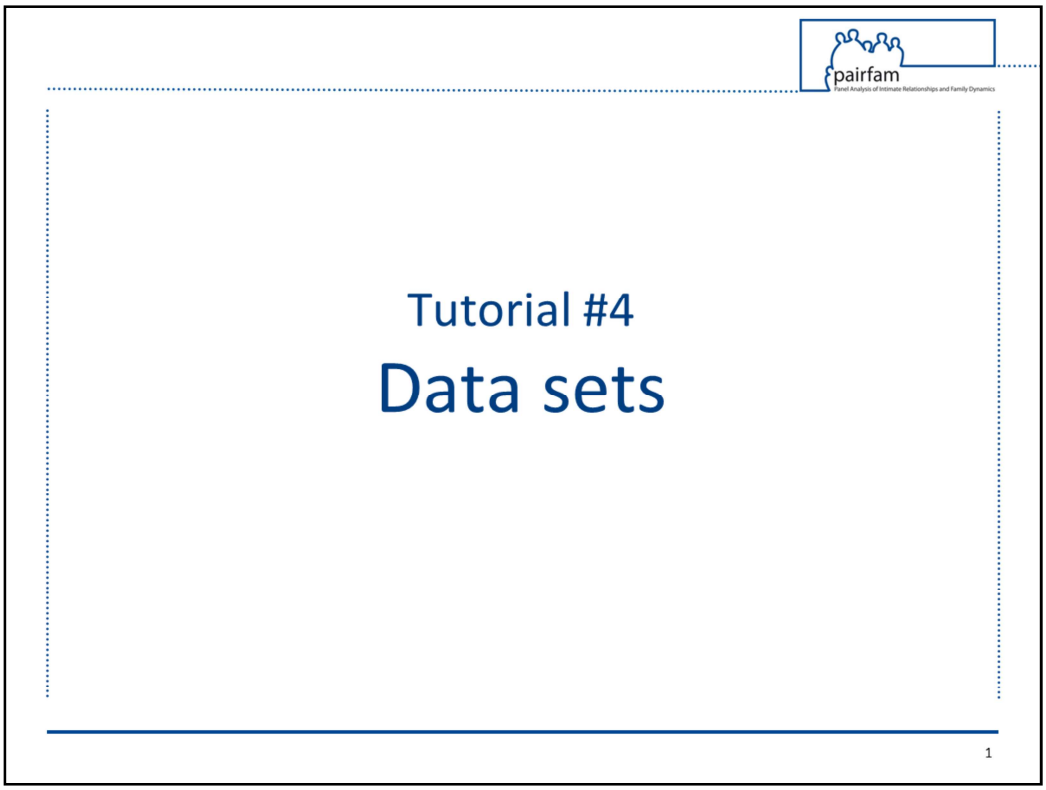
The corresponding video tutorials are available online:
https://www.youtube.com/playlist?list=PL7BcpOtSe5u_zQctYXz4ee79Zc9r4mfnr

