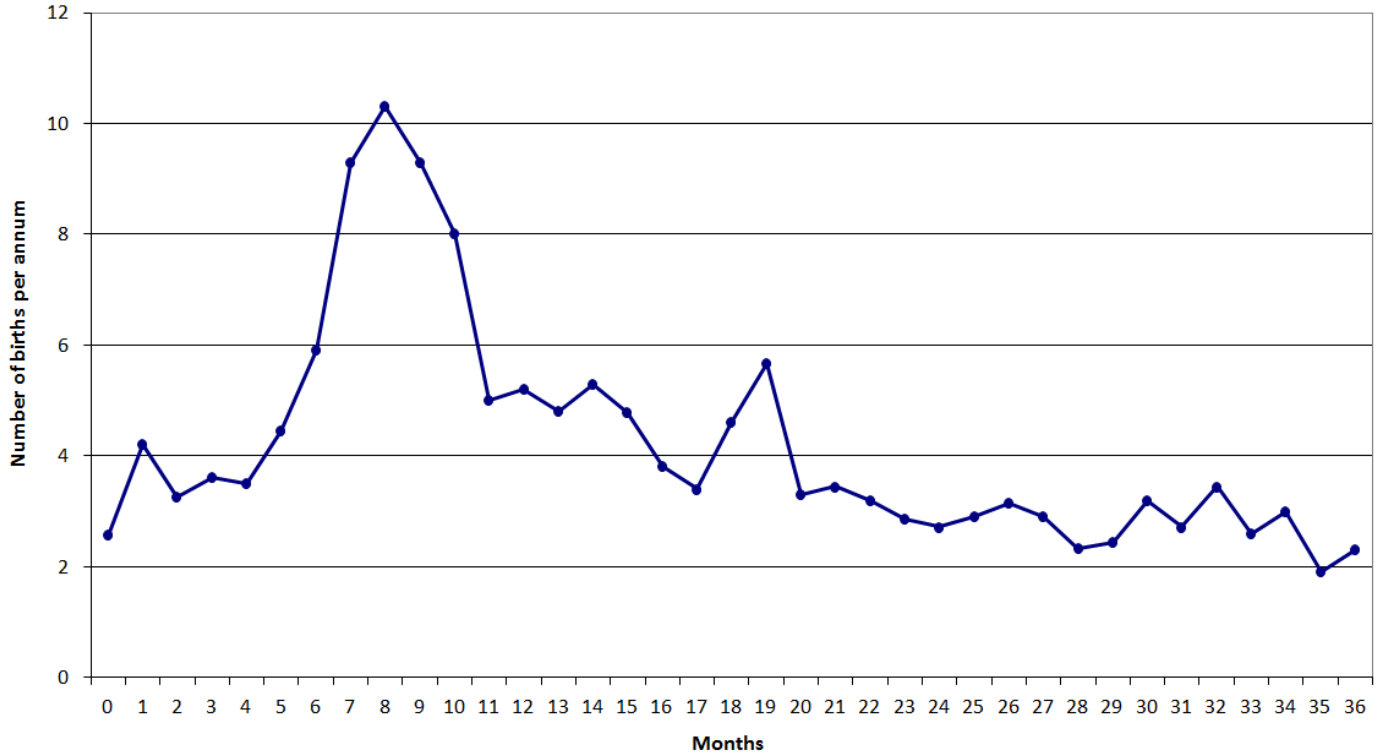


Source: Authors' analyses based on ONS LS.

