pairfam
Panel Analysis of Intimate Relationships and Family Dynamics

Partners and Babies

Partnership and Fertility Histories of the German Family Panel (pairfam/DemoDiff, waves 1-3) as an Event History Data Set

Christine Schnor, Sonja Bastin

January, 2014

Funded as long-term project by the German Research Foundation (DFG)

Pairfam Technical Paper

03

Partners and Babies

Partnership and Fertility Histories of the German Family Panel (pairfam/DemoDiff, waves 1-3) as an Event History Data Set

Authors: Christine Schnor[1] and Sonja Bastin[2]

Abstract:
The German Family Panel (pairfam and DemoDiff, waves 1-3, Release 3.1) provides two generated biographical data sets (*biochild.dta* and *biopart.dta*) that contain information on fertility and partnership histories. Before these data can be used for event history or sequence analyses, they must be transferred into a spell format. In this report, we explain how this transfer is made. We provide a STATA code (*Eventhistory.do)* that generates an event history data set that can be used for various kinds of event history and sequence analyses in the realm of fertility and partnership dynamics. With the generated *Eventhistory.dta*, it is easy to identify the timing of family-related events, like the formation, dissolution, and interruption of marriages and other types of partnerships; as well as the birth of children. In addition, further episode-specific information on the family arrangement is included. Transferring the data into a spell format involves two major tasks: bringing the partnership and fertility histories into convergent and compatible formats, and dealing with missing date information. Moreover, *Eventhistory.dta* includes indicator variables for imputed date information in *biopart.dta* and *biochild.dta,* as well as for individuals who did not experience any events in their partnership or fertility biographies. This technical report is accompanied by the STATA codes that generate the spell data, as well as two examples of analyses. In addition, an Excel file exemplifies the structure of the data set.

---

[1] Affiliations: Max Planck Institute for Demographic Research, Konrad-Zuse-Str. 1, D-18057 Rostock; Vrije Universiteit Brussel, Pleinlaan 5, B-1050 Brussels. Contact: christine.schnor@vub.ac.be

[2] Affiliation: Max Planck Institute for Demographic Research, Konrad-Zuse-Str. 1, D-18057 Rostock. Contact: bastin@demogr.mpg.de

Table of contents

1 Introduction **4**
2 General notes **8**
2.1 Definition of partnership and fertility 8
2.2 From *biochild* and *biopart* to an event history data set 10
3 Generation of the event history data **13**
3.1 General time-varying variables 13
3.2 Partnership biography 14
3.2.1 Partnership episodes 14
3.2.2 Order variables 16
3.2.3 Further partnership information 17
3.3 Fertility biography 19
3.3.1 Ordering of children 19
3.3.2 Fertility episodes 20
3.3.3 Episodes of living with children 21
3.3.4 Further information 22
4 Recommendations and summary **24**
4.1 Recommendations 24
4.2 Summary 27
5 Examples of analyses **28**
5.1 Transition to the first union 28
5.2 Transition to a separation after the first childbirth 30
6 Acknowledgements **34**
7 References **34**
8 Annex **35**

Index of Tables

Table 1: Files provided by the German Family Panel 6
Table 2: Definition of homosexual and heterosexual partnerships in *Eventhistory.dta* 18
Table 3: Overview of flag variables and possible applications 26

Index of Figures

Figure 1: Partnership dimensions 9
Figure 2: Transition to the first union, age 14 to age 24, Kaplan-Meier survival estimates 30
Figure 3: Transition to separation after first childbirth, from birth to age 8 (first child), Kaplan-Meier-Survival estimates, by marital status (time of first birth). 33
Figure 4: Piecewise constant exponential model, controlled for marital status (time-varying), from birth to age 8 of the first child, hazard ratios. 34

Index of Annex

Within Eventhistory package:

Eventhistory.do

Eventhistory_ReadMe.txt

Eventhistory_Example_id.xlsx

- Table I: *biopart.dta*
- Table II: *biochild.dta*
- Table III: *Eventhistory.dta*

Eventhistory _Example_analysis1.do

Eventhistory _Example_analysis2.do

biopart_PF.do

biopart_PF_IMP.do

biopart_DD.do

biopart_DD_IMP.do

At the end of this document:

Table IV: List of variables included in *Eventhistory.dta*

1 Introduction

Note to the user: Feel free to start immediately, or to read this report for more detailed information.

Please note that it is possible to start your own event analysis with STATA immediately by using *Eventhistory.dta*, which is generated by our *Eventhistory.do* file.[3] The delivered *Eventhistory_ReadMe.txt* document lists the steps you need to conduct to run *Eventhistory.do*. We further explain this procedure in the following paragraph ("How to retrieve *Eventhistory.dta*"). To learn more about our procedures or about how to perform individual modifications or to get an idea how to implement further waves (waves 4, 5 etc.), continue reading this report. Please note that we use STATA as software to construct *Eventhistory.dta*. The version STATA SE is needed to process the required number of variables.

How to retrieve Eventhistory.dta[4]

In order to run the *Eventhistory.do* file described here—which in turn generates *Eventhistory.dta*—small adjustments need to be made. Firstly, you need to gather several data sets and syntax files (see Table 1). Secondly, you need to rename some data file labels as we did (see *Eventhistory_ReadMe.txt*). Thirdly, small changes within the syntax of *biopart.do (*pairfam as well as DemoDiff*)* are necessary and recommendable. You need to enter your personal data path into the syntax. Further, for your convenience, we recommend that you introduce the command "set more off" at the beginning of the do file. *Biopart.do* uses variable name abbreviations in its commands. Thus, if you wanted to stop STATA from recognizing abbreviations, you would need to type >set varabbrev on< at the beginning of the *biopart.do* file and at the beginning of the *biochild.do* file. After making these adjustments, you can run the *Eventhistory.do* file.

---

[3] The use of data generated by *Eventhistory.do* should be indicated in your work by citing this report.

[4] To generate an event history data set, we split the data if an event occurs. These splits can consume a great deal of memory. We generally choose a memory setting of one gigabyte (g), which is enough to conduct all splits. The final *Eventhistory.dta* has a size of 113 megabytes (m). If the memory demand exceeds the user's capacities, we recommend commenting out the variable AGEANC (age of the respondent in years). This variable splits the data for each respondent by year. As the respondent's age is a central control covariate in most analyses, we decided to include the variable in the data. If you drop the generation of this variable, you will get a final data size of 55 megabytes.

Short description of the German Family Panel

The here described data base on the German Family Panel pairfam and its supplement DemoDiff. In the following we always refer to pairfam and DemoDiff jointly, when we mention the “German Family Panel”. Pairfam (Panel Analysis of Intimate Relationships and Family Dynamics) is a, multidisciplinary, longitudinal study for researching partner and family dynamics in Germany. It is coordinated by Josef Brüderl, Johannes Huinink, Bernhard Nauck, and Sabine Walper. The survey is funded as a long-term project by the German Research Foundation (DFG) (Huinink et al. 2011; Nauck et al. 2012). Pairfam had its first wave in 2008/2009, and is being conducted annually over the subsequent 14 years. The interview data are gathered from a nationwide random sample of anchor persons of the three birth cohorts 1971-73, 1981-83, and 1991-93. For the full data documentation, see Brüderl et al. (2013) and Huinink et al. (2011). Pairfam gathered information from respondents living in western and eastern Germany. DemoDiff (Demographic Differences in Life Course Dynamics in Eastern and Western Germany) is a supplementary study to pairfam. It only samples respondents of the birth cohorts 1971-1973 and 1981-1983 who lived in eastern Germany (excluding West Berlin) at time of first interview (2009/2010). Like pairfam, annual standardized personal interviews are conducted. The vast majority of the German population lives in the western part of the country. However, the German history raises the question whether family life in the former socialistic eastern part of Germany differs from the rest of the country. The oversampling of eastern Germans allows solid comparisons between the two regions with the German Family Panel. For detailed information on the conception of DemoDiff and the main differences to pairfam please see Kreyenfeld et al. 2011.

In its first wave (conducted 2008/2009 (pairfam) and 2009/2010 (DemoDiff), respectively), the German Family Panel collected retrospective data on the partnership and fertility biographies of the respondents, which are updated with each consecutive wave. These data include detailed information not only on episodes of co-residence with a spouse or partner, but also on partnership episodes that do not involve living together. Furthermore, retrospective information on biological children, as well as on non-biological children and the respondent's co-residential history with these children is available. Former partners can be identified as the second biological parent of children in the retrospective data. It is these partnership and fertility biographies – the partners and the babies - we focus on in this report.

The German Family Panel as an event history data set

This report describes how data on the partnership and fertility biographies of individuals can be brought together to form a single, coherent event history data set. This data set may be matched to further information on the anchor person, his or her children, and his or her parents, as well as prospective partners using the respective person identifiers (pid, cid, mid, fid, smid, sfid). *Eventhistory.dta* is based on the third release of data of the *pairfam* group, and therefore includes the pairfam waves 1 to 3 (Release 3.1) and the DemoDiff waves 1 and 2/3 (Release 2.0).[5] We base our code on files provided by the pairfam and DemoDiff group which are listed in Table 1. All files are either available as Scientific Use Files from the GESIS Data Archive or are provided in the Eventhistory package delivered by the DemoDiff group. For any questions please refer to the pairfam user service (support@pairfam.de).