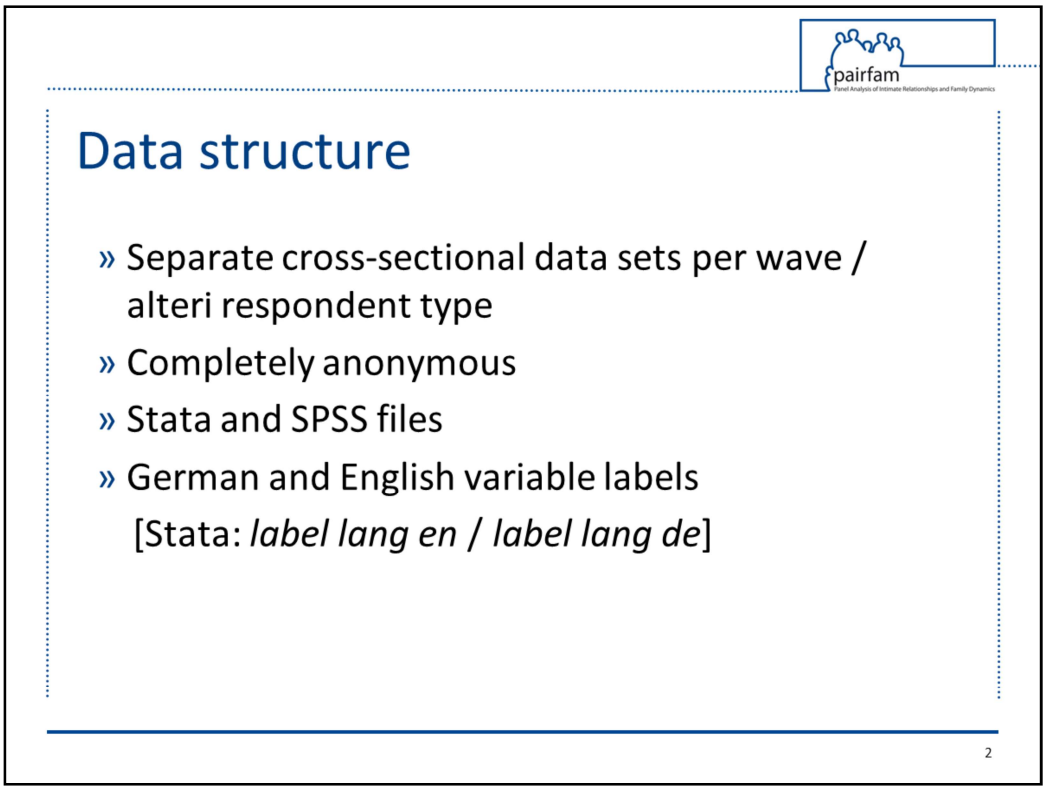


pairfam tutorial

4. Data sets

Kristin Hajek & Madison Garrett, May 2022

This fourth tutorial gives an overview of the data sets included in the pairfam Scientific Use File available to the scientific community.



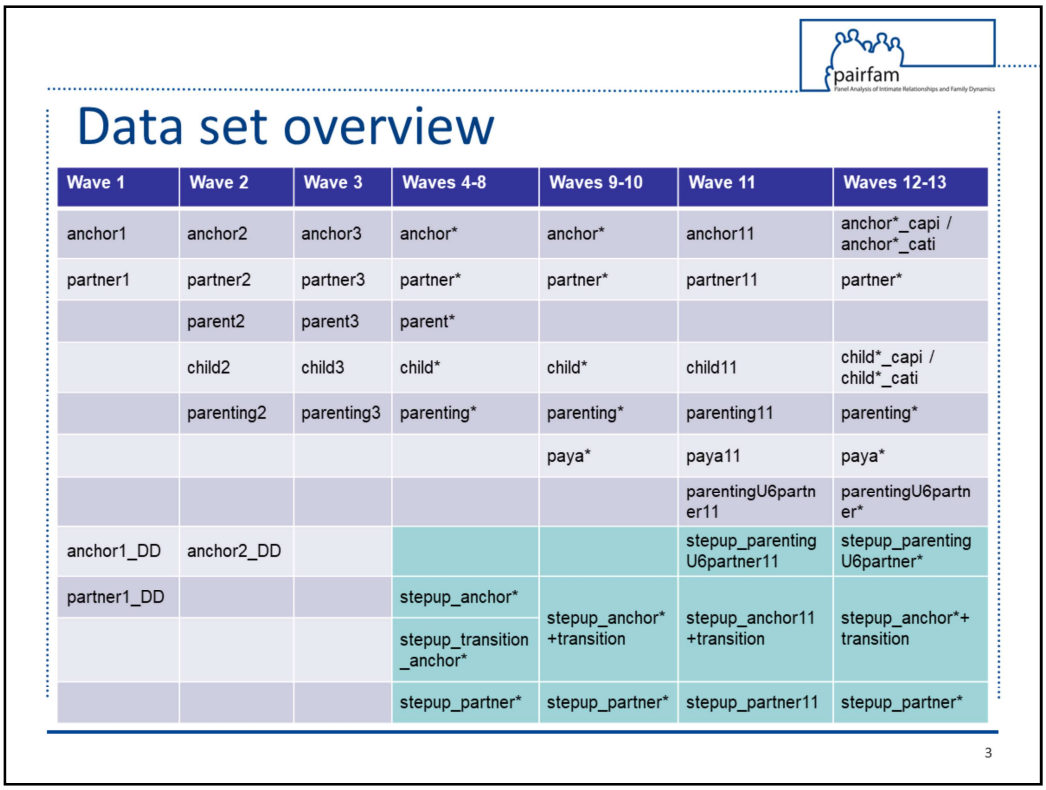
How are the *pairfam* data sets stored?

The *pairfam* data is stored in separate cross-sectional data sets per wave and per respondent type.

They are all completely anonymized so that individual respondents cannot be identified.

We provide the data sets in both Stata and SPSS format - with German and English variable labels.

The label language can be changed in Stata with the command label lang, followed by the abbreviated language: en for English and de for German.



Wave 1	Wave 2	Wave 3	Waves 4-8	Waves 9-10	Wave 11	Waves 12-13
anchor1	anchor2	anchor3	anchor*	anchor*	anchor11	anchor*_capi / anchor*_cati
partner1	partner2	partner3	partner*	partner*	partner11	partner*
	parent2	parent3	parent*			
	child2	child3	child*	child*	child11	child*_capi / child*_cati
	parenting2	parenting3	parenting*	parenting*	parenting11	parenting*
				paya*	paya11	paya*
					parentingU6partner11	parentingU6partner*
anchor1_DD	anchor2_DD				stepup_parentingU6partner11	stepup_parentingU6partner*
partner1_DD			stepup_anchor*	stepup_anchor* +transition	stepup_anchor11 +transition	stepup_anchor*+ transition
			stepup_transition_anchor*			
			stepup_partner*	stepup_partner*	stepup_partner11	stepup_partner*

Which data sets are available?

In wave 1 of the study, both anchor and partner data were collected. These data sets are named *anchor1* and *partner1*, respectively. Data from the anchor and partner surveys of wave 2 are named *anchor2* and *partner2*, and so on for each consecutive wave.

