CESifo Working Papers

The Parental Co-Immunization Hypothesis

Miguel Portela
Paul Schweinzer

CESifo
Center for Economic Studies & Ifo Institute

The Parental Co-Immunization Hypothesis

Abstract

We attempt to answer a simple empirical question: does having children make a parent live longer? The hypothesis we offer is that a parent's immune system is refreshed by a child's infections at a time when their own protection starts wearing thin. With the boosted immune system, the parent has a better chance to fend off whatever infections might strike when old and weak. Thus, parenthood is rewarded in individual terms. Using the Office for National Statistics Longitudinal Study (ONS-LS) data set following one percent of the population of England and Wales along four census waves 1971, 1981, 1991, and 2001, we are unable to reject this hypothesis. By contrast, we find in our key result that women with children have a roughly 8% higher survival probability than women without children.

Miguel Portela
Department of Economics & NIPE
University of Minho
Campus de Gualtar
Portugal – 4710-057 Braga
mangelo@eeg.uminho.pt

Paul Schweinzer
Department of Economics
University of York
Heslington
United Kingdom – York YO10 5DD
paul.schweinzer@york.ac.uk

1   Introduction

Children are born without significant defences against a large number of infections or diseases.[1] They acquire this immunity through exposure and consequently are often and repeatedly ill in their first years of life. Communal nursery, kindergarten and school attendance ensure that few major infections are missed. This protects the children in later life. The involved viruses and bacteria adopt and mutate over time, though, and therefore this early-life immunization does not last forever.

Our parental co-immunization hypothesis is that a fresh immunization at the adult stage is obtained through having or being around children. Thus, given that parents survive the initial exposure, they are better equipped for older age and live longer. More precisely, exposure to the pathogens which cause a child to acquire its own immunization have a secondary effect on the parents boosting the adults' own immune systems. We test this hypothesis on females only and subsequently verify that the male population does not exhibit significantly differing characteristics. As a sanity check, we show that a similar effect is significant for childless individuals in child caring and teaching professions which brings them in frequent and direct contact with young children.

Evolutionary speaking, it should be beneficial for the gene-pool to duplicate more than once with, perhaps, more than one partner, i.e., for any person to produce more than one child. Using the ONS-LS data set, we cannot confirm an effect of the number of children on the probability of dying from infectious diseases: having one or multiple children does not make a statistically significant difference. The only effect is between living together with children or not. But since a longer life span, in turn, also increases reproductive opportunities, re-immunization may benefit parents genetically and help illuminate the threshold to Medawar's 'Selection Shadow.'[2] This effect may be pronounced through the prolonged life expectancy of recent generations.

As discussed below, happiness, stress levels, etc may all have an impact on longevity. Using a statistical approach, we cannot fully disentangle our hypothesis from alternative and competing causal (behavioral) explanations: Specifically, we cannot rule out that the effects we find are in part due to different life styles individuals may adopt when having children. As we can control, however, for a parent's marital state, we hope that *some* behavioral effects which may affect a parent's risk of acquiring infections are already addressed by this variable. Moreover, the correlation we document for infectious diseases varies strongly across other diseases and alternative causes of death. Finally, as we can confirm the positive effect on life expectancy in individuals working with younger children (e.g., teachers) who do not have children themselves,

---

[1] Tollånes et al. (2008) show that babies born by Caesarean section have a 50% increased risk of developing asthma compared to babies born naturally. Emergency Caesarean sections increase the risk even further. This is probably because a Cesarean changes or postpones the bacterial colonization of a baby's stomach which is necessary for development of an immune system response. During the course of a vaginal birth, babies obtain their mother's vaginal, intestinal and perianal bacteria.

[2] This evolutionary beneficial immunological explanation is in line with the 'grandmother effect' as proposed by Lahdenperä et al. (2004). Kuningas et al. (2011) is a recent study into which genes control the presumed trade-off between fertility and lifespan.

we are confident that the regularity we document is not entirely spurious.

Literature

It is well documented that happier people live longer. Among the theories competing to explain this fact, prominent places are taken by individual wealth, marriage (e.g., Kaplan et al. (1994), Ikeda et al. (2007), or Wood et al. (2009)), or sports (e.g., Lahdenperä et al. (2004)). The purpose of our paper is to study the contribution of children.

More generally, our parental co-immunization hypothesis adds to (and contrasts in terms of prediction with) the three classic theories of the evolution of aging (and its relation to fertility): the *Mutation Accumulation Theory* due to Medawar (1946), the *Antagonistic Pleiotropy Theory* due to Williams (1957), and the *Disposable Soma Theory* due to Kirkwood (1977).[3] Each of these theories suggests a particular mechanism trading off longevity against reproduction but they all predict that, over a genotype's life span, there is a genetic trade-off between early reproduction and late fitness. Therefore, these theories usually associate an increased number of children with decreased lifespan because of a postulated balancing between the resources available for reproduction on the one hand and longevity on the other.[4]

A number of recent studies offer empirical evidence to substantiate this basic prediction of the classic theories. A recent analysis of the association between the number of children and the mortality of mothers is Dior et al. (2013). They concentrate on the effect the number of children have on individual causes of death while we study whether or not having children at all has an effect on the probability of dying of infections. In this setup, they observe higher mortality rates for mothers than for women without children: qualitatively, the opposite of our result. Similar positive associations between fertility and mortality are reported by Tabatabaie et al. (2011) on a cohort of Ashkenazi Jews. Doblhammer (2000) reports increased mortality risk for Austrian and Welsh early mothers while Helle et al. (2005) document no significant effects in a population of Sami women.

There is, however, also a number of recent studies that report effects which are at least partially compatible with our results.[5] For instance, Wang et al. (2013) investigate the genetic associations between post-reproductive lifespan and children ever born in the Framingham Heart Study data set. In this sample, they find a U-shaped impact of number of children on mortality. Having one or two children reduces maternal mortality while having more than a few increases maternal mortality. McArdle et al. (2009) employ genealogical data from an Amish community in Pennsylvania to document that high parity among men and later menopause among women may be markers for increased life span. Müller et al. (2002) study the relation

---

[3] For a beautiful introduction to the theory of aging see Fabian and Flatt (2011). A more detailed overview is http://en.wikipedia.org/wiki/Senescence#Evolutionary_theories. We should also point out that *Evolution of Lifespan* views such as Stearns (1992) are usually seen to (partially) counterbalance the lifespan shortening effects of the classical theories and are thus more in line with our results.

[4] See, for instance, Partridge and Barton (1993), Kirkwood and Austad (2000), or Flatt and Promislow (2007).

[5] Hurt et al. (2006) analyze the role that statistical and methodological errors can play in explaining some of this apparent empirical inconsistency.

between fertility and post-menopausal longevity in a historical French-Canadian sample of 1635 women and find that increased fertility is linked to increased rather than decreased post-reproductive survival. Specifically, they relate mortality to the age of the youngest child.

There is a group of recent studies which are a bit further removed from the central question behind our work but still impinge on some of its aspects. Helle et al. (2002), Hayward et al. (2013), and Helle and Lummaa (2013) are key results on the influence of early life circumstances on mortality. Berg et al. (2012) analyze both the impact of the economic conditions at births and the years leading to puberty on the individual fertility rate and, subsequently, examine the protective effect of fertility on mortality. They find that while women's health suffers from fertility during their reproductive period, fertility has a large, protective causal effect on female mortality in post-reproductive years. Gagnon et al. (2008) test the trade-off between fertility and longevity in frontier populations. They test the hypothesis whether increased reproduction reduces the chances for survival in old age and find a negative influence of parity and a positive influence of age at last child on postreproductive survival.[6]

This paper demonstrates the existence of an empirical regularity in the sample followed by the England and Wales census over four decades. We offer a hypothesis which explains these observations but we can only superficially comment on the medical plausibility of the proposed transmission mechanism itself. What we are certain of is that our view is not uncontroversial. For instance, Graham et al. (2007, p713) state that "the incidence of viral respiratory illness peaks in infancy and early childhood and steadily decreases with age because of changes in patterns of exposure and age-related acquisition of specific immunity to an increasing number of virus types encountered over time."[7]

2 Survival analysis

2.1 Single risk models

Survival analysis is the study of duration data. In the following discussion we assume that time is running continuously, and we therefore describe duration by a continuous random variable, denoted by T. In our setup, the duration data has information on the time from a well–defined starting point until the event of interest occurs, or until the end of the data collection process. Death is most adequately modeled as the probability of dying given that the person survived until that time, so that time until failure (duration or survival) models are most appropriate. Furthermore, we ignore the role of possible unobserved heterogeneity.

There are three functions critical to the analysis of time: *(i)* the density function, $f(t)$; *(ii)* the cumulative density function $F(t)$; and *(iii)* the survival function, $S(t) = 1 - F(t)$.[8] Knowing one of these functions means, at least in principle, than we can derive the other two.

---

[6] A recent a landmark paper on general ageing research is López-Otín et al. (2013).

[7] A recent survey of the epidemiology of viral respiratory infections is, for instance, Monto (2002).

[8] One should not see the function $f(t)$ as a probability since it can take values bigger than 1. We can see it as a function that describes how probability is distributed over the domain of T.

