

Missing data in the SHARE Job Episodes Panel: A potential cause and solution

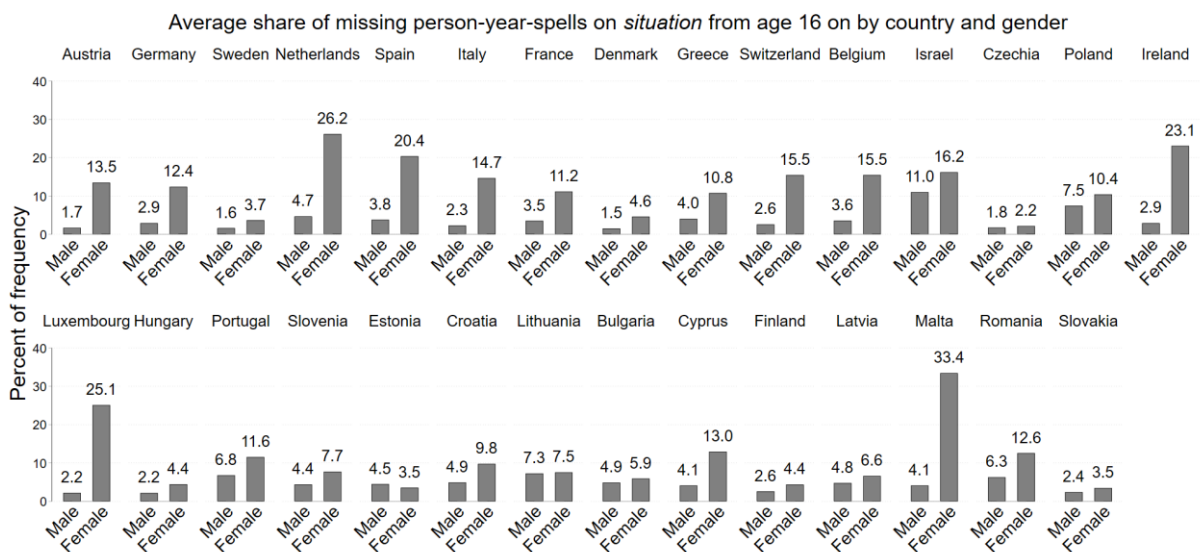
Missingness in the Job Episodes Panel

The Survey of Health, Ageing and Retirement in Europe (SHARE) is a cross-national panel study for individuals aged 50 and older covering 28 countries in the most recent waves (Bergmann u. a. 2019). SHARELIFE contains valuable and detailed retrospective data about individual life courses. Instead of the raw survey data, many researchers use the Job Episodes Panel (JEP), a generated dataset containing information on working life courses based on the retrospective SHARELIFE data (Brugiavini u. a. 2013, 2019). One of the particularly rich variables is the *situation* indicator containing the self-reported situation or activity individuals were in for each year, covering multiple categories out of labour force. Constructing such variables and making them publicly available is an important service for the academic community and improves the comparability of research since the data preparation is less heterogenous compared to when data preparation is based on individual decisions. However, the *situation* variable contains missing information for sizeable shares of person-year spells. This is problematic because missingness seems to be not random when looking at the two dimensions countries and gender.

Using the person-year-spells for ages above 15 of the whole JEP sample, the average share of missing observations of all person-year spells per person on the *situation* variable ranges from 1.5% in Denmark to 7.5% in Poland for men and from 3.5% in Estonia and Slovakia to 33.4% in Malta for women (figure 1). In many countries there is a considerable gender differences in relative missingness, while the level varies across countries. Due to this sizeable missingness, individual studies have to make decisions on how to proceed with such missing data (e.g. listwise deletion or imputations) if they are aware of them in the first place. These different approaches are likely to produce biased results since specific countries as well as specific groups of men and particularly women (see below) will have been more affected by large shares of missing information on their working lives.

When focussing on all person-year-spells the gender gaps become narrower (figure A1). However, in most empirical applications, the *situation* variable might be applied for employment biographies which is why I focus on the person-year spells for ages above 15 in the main figures.

Figure 1: Missingness before adjustment.



Own calculations based on whole sample of JEP release 7.1.0.

The Source of Missingness and a Potential Solution

The Job Episodes Panel (up until release 8.0.0) does not consider variables of the raw data containing information on the situations people were in after their last job (variables re035_*, re039_*, re039a_*). They are then asked whether the situation changed again and to report the second situation they were in after the change etc. until their situation does not change anymore. Most of the spells with missing information can be filled using this information of the raw data.

I impute only person-year-spells that are missing on the *situation* variable and that occur in the years after the last reported job (re035). Spells with missing information are filled for each year until the first year with non-missing information on the *situation* variable or until the situation changed again (re039a). If the situation changed, the same procedure is done for the n^{th} situation after the last job n^{th} times. The year in which the situation after the last job changed is filled with the new situation.