4.2 Is there evidence of an association between registering with a GP and the timing of a subsequent birth?

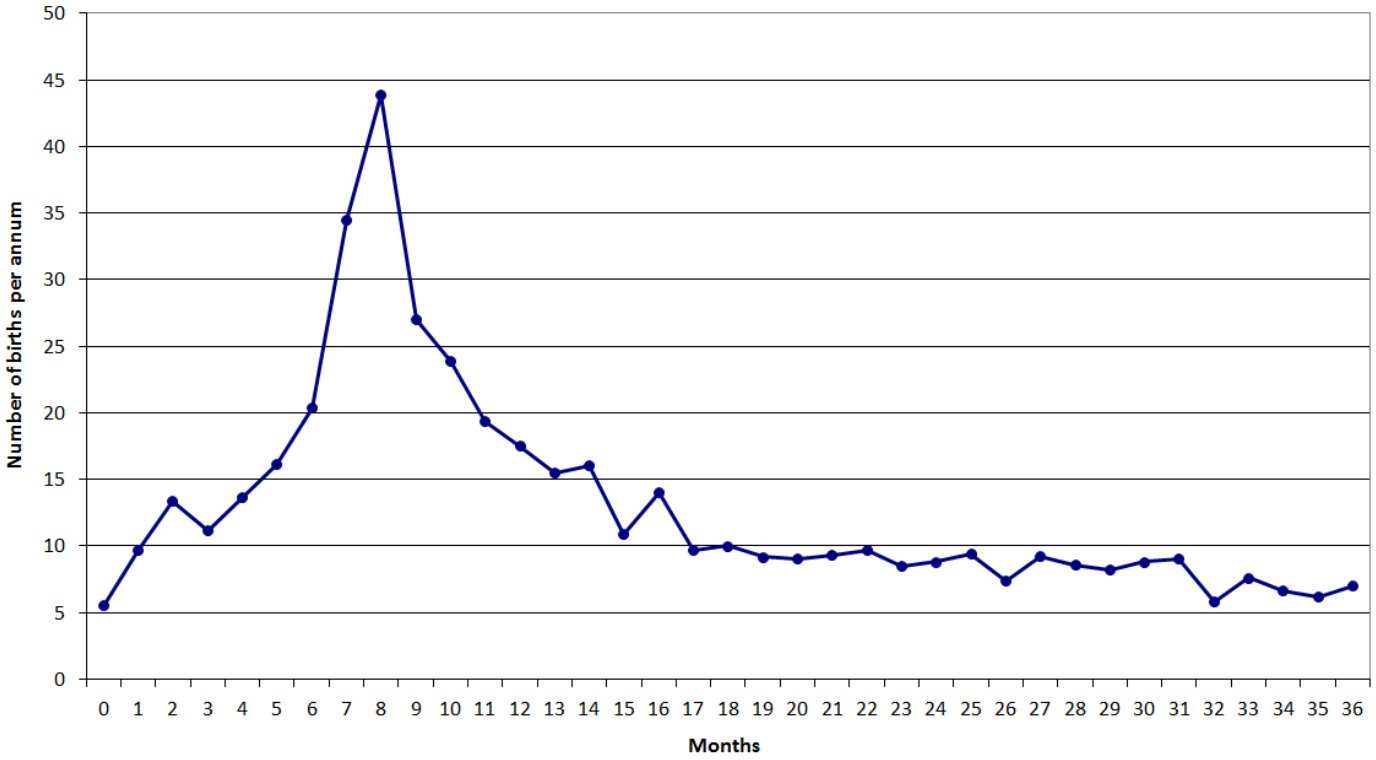
We calculate the number of births which occurred on average at each duration since GP registration in two periods: 1991-2000 (Figure 4) and 2001-2006 (Figure 5). Figure 4 (1991-2000) reveals a peak in first births to new entrants eight months after entry to the LS. Following this, there is a decline in the number of first births to a point 11 months after the registration on NHSCR. A second rise in first births is observed around the 18 months period, after which there is a gradual decline up to a point 36 months after entry. Figure 5 (2001-2006) shows a more pronounced peak in first birth numbers in the eighth month after entry to the LS. The number of first births falls from this peak up to about 18 months after registration, and at longer durations remains roughly constant. The pattern is clear. Migrant women are especially likely to register with a GP around the time they become pregnant, leading to a strong association between registration with a GP (and hence entry into the LS) and a subsequent birth. If the date of registration with a GP is used as a proxy for the date of migration, the fertility of recent migrants to England and Wales calculated using LS data will be overestimated.

Figure 4. Average number of first births per month after registration on the NHSCR (and entry to ONS LS) for the period 1991-2000



Source: Authors analyses based on ONS LS.