In wave 2, anchor respondents' parents and children were also surveyed, and their data is saved as the data sets *parent2* and *child2*. The parent survey was discontinued after wave 8; therefore, only *parent2* to *parent8* exist.

Furthermore, from wave 2 onwards, anchors and their partners received a paper parenting questionnaire including questions regarding the parenting of their children up to the age of 15 (*parenting2*, etc.).

Please note that for waves 12 and 13, due to methodological changes caused by the COVID-19 pandemic, the anchor and child data sets are divided according to survey mode, so that separate CAPI and CATI data sets are included.

The PAYA data sets (available for waves 9 to 13) contain parenting data from both anchor respondents and their partners.

The *parentingU6partner* data sets, available for waves 11 to 13, contain information from partner interviews about children under the age of six.

For waves 1 and 2, data of the *DemoDiff* subsample is stored in separate data sets, recognizable by the suffix *DD*. From wave 3 on, the *DemoDiff* subsample was

completely integrated into the *pairfam* data.

From wave 4 onwards, the *stepup_anchor* data sets contain data from former child respondents who reached the age of 15 and were integrated into the anchor survey.

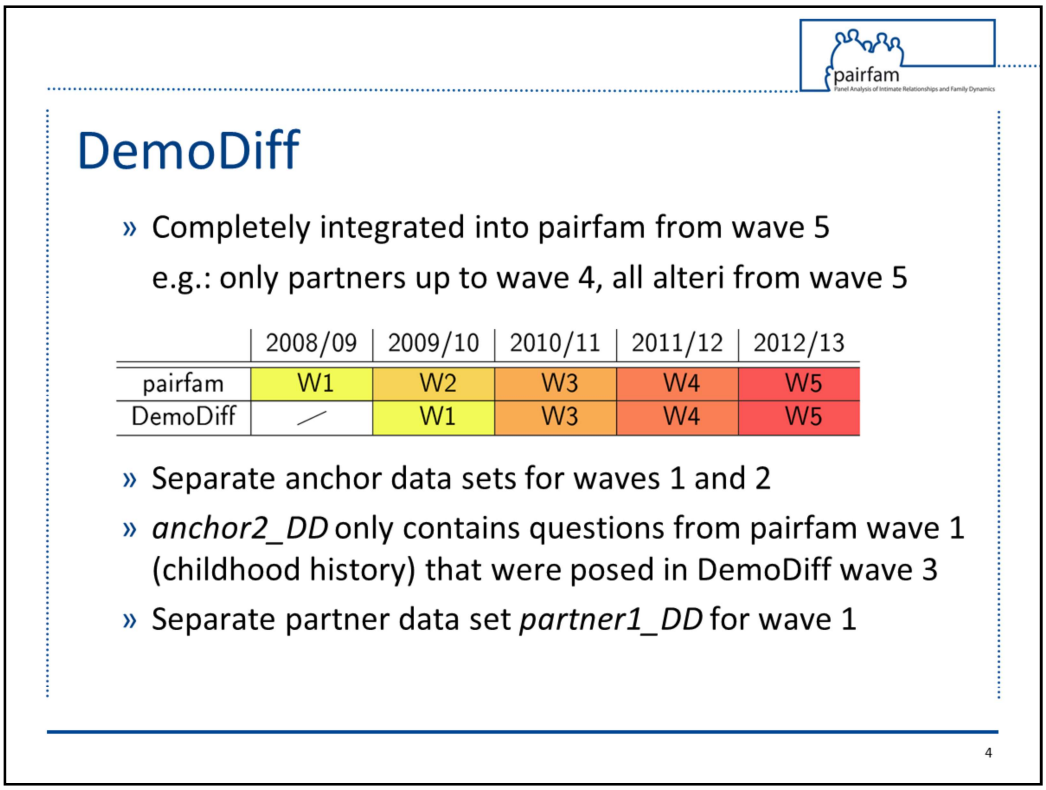
From wave 4 to 8, these “step-up” respondents also received an additional paper questionnaire in the first wave of participation as an anchor respondent covering retrospective questions. This data is stored in the respective *stepup_transition_anchor** data sets.

From wave 9 onwards, this transition module was incorporated into the anchor CAPI survey (as an add-on module for first-time anchor respondents), and data from these interviews are stored in a combined data set named *stepup_anchor+transition*.

Data from *step-up* respondents’ partners are stored in the data sets *stepup_partner4-13*.

Also available are *stepup_parentingU6partner* data sets from wave 11 to wave 13.

All *step-up* data sets are stored in a separate folder in the Scientific Use File.



Why are there separate *DemoDiff* data sets for wave 1 and wave 2?

