Repeatability of manual coding of cancer reports in the South African National Cancer Registry, 2010

NM Dube, BV Girdler-Brown, KS Tint, P Kellett

Nomathemba Dube, BSc, MSc(Med), MPH, Epidemiologist, School of Health Systems and Public Health, University of Pretoria
South African Field Epidemiology and Laboratory Training Programme
Brendan Girdler-Brown, FFPH, MMed, FCPHM, Extraordinary Professor, School of Health Systems and Public Health, University of Pretoria
Khin Tint, MB, BS, MPH MMedSc, Epidemiologist and Field Coordinator, School of Health Systems and Public Health, University of Pretoria
South African Field Epidemiology and Laboratory Training Programme, National Institute for Communicable Diseases
Patricia Kellett, DipMedTech, Operations Manager, South African National Cancer Registry, National Institute for Occupational Health
E-mail: nmgumede@yahoo.com
Keywords: cancer reports, manual coding, repeatability, Kappa score

Data validity is a very important aspect of cancer registries in ensuring data quality for research and interventions. This study focused on evaluating the repeatability of manual coding of cancer reports in the South African National Cancer Registry (NCR). This cross-sectional study used the Delphi technique to classify 48 generic tumour sites into sites that would be most likely ("difficult") and least likely ("not difficult") to give rise to discordant results among coders. Reports received from the Charlotte Maxeke Academic Hospital were manually recoded by five coders (2 301 reports, e.g. approximately 400 reports each) for intra-coder agreement; and by four coders (400 reports) for inter-coder agreement. Unweighted kappa statistics were calculated and interpreted using Byrts' criteria. After four rounds of the Delphi technique, consensus was reached on the classification of 91.7% (44/48) of the sites. The remaining four sites were classified according to modal expert opinion. The overall kappa was higher for intra-coder agreement (0.92) than for inter-coder agreement (0.89). "Not difficult" tumour sites reflected better agreement than "difficult" tumour sites. Ten sites (skin other, basal cell carcinoma of the skin, connective tissue, other specified, lung, colorectal, prostate, oesophagus, naso-oropharynx and primary site unknown) were among the top 80% misclassified sites. The repeatability of manual coding at the NCR was rated as "good" according to Byrts' criteria. Misclassified sites should be prioritised for coder training and the strengthening of the quality assurance system.

Ⓟ Peer reviewed. (Submitted: 2012-06-07. Accepted: 2012-10-12.) © SAJEI

South Afr J Epidemiol Infect 2013;28(3):157-165

Introduction

The South African National Cancer Registry (NCR) was established in 1986[1] up to the time of this study, as a passive, pathology-based surveillance registry. In 2010, the registry received approximately 80 000 laboratory-confirmed cancer case reports, including some duplications, from 84 laboratories countrywide (Kellett P, 2001, personal communication). These case reports are normally received in a form known as Systematized Nomenclature of Medicine. Coders employed by the NCR recode these cancer reports according to the International Classification of Diseases for Oncology, third edition, an extension of the International Statistical Classification of Diseases and Related Health Problems, tenth edition. Each cancer is then allocated to one of 48 mutually exclusive "sites". These sites are used for international incidence reporting and comparison purposes.

Manual coding requires specialised experience, training and skill. The terms used in the registration documents are not always the same as those used in the coding manuals, and reports may be written or phrased in an ambiguous way. It is important to ensure correct and repeatable coding of cancer reports in order to prevent the reporting of misleading information about cancer rates.

Several studies have examined the accuracy of cancer registration data in various countries,[2-4] although they did not evaluate the aspect of repeatability. These studies have either been tumour-specific, health board-related, or both. The repeatability of manual coding at the NCR has never been evaluated, yet this attribute is essential in assessing the validity of results published by the registry. The aim of this study was to evaluate the inter-observer and intra-observer repeatability of the NCR coding for the 48 tumour sites.

Method

Study design and study setting

This was a cross-sectional study, performed at the NCR,

situated at the National Institute for Occupational Health in Johannesburg.

