Medicines & Healthcare products Regulatory Agency

CPRD More Dimensions to Data

Cancer Registration Data Dictionary (Set 17)

Version 8.1

Date: 18 April 2019

NHS National Institute for Health Research

Documentation Control Sheet

Over time, it may be necessary to issue amendments or clarifications to parts of this document. This form must be updated whenever changes are made.

Version	Affected Areas Summary of Change	Prepared By	Reviewed By
7.0	Created separate data documentation and data dictionary files	Helen Strongman	Rachael Williams
7.1	Add tumour identifier to patient file for clarity (this has been available since set 14)	Helen Strongman	Eleanor Yelland
8.0	Refresh for Set 16	Rachael Williams	Eleanor Yelland
8.1	Review for Set 17	Eleanor Yelland	Helen Booth

Version 7.0

- Cancer Registration Data dictionary separated from NCRAS documentation. No changes to data structure between set 13 (up to 2014) and set 14 (up to 2015)

Version 7.1

- Added tumour identifier to patient file for clarity (this has been available since set 14)

Version 8.0

- Refreshed for Set 16

Version 8.1

- Refreshed for Set 17
- Minor updates to variable descriptions

CPRD Cancer Registration Data Structure

1. Patient - One row per patient

Column description	*Column name*	*Details*	*Field Type*	*Valid Content*
CPRD patient Identifier	e_patid	Unique patient identifier based on CPRD primary care data – pseudonymised. In some cases, the same person may have multiple patient IDs.	ID	Number
CR patient Identifier	e_cr_patid	Unique patient identifier based on NCRAS data – pseudonymised. In some cases, the same person may have multiple patient IDs. Patient IDs will be retained even after two patient records are found to be the same person.	ID	Number
Tumour count	tumourcount	Count of every tumour associated with this e_cr_patid	NUMBER	Number
Big tumour count	bigtumourcount	Count of every tumour associated with this e_cr_patid in range C00-97 excluding C44	NUMBER	Number

2. Tumour table – one row per patient per tumour

Column description	*Column name*	*Details*	*Field Type*	*Valid Content*
CPRD patient Identifier	e_patid	Unique patient identifier based on CPRD primary care data – pseudonymised. In some cases, the same person may have multiple patient IDs.	ID	Number
CR patient Identifier	e_cr_patid	Unique patient identifier based on NCRAS data – pseudonymised. In some cases, the same person may have multiple patient IDs. Patient IDs will be retained even after two patient records are found to be the same person.	ID	Number
CR tumour identifier	e_cr_id	Unique tumour identifier based on NCRAS data – pseudonymised.	ID	Number
Year of Birth	year_dob	Year portion of date of birth as recorded in the cancer registry data, where available (NB: year of birth is available from GP records for all research acceptable patients)	NUMBER	yyyy
Age at diagnosis	age	Age in years at diagnosis, rounded down to full years	NUMBER	Number or blank
Age group at diagnosis	fiveyearageband	Age at diagnosis in 5-year groupings	TEXT	0 - 4 YRS | 5 - 9 YRS | 10 - 14 YRS | 15 - 19 YRS | 20 - 24 YRS | 25 - 29 YRS | 30 - 34 YRS | 35 - 39 YRS | 40 - 44 YRS | 45 - 49 YRS | 50 - 54 YRS | 55 - 59 YRS | 60 - 64 YRS | 65 - 69 YRS | 70 – 74 YRS | 75 - 79 YRS | 80 - 84 YRS | Blank

Sex	sex	Sex of the patient at diagnosis as recorded in the Cancer Registration data (NB: sex is available from GP records for all research acceptable patients)	NUMBER	0=Not known, 1=Male, 2=Female, 9=Not specified
Ethnic Origin	ethnicity	Follows 2001 census definition. Data fields in linked Hospital Episode Statistics containing ethnicity data were used. The most common, or if not most common, the most recent known, ethnicity was taken.	TEXT	A = (White) British, B =(White) Irish, C = Any other White background, D = White and Black Caribbean, E = White and Black African, F = White and Asian, G = Any other mixed background, H = Indian, J = Pakistani, K = Bangladeshi, L = Any other Asian background, M = Caribbean, N = African, P = Any other Black background, R = Chinese, S = Any other ethnic group, Z = Not stated, X = Not Known
Diagnosis date	diagnosisdatebest	Diagnosis date of the patient, as defined by the UKACR	DATE	dd/mm/yyyy
Month of diagnosis	diagnosismonth	Month of diagnosis	NUMBER	mm
Diagnosis year	diagnosisyear	Year of diagnosis	NUMBER	yyyy
Date of diagnosis imputed flag	diagnosisdateflag	Imputation of dates follows rules agreed by UKACR DQAR sub-group (August 2010). Blank field indicates that date imputation did not occur.	NUMBER	0 = date fully specified, 1 = only month and year specified, 2 = only year specified, 3 = none of the above