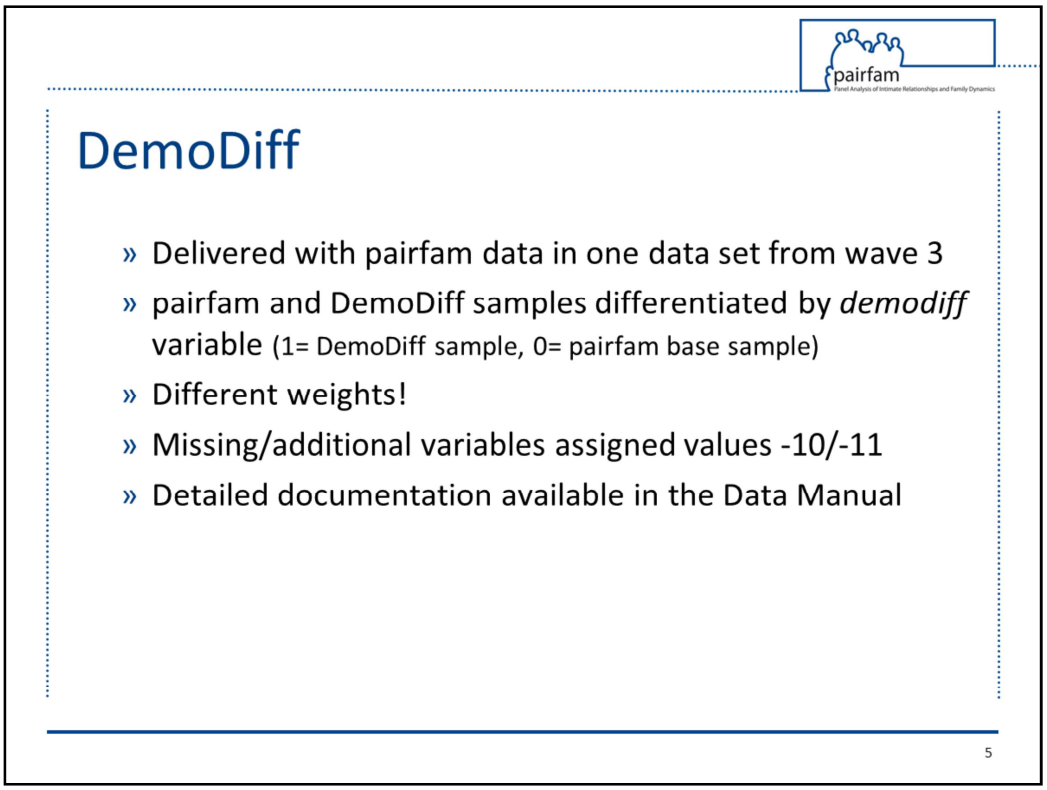
As previously mentioned, *DemoDiff* began as a separate project and was later integrated into *pairfam* from wave 5.

Up until wave 4, only partners of *DemoDiff* respondents were surveyed, not their parents and children. After the integration of DemoDiff into pairfam in wave 5 onwards, all available alteri were surveyed.

The *DemoDiff* questionnaire was largely similar to the *pairfam* questionnaire. However, *DemoDiff* began one year after *pairfam*. While pairfam's first wave took place at the end of 2008/the beginning of 2009, wave 1 of *DemoDiff* took place in 2009/2010.

In order to integrate *DemoDiff* respondents into the *pairfam* rhythm, *DemoDiff* skipped its second wave and instead implemented the *pairfam* wave 3 questionnaire, including the retrospective childhood history module from wave 2. The retrospective childhood history module is stored as *anchor2_DD*.

Due to the complexity of the integration, *pairfam* and *DemoDiff* data sets are separate for waves 1 and 2. The *DemoDiff* wave 1 partner data set is also stored separately as *partner1_DD*.

