EPHOR

H2020 program	The Exposome Project for Health and Occupational Research
Grant agreement number	874703
Title	Mega Cohort Occupational Variables Harmonisation Protocol
Date	December, 2022
Responsible author	Ingrid Sivesind Mehlum
E-mail	ingrid.s.mehlum@stami.no
Co-authors	Karina Undem (STAMI)
	Calvin Ge (TNO)
	Michelle Turner (ISGlobal)

Background:

EPHOR (2020-2024) (https://www.ephor-project.eu/) seeks to lay the groundwork for evidence-based and cost-effective prevention of work-related disease and improving health at work and over the lifecourse, by developing a working life exposome toolbox, for researchers, occupational health practitioners, and policy makers. Within EPHOR, a mega cohort is being constructed based on large-scale pooling of data from multiple individual cohort studies, resulting in one of the largest pooling of occupational cohorts ever undertaken. Pooling of cohorts is essential to achieve large sample sizes to: 1) Move away from single exposure, single disease evaluations to the exposome-based exploration of (combinations of) risk factors in relation to multiple non-communicable diseases (NCDs), including rare exposures and rare diseases; and 2) Identify vulnerable life stages (e.g. young adult life, reproductive life, ageing life) and population subgroups (e.g. gender, socio-economic groups) in which these risk factors may result in more pronounced health effects. **The EPHOR mega cohort is currently comprised of 29 participating cohort studies** (though new cohorts may join) (Annex 1, see also google drive “Working Groups and Mailing Lists” for an up to date list of cohorts and contact persons). For meta-data information on the individual cohort studies, see the online Inventory of Occupational Cohort Studies (https://occupationalcohorts.net/). A summary overview of available cohorts and variables can be found in the google drive “Cohort Summary Information”. EPHOR will also work towards developing a more formal variable catalogue over time.

The overall objective of the EPHOR mega cohort is to evaluate occupational risk factors and the occurrence or severity of major NCDs over the life course, including cancers, respiratory, cardiovascular, metabolic and neurodegenerative diseases and musculoskeletal and mental disorders, as well as multimorbidity and work participation. The EPHOR mega cohort will provide new evidence of the impact of occupational exposures on the risk of major NCDs, through both **systematic** (targeted) and **agnostic** (exploratory) analyses of occupational exposures and risk factors across the lifecourse. EPHOR is also, in parallel, developing a dynamic **EuroJEM** (European Job-Exposure Matrix) permitting standardised assessment of multiple exposures in large populations across Europe. Initially, emphasis will be placed on analyses complementary to work performed in other EPHOR Work Packages.

Epidemiological analyses currently planned in EPHOR include descriptive and analytic studies by: 1) Job title; 2) EuroJEM defined exposure, also supplemented with other European JEMs; 3) Hierarchical analyses by job title and EuroJEM defined exposure; 4) Other measured/estimated exposures (on selected topics). Epidemiological analyses in the EPHOR mega cohort are planned using meta-analysis. EPHOR will also work towards a pooled decentralised infrastructure, where possible. DataSHIELD is an infrastructure and series of R packages that enables the safe, remote, and non-disclosive analysis of sensitive research data (see google drive “Datashield”).

The purpose of this mega cohort occupational variable harmonisation protocol is to provide a protocol for harmonisation of occupational variables in each participating cohort for participation in mega cohort analyses related to Job title and EuroJEM defined exposure. Protocols for harmonisation of outcome or other occupational or non-occupational variables will be provided as separate documents.

Occupational Variables, Overview:

Please see the Occupational Variable table below which provides details regarding the harmonised variables: variable name, label, format, unit for categorical variables, description, value definitions, comments, and vocabularies. **Only occupational codes and start and end year of each job entry are mandatory variables here.**

To help with the harmonisation of occupational coding, R scripts have been developed for crosswalking and checking of job codes.

Please contact Ingrid Sivesind Mehlum, STAMI: ingrid.s.mehlum@stami.no and Karina Udem, STAMI: Karina.Undem@stami.no if any comments or questions. For questions regarding R scripts themselves, please contact: Calvin Ge, TNO: calvin.ge@tno.nl.