Each analyzed subject is characterized by *(i)* survival time, or spell, *(ii)* status at the end of the survival time (event occurrence or censored), and, in some cases, *(iii)* the study group (s)he is in. In our case the groups are alive, death by infection and death for other reason (later, other reasons are split in different causes of death).[9]

The hazard function, λ, is a key concept in survival analysis and is defined as the rate of failure at a point in time t, given survival until that time

$$\lambda(t, x) = \lim_{dt \to 0} \frac{\Pr(t \leq T < t + dt | T \geq t, x)}{dt}, \tag{1}$$

where T denotes the random variable length of stay, measured in continuous time, and x is a vector of explanatory variables consisting of individual characteristics. Among the possible alternative interactions between T and x proposed, the most popular in the length of stay literature is the proportional hazards (PH) specification.

The most commonly used semiparametric duration model is the Cox PH model. Cox suggest a likelihood procedure (partial likelihood) to estimate the relationship between the hazard rate and explanatory variables in the following general proportional hazards model

$$\text{Cox: } \lambda(t, x, \beta) = \lambda_0(t) \exp(x'\beta). \tag{2}$$

In Cox's model we do not need to make assumptions about the functional form of the baseline hazard function, $\lambda_0(t)$. As a result, and as Lee and Wang (2003) put it, "the ratio of the risk of dying of two individuals is the same no matter how long they have survived." The model, defined as

$$\log \lambda_i(t) = \beta_0(t) + \beta_1 x_{1i} + \ldots + \beta_r x_{ri}, \tag{3}$$

leaves the baseline hazard function $\beta_0(t) = \log \lambda_0(t)$ unspecified. This way, the model is semiparametric. This results from the fact that the baseline hazard can take any form, and that the covariates enter the model linearly. The baseline hazard does not depend on covariates, but only on time, and the covariates are time-constant. As a result we have the proportional hazard assumption.

The fact that the Cox model does not estimate the baseline hazard, $\lambda_0(t)$, is both an advantage and a disadvantage. For two observations i and j, the hazard ratio

$$\begin{aligned} \frac{\lambda_i(t)}{\lambda_j(t)} &= \frac{\lambda_0(t) \exp(\beta_1 x_{1i} + \ldots + \beta_r x_{ri})}{\lambda_0(t) \exp(\beta_1 x_{1j} + \ldots + \beta_r x_{rj})} \\ &= \exp\left[\sum_{l=1}^{r} \beta_l (x_{li} - x_{lj})\right] \end{aligned} \tag{4}$$

---

[9] Censoring means that the total survival time for that subject cannot be accurately determined. This could happen because the subject drops out, is lost to follow-up, or because the study ends before the subject experiences the event of interest. In this case the individual survived at least until the end of the study, which means that there is no knowledge of what happened thereafter. As such we face right censoring as the individual is removed from the study before the event occurs.

is independent of time t. This implies that the Cox model is a proportional hazards model. However, this property comes at a cost. The efficiency of the estimates is reduced as this approach discards information regarding actual failure times and uses only their rank order.

Alternatively, the hazard function can be restricted to a multiplicative form and defined as

$$\lambda(t, x, \beta, \theta) = \lambda_0(t, \theta) \phi(x, \beta), \tag{5}$$

where λ_0 is the baseline hazard function and depends both on time, and, compared to Cox's model, on an additional parameter, θ. This parameter is a vector of auxiliary parameters characterizing the distribution of T. β is a vector of unknown coefficients associated with x and $\phi(x, \beta)$ is a proportionality factor which does not depend on duration. $\phi(x, \beta)$ is a non-negative function of the covariates. With proportional hazards the effects of the regressors on the conditional probability of failure do not depend on duration. The baseline hazard function summarizes the pattern of duration dependence, and alternative specifications of the baseline function lead to different hazard functions.

The impact of the covariates on the hazard function can also be estimated using parametric techniques, which require potentially restrictive assumptions regarding the functional form of the baseline hazard function. Parametric models are useful because of their predictive and extrapolative capabilities, as well as the possibility for quantification of the effects of covariates. Within parametric models, the Exponential and Weibull models are common solutions. These models are defined as

$$\text{Exponential: } \lambda(t, x, \beta) = \lambda_0 \exp(x'\beta), \tag{6}$$

and

$$\text{Weibull: } \lambda(t, x, \beta, \alpha) = \alpha t^{\alpha-1} \exp(x'\beta), \tag{7}$$

respectively. The exponential model assumes a constant baseline hazard for each patient, while the baseline hazard for the Weibull model is strictly increasing or decreasing depending on the value of α.[10] "The exponential distribution is the only one that has the lack of memory property that the distribution of the residual lifetime, after truncation, is the same as the original distribution" (Hougaard, 2000).

In a regression analysis environment, a parametric model based on the exponential distribution may be written as

$$\log \lambda_i(t) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_r x_{ri}. \tag{8}$$

The constant β_0 can be interpreted as some form of log–baseline hazard. For the Weibull model this regression becomes

$$
\begin{aligned}
\log \lambda_i(t) &= ln(\alpha) + (\alpha - 1)ln(t) + \beta_1 x_{1i} + \ldots + \beta_r x_{ri} \\
&= \beta_0 + (\alpha - 1)ln(t) + \beta_1 x_{1i} + \ldots + \beta_r x_{ri}.
\end{aligned}
\tag{9}
$$

---

[10] The parameter α assumes only positive values. If $\alpha > 1$ then the hazard function increases monotonically; if $\alpha < 1$ then it decreases monotonically; and if $\alpha = 1$ the model collapses to the exponential case.

2.2 Competing risks models

Within our setup, we can clearly distinguish, for example, three possible causes of death: infections, cancer, and heart disease (the censored observation is alive). It follows, naturally, that as more individuals die from infections, there are fewer individuals at risk to die from cancer. It is the case that individuals face multiple causes of death, and as such the number of deaths for a particular cause will influence the estimate of the probability of dying due to the cause under scrutiny.

In this setting one needs to deal, simultaneously, with different competing events. The models discussed above must be adapted in order to deal with the fact that the number of failures from any competing risk (of failure) will condition the number of failures from the main failure, which, in turn, implies changes in the estimate of the probability of failure. Failures from any competing risk reduce the number of individuals at risk of failure from the cause under analysis (Gooley et al. (1999)). Competing risks are events that occur instead of the failure event of interest, which implies that we cannot treat them as censored. It follows that a competing risks framework becomes a natural solution for our estimation strategy. Two advantages follow from the use of a competing risks model. The theory behind the model allows both the computation of hazard functions where individuals can die due to multiple causes, and the computation of probabilities of death according to different values of the covariates.

Under the existence of competing risks we want to focus our attention on cause–specific hazards, as compared to standard hazards. A cause–specific hazard is the instantaneous risk of failure from a specific cause given that failure (from any cause) did not yet happened. We can see our problem as one where we have two cause–specific hazards: one for death by an infection and one for death by other cause. For the sake of simplicity, we focus below on this particular setting of the problem at hand. The analysis can easily be extended to a situation where one has three or more causes of death.

When we have competing events, we need to focus on the cumulative incidence function (CIF) rather than the survival function. A CIF is just the probability that a specific type of event is observed before a given time. In our analysis, we have two CIFs; one for death by infection and one for death due to other causes. For example, the CIF for death by infection at 60 years is just the probability of death by infection before an age of 60. CIFs begin at time zero and increase to an upper limit equal to the eventual probability that the event will take place, but this is not equal to one because of competing events. Mathematically, the CIF for death by infection is a function of all cause–specific hazards.

So, in a competing risks setting, a Kaplan-Meier curve is inadequate for three reasons.[11] First, it fails to acknowledge that death by infection may never occur. Second, the Kaplan–Meier solution does not take into account dependence between competing events. Third, facing competing risks, it is better to reverse the temporal ordering of the question. This implies that using Kaplan–Meier demands too much of the data; it requires *(i)* independent risks and *(ii)* a

---

[11] The Kaplan-Meier estimator is often used for estimating the survival function from lifetime data in medical research. For discussions see, for instance, Clark et al. (2003).

setup where the competing event does not occur. Berry et al. (2010) summarize the argument: "Kaplan-Meier survival analysis and Cox proportional hazards regression [...] can overestimate risk of disease by failing to account for the competing risk of death."

The CIF gives the proportion of individuals at time t who have died from cause k accounting for the fact that patients can die from other causes. For example, the CIF for death due to infections depends not only on the hazard for death by infection but also on the remaining hazards associated with other causes of death. This implies that it is no longer possible to define a direct relation between cause–specific hazard rate and the probability of death. Although nonparametric estimation of CIFs is flexible, it cannot be adjusted for relevant regressors, as they are associated with the cause–specific hazard. The efficient (and correct) way to run CIF covariate analysis is to implement a competing risks regression, according to the model of Fine and Gray (1999).

Fine and Gray (1999) propose an alternative to cause–specific hazards: a model for the hazard of the subdistribution for the failure event of interest, known as the sub–hazard. Unlike cause-specific hazards, discussed above, there is a one-to-one correspondence between sub–hazards and CIFs for respective event types; that is, the CIF for local relapse is a function of only the subhazard for local relapse. Covariates affect the sub–hazard proportionally, similar to the Cox regression. The authors propose a transformation of the Cox model associated with a direct transformation of the CIF. From the relation between the hazard and survival functions, Fine and Gray (1999) define a subdistribution function.