Table 1: Files provided by the German Family Panel

Pairfam	DemoDiff	
anchor1.dta	*anchor1_DD.dta*	Provided by pairfam user service
anchor2.dta	*anchor2_DD.dta*	
anchor3.dta		
biopart.dta	*biopart.dta*	
biochild.dta	*biochild.dta*	
biopart_PF.do	*biopart_DD.do*	Provided by DemoDiff group within the *Eventhistory* package
biopart_PF_IMP.do	*biopart_DD_IMP.do*	

[5] As the first wave of DemoDiff started one year after the first pairfam waves, the DemoDiff coordinators decided to merge the questionnaires of the second and third wave to allow the synchronization with pairfam.

We validate our work in three steps. First, we illustrate the structure of the data sets *biopart.dta, biochild.dta*, and *Eventhistory.dta* with an example id (see *Eventhistory_Example_id.xlsx*). This allows the user to compare easily the information in *biopart.dta* and *biochild.dta* with our generated *Eventhistory.dta*. Second, we provide two examples of how the data may be used (see *Eventhistory_Example_analysis1.do* and *Eventhistory_Example_analysis2.do*. Third, we provide the STATA syntax file *Eventhistory.do*, which creates the *Eventhistory.dta*, which in turn makes our work completely comprehensible. The files *Eventhistory.do* and *Eventhistory.dta* may be used by other users by citing this document. However, we accept no responsibility for errors that may have arisen during the coding procedures.

The original data sets (*biopart.dta* and *biochild.dta*) contain a wide range of date variables. This guarantees that the user has access to the data that have been manipulated the least and that have a variety of potential uses. However, because the data structure is complex, the data need to be edited extensively before analyses like event history or sequence analyses can be conducted. Our aim is to create a data set that allows users to conduct duration analyses immediately, without a major restructuring of the data. We improve the manageability of the data by transforming all of the available date information into time-varying variables. When appropriate, further information on partnership and fertility is also included in this time-varying manner by linking it to the date information. This data set thus offers users the opportunity to analyze easily a variety of research topics, including fertility behavior, union formation and dissolution, and the process by which partnerships are established.

We very carefully clean the fertility and partnership information to provide detailed and consistent biographies, and to flag any imputed date information in *biopart.dta* and *biochild.dta*. In order to generate a single, coherent event history data set, we bring the partnership and fertility histories into compatible formats and add the population at risk (persons without partnership or children information). However, transferring the data into spell format requires us to simplify the data in some instances. We aim to generate clean and consistent fertility careers, which is why we focus on the *biological* children of the anchor person.[6] Thus, for researchers who are particularly interested in non-biological children, these data might be of limited value.

Structure of the Report

Chapter 2 defines our concept of partnership and fertility. We then give an overview of how fertility and partnership histories are provided in the German Family Panel, and the factors we consider when we generate *Eventhistory.dta*. In Chapter 3, we illustrate how we generate the spell data and describe in more detail the variables included. In Chapter 4, we summarize the benefits of *Eventhistory.dta* and offer advice on how users should handle the provided flag variables. Finally, we provide two examples of analyses in Chapter 5.

2 General notes

2.1 Definition of partnership and fertility

The German Family Panel includes different partnership dimensions. First, it provides information on whether the respondent has a partner. We define this dimension as a "union." Second, it provides information on the partner with whom the respondent co-resides. This living arrangement is called "cohabitation." Third, information about marriage is included in the data. These three dimensions of union, cohabitation, and marriage are included in the concept of "partnership." Thus, when we refer to issues that are relevant to any of these three dimensions, we use the term "partnership."

In the interviews, the respondents are asked about the length of their union, cohabitation, or marriage. This date information leads to different combinations of the partnership dimensions, as we illustrate in Figure 1; namely,

1: having a relationship outside of marriage and cohabitation (a so-called "living apart together relationship");

2: having a co-residing non-marital relationship;

3: having a co-residing marital relationship;

[6] For non-biological children more limited information like episodes of co-residence with the anchor person is considered.

4: being married to the partner and living in separate households;

5: "still" cohabiting, without having a relationship with the (former) partner;

6: "still" cohabiting with the spouse, without being in a union with the spouse; and

7: "still" being married, but living apart from the former partner.

These are the main partnership categories that concentrate on the relationship to a single partner. However, over the life course, most people enter into relationships with various partners. This leads to different combinations of the partnership categories. For example, a person can still be married to a former partner (Category 7), but already have a new non-marital relationship with a separate household (Category 1). Another possibility is that a person is engaged in two simultaneous partnerships (both category 1). These multiple partnership statuses are identified by the German Family Panel and are also considered in *Eventhistory.dta*.

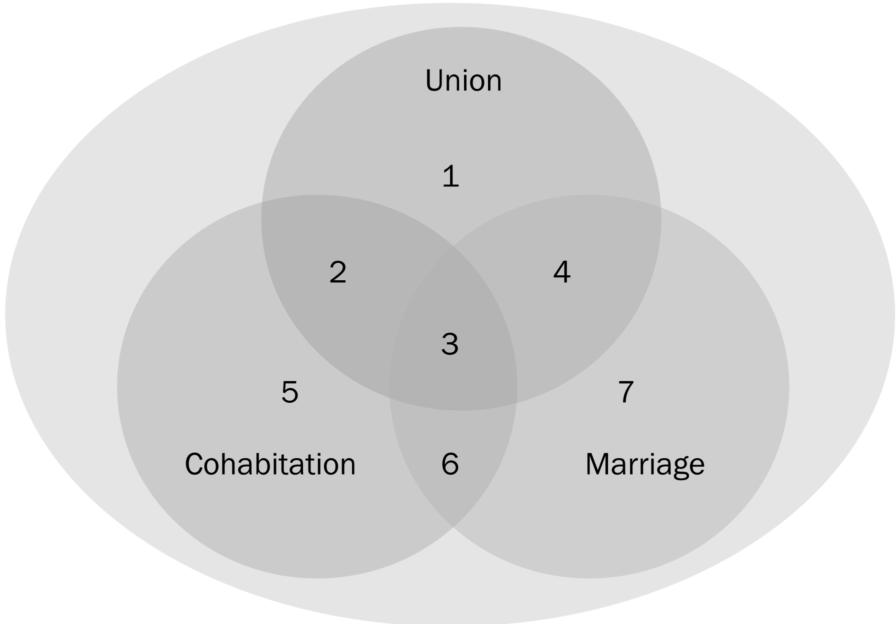

Figure 1: Partnership dimensions

The German Family Panel collects a range of information on the biological children, stepchildren, foster children, and adoptive children of the respondent[7]. In *Eventhistory.dta*, we provide information on the date of birth, the sex, the identity of the second biological parent, and the residence of each *biological* child of the respondent. For all other children (step-, foster, and adoptive children), we provide information on their co-residence with the respondent.

[7] Information on non-biological children is only available for children who have ever lived in the respondent's household.

2.2 From biochild and biopart to an event history data set

In order to facilitate analyses with the partnership and fertility biographies, the German Family Panel data group offers two files that provide fertility and partnership histories (*biopart.dta* and *biochild.dta*), as well as the Stata codes that generate these files from the original data (*biopart.do* and *biochild.do*[8]). This means that the fertility and partnership histories are already cleaned to some extent, as the data have already been checked for major inconsistencies (such as cases in which the end of a partnership was dated before the start, or in which the partners' first meeting is dated after they started their relationship). The data are provided in long format, which means that all of the information is stored in one row per partnership or per child. The partners (rows) are ordered according to the stated start date of the relationship, with the last row being the most current partnership. For the fertility history, there is one row for each child ordered according to the birth dates, with the youngest child in the last row.[9] For each child, there is also an additional row for each wave. This format of *biochild.dta* is referred to as the "long-long" format (Brüderl et al. 2013: 51).

These formats have some drawbacks when they are used for joint analyses of fertility and partnership events. In *Eventhistory.dta*, we address these difficulties.
1. With their wide range of date variables, *biopart.dta* and *biochild.dta* are rather complex data sets. Thus, they have to be brought into a spell format before any kind of duration analyses can be conducted.[10] In contrast, *Eventhistory.dta* features several rows per respondent, with each referring to a specific point in time in the life of the respondent, and showing whether he or she was in a union, cohabitation, or marriage, or was a parent at the respective point in time. One advantage of this format is that it

---

[8] The exact file labels vary depending on wave, release and whether it refers to pairfam or DemoDiff data.

[9] Note that respondents were asked to report all the children the respondent ever had. These are defined as all biological children, regardless of whether the respondent ever lived with them or not, and all other children, like adoptive, foster, or stepchildren, provided the respondent has ever lived with them.

[10] Spell data sets include a separate episode (row) for each event that occurs in the respondent's biography. These "one row per event" data show all of the specific information for each defined episode in the life course of the respondent.

is easy to identify the start and end of partnerships, as well as to consider information on union or cohabitation interruptions (so-called "breaks") and overlapping partnership episodes. This information might otherwise be neglected.

2. *Biochild.dta* and *biopart.dta* are stored separately and differently. The long-long format of *biochild.dta* cannot be used directly for event history analysis. We therefore bring the fertility data from long-long into a long-wide format.[11] The advantage of using this format is that the fertility data are in the same format as the partnership data (*biopart.dta)*. This enables us to merge the fertility with the partnership histories. This step is important, as most fertility analysis is directly connected to the partnership dimension.

