National Cancer Data Repository 1990-2010

Release notes and data definitions

Version 5.2

Revision History

Date	Version	Description	Author
1.10.2012	V1.0	Draft content for release notes	Kath Yates
08/10/12	V2.0	Add release contacts	Kath Yates
15/10/12	V3.0	Amendments & caveats	Catherine O'Hara
16/10/12	V4.0	Final appendices added	James Thomas
16/10/12	V5.0	Editing	Sean McPhail
23/10/12	V5.2	Version for Web	Kath Yates and Sean McPhail

Contents

Introduction

Background

The 1990–2010 England NCDR Analysis Dataset brings together data from each of the England Cancer Registries for the period of 1990 to 2010.

The data consists of tumour level records submitted to ONS by the England Cancer Registries together with a further sub-set of data covering additional data fields required for analysis purposes. The ONS data set has already been collated, cleaned and uses standardised data items (as submitted by Cancer Registries), however there may be caveats within this document where data has been updated locally in the Cancer Registry after the date of submission and close of the ONS dataset – these variances are noted below.

Document Purpose

The document is intended to detail the content of the 1990–2010 England NCDR Analysis Dataset. The addition of a Celtic Countries NCDR Analysis Dataset will follow this publication. As these release notes are used, there may be further iterations which reflect updated amendments – these will be shared with the cancer registry leads.

Dataset description

Production Process

The NCDR 2010 build is based around the ONS Cancer Dataset, with additional information coming, where available, from a merged dataset of cancer registry data.

Data was requested for the 8 English cancer registries conforming to the data specification contained in this data release pack as "National_Cancer_Data_Repository_1985_-_2010_v1.1.pdf". (The 3 Celtic nation's data specification will be added once final verification has been completed). On receipt each of these datasets were checked for conformity with the specification. Non-conformities were reported back to the supplying registry and either updated centrally or updates of data were supplied. A detailed account of the non-conformities and remedial action is contained in the Regional Completeness report in the release folder.

Once all issues had been resolved the individual English datasets were merged into a single dataset. Additional fields to identify the registry supplying, ICD10-ICDO3 topography/ morphology and records for exclusion were included. The ONS cancer dataset was taken for the period covering 1990–2010 as the base table for this dataset. This would ensure agreement between this and other nationally analysed datasets. Data from the merged registry dataset was linked on the first 2 characters of the ONS_Number and the first two characters of the CANREG field, and the CANREGNO and last 7 characters of the ONS_number. This allowed the changes in registry boundaries and takeovers to be accounted for. In a small number of cases (18) this method of matching returned the incorrect record, these were rectified using the full CANREG field for updating.

The current Cancer Postcode Directory (CPD) and Deprivation indices were downloaded from the UKACR member's website. The CPD file was linked to the data file by postcode as supplied in ONS data. From the CPD file the Lower Super Output Area was used to return the IMD quintile for 2004, 2007 and 2010.

To produce a dataset with a consistent ICD coding pre 1995 tumours (ICD9) reported in the data were re-coded to ICD-10 using the algorithm contained in Appendix 1. All potentially person identifiable data items were removed from the data and dates were coded into Month/Year aggregates with date intervals required for analysis calculated and added. Finally, patient/tumour identifiers were recreated for this dataset to ensure that the data would be anonymised when supplied to potential users.

Where data items were available in more than one dataset, data were assessed to firstly ensure agreement with other national analyses and then in terms of completeness of data availability. The source of data is available in the data fields table below.

Data File

The data file provided is a pipe delimited text file consisting of 91 columns. The file contains a total of 6998516 records for data covering the period 1990 to 2010 inclusive. The basic profile characteristics of each data item are given in the following table. Field names and data definitions follow the NCDR data specifications definitions.

Col. No.	Field Name	Source	datatype (example format)
1	Patient_Identifier	Calculated	an15 (NCIN00000000001)
2	ons_number_Anonymised		n9 (100000000)
3	DOB_MONTH	ONS	n2
4	DOB_YEAR	ONS	n4
5	DOD_MONTH	ONS	n2
6	DOD_YEAR	ONS	n4
7	AgeAtDeath_YEARS	Calculated	n3
8	AgeAtDeath_5_YEAR_GROUP	Calculated	an9 (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)
9	Embarkation_date_MONTH	Calculated	n2
10	Embarkation_date_YEAR	Calculated	n4
11	MIND	ONS	AS ONS Data specification
12	SEX	ONS	AS ONS Data specification
13	TRACEIND	ONS	AS ONS Data specification
14	diag_date_MONTH	Calculated	n2
15	diag_date_YEAR	Calculated	n4
16	AgeAtDiagnosis_YEARS	Calculated	n3
17	AgeAtDiagnosis_5_YEAR_GROUP	Calculated	an9 (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)
18	DiagnosisToDeath_Days	Calculated	n4
19	Site4	ONS	an5
20	Site4_ICD10_Recoded	Calculated	an5
21	STAGE	ONS	AS ONS Data specification
22	STATIND	ONS	AS ONS Data specification
23	Type5	ONS	an6
24	SOA1	CPD	an9 (E00000001)
25	SOA2	CPD	an9 (E00000001)
26	CPD_CANREG	CPD	an5 (Y0101)
27	CANNET	CPD	an3 (N01)
28	ukacr_la	CPD	an4 (99AA)
29	ukacr_pct	CPD	an3
30	ukacr_cnet	CPD	an3
31	ukacr_sha	CPD	an3
32	ukacr_creg	CPD	an5
33	ukacr_gor	CPD	an1
34	ukacr_cty	CPD	an4

