Title

Harmonizing work history data in epidemiologic studies with overlapping employment records.

Short title

Harmonizing overlapping employment records

Complete names (full first and last) and academic degrees of all authors

Jo Steinson Stenehjem*[1,2], PhD; Ronnie Babigumira*[1], MSc; Melissa C Friesen[3], PhD; Tom Kristian Grimsrud[1], MD, PhD.

Author's institutions (linked to names with superscript letters or numbers)

[1]Department of Research, Cancer Registry of Norway, Oslo, Norway
[2]Oslo Centre for Biostatistics and Epidemiology, University of Oslo, Oslo, Norway
[3]Occupational and Environmental Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, USA
***Co-first authors**

Institution at which the work was performed

The Cancer Registry of Norway and the University of Oslo

Name, mailing address, and email address for the corresponding author

Jo S Stenehjem, PhD
Cancer Registry of Norway, P.O. box 5313 Majorstuen, N-0304 Oslo, Norway
E-mail: jo.stenehjem@kreftregisteret.no
Phone: +4722451300
Fax: +4722451370

Authors' contributions: Authors must provide a statement identifying which authors participated in the a) conception or design of the work; b) the acquisition, analysis, or interpretation of data for the work; c) drafting the work or revising it critically for important intellectual content; d) final approval of the version to be published; and e) agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. [Please describe each author's contribution. Do NOT merely list the above letters; they are for guidance, not shorthand for crafting a statement.]

JSS and RB conceived the study, drafted the manuscript and are responsible for management and analysis of the data. RB developed the conceptual framework and conducted the data management and analysis. TKG is PI of the project and contributed with expertise in occupational epidemiology. MCF contributed with expertise in exposure assessment. TKG and MCF reviewed the manuscript and revised it critically for important intellectual content. All authors approved the final version for submission, and JSS and TKG are the guarantors. Each author believes that the manuscript represents honest work and accept the responsibility of its content. The authors declare that they have no financial or other relationships that might lead to a conflict of interest.

Acknowledgements: Authors may acknowledge contributors to the article but should not cite funding sources, which go in a following section.

We wish to thank our former Department Head Aage Andersen and Researcher Leif-Åge Strand for conducting the offshore survey, Advisor Tone Eggen for extracting the detailed work histories and mapping of job titles, and Statistician Haris Fawad for statistical advice (Cancer Registry of Norway). Further, we wish to thank Industrial Hygienists Prof. Magne Bråtveit, Dr. Jorunn Kirkeleit and Dr. Bjørg Eli Hollund (University of Bergen and Haukeland University Hospital, Norway), and Prof. John Cherrie (Heriot-Watt University, Edinburgh, UK) for development of the JEMs.

Funding: Authors cite funding source for work that is described in the article. Please list all grant information here using the following format: Grant sponsor:______; Grant number:_________. If the authors identify no funding source, then the following statement will be inserted: *The authors report that there was no funding source for the work that resulted in the article or the preparation of the article.*
This work was funded by a grant from the Research Council of Norway's PETROMAKS2 program (grant no 280537) to the Cancer Registry of Norway (JSS, RB, and TKG). MCF was funded by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, at the U.S. National Cancer Institute of the National Institutes of Health.

Institution and Ethics approval and informed consent: Authors must name the institution at which the work was performed, and cite human subjects' institutional review and approval, describing whether there was verbal or written informed consent. If there was no ethics review and approval and/or no informed consent, the authors state these facts and describe why approval and/or consent were not required or relevant. [A blinded version of this information (no institution and/or author names) should be included in the Methods section of the paper.]
This work was performed at the Cancer Registry of Norway and at the University of Oslo. Study participants signed an informed consent as part of this study's questionnaire survey. Necessary legal and ethical approvals were obtained from the Norwegian Data Inspectorate, the Regional Committee for Medical Research Ethics, and the Norwegian Directorate of Health.

Disclosure (Authors): The authors disclose conflicts of interest per AJIM policy. If the authors disclose a conflict, AJIM will also include a weblink to the authors' disclosure statements submitted to AJIM. If there is no conflict, the following statement should be inserted: *The authors declare no conflicts of interest.*
The authors declare no conflicts of interest.