Basis of Diagnosis	basisofdiagnosis	Basis of the diagnosis data (e.g. Death Certificate; Clinical; Clinical Investigation; Specific tumour markers; Cytology; Histology of a metastases; Histology of a primary tumour; Unknown)	NUMBER	<u>Non-microscopic</u> 0 = Death certificate 1 = Clinical: Diagnosis made before death without (2-7) 2 = Clinical investigation: Includes all diagnostic techniques without a tissue diagnosis 4 = Specific tumour markers: Includes biochemical and/or immunological markers which are site specific <u>Microscopic</u> 5 = Cytology: Examination of cells whether from a primary or secondary site, including fluids aspirated using endoscopes or needles. Also including microscopic examination of peripheral blood films and trephine bone marrow aspirates 6 = Histology of a metastases: Includes autopsy specimens 7 = Histology of a primary tumour: Includes all cutting and bone marrow biopsies. Also includes autopsy specimens of a primary tumour 9 = Unknown, e.g. PAS or HISS record only
Death certificate only	dco	Whether basis of diagnosis of the tumour was death certificate only	TEXT	Y = Yes, N = No
Site (recoded, 4 characters)	site_icd10_O2	Site of the cancer mapped to a 4-character ICD-10-O2 code (from 1995 only)	TEXT	Valid 4 digit ICD-10 codes in the range C00-D48 plus D76, E85, O01, Q85 or blank
Site (recoded, 3 characters)	site_icd10_O2_3char	Site of the cancer mapped to a 3-character ICD-10-O2 code (from 1995 only)	TEXT	Valid 3 digit ICD-10 codes in the range C00-D48 plus D76, E85, O01, Q85 or blank

Site of primary neoplasm	site_coded	Site of the cancer, in the coding system that the tumour was originally coded in. This variable (or site_coded_3char) should be selected if data prior to 1995 are being requested.	TEXT	
Site of primary neoplasm (3 digits)	site_coded_3char	Three-digit version of site_coded. This variable (or site_coded) should be selected if data prior to 1995 are being requested.	TEXT	
Coding system	coding_system	The coding system used to register the tumour (e.g. ICD10/O-2). This should be requested if site_coded or site_coded_3char are being requested.	NUMBER	1 = ICD-8, 2 = ICD-9, 3 = ICD-10/O-2, 4 = ICD-10/O-3, 5 = ICD-O-3, 6 = ICD-7, 7 = ICD-8pre1971, 8 = ICD-O-2, 9 = ICD-O, 10 = ICD-O-3 (2011), 11 = ICD-10rev4/O-2, 12 = MOTNAC, 14 = SNOMED/O(TCR), 15 = SNOMED/O-1, 16 = SNOMED/O-2, 17 = SNOMED/O-3
Morphology	morph_coded	Morphology of the cancer, in the coding system that the tumour was originally coded in, describing the cell type of malignant disease determined before the start of treatment. The relevance of tumour morphology differs across tumour site.	TEXT	
Morphology (recoded)	morph_icd10_O2	Morphology of cancer mapped to ICD-10-O2	NUMBER	Number 8000-9990 or blank
Behaviour	behaviour_coded	Behaviour of the cancer	NUMBER	

Behaviour (recoded)	behaviour_icd10_O2	Behaviour of cancer mapped to ICD-10-O2	TEXT	0, 1,2,3,5,6,9,XXX,XXXX, blank
Histology	histology_coded	Histology code	TEXT	
Grade	grade	Records the grade of the tumour, for tumours that are graded on a simple numeric 1-3 or 1-4 scale. In tumours containing several areas of different grade, the grade of the predominant component is recorded.	TEXT	GX = Grade of differentiation is not appropriate or cannot be assessed G1 = Well differentiated G2 = Moderately differentiated G3 = Poorly differentiated G4 = Undifferentiated / anaplastic
Tumour size	tumoursize	Diameter of a tumour in mm, largest if more than one.	NUMBER	Number or blank
Nodes excised	nodesexcised	Number of nodes excised	NUMBER	Number or blank
Nodes involved	nodesinvolved	Number of nodes involved	NUMBER	Number or blank