Instructions on Harmonising Occupational Variables:

Introduction to ISCO:

The first version of the International Standard Classification of Occupations (ISCO) was prepared by the International Labour Organisation (ILO) in 1958 (ISCO-58) and has been revised: ISCO-68, ISCO-88, and ISCO-08 (the current (2008) version). Eurostat has developed and promoted a European variant, ISCO-88(COM), which differs somewhat from the international ISCO variant, e.g., by adding "*Public Service Administrative Professionals*" as a new "minor group" (the 3-digit level) and removing occupations that are not typical in Europe. The national occupational standards in European countries are usually based on the European standard. There are also other ISCO variants, such as ISCO-88(CIS) for the Commonwealth of Independent States.

The ISCO classification was developed for official statistics purposes. National Labour Market Administrations sometimes further adjust the ISCO classifications to satisfy local needs and national conditions. For instance, in some countries, country-specific occupations have been added or removed, or extra digits have been added to the existing codes (e.g., the usual 4-digit ISCO-88(COM) may have as many as 7 digits in a national variant).

National variants of the same ISCO standard may therefore differ, even if they are based on the same international and European ISCO variant. The number of occupational groups may differ, but the same ISCO codes may also be used for different occupations in different countries. This hampers comparisons and complicates pooling of data from different countries. There are also substantial differences between different revisions of the ISCO classification. Occupational groups may be added or removed, divided into more groups, or combined into fewer groups, which complicates pooling of data from different time periods.

To facilitate comparison and pooling, crosswalks have been created between ISCO *versions* (different revisions over time) and between some *variants* of the same ISCO version, e.g., between national variants and the corresponding European ISCO variant.

Occupational coding for the EPHOR mega cohort, using the European ISCO-88(COM):

The EPHOR mega cohort will include cohorts from different countries and time periods which may use different study designs and different types of occupational information, e.g., cohorts may be population-based, industry-based or registry-based, and information about occupation may be self-reported job titles (possibly coded later) or registry-based occupational codes. Occupational titles and codes will need to be harmonised to the same ISCO version and variant. In EPHOR, we will use **ISCO-88(COM) codes at the 4-digit level** for both the mega cohort and the EuroJEM. The official ISCO-88(COM) is available from Eurostat: https://ec.europa.eu/eurostat/documents/1978984/6037342/ISCO-88-COM.pdf.

Crosswalks between ISCO-88(COM) and some other occupational classifications can be downloaded from the EPHOR mega cohort google drive.

For occupational classifications where crosswalks have not been developed and for cohorts using job titles instead of occupational codes, manual coding to ISCO-88(COM) may be necessary. Correct coding of occupations at the 4-digit level may require more information than just the job title. Information about tasks and duties is often necessary to give an accurate classification but is rarely available. Information about education can be used to distinguish between occupations in Major groups *2 Professionals* and *3 Technicians and associate professionals*, and information about industry can be helpful to distinguish between occupations in Major groups *7 Craft and related trades workers* and *8 Plant and machine operators and assemblers*.

Harmonisation of occupation by each cohort – step-by-step

How is occupation classified in the cohort?

1. Coded in the European ISCO-88(COM)
2. Coded in the international (ISCO-88)
3. Coded in a coding system largely based on ISCO-88(COM) or ISCO-88, including national variants
4. Coded in any other version of ISCO (58, 68 or 08)
5. Coded in a coding system largely based on any other version of ISCO (58, 68 or 08), including national variants
6. Coded in a coding system that is not ISCO and not based on ISCO
7. Job titles plus additional occupation-related variables (e.g., industry, job tasks)
8. Job titles only

If occupational codes are used, it is important to confirm the coding system (ISCO or non-ISCO), as well as version and variant. This may be known by the cohort owner. If a national variant of ISCO is used, e.g. if the data are compiled from a national register, the national statistics office/bureau will likely know which version and variant are used and may also have developed a crosswalk between the national and the corresponding European/international variants.

Harmonisation of occupational codes

1. If occupation is coded using the European ISCO-88(COM), please check that this is correct by comparing the list of occupations in the cohort with Eurostat's list ISCO_88_COM (europa.eu). An R script is available in the shared drive for checking if all job codes within the dataset are in the correct coding of ISCO-88(COM).

2. If occupation is coded using the international (ISCO-88) variant, occupational codes should be recoded using a crosswalk between ISCO-88 and ISCO-88(COM), which can be

downloaded from the shared drive folder. There is also an R script available in the shared folder for crosswalking between ISCO-88 international and ISCO-88(COM).