Col. No.	Field Name	Source	datatype (example format)
35	RegistryDataAvailability	Calculated	an1
36	REG_ethnicity	REG	As Registry data request specification
37	REG_dco_flag	REG	As Registry data request specification
38	REG_extra_regional	REG	As Registry data request specification
39	REG_cod_1a	REG	As Registry data request specification
40	REG_cod_1b	REG	As Registry data request specification
41	REG_cod_1c	REG	As Registry data request specification
42	REG_cod_2	REG	As Registry data request specification
43	REG_place_of_death	REG	As Registry data request specification
44	REG_site4	REG	As Registry data request specification
45	REG_morphology_system	REG	As Registry data request specification
46	REG_type5	REG	As Registry data request specification
47	REG_basis_code	REG	As Registry data request specification
48	REG_screening_status	REG	As Registry data request specification
49	REG_screening_category	REG	As Registry data request specification
50	REG_tumour_size	REG	As Registry data request specification
51	REG_grade	REG	As Registry data request specification
52	REG_grade_description	REG	As Registry data request specification
53	REG_gleason_grade	REG	As Registry data request specification
54	REG_Laterality	REG	As Registry data request specification
55	REG_nodes_examined	REG	As Registry data request specification
56	REG_nodes_positive_yn	REG	As Registry data request specification
57	REG_nodes_positive	REG	As Registry data request specification
58	REG_mets	REG	As Registry data request specification
59	REG_DUKE_stage	REG	As Registry data request specification
60	REG_FIGO_stage	REG	As Registry data request specification
61	REG_CLARK_level	REG	As Registry data request specification
62	REG_NPI_score	REG	As Registry data request specification
63	REG_Breslow	REG	As Registry data request specification
64	REG_TNM_clin	REG	As Registry data request specification
65	REG_t_clin	REG	As Registry data request specification
66	REG_n_clin	REG	As Registry data request specification
67	REG_m_clin	REG	As Registry data request specification
68	REG_UICC_version_clin	REG	As Registry data request specification
69	REG_neoadjuvant_flag_path	REG	As Registry data request specification
70	REG_tnm_path	REG	As Registry data request specification
71	REG_t_path	REG	As Registry data request specification
72	REG_n_path	REG	As Registry data request specification
73	REG_m_path	REG	As Registry data request specification
74	REG_UICC_version_path	REG	As Registry data request specification
75	REG_TNM_int	REG	As Registry data request specification
76	REG_t_int	REG	As Registry data request specification
77	REG_n_int	REG	As Registry data request specification
78	REG_m_int	REG	As Registry data request specification
79	REG_UICC_version_int	REG	As Registry data request specification
80	REG_CIS_stage	REG	As Registry data request specification
81	REG_surgeryTherapy	REG	As Registry data request specification
82	REG_RT	REG	As Registry data request specification

Col. No.	Field Name	Source	datatype (example format)
83	REG_CT	REG	As Registry data request specification
84	REG_HormoneTherapy	REG	As Registry data request specification
85	REG_ICDO3_topography	REG	As Registry data request specification
86	REG_ICDO3_morphology	REG	As Registry data request specification
87	REG_ExclusionFlag	REG	As Registry data request specification
88	REG_RegsitrySupplying	REG	As Registry data request specification
89	2004Quintile	UKACR LSOA2004-07-10 Income Deprivation	1 - least deprived, 2, 3, 4, 5 - most deprived
90	2007Quintile	UKACR LSOA2004-07-10 Income Deprivation	1 - least deprived, 2, 3, 4, 5 - most deprived
91	2010Quintile	UKACR LSOA2004-07-10 Income Deprivation	1 - least deprived, 2, 3, 4, 5 - most deprived

Source for NCDR output

Many of the data items contained in this dataset are available in both Registry and ONS datasets. In these cases they were taken entirely from one or the other. The table above show the shows the source dataset used for each field as either “ONS” (sourced from ONS), “REG” (sourced directly from Registry), “Calculated” (calculated from other fields) or “UKACR LSOA2004-07-10 Income Deprivation” (calculated as described above). ONS was chosen for the person and tumour identification terms to give total counts by tumour consistent with published ONS data. As the data from registry is updated with time many of the non-identification data items had better completeness in the registry dataset and were therefore used.