Figure 5. Average number of first births per month after registration on the NHSCR (and entry to ONS LS) for the period 2001-2006



Source: Authors analyses based on ONS LS.

4.3 Is there is a relationship between the mode of entry to the ONS LS (between 1991 and 2001) and fertility after the 2001 census?

Table 1 presents total fertility rates (TFRs) for 2001-2004 for the four migrant groups listed in section 3.2. This time period was selected to enable the use of details on LS members recorded at the 2001 census; it allows a period of exposure long enough to provide reliable rates, while being short enough to minimise the impact of attrition.

Group 1, women continuously resident between 1991 and 2001, have a TFR of around 1.5 in all four years. Group 2 is a set of migrants between April 2000 and April 2001. There is a decrease in the TFR from 2001 – the rate is 2.1 in 2001 before dropping to 1.1 in the other years. For each year, the number of women resident remains roughly the same. Compared with the continuously resident women, these recent migrants have higher fertility in 2001, a period 8-21 months after the date of migration, but lower fertility thereafter. Information about these women's date of migration comes from the census, not from GP registration, and is therefore not associated with reproduction. Therefore the higher fertility in 2001 is probably genuine, and reflects the fact that fertility tends to be high among recent immigrants (Toulemon, 2004). The abrupt decline in fertility in 2002 and onwards is for two

reasons: first, the high fertility among these women in 2001 will, for biological reasons, tend to be followed by a period of relatively few births; and, second, it is likely that attrition due to out-migration is more common among recent migrants than among those continuously resident (Group 1). As the LS is poor at capturing emigration in a timely fashion (Hattersley, 1999), the denominators of the fertility rates for years 2002, 2003 and 2004 will include an increasing number of women who have, in fact, departed from England and Wales.

Group 3 are LS members who migrated to England and Wales between 1991 and 2000 and reported at the 2001 census that they were not overseas 12 months before. The group has a high level of fertility compared to the continually resident sample and the recent migrants as of the 2001 census. Among this specific group, the fertility rate is high in 2001 and 2002 before dropping for 2003 and 2004. The difference between this group and Group 2 is that the members of Group 3 have, in 2001, been in England and Wales for some time, so that the rate of attrition between 2001 and 2004 is probably lower. The fact that they have higher fertility than Group 1 is probably associated with their status as recent migrants.

Group 4 has the highest fertility 2001-2004. This group was selected based on registration with a GP in the April 2000-April 2001 period, and reporting at the 2001 census that they were not living overseas 12 months before the 2001 census (April 2000). This indicates that there was a lag between entry to England and Wales (before April 2000) and registration with a GP (between April 2000 and April 2001). It is among these women that the association between registration with a GP and fertility is likely to be strongest, for we know that these women had been resident in England and Wales for some time before entering the LS through registration with a GP. In 2001, their TFR was 4.8. It seems from this that registration takes place in relation to intentions for subsequent fertility. Following the high TFR in 2001, there is a decline to 2.0 in 2002 and then 2.6 and 2.2 in 2003 and 2004 respectively. The timing of registration with a GP was associated with the date of conception of the first child; many of these women will have gone on to have a second child relatively soon after their first, which accounts for the continued high TFR of this group in 2002, 2003 and 2004. There is even some evidence of a two-year spacing between the first and the second child, in that the TFR is higher in 2003 than in either 2002 or 2004.

Table 1. Total fertility rates for 2001-2004 by group at the 2001 census

Sample	2001	2002	2003	2004
Group 1 – continually resident persons, 1991-2001	1.54	1.49	1.56	1.62
Group 2 – LS members at the 2001 census overseas 12 months before	2.12	1.05	1.07	1.07
Group 3 – LS members entering via NHSCR 1991-April 2000, not overseas 12 months before 2001 census	2.20	2.20	1.99	1.89
Group 4 – LS members entering via NHSCR April 2000-April 2001, not overseas 12 months before 2001 census	4.77	2.01	2.63	2.16

Source: Authors analyses based on ONS LS.