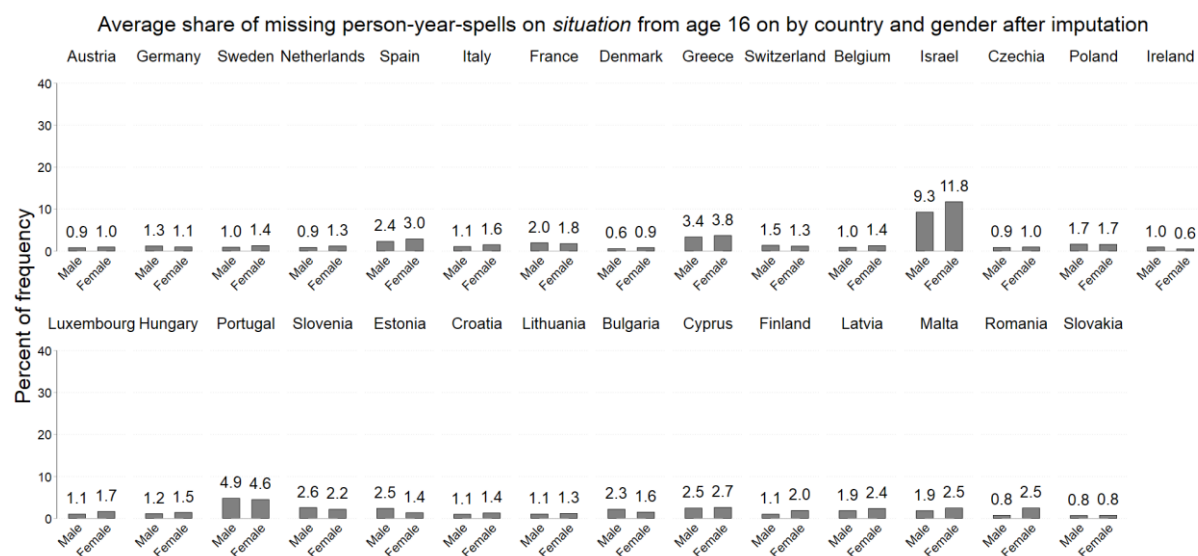
The code for filling the missing information is [available online](#).

Reduction in Missingness & country- and gender-specific patterns

Figure 2 shows the share of missingness by country and gender after filling the missing information as described above. The level of relative missingness drops down to on average below 2% of all person-year spells for most countries. The improvement differs across countries and gender, but overall, the gender gaps in missingness are reduced to a minor level.

When considering all person-year spells and not only those for ages 16 onwards the share of missing data drops down to an average of 7-9% for most countries (figure A2). This is because most respondents have missing information in their first 4 years of life which cannot be filled by the applied procedure – 81% of the remaining missing information on the *situation* variable occur during the first 15 years of age. And a large share of the remaining missing information is due to gaps between education and the first situation after the end of the education phase.

Figure 2. Missingness after adjustment.

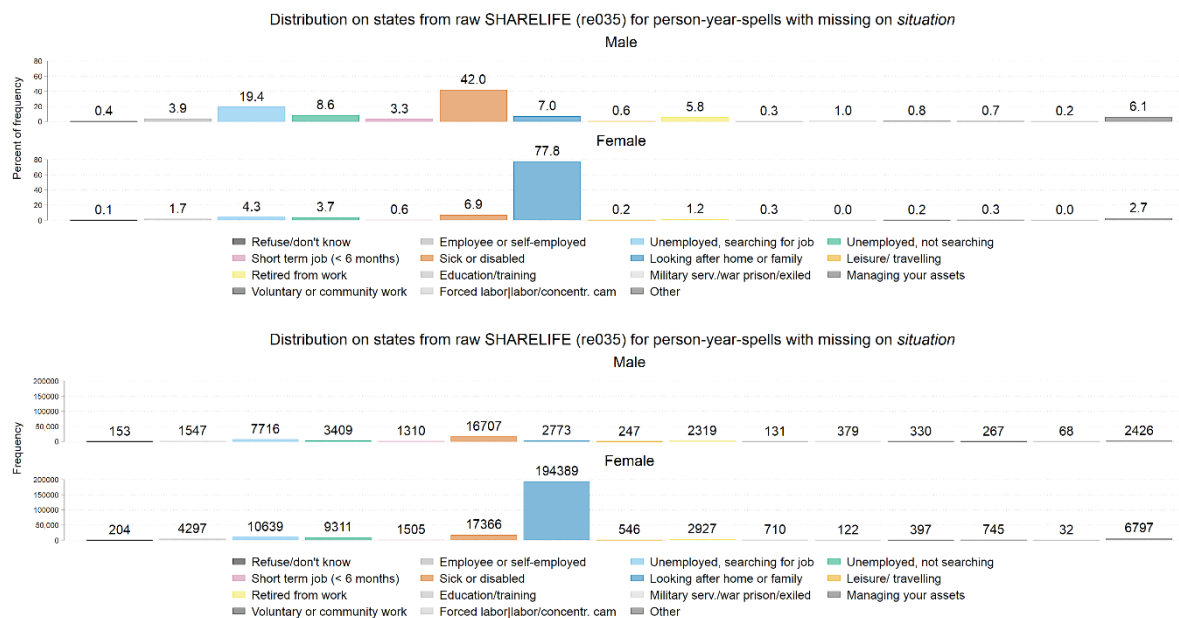


Own calculations based on whole sample of JEP and SHARELIFE release 7.1.0.