Disclaimer: Authors may add a disclaimer in this section. If there is no disclaimer, the following word should be inserted: *None*
None

Word Counts: 159 words (abstract) and 2781 (main text)

ABSTRACT

Background: Work history data often require major data management including handling of overlapping jobs to avoid overestimating exposure before exposure linkage to job-exposure matrices (JEMs) is possible.

Methods: In a case-cohort study of 1825 Norwegian offshore petroleum workers, 3979 jobs were reported (mean duration 2417 days/job; maximum 8 jobs/worker). Each job was assigned to one of 27 occupation categories. Overlapping jobs of the same category (1142 jobs) were collapsed and overlapping jobs of different categories (1013 jobs) were split. The resulting durations were weighted by a factor accounting for the number of overlapping jobs.

Results: Collapsing overlapping jobs within the same category resulted in 3295 jobs (mean 2629 days/job). Splitting overlapping jobs of different categories increased the number to 4239 jobs (mean 2043 days/job), while the total duration in days dropped with 10%.

Conclusions: We demonstrated that overlapping employment data structures can be harmonized in a systematic and unbiased way, preparing work history data for linkage to several JEMs.

Keywords: Work history, employment spells, exposure assessment, job-exposure matrices, occupational epidemiology.

INTRODUCTION

In epidemiologic studies of occupational risk factors, individual exposure estimation often relies on work-histories comprising data on job title, and start and stop dates for each individual's employments. Work histories are commonly obtained from either company or union employment records in industry-based studies, or from self-reported questionnaires in population-based studies.(1) Although work history data have been proven to be fairly accurate,(2-4) they often present major data management challenges. Mapping job titles into standardized categories, and harmonization of complex employment data structures are prerequisite tasks before application of job-exposure matrices (JEMs) or other exposure assessment approaches is possible.

Various methods have been used to ease and improve the laborious way of manually mapping of job titles into aggregate categories or occupational codes, which form the basis of identifying occupational risk factors in epidemiological studies. In the early 1990s, Loomis et al. used a combination of computer algorithms and expert judgment to assign individual job titles into 28 job categories in a study of 135,000 electric power industry workers.(5) More recently Russ et al., showed how computer-based coding of free-text job descriptions can be used to efficiently identify occupations in epidemiological studies.(6) However, the data management phase that follows mapping job titles into aggregate categories, namely the handling of overlapping jobs has, to our knowledge, not been described previously in the occupational epidemiology literature.

From a data management point of view, the ideal data structure would be cascading work-histories in which, at any given period, an individual is either unemployed or employed, and that the preceding job stops before the new one starts (*i.e.* no overlapping jobs). However, real life work-histories are more complex and often comprise multiple overlapping jobs. This is particularly the situation for industries with long touring patterns or where parallel positions are common (*e.g.* offshore petroleum, farming, and shipping industries), resulting in overlapping jobs.(7) If jobs with overlapping employment periods are not resolved, exposures can be grossly overestimated as the overlapping duration would be counted multiple times. A manual cleaning procedure may also be prone to personal preferences, random judgement, and errors.

In the present study we handled work history data from a case-cohort dataset of 1825 Norwegian offshore petroleum workers who reported up to 8 jobs per worker. We demonstrate how overlapping employment periods were harmonized by collapsing jobs within the same category and by splitting jobs of different categories into proportionally equal parts before linkage to JEMs.

METHODS

Study population

In 1998, the Cancer Registry of Norway conducted a questionnaire-based survey among active and former offshore petroleum workers and established a cohort of 25,413 men who confirmed that they had worked offshore on the Norwegian continental shelf for at least 20 days between January 1, 1965 and December 31, 1998 (inclusion criterion). The workers were asked to provide details of all (or 8 most recent) employments. The questionnaire was limited to 8 jobs per worker based on consensus in the project reference group (*i.e.* experts from the petroleum industry, unions and the Norwegian Petroleum Safety Authorities), that few workers would have more jobs to report. For petroleum workers with more than two offshore jobs, information had to be extracted manually from the questionnaires. In order to limit costs, this was done for a random subsample of the cohort (*i.e.* subcohort), and for all skin cancer patients according to a stratified case-cohort design.(8)