3. Respondents without partnership experience and childless persons are not included in *biochild.dta* or *biopart.dta*. This is a disadvantage for event history analyses because these estimations refer to a population at risk. If not the entire population is included, censored episodes cannot be taken into account appropriately. One example is the transition to the first birth. Neglecting childless people would lead to an underestimation of the amount of time that elapses until the first birth for the whole population, as censored episodes are not included in the sample. Hence, we include all of the respondents in *Eventhistory.dta*, not just the ones who have experienced an event.[12]

4. *Biochild.dta* and *biopart.dta* include imputed date information if the monthly dates are missing, but information on the year is given. These imputations cannot be identified. However, the identification of imputations is important for any kind of analysis, as the results may depend on the imputation mechanism. This appears to be especially important in the context of family dynamics, as events are closely related and often occur within a narrow time frame. We therefore flag cases in which the dates have been imputed in *biopart.dta* and *biochild.dta*. These flag variables are denoted with the prefix "IMP" and are coded as (1) in cases in which only information on the year is available, and (2) in cases in which information on the season is

---

[11] After this procedure, we have only one line for each child, and the information from different interview times is stored in separate variables.

[12] Technically, we access the original data set *anchor1.dta* and extract several variables (e.g., the dates of birth and interviews), which otherwise are available only for individuals who had ever reported having children or partnerships.

available.[13] The structure of *biopart.do* implies that these variables had to be introduced for both pairfam and DemoDiff with a modified *biopart.do* named *biopart_*_IMP.do*[14] (see page 6). Imputed date information regarding children is flagged separately for each biological child,[15] and aggregately for non-biological children (i.e., without referring to specific non-biological children).[16]

5. In *biopart.dta* and *biochild.dta*, no imputation is carried out if the year of the date is missing. These missing dates present difficulties in the event history data, which are based on date information. For example, a missing separation date or interruption date would lead us to assume that the partnership continues, because the data cannot be split at the (unknown) time of separation. We have therefore decided to recode the whole episode to missing if either the start or the end date is missing. Partnership episodes with nonexistent year information are flagged. The time-varying partnership flag variables are denoted with the prefix "FLAG" and are coded as (1) if the start or end date information is missing and (2) if the information on breaks is missing.[17] If at least one year of the birth information for the biological children is missing, the complete ids are tagged. To indicate that year of birth information on children is missing no flag variables are introduced, but the respective variables are coded as (-7) "incomplete information." We consequently set these variables for the whole id, and not just for specific episodes, to (-7). For a full list of the respective variables, see Table 3 in Chapter 4.

---

[13] *Biopart.dta* offers several flag variables that mark inconsistencies in the partnership history. These variables may tag inconsistencies that caused the imputation procedure because they are generated after the random imputation process. We therefore checked the relevant dates if an imputation took place and recoded the respective dates.

[14] Missing month information only occurs in wave 1. Starting with wave 2, respondents could provide only concrete dates. This difference is due to the different interview methods in waves 1 and 2.

[15] Specific variables indicate imputations of the date of birth (IMP_dobbiok{1-10}) and death (IMP_dodbiok{1-10}) of a biological child, as well as imputations of the beginning and the end of co-residence (IMP_beglivbiok{1-10}, IMP_endlivbiok{1-10}).

[16] IMP_beglivnonbiok, IMP_endlivnonbiok.

[17] The time-varying flag variable FLAG_M_UNION marks the episode in which the partnership might have taken place. If the start or end of a union is missing, we flag the period between the known dates with (1) if the missing partnership episode is of a higher order, and from birth onwards if information on the first union is missing. If a union interruption is missing, we flag the respective union with (2). A missing cohabitation episode is flagged for the respective union episode, if it is known (FLAG_M_COHAB). Otherwise, we flag the missing cohabitation for the same episode as the missing union episode. The same strategy is also applied to missing marriage episodes (FLAG_M_MARR).

Note to the user: Comparing the datasets by an example id

An example id in the appended Excel document "Eventhistory_Example_id.xlsx" shows in what form the biography information is available in *biopart.dta* and *biochild.dta* (Table I and Table II) and how it is available in *Eventhistory.dta* (Table III). The variables are shaded in different colors, which facilitates a comparison of the data structures.

3 Generation of the event history data

In the following, we describe in detail how we generate *Eventhistory.dta* using several data files provided by the German Family Panel data group (see Table 1). Basically this data bases on release 3.1 (pairfam) and release 2.0 (DemoDiff).

The essential feature of an event history data set is the time-varying information, which is generated by episode splitting. This "one row per event" format is also known as the "spell data" format. We set the start of the processing time at the birth of the anchor person. As events can only be identified in episodes after the event actually occurred, events that happened in the month of the interview would not be able to be identified if we were to end the processing time in the month of the interview. Thus, we censor episodes one month *after* the date of the last available interview, which allows us to account for events that happened in the month of the interview. The disadvantage of this procedure is that censored episodes may be overestimated by up to one month each.

We describe the editing separately for general information, (3.1) information on the partnership (3.2.), and information on the fertility (3.3) history, as specific strategies apply in each case.

3.1 General time-varying variables

We include in the data two general time-varying variables: one that marks the age of the person, and one that marks the timing of the interviews in the histories.

Age of anchor

The age variable (AGEANC) shows the age (in single years) of the respondent starting from age 14.

Timing of interviews

The second general time-varying variable marks the timing of the interviews within the individual biographies (INT). INT splits the episode before and after an interview, showing the month of each interview as a separate episode. This makes it easy to match prospective, wave-specific information of the German Family Panel to the relevant episodes. For example, responses from the questionnaires of the anchor, partner, child, or parents can be matched with the information of the respondent to provide a dyadic perspective. Also, it is important to note that all panels, including the German Family Panel, suffer from panel attrition. For several reasons (for example selectivity issues), it might be useful to only include information from specific waves in the analysis. This can be easily done with the variable INT.

3.2 Partnership biography

In this section, we describe the union, cohabitation, and marriage information in *Eventhistory.dta*. First, we explain the structure of the partnership episodes (3.2.1). Second, we focus on the order variables implemented in *Eventhistory.dta* (3.2.2). Third, we present any additional information that is given on partnership episodes in *Eventhistory.dta* (3.2.3).

3.2.1 Partnership episodes

Pairfam includes information on the partnership episodes of the anchor, and considers eventual temporary disruptions and later reunions with the same partner. *Biopart.dta* draws on this information by defining the earliest start, most recent ending, and possible interruption dates of the relationships with each partner of the anchor person. Within this concept, the focus is on the earliest start and the most recent ending of a partnership. The information on interruptions (start date and end date of the first break, the second break, the third break, etc.) is stored in separate variables, and is therefore likely to be neglected in analyses. We have decided not to distinguish between temporal and final disruptions in the *central* partnership variables in *Eventhistory.dta*, as this means that partnership durations are less likely to be overestimated than with the *biopart.dta* structure. Referring to the different partnership dimensions in pairfam, we include time-varying partnership variables that indicate whether the respondent was in a union, a cohabitation, or a marriage at the respective points in time.[18] Because we are using the event history approach, we consider only episodes in which both the start *and* the end date are available. Episodes with missing information are flagged (FLAG*) (see Section 2.2). The information on whether and how long the respondent was in a union is stored in a single variable (UNION), which distinguishes between (0) "no union" and (1) "in union." This variable takes the earliest start date and the latest end date of each union, *as well as* episodes of union interruptions into account. The existence of a union break is indicated by the variable UBREAKORDER. The variable COHAB distinguishes between (0) "not in cohabitation" and (1) "cohabiting". Episodes in which the cohabitation is temporarily disrupted are indicated by CBREAKORDER. The variable MARR has the values (0) "single," (1) "married," or (2) "divorced." There is no information on repeated marriage episodes with the same partner in the retrospective data. The structure is visualized in the Example id (*Eventhistory_Example_id.xlsx*).

Inconsistent partnership episodes

Biopart.dta offers four flag variables that mark inconsistencies in the partnership biography. As these consistencies never occur simultaneously in a single episode, we aggregated them in one time-varying variable (BIOPARTFLAG). Inconsistencies are coded to (1) if the marriage starts earlier than the partnership, which refers to the flag variable *biopartflag1* in *biopart.dta*. In *biopart.dta*, the flag variable *biopartflag2* identifies overlapping cohabitation episodes with different partners. In *Eventhistory.dta*, we do not consider such episodes as inconsistent per se, but include variables that indicate the union or cohabitation order of the simultaneous partner in case of overlapping episodes (UNIONORDER_SIM and COHABORDER_SIM, see Section 3.2.2). With these variables in place, the flag variable *biopartflag2* is no longer needed. We have therefore decided to drop this information. Corresponding to

[18] Please note that we first had to split each partnership episode, with reference to the duration variables, to generate time-varying partnership variables for each episode. The information of the respective partnership episodes is then concentrated into a few central variables.

the variables *biopartflag3* and *biopartflag4* in *biopart.dta*, the time-varying flag variable BIOPARTFLAG in *Eventhistory.dta* has the value (3) if the current marriage started before the previous marriage had been terminated, and the value (4) if the age of the partner had been misreported (being less than 10 years old). We further sought to ensure that inconsistencies are not artificially created in the imputation process of missing month information. With the help of our generated flag variables (IMP*), which mark random imputations in the date variables, we identified these episodes and recoded the respective start and end dates in a manner that eliminated overlapping episodes.[19] Consequently, these episodes are not marked as inconsistent.