Figure 3 shows the categories of the *situation* variable that have been filled in relative and absolute terms by gender and pooled over all countries. 77.8% of the missing data (a total of 194,383 person-year spells) for women is due to unpaid care work (*Looking after home and family*) they engaged in after their last job. Thus, not considering the situation respondents were in after their last job produces particularly large shares of missing spells for women engaged in care work. This is most likely due to the traditionally gendered division of labour in most of the observed countries, in which women are

more likely to drop-out of the labour market and retire earlier due to care duties (e.g. Carr u. a. 2016; Zagel und Van Winkle 2020). Thus, a large share of unpaid care work is obscured in the original situation variable. For women, small shares of the previously missing data represent being *sick or disabled* as well as unemployment periods – spells that were underestimated in the JEP data so far too.

Figures 3. Distribution of working states that have been filled for all countries by gender.



Own calculations based on whole sample of JEP and SHARELIFE release 7.1.0. Some small categories merged.

For men on the other hand, a majority of previously missing data is due to being *sick or disabled* (42%), followed by *unemployed and searching for a job* (19.4%) as well as *unemployed but not searching for a job* (8.6%) and *looking after home and family* (7%). Thus, even for men, specific working spells that mostly deviate from employment were concealed in the original situation variable.

However, for both, men and women, the situation that is filled and thus was previously neglected differ across countries. Figures A3 show the distribution of working states that have been filled for each country by gender. For women, in all countries the highest share of missing spells was filled with care work. However, the magnitude differs. Whereas in Spain 91% of all missing spells of women are due to *looking after home and family*, in Poland it is 44%. For example, for men in Italy, roughly as many spells of *unemployed and searching for a job* as well as *being sick or disabled* were filled (27% each), whereas in the Netherlands 67% of spells were filled with *being sick or disabled* alone. For men in Slovenia, only 9% of the previously missing social situation spells are due to being sick or disabled, while 35% are due to *unemployed and searching for a job* and 27% due to *unemployed and not searching for a job*.

Summary and use

The missingness of the *situation* variable is not random, but very gendered. Furthermore, the situations concealed by the missing information in the original *situation* variable are highly skewed as well, mostly covering rather precarious situations beyond employment.

Depending on how researchers have dealt with this missingness in their analysis, they might have created skewed data, if they, for example, excluded respondents with large shares of missing data. Particularly, they might have excluded selective individuals such as sick/disabled or unemployed men and female unpaid caregivers. The figures presented here aim to give researchers a first idea whether or not the missingness issue might be problematic for their studies.

Other variables provided by the Job Episodes Panel that are originally based on the same variables as the *situation* variable (Brugiavini u. a. 2013) might be updated based on the new *situation* variable as well. This is particularly the case for the *unemployed* and *retired* dummies. The *working* dummy on the other hand should not be adjusted based on the new *situation* variable as it was constructed from another set of variables.

The team of the Job Episodes Panel will address this issue, among others, and publish a new version of the Job Episodes Panel including some other improvements with the next SHARE release. Until this release is publicly available, scholars who used or currently use the Job Episodes Panel and the *situation* variable might still want to use the (almost) full retrospective working/activity lives. This might be relevant for ongoing projects, but also for replications of already finished studies for which the *situation* variable was relevant.

The code provided can be applied to fill most of the missing person-year spells on the *situation* variable and is meant to be used only as transition solution until the fully revised Job Episodes Panel is available.

Acknowledgements

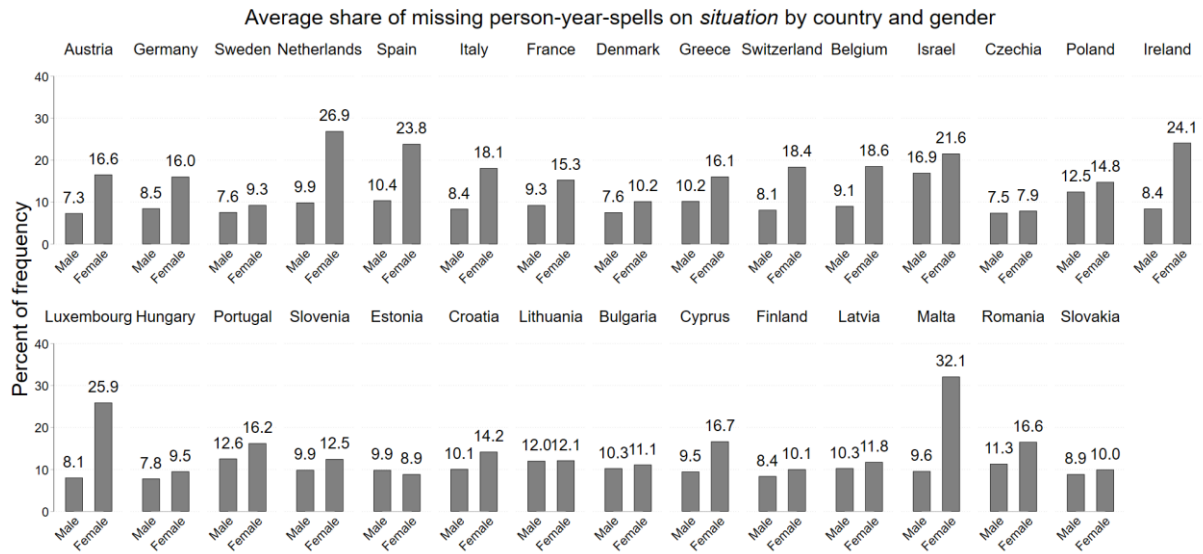
I want to thank Christiaan Monden, Rinaldo Naci, Kent Troutman and Zachary Van Winkle for their helpful feedback.