The study design and study population have been described in detail in previous publications on skin cancer risk associated with exposure to hydrocarbons, ionizing radiation, and ultraviolet radiation.(9, 10) In the present paper, we used the same case-cohort data set of 1825 workers (including 182 skin cancer cases and 1643 subcohort members) with individual work history data (start year, stop year, job title). A total of 36 workers (1.97%) in the case-cohort dataset reported 8 jobs, meaning that the fraction with potentially >8 jobs was small, and hence that the risk of missing employment data was small. The self-reported job titles in the case-cohort set were mapped into 27 aggregate job categories modified from the Standardized Occupational

Coding system based on communication with the project reference group. These job categories were used to develop JEMs (*e.g.* for hydrocarbons and radiation), specifically for this cohort.(11)

Study participants signed an informed consent when returning the questionnaire. Necessary legal and ethical approvals were obtained from Norwegian Data Inspectorate, the Regional Committee for Medical Research Ethics, and the Norwegian Directorate of Health. Fictional job categories were used in two work history examples to not disclose the identity of these workers.

Conceptual framework

To illustrate the structure and management of our data, consider the fictional illustration of an individual's full work history (Figure 1) comprising four jobs (distinct categories coded as 1, 2, 3 and 4). The underlying data structure, known as event-time data structure,(12) has two components (1) an event and (2) the event-time. For the present study, an event is the job and the event-time refers to the duration for each job. Each horizontal line represents a "normal" continuous employment period and does not include significant breaks (such as unemployment or prolonged sick leave). The prolonged breaks would be gaps in the dataset. Drawing on Blossfeld and Cox, we use three concepts to frame the data management challenges of work history data (as displayed in Figure 1), namely *states*, *spells*, and *duplicates.*(12, 13)

States

State refers to a unique description of what a person is doing at any given point in their work history. One can only be in one of three states at a time: s*tate 1* unemployed (Figure 1, panel B), *state 2* employed with no overlapping jobs (panels A and E), or, *state 3* employed with overlapping jobs (panels C and D). We did not address *state 1* (gaps) because we did not assign any exposure to the periods of unemployment. In this paper we focus on *state 3* and how to resolve overlaps.

Spells

Cox defined spells as periods that are homogenous in some sense.(13) In our setting, spells can be thought of as nuanced states where a spell indicates the job's employment period. The concept of spells is useful to identify and isolate the employment periods by whether or not they overlap.

Figure 1 illustrates that a (fictional) worker may have multiple spells. The 5 spells for this worker are denoted by the letters above the graph. *A* is the first employment spell for job 1 (here recorded twice), *B* is an unemployment spell, *C* is the first employment spell for jobs 2 and 4, *D* is the second spell for jobs 2 and 4 and the first spell for job 3, and *E* is the second spell for job 3. The data management task is to clearly delineate where complex spells start and stop. By taking into account and adjusting for the fact that the expanded data represents shared exposure between the overlapping jobs in the section, exposure estimates are not overestimated.

Duplicates

Finally, duplicates are defined as spells with identical job categories and identical start and stop dates. In Figure 1, we have two entries under job category 1. These are typically errors made by respondents as they filled in the questionnaire, or, later, during data entry, and it needs to be treated to avoid a doubling of the exposure.

Data management and harmonization

The first task was to identify and remove jobs with duplicate spells, as defined above. Having removed duplicates, we split the data into two parts, the first half comprised data in s*tate 2* (no overlaps) and the second, data in *state 3* (overlaps).

The overlapping periods in *state 3* were further classified in two, (1) "overlaps between same job categories", which is the case when two or more spells overlap but they belong to the same job category, and (2) "overlaps between different job categories", which is when two or more spells, from different job categories, overlap. The "overlaps between same job categories", which were primarily the result of the mapping into aggregate job categories, were collapsed into a single spell with the start and stop being the earliest and the latest dates, respectively, in the overlapping contiguous spells.