3.2.2 Order variables

Biopart.dta includes a variable that gives information about the ordering of the different partners according to the start date of the partnerships ("index"). If the start date of a partnership is missing, this variable relies upon the order in which the partnerships were listed during the interview, and assumes this to be the chronologically correct ordering of partnerships. This "index" variable refers to the union dimension. We rely upon this index variable and generate a variable that indicates in a time-varying manner the order of the union partner. UNIONORDER shows the respective order number or has the value (0) "no partner" if no union is ongoing in the respective episode. Additionally, we provide information about the order number of the union partner with whom the respondent cohabited or was married to by the variables UNIONORDER_COHAB and UNIONORDER_MARR. The example id (see Table III in the appended Excel document *Eventhistory_Example_id.xlsx*) illustrates what the variables look like.

As further order variables, we include information about the order of cohabitations and marriages. This information is, for example, essential to an analysis that is restricted to first cohabitations or first marriages. Therefore, we construct indexes for cohabitation and marriage by ordering the data according to the cohabitation and the marriage histories, respectively.[20] In *Eventhistory.dta* these index variables are time-

---

[19] We assume that this is more likely. In any case, all of the episodes based on imputed dates can be identified with our provided flag variables.

[20] While episodes with start dates can be sorted correctly, episodes with missing start dates are problematic because they cannot be taken into account easily. If, for example, the date of the formation of the first household is missing, but the date of the formation of the second household exists, it is possible to assume that the second formation is the first unless effort is applied to sorting

varying and labeled COHABORDER and MARRORDER. They refer to the order of domestic partners and spouses (as can be seen from the Example id in *Eventhistory_Example_id.xlsx*.).

Some of the respondents reported having overlapping union and cohabitation episodes. Overlaps can occur during the transition to a new partnership (a partnership starts before the previous one has ended), but can also take place within a partnership (a partnership starts and ends while the previous one lasts). Thus, the variables UNIONORDER_SIM and COHABORDER_SIM indicate overlapping partnership episodes. The main order variables (UNIONORDER, COHABORDER) show the order number of the new partner, while the order number of the previous partner is shown in the UNIONORDER_SIM/ COHABORDER_SIM variable. To ensure that these overlaps are not produced by the random imputation procedure, we checked cases with overlaps and imputations. Episodes are recoded if they have overlaps of less than four months and only information about the season is available, or if they have overlaps of up to 12 months and only yearly information is given.

3.2.3 Further partnership information

Homosexuality – a partnership dimension

We define homosexuality as a partnership dimension that may vary across different partnerships. The respective variable HOMOSEX indicates for each episode whether the respondent lives in an opposite-sex union (1), in a same-sex union (2), or in no union at all (0).

Biopart.dta offers a variable that indicates homosexuality as a time-constant trait.[21] In contrast, we assume homosexuality to be an individual characteristic that may vary across time.[22]

---

out the episodes with the missing dates. If the start dates are missing, we assume that the order in which the cohabitations or marriages were reported in the interview is correct. To consider this order, we ascribed an imputed start date to episodes with missing start dates. Note that the sole purpose of the imputed values is to make the respective episode count, and that it will be recoded to missing afterwards. We imputed the start of the cohabitation or marriage using the date of union formation, if it was not missing as well. Otherwise, we recoded the missing cohabitation start date to the cohabitation end date. If the required information was missing, we ordered the missing dates according to the "index" category.

[21] The syntax used to compute the variable is contained in the Stata do file *homosex.do*. The information on homosexuality is taken from the anchor interview, combining information from waves 1 and 2. The anchor is defined as being homosexual if he or she had a same-sex partner in

In order to exemplify the benefits of the variable HOMOSEX, we show the union trajectories of two example ids (Table 1). The original variable "homosex" in *biopart.dta* categorizes the first example id constantly as homosexual because she is living in a same-sex partnership at the time of interview, even though in the past she had a heterosexual relationship. The second example id has no partnership at the time of interview, but had same-sex as well as opposite-sex unions in the past. As the question about sexual orientation was not answered in the questionnaire by this respondent, the original variable "homosex" in *biopart.dta* is marked as missing, in contrast to the new variable HOMOSEX. Thus, HOMOSEX emphasizes the sexual *practice* of a person.

Id	*Start (_t0)*	*End (_t)*	*UNION-ORDER*	*HOMOSEX*	*For comparison: homosex (by pairfam group)*
715391000	*0*	*221*	*0*	*"no relationship"*	*"homosexual"*
	221	*238*	*1st partner*	*"heterosexual relationship"*	*"homosexual"*
	238	*280*	*0*	*"no relationship"*	*"homosexual"*
	280	*320*	*2nd partner*	*"homosexual relationship"*	*"homosexual"*
616520000	*0*	*173*	*0*	*"no relationship"*	*"incomplete data"*
	173	*213*	*1st partner*	*"heterosexual relationship"*	*"incomplete data"*
	213	*219*	*0*	*"no relationship"*	*"incomplete data"*
	219	*241*	*2nd partner*	*"homosexual relationship"*	*"incomplete data"*
	241	*280*	*0*	*"no relationship"*	*"incomplete data"*
	280	*336*	*3rd partner*	*"heterosexual relationship"*	*"incomplete data"*
	336	*337*	*0*	*"no relationship"*	*"incomplete data"*

Table 2: Definition of homosexual and heterosexual partnerships in Eventhistory.dta

Marriage ceremony

MARCER shows the type of wedding ceremony for each marriage while it lasts. During unmarried episodes, this variable has the value (-3) "does not apply". Analogous to *biopart.dta*, we distinguish between having had a civil ceremony (1), a religious and a civil ceremony (2), or only a religious ceremony (3). Please note that the data include the date of marriage formation, but we do not know whether this information refers to the religious or the civil ceremony.

wave 2. If the anchor person did not have a same-sex partner in wave 2, information from wave 1 (anchor's reported homosexual preference or a same-sex partnership) was added (Brüderl et al. 2011).

[22] We therefore decided to drop the "homosex" variable offered in *biopart.dta*.

Death of a partner

DEADPARTNER shows in each time period whether the respondent experienced a partnership that ended through the death of a partner. The variable is either zero for "no death of partner," or shows the order number of the partner who died. If the respondent did not remember the year of death of one partner, the variable has the value (-7) "incomplete information" for the whole id.

Matching prospective partner information

There is no information on the characteristics of previous partners in the German Family Panel[23]. However, pairfam offers rich information on the partnerships that are ongoing at the time of an interview. Some information is collected through the anchor person and some through the partner questionnaire. We decided to include the partner id in *Eventhistory.dta* to simplify the matching with the partner data. We included this information in a time-varying way: whenever the anchor person shows an episode in which he or she is in a relationship with a partner who is the current partner in wave 1 or 2, this episode is assigned the partner id of this partner (PID). This is illustrated in the example id's history (Table III in the appended Excel document *Eventhistory_Example_id.xlsx*). Thus, it is easy to identify for which relationship episodes dyadic analyses can be conducted.

3.3 Fertility biography

In this section, we first provide details on the ordering of children in *Eventhistory.dta*. Second, we explain the generation and content of the variables on the actual fertility history (3.3.2), on episodes of co-residence with children (3.3.3), and on further variables regarding the children of the anchor person (3.3.4).

3.3.1 Ordering of children

We order all biological children according to their dates of birth. The birth dates of non-biological children are not considered in the ordering procedure, as we

[23] The only information provided is the sex and the relationship of the partner to the children of the anchor person.

concentrate on the fertility behavior of the respondents.[24] Thus, order-specific information is available for all biological children. The respective variables carry the letters "**BIOK**" for "**bio**logical **k**id" in their name.

Thus, the order of children in *Eventhistory.dta* deviates from the ordering in *biochild.dta*, in which non-biological children are also taken into account when ordering the children. The reordering of children makes it necessary to include a variable that indicates the original number of each biological child, as it is stated in the anchor data set (NUMBERBIOK{1-10}). This time-varying variable works as an identifier and allows to match further child-specific information from other pairfam data sets.

Second, if respondents have children with missing year-of-birth information, these children cannot be ordered. The denoting of childless episodes and the correct ordering of biological children are therefore not possible in these cases[25]. Thus, if at least one of the biological children has a missing year of birth, the whole fertility biography of the respondent is flagged as incomplete (see section 2.2). These respondents should not be considered in family-related analyses. The respective temporary variable "FLAG_M_bio_dobk" indicates incomplete fertility biographies. In the final *Eventhistory.dta*, several variables have the value (-7) "incomplete information" whenever the fertility biography of the person is incomplete (see Table 8).

3.3.2 Fertility episodes

Fertility episodes refer to the timing of the births of biological children. Thus, our central fertility variable is the age of the biological children of the respondent. AGEBIOK{1-10} shows the age of the respective child in each episode, starting nine months prior to the date of birth (pregnant (1))[26]. Thereafter, the episodes are split at

[24] For practical reasons, we have assigned non-biological children imputed birth dates that order them after all of the biological children. These dates are later recoded to -3.

[25] In contrast, in *biochild.dta* children with missing information on their year of birth are placed after all of the children with known dates of birth. The advantage of this procedure is that other child-specific information still is available, even though the year of birth is not.

Fertility analysis is usually strongly dependent on the reliable ordering of children. This is why we have decided that it would not be appropriate to randomly choose an adapted ordering or to rely on the order in which the children were reported during the interview, as we did in the case of missing dates in the partnership biography.

[26] The duration of pregnancy is a proxy (nine months prior to date of birth) and does not rely on any stated date of conception by the respondent. As pregnancies sometimes end earlier than after nine months, the duration is overestimated in this data set. The lack of exact information of conceptions

each birthday of the child (zero years old (2), one year old (3), and so on). Childless episodes are coded as (0). Respondents with missing birth year information in any of the children's years of birth are coded as (-7) "incomplete information." We provide a further variable that indicates the age of the youngest child of the respondent for each episode (AGEBIOK_YNG).