References

- Bergmann, Michael, Thorsten Kneip, Giuseppe De Luca, und Annette Scherpenzeel. 2019. „Survey participation in the Survey of Health, Ageing and Retirement in Europe (SHARE), Wave 1-7“. *Munich: Munich Center for the Economics of Aging*.
- Brugiavini, Agar, Danilo Cavapozzi, Giacomo Pasini, und Elisabetta Trevisan. 2013. „Working Life Histories from SHARELIFE: A Retrospective Panel“. 14.
- Brugiavini, Agar, Cristina E. Orso, Mesfin G. Genie, Rinaldo Naci, und Giacomo Pasini. 2019. „Combining the Retrospective Interviews of Wave 3 and Wave 7: The Third Release of the SHARE Job Episodes Panel“.
- Carr, Ewan, Emily T. Murray, Paola Zaninotto, Dorina Cadar, Jenny Head, Stephen Stansfeld, und Mai Stafford. 2016. „The Association Between Informal Caregiving and Exit From Employment Among Older Workers: Prospective Findings From the UK Household Longitudinal Study“. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* gbw156. doi: 10.1093/geronb/gbw156.
- Zagel, Hannah, und Zachary Van Winkle. 2020. „Women’s Family and Employment Life Courses Across Twentieth-Century Europe: The Role of Policies and Norms“. *Social Politics: International Studies in Gender, State & Society* . Online first. doi: 10.1093/sp/jxz056.

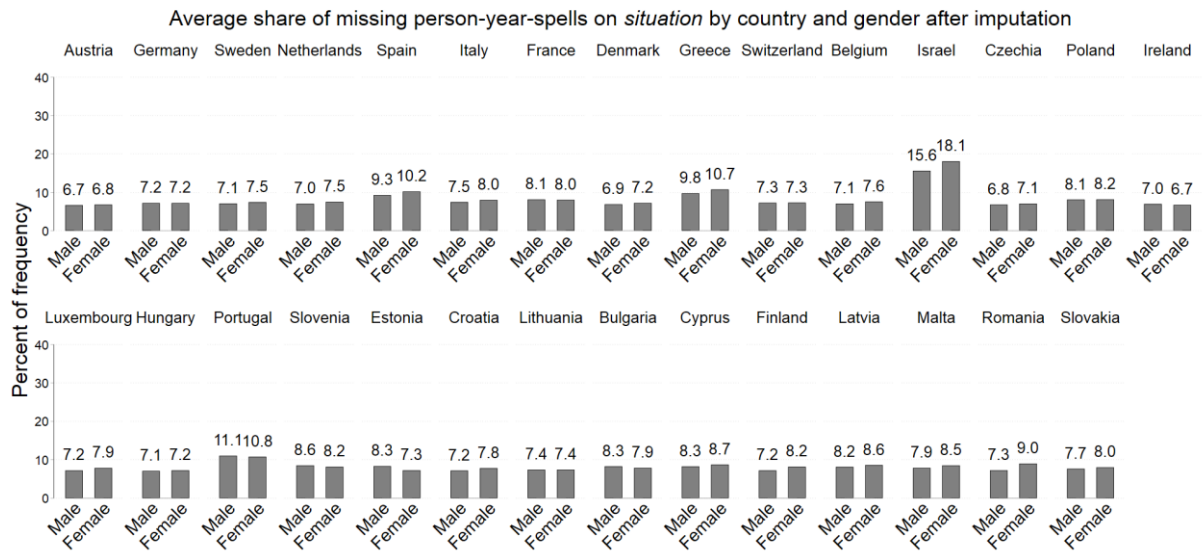
APPENDIX

Figure A1. Share of missingness for all person-year spells before filling missing information



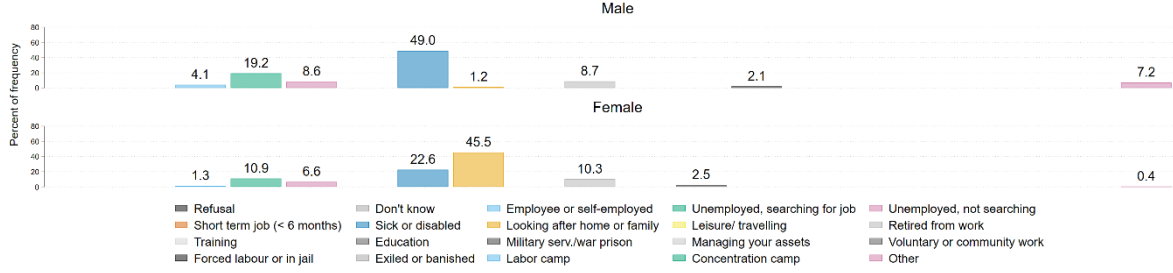
Own calculations based on whole sample of JEP release 7.1.0.

