Appendix 2

Calibration of the dataset

In a sample survey such as in the present study, estimates are always subject to selection errors, because only a subset (sample) of the population is studied. Another error occurs if we fail to get answers from all people in the sample (dropout) and if those who do not respond differ from the respondents with respect to survey variables.

To facilitate interpretation of the data, it is valuable to assess the size of these errors. Of the aforementioned error types, only the magnitude of the selection error can be estimated using selection information. Knowledge about the dropout error can usually only be obtained in an indirect and approximative way by utilising registry variables. Both selection errors and dropout errors can be reduced by using an effective enumeration procedure.

In the present study, both selection errors and dropout errors have been analysed and corrected for through the creation of a weighting variable, which was applied when deriving the frequency statistics presented in the results section. Statistics Sweden conducted the analyses and created the weighting variable.

The first step in creating the weighting variable was to establish the auxiliary variables that most accurately could correct for the selection and dropout errors (Lundström och Särndal, 2001). These auxiliary variables were selected according to three criteria:

1. The auxiliary variable correlates well with the probability of response. This is the most important criterion as it leads to a reduction in the skewness due to dropout for all estimates
2. The auxiliary variable correlates well with important target variables. If so, the dropout bias and variance decrease for those estimates on which these target variables are based
3. The auxiliary variable delimits important reporting groups. This leads to reduced variance in estimates for these groups.

Conceivable auxiliary variables were derived from the Swedish Total Population Register (RTB), the Longitudinal Integrated Database for Health Insurance and Labour Market Studies (LISA), the Income and Taxation Register (IoT), the Swedish Education Register (UREG) and the The National Patient Register. Table 1 gives the variables used and their categories.

Table 1 Conceivable auxiliary variables

Variable name	Categories
Gender	1 = Male 2 = Female
Age	1 = 18-29 2 = 30-39 3 = 40-49 4 = 50-59 5 = 60-69 6 = 70-75
Country of birth	1 = Sweden 2 = Other European country

	3 = Outside Europe
Marital status	1 = Married or registered partner 2 = Others
Income (year 2010, thousands of SEK*)	1 = None (or not included in IoT 2010) 2 = 85 - 159 3 = 160 - 309 4 = 310 -
Educational level	1 = Below secondary (incl. unknown/missing) 2 = Secondary 3 = Post-secondary
Region of residence in Sweden	1 = Suburbs of major cities (Grp1) 2 = Suburbs of large cities (Grp2) 3 = Sparsely populated municipalities (Grp3) 4 = Municipalities in sparsely populated region (Grp4) 5 = Municipalities in densely populated region (Grp5) 6 = Commuter municipalities (Grp6) 7 = Major cities (Grp7) 8 = Large cities (Grp8) 9 = Tourism municipalities (Grp9) 10 = Manufacturing municipalities (Grp10)
Social welfare benefits	1 = Collection of social welfare benefits 2005-09 2 = No collection of social welfare benefits 2005-09
Early retirement benefits	1 = Collection of early retirement benefits 2005-09 2 = No collection of early retirement benefits 2005-09
Unemployment	1 = Unemployed any time 2005-09 2 = Not unemployed 2005-09
Illness	1 = Illness registered 2005-09 2 = No illness registered 2005-09
Work-related injury	1 = Work-related injury 2005-09 2 = No work-related injury 2005-09
Primary care utilisation	1 = 0-3 visits during 2006-10 2 = 4-11 visits during 2006-10 3 = more than 11 visits during 2006-10
Hospitalisation	1 = No hospital admissions during 2006-10 2 = One hospital admission during 2006-10 3 = More than one hospital admission during 2006-10

*Average value of one SEK was 0.14 USD in 2010.

Criterion 1

In order to assess whether the presumptive auxiliary variables fulfilled critertion 1, correlations were examined between the dichotomised variable responder/non-responder (dropout) and each auxiliary variable (Table 2). The design weight (N_h/n_h) was applied in these estimations.

Table 2 Percent of individuals responding to the survey according to possible auxiliary variables