Regional completeness report

This report, included in the release folder as file “Regional Completeness Report_2010.xls” shows the validation rules and counts of data received by Registry. This shows the data as supplied before appending onto ONS data, hence it shows numbers of records supplied and the counts for data values supplied. It could be used for comparison of data held by registry and that data available in final release.

De-identification of the data

Patient identifiable data items and sensitive data items have been removed. The data remaining is still a rich level of tumour data and should therefore be considered sensitive. Each patient has been given a unique patient identifier, separate from those in the provided datasets. A separate linkage file of patient identifiers to identifiers provided by the source registry is available in the event of queries.

Data from the Celtic Countries is supplied in de-identified format and does not use the ONS dataset as a baseline.

Extra-Regional cases and geography fields

The repository contains only those cases that the contributing registry reported and resident at the time of diagnosis as judged by linking to the UKACR Cancer Postcode Directory File ,

September 2012, “CPD201209_NHSPD201208_TXT”, available from the UKACR website member’s area.

The English data has the geographical areas fields re-calculated from the above postcode file (PCT and GOR are English only data items.). Area codes for Northern Ireland, Scotland and Wales have been calculated from the pseudo-postcode data provided by each registry.

Governance and ownership of the data

Creation of the NCDR 2010 was a joint project between the NCIN, UKACR and ONS. However the ultimate source for the data is the UK cancer registries and the UKACR should be considered the data owners.

Governance and data release should proceed under UKACR policies.

Notes and Warnings

Dates of Birth, Diagnosis and Death

Full dates of birth, diagnosis and death if present have been removed and replaced with separate year and month fields. Single year age and 5 year age group were calculated prior to this. To ensure analysis is still possible, time difference in days between diagnosis and death, age at diagnosis and age at death have also been calculated prior to removal of dates.

Deprivation

Deprivation values have been calculated for England and Wales in line with guidance produced by UKACR in 2011. 3 fields have been included containing information on deprivation scores for 2004, 2007 and 2010. The deprivation scores are available in the file from the UKACR member's website in a file named "UKACR LSOA2004-07-10 Income Deprivation.txt".

Gleason Score

The NCDR team used the Combined Score (as per NCDR2009 dataset specification) as part of the submission process for this year of data. The additional data field showing single value / tertiary components will be added to the NCDR 2011 dataset.

Birth Date

This data field birth_date_flag will be added to NCDR 2011 dataset. Dates of birth are supplied from ONS data file for this anonymised release of data

ICD10-O2 – ICD10-O3

Tumours coded using ICD10 (topology) and O2 (morphology) codes for all registries are available in the ONS core dataset.

Variances between the ONS data submissions and the locally recorded Cancer Registry data have been noted with additional data fields added to allow local interpretation of data at ICD10-O3 where this has been submitted.

Two additional data fields have been added to the NCDR table – ONS ICD10-O2 will be used as the baseline codes, however if a cancer registry has recorded ICD10-O3 then the ICD10 site code and ICDO3 morphology code will also appear in the two additional data fields. Counts of availability of ICDO3 data by registry and or ICD 10 Code are contained in Appendix 2 and 3 respectively.

Before 1995 data were submitted by cancer registries to ONS in ICD9 and have not been subsequently converted to ICD10. For robust comparisons over time these data need to be converted to ICD10 (this needs to be converted to be useful for CIS / E-atlas etc.) As there is no official method for mapping from ICD9 to ICD10, methods used by ECRIC regional

registry were followed to recode ICD9 codes to ICD10 these methods are contained in Appendix 1

This recoding is not exact as ICD10 allows greater detail to be recorded; as such Appendix 2 shows a breakdown of recoded ICD10 codes by year to allow comparison of changes incidence by year.

The ICD09/ICD10 coding supplied by ONS is still available in this data file and can be used for comparison to recoded data.

Initial matching of ONS number/ Cancer Registry dataset

The current match is 96.8% to the ONS code. The table below shows the breakdown by year.

