EUROCARE-5 DATA QUALITY CHECK PROCESS

The EUROCARE-5 database

All incidence and follow up data files arrived by June 2011 have been checked and loaded in the EUROCARE-5 database. Data files were transmitted through the IARC gateway or via a direct upload on the EUROCARE web portal or through both means. The database presently includes information on more than 20 million cancer cases diagnosed from 1978 to 2007 in 30 European countries

Data standardization

Previous to be imported in the management system, data files were standardised in order to have the same record structure for all the registries. For example, the field *life status* had to be derived from three fields included in the IARC protocol (base of diagnosis, life status, autopsy). The information on the dates' *day* was imputed for registries not supplying it and was computed for those providing follow-up duration in days (see **Appendix 1** for details). *Topography* and *morphology* had to be converted to ICDO-3 coding system for the very few registries providing data according to ICD-9, ICD-10 or ICDO-2 systems.

Data checks

Automated procedures checked data fields in each case record, including benign and in-situ cases, and those diagnosed before the current study period (2000-2007). **Table 1** shows the full list of *errors* and *warnings* generated by the EUROCARE-5 check procedures. A specific error code is associated to each type of error/warning.

The consistency of each field was checked by comparison with the range of validity stated in the EUROCARE-5 protocol. Topography and morphology codes were checked against ICD-O-3 lists. Out of range values are considered as *errors.*

The consistency of combinations of two or more fields was also checked and concerned:

- Consistency between *dates* of birth, diagnosis and follow-up.
- Consistency of *site-morphology* combinations. The standard IARC criteria, as described in IARC Technical report n. 42, were applied first, followed by additional EUROCARE criteria.
- Consistency of *age-site, age-morphology, sex-site, and sex-morphology* combinations. Unlikely combinations were checked against IARC criteria.
- Consistency of *morphology-behaviour* combinations. Combinations not listed in ICD-O-3 classification were flagged as unlikely.
- Consistency of *stage information* (EOD and TNM, EOD and condensed TNM, TNM and condensed TNM, site and EOD, behaviour and stage, number of metastatic nodes and stage)

Most inconsistencies among fields combinations are classified as *warnings*, which means possible but not certain error. Inconsistencies on stage information were in general classified as *errors* (see **Table 1**).

Registries' revision of errors and inconsistencies

The records flagged by the data checking process are sent back to the registries for their further revision and correction. Revision of individual records has not been required for all records or warnings, depending on several criteria such as the presence of microscopic verification (non MV are more frequently sent), behaviour (non malignant are in general not sent), age (a few number of childhood cancers are also accepted for adolescents and young adults), and the frequency a given problem is met in a registry file. Furthermore, whenever possible corrections were recovered from the revisions already provided for the EUROCARE-4 run.

Appendix 1

Calculations and imputations in the dates

The EUROCARE-5 protocol includes as optional the *day* of diagnosis and of follow up closure. Some registries were not able to provide such information. For these registries the days in dates were imputed in order to have the same data structure for all the registries. The following rules were used for days imputations:

1. day of diagnosis and day of death were set equal to 15. For cases censored alive the follow up closure day was set equal to 31 if the corresponding month was December (the wide majority of cases), or 15 otherwise.
2. day of birth was always set at 15.

When the registries provided in addition to month and year also *age at diagnosis and disease duration in exact days*, the full dates of diagnosis and end of follow-up were imputed at random with the constraint of satisfying the given time interval and the months in which diagnosis and death/censoring occurred.

When *month* information was missing in the date of *birth* or in the date of *diagnosis*, the month was imputed according to the following algorithm:

1. year of birth < year of diagnosis → month of birth=7
2. year of birth = year of diagnosis → month of birth=month of diagnosis/2
3. year of diagnosis < year of follow up → month of diagnosis=7
4. year of diagnosis=year of follow up → month of diagnosis=month of follow up/2

No imputation was done for month of follow up.

Table1
EUROCARE-5 data checks list (error codes 036-100)