Figure A2. Share of missingness for all person-year spells after filling missing information



Own calculations based on whole sample of JEP and SHARELIFE release 7.1.0.

Distribution on states from raw SHARELIFE (re035) for person-year-spells with missing on *situation* in Czechia



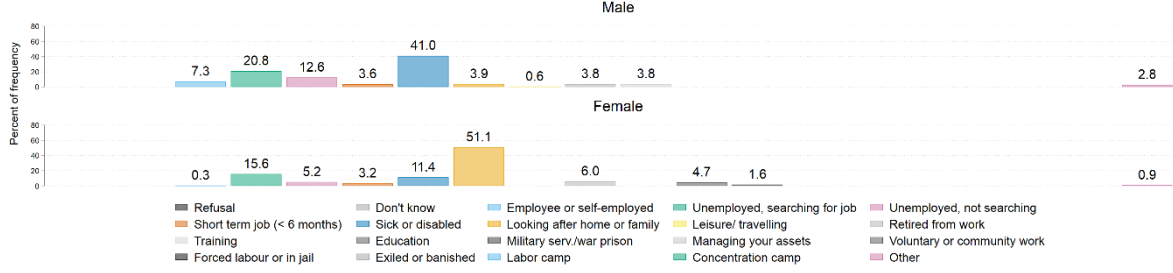
Distribution on states from raw SHARELIFE (re035) for person-year-spells with missing on *situation* in Denmark



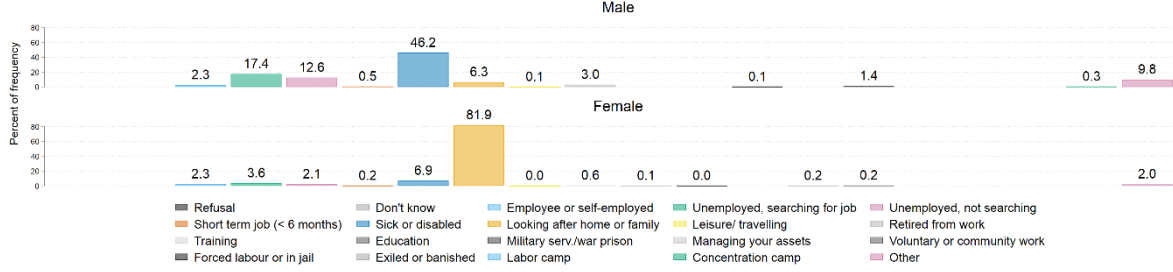
Distribution on states from raw SHARELIFE (re035) for person-year-spells with missing on *situation* in Estonia



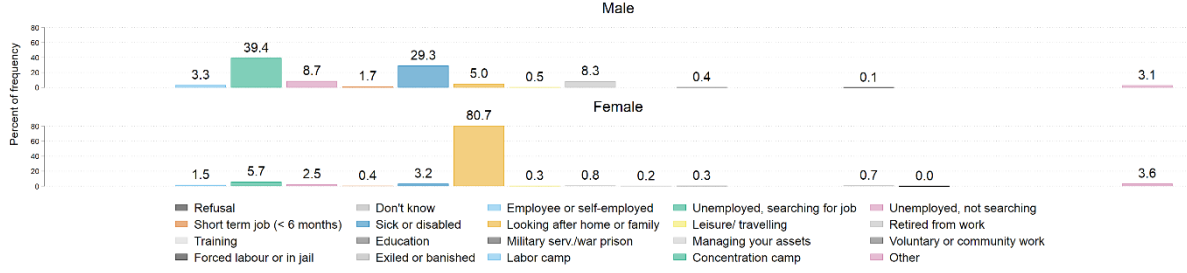
Distribution on states from raw SHARELIFE (re035) for person-year-spells with missing on *situation* in Finland



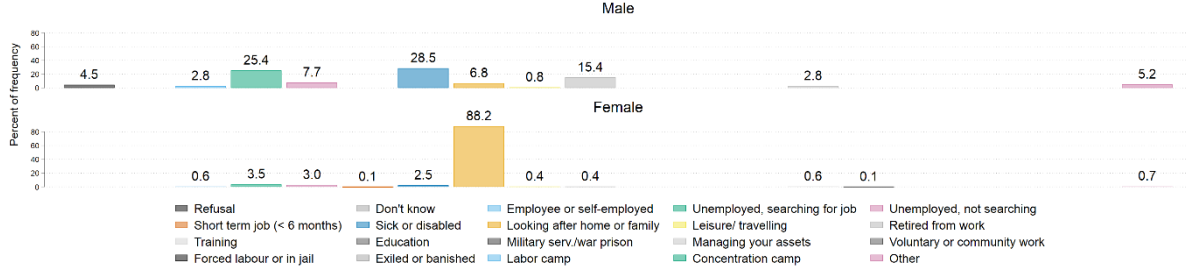
Distribution on states from raw SHARELIFE (re035) for person-year-spells with missing on *situation* in France