To resolve the "overlaps between different job categories", we identified the start and stop of each spell with complete or partly overlap and then split the data, leaving behind an expanded dataset inclusive of spells with exact overlap. We then weighted the duration for each spell in the overlapping spans by an adjustment factor based on the total number of overlapping spells in the span, so that the resulting exposure within each employment spell would be derived according to the following the formula:

$$E_{jst} = \frac{T_s}{J_s} * d_{jt}$$

Where

E_{jst} = Exposure (*E*), job (*j*), spell (*s*), and time (*t*) specific exposure

T_s = Duration of spell

J_s = Number of jobs in spell

d_{jt} = Job and time specific exposure ratings

Data management was performed using Stata version 15.1 (StataCorp, College Station, TX, USA), and the Stata module –splitit–.(14)

RESULTS

Table 1 shows descriptive statistics of the work history by stage of data harmonization. Because the number of jobs was constant but the spell length changed with stage of harmonization, the column “Before data cleaning” shows statistics for *jobs*, but the columns “After collapsing” and “After splitting” show statistics for *employment spells*. After removal of duplicates (n=623), the total number of jobs before data cleaning was 3979, with 2.2 jobs per worker on average and 2417 days of average duration per job. After collapsing 1142 overlapping employment spells within the same job category, the number of spells reduced by 684 to a total of 3295 and an average of 1.8 spells per worker. The duration, however, increased to an average of 2629 days per employment spell because the number of spells was reduced. After splitting the 1013 employment spells that were overlapping between different job categories, the total number of employment spells increased by 944 to 4239 spells, with an average of 3.0 spells per worker, and reduced the average duration of 2043 days per employment spell. The total duration in the dataset dropped by 10% from before data cleaning (9,618,646 days) to after splitting (8,658,953 days).

Figure 2 shows an example of overlapping work history between same categories. The original data yielded four jobs as an electrician for this worker. During harmonization, we identified the earliest start and the latest stop of these four jobs, and then collapsed them into one period spanning the full length.

Figure 3 shows an example of overlapping work history between different job categories. This worker’s original data showed four jobs of different categories; industrial cleaner, control room operators, catering workers and radio employees. During harmonization, the jobs with overlapping time periods were split into spells of equal duration, resulting in new rows for each overlapping time period.

Table 2 shows the data underlying Figure 3. The worker had 4 jobs spanning the period 1975–1996. To link this work history to the JEM ratings, we expanded the data from the paired (*i.e.* start, stop, job title) to a time series format. The crude pre-harmonization data shows the number of days for each job by year as well as the associated exposure. Without taking into account the fact that some of these jobs overlapped in some periods, the total duration would be summed to 14,144 days compared to 7665 days after adjustment. The exposure would be summed up to 45, 71, 18, and 62 for benzene, crude oil, ionizing radiation, and mineral oil, respectively. After accounting for the overlaps using the approach we described earlier, we see that the pre-harmonization exposure ratings are on average 73 % higher (range 47-89) than the adjusted values. The post-harmonization data structure, made it convenient to calculate exposure duration, cumulative exposure, and average intensity of exposure for each of the four agents without overestimating exposure due to overlaps.

DISCUSSION

Complex work histories are an important issue in any epidemiological study involving exposure assessment of occupational risk factors. The issue is especially likely to arise in industries where workers can have multiple jobs running concurrently (often with different employers) or among shift workers, resulting in overlapping employment periods. It will also arise when work histories include secondary part- or full-time jobs. In the present paper, we sought to address this by applying a conceptual framework and a systematic procedure for handling work history data with overlapping jobs. Assigning ratings from different JEMs to overlapping jobs first required collapsing jobs within the same category and then splitting jobs overlapping between different categories before we were able to assign the JEM-rating to the correct duration and time period for each job.

To demonstrate how we handled overlaps, we used data from a cohort of Norwegian offshore petroleum workers as examples. In our cohort, the vast majority (65%) was employed in a contracting company, 32% in an operating company, and 3% did not report on type of company.(7, 15) Contractors usually performed highly specialized tasks (*e.g.* industrial cleaning, drilling, electrical work) that may have lasted from days to months serving different oil and gas companies within the same time period. Such work schedules may thus have led to what we have termed spells that were “overlapping between same job categories”. Workers who had parallel employments in different categories, what we termed “overlapping between different categories”, may have been more common on the larger platforms. Larger platform clusters (*e.g.* the Ekofisk complex) often had drilling, production activities and accommodation facilities, requiring the labour force to perform several operations within a very constrained physical area. Hence, the issue of overlapping employment spells may arise particularly in industries where parallel positions are common or when participants may have more than one employer at a time. However, the conceptual approach we present is generic and will apply to any handling of event-history data with overlapping spell structures.(13)

The main advantage of this approach is that overlaps are handled following a systematic procedure, limiting the potential for exposure overestimation that would result from overlapping duration being counted multiple times, as illustrated by the crude and adjusted example in Table 2. In turn, overestimation and misclassification of exposure could lead to biased risk estimates of disease. The direction of such bias would depend whether or not exposure was differentially misclassified, and the number of exposure categories.(16-18) An alternative approach with manual resolution of overlapping spells would easily lead to errors and be prone to personal preferences and misjudgment.