5. Conclusions

It is common for life history research to combine data from different sources to produce comprehensive life history data. Linkage of administrative data can often provide high quality, detailed and timely data, but exposure to risk and event likelihood can be related in combined datasets. The ONS LS is one such data set composed from multiple data sources. This research has sought to identify if there is an association between GP registration and subsequent fertility among migrants entering the ONS LS.

The three research questions covered the ability of migrants to be captured in NHS systems and enter the ONS LS, the biases in measuring duration between the entry of female migrants to the data set and a subsequent birth, and the fertility of recent migrants in the period after the 2001 census. It was shown that, during the five years before the 2001 census, the ONS LS generally collected more female migrants through registration with a GP than at the census. However, this does not necessarily mean that the female migrants who enter the LS through NHSCR registration enter at the same time as their migration event, and the analysis strongly suggested an association between the date of registration and the initiation of reproduction. Indeed, it is clear that becoming pregnant leads to a surge in GP registrations and hence in entries to the LS (shown by the peak in the number of births after 7-9 months). This is confirmed by the third piece of analysis presented in this paper, in which fertility during the period 2001-2004 was estimated for four groups of women classified according to their recorded migration history. The group of women who were known to have registered with a GP shortly before 2001 but had migrated earlier (Group 4) exhibited very high fertility in the calendar year 2001. For these women, it seems clear that their

registration with a GP was triggered by their intention to become pregnant shortly, or by the fact that they had become pregnant.

Despite this, our analysis does suggest that the fertility of migrants in the years following their arrival in England and Wales is higher than that of non-migrants. A comparison of non-migrants (Group 1) with migrants (Group 3) suggested that those who migrated between 1991 and 2000 had fertility in the period 2001-2004 up to about 40 per cent higher than women who had been continuously resident in England and Wales between 1991 and 2001. Because Group 3 includes only those migrating between 1991 and April 2000 and registering with a GP, bias arising from the relationship between GP registration and subsequent fertility is excluded from these results. Results for Group 2 are key as this sample has been selected based on the response at the 2001 census indicating that the LS members were overseas 12 months before, and for this group we see an immediately higher TFR for 2001 compared to the subsequent years. Therefore, preliminary evidence suggests a high level of fertility among recent migrants, supporting the findings of Toulemon (2004).

At the 2011 census, for the first time since the 1971 census, a question was asked on the date of migration to the United Kingdom, which should provide a more precise date of migration. To further improve information in the ONS LS on the date of migration, a question asking for the date of migration or first permanent residence at the GP registration stage could be asked and included in the dataset. Such a question would also be beneficial for further linkage of administrative data and improving estimates of international migration to England and Wales.

The findings of this research highlight the potential bias which can be introduced into analyses of the risk of life course events, using data sets assembled from several sources. Combining data sources to produce life history data requires an understanding of the source of each data, and associations between the events in one source and appearance in another source.

Acknowledgements

This research has been facilitated by funding from the Economic and Social Research Council (ESRC) (studentship number ES/G018766/1).

The authors wish to acknowledge user support from Chris Marshall and staff at the Centre for Longitudinal Study Information and User Support (CeLSIUS) and assistance from the Office for National Statistics Longitudinal Study Development Team. The authors would like to thank the two anonymous reviewers for their helpful comments.

The permission of the Office for National Statistics to use the Longitudinal Study is gratefully acknowledged, as is the help provided by staff of the Centre for Longitudinal Study Information & User Support (CeLSIUS). CeLSIUS is supported by the ESRC Census of Population Programme (Award Ref: RES-348-25-0004). The authors alone are responsible for the interpretation of the data. Census output is Crown copyright and is reproduced with the permission of the Controller of Her Majesty's Stationery Office and the Queen's Printer for Scotland. Clearance number LS30106.