Error Code	Severity	Error Type	Description
036	error	Sex code invalid	Valid values: (1,2,9)
037	error	Date of birth invalid	
038	error	Date of diagnosis invalid	
039	error	Registration date invalid	
040	error	Date of FU invalid	
041	warning	Age > 120	
042	error	Dates' sequence	Birth ≤ Diagnosis ≤ FU
043	error	Vital status code invalid	Valid values: (1-5)
044	error	Site code invalid	Valid ICDO-3-T
045	error	Histological confirmation invalid	Valid values: (1-4,9)
046	error	Morphology code invalid	Valid ICDO-3-M
047	error	Summary extent of disease invalid	Valid values: (1-5,9)
048	error	ID code duplicated	same ID code and multiple tumour indicator
049	error	Female/site combination	
050	error	Male/site combination	
051	warning	Hepato- & retinoblastoma AGE > 5	
052	warning	NGGC & Wilms AGE > 9	
053	warning	Neuroblastoma AGE > 9	
054	warning	Hodgkin AGE < 3	
055	warning	Ewing AGE < 4	
056	warning	Carcinoma NOS AGE< 5	
057	warning	Osteo- & chondrosarcoma AGE < 6	
058	warning	Thyroid and nasopharinx AGE < 6	
059	warning	Renal carcinoma AGE < 8	
060	warning	Hepatic carcinoma AGE< 8	
061	warning	Gonadal carcinoma AGE < 15	
062	warning	Mesothelioma AGE <15	
063	warning	Placental tumours AGE< 15 or > 45	
064	warning	Unlikely sites AGE < 20	
065	warning	Unlikely site and morphology AGE < 20	
066	warning	Unlikely site and morphology AGE < 20	
067	warning	Unlikely morphology AGE < 26	
068	warning	Prostate adenocarcinoma AGE < 30	
069	warning	Unlikely at AGE > 14	
070	warning	Unlikely site/morphology (IARC criteria)	
071	warning	Unlikely site/morphology (added EUROCARE criteria)	
072	warning	Behavior/morphology combination	Combination not included in ICDO-3 list
073	error	Patient identification code	
074	error	Multiple tumour code	
075	error	TNM:N invalid values	
076	error	TNM:T invalid values	
077	error	TNM:M invalid values	

(continue)

Error Code	Severity	Error Type	Description
078	error	Condensed TNM: T	Valid values: (1,2,blank/null)
079	error	Condensed TNM: N	Valid values: (0,1,blank/null)
080	error	Condensed TNM: M	Valid values: (0,1,blank/null)
081	error	Size of tumour in millimetres	Valid values: (0-999, blank/null)
082	error	Number of examined nodes	Valid values: (0-999, blank/null)
083	error	Number of metastatic nodes	Valid values: (0-999, blank/null)
084	error	'C' factor	Valid values: (1,2,9,blank/null)
085	error	Surgery with curative intent	Valid values: (1,2,9,blank/null)
086	error	Chemotherapy with curative intent	Valid values: (1,2,9,blank/null)
087	error	Radiotherapy with curative intent	Valid values: (1,2,9,blank/null)
088	error	Other therapy	Valid values: (1,2,9,blank/null)
089	error	Symptomatic treatment	Valid values: t (1,2,9,blank/null)
090	error	Underlying cause of death	Valid values: (codes ICD-9 o ICD-10)
91	warning	Inconsistent site and EOD	site code=809 and (N=0, M=0)
	warning		site code=809 and (condensed N=0, condensed M=0)
	warning		site code=809 and summary extent of disease = 1
	warning		site code=76* and summary extent of disease = 1
	warning		site code=77* and N=0
	warning		site code=77* and condensed N=0
92	error	Inconsistent behaviour and stage	behavior > 2 and T=is
	error		behaviour=6 and M=0
	error		behaviour=6 and condensed M=0
	error		behaviour=6 and summary extent of disease = local
93	error	Inconsistent EOD=1 and TNM	summary extent of disease=1 and first digit(N)=1,2,3,4
	error		summary extent of disease=1 and first digit(M)=1,2,3,4
94	error	Inconsistent EOD=1 and condensed TNM	summary extent of disease=1 and condensed T=2
	error		summary extent of disease=1 and condensed N=1
	error		summary extent of disease=1 and condensed M=1
95	error	Inconsistent EOD=2 and TNM	summary extent of disease=2 and first digit(M)=1,2,3,4
	error		summary extent of disease=2 and (T=is, or first digit(T)=0,1) and N=0, M=0
96	error	Inconsistent EOD=2 and condensed TNM	summary extent of disease=2 and condensed M=1
97	error	Inconsistent EOD=3 and TNM	summary extent of disease=3 and M=0
98	error	Inconsistent EOD=3 and condensed TNM	summary extent of disease=3 and condensed M=0
99	error	Inconsistent TNM and condensed TNM	first digit(N)=1,2,3,4 and condensed N=0
	error		N=0 and condensed N=1
	error		M=0 and condensed M=1
	error		first digit(M)=1,2,3,4 and condensed M=0
100	error	Inconsistent N+ and stage	number of metastatic nodes >0 and N=0
	error		number of metastatic nodes >0 and condensed N=0
	error		number of metastatic nodes >0 and summary extent of disease=1