3.3.3 Episodes of living with children

Information on the episodes during which the anchor person lived or did not live with children is surveyed differently in waves 1 and 2. In wave 1, respondents were asked to list all previous episodes of co-residence if they were not living with the respective child at the time of interview. If the respondent and his/her child co-resided at the time of interview, only the starting date of co-residence was surveyed. That means that interruption dates were not reported. This results in disparate retrospective data on co-residence with children.

Starting with wave 2, no information on the dates when children move in or out is being collected. Instead, the survey only shows whether children are *currently* living in the anchor's household. Apart from that, no information on cohabitation episodes of dead children is included in *biochild.dta*. Due to these restrictions, we had to make several assumptions about the cohabitation history of the anchor with biological and non-biological children.[27]

---

further leads to the fact that we only consider terminated pregnancies. If respondents state that they are pregnant during the latest interview this is not considered in *Eventhistory.dta.*

27

1. Cohabitation breaks with children are only available for a selective group of respondents, which is why we only consider the first reported episode of living together with each child of all of the respondents.
2. Deceased children were living with the anchor person from the date of birth until the date of death.
3. Children who were living with the anchor person in wave 1 had not moved out between the date when they first moved in and the date of the first interview.
4. Children who were not living with the anchor person in wave 1 had lived with the anchor person only once before the first interview. After wave 1, prospective data provide information about further episodes of co-residing with children.
5. Respondents with children born after wave 1 who were living with these children at the time of the second interview had been living with these children since birth.
6. Respondents with children born after wave 1 who were not co-residing with these children in wave 2 had never co-resided with these children.
7. Non-biological children who were first reported in the second interview and who were living with the respondent at the time of the second interview had moved in with the respondent in the month of second interview.

Based on these assumptions (and reported dates), LIVBIOK{1-10} shows for each episode whether the respondent is living with a specific biological child (1) or not (0). Respondents with missing information on the year of the beginning or ending of co-residence are coded as "incomplete information" (-7).
LIVKIDS is the only variable that contains information on non-biological children. It shows for each episode, independent of the order of the children, whether the respondent shares a household with biological children only (1), with non-biological children only (2), with both biological and non-biological children (3), or with no children at all (0). This variable also shows whether there is missing information on co-residence with children (-7). The example id (see Table III in the appended Excel document *Eventhistory_Example_id.xlsx*) illustrates what the variables on co-residence with biological and non-biological children look like.

3.3.4 Further information

Death of a child
DEADBIOK shows in each time period whether the respondent experienced a death of a biological child. The variable is either (0) "no child died" or shows the order number of the biological child who died. If the respondent does not remember the year of the death of a child, the variable has the value (-7) "incomplete information" for the whole id.

Sex of a child
BIOSEXK{1-10} shows the sex of each biological child from the time the child was conceived (nine months prior to birth).

Order of surveyed child
As described, the ordering of children in *Eventhistory.dta* deviates from the ordering in *biochild.dta* (see Section 3.2.2). This means that we have to adapt the categories of the variable that shows the order number of the surveyed child (named *surveykid* in

8. Children who had moved out after the first interview moved out in the month of the second interview.

biochild.dta). CAPIBIOK shows the *Eventhistory.dta* order of the biological child, which was surveyed via the children's questionnaire. The variable has the value (-3) "does not apply" in pre-conception episodes, and the value (-7) "incomplete data" if the fertility biography is incomplete.

Partner order of second biological parent

The variable pno in *biochild.dta* is recoded because the categories deviate from the partner ordering in *Eventhistory.dta.*[28] The variable indicates the partner who is the second biological parent. In *Eventhistory.dta*, the variable UNIONORDER_BIOK{1-10} shows for each biological child the partner order number of the second biological parent. If the second biological parent is not reported in the partnership history, this is marked as (97) "another person." Again the variable has the value (-3) "does not apply" in episodes in which the child has not yet been conceived, and the value (-7) "incomplete data" if the fertility biography is incomplete. For the example id (Table III in the appended Excel document *Eventhistory_Example_id.xlsx*), we see that UNIONORDER_BIOK1 shows that the third partner is the second biological parent of the first biological child.[29]

[28] In pairfam has the value zero when the current partner of wave 1 is the second biological parent of a respective child. Starting with wave 2, the current partner who had also been the current partner in the previous wave is assigned the number one; the current new partner, the number two; and partners who had been partners between two interviews are assigned the numbers three, four, five, etc. (Brüderl et al. 2013). These differences in coding between waves 1 and 2 also appear to suggest that a lot of respondents reported in wave 2 that none of the stated partners, but rather "another person," is the second biological parent of the child. In fact, the response of "another person" in wave 2 means that no current partner or partner from a relationship that took place between waves 1 and 2 is the second biological parent. In *Eventhistory.dta*, we show one consistent variable that has the running partner number for the second parent. However, for children born before wave 1, we only account for the information given in wave 1. Information on the second biological parent given in wave 2 is only used for children born after wave 1. Thus, we do not consider revisions of the respondents regarding this information.

[29] Incidentally, the example id (see Excel table *Eventhistory_Example_id_xlsx* in the appendix) is one of the rare cases that involve a revision of the respondent between the two waves (see Footnote 24). For wave 1, it is reported that "the current partner in wave 1" is the second biological parent of the first child. In contrast, it is reported for wave 2 that "another person" is the second parent, even though the respondent's current partner in wave 1 was still the current partner of the respondent (see Table II). In order to limit the coding complexity, we accept this kind of potential misreporting and assume that the third partner is the second biological parent, as was stated in wave 1 (see Table III).

4 Recommendations and summary

4.1 Recommendations

In Table 3 we list the generated flag variables, and provide recommendations about whether to drop or to keep flagged individuals or episodes. An episode or respondent is either flagged by a separate flag variable or by a separate category in a respective event variable. In the latter case, we follow the flagging strategy of *biopart.dta* and *biochild.dta* and use (-7) as the flag category for missing information. We distinguish three different kinds of flagged information:

First, variables can mark missing information. We distinguish between completely missing dates and missing dates that have been imputed in *biochild.dta* or *biopart.dta*. If the information on the date of birth is completely missing, not only are selected episodes flagged, but all of the parity-specific information is marked as (-7) “incomplete information,” as in these cases no correct ordering of children is possible. If the year of co-residence with a specific child is missing, the co-residential biography with the respective child is coded as (-7). If the date of death of any child or any partner is missing, the respective variables (DEADBIOK and DEADPARTNER) are set to (-7) for the whole id. If at least the year of an event is available, the date is imputed. Variables that mark such imputed missing information have the prefix “IMP” and distinguish between whether (1) “only season information” or (2) “only year information” was given in the interview.

Second, we flag inconsistencies in the data. BIOPARTFLAG refers to different inconsistency flags of *biopart.dta*. It marks episodes that are probably misstated. *Biopart.dta* also defines cohabitation episode overlaps as inconsistent. We mark these episodes with the variable COHABORDER_SIM. The variable shows not only whether an overlap exists, but also with whom the respondent reported cohabiting simultaneously. As it is not clear whether these overlaps are due to a misstatement of the dates, we do not define them as inconsistent per se. Similarly, partnership episodes may overlap, which is indicated in the variable UNIONORDER_SIM.

Third, the flag variables mark episodes that can contradict general assumptions about fertility and partnership behavior. Same-sex unions should be dropped if theoretical assumptions rely on opposite-sex unions. The death of a partner can be mistakenly interpreted as a separation if the respective flag variable DEADPARTNER is not taken into account. The variables that show the age and co-residence of a child do not mark the child's death. Thus, we would have assumed the child was getting older if we had not considered the flag variable DEADBIOK. Similarly, we might have incorrectly assumed that co-residence with a child ended with the child moving out, instead of with his or her death. Thus, DEADBIOK needs to be taken into account in the respective analyses.

Table 3 is constructed in the following way. In the first column, the variable's label or the relevant category is listed. The second column defines what is flagged by the respective variables or categories. The third column contains an explanation of the consequences if the flag variable is ignored. In the fourth column, a recommendation is provided for cases in which the flagged information is sensitive for the analysis.

