Matching lifecourses for causal inference: The effect of early retirement on health using (Swedish) register data.

Nicola Barban, Ingrid Svensson, Francesco Billari, Xavier de Luna, Emma Lundholm

May 1, 2012

Extended abstract prepared for the Lausanne Conference On Sequence Analysis – University of Lausanne, June 6th-8th 2012.
Please do not quote.

1 Introduction

Life expectancy is increasing steadily in developed countries. Governments are seeking to increase the proportion of elderly people in paid employment to balance the ratio of employed people over dependent ones (Vaupel, 2006).

This led to a considerable debate about the timing of retirement and its influence on health: is early retirement good or bad for your health? Several studies have shown that retirement at younger age has adverse effects on health (Westerlund et al., 2010; Hult et al., 2010; Burdorf, 2010). However, selection into retirement may obscure the effect of retirement on health. The individual decision to retire can be influenced by previous health trajectory, marital status and widowhood, social relations with relatives and work career. Moreover, the transition to retirement has become blurred, and the actual range of retirement age has expanded, making the transition "longer and fuzzier" (Han and Moen, 1999). As a result, retirement is becoming more "destandardized" and "deinstitutionalized" (Guillemard and Rein, 1993,?) with people anticipating retirement entering periods of inactivity or reducing their labor supply. Starting from this theoretical framework, we develop a new matching approach to investigate the causal effect of age at retirement on later health outcomes. Standard matching estimators (Rosembaum and Rubin, 1983) based on propensity score pair each treatment participant with a single (or multiple) non-treated participant based on a set of observed characteristics. However, we claim that selection into treatment can be affected by the trajectories of a set of observed characteristics before treatment. For this reason, using sequence analysis with Optimal Matching (OM) (Abbott, 1995), we develop a matching procedure based on the trajectory before treatment. Our method use an extension of nearest neighborhood matching estimator using OM distances. In this way we matched individuals with the most similar trajectory before retirement. We use Swedish register data and we restrict the analysis to the cohorts of people born in Sweden during 1935-1946. Our measure of outcome is the average days of hospitalization 5 years after retirement. We conduct separate analysis for different age at retirement, focusing on retirement between age 60 and 65. Our preliminary results confirm that early retirement is associated with poorer health outcomes. Once we control for selection issues the negative effect of retirement is negligible except for men and women who retire at age 60.

2 Descriptive analysis of early retirement and health

2.1 Data

Data come from the Linnaeus Database which is a longitudinal dataset developed within the Ageing and Living Conditions Research Programme (ALC) at the Centre for Population Studies at Umeå University. The Linnaeus database was created in order to facilitate studies on large-scale population registers concerning i.a. the relationship between socioeconomic conditions and health from an ageing perspective. The Database links nationwide longitudinal data from various registers from Statistics Sweden and the

National Board for Health and Welfare with two regional longitudinal datasets, the Betula investigation (Nilsson *et al*, 1997 and 2004) and the Västerbotten Intervention Programme (VIP) (Norberg *et al,* 2010). Betula is a longitudinal study on aging, memory and dementia and VIP is a longitudinal community intervention programme with the aim of reducing morbidity and mortality from CVD and diabetes. Yearly data such as death causes, hospitalization and socioeconomic conditions are available on an individual level from 1990 to 2006 in the Linnaeus Database as well as links between parents, siblings, children, partners, and in-laws. The database also includes geographical coordinate for the individual's residence and work place. For a more detailed description of the Linnaeus Database, see Malmberg *et al* (2010).

In this study we focus on individuals born between 1935 and 1946 that lived in Sweden during 1990. In the Linnaeus Database there is yearly information about these individuals from 1990 until 2006 given that they live in the country during at least one year during the period.

Besides the information about country of origin, year and place of birth for each individual in the study, data on different sources of income are included. In particular, salary, self-employment income, unemployment benefit, sick-leave benefits, occupational pension, old-age pension, disability pension. Moreover, the register contains detailed information about marital status and educational level. These data are linked to the Inpatient Register where it is possible to collect information on days in hospital and diagnosis at each enrollment in a hospital. Last, data are linked to the Cause of Death Register where the year and cause of death fare available.

There is a linkage to the individual's partner for those having one, and also yearly information of the partner's country of origin, education and income. For the individual there are likewise links to own and partner's biological or adopted children, parents and siblings. There are geographic coordinates for the residence for every person in the dataset, making it possible to get a picture of the family social network the person is living in by looking at the distance between the individual and her or his closest family.

