Appendix 2. Editing for consistency of data items

V. van Holten

Cancer Statistics Branch, National Cancer Institute,
Bethesda, MD 20205, USA

Editing for consistency of data items can take place either before or after coding. The use of so-called intelligent data entry terminals to edit data as they are keyed is becoming increasingly popular. This process allows the data to be checked for legitimate codes and for consistency between fields as entered. This has the additional advantage of allowing discrepancies to be checked and corrected before the data are added to the permanent data-base. However, this procedure requires the person keying the data to have the ability either to correct the problem or to save the record until the problem can be resolved by someone else. The editing and consistency checking may also be done after all data are entered (in batch mode). In addition, the computer can also be used to check for consistency between incoming data and data previously reported to the registry. When discrepancies are found, list(s) can be created so that someone in the registry can resolve the problems, or algorithms can be established which will allow previously submitted data to be updated by the computer without requiring human intervention.

Since one of the most important functions of a registry is the consolidation of information from a variety of sources for a given patient, every effort must be expended to ensure that the aggregated data are internally consistent. Examples are given here of consistency checks between data items both within a single record and among multiple records submitted to the United States Surveillance, Epidemiology, and End Results (SEER) Program. In brief, the SEER Program is a consortium of 13 population-based central cancer registries which report data in coded format to the US National Cancer Institute (NCI) on an annual basis. Since NCI receives data in a coded format only, all problems uncovered must be referred to the individual registries for resolution. For each independent primary cancer one data record is submitted to the NCI. However, for persons having more than one independent primary cancer, the patient registration number is the same for each data record submitted; hence it is possible to check for consistency between records as well.

Below are listed some of the 50 editing procedures (edits) currently being utilized by the SEER Program. These edits were selected because they should be useful to any registry. They pertain to fields that are collected by most cancer registries. The other edits maintained by the SEER Program apply to fields unique to the SEER data-base (e.g., registry identifier) or to reporting requirements that have changed since SEER began collecting data.

Some of these edits were designed so that combinations of codes usually expected are considered correct and those not usually expected as incorrect. Thus, rare combinations of the fields which are correct can be marked as incorrect. In order to retain the information for these rare cases, a separate field, an override flag, is created for the edit. When the override flag is set to 'on', the case is no longer considered to contain a discrepancy. For example, the edit for age and primary site could be overridden for the rare case of an 18-year-old with invasive cervical cancer. However, review of any case found in the future would be required.

SEER inter-field edits

Table 1 lists some of the inter-field edits used by the SEER Program.

Table 1. SEER inter-field edits (Field A in conjunction with field B)

(Note: the edits described below refer to the 1976 edition of ICD-O)

Field A Item name[1]	Field B Item name[1]	Editing criteria
Type of reporting source (item 35)	Follow-up (item 31)	If this is an autopsy or death certificate only case, then follow-up status must be dead.
Type of reporting source (item 35)	Cause of death (item 33)	If this is an autopsy or death certificate only case, then cause of death must be specified.
Age at incidence date (item 9)	Marital status (item 8)	If age <15 years, then marital status must be single.
Age at incidence date (item 9)	Date (year/month) of birth (item 5) and incidence date (year/month) (item 16)	Age must equal calculated age, where calculated age = ((incidence year × 12 + incidence month) − (birth year × 12 + birth month))/12
Date of birth (item 5)	Incidence date (item 16)	Date of incidence must be equal or greater than date of birth.
Sex (item 4)	Primary site (item 20)	Primary site codes for female breast (174.-) and female genital organs (179.9–184.9) are invalid for males.
		Primary site codes for male breast (175.9) and male genital organs (185.9–187.9) are invalid for females
Age at incidence date (item 9)	Primary site (item 20) and histological type (item 21)	If override flag is set to 'on' indicating case has been previously reviewed, no further checks are performed unless one of the other fields involved in the edit has been modified.
	Override flag for age/site edit	If age <5 years, the primary site cannot be: Cervix uteri (180.-) Prostate (185.9).
		If age <20 years, then primary site cannot be: Oesophagus (150.-) Small intestine (152.-)