Variable name	**What they mark**	**Consequence**	**Recommendation**
FLAG_M_UNION, FLAG_M_COHAB, FLAG_M_MARR	A missing partnership episode in the data	A partnership episode is mistakenly reported as partnerless/not cohabiting/not married.	Drop episodes/ids in analyses that refer to the partnership status at a single point in time, e.g., at childbearing.
AGEBIOK{1-10}==-7 AGEBIOK_YNG{1-10}==-7	Fertility history is missing because the year of birth of at least one biological child is not known.	The fertility history of the respective id cannot be used.	Drop id for fertility analyses.
SEXBIOK{1-10}==-7, NUMBERBIOK{1-10}==-7, UNIONORDER_BIOK{1-10}==-7 CAPIBIOK==-7	Information on the child is missing because the fertility history is missing	Information on the child cannot be used.	Drop id for child-related analyses.
DEADBIOK==-7	Information on the year of death for any biological child is not known.	It is not possible to determine whether all of the children are alive.	Drop ids for analyses in which it is relevant to know whether the children are alive.
DEADPARTNER==-7	Information on the year of death of the partner is not known.	It is not possible to determine whether the dissolution was due to separation or to the death of one of the partners.	Drop episodes for the separation analyses.
LIVBIOK{1-10}==-7	Information on the year of co-residence with a specific biological child is not known.	The co-residence history of specific biological child and anchor is not clear.	Drop ids in analyses that refer to co-residence with specific biological child.
LIVKIDS==-7	Information on the year of co-residence with at least one	The co-residence history with children is not clear.	Drop ids in analyses that refer to the co-residence with

	child (biological or not) is not known.		children.
IMP_UNION, IMP_SEP, IMP_COHAB, IMP_COHABend, IMP_MARR, IMP_DIV IMP_dobbiok{1-10} IMP_dodbiok{1-10} IMP_beglivbiok{1-10} IMP_endlivbiok{1-10} IMP_beglivnonbiok IMP_endlivnonbiok	An imputed month in the date	The date is imprecise because it refers to information on the season or year.	Drop episodes/ids in analyses that refer to precise dates, e.g., the timing of marriage relative to childbearing.
BIOPARTFLAG==1 | BIOPARTFLAG==3	Inconsistencies in the marriage history	The duration of marriage is probably misstated.	Drop episodes in analyses that refer to marriage.
BIOPARTFLAG==4	Inconsistencies in the birth date the of partner	Wrong age of the partner	Drop the episode in analyses that refer to the partner's birth date.
UNIONORDER_SIM	Union overlap	The union duration of the previous partnership is underestimated.	Decide which partnership should be followed.
COHABORDER_SIM	Cohabitation overlap	The cohabitation duration of the previous cohabitation is underestimated.	Decide which cohabitation should be followed.
HOMOSEX==2	Same-sex unions	A partnership may be mistakenly assumed to be opposite-sex.	Drop episodes in analyses that refer to opposite-sex assumptions.
DEADPARTNER=={1-?}	Respective partner died	It is mistakenly assumed that the partnership ended by separation.	Drop episodes for separation analyses.
DEADBIOK=={1-10}	Respective child died	It is mistakenly assumed that all of the biological children are alive.	Drop ids for analyses in which it is relevant that the children are alive.

Table 3: Overview of flag variables and possible applications

The fourth column further mentions whether the respective episode or the whole individual should be dropped.[30]

The problematic episode can be excluded from the analysis by dropping the flagged episode. For variables that mark imputed dates in the union or cohabitation biography, it is important to consider the break information because imputations are not flagged during union or cohabitation breaks.

[30] The problematic individual can be excluded from analysis by using the following command:

```
.     Sort id
.     by id: egen NewVariable=max(FlagVariable)
.     drop if NewVariable=[problematic value]
```

For more information, please see the do files of the example analyses (especially *Eventhistory_Example_analysis2.do*).

4.2 Summary

With *Eventhistory.do*, we provide a syntax that facilitates the use of the rich biographic information in pairfam. The attached STATA do file *Eventhistory.do* enables the pairfam user to generate the event history data set *Eventhistory.dta*, which contains the fertility and partnership biographies of the first three waves of the German Family Panel pairfam (release 3.1) and the first two waves of DemoDiff (release 2.0). Referring to the original data sets and the syntaxes, the do file *Eventhistory.do* transforms the fertility and partnership information into spell data. The structure of *Eventhistory.dta* is "one row per event." All of the relevant information is considered in time-varying variables. *Eventhistory.dta* also includes individuals without children or partnership experience in the data - the so-called "risk population". Furthermore, *Eventhistory.dta* enables the pairfam user to identify date information that was imputed in *biochild.dta* or *biopart.dta*.

Thus, the data are very flexible and ready to be used. The pairfam user can easily conduct empirical analyses on a wide range of research topics. With *Eventhistory.dta* fertility and partnership behavior can be easily explored by applying empirical methods like event history or sequence analyses. *Eventhistory.dta* might not be appropriate for every research question concerning fertility and partnership behavior because information on specific children is only available for biological children of the anchor person.

Eventhistory.dta can be matched to further information on the anchor person, his or her children, his or her parents, and his or her partners by their respective personal identifiers.

Eventhistory.dta was developed as part of the authors' dissertation. We plan to include in a future version information on the employment biographies of the respondents. The use of data generated by *Eventhistory.do* should be indicated in your work by citing this report. We accept no responsibility for errors that may have arisen during the coding procedures.

Please contact the authors if you have any questions.

5 Examples of analyses

In order to illustrate how *Eventhistory.dta* may be used, we provide two examples of event history analyses.[31] The first example (see *Eventhistory_Example_analysis1.do*) describes the transition to the first union. The process of first union formation is a central event during adolescence for young men and women. We show in Kaplan-Meier survival estimates the percentage of males and females who experience a first union between the ages of 14 and 24. In this example, we draw special attention to the use of flag variables.

The second example (see *Eventhistory_Example_analysis2.do*) focuses on the transition to a partnership separation after the first child is born. The stability of couples with children is of central concern because it has a strong impact on the living conditions of parents and children. The subject of the analyses requires that the population at risk is restricted to parents; childless episodes are excluded. We consider only individuals who were in a union at the time they had their first child, because they make up the population who are at direct risk of separation after childbirth. In addition to these basic and necessary restrictions, we can apply more rigorous restrictions. The usefulness of such restrictions depends on the specific research question and its theoretical framing. In this example, we want to emphasize different levels of analysis restrictions. Again, Kaplan-Meier survival estimates are shown, and time-varying effects are also visualized. In this example, we also show the options offered by *Eventhistory.dta* for selecting specific populations for analyses.

The STATA do files of both examples are appended to the technical report.

5.1 Transition to the first union

To model the transition to the first union (see the appended file *Eventhistory_Example_analysis1.do*) we use *Eventhistory.dta*. As a first step, we define the start of the process time. We want to start the modeling with age 14, but we

[31] We cannot provide a full explanation of applied event history methods here. For an introduction to event history techniques, please see Blossfeld/Golsch/Rohwer (2007).

also want to consider union experience before age 14. We therefore assign a very short process time to respondents who had already had a union before age 14. Afterwards, the episodes prior to age 14 are dropped. This allows us to distinguish sex differences at the initial level from those that occur after age 14 in the observation period.

We then adjust the survival time variables START and END to the observation start (age 14). The event is defined as the date of the formation of the first union. We drop episodes after the transition to the first union, because the respective persons are then no longer at risk. Additionally, we drop episodes of higher order unions. The episode is censored if the person has not formed a union by age 24. Further, it is censored at the time of the second interview (or at the time of the first interview if the person did not participate in wave 2).

Figure 2 shows the results of the Kaplan-Meier survival estimates for males and females. The results of the dashed lines include persons with a missing union formation date. These are considered as censored episodes. Individuals with a missing first union are included in the data until either the time of the interview, age 24, or the formation of a subsequent union. As a consequence, the survival curve may be overestimated. The continuous lines show the transition to the first union only for persons with known dates; that is, after the individuals with missing first union formation dates have been dropped. We see that the omission of these individuals leads to a slightly lower survival curve.

Women and men show similar shares of first unions before age 14. However, women between the ages of 16 and 19 are more likely than men to transition into a first union. Men catch up later, but still lag behind slightly at age 24.

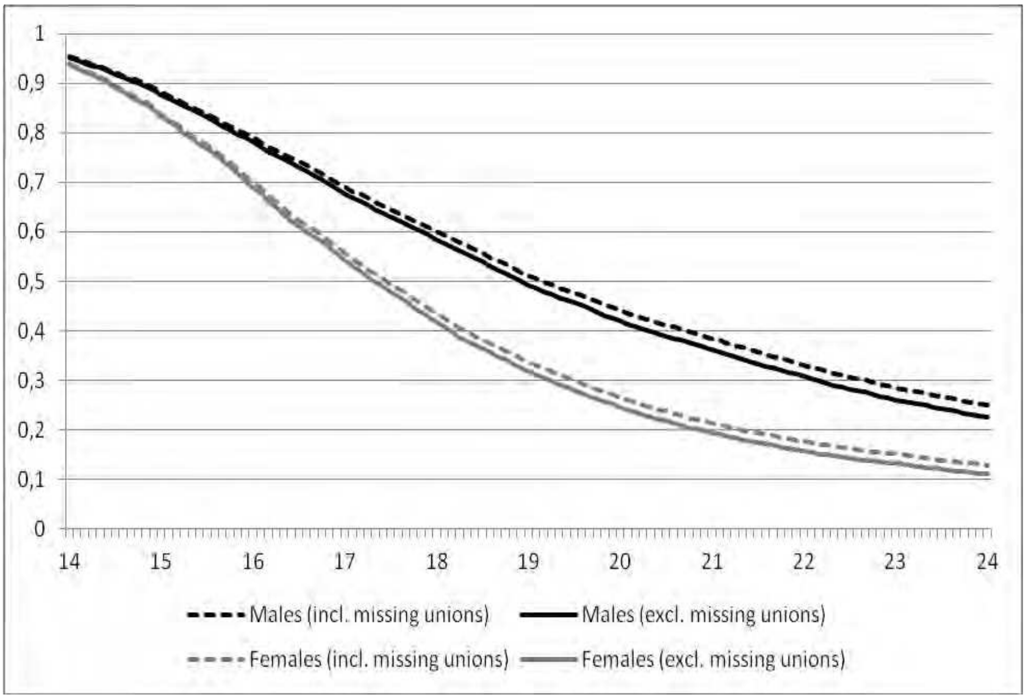

Figure 2: Transition to the first union, age 14 to age 24, Kaplan-Meier survival estimates

5.2 Transition to a separation after the first childbirth

In this example (see *Eventhistory_Example_analysis2.do*), we again use *Eventhistory.dta*. We generate a variable that indicates whether a person has a partner at the time his or her first child is born. We keep only persons who are at risk of experiencing the event in question (basic restrictions).