Although it is not at the core of our analysis, it is important to stress that the difference between cause-specific and subdistribution hazards is the risk set. For the cause-specific hazards, the risk set decreases each time there is a death from another cause (censoring), while under subdistribution hazards those who die due to another cause remain in the risk set and are given a censoring time, larger than all event times. Cause–specific hazard ratios give us a relative measure, where we can use standard survival analysis methods. However, covariates may not be associated with the cause–hazard. With the subdistribution hazards we account for competing events by altering the risk set. As there is a direct link between the subdistribution hazard and the CIF, one can compute the regressors' effects.[12]

Setup of competing risks models

In a general setting, for each individual in a competing risks model, the type of failure is specified by J, with values ranging from 1 to k. The random duration variable is defined by T. We assume, within our analysis, that there exists only one period of duration. The spell ends when individuals leave for one of the k possible states (states = failure type). The states are mutually exclusive and exhaustive, and are identified by the index j, where $j = 1, ..., k$.[13]

There are k random variables, $T^{(1)}, T^{(2)}, \dots, T^{(k)}$, corresponding to the existing states. These

---

[12] See Fine and Gray (1999) for a further discussion of semiparametric proportional hazards model for the subdistribution.

[13] In the presentation of the setup of competing risks models we follow closely the discussion in Sá et al. (2007).

variables can be interpreted as latent durations. These are abstract time periods used in the construction of the models, where $T^{(j)}$ is the time to failure to state j after the elimination of all other possible states. For each point in time, entry into a certain state is dictated by the smallest latent time period (the smallest $T(j)$). The time to failure can be specified as

$$T = min[T^{(1)}, \ldots, T^{(k)}], \tag{10}$$

where and $J = j$, if $T = T(j)$. For each individual, only one $T^{(j)}$ is observed and others are considered censored. We will have a competing risks model with independent risks under the assumption that the random variables $T^{(1)}, \ldots, T^{(k)}$ are independent.

It is possible to estimate conditional and unconditional probability functions that characterize the variables T and J. The expression

$$\lambda_j(t, x) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt, J = j | T \geq t, x)}{dt} \tag{11}$$

is the transition intensity into state j. These functions are designated as cause–specific hazard functions; they can be empirically interpreted as the fraction of survivors at time t that subsequently leave for state j. If one assumes a proportional hazards specification, the cause–specific hazard functions can be defined as

$$\lambda_j(t, x, \beta_j, \alpha_j) = \lambda_{0j}(t, \alpha_j) \exp(x'\beta_j), \ \ j = 1, ..., k, \tag{12}$$

where the risk-specific baseline hazard function is $\lambda_{0j}(t, \alpha_j)$. The parameters β_j and α_j are allowed to freely vary across the k failure types. Alternative distributions for the cause-specific baseline hazard lead to different cause-specific hazard functions. For example, if a Weibull baseline hazard is assumed, then the hazard function becomes

$$\lambda_j(t, x, \beta_j, \alpha_j) = \alpha_j t^{\alpha_j - 1} \exp(x'\beta_j). \tag{13}$$

It follows that we can estimate a set of coefficients for each of the competing risks.

Finally, the log-likelihood function within a competing risks framework can be expressed as

$$\ln L = \sum_{j=1}^{k} \left[\sum_{i=1}^{n} d_i \ln f(t_i, \beta_j, x_i, \theta_j) + \sum_{i=1}^{n} (1 - d_i) \ln S(t_i, \beta_j, x_i, \theta_j) \right]. \tag{14}$$

For a more detailed discussion of competing risks models, see, for instance, Cox (1959), David and Moeschberger (1978), Prentice et al. (1978), Lancaster (1990), and Kalbfleisch and Prentice (2002).

3 Data and model specification

3.1 Data selection and description

We select a sample of 155,062 women that were born before and are alive in 1971, and whose age throughout the sample is bounded between 16 and 85 years. The variable `Young` equals 1 if the woman had an own child.[14]

Table 1: Descriptive statistics

Variable	Mean	St.Dev.	Min.	Max.
Age	57·25	15·87	16	85
Died	0·30		0	1
Young	0·80		0	1
WChildren	0·04		0	1
Married	0·83		0	1
Occupation	0·61		0	1
House	0·80		0	1

The total number of observations is 155,062.
Source: ONS-LS.

Table 2: Health status

Status	Frequency	Share
Alive	108,717	70.11
Heart Diseases	13,563	8.75
Infections	5,318	3.43
Cancers	13,410	8.65
Other Diseases	5,321	3.43
Accidents, Hom., Suic.	1,111	0.72
Errors, Open, Others	7,622	4.92

Status indicating alive, or cause of death.
Source: ONS-LS.

From Table 1 we observe that average sample age is about 57 years. Detailed descriptives on age show us that 10% of the individuals have age equal or above 79 years; 10% of the sample is at most 35 years old. `Age` is either most recent age for alive individuals, or age at death.

In our sample, 30% of the individuals have died (`Died`). Details on the causes of death are provided in Table 2. The most common causes of death are Heart Diseases and Cancers. Looking to the event of interest for the current paper, 3.4% of the individuals in the sample died due to infections, which accounts for 11.5% of the deaths.[15] about 80% of our sample had young children, `Young`. The share of individuals that have at some point in their lives worked with young children, `WChildren`, is 4%. The share of individuals who were married, `Married`, were in white collar professions, `Occupation`, or own a house, `House`, are 83%, 61%, and 80%, respectively.

3.2 Empirical model

Death age is our duration variable, T. Cause of death, J, is equal to 0 if the observation is censored (the individual is alive), 1 if the individual dies from an infection and, 2 if (s)he dies

---

[14] For the robustness checks discussed in Section 4.3.3 we use a sample of 182,895 men. In this case `Young` equals 1 if they lived with a child in a common household at some point through their life span.

[15] The ONS-LS data set uses International Classification of Diseases (ICD) codes to categorize the main and, if applicable, contributory reasons of death. These codes come in several revisions of which 8,9, and 10 are relevant for the census waves we study; for details see World Health Organization (2010). The exact definition of infectious disease we use is the following combination of ICD-9 codes (and their earlier and later equivalents): Intestinal Infectious Diseases 001–139, Chronic Obstructive Pulmonary Disease 490–496, Occupational or Environmental Lung Disease 500–508, Other Diseases of Respiratory System 510–519. The other reasons for death listed in our tables (i.e., Heart Disease, Cancer, Other, Accident & Homicide & Suicide, and Error) are defined similarly according to the ICD system.

from other causes.[16] Individual characteristics include having young children in 1971 (`Young`), worked with children (`WChildren`=1 if ever worked with children), married (`Married`=1 if the individual was ever married), occupation (`Occupation`=1 if it is at some point in time a white collar worker), and house ownership (`House`=1 if the individual owns a house). The last two variables proxy for education and income, while the first two are our variables of interest.

The regressors used are the same in all specifications. Thus,

$$x'\beta = \beta_1 \texttt{Young} + \beta_2 \texttt{WChildren} + \beta_3 \texttt{Married} + \beta_4 \texttt{Occupation} + \beta_5 \texttt{House} \qquad (15)$$

where, depending on the specification, the constant may be explicitly added to the model or defined implicitly by a set of dummy variables. Coefficient estimates are then interpreted as the impact of each variable on the (conditional) probability of death and, consequently, on the age at death. For example, a negative estimate for β_1 indicates that, everything else constant, individuals with children show lower death probabilities and hence are more likely to stay alive.

4 Empirical analysis

4.1 Approach from a nonparametric perspective

In order to define a relatively homogenous group of individuals, we only consider the sample of women for most of our analysis. Later we run robustness checks with a sample of males.
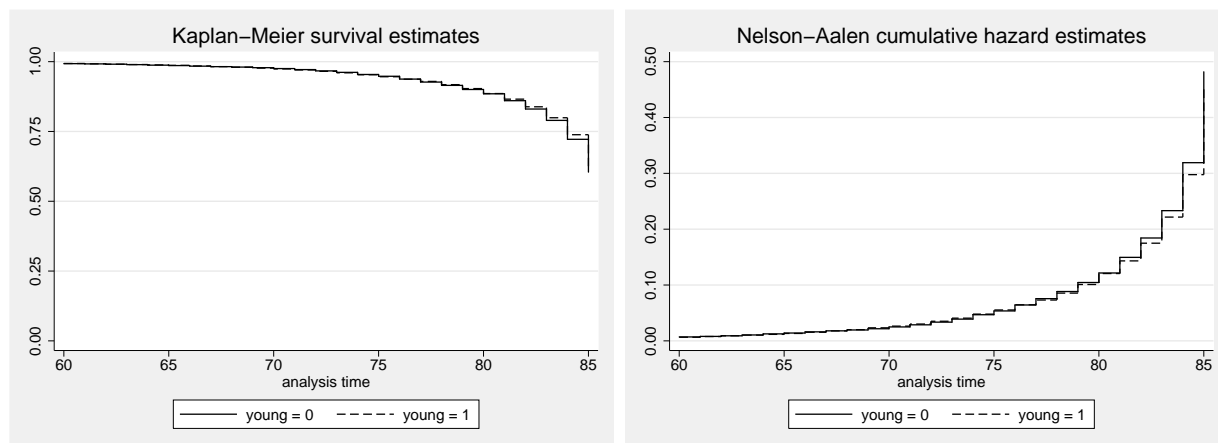

Figure 1: Left: Nonparametric: Kaplan-Meier survivor function. Right: Nonparametric: cumulative hazard. Sample where we define deaths by other causes as censored. Source: ONS-LS.