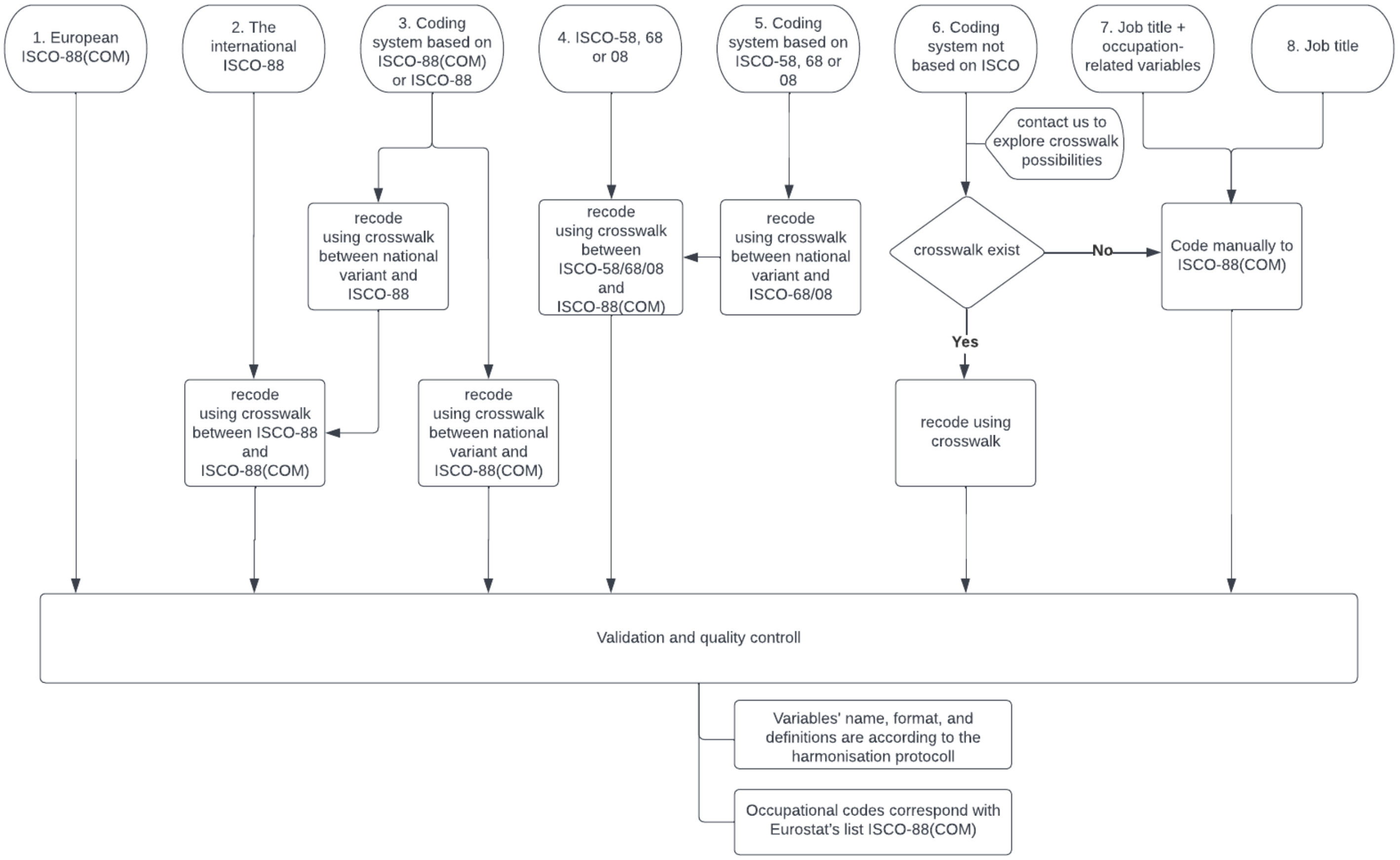
3. If occupation is coded using a national variant of ISCO-88, the job coding must be crosswalked to ISCO-88 (COM). You may do so using crosswalks already available for your national versions of ISCO-88, or use the R script available in the shared folder if the specific crosswalk is already available.

4. If occupation is coded in ISCO-58, ISCO-68 or ISCO-08, the job coding must be crosswalked to ISCO-88(COM). This may be done manually using available crosswalks from ILO or R scripts available in the shared folder that performs crosswalk first between ISCO-68/08 to ISCO-88, then ISCO-88 to ISCO-88(COM).