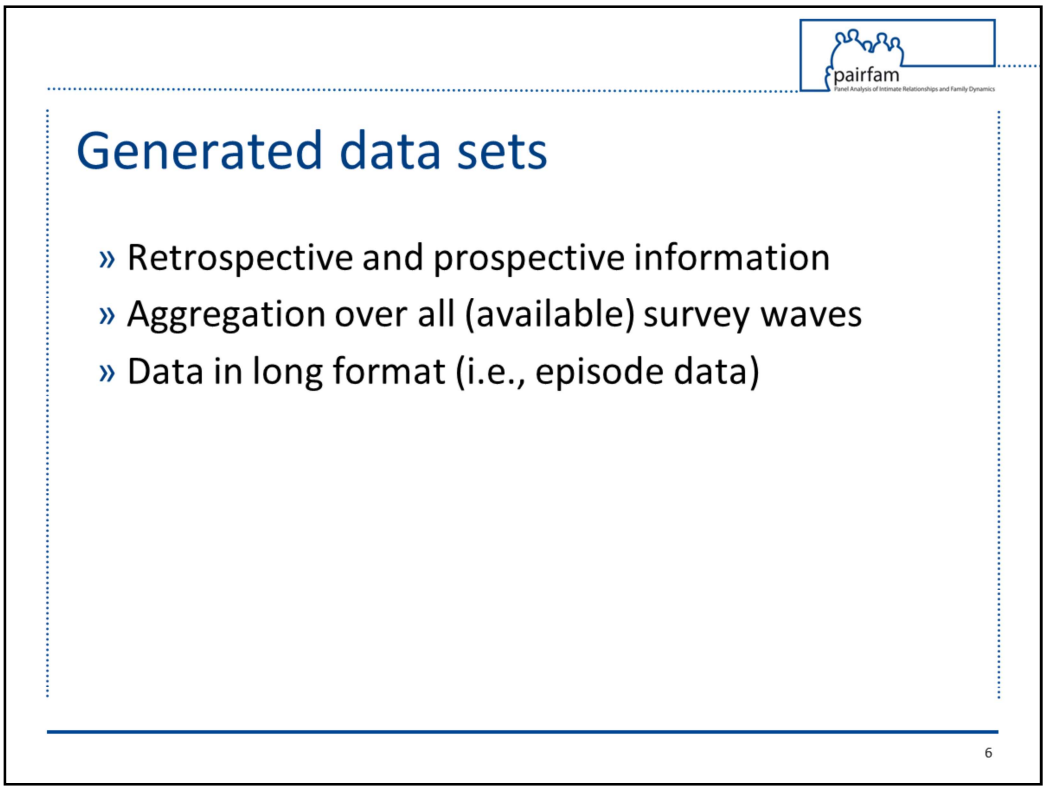


From wave 3 onward, *pairfam* and *DemoDiff* data are delivered in one integrated data set differentiated by the variable *demodiff*.

Design weights are available for the different samples, as well as combined weights that allow for analyses of the full data set.

Missing and additional variables both in the *pairfam* and *DemoDiff* data are marked with the missing codes -10 and -11.

Detailed documentation of the differences between *pairfam* and *DemoDiff*, their integration, and weighting procedures can be found in the Data Manual.

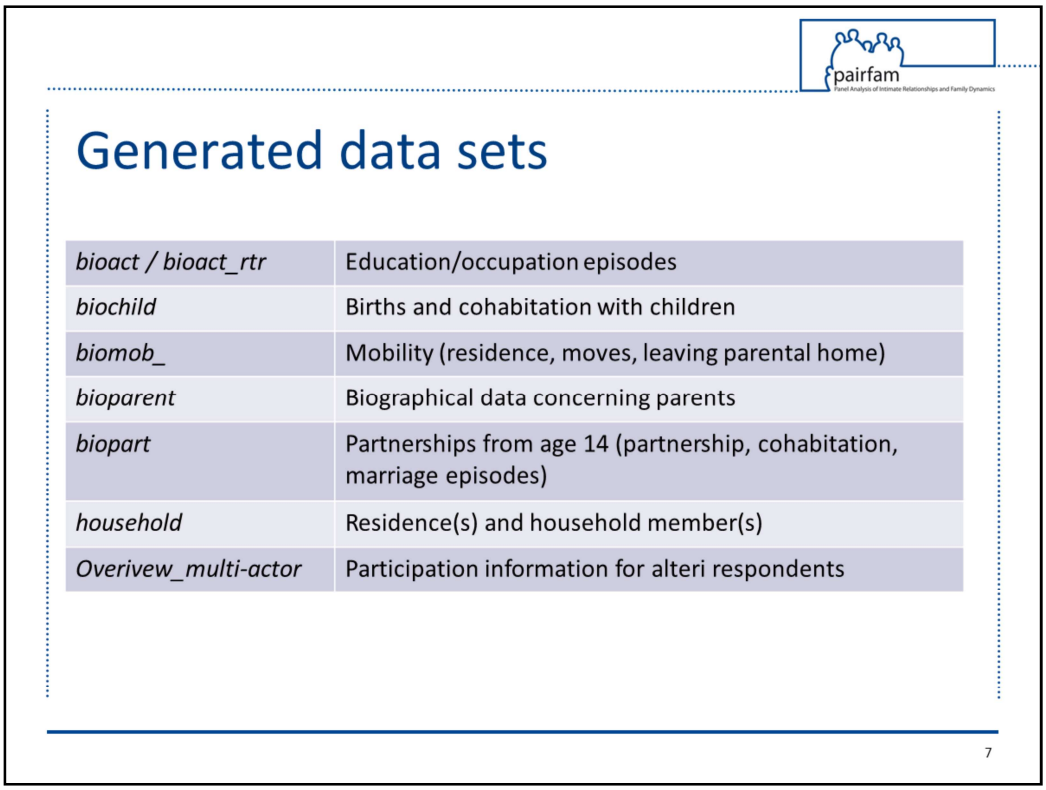


What are the generated data sets?

Information on anchor respondents' life history domains as well as their parents are compiled into separate user-friendly data sets that allow for a more convenient analysis of biographical information.

These generated data sets combine retrospective and prospective information across all survey waves.

The data is stored in long format, meaning one row for each episode (for example: one row for each partnership).



The following data sets were generated to facilitate analyses:

bioact contains monthly information on educational and occupational activities from the month of the respondent's first interview.

bioact_rtr covers retrospective information collected in wave 3 on the anchor respondent's education and occupation from the age of eighteen.

The data set *biochild* is a panel data set that includes additional retrospective episode data collected in wave 1 covering births to anchor respondents and their cohabitation episodes with children.