2.2 Treatment: early retirement

Since the interest of this study is to look at the effect of the age at retirement on health outcomes, the definition of the time of retirement is essential. We have used the year of retirement as the first year the annual income from pension exceeds the income from annual labour earnings. In the income from labour earnings, we have included transfers connected to unemployment and labour market measures. These kinds of transfers are not given to individuals after the age of 65. This way to define retirement is in concordance to that of de Luna et al. (2010). Even though the transition to retirement has become blurred, and the actual range of retirement age has expanded, making the transition "longer and fuzzier" (Kohli and Rein 1991; Han and Moen,1999), we have defined retirement as an absorbing state so that an individual, once retired, is assumed to be retired for good.

2.3 Measure of health outcome

One way to see if the age of retirement affects health is to look at the number of days in hospital after retirement. As shown in Figure 1, early retirement is associated with lower health outcome both before and after retirement. Also, the two groups have specific trajectories, with an increase of number of days in hospital in the years before retirement.

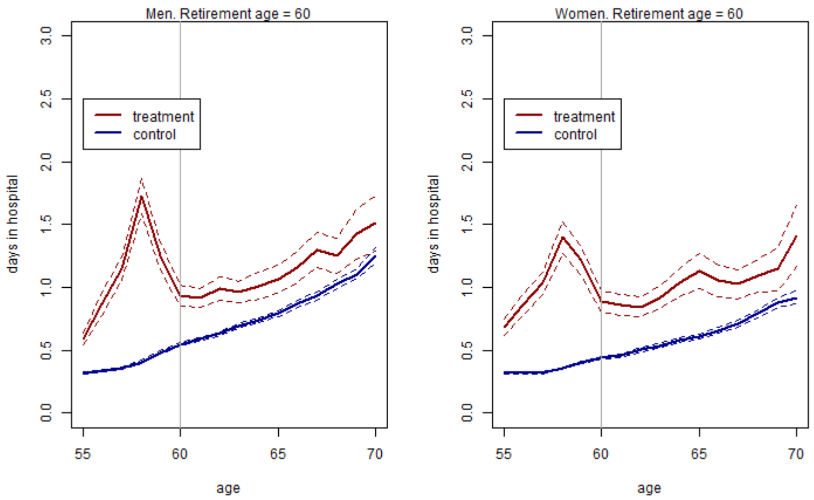
In the Inpatient Register there is information on number of days in hospital for those who have been enrolled in a Swedish hospital. Since part of the cohorts die and others leave Sweden during the observed period, we have used the mean number of days in hospital after the year of retirement.

3 Designing the study by matching life trajectories for causal inference

3.1 Model and parameter of interest

A widely used framework for causal reasoning is the potential outcome framework originally due to Neyman (1923) and developed to observational studies by Rubin (1973). Consider a binary treatment variable T (retiring at a given age a or not) and an outcome of interest (in our study a measure of

Figure 1: Number of days in hospital before and after retirement. Age at retirement 60, Men and Women