We start by considering a nonparametric estimation. At this stage, we consider three states for an individual's life condition: *(i)* alive; *(ii)* death by infection; *(iii)* death by other causes. Assume, for now, that the status is death by infections, and redefine other causes of deaths as alive (Left, Figure 1). The differences in the survivor function under Kaplan-Meier are small, with the survivor function for women with children slightly higher. In the right panel of Figure 1 we observe the corresponding cumulative hazard. (See the left and right panels of Figure 2 for

---

[16] Later we will expand the set of alternative causes of death according to Table 2.

smoothed versions of the cumulative hazard, based on Epanechnikov and Gaussian smoothing functions, respectively). Implementing a log-rank test for equality of survivor functions, we obtain a χ^2 statistic with 1 degree of freedom of 2.42, with a corresponding `p-value` of 0.12; i.e,. marginally, we do not reject the null hypothesis that both functions across the `Young` status are equal. Combining this information, although the evidence is not conclusive, it points, to some extent, to longer survival for women with children, or those who lived with children.
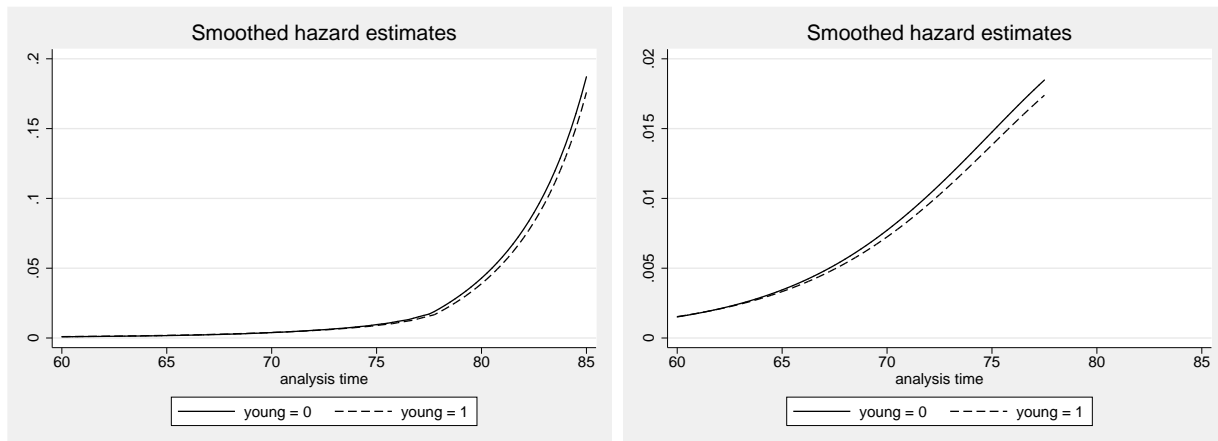

Figure 2: Left: Nonparametric: cumulative hazard, Epanechnikov smooth. Right: Nonparametric: cumulative hazard, Gaussian smooth. Sample where we define deaths by other causes as censored. Source: ONS-LS.

We next drop observations corresponding to death by other reasons. Now, the Kaplan-Meier and the Nelson-Aalen cumulative hazard are represented in Figure 3 which (as Figures 1), separate females according to the `Young` status. See Figure 4 for the smoothed versions. Once we opt for dropping observations corresponding to occurrences of death by other reasons, we do observe a stronger separation of survival and cumulative hazard according to the children status. The log-rank test for equality of survivor functions shows a χ^2 statistic of 173.41, with a corresponding `p-value` of approximately 0; i.e., we reject the null hypothesis that both functions across the `Young` status are equal. These results align with our hypothesis that life length can vary between those who had children, and those who had not. Women who had children (or lived with children) show both a higher survival rate, as well as a lower hazard rate at each age, indicating that, on average, they live longer.

We run a counterfactual analysis by considering that the event of interest is death by other causes. First, replicating the strategy designed above, we set those who died by infections as alive (although an incorrect procedure, this might give us a hint for what to expect when we move to the correct procedure). The Kaplan-Meier survivor function is represented in the left panel of Figure 5. If, as before, we drop the alternative death cause, which in the current case is death by infections, we obtain the right panel of Figure 5. Both figures seem to corroborate the key message of our paper; i.e., that the children status seems to matter for death by infections which it does not, or at least not to the same degree, for deaths by other reasons. This implies what we observe in both figures: survival by children status is indistinguishable. Performing the log-rank test for equality of survivor functions, we do, however, reject the null that both
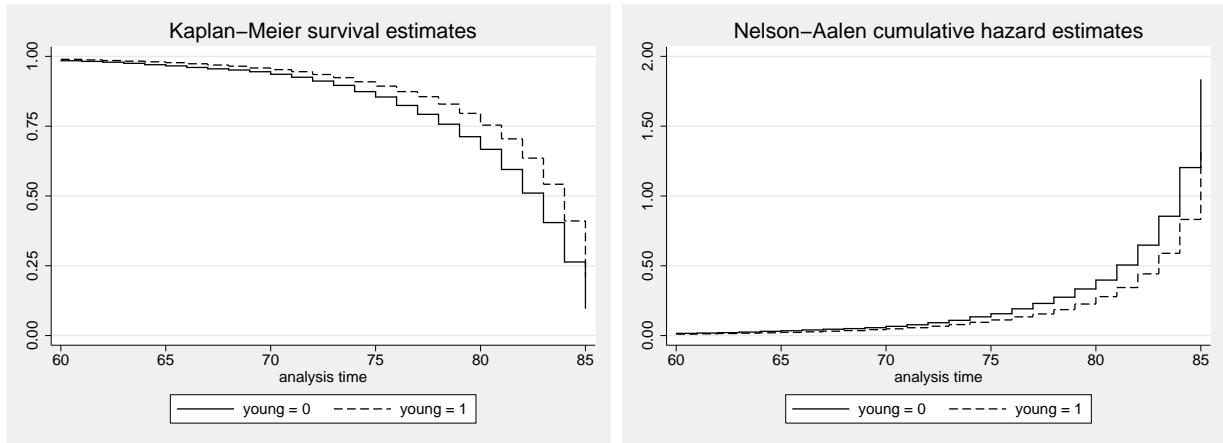
Figure 3: Left: Nonparametric: Kaplan-Meier survivor function. Right: Nonparametric: cumulative hazard. Sample where we eliminate deaths by other causes besides infection. Source: ONS-LS.
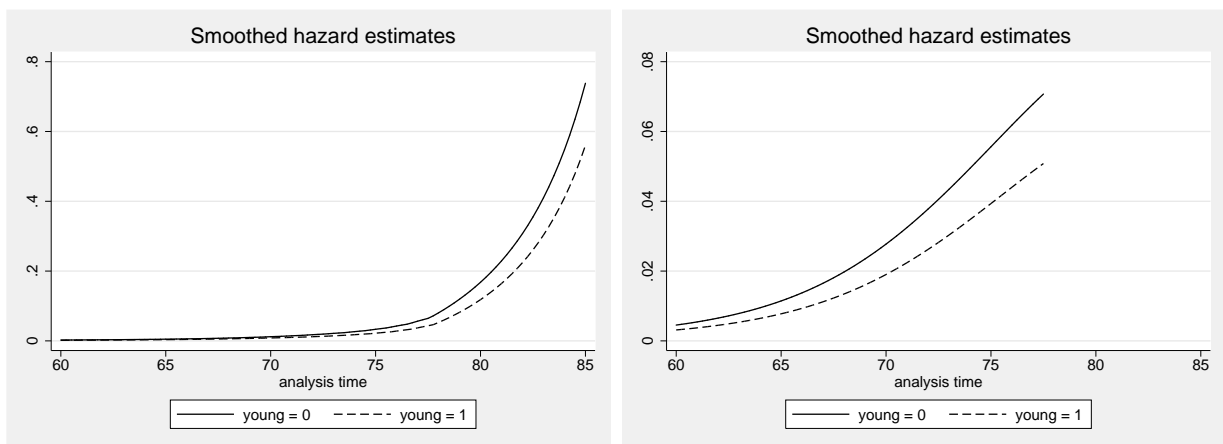

Figure 4: Left: Nonparametric: cumulative hazard, Epanechnikov smooth. Right: Nonparametric: cumulative hazard, Gaussian smooth. Sample where we eliminate deaths by other causes besides infection. Source: ONS-LS.

survivor functions are equal.[17] We discuss later why for some other causes of death, besides infections, we might still find a statistical difference across Young status. Essentially, we argue that this might be due to behavioral differences.

4.2   Semiparametric results

We now move to a semiparametric analysis and present results for the Cox proportional hazard model. The failure occurs when infections are the cause of death, and we treat, in a first stage, death by other causes as alive. The left panel of Figure 6 shows our first results. We observe that those who do not have children nor work with children face a higher hazard rate when compared to those with children or working with kids. If one drops observations related with death by other causes instead of treating these individuals as alive, we get the result in the right panel of Figure 6. The results point in the same direction as in the nonparametric analysis. I.e., women with children, or working with children live longer.