5. If occupation is coded using a national variant of ISCO-58, ISCO-68 or ISCO-08, the occupational codes should first be recoded to the respective international ISCO-version, then crosswalked from the international version to ISCO-88(COM) following step 4.

6. If occupation is coded in a coding system that is not ISCO and not based on ISCO, crosswalks may also be available. Please contact us to explore crosswalk possibilities.

7. & 8. If no occupational coding is used and only job titles (and additional occupation-related variables) are available, jobs should be coded into ISCO-88(COM) manually. See Eurostat's list of ISCO_88_COM (europa.eu). If manual coding is not feasible, please let us know.


1. European ISCO-88(COM)
2. The international ISCO-88
3. Coding system based on ISCO-88(COM) or ISCO-88
4. ISCO-58, 68 or 08
5. Coding system based on ISCO-58, 68 or 08
6. Coding system not based on ISCO
7. Job title + occupation-related variables
8. Job title
contact us to explore crosswalk possibilities
recode using crosswalk between national variant and ISCO-88
recode using crosswalk between ISCO-58/68/08 and ISCO-88(COM)
recode using crosswalk between national variant and ISCO-68/08
crosswalk exist
No
Yes
Code manually to ISCO-88(COM)
recode using crosswalk between ISCO-88 and ISCO-88(COM)
recode using crosswalk between national variant and ISCO-88(COM)
recode using crosswalk
Validation and quality controll
Variables' name, format, and definitions are according to the harmonisation protocoll
Occupational codes correspond with Eurostat's list ISCO-88(COM)

Formatting occupation variables

The table below shows the final harmonised occupational variables and the format of these variables needed.

If there are ISCO-88(COM) *codes with less than 4 digits*, include the most specific codes that are available, i.e., with the most digits (3-, 2- or 1-digit level with trailing zero(s) so that the job code always contains 4 characters). Note that the codes for major occupational group *0 Armed forces and unspecified*, start with 0.

If there are ISCO-88(COM) *codes with more than 4 digits*, include only the first 4 digits, but add in a comment about how many digits are available.

Occupation should be coded with one line per job with start and end year.

Validation and quality control of occupational coding

Check that the variables' name, format, and definitions are according to the harmonised occupational variable table below.

Check if occupational codes correspond with Eurostat's list ISCO_88_COM (europa.eu), including checking for incorrect values. An R-script is available in the shared drive for checking any incorrect and ISCO-88(COM) occupational codes.

Other occupational variables

In addition to occupational code, the EPHOR mega cohort will harmonise a few other occupational variables. For instance, ISCO-88(COM) does not distinguish between employers and employees (i.e., employment status), only tasks and duties. Employment status classification is therefore included as a separate variable, to be included, if available. Please see the Occupational Variable table below which provides details regarding the harmonised variables: variable name, label, format, unit for categorical variables, description, definitions, and additional comments and vocabularies.

R and RStudio

The latest versions of R and RStudio are needed to run the R scripts in Google Drive. Information about R and RStudio is available in Appendix A of the eBook: “Hands-On Programming with R” by Garret Grolemund (A Installing R and RStudio | Hands-On Programming with R (rstudio-education.github.io)).

Harmonised Occupational Variables

Name	Label	Format	unit	Description	Values	Comments	Vocabularies
Start_year	***Start year***	***numeric***	***Year***	***Start year of job entry***			
End_year	***End year***	***numeric***	***Year***	***End year of job entry***			
Occup	***European ISCO-88(COM)***	***factor***		***ISCO-88(COM) Occupation***		***Additional information about type of measure (e.g., current, lifetime, longest held, etc.)***	***ISCO-88(COM)***
empl_status	*Employment status*	*factor*		*Employment status*	*1 = employed* *2 = self-employed*		
Whours	*Working hours*	*numeric*	*hours*	*Number of hours per week*			
Permanency	*Job permanency*	*factor*		*Permanency of the job according to the job contract*	*1 = Permanent* *2 = Temporary*		

* Variables in **bold** are mandatory variables. Cohorts may include variables in *italics* if this information is available.

Example of Table of Harmonised Variables

studyperson_ID	**start_year**	**end_year**	**occup**	empl_status	whours	permanency
1	2003	2016	5163	1	40	1
2	2005	2007	2223	1	40	1
2	2007	2012	2223	2	.	.
3	2006	2007	4212	1	40	2
3	2007	2008	4212	1	40	2
3	2009	2009	4000	1	20	2
3	2009	2010	4000	1	5	.
3	2009	2012	3419	1	30	1