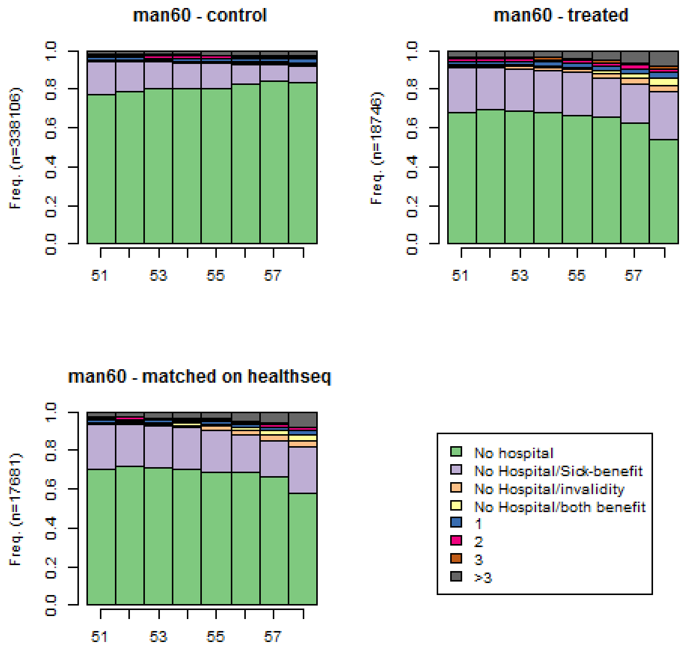


health after treatment, see above). Then, two outcome variables, called potential outcomes, are defined for each unit in the study, the outcome under treatment (unit retires at age a), $Y(0)$, and the outcome without treatment (unit does not retire at age a), $Y(1)$. The difference $Y(1) - Y(0)$ is then interpreted as the causal effect at the unit level. This effect is not identified since for each unit either $Y(0)$ or $Y(1)$ is unobserved, because any unit in the study will be either treated or not. On the other hand, under certain conditions average causal effects, $E(Y(1)-Y(0))$, may be identified, where the expectation is taken over a well defined population of interest. Our study will focus on the average causal effect of early retirement for those actually retiring early, i.e. $\tau = E(Y(1)-Y(0) \mid T = 1)$. This parameter has several advantages: it is relevant for the individuals that have actually retired early (answering the question: what would have been their average health would they have retired later?); it is more realistically identifiable (needs less conditions for identification than, e.g., the unconditional average causal effect $E(Y(1) - Y(0)$, i.e. averaging over the population from which the sample is drawn); and it is easier to estimate in situations where many controls are available compared to the number of treated individuals as it is the case in our study. The parameter τ is identified under the following conditions. First no interference are allowed, that is the potential outcomes of any unit in the study is not affected by the retirement decision of other units. This condition called Stable Unit Value Assumption (e.g. Rubin, 1991) seems reasonable in our case, at least for individuals which are not partners and we therefore make separate analysis for women and men. Also for indentification purposes, we need to have access to a collection of background information $\mathbf{X}$, which is not affected by treatment T, and such that $(Y(0), Y(1), T, \mathbf{X})$ has a joint distribution such that $Y(0) \perp T \mid \mathbf{X}$ and $\Pr(T = 0 \mid \mathbf{X}) > 0$. The latter condition is called strong ignorability, here, of the decision to retire early, and it requires, for instance, that all background information affecting both $Y(0)$ and T is observed. This two latter conditions ensure that we are able to design a study by conditioning on the necessary background information in order to obtain an estimator of the causal effect τ. The conditional independence statement is often called unconfoundedness assumption since the background information $\mathbf{X}$ needed in the conditioning set to obtain independence is called confounding information. The unconfoundedness assumption is a strong condition and conclusion of observational studies must be interpreted with care. We have the oportunity in this study to have access to rich background information through socio-economics and health registers and we show in this paper how this information can be exploited using sequence analysis techniques.

3.2 Classical matching designs

Matching estimators are widely used for the nonparametric evaluation of average causal effects; see, e.g., Dehejia and Wahba (1999); Imbens (2004); Imbens and Wooldridge (2008). Matching estimators are

Figure 2: Distribution plot health trajectories before and after matching. Men, age at retirement 60



defined with respect to a matching design.

Assume that we have a random sample of N_1 units from the treated population (have retired early at age a) and a random sample of N_0 units from the control population (have not yet retired at age a), with $N = N_0 + N_1$. Then, an exact matched design is constructed by, for each treated unit $i = 1, \ldots, N_1$, picking (with or without replacement) M control units $j = 1, \ldots, M$ such that for all j, $\mathbf{X}_j = \mathbf{X}_i$. Exact matching where the information set $\mathbf{X}$ is equal for the treated and the matched controls is seldom possible in practice due to the dimensionality of $\mathbf{X}$ and/or to the fact that some of the variables in $\mathbf{X}$ are continous valued. When exact matching is not feasable one need to rely on a measure of similarity (in $\mathbf{X}$) between a treated and a control. Thus, when $\mathbf{X}$ is a vector of variables several matched designes have been proposed in the literature, including using the Mahalonobis distance (Mahalanobis matching), the propensity score (Rosenbaum and Rubin, 1983), the prognostic score (Hansen, 2008).

Available matching algorithms are designed for the situation that $\mathbf{X}$ is a vector of variables without the possibility to take into account information carried by life trajectories (discrete time stochastic processes).