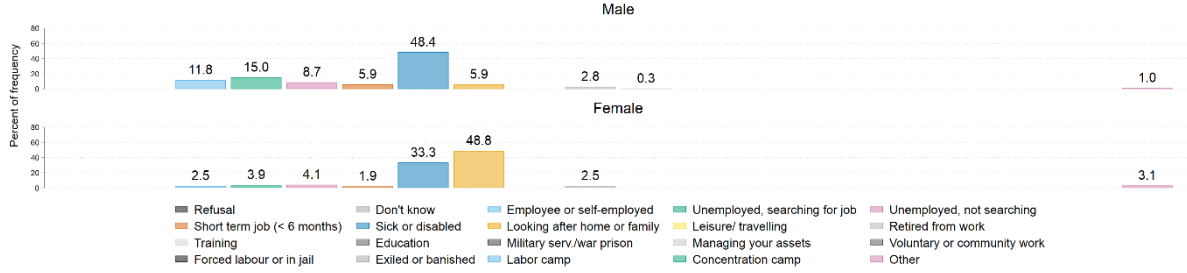
Distribution on states from raw SHARELIFE (re035) for person-year-spells with missing on *situation* in Germany



Distribution on states from raw SHARELIFE (re035) for person-year-spells with missing on *situation* in Greece



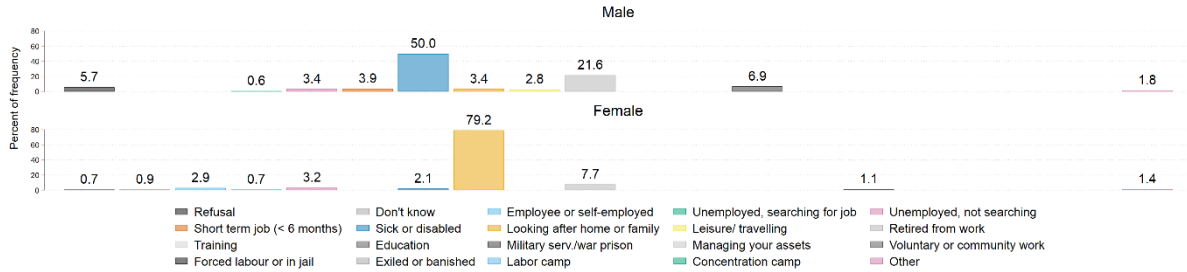
Distribution on states from raw SHARELIFE (re035) for person-year-spells with missing on *situation* in Hungary



Distribution on states from raw SHARELIFE (re035) for person-year-spells with missing on *situation* in Ireland



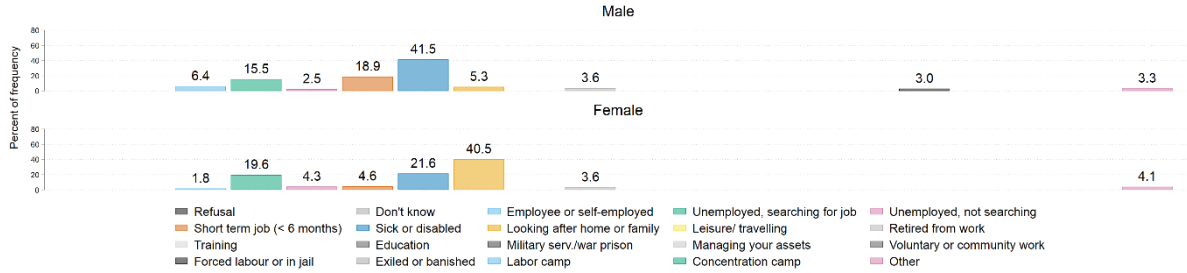
Distribution on states from raw SHARELIFE (re035) for person-year-spells with missing on *situation* in Israel



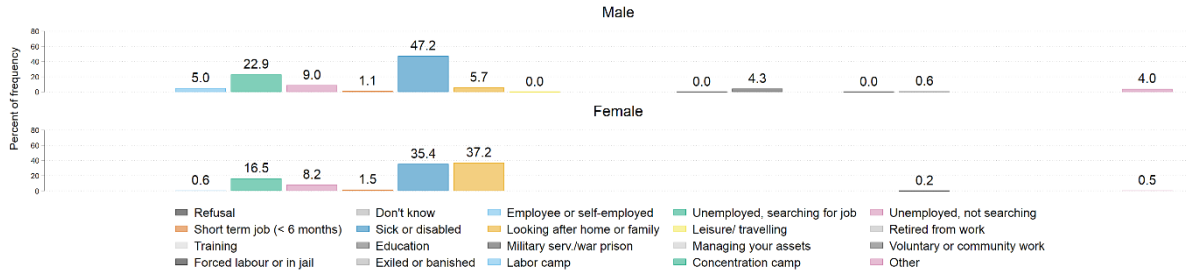
Distribution on states from raw SHARELIFE (re035) for person-year-spells with missing on *situation* in Italy