Laterality (side) for paired organs	laterality	For paired sites, for e.g. the tonsils, if there is a tumour in one side, the laterality of that side, left or right, is recorded. For some paired sites, if there are tumours in both sides then two tumours are registered, one a left and the other a right. If there is a tumour in both sides (and they have other factors such as morphology the same) then only one registration is made, and the laterality is coded as bilateral. If the site of the primary cancer is not part of a pair, then laterality is coded as not applicable.	TEXT	L = Left, R = Right, M = Midline, B = Bilateral, 8 = Not applicable, 9 = Not Known
Multifocal	multifocal	Whether or not the tumour is multifocal	TEXT	N= No, Y = Yes, 8 = Not applicable, 9 = Not known
Oestrogen receptor status	er_status	Oestrogen receptor status of the tumour	TEXT	N = negative, P = positive, X = not performed
Oestrogen receptor score	er_score	Oestrogen receptor score of the tumour	TEXT	ER Allred score (range 0, 2-8)
Progesterone receptor status	pr_status	Progesterone receptor status of the tumour	TEXT	N = negative, P = positive, X = not performed
Progesterone receptor score	pr_score	Progesterone receptor score of the tumour	TEXT	ER Allred score (range 0, 2-8)
HER2 status	her2_status	HER2 status of the tumour	TEXT	N = negative, P = positive, X = not performed
NPI score	npi	Nottingham Prognostic Indicator score (not the derived stage) for prognosis following surgery for breast cancer. Combines grade, axillary node involvement and tumour size.	NUMBER	Number (two decimal places) or blank

Dukes' stage	dukes	Used for colorectal cancer	TEXT	A = Dukes' A: Tumour confined to wall of bowel, nodes negative B = Dukes' B: Tumour penetrates through the muscularis propria to involve extramural tissues, nodes negative C1 = Dukes' C1: Metastases confined to regional lymph nodes (node/s positive but apical node negative) C2 = Dukes' C2: Metastases present in nodes at mesenteric artery ligature (apical node positive) D = Dukes D: Metastatic spread outside the operative field 99 = Not Known
FIGO stage	figo	Used for cervical cancer	TEXT	0, 1, 1a, 1a1, 1a2, 1b, 1b1, 1b2, 1c, 1c1, 1c2, 1c3, 2, 2a, 2a1, 2a2, 2b, 2c, 3, 3a, 3b, 3c, 3c1, 3c2, 4, 4a, 4b, I, IA, IA1, IA2, IB, IB1, IB2, IC, II, IIA, IIA2, IIB, IIC, III, IIIA, IIIB, IIIC, IIIC1, IIIC2, IV, IVA, IVB, blank
Clark's level	clarks	Used for melanoma of the skin	TEXT	1, 2, 3, 4, 5, blank
Breslow thickness	breslow	Used for melanoma of the skin – measured in millimetres	TEXT	Number or range, x, or blank
Gleason primary pattern	gleason_primary	Used for prostate cancer - the grade that comprises most of the tumour volume is called the "primary pattern"	NUMBER	1-5, 8 = not applicable

Gleason secondary grade	gleason_secondary	Used for prostate cancer - if additional grades present, the highest grade (biopsy) or the second most extensive grade (TURP and radicals). If none present, primary and secondary grades are the same.	NUMBER	1-5, 8 = not applicable
Gleason tertiary	gleason_tertiary	Value of any different third grade in addition to the primary and secondary grades	NUMBER	1-5, 8 = not applicable
Gleason combined	gleason_combined	Gleason grade format follows that of NCDS: www.ic.nhs.uk/webfiles/Services/Datasets/cANCER/appurological.doc	NUMBER	2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, blank
T stage (Imaging)	t_img	UICC code classifying the size and extent of the primary tumour before treatment	TEXT	UICC code
N stage (Imaging)	n_img	UICC code classifying the absence or presence and extent of regional lymph node metastases before treatment	TEXT	UICC code