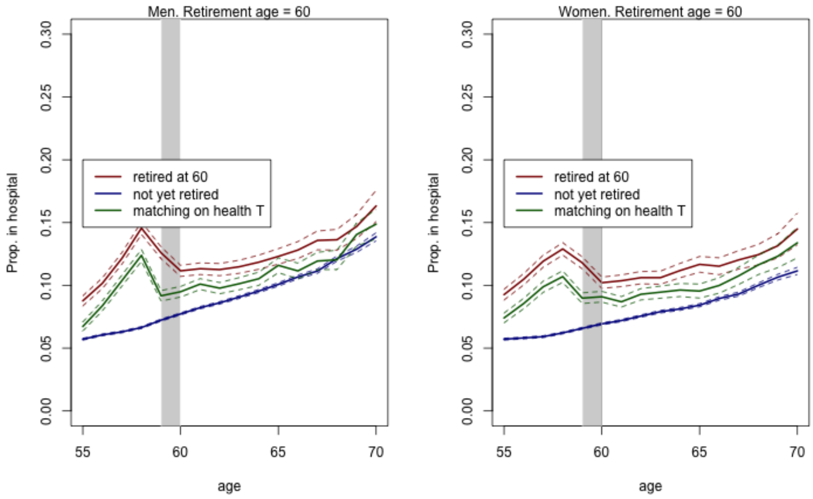
3.3 Using sequence analysis to match on life course trajectories

In this paper, we apply Optimal Matching in order to identify individuals with similar pre-retirement trajectory. For each retired individual at age a, we select a control individual (who retires at age $z > a$) with the most similar pre-retirement trajectory. Sequence dissimilarities are based on the categorical time series between age 55 and age at retirement. The state-space is composed by the combination of days in hospital (0 days, 1, 2 , 3+ days in hospital for each year of observation); sick-leave benefits, and other health benefits received during each year of observation.

Individuals are matched exactly by birth cohorts (to avoid specific birth cohort characteristics, e.g. changes in policy), educational level (low, medium, high) and marital status at the year of retirement.

We adopt alternative matching strategies based on sequence analysis and propensity scores and we compare the results.

Figure 3: Number of days in hospital before and after retirement. Retired group, control group and matched group. Age at retirement 60, Men and Women. Matching on health trajectories.



4 Preliminary results

Our preliminary results indicate that matching on trajectories can be used in a causal analysis framework. As shown in Figure 3, The matched group represents individuals who have not retired yet at age 60 but with "similar" health trajectories with the retired group. Our preliminary results show that the two groups have similar trajectories before and after retirement. This suggests that sequence analysis is able to capture specific group characteristics based only on the life course. Unlike other matching approaches that control for characteristics fixed in time, our strategy take into account the entire lifecourse. In the complete paper, we compare this results with the one obtained by using propensity score matching estimators.

References

Abbott, A., 1995. Sequence analysis: new methods for old ideas. *Annual Reviews in Sociology*, 21(1):93–113.

Burdorf, A., 2010. Is early retirement good for your health? *BMJ (Clinical research ed)*.

de Luna, X., A. Stenberg, and O. Westerlund, 2010. Can adult education delay retirement from the labour market? *Working Paper Series*.

Dehejia, R. and S. Wahba, 1999. Causal effects in, nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062.

Guillemard, A.-M. and M. Rein, 1993. Comparative Patterns of Retirement: Recent Trends in Developed Societies. *Annual Reviews in Sociology*, 19:469–503.

Han, S. and P. Moen, 1999. Clocking out: Temporal patterning of retirement. *American journal of sociology*.

Hult, C., M. Stattin, U. Janlert, and B. Järvholm, 2010. Timing of retirement and mortality - A cohort study of Swedish construction workers. *Social Science & Medicine*, 70(10):1480–1486.

Imbens, G. M. and J. M. Wooldridge, 2008. Recent Developments in the Econometrics of Program Evaluation. page 305.

Imbens, G. W., 2004. Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review. *The Review of Economics and Statistics*, 86(1):4–29.

Rosembaum, P. and D. Rubin, 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Vaupel, J. W., 2006. Redistributing Work in Aging Europe. *Science*, 312(5782):1911–1913.

Westerlund, H., J. Vahtera, J. E. Ferrie, A. Singh-Manoux, J. Pentti, M. Melchior, C. Leineweber, M. Jokela, J. Siegrist, M. Goldberg, M. Zins, and M. Kivimäki, 2010. Effect of retirement on major chronic conditions and fatigue: French GAZEL occupational cohort study. *BMJ (Clinical research ed)*, 341:c6149.