---

[17] The χ^2 statistics are 11.23 (p-value = 0.0008) and 36.20 (the p-value is approximately 0), respectively.
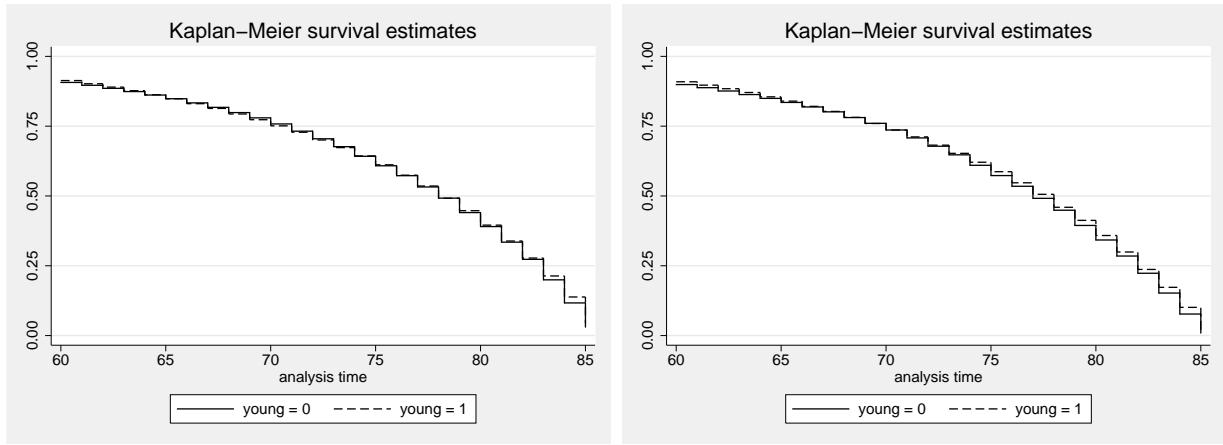
Figure 5: Left: Nonparametric: Kaplan-Meier survivor function. Sample where we define death by an infection as censored. Right: Nonparametric: Kaplan-Meier survivor function. Sample where we eliminate deaths by infection. Source: ONS-LS.
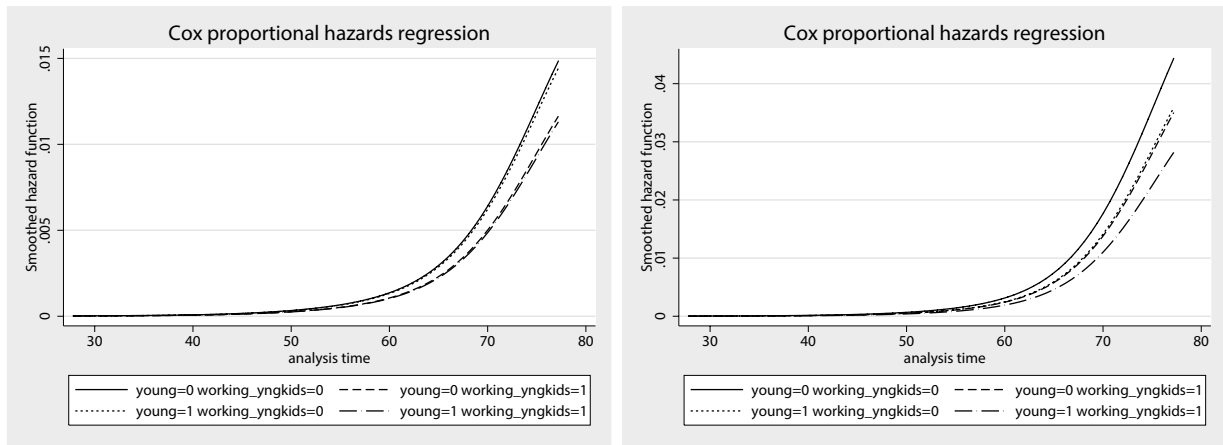
Figure 6: Left: Cox proportional hazard regression. Other Causes of death are treated as censored. Right: Cox proportional hazard regression. Observations for other causes of death are dropped. Source: ONS-LS.

In Table 4 (in the appendix) we present the estimation results for the Cox, Exponential and Weibull models. Models (1) and (2) define the failure event as death by an infection, while models (3) and (4) consider the failure as death by other cause besides infections. Models (1) and (3) treat the other cause of death as censored (alive), while models (2) and (4) drop observations associated with other causes of death. Under models (3) and (4) other cause of death only includes death by an infection.

All models are statistically significant as shown by the likelihood ratio tests. Being at some point in time a white collar worker or owning a house is associated with higher life expectancy. Married status is either associated with higher life expectancy, or statistically insignificant for the determination of the hazard rate.

The key variables of interest for the analysis are `Young` and `WYoung`. Models (1) and (2), using failure as death by infection, are the main semiparametric estimations, while models (3) and (4) work as counterfactual analysis. Under Model (1), the mistake we make is that death by other causes is treated as a censored observation. Under this restriction, and estimating a

14

Cox model, we observe that the hazard is 22% lower than the baseline hazard for those who work with children. Although the hazard rate is slightly lower for women with children, this difference is not statistically different from zero (the hazard ratio is not statistically different from 1). Excluding the observations corresponding to death by other causes from the analysis (Model (2), column Cox), we observe that the hazard for women is about 21% lower than the base line hazard, while for those who worked with children it is about 23% lower. In both cases the hazard ratios are statistically different from 1, and very close to each other. Looking to the counterfactual analysis, Models (3) and (4), and still focusing at the Cox estimations, we observe that the hazard ratios for the regressors of interest increased substantially. I.e., the impact of having a child, or working with children, is substantially lower when compared to the results under Models (1) and (2). The hazard ratio for women with children is under 4%; and for those working with children is 8%. Keep in mind that these results are incorrect following the discussion above. Still, the two sets of results do not reject our hypothesis.

4.3 Parametric results

4.3.1 Base estimations
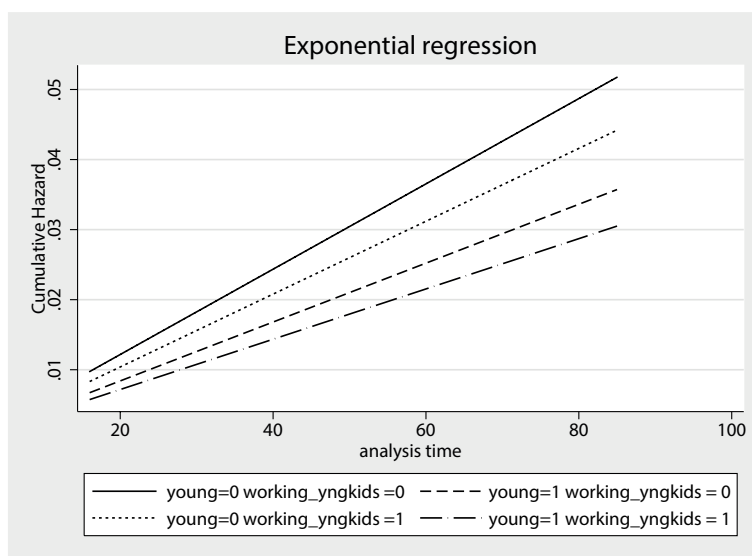
Figure 7: Parametric estimation. Source: ONS-LS.

Implementing a parametric estimation with a single risk, Figure 7, we are now clearly able to observe a distinction between individuals without kids (or contact with kids), and those who either had children and/or worked with children. Being in the presence of children significantly decreases the hazard of death.

Estimation results are provided in Table 4 (in the appendix), columns Exponential and Weibull. Looking to the log likelihood, it is always the case that the Weibull model presents the lower absolute value. Given the topic of our analysis, probability of dying, *a priory* the Weibull model seems more appropriate; i.e., the failure rate is expected to increase with time if there is an "aging" process. In all models the hypothesis that $ln(p)$ equals zero is rejected.

As such, we focus this section of the discussion the Weibull parametric estimation. Across all models the Weibull estimations confirm what we already observed with the Cox semiparametric estimations. I.e., women who have children or work with children live longer. Additionally, the counterfactual analysis strengthens our hypothesis as the effect hypothesized is minimal or non-existent for other diseases.

4.3.2 Competing risks

Figure 8 introduces the results for the competing risks estimations. In all figures we have represented the cumulative incidence of four situations: *(i)* women without children who never worked with children, *(ii)* women without children who worked with children; *(iii)* women with children but who never worked with children; and, finally, *(iv)* women with children who worked with children.

Figure 8 Left is obtained after an estimation of a competing risks model where failure is defined by death due to an infection, while the competing event is death by other causes, in which case all other causes are aggregated in a single category. The right figure is the result of an estimation where we reverse the roles of the previous two sets of causes of death: failure is death by other causes, while the competing event is death by an infection. There are clear differences between the two figures. While in the right hand figure, the counterfactual situation, there seems to be no distinction between the four cumulative incidences, on the left it seems clear that there is a separation of the different cumulative incidences. Women with children and those working with children live longer; women without children and those who never worked with children die at younger ages.