References

Adelstein, A. M. (1976) Policies of Office of Population Censuses and Surveys – Philosophy and Constraints, *British Journal of Preventive & Social Medicine*, *30*, 1-10. Retrieved from http://www.jstor.org/stable/25565877

Blackwell. L., Lynch, K., Smith. J. & Goldblatt, P. (2003) *Longitudinal Study 1971-2001: Completeness of Census Linkage, Series LS no. 10.*, Office for National Statistics, London. Retrieved from http://celsius.lshtm.ac.uk/documents/LS10.pdf

Ford, D. V., Jones, K. H., Verplancke, J. P., Lyons, R. A., John, G., Brown, G., Brooks, C. J., Thompson, S., Bodger, O., Couch, T., & Leake, K. (2009) The SAIL Databank: building a national architecture for e-health research and evaluation, *BMC Health Services Research, 9*: 157. doi: 10.1186/1472-6963-9-157

Hattersley, L. (1999) *LS User Guide 18 – International migration data in the Longitudinal Study*, Office for National Statistics, LS Unit, London. Retrieved from http://www.celsius.lshtm.ac.uk/documents/userguide18.pdf

Hattersley, L. & Creeser, R. (1995) *Longitudinal Study 1971-1991 - History, organisation and quality of data*, HMSO, London. Retrieved from http://celsius.lshtm.ac.uk/documents/LS%20No.7%20Hattersley%20&%20Creeser%201995.pdf

Lyons, R. A., Jones, K. H., John, G., Brooks, C. J., Verplancke, J. P., Ford, D. V., Brown, G., & Leake, K. (2009) The SAIL databank: linking multiple health and social care datasets, *BMC Health Services Research, 9*: 3. doi: 10.1186/1472-6947-9-3.

Murphy, M. (2009) Where have all the children gone? Women's reports of more childlessness at older ages than when they were younger in a large-scale continuous household survey in Britain, *Population Studies*, *63*, 115–133. doi: 10.1080/00324720902917238

Ní Bhrolcháin, M., Beaujouan, E. & Murphy, M. (2011) Sources of error in reported childlessness in a continuous British household survey. *Population Studies*, 65, 305-318. doi: 10.1080/00324728.2011.607901

Rahman, N. & Goldring, S. (2006) Factors Associated with household non-response in the Census 2001, *Survey Methodology Bulletin*, *59*, 11-24. Retrieved from http://www.s3ri.soton.ac.uk/isi2007/papers/Paper13.pdf

Ralphs, M. & Staples, V. (2012) *Exploring the Challenges of Using Administrative Data*. Office for National Statistics. Retrieved from http://www.ons.gov.uk/ons/about-ons/what-we-do/programmes---projects/beyond-2011/news/reports-and-publications/index.html

Toulemon, L. (2004) Fertility among immigrant women: new data, a new approach, *Population & Societies 400, 1-4.* Retrieved from http://www.ined.fr/en/resources_documentation/publications/pop_soc/bdd/publication/540/

Tromans, N., Natamba, E., & Jefferies, J. (2009) Have women born outside the UK driven the rise in UK births since 2001? *Population Trends, 136***,** 28-42. Retrieved from http://www.ons.gov.uk/ons/rel/population-trends-rd/population-trends/no--136--summer-2009/have-women-born-outside-the-uk-driven-the-rise-in-uk-births-since-2001-.pdf

Waller, L., Berrington, A., & Raymer, J. (2012) *Understanding recent migrant fertility in the United Kingdom*, Centre for Population Change Working Paper Number 27. Retrieved from http://cpc.geodata.soton.ac.uk/publications/2012_Understanding_recent_migrant_fertility_WP27_Waller_et_al.pdf

Zumpe, J., Dormon, O., & Jefferies, J. (2012) *Childbearing among UK born and non-UK born women living in the UK*, 25 October 2012. Office for National Statistics. Retrieved from http://www.ons.gov.uk/ons/dcp171766_283876.pdf

Endnotes

[1] Change in address 12 months before the census (internal and international migration) has been identified as being related to higher census non-response (Rahman and Goldring, 2006).

[2] A small number of individuals gave LS dates of birth at the 2001 census and were subsequently found to have been registered before the census date on the NHSCR but with a different date of birth.

These people have been excluded from analyses reported for Questions 1 and 3 as, because of their 'date of birth discrepancy', they could not have entered the LS at a date before the census.