An important assumption of the splitting approach we present is that that each overlapping employment spell contributes with an equal proportion of the time used in each job category. That is, if a worker has four overlapping employment spells over a given time period, we assume that he or she is using 25% of the time in each spell. This assumption is not likely to hold for all workers in our cohort, and will in such cases misclassify exposure when the work history is linked to JEMs. However, this misclassification should be nondifferential because neither case status nor exposure status is taken into account in the duration weighting. Also, in other settings where workers have secondary jobs contributing significantly to the overall exposure time, the number of hours worked per day should be recorded as part of the work history, and be factored into the exposure metric calculations.

Another approach, as suggested by Kröger, would be to rank the different overlapping spells,(19) and use the duration from the most relevant spell with respect to the exposure and disease in question to generate exposure duration for time periods with overlaps. For our recent paper on exposure to hydrocarbons and ionizing radiation in relation to skin cancer risk,(10) we considered the ranking approach in the initial phase of data management. However, since we applied four different JEMs (benzene, crude oil, mineral oil and ionizing radiation) to the work history data, a ranking approach would give priority to one exposure over the other when JEM ratings differed between overlapping spells of different job categories. We made an *a priori* decision of giving each exposure equal priority and therefore opted for splitting instead of ranking. Also, with the splitting approach, the work history data were harmonized and ready in one procedure for linkage to any JEM and no exposure-specific considerations to the handling of work history data were needed.

In summary, we show how overlapping employment spell structures can be harmonized in a way that minimizes bias and prepares work history data for linkage to several JEMs, using data from a cohort of Norwegian offshore petroleum workers to give examples. This systematic procedure is thought to be a supplement to existing methodological tools that handles mapping of job titles into aggregate categories.

REFERENCES

1. Friesen MC, Lavoue J, Teschke K, van Tongeren M. Occupational exposure assessment in industry- and population-based epidemiologic studies. In: Nieuwenhuijsen MJ, editor. Exposure Assessment in Environmental Epidemiology 2nd Edition. 2nd ed. Oxford, England: Oxford University Press; 2015.
2. Kromhout H, Vermeulen R. Application of job-exposure matrices in studies of the general population-some clues to their performance. European Respiratory Review. 2001;11(80):80-90.
3. Teschke K, Olshan A, Daniels J, De Roos A, Parks C, Schulz M, et al. Occupational exposure assessment in case–control studies: opportunities for improvement. Occupational and environmental medicine. 2002;59(9):575-94.
4. Bourbonnais R, Meyer F, Theriault G. Validity of self reported work history. British journal of industrial medicine. 1988;45(1):29-32.
5. Loomis DP, Peipins LA, Browning SR, Howard RL, Kromhout H, Savitz DA. Organization and classification of work history data in industry-wide studies: An application to the electric power industry. American journal of industrial medicine. 1994;26(3):413-25.
6. Russ DE, Ho K-Y, Colt JS, Armenti KR, Baris D, Chow W-H, et al. Computer-based coding of free-text job descriptions to efficiently identify occupations in epidemiological studies. Occupational and environmental medicine. 2016.
7. Stenehjem JS, Friesen MC, Eggen T, Kjaerheim K, Bratveit M, Grimsrud TK. Self-reported Occupational Exposures Relevant for Cancer among 28,000 Offshore Oil Industry Workers Employed between 1965 and 1999. Journal of occupational and environmental hygiene. 2015;12(7):458-68.
8. Borgan O, Langholz B, Samuelsen SO, Goldstein L, Pogoda J. Exposure stratified case-cohort designs. Lifetime data analysis. 2000;6(1):39-58.
9. Stenehjem J, Robsahm T, Bråtveit M, Samuelsen S, Kirkeleit J, Grimsrud T. Ultraviolet radiation and skin cancer risk in offshore workers. Occupational Medicine. 2017;67(7):569-73.
10. Stenehjem JS, Robsahm TE, Bråtveit M, Samuelsen SO, Kirkeleit J, Grimsrud TK. Aromatic hydrocarbons and risk of skin cancer by anatomical site in 25 000 male offshore petroleum workers. American journal of industrial medicine. 2017;60(8):679-88.
11. Steinsvåg K, Bratveit M, Moen BE. Exposure to carcinogens for defined job categories in Norway's offshore petroleum industry, 1970 to 2005. Occupational and environmental medicine. 2007;64(4):250-8.
12. Blossfeld H-P. Event history analysis with Stata: Psychology Press; 2012.
13. Cox NJ. Speaking stata: Identifying spells. Stata Journal. 2007;7(2):249-65.
14. Erhardt K, Kuenster R. SPLITIT: Stata module to split chronological overlapping spells in spell data. 2017.
15. Strand LÅ, Andersen A. Kartlegging av kreftrisiko og årsaksspesifikk dødelighet blant ansatte i norskoffshore virksomhet – innsamling av bakgrunnsdata og etablering av kohort. [Identification of cancer risk and cause-specific