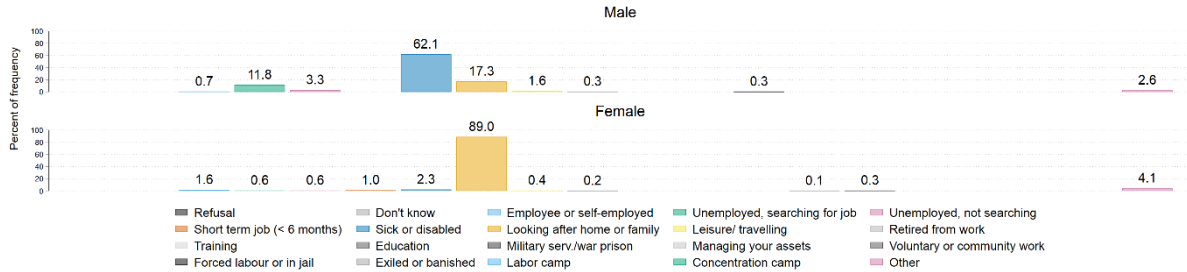
Distribution on states from raw SHARELIFE (re035) for person-year-spells with missing on *situation* in Latvia



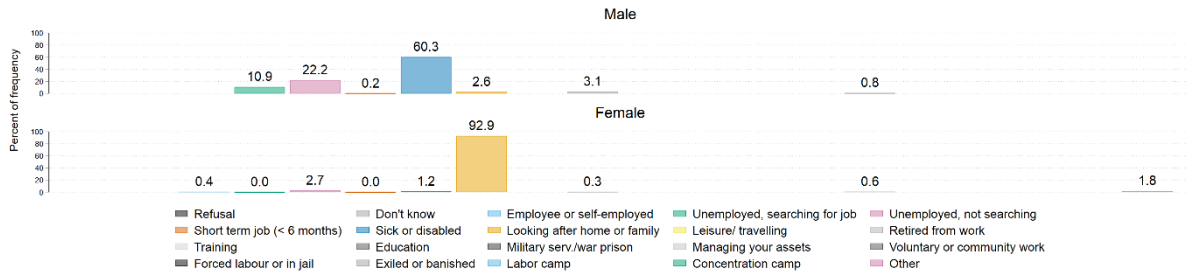
Distribution on states from raw SHARELIFE (re035) for person-year-spells with missing on *situation* in Lithuania



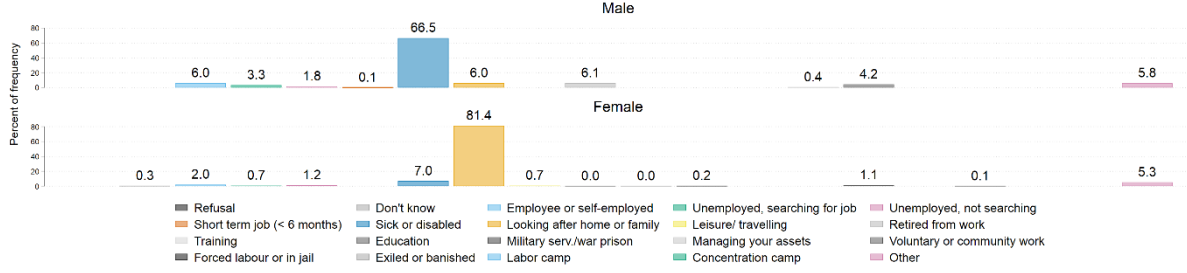
Distribution on states from raw SHARELIFE (re035) for person-year-spells with missing on *situation* in Luxembourg



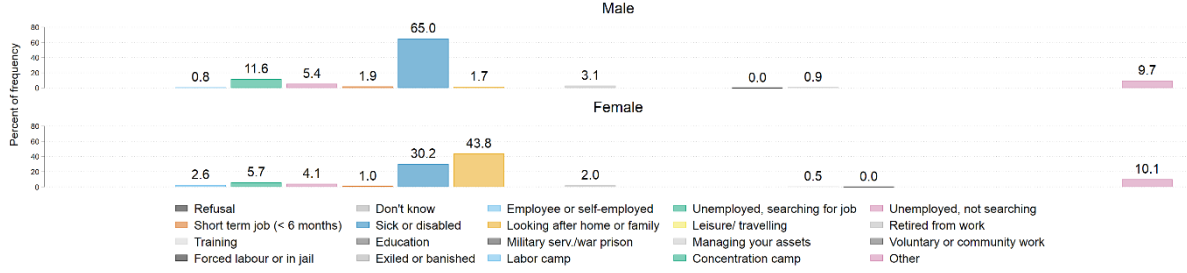
Distribution on states from raw SHARELIFE (re035) for person-year-spells with missing on *situation* in Malta