Table 1 — continued

(Note: the edits described below refer to the 1976 edition of ICD-O)

Field A Item name[1]	Field B Item name[1]	Editing criteria
		Colon (153.-); (histological type is not carcinoid (M8240–8244) Rectum, rectosigmoid junction, anal canal and anus, NOS (154.-) Gallbladder and extrahepatic bile ducts (156.-) Pancreas (157.-) Lung and bronchus (162.—) if histological type is not carcinoid (M8240–8244) Pleura (163.-) Breast (174.-, 175.9) Uterus NOS (179.9) Cervix uteri (180.-) with invasive behaviour Corpus uteri (182.-).
		If age <30 years, histological type cannot be: Multiple myeloma (M9730) Chronic lymphocytic leukaemia (M9823) Chronic monocytic leukaemia (M9863) Monocytic leukaemia, NOS (M9890).
		If age <45 years, primary site cannot be prostate (185.9) with histological type of adenocarcinoma (M8140).
		If age >5 years, primary site cannot be eye (190.-) with a histological type of retinoblastoma (M9510–9512).
		If age <15 years or >45 years, then primary site cannot be placenta (181.9) with a histological type of choriocarcinoma (M9100).
Incidence date (item 16)	Date cancer-directed therapy started	Incidence date must be the same as or before date cancer-directed therapy started.
Incidence date (item 16)	Date of last contact (item 30)	Incidence date must be the same as or before date of last follow-up or death.
Primary site (item 20)	Histological type (item 21)	This edit is defined in the Site/Histology Validation Edit description.
Primary site (item 20) and histological type (item 21)	Extent of disease (items 23–25)	This edit is performed according to the extent of disease codes allowed for each primary site and histological type combination.
Primary site (item 20)	Laterality (item 28)	The following ICD-O sites must have a valid laterality code: 142.0 Parotid gland 142.1 Submaxillary gland 142.2 Sublingual gland 146.0 Tonsil

Table 1 — continued

(Note: the edits described below refer to the 1976 edition of ICD-O)

Field A Item name[1]	Field B Item name[1]	Editing criteria
		146.1 Tonsillar fossa 146.2 Tonsillar pillar 160.1 Eustachian tube Middle ear 160.2 Maxillary sinus 160.4 Frontal sinus 162.3 Lung, upper lobe 162.4 Lung, middle lobe 162.5 Lung, lower lobe 162.8 Lung 162.9 Lung, NOS 163.- Pleura 170.4 Long bones of upper limb and scapula 170.5 Short bones of upper limb 170.7 Long bones of lower limb 170.8 Short bones of lower limb 171.2 Soft tissue of upper limb and shoulder 171.3 Soft tissue of lower limb 173.1 Eyelid 173.2 External ear 173.3 Skin of face 173.5 Skin of trunk 173.6 Skin of arm and shoulder 173.7 Skin of leg and hip 174.- Female breast 175.9 Male breast 183.0 Ovary 183.2 Fallopian type 186.0 Undescended testis 186.9 Testis 187.5 Epididymis 187.6 Spermatic cord 189.0 Kidney 189.2 Ureter 190.- Eye 194.0 Suprarenal gland 194.5 Carotid body
Behaviour (item 2)	Most valid basis of diagnosis (item 17)	If behaviour code is *in situ*, then there must be a positive histological confirmation.
Behaviour (item 22)	Extent of disease (items 23–25)	If behaviour is *in situ*, then extent of disease must be *in situ*.
Date cancer-directed therapy started	First course of cancer-directed therapy (item 29)	Date of therapy must contain a valid date if first course of therapy indicates therapy was performed.