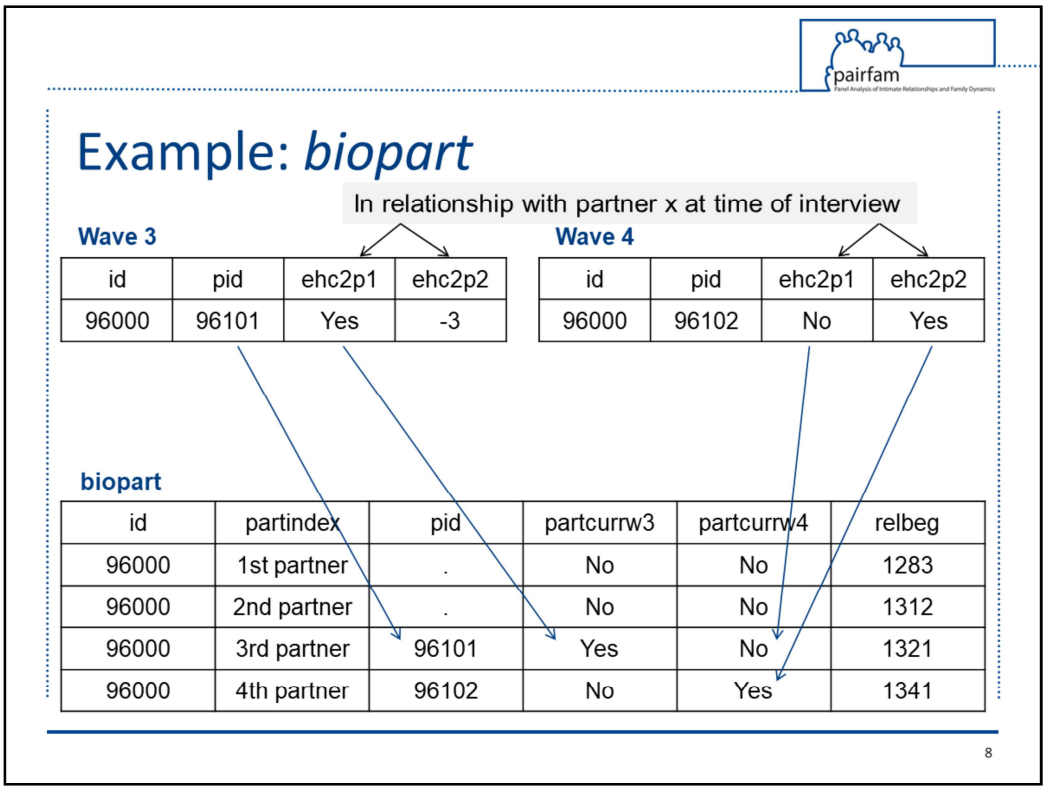
The *biomob* data sets are panel data sets with both prospective and retrospective information on anchor respondents' mobility, residences, and moves.

bioparent provides prospective and retrospective biographical data for the anchor respondents' parents, while *biopart* covers all partnerships from the age of 14.

The data set *household* is a panel data set containing information on the anchor respondent's residence and household members at the time of the interview for waves 1-3 only.

The data set *Overview_multi-actor* provides an overview of all alteri respondent participation in the partner, child, parenting, and parent surveys in each survey wave.

The generation of these data sets is described in detail in the Data Manual, and the syntax for their generation is included in the Scientific Use File.



To illustrate the differences between the survey and generated data sets, let's have a look at the generated data set *biopart*.

The anchor data sets of wave 3 and wave 4 include the following information: *id* is the person identifier for the anchor respondent and *pid* is the partner identifier. *ehc2p1* and *ehc2p2* provide information on the anchor respondent's relationship status with partner 1 and partner 2, respectively, at the time of the interview.

Here, this particular anchor respondent was in a relationship with partner 1 in wave 3, who had the partner id 96101. A second partner did not exist, resulting in the missing code -3 for the variable *ehc2p2*.

In wave 4, the anchor respondent was no longer in a relationship with partner 1, and reported a new partner – partner 2 – who was allocated the *pid* 96102.

Now let's have a look how this information is presented in the biopart data set.