Year	0 - Registry Data not Available		1 - Registry Data Available		Grand Total
1990	28753	(11.1%)	231405	(88.9%)	260158
1991	28485	(10.8%)	235135	(89.2%)	263620
1992	29096	(10.6%)	245673	(89.4%)	274769
1993	24782	(9.1%)	246525	(90.9%)	271307
1994	27891	(9.9%)	252968	(90.1%)	280859
1995	2068	(0.7%)	285752	(99.3%)	287820
1996	2082	(0.7%)	289868	(99.3%)	291950
1997	2409	(0.8%)	302504	(99.2%)	304913
1998	2859	(0.9%)	310856	(99.1%)	313715
1999	7556	(2.3%)	318493	(97.7%)	326049
2000	4063	(1.2%)	326972	(98.8%)	331035
2001	2782	(0.8%)	334181	(99.2%)	336963
2002	2020	(0.6%)	335163	(99.4%)	337183
2003	2271	(0.7%)	342347	(99.3%)	344618
2004	2522	(0.7%)	355576	(99.3%)	358098
2005	2323	(0.6%)	366396	(99.4%)	368719
2006	2571	(0.7%)	379203	(99.3%)	381774
2007	8209	(2.1%)	387532	(97.9%)	395741
2008	11200	(2.7%)	410549	(97.3%)	421749
2009	15190	(3.5%)	414897	(96.5%)	430087
2010	13546	(3.2%)	403843	(96.8%)	417389
Grand Total	222678	(3.2%)	6775838	(96.8%)	6998516

Due to changes in registry boundaries and merging of registries over the time period of these datasets, the ONS supplied [CANREG] field and the first 4 characters of the [ONS_number] supplied did not match as well as if the first 2 characters of [ONS Number] and [CANREG] were used for matching. However in 18 cases there were duplicates caused using this method. This occurred as the registries had supplied more than one registration with the same CANREGNO but different [CANREG]'s. For cases identified with this situation, data was re-linked using all 4 characters of [CANREG]

Exclusion Flag

The Exclusion Flag allows the cancer registry leads to identify corrections (the flag will then be used when extraction is required for UKCIS / CCT / E-Atlas or any required SSCRG work if relevant). This will be added as a standard item for the NCDR2011 dataset request.

87 cervical cases incorrectly recorded in ONS dataset have been updated for this data release and the exclusion flag is available for these cases (these were amended manually for the inclusion of the 2009 NCDR in the UKCIS - although a small number it represents a large proportion of cervical screening cases).

Regional completeness report

This report, included in the release folder as file “Regional Completeness Report_2010.xls” shows the validation rules and counts of data received by Registry. This shows the data as supplied before appending onto ONS data, hence it shows numbers of records supplied and the counts for data values supplied. It could be used for comparison of data held by registry and that data available in final release.

Caveats due to use of ONS data from 1990 onwards

1. As the earliest year of data available in the ONS source used was 1990 the NCDR2010 is unable to do trend analysis back to 1985 or populate the UKCIS that far back - these capabilities will return once all Cancer Registries are using Encore, and 20 years of historical data is sufficient for most purposes.

2. The 1990-2010 ONS dataset codes the tumour type using ICD9 codes for 1990-1994 – these have been converted into ICD10-O2 codes in order to populate the UKCIS and CCT. The algorithm for conversion is listed in Appendix 1. Examination of this algorithm will allow cancer registries to see if it differs in any way from mapping done locally in registries which could lead to differences in the resulting ICD10 code.

Cancer Registry Caveats

The following cancer registries would like users of the data to consider the following caveats when using the NCDR2010 data.

NYCRIS – caveats	*None received.*

Thames caveats	*The Thames team has reviewed mapping of ONS / Thames (to ensure conversion of codes recorded in ICD10/03 specifically around Ovary / Haematology).*

WMCIU caveats	*Note - It's not just CIS, E-Atlas etc that needs tumours coded in a consistent coding scheme. For analysts to use data prior to 1995 they will also need the data in a single classification system. I imagine most analysts (and most people who were not working in cancer registration in the 1990s) will have no familiarity with ICD9 coding.*

SWCIS - caveats	*None received.*

Oxford - caveats	*None received.*

Encore - caveats	
NWCIS	Data on 87 cases diagnosed as malignant cervical cancers and submitted to ONS have subsequently been revised to CIN IIs and CINIIIs. These have been removed from the regional cancer registry dataset and notified to ONS but not in sufficient time for the ONS extract to be modified. Therefore these remain within the NCDR but are flagged to be excluded under the exclusion flag
NWCIS	A number of cases have been coded in ICD10 as C80 within the ONS data but appear as ICD10 C78-C79 in the regional cancer registry. Patient level information are available on these cases on request from NWCIS.
NWCIS	Cause of death data for a number of cases were recorded as text within NWCIS' old cancer registration system and it has not been possible to convert these to cause of death codes. Therefore these have been excluded from the NCDR. These are available on request from NWCIS.
NWCIS	NWCIS' historical methods for flagging DCOs is not compatible with Encore and current DCO flags being derived from Encore have not yet been tested. Therefore for this NCDR the ONS DCO flag has been used for NWCIS cases

Queries raised by Registry staff using the data

Please email any queries regarding data to NCDR2010@ncin.org.uk. All queries will be addressed by the NCDR team and summarised in a regular email output. The core NCDR team will review any queries raised, and workarounds or solutions suggested to ensure that these are addressed prior to publication of the NCDR2011 dataset specification.