M stage (Imaging)	m_img	UICC code classifying the absence or presence of distant metastases before treatment	TEXT	0 = no distant metastasis 1, 1a, 1b, 1c, 1e = distant metastasis X = unknown
Stage (Imaging)	stage_img	Combination of imaging T, N and M in “t_img”, “n_img” and “m_img”. Includes Ann Arbor staging for lymphomas. NB: It is not guaranteed that data from the individual t_img, n_img and m_img variables have been combined into this variable, so they should be used in parallel.	TEXT	
Stage system (Imaging)	stage_img_system	Version of the TNM classification of malignant cancers used to stage the tumour for the imaging TNM values	NUMBER	5 = 5th, 6 = 6th, 7 = 7th, 20 = UICC 5, 21 = UICC 6, 22 = UICC 7, 23 = AJCC 7, 24 =Unknown
T stage (Pathological)	t_path	UICC code classifying the size and extent of the primary tumour based on the evidence from a pathological examination	TEXT	UICC code
N stage (Pathological)	n_path	UICC code classifying the absence or presence and extent of regional lymph node metastases before treatment based on the evidence from a pathological examination	TEXT	UICC code
M stage (Pathological)	m_path	UICC code classifying the absence or presence of distant metastases before treatment based on the evidence from a pathological examination	TEXT	0, 1, 1a, 1b, 1c, 1e, 2, 3, 4, 9, X, blank

Stage (Pathological)	stage_path	Combination of pathological T, N and M in "t_path", "n_path" and "m_path". Includes Ann Arbor staging for lymphomas. NB: It is not guaranteed that data from the individual t_path, n_path and m_path variables have been combined into this variable, so they should be used in parallel.	TEXT	0, 0A, 0IS, 1, 1A, 1A1, 1A2, 1B, 1B1, 1B2, 1C, 1E, 2, 2A, 2B, 2C, 2E, 3, 3A, 3B, 3C, 3E, 4, 4A, 4B, 4C, 5, 6, ?, U, X, blank
Stage system (Pathological)	stage_path_system	Version of the TNM classification of malignant cancers used to stage the tumour for the pathological TNM values	NUMBER	5, 6, 7, 20, 21, 22, 23,24, blank
Stage (Pathological pre-treatment)	stage_path_pretreated	Pathological stage at diagnosis recorded prior to treatment	TEXT	Y = Yes, N = No
T stage (Best)	t_best	T stage flagged by the registry as the best	TEXT	
N stage (Best)	n_best	N stage flagged by the registry as the best	TEXT	
M stage (Best)	m_best	M stage flagged by the registry as the best	TEXT	
Stage (Best)	stage_best	Combination of best T, N and M in "t_best", "n_best" and "m_best". Includes Ann Arbor staging for lymphomas. NB: It is not guaranteed that data from the individual t_best, n_best and m_best variables have been combined into this variable, so they should be used in parallel.	TEXT	0, 0A, 0IS =Stage 0 1, 1A, 1A1, 1A2, 1B, 1B1, 1B2, 1C, 1E = Stage 1 2, 2A, 2A1, 2A2, 2B, 2C, 2E, 2S = Stage 2 3, 3A, 3B, 3C, 3E, 3S = Stage 3 4, 4A, 4B, 4C, 4S = Stage 4 6 = not stageable ? = insufficient information U = unstageable, X = not staged
Stage system (Best)	stage_best_system	Version of the TNM classification of malignant cancers used to stage the tumour for the best TNM values	NUMBER	5 = 5th, 6 = 6th, 7 = 7th, 20 = UICC 5, 21 = UICC 6, 22 = UICC 7, 23 = AJCC 7, 24 =Unknown

Excision margin	excisionmargin	Whether the surgical excision margin finding was clear of the tumour and if so, by how much	TEXT	01 = Excision margins are clear (distance from margin not stated) 02 = Excision margins are clear (tumour >5mm from the margin) 03 = Excision margins are clear (tumour >1mm but less than or equal to 5mm from the margin 04 = Tumour is less than or equal to 1mm from excision margin, but does not reach margin 05 = Tumour reaches excision margin 06 = Uncertain 07 = Margin not involved =>1mm 08 = Margin not involved <1mm 09 = Margin not involved 1-5mm 98 = Not applicable 99 = Not Known
Screen detected	screendetected	Whether or not the tumour was detected by a screening programme	TEXT	N = No, Y = Yes, 8 = Not applicable, 9 = Not known
Screening status	screeningstatuscosd_code	Screening status	TEXT	1 = screen-detected, 2 = interval cancer, 4 = lapsed attender, 5 = never attended, 6 = never invited, 9 = not known, NM = not mapped
Screening status (detailed)	screeningstatusfull_code	The value of the sub-classification of the screening flag. Populated when the screening status is “Other”. For breast screening service: www.cancerscreening.nhs.uk/breastscreen/publications/nhsbsp62.pdf (page 4, section 2). For cervical screening service see: www.cancerscreening.nhs.uk/cervical/publications/nhscsp28.pdf (page 37)	TEXT	