Table 5 (in the appendix) shows the estimation results for the different competing risks models that we discuss below. Definition A stands for the situation where causes of death are aggregated in two groups: infections and other diseases. Definition B occurs when we disaggregate other causes of death according to the categories defined in Table 2. All models estimated are statistically significant. Reinforcing the results discussed above, being married, being white collar or owning a house are all factors associated with longer lives. Focusing on the variables of interest, and looking to the key model in which failure is defined by death due to infection, Model (1), we conclude that women who either worked with children or had children live longer. Having children decreases the hazard rate by almost 8%, while working with children is associated with a reduction of the hazard by about 20%. Combining both conditions is associated with a reduction in the hazard of about 29%. The counterfactual, column (2), indicates that the effects of the key variables are much smaller. We face a combined effect below 8% against 21%; i.e., 21 percentage points smaller. At the same time, a reduction in the statistical significance occurs. This parametric interpretation is, naturally, aligned with the observations of Figure 8.

In a second stage we show the results for Definition B in Table 5, where we further disaggregate the causes of death. Under columns (3) to (8) we define the failure event as death by other cause besides infection, namely Heart Diseases, Cancers, Other Diseases, Accidents, Homicides,
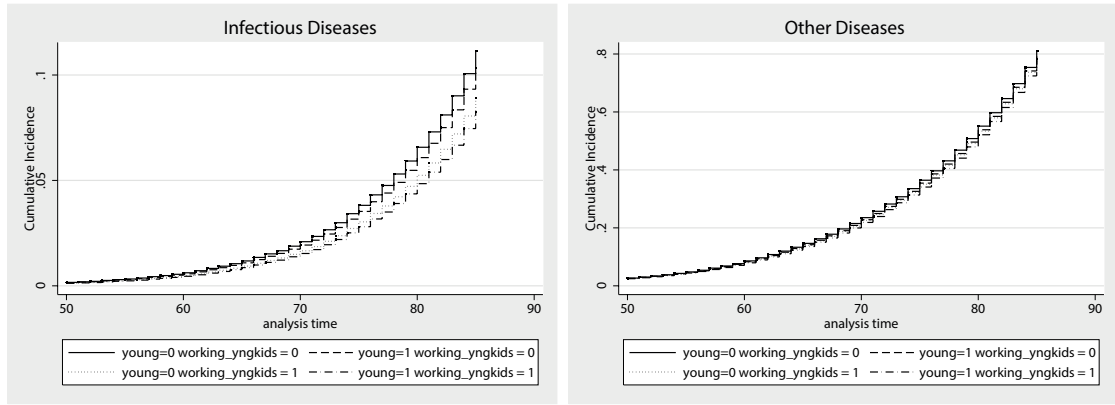
Figure 8: Left: Competing risks. Main risk: Infectious Diseases. Right: Competing risks. Main risk: Other Diseases. Source: ONS-LS.

Suicides, and Errors, Open or Other causes, respectively. In each case, the competing events are either infections and all other causes, columns (a), or just other causes as we drop the observations associated to death by an infection, columns (b). Columns (b) can be viewed as robustness checks for counterfactuals that are in themselves robustness checks.

Under Definition B, working with children is not coupled with the length of life. Except for Other Diseases, Column (5), condition (b) where we drop the observations associated with death by other causes, we face a marginally significant result at the 10% level of significance. Combining this result with the one under Column (1), our key result, we conclude in favor of our hypothesis: contact with children matters for death by an infection, but not for death by other causes, as predicted by our hypothesis.

Finally, we focus on the effect of having children on the hazard of dying by other causes. From Table 5 we can observe that this covariate does not matter for death by Cancer, column (4), neither for that by Errors, Open, or Other causes. This is the expected result. According to our hypothesis, having children should be uncorrelated with the timing of death by other cause besides an infection. The critical results, in the sense that they seem to be in dissonance with our hypothesis, are those under columns (3), (5) and (6). Death by Heart Diseases, Other Diseases, and Accidents, Homicides and Suicides, respectively. We argue that the results under these columns do not reject our hypothesis as they can be mainly attributed to behavioral changes.

Having children may be positively associated with improved health awareness (for instance, anecdotal evidence suggests that many adults stop smoking when they become parents), which would imply a lower hazard for dying due to Heart Diseases.[18] The same behavioral explanation may, to some extent, also be true for infectious diseases. Fewer parents may choose life styles which lower their defences against infections and, therefore, have a reduced mortality due to

---

[18] Willyard (2013), for instance, lists "smoking, high cholesterol, high blood pressure, diabetes, obesity and lack of exercise" as some of the major risk factors leading to heart disease. Many of these risks are to some extent behaviorally determined. While cessation of smoking benefits life expectancy, there are also negative influences. Fahrenwald and Walker (2003) report that "cross-sectional studies indicate that women with children are more sedentary than women without children." Since they argue that "physical activity reduces the risk not only of premature mortality, but also of coronary heart disease, hypertension, colon cancer, and type 2 diabetes," a sedentary lifestyle may increase overall mortality of mothers.

infections because of behavioral adjustments. This is unlikely, however, to explain the whole effect we observe for at least two reasons: *(i)* the group of those who work with young children *without having own children* show similarly improved survival probabilities although they have no similarly systematic reason to adjust their life style and *(ii)* the inclusion of an individual's marital state should already capture some of this behavioral effect.

Similarly, women with children may be less prone to commit suicide or to be involved in either car accidents, or death by an homicide. All these reasons may be, at least partially, the result of behavioral decisions that can be influenced by the fact that the women has a child. The last result that needs clarification is the one under column (5), death by Other Diseases. This classification is not defined in a precise way, and, we argue, it might include factors related to infections, not reported under column (1), or again causes that can be imputed to behavioral adaptations correlated with having children.

4.3.3 Additional robustness check: the male sample

Table 6 replicates Table 5 (both in the appendix) for the sample of males. Column (1) shows a much stronger effect of children on mens' hazard rate: 23% against 8%. The effect of working with children is also much stronger when compared with the effect on women. Disaggregating the different causes of death, columns (3) to (7), we conclude that the results under column (3) corroborate comparable results for females. Under column (4) we find that men with children have a lower hazard of dying by Heart diseases. For Other Causes of death, column (5), having children apparently does not determine the hazard of dying. For males, having children doubles the hazard of dying from an accident, Homicide or Suicide. We have no ready explanation for this result and, again, tend to a behavioral interpretation. Finally, in column (7) 'dying from Errors, Open and Others,' men face a higher, and statistically significant, hazard rate. However, given the open definition of this category, we do not consider this critical for our main result. When looking to the effect of working with children, one must take into account the fact that there are relatively few men performing this job. Nevertheless, the results we have are similar to those we found for women.

Concluding remarks

The key hypothesis of our paper, that having children makes a parent live longer, is not rejected by the data. Since the percentage of deaths due to infectious diseases is relatively small in developed countries (see Table 2), it would be most interesting to compare our results to those of a complementary study for a developing country where the population percentage dying from infections is higher. Unfortunately, it seems that such data is not readily available.[19] Further testing of our hypothesis could be done, for instance, in the wake of major immunization programs (which should be less effective for parents than for adults without children) or on the

---

[19] We found that data sets comparable to that of ONS-LS are collected in the Scandinavian countries. These seem to be, however, not accessible to outside researchers.

victims of major epidemics. Did, for instance, fewer WWI fathers die from the Spanish flu than soldiers without children? Does it make a difference if not the case of having children versus no children is studied but the (marginal) effects of the second, third etc child on mortality are analyzed? Should we expect similar effects in grandparents if they are looking after their grand children? In all cases that we considered the behavioral implications of parenthood are difficult to fully disentangle from the hypothesized pathological transmission mechanism. Hence, we are confident that our findings will spur vigourous debate.

References

BERG, G., S. GUPTA, AND F. PORTRAIT (2012): "Do Children Affect Life Expectancy? A Joint Study of Early life Conditions, Fertility and Mortality," *Population Association of America*, 2010 Annual Meeting.

BERRY, S. D., L. NGO, E. J. SAMELSON, AND D. P. KIEL (2010): "Competing Risk of Death: An Important Consideration in Studies of Older Adults," *Journal of the American Geriatrics Society*, 58, 783–7.

CLARK, T., M. BRADBURN, S. LOVE, AND D.G.ALTMAN (2003): "Survival Analysis Part I: Basic concepts and first analyses," *British Journal of Cancer*, 89, 232–38.

COX, D. R. (1959): "The analysis of exponentially distributed lifetimes with two types of failure," *Journal of the Royal Statistical Society, Series B*, 21, 411–21.

DAVID, H. A. AND M. L. MOESCHBERGER (1978): *The theory of competing risks*, London, UK: Griffin.

DIOR, U. P., H. HOCHNER, Y. FRIEDLANDER, R. CALDERON-MARGALIT, D. JAFFE, A. BURGER, M. AVGIL, O. MANOR, AND U. ELCHALAL (2013): "Association between number of children and mortality of mothers: results of a 37-year follow-up study," *Annals of Epidemiology*, 23.

DOBLHAMMER, G. (2000): "Reproductive history and mortality later in life: A comparative study of England and Wales and Austria," *Population Studies*, 54, 169–76.

FABIAN, D. AND T. FLATT (2011): "The Evolution of Aging," *Nature Education Knowledge*, 3, 9.