mortality among employees in the Norwegian offshore oil industry – data retrieval and cohort establishment]. Research report. Cancer Registry of Norway; 2001. Report No.: 1501-5831.
16. Dosemeci M, Wacholder S, Lubin JH. Does nondifferential misclassification of exposure always bias a true effect toward the null value? American journal of epidemiology. 1990;132(4):746-8.
17. Wacholder S, Dosemeci M, Lubin JH. Blind assignment of exposure does not always prevent differential misclassification. American journal of epidemiology. 1991;134(4):433-7.
18. Wacholder S, Hartge P, Lubin JH, Dosemeci M. Non-differential misclassification and bias towards the null: a clarification. Occupational and environmental medicine. 1995;52(8):557.
19. Kröger H. Newspell—Easy management of complex spell data. Stata J. 2015;15:155-72.

TABLES

Table 1. Descriptive statistics of work history stratified by stages of data harmonization in a case-cohort dataset of 1825 male Norwegian offshore petroleum workers.

	Before data cleaning[a]	After collapsing[b,c]	After splitting[b,d]
Jobs[a] / Employment spells[b]			
Total number in dataset, n	3979	3295	4239
Number per worker, mean (range)	2.18 (1-8)	1.81 (1-8)	2.99 (1-24)
Total duration (days) in dataset, n	9,618,646	8,661,077	8,658,953
Duration (days), mean (range)	2417 (30-12236)	2629 (30-12236)	2043 (0-12236)
By job category, n (%)			
Catering main category	129 (3.24)	110 (3.34)	142 (3.35)
Catering workers	97 (2.44)	82 (2.49)	102 (2.41)
Chefs	84 (2.11)	57 (1.73)	71 (1.67)
Control room operators	98 (2.46)	89 (2.7)	112 (2.64)
Could not be categorized	41 (1.03)	36 (1.09)	55 (1.3)
Deck crew	355 (8.92)	292 (8.86)	361 (8.52)
Derrick employees	72 (1.81)	66 (2)	99 (2.34)
Drill floor crew	254 (6.38)	211 (6.4)	285 (6.72)
Drillers	153 (3.85)	118 (3.58)	154 (3.63)
Drilling main category	122 (3.07)	114 (3.46)	188 (4.44)
Electric instrument technicians	107 (2.69)	91 (2.76)	122 (2.88)
Electricians	266 (6.69)	205 (6.22)	223 (5.26)
Health, office and administration personnel	304 (7.64)	238 (7.22)	324 (7.64)
Insulators	30 (0.75)	27 (0.82)	34 (0.8)
Invalid answer	4 (0.1)	4 (0.12)	7 (0.17)
Laboratory engineers	3 (0.08)	3 (0.09)	3 (0.07)
MWD and mud loggers/engineers	25 (0.63)	22 (0.67)	25 (0.59)
Machinists	134 (3.37)	112 (3.4)	139 (3.28)
Maintenance main category	607 (15.26)	505 (15.33)	648 (15.29)
Maritime workers	94 (2.36)	89 (2.7)	114 (2.69)
Mechanics	240 (6.03)	188 (5.71)	248 (5.85)
Non-destructive testing	13 (0.33)	11 (0.33)	14 (0.33)
Plumbers and piping engineers	81 (2.04)	66 (2)	84 (1.98)
Process technicians A	18 (0.45)	15 (0.46)	24 (0.57)
Process technicians B	160 (4.02)	137 (4.16)	167 (3.94)
Production main category.	104 (2.61)	101 (3.07)	129 (3.04)
Radio employees	97 (2.44)	75 (2.28)	82 (1.93)
Scaffold crew	40 (1.01)	32 (0.97)	37 (0.87)
Shale shaker operators	1 (0.03)	1 (0.03)	1 (0.02)
Sheet metal workers	21 (.53)	16 (0.49)	24 (0.57)
Surface treatment (painters)	70 (1.76)	55 (1.67)	65 (1.53)
Turbine operators and hydraulics technincians	5 (0.13)	5 (0.15)	8 (0.19)
Welders	108 (2.71)	85 (2.58)	106 (2.5)
Well service crew	42 (1.06)	37 (1.12)	42 (0.99)
Catering main category	129 (3.24)	110 (3.34)	142 (3.35)