Table 1 — continued

(Note: the edits described below refer to the 1976 edition of ICD-O)

Field A Item name[1]	Field B Item name[1]	Editing criteria
Follow-up status (item 31)	Cause of death (item 33)	If follow-up status is alive, then cause of death must be 0000 (alive); if follow-up status is dead, then cause of death must not be 0000
Date of last contact (item 30)	Date cancer-directed therapy started	Date of cancer-directed therapy must be the same as or before date of last contact.

[1] The item numbers in parentheses refer to the corresponding items of patient information described in Chapter 6

SEER site/histology validation edit

When both the site and the histological type (including behaviour code) are found to be valid, a site/type combination edit will be performed. The site/type edit checks for allowable histology codes for each site. The site/histology validation list designates each site and the histology codes (four-digit) that are considered valid for each site—an example, for lip cancer, is shown in Table 2. The list frequently specifies ranges of site codes but only valid site codes within the range are applicable. The histology terms on the site/histology validation list are not necessarily the preferred terms specified in the International Classification of Diseases for Oncology (ICD-O). A diagnostic message will be generated for those cases for which the histology code is not specified as valid for the site code. If, after review of the case, the site/histology combination is found to be correct, the override flag should be set to 'on'.

In addition, the following combinations are invalid:

(1) *An unknown or ill-defined site with a histology that has an in situ behaviour code*:

149.9 Ill-defined sites in lip, oral cavity and pharynx
159.9 Ill-defined sites within digestive organs and peritoneum
165.9 Ill-defined sites within respiratory system
179.9 Uterus, NOS
184.9 Female genital tract, NOS
187.9 Male genital tract, NOS
189.9 Urinary system, NOS
192.9 Nervous system, NOS
194.9 Endocrine gland, NOS
195.- Other ill-defined sites
199.9 Unknown primary site

(2) *Ill-defined sites (195.-) with histologies specifying melanoma (M8720–8790)*

Table 2. Example of a SEER site/histology validation list

Site		Histology (four-digit)
Lip	400–409	8000/3 Neoplasm, malignant
		8001/3 Tumor cells, malignant
		8002/3 Malignant tumor, small cell type
		8003/3 Malignant tumor, giant cell type
		8004/3 Malignant tumor, fusiform cell type
		8010/2 Carcinoma-in-situ, NOS
		8010/3 Carcinoma, NOS
		8011/3 Epithelioma, malignant
		8012/3 Large cell carcinoma, NOS
		8020/3 Carcinoma, undifferentiated type, NOS
		8021/3 Carcinoma, anaplastic type, NOS
		8022/3 Pleomorphic carcinoma
		8030/3 Giant cell and spindle cell carcinoma
		8031/3 Giant cell carcinoma
		8032/3 Spindle cell carcinoma
		8033/3 Pseudosarcomatous carcinoma
		8034/3 Polygonal cell carcinoma
		8050/2 Papillary carcinoma-in-situ
		8050/3 Papillary carcinoma, NOS
		8051/3 Verrucous carcinoma, NOS
		8052/3 Papillary squamous cell carcinoma
		8070/2 Squamous cell carcinoma-in-situ, NOS
		8070/3 Squamous cell carcinoma, NOS
		8071/3 Sq. cell carc., ker. type, NOS
		8072/3 Sq. cell carc., lg. cell, non-ker.
		8073/3 Sq. cell carc., sm. cell, non-ker.
		8074/3 Sq. cell carc., spindle cell
		8075/3 Adenoid squamous cell carcinoma
		8076/2 Sq. cell carc.-in-situ
		8076/3 Sq. cell carc., micro-invasive
		8081/2 Bowen's disease
		8082/3 Lymphoepithelial carcinoma
		8140/2 Adenocarcinoma-in-situ
		8140/3 Adenocarcinoma, NOS
		8141/3 Scirrhous adenocarcinoma
		8143/3 Superficial spreading adenoca.
		8200/3 Adenoid cystic carcinoma
		8201/3 Cribriform carcinoma
		8260/3 Papillary adenocarcinoma, NOS
		8261/2 Adenoca. in situ in villous adenoma
		8261/3 Adenocarcinoma in villous adenoma
		8262/3 Villous adenocarcinoma
		8263/2 Adenoca. in situ in tubulovillous adenoma
		8263/3 Adenocarcinoma in tubulovillous adenoma