FAHRENWALD, N. L. AND S. N. WALKER (2003): "Application of the Transtheoretical Model of Behavior Change to the Physical Activity Behavior of WIC Mothers," *Public Health Nursing*, 20, 307–17.

FINE, J. P. AND R. J. GRAY (1999): "A Porportional Hazards Model for the Subdistribution of a Competing Risk," *Journal of the American Statistical Association*, 94, 496–509.

FLATT, T. AND D. PROMISLOW (2007): "Physiology: Still pondering an age-old question," *Science*, 318, 1255–6.

GAGNON, A., K. R. SMITH, M. TREMBLAY, H. VÉZINA, P.-P. PARÉ, AND B. DESJARDINS (2008): "Is There a Trade-off between Fertility and Longevity?" *University of Western Ontario*, PSC Discussion Paper #08-05.

GOOLEY, T. A., W. LEISENRING, J. CROWLEY, AND B. E. STORER (1999): "Estimation of Failure Probabilities in the Presence of Competing Risks: New Representations of old Estimators," *Statistics in Medicine*, 18, 695–706.

GRAHAM, N. M. H., K. E. NELSON, AND M. C. STEINHOFF (2007): "The Epidemiology of Acute Respiratory Infections," in *Infectious Disease Epidemiology*, ed. by K. E. Nelson and C. M. Williams, Sudbury, Mass: Jones & Bartlett, second ed.

HAYWARD, A. D., I. J. RICKARD, AND V. LUMMAA (2013): "Influence of early-life nutrition on mortality and reproductive success during a subsequent famine in a preindustrial population," *Proceedings of the National Academy of Sciences*, forthcoming.

HELLE, S. AND V. LUMMAA (2013): "A trade-off between having many sons and shorter maternal post-reproductive survival in pre-industrial Finland," *Biology Letters*, 9.

HELLE, S., V. LUMMAA, AND J. JOKELA (2002): "Sons reduced maternal longevity in preindustrial humans," *Science*, 296, 1085.

——— (2005): "Are reproductive and somatic senescence coupled in humans? Late, but not early, reproduction correlated with longevity in historical Sami women," *Proceedings of the Royal Society*, 272, 29–37.

HOUGAARD, P. (2000): *Analysis of multivariate survival data*, New York: Springer.

HURT, L., C. RONSMANS, AND S. THOMAS (2006): "The effect of number of births on womens mortality: systematic review of the evidence for women who have completed their childbearing," *Population Studies*, 60, 55–71.

IKEDA, A., H. ISO, H. TOYOSHIMA, Y. FUJINO, T. MIZOUE, T. YOSHIMURA, Y. INABA, A. TAMAKOSHI, AND J. GROUP (2007): "Marital status and mortality among Japanese men and women: the Japan Collaborative Cohort Study," *BMC Public Health*, 7.

KALBFLEISCH, J. D. AND R. L. PRENTICE (2002): *The Statistical Analysis of Failure Time Data*, New York: John Wiley & Sons Ltd.

KAPLAN, G., T. WILSON, R. COHEN, J. KAUHANEN, M. WU, AND J. SALONEN (1994): "Social functioning and overall mortality: prospective evidence from the Kuopio Ischemic Heart Disease Risk Factor Study," *Epidemiology*, 5, 495–500.

KIRKWOOD, T. (1977): "Evolution of aging," *Nature*, 270, 301–4.

KIRKWOOD, T. AND S. AUSTAD (2000): "Why do we age?" *Nature*, 408, 233–38.

KUNINGAS, M., S. ALTMÄE, A. UITTERLINDEN, A. HOFMAN, C. VAN DUIJN, AND H. TIEMEIER (2011): "The relationship between fertility and lifespan in humans," *Age*, 33, 615–22.

LAHDENPERÄ, M., V. LUMMAA, S. HELLE, M. TREMBLAY, AND A. F. RUSSELL (2004): "Fitness benefits of prolonged post-reproductive lifespan in women," *Nature*, 428, 178–81.

LANCASTER, T. (1990): *The econometric analysis of transition data*, Cambridge, UK: Cambridge University Press.

LEE, E. T. AND J. W. WANG (2003): *Statistical Methods for Survival Data Analysis*, New York: John Wiley & Sons Ltd.

LÓPEZ-OTÍN, C., M. A. BLASCO, L. PARTRIDGE, M. SERRANO, AND G. KROEMER (2013): "The Hallmarks of Aging," *Cell*, 153, 1194–217.

MCARDLE, P. F., T. I. POLLIN, J. R. OCONNELL, J. D. SORKIN, R. AGARWALA, A. A. SCHAÄFFER, E. A. STREETEN, T. M. KING, A. R. SHULDINER, AND B. D. MITCHELL (2009): "Does Having Children Extend Life Span? A Genealogical Study of Parity and Longevity in the Amish," *Journal of Gerontology*, 61A, 190–95.

MEDAWAR, P. B. (1946): "Old age and natural death," *Modern Quarterly*, 1, 30–56.

MONTO, A. S. (2002): "Epidemiology of Viral Respiratory Infections," *The American Journal of Medicine*, 112, 4S–12S.

MÜLLER, H., J. CHIOU, J. CAREY, AND J. WANG (2002): "Fertility and Lifespan: Later Children Enhance Female Longevity," *J Gerontol A Biol Sci*, 57, 202–6.

PARTRIDGE, L. AND N. BARTON (1993): "Optimality, mutation and the nature of ageing," *Nature*, 362, 305–11.

PRENTICE, R. L., J. D. KALBFLEISCH, A. V. J. PETERSON, N. FLOURNOY, V. T. FAREWELL, AND N. E. BRESLOW (1978): "The analysis of failure times in the presence of competing risks," *Biometrics*, 34, 541–54.

SÁ, C., C. E. DISMUKE, AND P. GUIMARÃES (2007): "Survival analysis and competing risk models of hospital length of stay and discharge destination: the effect of distributional assumptions," *Health Services and Outcomes Research Methodology*, 7, 109–24.

STEARNS, S. (1992): *The Evolution of Life Histories*, Oxford, UK: Oxford University Press.

TABATABAIE, V., G. ATZMON, S. N. RAJPATHAK, R. FREEMAN, N. BARZILAI, AND J. CRANDALL (2011): "Exceptional longevity is associated with decreased reproduction," *AGING*, 3, 1,202–5.

TOLLÅNES, M., D. MOSTER, A. DALTVEIT, AND L. IRGENS (2008): "Cesarean Section and Risk of Severe Childhood Asthma: A Population-Based Cohort Study," *The Journal of Pediatrics*, 153, 112–6.

WANG, X., S. G. BYARS, AND S. C. STEARNS (2013): "Genetic links between post-reproductive lifespan and family size in Framingham," *Evolution, Medicine, and Public Health*, forthcoming.

WILLIAMS, G. (1957): "Pleiotropy, natural selection, and the evolution of senescence," *Evolution*, 11, 398–411.

WILLYARD, C. (2013): "Pathology: At the heart of the problem," *Nature*, 493, doi:10.1038/493S10a.

WOOD, R. G., S. AVELLAR, AND B. GOESLING (2009): *Effects of Marriage on Health: A Synthesis of Recent Research Evidence*, New York: Nova Science Publishers Inc.

WORLD HEALTH ORGANIZATION (2010): *ICD-10: International statistical classification of diseases and related health problems (10th Rev. ed.)*, Genvea: World Health Organization.

Appendix

Table 3: Health status by working with children and having children

	Health status	Females		Males	
		No–Children	Children	No–Children	Children
No–Work with Children	Alive	15, 931	87, 758	9, 582	84, 633
	Heart Diseases	4, 185	9, 083	22, 379	9, 698
	Remaining diseases	[8, 720]	[23, 236]	[33, 954]	[19, 247]
	Infections	1, 578	3, 646	8, 065	3, 048
	Remaining diseases	[11, 327]	[28, 673]	[48, 268]	[25, 897]
	Cancers	3, 209	9, 828	13, 908	7, 348
	Remaining diseases	[9, 696]	[22, 491]	[42, 425]	[21, 597]
	Other Diseases	1, 535	3, 665	4, 427	2, 456
	Remaining diseases	[11, 370]	[28, 654]	[51, 906]	[26, 489]
	Accidents, Hom., Suic.	375	709	1, 118	1, 345
	Remaining diseases	[12, 530]	[31, 610]	[55, 215]	[27, 600]
	Errors, Open, Others	2, 023	5, 388	6, 436	5, 049
	Remaining diseases	[10, 882]	[26, 931]	[49, 897]	[23, 896]
Work with Children	Alive	963	4, 065	317	1, 933
	Heart Diseases	181	114	314	153
	Remaining diseases	[362]	[464]	[379]	[306]
	Infections	57	37	66	27
	Remaining diseases	[486]	[541]	[627]	[432]
	Cancers	158	215	150	125
	Remaining diseases	[385]	[363]	[543]	[334]
	Other Diseases	50	71	52	49
	Remaining diseases	[493]	[507]	[641]	[410]
	Accidents, Hom., Suic.	11	16	14	17
	Remaining diseases	[532]	[562]	[679]	[442]
	Errors, Open, Others	86	125	97	88
	Remaining diseases	[457]	[453]	[596]	[371]

Health status indicates if the individual is alive, or the cause of death. Source: ONS-LS.