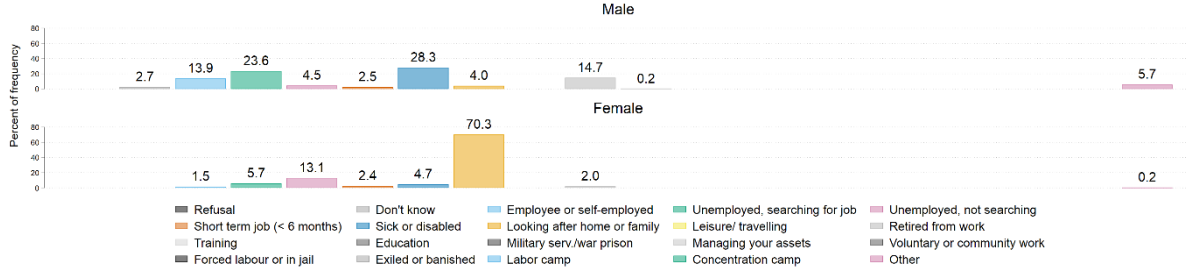
Distribution on states from raw SHARELIFE (re035) for person-year-spells with missing on *situation* in Netherlands



Distribution on states from raw SHARELIFE (re035) for person-year-spells with missing on *situation* in Poland



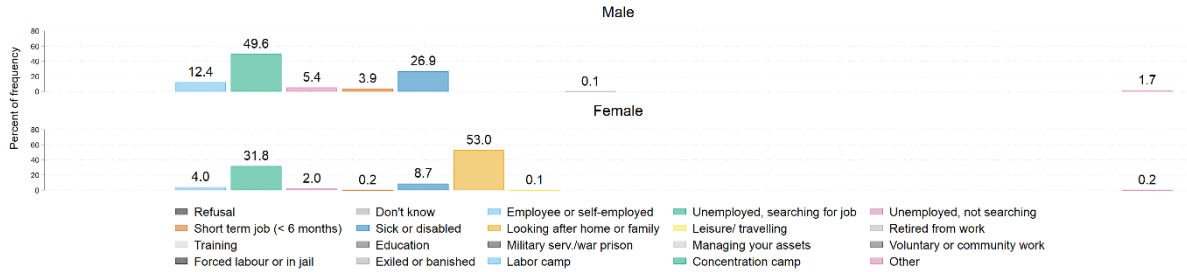
Distribution on states from raw SHARELIFE (re035) for person-year-spells with missing on *situation* in Portugal



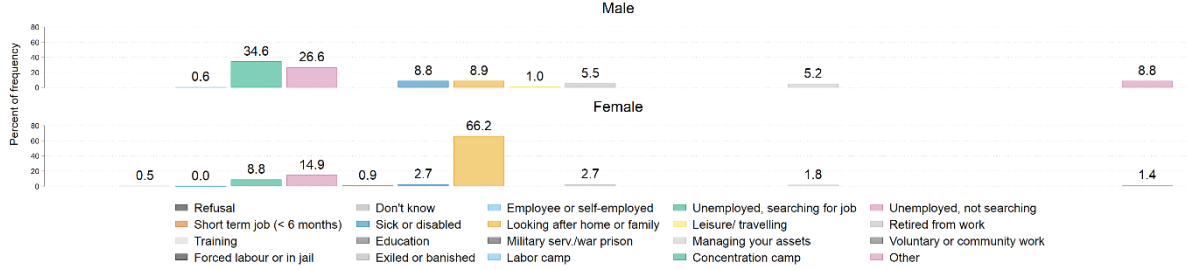
Distribution on states from raw SHARELIFE (re035) for person-year-spells with missing on *situation* in Romania



Distribution on states from raw SHARELIFE (re035) for person-year-spells with missing on *situation* in Slovakia



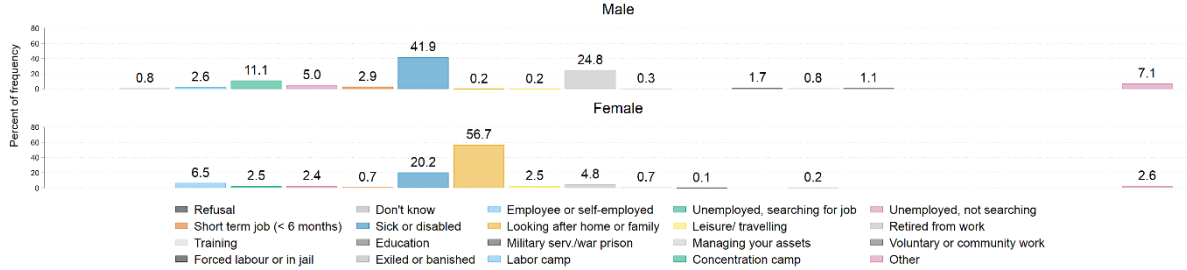
Distribution on states from raw SHARELIFE (re035) for person-year-spells with missing on *situation* in Slovenia



Distribution on states from raw SHARELIFE (re035) for person-year-spells with missing on *situation* in Spain



Distribution on states from raw SHARELIFE (re035) for person-year-spells with missing on *situation* in Sweden



Distribution on states from raw SHARELIFE (re035) for person-year-spells with missing on *situation* in Switzerland

