2.5   Data Validation Routines

Validation of data should be done using the EDMP before data is transmitted to Central Registry.

1   Essentials

- The local identity number (variable numloc) within a registry cannot be duplicated
- Data with values outside the accepted range (as described in Chapter 2.2.1b) will not be imported into EDMP.

2   Duplication checks

- Cases with the same values for 4 key variables (date of birth, sex, birthweight +/- 100 g, maternal age) should be checked as possible duplicates. If value= unknown for any of the 4 matching variables, then case is not matched on that specific variable.
- If case is a twin or higher order multiple birth (NBRBABY=2, 3, 4, 5, 6 or 7), then no duplicate check is carried out.
- It is not sufficient to rely on matching the name of the baby or mother for finding duplicates

3   "Core" information

All babies must have local identification number, date of birth, sex, number of babies delivered, number of malformed cases in multiple set, type of birth, civil registration status, birthweight, length of gestation, survival status, age of mother, when malformation was discovered, first surgery and at least one malformation or syndrome code. Before sending data to the Central Registry any cases with this "core" information lacking should be reviewed to find out if it is possible to complete the missing data.

4   Range Error Checks

Unusual values should be verified eg.

a)   mother's age outside the range 15 to 50
b)   total previous pregnancies greater than 12

5   Logical validation

The following checks of the logical relation between variables are suggested. Sometimes these checks only indicate unusual but possible relationships between different items of information (for example a livebirth at 19 weeks gestation). The more unusual the information, the more likely that there is a coding error. Therefore, these cases should be checked to make sure that the information is correctly coded.

- <u>Baby and Mother</u>

1. If SEX=3, then the code for indeterminate sex (Q56) MUST be entered in malformation field.
2. If SEX=9 (unknown), then PM must be 3 (not performed) or 4 (macerated fetus) or 9 (Not known).
3. If NBRBABY=2, 3, 4, 5, 6 or 7, then NBRMALF must be entered.
4. If both twins of a twin pair are malformed (NBRMALF=2), then PREVSIB=1 (for both twins).
5. If both twins of a twin pair are malformed (NBRMALF=2), then the local ID number of the co-twin should be entered in SIB1.
6. If NBRBABY=2, 3, 4, 5, 6 or 7, then specify twin or multiple type of birth in SP_TWIN field.

- Type of birth
  1. Type of birth, length of gestation and birthweight should be compatible according to the definitions used by the local registry (see instructions)
  2. If type of birth =2, 3 or 4 (SB/ SA/ or TOPFA), then SURVIVAL must be 2 (No)
  3. If type of birth =4 (TOPFA), then WHENDISC must be 6 (prenatal)

- Gestational age, and Birthweight
  1. Maximum birthweights for gestational age are usually:

20-22 weeks	750 g
23-25 weeks	1000 g
26-27 weeks	1500 g
28-31 weeks	2000 g
32-35 weeks	3500 g
36-37 weeks	4000 g
38+ weeks	6000 g

     Birthweights outside these values should be checked
  2. Birthweights less than 500g should be verified if coded as a live or stillbirth

- Death
  1. If survival beyond a week of age =1 (yes), then TYPE =1 (livebirth).
  2. If survival beyond a week of age =2 (no), and TYPE =1 (livebirth), then date of death should be known and should be within one week of birth

- Parental age
  If date of birth of mother (father) is known, the age of mother (father) must also be completed. The age of mother (father) must be the number of completed years between the date of birth of the mother (father) and the date of birth of the baby.

- Previous reproductive history
  1. Total pregnancy validation – total pregnancies may not be equal to total births (twins=2 births, but 1 pregnancy)
  2. Implausible combinations of maternal age and number of previous pregnancies are age 15 or less with 2 or more previous pregnancies, or age 16-19 with 3 or more previous pregnancies.

- Diagnosis
  1. If "when discovered" =6 (prenatal) then AGEDISC should be completed.
  2. If "when discovered" =6 (prenatal) and "condition at discovery" =2 (dead), then type of birth should be a spontaneous abortion (code 3) or a stillbirth (code 2).
  3. If "when discovered"=7 (at abortion 7), then the type of birth should usually be a spontaneous abortion (code 3).
  4. If PM= 1, 2 or 4 (performed) and TYPE =1 (livebirth), then DATE OF DEATH must be entered
  5. If "when discovered" =6, then FIRST POSITIVE PRENATAL TEST must be coded as 1-7, 9 or 11.

- Malformation codes
  1. If OMIM code is entered, there should be a valid ICD code entered in the Syndrome field
  2. All syndrome codes MUST be specified in the SP_SYNDROME field.
  3. If unspecified malformation code is entered, further information MUST be given in the SP_MALF fields.

- Specific malformation coding rules
  1. There is only one valid 2-digit ICD10 code, and that is microcephaly (Q02). All other malformations coded in ICD10 must have at least one letter (eg. Q) followed by 3 digits.
  2. If code is patent ductus Q250, gestational age must be at least 37 weeks in a livebirth to be counted in this anomaly subgroup
  3. If code is Q53, Q54, Q55 (male genital organs) sex must be 1
  4. If code is Q50, Q51, Q52 (female genital organs) sex must be 2
  5. If code Q96 and Q97 sex must be 2
  6. If code Q98 sex must be 1
  7. If code Q00 (anencephalus) "survival first week" must be 2
  8. If code Q05 (spina bifida) no separate code for hydrocephalus (Q03)
  9. If code Q601 (bilateral renal agenesis) or Q606 (Potter syndrome) "survival first week" must be 2
  10. Code Q897 or Q899 cannot be used as the only malformation code

- Family history
  1. If CONSANG=8 (other relation), text information MUST be specified in the SP_CONSANG field.
  2. If SIBANOM=1, 2 or 3 and PREVSIB=1, specify the local identification number in SIB1, SIB2 or SIB3.

6   Frequency checks

Before sending a batch of data to EUROCAT central registry, produce some frequency tables to ensure that the quality of the information corresponds to the aims of the local registry. NB: This check is not automatically performed by EDMP.

- A high frequency of unknown values for any variable should prompt an investigation of how the recording of the variable can be improved, and the registry should communicate with Central Registry concerning how the variable can be used in analyses of data, or if there is selection bias in the distribution of known values.
- It may be useful to check that all malformation codes which have been used only once are valid codes.
- A high frequency of poorly specified malformation codes should prompt investigation.
- The number of cases where "total previous pregnancies" has been coded "0" should correspond approximately to the number of cases expected from the proportion of primiparous mothers in the population.
- Cross-tabulation of maternal age and number of previous pregnancies should show a distribution roughly corresponding to the distribution in the total birth population.

Data Validation Checking

The records will then be validated and given a data validation status. Which are as follows:

1) Core data incomplete or invalid, non-core data valid or invalid
2) Core data complete & valid, non-core data invalid
3) Core data complete & valid, non-core data valid