[a]Shows statistics for jobs

[b]Show statistics for employment spells

[c]Shows statistics after collapsing overlapping employment spells within the same job category

[d]Shows statistics after splitting and weighting employment spells that were overlapping between different job-categories

Table 2. Data structure pre- and post-harmonization*. This example is based on the same individual's work history showed in Figure 3; an individual in a cohort of Norwegian offshore petroleum workers. Fictional values were used to maintain data confidentiality.

			Pre-harmonization Crude exposure data					Post-harmonization Adjusted exposure data				
ID	Year	Job Category	Days	Benz.	Cru. Oil	Min. Oil	Ion. Rad.	Days	Benz.	Cru. Oil	Min. Oil	Ion. Rad.
0000	1975	Industrial cleaner	184	0,71	1,51	0,50	1,01	184	0,71	1,51	0,50	1,01
0000	1976	Industrial cleaner	366	1,40	3,00	1,00	2,00	366	1,40	3,00	1,00	2,00
0000	1977	Industrial cleaner	365	1,40	3,00	1,00	2,00	365	1,40	3,00	1,00	2,00
0000	1978	Industrial cleaner	365	1,40	3,00	1,00	2,00	365	1,40	3,00	1,00	2,00
0000	1979	Industrial cleaner	365	1,40	3,00	1,00	2,00	365	1,40	3,00	1,00	2,00
0000	1980	Industrial cleaner	366	1,40	3,00	1,00	2,00	243	0,93	2,00	0,67	1,33
0000	1980	Control room operators	184	0,23	0,47	0,14	0,37	61	0,08	0,16	0,05	0,12
0000	1980	Catering workers	184	0,96	1,01		1,01	61	0,32	0,34		0,34
0000	1981	Industrial cleaner	365	1,40	3,00	1,00	2,00	122	0,47	1,00	0,33	0,67
0000	1981	Control room operators	365	0,45	0,93	0,27	0,73	122	0,15	0,31	0,09	0,24
0000	1981	Catering workers	365	1,90	2,00		2,00	122	0,63	0,67		0,67
0000	1982	Industrial cleaner	365	1,40	3,00	1,00	2,00	122	0,47	1,00	0,33	0,67
0000	1982	Control room operators	365	0,45	0,93	0,27	0,73	122	0,15	0,31	0,09	0,24
0000	1982	Catering workers	365	1,90	2,00		2,00	122	0,63	0,67		0,67
0000	1983	Industrial cleaner	365	1,40	3,00	1,00	2,00	122	0,47	1,00	0,33	0,67
0000	1983	Control room operators	365	0,45	0,93	0,27	0,73	122	0,15	0,31	0,09	0,24
0000	1983	Catering workers	365	1,90	2,00		2,00	122	0,63	0,67		0,67
0000	1984	Industrial cleaner	366	1,40	3,00	1,00	2,00	122	0,47	1,00	0,33	0,67
0000	1984	Control room operators	366	0,45	0,93	0,27	0,73	122	0,15	0,31	0,09	0,24
0000	1984	Catering workers	366	1,90	2,00		2,00	122	0,63	0,67		0,67
0000	1985	Industrial cleaner	365	1,40	3,00	1,00	2,00	122	0,47	1,00	0,33	0,67
0000	1985	Control room operators	365	0,45	0,93	0,27	0,73	122	0,15	0,31	0,09	0,24
0000	1985	Catering workers	365	1,90	2,00		2,00	122	0,63	0,67		0,67
0000	1986	Industrial cleaner	90	0,34	0,74	0,25	0,49	30	0,11	0,25	0,08	0,16
0000	1986	Control room operators	365	0,45	0,93	0,27	0,73	168	0,21	0,43	0,12	0,33
0000	1986	Catering workers	365	1,90	2,00		2,00	168	0,87	0,92		0,92