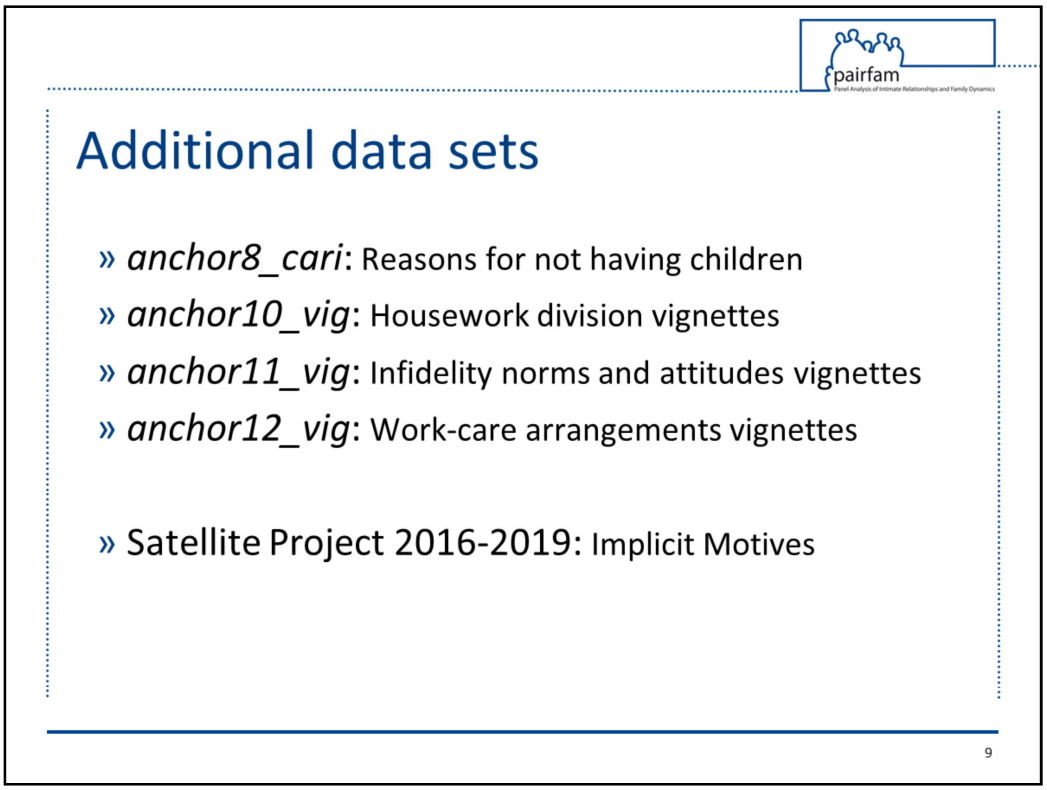
We can see that the anchor respondent with the *id* 96000 reported four different partnerships. Each relationship is stored in a different row. The first two partners do not have a *pid*, indicating that these relationships took place before the first interview. All partnerships from age 14 up until the first interview were retrospectively recorded in an add-on module and integrated into the *biopart* data set. Partners reported during the panel have a *pid*.

Here, the third partner has the *pid* 96101 and was listed as the anchor respondent's current partner in wave 3, which is visible in the variable *partcurrw3*.

The fourth partner with the *pid* 96102 was in a relationship with the anchor respondent in wave 4, stored in the variable *partcurrw4*.

The variable *relbeg* contains the month of the start of each relationship. In order to facilitate duration calculations, dates within generated data sets are stored as a combination of both month and year as numerical values representing the number of months that have passed since January 1900.

A helpful description of how to convert this information into the original month and year is included in the Data Manual.



Are there any additional data sets?

Several special surveys and satellite projects were conducted throughout the panel alongside the standard questionnaires, and this data is also delivered as part of the Scientific Use File.

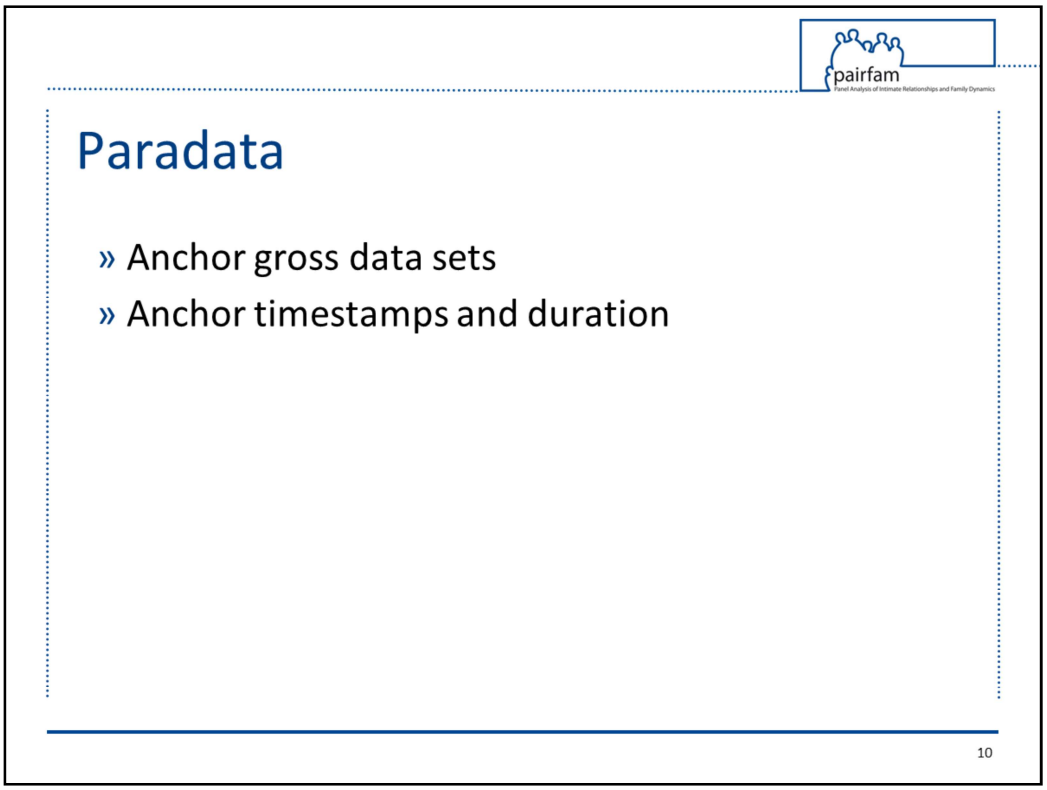
For example, the standard anchor questionnaire for waves 1-7 entails questions about reasons for not having children (or additional children). In wave 8, an additional qualitative audio recording covering this topic was introduced. These recordings were transcribed with the MAXQDA software and are saved in the data set *anchor8_cari*. Details on the coding strategy are available in Technical Paper No. 10.