Table 2 — continued

Site	Histology (four-digit)
	8430/3 Mucoepidermoid carcinoma
	8480/3 Mucinous adenocarcinoma 8481/3 Mucin-producing adenocarcinoma
	8720/3 Malignant melanoma, NOS 8721/3 Nodular melanoma 8722/3 Balloon cell melanoma
	8730/3 Amelanotic melanoma
	8743/3 Superficial spreading melanoma
	8771/3 Epithelioid cell melanoma 8772/3 Spindle cell melanoma, NOS 8775/3 Mixed epithel. & spindle cell melan.
	8940/3 Mixed tumor, malignant, NOS 8941/3 Carcinoma in pleomorphic adenoma
	9140/3 Kaposi's sarcoma

SEER inter-record edits

Whenever a patient has more than one record on file, the following fields will be edited to ensure proper consistency between records:

Place of birth (item 7)
Must be equal on all records

Date of birth (item 5)
Must be equal on all records

Ethnic group (item 11)
Must be equal on all records

Sex (item 4)
Must be equal on all records

Sequence number (tumour identification)

(1) when there is more than one record for a patient, no record may contain a zero or unknown in sequence number

(2) sequence numbers must be unique

Sequence number/date of incidence

The tumour sequence numbers must reflect the chronological sequence of the incidence of the primaries. Thus the primary assigned a sequence number of '1' must have a date of incidence = or < the date of incidence of the primary assigned sequence number '2', etc.

Date of follow-up or death
Must be equal on all records

Follow-up status (item 31)

(1) If the patient is indicated as being 'alive' in any one record, the other records must also specify 'alive'

(2) If the patient is indicated as being 'dead' in any one record, the other records must also specify 'dead'

Distinguishing multiple primaries from duplicate registrations

An editing procedure may be adopted to check primaries which have been reported as independent, but are really only one. This type of edit will apply only to invasive cases, and its format will be determined by the definition used for 'Multiple Tumours'. If that in Chapter 7 is used, the two cases must be in the same histological group (Groups 1, 2, 3, 5, 6 or 7 of Table 2, Chapter 7).

Records are marked for review whenever one site specifies an ill-defined site or an NOS site and the other site specifies a specific subsite. If, after review, the cases are determined to be independent primaries, an override flag is used to indicate the primaries for the person have been reviewed and found to be correct. Table 3 specifies the combinations of primary sites for which review is required.

Table 3. Combinations of primary sites for which review is required

Ill-defined or NOS site	Specified site
149.9 Ill-defined sites in lip oral cavity and pharynx	140.0–149.8
159.0 Intestinal tract, NOS	150.0–158.9
159.8 Overlapping sites of digestive system	150.0–158.9
159.9 Ill-defined sites of digestive system and peritoneum	150.0–158.9
165.0 Upper respiratory tract	160.0–163.9
165.8 Overlapping respiratory and intrathoracic sites	160.0–164.9
165.9 Ill-defined sites of respiratory system	160.0–163.9
184.9 Female genital organs, NOS	179.9–184.8
187.9 Male genital organs, NOS	185.9–187.8
189.9 Urinary system, NOS	188.0–189.8
194.8 Multiple endocrine glands	193.9–194.6
194.9 Endocrine gland, NOS	193.9–194.8
1AA.8[a] Any overlapping site code	1AA.x[a],[b] any associated subsite
1BB.9[c] Any NOS site	1BB.x[b],[c] any associated subsite

[a] AA is any two-digit number in the range 40–99 except 46, 51, 54, 58, 62, 70, 74, 80, 83, 87 and 91

[b] x is any one-digit number

[c] BB is any two-digit number in the range 40–99