0000	1987	Control room operators	365	0,45	0,93	0,27	0,73	183	0,23	0,47	0,14	0,37
0000	1987	Catering workers	365	1,90	2,00		2,00	183	0,95	1,00		1,00
0000	1988	Control room operators	366	0,45	0,93	0,27	0,73	183	0,23	0,47	0,14	0,37
0000	1988	Catering workers	366	1,90	2,00		2,00	183	0,95	1,00		1,00
0000	1989	Control room operators	365	0,45	0,93	0,27	0,73	183	0,23	0,47	0,14	0,37
0000	1989	Catering workers	365	1,90	2,00		2,00	183	0,95	1,00		1,00
0000	1990	Control room operators	365	0,39	0,93	0,27	0,73	183	0,20	0,47	0,14	0,37
0000	1990	Catering workers	365	1,60	2,00		2,00	183	0,80	1,00		1,00
0000	1991	Control room operators	365	0,39	0,93	0,27	0,73	183	0,20	0,47	0,14	0,37
0000	1991	Catering workers	365	1,60	2,00		2,00	183	0,80	1,00		1,00
0000	1992	Control room operators	178	0,19	0,45	0,13	0,35	89	0,09	0,23	0,07	0,18
0000	1992	Catering workers	178	0,78	0,97		0,97	89	0,39	0,48		0,48
0000	1992	Radio employees	184	0,35	0,50	0,50	1,01	184	0,35	0,50	0,50	1,01
0000	1993	Radio employees	365	0,70	1,00	1,00	2,00	365	0,70	1,00	1,00	2,00
0000	1994	Radio employees	365	0,70	1,00	1,00	2,00	365	0,70	1,00	1,00	2,00
0000	1995	Radio employees	365	0,70	1,00	1,00	2,00	365	0,70	1,00	1,00	2,00
0000	1996	Radio employees	181	0,35	0,50	0,50	0,99	181	0,35	0,50	0,50	0,99
		Sum	14144	45	71	18	62	7665	24	40	12	36

Abbreviations: Adj. = Adjusted; Benz. = Benzene; Cru. = Crude; Ion. = Ionizing; Min. = Mineral.

*Splitting employment spells that were overlapping between different job-categories

FIGURE LEGENDS

Figure 1. Fictional illustration of an individual's full work history comprising four employments (1, 2, 3 and 4). The dark area indicate unemployment (state 1), and the light grey areas indicate employment (states 2 & 3). The panels (A, B, C, D and E) indicate duration (or start and stop) of each employment spell. Panel A shows duplicates. Formula E_{jst} in panel C and D: Exposure (*E*), job (*j*), spell (*s*), and time (*t*) specific exposure. J_{2C} and J_{2D} indicate that employment spell C and D constitute the 2nd job.

Figure 2. Transition from original data with overlapping work history between same job categories to harmonized data where overlaps are collapsed into one employment spell. This example is based on the work history of an individual in a cohort of Norwegian offshore petroleum workers, but show fictional values to maintain data confidentiality.

Figure 3 Transition from original data with overlapping work history between different job categories to harmonized data where overlaps are split into equal employment spells. This example is based on the work history of an individual in a cohort of Norwegian offshore petroleum workers, but show fictional values to maintain data confidentiality.