In wave 10, a factorial survey experiment with vignettes was conducted within the CASI section of the anchor questionnaire to evaluate the division of housework and paid labor in 3 hypothetical partnerships. The goal of this sub-study was to explain gendered work distributions within couples by disentangling how individual dimensions (such as financial resources, gender, and family status) influence evaluations of fairness. Detailed information on this vignette study are available in Technical Paper No. 14.

In wave 11, a factorial survey experiment focusing on infidelity in intimate relationships was implemented. Further information on this vignette study are available in Technical Paper No. 18.

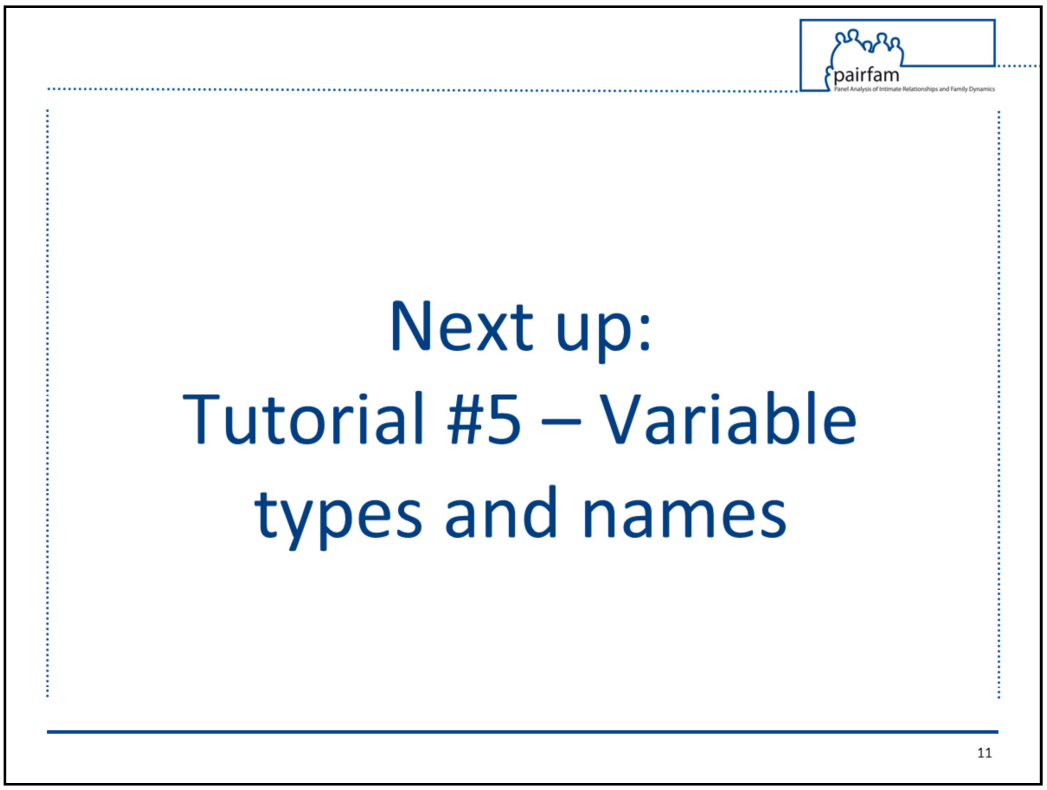
Wave 12 included a factorial survey experiment examining normative judgements of work-care arrangements. The study focused on the combination of mothers‘ and fathers‘ employment and use of day care for their children. A detailed description of the study is available in Technical Paper No. 19.

The Implicit Motives Project was a satellite project that ran from 2016 to 2019. The goal of the project was to research the interplay of partner-related explicit and implicit motive dispositions with relationship quality in German couple relationships. Data and documentation are stored in a separate folder within the Scientific Use File.



Are there additional paradata sets available?

We added information on the survey process and the interviewer to the data set for all valid and completed interviews. In addition, the *paradata* directory in the Scientific Use File contains the gross anchor data set and anchor timestamps and duration. The gross data set includes all cases contacted for an additional interview. The following fieldwork information is provided: number of contact attempts, interviewer identification numbers, information on the place of residence, final processing status and interview mode. The data sets *time+duration_W** contain information on the interview day, timestamps, and duration of individual modules. Please note that these are raw data sets that have not been reviewed or processed. Further information is included in the Data Manual.



This was a brief description of all the available data sets included in the *pairfam* Scientific Use File.

The next tutorial will discuss available variables types and their naming convention.