Table 4: Semiparametric and parametric analysis – Cox, Exponential and Weibull regressions – females

	Risk = Infections						Risk = Other diseases					
	Model (1)			Model (2)			Model (3)			Model (4)		
Variables	Cox	Exponential	Weibull	Cox	Exponential	Weibull	Cox	Exponential	Weibull	Cox	Exponential	Weibull
Young	0·971	0·690***	0·973	0·789***	0·499***	0·787***	0·974**	0·715***	0·973**	0·962***	0·698***	0·959***
	(0·036)	(0·025)	(0·036)	(0·028)	(0·017)	(0·028)	(0·013)	(0·009)	(0·013)	(0·013)	(0·009)	(0·013)
WYoung	0·778**	0·854	0·787**	0·768**	0·889	0·767**	0·940*	1·033	0·943*	0·924**	1·031	0·929**
	(0·083)	(0·091)	(0·084)	(0·082)	(0·094)	(0·082)	(0·031)	(0·033)	(0·031)	(0·030)	(0·033)	(0·030)
Married	0·992	0·790***	0·961	0·729***	0·693***	0·676***	1·006	0·854***	0·974*	0·982	0·839***	0·948***
	(0·038)	(0·029)	(0·037)	(0·028)	(0·025)	(0·025)	(0·014)	(0·012)	(0·014)	(0·014)	(0·012)	(0·013)
Occupation	0·773***	0·406***	0·777***	0·620***	0·273***	0·613***	0·855***	0·476***	0·857***	0·831***	0·457***	0·833***
	(0·024)	(0·013)	(0·024)	(0·020)	(0·009)	(0·019)	(0·009)	(0·005)	(0·009)	(0·009)	(0·005)	(0·009)
House	0·611***	0·501***	0·614***	0·508***	0·346***	0·508***	0·731***	0·604***	0·736***	0·703***	0·579***	0·707***
	(0·018)	(0·015)	(0·018)	(0·015)	(0·010)	(0·014)	(0·008)	(0·007)	(0·008)	(0·007)	(0·006)	(0·007)
Log likelihood	-52816	-20734	-8575	-47427	-17187	-3237	-422420	-82346	-7881	-417980	-79690	-3586
LR test	481***	2,558***	475***	1,316***	5,533***	1,434***	1,407***	12,079***	1,398***	1,863***	13,709***	1,879***
PH test	161·04***			349·94***			606·63***			642·86***		
$ln(p)$			2·471***			2·549***			2·154***			2·170***
Observations		155,062			114,035			155,062			149,744	
Failures		5,318			5,318			41,027			41,027	

Significance levels: *: 10%, **: 5%, ***: 1%. The dependent variable is age. Standard errors in parentheses. Under each model we report the hazard ratio. The estimation procedure is defined in each column. Risk = Infections – the failure is defined as death by an infection; Risk = Other diseases – the failure is defined as death by other causes besides infection. Model (1) – death by other causes is defined as censored; Model (2): deaths by other causes are dropped from the sample; Model (3) – death by infection is defined as censored; Model (4) deaths by infection are dropped from the sample. Source: ONS-LS.

Table 5: Competing risks analysis – females

Variables	Definition A		Definition B									
	(1)	(2)	(3)		(4)		(5)		(6)		(7)	
			(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)
Young	0·923**	0·966***	0·876***	0·867***	0·995	0·985	0·911**	0·903***	0·770***	0·760***	1·015	1·006
	(0·033)	(0·013)	(0·020)	(0·019)	(0·024)	(0·024)	(0·033)	(0·033)	(0·059)	(0·058)	(0·030)	(0·030)
WYoung	0·792**	0·953*	0·962	0·947	1·000	0·986	0·868	0·855*	0·797	0·790	0·945	0·930
	(0·083)	(0·028)	(0·055)	(0·054)	(0·053)	(0·052)	(0·081)	(0·080)	(0·156)	(0·157)	(0·067)	(0·066)
Married	0·862***	0·976*	0·799***	0·785***	1·085***	1·064**	0·907**	0·893***	0·610***	0·598***	0·942*	0·929**
	(0·032)	(0·013)	(0·019)	(0·018)	(0·029)	(0·028)	(0·035)	(0·034)	(0·048)	(0·047)	(0·030)	(0·030)
Occupation	0·749***	0·858***	0·683***	0·663***	0·904***	0·878***	0·852***	0·832***	0·757***	0·733***	1·040	1·018
	(0·024)	(0·009)	(0·013)	(0·013)	(0·017)	(0·017)	(0·026)	(0·025)	(0·051)	(0·050)	(0·026)	(0·025)
House	0·699***	0·772***	0·774***	0·739***	0·776***	0·742***	0·837***	0·805***	0·847**	0·815***	0·982	0·944**
	(0·020)	(0·008)	(0·014)	(0·013)	(0·015)	(0·014)	(0·025)	(0·024)	(0·057)	(0·055)	(0·025)	(0·024)
Wald $\chi^2(5)$	404.92***	1,154.38***	1,218.05***	1,511.00***	267.22***	387.92***	122.52***	168.90***	153.52***	175.55***	7.46	14.84**
Observations	155,062	155,062	155,062	149,744	155,062	149,744	155,062	149,744	155,062	149,744	155,062	149,744
No. Failures	5,318	41,027	13,563	13,563	13,410	13,410	5,321	5,321	1,111	1,111	7,622	7,622
No. Competing	41,027	5,318	32,782	27,464	32,935	27,617	41,024	35,706	45,234	39,916	38,723	33,405

Significance levels: *: 10%, **: 5%, ***: 1%. The dependent variable is age. The models are estimated by competing risks procedures. Standard errors in parentheses. Definition A: the cause of death is defined in two categories - (1) infection as the main risk, and other causes as the competing risk, and (2) the reverse, where the main risk is other causes of death. Definition B: other causes of death are split in (3) Heart Diseases, (4) Cancers, (5) Other Diseases, (6) Accidents, Hom., Suic., and (7) Errors, Open, Others. In columns (a) we keep the full set of observations, implying that for each alternative cause of death infection becomes a competing risk. In columns (b) we drop the observations corresponding to infections when analyzing a particular risk of death. Source: ONS-LS.

Table 6: Competing risks analysis – males

Variables	Definition A		Definition B									
	(1)	(2)	(3)		(4)		(5)		(6)		(7)	
			(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)
Young	0·775***	1·220***	0·770***	0·742***	0·929***	0·898***	1·034	1·002	2·013***	1·908***	1·483***	1·446***
	(0·017)	(0·010)	(0·009)	(0·009)	(0·014)	(0·013)	(0·027)	(0·026)	(0·090)	(0·085)	(0·028)	(0·027)
WYoung	0·748***	0·923***	0·985	0·962	0·887**	0·866**	0·935	0·915	0·925	0·910	0·932	0·912
	(0·078)	(0·025)	(0·045)	(0·043)	(0·054)	(0·052)	(0·094)	(0·092)	(0·169)	(0·166)	(0·069)	(0·067)
Married	0·801***	0·894***	1·076***	1·036**	1·269***	1·230***	0·671***	0·648***	0·246***	0·242***	0·742***	0·715***
	(0·019)	(0·009)	(0·017)	(0·016)	(0·026)	(0·025)	(0·021)	(0·020)	(0·011)	(0·011)	(0·019)	(0·018)
Occupation	0·725***	0·967***	0·968***	0·923***	0·913***	0·874***	1·028	0·989	0·752***	0·727***	1·072***	1·031
	(0·016)	(0·007)	(0·012)	(0·011)	(0·014)	(0·013)	(0·027)	(0·026)	(0·035)	(0·034)	(0·022)	(0·021)
House	0·750***	0·843***	0·884***	0·838***	0·803***	0·763***	0·941***	0·902***	0·747***	0·710***	1·116***	1·070***
	(0·015)	(0·007)	(0·011)	(0·010)	(0·012)	(0·011)	(0·025)	(0·024)	(0·035)	(0·033)	(0·023)	(0·022)
Wald $\chi^2(5)$	1,103.31***	1,122.84***	731.79***	1,215.56***	497.05***	766.20***	182.91***	247.90***	1166.95***	1245.59***	564.25***	492.96**
Observations	182,895	182,895	182,895	171,689	182,895	171,689	182,895	171,689	182,895	171,689	182,895	171,689
No. Failures	11,206	75,224	32,544	32,544	21,531	21,531	6,984	6,984	2,494	2,494	11,670	11,670
No. Competing	75,224	11,206	53,886	42,680	64,889	53,693	79,446	68,240	83,936	72,730	74,760	63,554

Significance levels: *: 10%, **: 5%, ***: 1%. The dependent variable is age. The models are estimated by competing risks procedures. Standard errors in parentheses. Definition A: the cause of death is defined in two categories - (1) infection as the main risk, and other causes as the competing risk, and (2) the reverse, where the main risk is other causes of death. Definition B: other causes of death are split in (3) Heart Diseases, (4) Cancers, (5) Other Diseases, (6) Accidents, Hom., Suic., and (7) Errors, Open, Others. In columns (a) we keep the full set of observations, implying that for each alternative cause of death infection becomes a competing risk. In columns (b) we drop the observations corresponding to infections when analyzing a particular risk of death. Source: ONS-LS.