Date of first recorded treatment event	date_first_event	Date of first recorded treatment event	DATE	ddmmyyyy
Date of first recorded surgery	date_first_surgery	Date of first recorded surgery event	DATE	ddmmyyyy
Catchment area code	creg_code	Code for the cancer registry catchment area the patient was resident in when the tumour was diagnosed	TEXT	Y0801=Thames Cancer Registry Y0201=Northern & Yorkshire Cancer Registry & Information Service Y0301=Trent Cancer Registry Y1201=West Midlands Cancer Intelligence Unit Y0401=Eastern Cancer Registration & Information Centre Y1701=North West Cancer Intelligence Service Y1001=South West Cancer Intelligence Service Y1101=Welsh Cancer Intelligence & Surveillance Unit Y0901=Oxford Cancer Intelligence Unit Z9999=blank
Route to diagnosis	route_code	The code assigned to a route for the purpose of the algorithm. Note: available for cancers diagnosed in 2006-2014. See BJC publication.		
Finalised route code	final_route	The published route with all dataset types accounted for. Note: available for cancers diagnosed in 2006-2014. See BJC publication.		

3. Treatment table

Column description	*Column name*	*Details*	*Field Type*	*Valid Content*
CPRD patient Identifier	e_patid	Unique patient identifier based on CPRD primary care data – pseudonymised.	ID	Number
CR patient Identifier	e_cr_patid	Unique patient identifier based on NCRAS data - pseudonymised. In some cases the same person may have multiple patient IDs. Patient IDs will be retained even after two patient records are found to be the same person.	ID	Number
CR tumour identifier	e_cr_id	Unique tumour identifier based on NCRAS data – pseudonymised.	ID	Number
Number of tumours	number_of_tumours	Number of tumours affected by this event	NUMBER	
Type of event	eventcode	Type of event	TEXT	01a = Surgery – curative, 01b = Surgery - not curative, 01z = Surgery etc. - type unknown, 02 = Cytotoxic Chemotherapy, 03 = Hormone Therapy, 05 = RT – Teletherapy, 06 =RT – Brachytherapy, 15 = Immunotherapy, 97 = Other Treatment, 99 = Treatment unknown, CTX = CT – Other, IM = Imaging, RTX = RT - Other/NK
Treatment date	eventdate	Date the treatment took place	DATE	ddmmyyyy
Treatment month	eventmonth	Month the treatment took place	NUMBER	mm
Treatment year	eventyear	Year the treatment took place	NUMBER	yyyy

Treatment within six months	within_six_months_flag	Whether treatment was within six months of date of diagnosis	NUMBER	0 = No, 1 = Yes
Treatment after six months	six_months_after_flag	Whether treatment was after six months from date of diagnosis	NUMBER	0 = No, 1 = Yes
OPCS code	opcs4_code	Operations, procedures and interventions coded	TEXT	OPCS4 code
OPCS name	opcs4_name	Name of operation, procedure or intervention	TEXT	
Radiotherapy code	radiocode	Radiotherapy type	TEXT	1 = 1 + 2, 2 = 1 + 4, 3 = Brachytherapy, 4 = External beam, 5 = Intracavitary or interstitial, 8 = Other, B = Radioactive isotopes, X = Unknown / inapplicable
Imaging code	imagingcode	Imaging code	TEXT	
Imaging site	imagingsite	Site on body where imaging occurred	TEXT	
Lesion size	lesionsize	The size in millimetres of the diameter of a lesion, largest if more than one (histology)	NUMBER	Number or blank
Chemotherapy drug	chemo_drug	Name of chemotherapy drug	TEXT	Please note this is a non-mandated text field that may not be complete or contain the specific drug name.