In the second step, we seek to clean our sample of potential inconsistencies due to the random imputation of missing months. A missing month is imputed if the date information is restricted to the season or the year of the event. Imputations have been flagged. These imputed dates can influence our results because the union status at the first birth may be unclear. A very rigorous option would be to drop all persons with imputed months from the observation. But this strategy could lead to a bias in the results, as it is possible that separated people, in particular, did not give exact monthly information about their former partner. We would then have overestimated the stability of the partnership. Alternatively, we can check whether the imputed birth date of the first child and the union formation and separation dates occurred in a time range that makes it likely that the union status at birth is not clear. We have chosen this option because it minimizes the possibility of a bias arising in the results. To check the time range, we generate flag variables that indicate the union duration *until* the first birth and the union duration *after* the first birth. Individuals should be dropped from analysis if the time range is less than 12 months in imputed cases in which only information on the year is available, and if the time range is less than four months in cases in which information on the season is also available (Please see the attached file *Eventhistory_Example_analysis2.do* for more detailed information).

In the third step, we outline some of the restrictions that might be necessary, depending on the research question.

1. SEX: Fertility analyses are usually concentrated on women because their fertility history is assumed to be more reliable. If you wish to restrict the analysis to females, males can be dropped.
2. AGE AT FIRST BIRTH: A very young age at first birth can refer to a selective life course. Furthermore, outliers can bias the model results. It might therefore be useful to restrict the population at risk to a defined time frame. We have decided to drop individuals who were under age 18 when they had their first child.
3. HOUSEHOLD COMPOSITION: Some theoretical frameworks rely on the household dimension when defining the family, while other focus on the "classical" family, which consists of a co-residing biological family. With our data set, it is possible to identify individuals who
   - do not co-reside with their child or their partner,
   - also live with non-biological children in the household, and
   - have a partner who is not the second biological parent of the child.
4. HOMOSEXUAL PARTNERSHIPS: The family formation patterns of people with a same-sex partner differ from those of opposite-sex couples. It can be assumed that their family life is selected, which might affect partnership stability.
5. MULTIPLE PARTNERSHIPS: Some people have simultaneous partnership episodes with different partners. These multiple partnerships should be considered for analysis. Selected individuals reported having more than one partner when they had their first child. We have decided to drop these persons from the sample.

These aspects represent only a selection of possible forms of information that may be considered for sample restriction. They rely on *Eventhistory.dta*, but it is possible to take into account information from the anchor, partner, or child data by merging the data sets through the respective id variables. We demonstrate the merging strategy by adding the anchor data to the event history data set. We have decided to keep just the information on the birth cohorts in the data, but in general it is possible to include other time-constant determinants in the manner presented. Members of the youngest cohort were born between 1991 and 1993. We have decided to drop this cohort because of their young ages. In the final sample, we have 2,031 mothers and 304 separations. The total analysis time at risk amounts to 133,776.

After the restriction procedure, we adjust the survival time variables START and END to the observation start (date of birth of the first child). The event is defined as the date of the first union dissolution after childbirth. We drop episodes after the transition to a separation because the respective persons are then no longer at risk. The event is censored if the person does not experience a union dissolution by the time the child reaches age eight. Further, it is censored at the time of the second interview (or at the time of the first interview if the person did not participate in wave 2) or in case of the partner's death.

Figure 3 shows the transition to a separation after the first childbirth for the period from the birth of the first child until age eight in the Kaplan-Meier survival estimates. These estimates show the proportion of the women who remain partnered during the observation period. We distinguish between women who were married when they had their first child and those who were not. The results show that women who were married when they had their first child were less likely to have experienced a separation: 14 percent of the married women separated from their partners in the first eight years after they gave birth, compared to 35 percent of unmarried women.

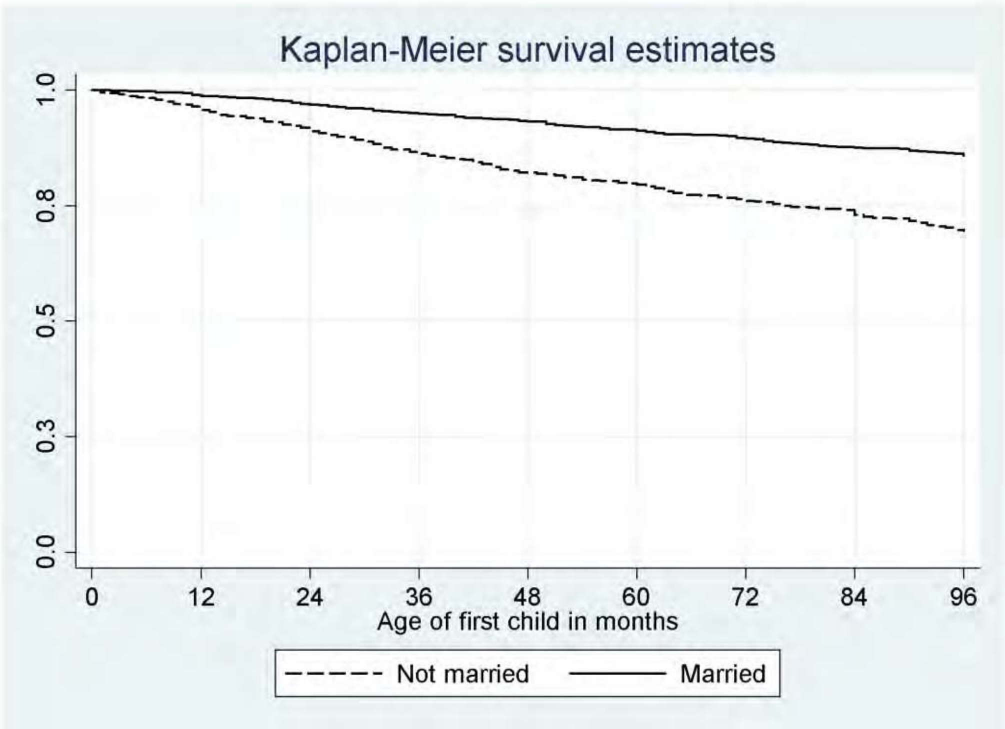

Figure 3: Transition to separation after first childbirth, from birth to age 8 (first child), Kaplan-Meier-Survival estimates, by marital status (time of first birth).

The illustration with Kaplan-Meier estimates is restricted to time-constant covariates. Figure 4 therefore refers to the marital status at a single time point: the time of the first childbirth. But it is possible to question how marital status affects the risk of separation by also considering marriages that took place after a couple's first child was born. We illustrate the effect of marital status if it is considered time-varyingly in a piecewise constant exponential model (further information is available from the file *Eventhistory_Example_analysis2.do*).

Figure 4 shows that, when marriages after the birth of the first child are also considered, being married has a positive effect on union stability compared to being unmarried. The risk of separation remains stable for married women in the observation period. Being unmarried is related to a considerable decrease in stability, except in the fourth to fifth year after family formation. Thus, as differences in the level of union stability by marital status increase after childbirth, a time-constant consideration of the marital status at childbirth would underestimate the impact of marriage within the regarded period.

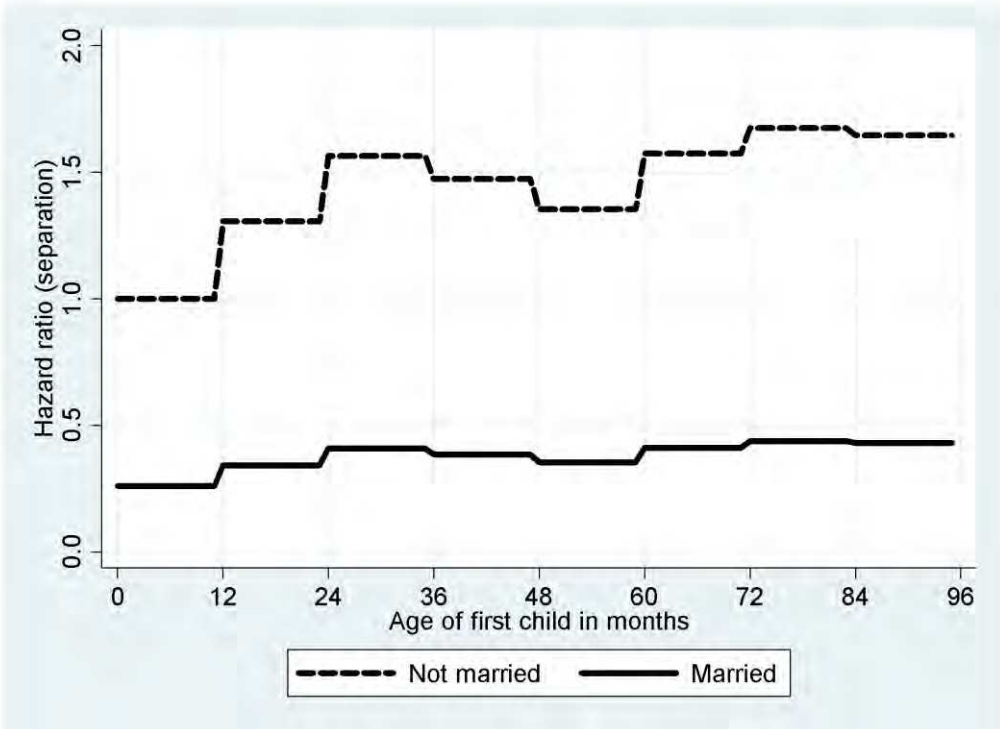

Figure 4: Piecewise constant exponential model, controlled for marital status (time-varying), from birth to age 8 of the first child, hazard ratios.