Variable name	Categories	% responding

Appendix 2

Gender	1 = Male 2 = Female	46.7 57.1
Age	1 = 18-29 2 = 30-39 3 = 40-49 4 = 50-59 5 = 60-69 6 = 70-75	40.9 47.0 49.3 58.4 62.7 58.5
Country of birth	1 = Sweden 2 = Other European country 3 = Outside Europe	55.9 39.1 24.4
Marital status	1 = Married or registered partner 2 = Others	58.9 46.6
Income (year 2010, thousands of SEK*)	1 = None (or not included in IoT 2010) 2 = 85 - 159 3 = 160 - 309 4 = 310 -	32.1 42.7 54.7 62.4
Educational level	1 = Below secondary (incl. unknown/missing) 2 = Secondary 3 = Post-secondary	39.2 49.8 62.0
Region of residence in Sweden	1 = Suburbs of major cities (Grp1) 2 = Suburbs of large cities (Grp2) 3 = Sparsely populated municipalities (Grp3) 4 = Municipalities in sparsely populated region (Grp4) 5 = Municipalities in densely populated region (Grp5) 6 = Commuter municipalities (Grp6) 7 = Major cities (Grp7) 8 = Large cities (Grp8) 9 = Tourism municipalities (Grp9) 10 = Manufacturing municipalities (Grp10)	48.3 50.7 52.8 51.6 54.2 51.6 53.3 49.3 54.1 53.3
Social welfare benefits	1 = Collection of social welfare benefits 2005-09 2 = No collection of social welfare benefits 2005-09	31.1 54.0
Early retirement benefits	1 = Collection of early retirement benefits 2005-09 2 = No collection of early retirement benefits 2005-09	44.9 52.5
Unemployment	1 = Unemployed any time 2005-09 2 = Not unemployed 2005-09	46.1 53.9
Illness	1 = Illness registered 2005-09 2 = No illness registered 2005-09	54.8 50.9
Work-related injury	1 = Work-related injury 2005-09 2 = No work-related injury 2005-09	51.3 51.8
Primary care utilisation	1 = 0-3 visits during 2006-10 2 = 4-11 visits during 2006-10 3 = more than 11 visits during 2006-10	52.0 52.0 50.0

Appendix 2

Hospitalisation	1 = 0 hospital admissions during 2006-10 2 = 1 hospital admission during 2006-10 3 = >1 hospital admission during 2006-10	51.8 52.6 50.6

Variables demonstrating large differences in response rates are strong candidates as auxiliary variables. The results indicated that gender, age, country of birth, marital status, income, educational level, social welfare benefits, early retirement benefits, and unemployment were potential auxiliary variables.

Criterion 2
Seven target variables considered particularly important for the study were chosen to assess covariation. The variables were dichotomised for the analysis.

Table 3 Target variables

Variable	Notation
Sexual violence before age 15 perpetrated by an adult	M1
Sexual violence before age 15 perpetrated by a peer	M2
Physical violence before age 15 perpetrated by an adult	M3
Physical violence before age 15 perpetrated by a peer	M4
Sexual violence in adulthood (18+)	M5
Psychological violence in adulthood (18+)	M6
Physical violence in adulthood (18+)	M7

The percentage that met the conditions for the target variable was estimated in different accounting groups with respect to each auxiliary variable. In the estimation, the design weight was corrected for dropout within strata. The weight N_h/m_h was used Instead of the design weight (N_h/n_h), where m_h denotes the number of responders in stratum h.

Gender was found to be closely correlated to the target variables. Women were significantly more likely to report variables M1, M2, M5 and M6, while men reported M3, M4 and M7 to a greater extent. Most target variables decreased in frequency with increasing age. With respect to marital status, married respondents were less exposed than non-married respondents. Those with higher levels of education appear to be more exposed compared to less highly educated respondents. Having received social welfare benefits, early retirement or unemployment benefits was associated with higher levels of exposure to the target Variables.

Income showed small variations with respect to the target variables, as did region, illness, work-related injury and both primary care utilisation and hospitalisation.

The auxiliary variables deemed to fulfil criterion 2 were thus gender, age, country of birth, marital status, educational level, social welfare benefits, early retirement benefits and unemployment.

Criterion 3
The auxiliary variables gender and age delimit important reporting groups for the present study and were therefore included in the auxiliary vector.

Appendix 2

Creation of the weighting variable

The information gained from the auxiliary variables was included in order to create an estimator that reduces the effects of selection and dropout error by creating an auxiliary vector $\mathbf{x}_k$ that describes the categories of *gender + age + country of birth + marital status + educational level + social welfare benefits + early retirement benefits + and unemployment* to which individual *k* belongs. From register data, the following auxiliary sums were then calculated:

$$\sum_{s_d} \bar{d}_k \mathbf{x}_k$$

The calibration estimator was thus

$$\hat{Y}_{wd} = \sum_r d_k v_k y_{dk}$$

where

$$d_k = N_h / n_h$$

The weighting variable v_k used to compensate for errors due to selection and dropout was as follows:

$$v_{sk} = 1 + c_k \left(\sum_s d_k \mathbf{x}_k - \sum_r d_k \mathbf{x}_k \right)' \left(\sum_r d_k c_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \mathbf{x}_k$$

Reference:

Lundström S. and Särndal C.-E. (2001). Estimation in the Presence of Nonresponse and Frame Imperfection. Stockholm: Statistics Sweden.