6 Acknowledgements

The authors would like to thank Michaela Kreyenfeld, Rainer Walke, Tom Hensel, Rüdiger Lenke and Daniel Fuß for their valuable suggestions on the syntax and this report.

7 References

Blossfeld, Hans-Peter/Golsch, Katrin/Rohwer; Götz (2007): *Event History Analysis with Stata*. Mahwah (NJ): Erlbaum.

Brüderl, Josef/Hajek, Kristin/Huyer-May, Bernadette/Ludwig, Volker/Müller, Bettina/Müller, Ulrike/Passet, Jasmin/Pforr, Klaus/Scholten, Mirte/ Schütze, Philipp/Schumann, Nina (2013): *pairfam Data Manual. Release 4.0.* University of Munich.

Huinink, Johannes/Brüderl, Josef/Nauck, Bernhard/Walper, Sabine/Castiglioni, Laura/Feldhaus, Michael (2011): Panel Analysis of Intimate Relationships and Family

Dynamics (pairfam): Conceptual framework and design. In: *Zeitschrift für Familienforschung/Journal of Family Research, 23, 1*, S. 77-101.

Kreyenfeld, Michaela/Walke, Rainer/Salzburger, Veronika/Schnor, Christine/Bastin, Sonja/Kuhnt, Anne-Kristin (2011): *DemoDiff – Wave 1. Supplement to the pairfam data manual*. Rostock: MPIDR (MPIDR Technical Report 2011-014).

Nauck, Bernhard/Brüderl, Josef/Huinink, Johannes/Walper, Sabine (2012): *Beziehungs- und Familienpanel (pairfam).* GESIS Datenarchiv, Köln. ZA5678 Datenfile Version 3.0.0, doi:10.4232/pairfam.5678.3.0.0.

StataCorp. (2009): *Stata Statistical Software: Release 11*. College Station, TX: StataCorp LP.

8 Annex

- *Eventhistory.do* (STATA file)
- *Eventhistory_ReadMe.txt*
- *Eventhistory _Example id.xlsx* (Excel file)
- *Eventhistory _Example_analysis_1.do* (STATA file)
- *Eventhistory _Example_analysis_2.do* (STATA file)
- *biopart_PF.do* (STATA file)
- *biopart_PF_IMP.do* (STATA file)
- *biopart_DD.do* (STATA file)
- *biopart_DD_IMP.do* (STATA file)
- List of variables included in *Eventhistory.dta* (see table IV below)

Table IV: List of variables included in Eventhistory.dta

Variable	Variable label	Values	Value labels
General information			
id	Person number anchor	*Person number*	--
START	Beginning of episode in months since birth of anchor	--	--
END	End of episode in months since birth of anchor	--	--
sex	Sex anchor	1 2	Male Female
dob	Date of birth anchor *(in months since January 1900)*	*date*	--
AGEANC	Age of anchor *(in years)*	0 14 15 …	Below 14 years old 14 years old 15 years old …
INT	Before/at/after respective interview	0 1 2 3	Before 1st interview Month of 1st interview Between waves 1 and 2 Month of 2nd interview
Information on partnerships			
UNION	Union status	0 1	No partner In union
UNIONORDER	Order of unions (shows order of later partners in simultaneous cases)	0 1 2 …	No partner 1st partner 2nd partner …
UNIONORDER_SIM	Union order for simultaneous unions: shows order of 'previous' partner	0 1 2 …	No *(simultaneous)* partner 1st partner 2nd partner …
UBREAKORDER	Order of union breaks	0	No break

	within one union	1 2 …	1st union break 2nd union break …
pid	Person number partner; if he/she was a partner at interview	. *Person number*	No person number available --
HOMOSEX	Sexual orientation within union	0 1 2	No union Heterosexual union Homosexual union
DEADPARTNER	Death of partner	0 1 2 …	No death of partner 1st partner died 2nd partner died …
COHAB	Cohabitation status	0 1	No domestic partner Domestic partner
COHABORDER	Order of cohabitations (shows additional/'later' cohabitation partners in simultaneous cases)	0 1 2 …	No domestic partner 1st domestic partner 2nd domestic partner …
COHABORDER_SIM	Cohabitation order for simultaneous cohabitations: shows order of 'previous' cohabiting partner	0 1 2 …	No *(simultaneous)* domestic partner 1st domestic partner 2nd domestic partner …
CBREAKORDER	Order of cohabitation breaks within one cohabitation	0 1 2 …	No break 1st cohab break 2nd cohab break …
UNIONORDER_COHAB	Union order number of cohabiting partner	0 1 2 …	No domestic partner 1st partner is domestic partner 2nd partner is domestic partner …
MARR	Marriage status	0 1 2	Single Married Divorced
MARRORDER	Order of marriages	0 1	No spouse 1st spouse

		2	2nd spouse
		…	…
UNIONORDER_MARR	Union order number of married partner	0	Not married
		1	1st partner is spouse
		2	2nd partner is spouse
		…	…
MARCER	Type of wedding ceremony	-7	Incomplete data
		-3	Does not apply
		1	Only a civil ceremony
		2	A civil and a religious ceremony
		3	Only a religious ceremony
Information on children			
AGEBIOK{1-10}	Age of 1st (2nd, 3rd etc.) biological child	-7	Incomplete information
		0	Childless
		1	Pregnant
		2	0 years old
		3	1 year old
		…	…
AGEBIOK_YNG	Age of youngest biological child	-7	Incomplete information
		0	Childless
		1	Pregnant
		2	0 years old
		3	1 year old
		…	…
LIVBIOK{1-10}	Co-residence with 1st (2nd, 3rd etc.) bio. child	-7	Incomplete information
		0	Not living with child
		1	Living with child
LIVKIDS	Co-residence with biological and/or non-biological children	-7	Incomplete information
		0	Living without children
		1	Living only with biological children
		2	Living only with non-biological children
		3	Living with biological and non-biological children
UNIONORDER_BIOK{1-10}	Partner number of 2nd bio. parent of 1st (2nd, 3rd etc.) child	-7	Incomplete information
		-3	Does not apply
		1	1st partner

		2	2nd partner
		…	…
		97	Another person
CAPIBIOK	Parity of surveyed bio. child	-7	Incomplete information
		-3	Does not apply
		1	1st bio. child
		2	2nd bio. child
		…	…
SEXBIOK{1-10}	Sex of 1st (2nd, 3rd etc.) bio. child	-7	Incomplete information
		-3	Does not apply
		1	Male
		2	Female
NUMBERBIOK{1-10}	Original order number of 1st (2nd, 3rd etc.) bio. child	-7	Incomplete information
		-3	Does not apply
		1	1st reported child
		2	2nd reported child
		…	…
DEADBIOK	Death of which child	-7	Incomplete information
		0	No child died
		1	1st child died
		2	2nd child died
		…	…
cid	Person number CAPI-kid	.	*No person number available*
		Person number	--
Information on inconsistent, missing and imputed data			
BIOPARTFLAG	Flag inconsistencies in the partnership biographies (biopart)	0	No inconsistencies
		1	Marriage earlier than beginning of relationship
		3	Beginning current and end previous marriage
		4	Year of birth partner
FLAG_M_UNION	Missing union episode	0	No missing
		1	Missing union episode
		2	Missing union break episode
FLAG_M_COHAB	Missing cohabitation episode	0	No missing
		1	Missing cohabitation episode
		2	Missing cohabitation

			break episode
FLAG_M_MARR	Missing marriage episode	0 1	No missing Missing marriage episode
IMP_UNION	Imputed union start date	0 1 2	No imputation Only year information Only season information
IMP_SEP	Imputed union end date	0 1 2	No imputation Only year information Only season information
IMP_COHAB	Imputed cohabitation start date	0 1 2	No imputation Only year information Only season information
IMP_COHABend	Imputed cohabitation end date	0 1 2	No imputation Only year information Only season information
IMP_MARR	Imputed wedding date	0 1 2	No imputation Only year information Only season information
IMP_DIV	Imputed divorce date	0 1 2	No imputation Only year information Only season information
IMP_dobbiok{1-10}	Imputed month in the date of birth of 1st (2nd, 3rd etc.) bio. child	0 1 2	No imputation Only year information Only season information
IMP_dodbiok{1-10}	Imputed month in the date of death of 1st (2nd, 3rd etc.) bio. child	0 1 2	No imputation Only year information Only season information
IMP_beglivbiok{1-10}	Imputed month in the start date of co-residence with 1st (2nd, 3rd etc.) bio. child	0 1 2	No imputation Only year information Only season information
IMP_endlivbiok{1-10}	Imputed month in the end date of co-residence with 1st (2nd, 3rd etc.) bio. child	0 1 2	No imputation Only year information Only season information
IMP_BEGLIVnonbiok	Imputed month at the start of co-residence with non-biological children	0 1	No imputation Only year information
IMP_ENDLIVnonbiok	Imputed month at the end	0	No imputation

	of co-residence with non-biological children	1	Only year information
Person identifiers of anchor's parents			
mid	Person number mother	.	*No person number available*
		Person number	--
fid	Person number father	.	*No person number available*
		Person number	--
smid	Person number stepmother	.	*No person number available*
		Person number	--
sfid	Person number stepfather	.	*No person number available*
		Person number	--