The Delphi technique and panel of experts

The Delphi technique was used to identify tumour sites that were either "more likely" or "less likely" to give rise to discordant results among coders. These tumour sites were termed as "difficult" and "not difficult". "Difficult" tumour sites were defined as the 24 cancer sites (out of 48) for which the manual coding was deemed to be relatively more complex and demanding, with mistakes in the site classification (possibly) to be more likely as a result. "Not difficult" tumour sites were the 24 sites that were deemed by the Delphi panel to be less difficult to classify.

Three coding experts based at the NCR were chosen to constitute the Delphi panel. The selection of these coding experts was based on their willingness to participate, as well as their level of knowledge and experience in the field of cancer coding. The Delphi panel comprised a natural scientist working at the NCR, the acting registry manager, and the registry quality assurance officer. The average length of professional experience of the experts in the cancer coding field was 9.7 years.

The purpose of the panel of coding experts was for them to reach a consensus on which 24 tumour sites, of 48, were the more "difficult" to code. The remainder would then constitute the "not difficult" group. These 48 cancer sites are those used by the NCR for the purposes of international reporting. The panel of experts met before the implementation of the Delphi technique began and before their role in the study was explained. They were asked to not discuss the ranking of cancer tumour sites among themselves, or to ask for advice from one another in this regard. During the Delphi procedure, the experts submitted their classifications, as e-mail attachments, to the researcher via a moderator. The moderator removed the names of the experts from the rankings before sending them on to the researcher.

The researcher then made a list of those sites for which no complete agreement on the classification existed among the experts. This list was then returned to the panel members, who were asked to reflect on the classification, and to classify the sites again. This time, the set of sites was reduced, as those on which there was agreement had been removed from consideration. The process was repeated until agreement was reached on 44 sites. The remaining four sites were classified according to the final round modal (most common) response.

Study population

The study population consisted of a convenience sample of 2 301 cancer reports from Charlotte Maxeke Academic Hospital in Johannesburg. These unique pathology reports were of cancer cases diagnosed between July and November 2009. The Charlotte Maxeke Academic Hospital was chosen as the source of reports because experience at the NCR had shown that cancer reports received from this facility were likely to contain a wide variety of tumour sites (Kellett P, 2011, personal communication).

Reports that referred to non-malignant conditions, or those that had accidentally been coded only once (for intra-coder agreement), were excluded from the analyses. Reports with unintelligible codes or those pertaining to foreign residents were also excluded.

Coders

All five coders who routinely recoded cancer reports at the NCR were asked to participate in this study and sign informed consent forms. They were told that participation was voluntary, and that if they agreed to participate, they could withdraw from the study at any time. The names of the coders would not be revealed to the NCR management in any way that would link the individuals to a particular repeatability score. All five coders had received training in coding provided by the quality assurance officer, and had been coding cancer reports for between nine months and 10 years.

Recoding of cancer reports

The procedures used to assess intra- and inter-coder agreement differed and are described below.

Assessing intra-coder agreement

The 2 301 cancer reports were routinely coded for the first time in September 2010, when each of the five coders was assigned approximately 400 reports made in a selected month between July and November 2009. The coders were unaware that the batches of reports that they had been given were part of the study. Coders were assigned unique identification numbers by the quality assurance officer, and the data supplied to the researcher were identified only by these unique numbers, to prevent the researcher from linking the results to specific coders.

The recoding of these reports was performed three months later, without the coders being aware that these were repeat codings. A period of three months between the two coding sessions was decided upon to minimise the coders' ability to recognise the batch of reports and remember what had been coded the first time. The first coding given by a coder for each report in September 2010 was used to classify a site as "difficult" or "not difficult".

Assessing inter-coder agreement

Two hundred "difficult" and 200 "not difficult" cancer reports were randomly selected from the 2 301 reports that had been routinely recoded in September 2010. Stata® version 11[5] was used to generate the random numbers for these selections. The 400 selected reports were duplicated, and each coder

was given the same batch of reports to code. The coders were not told that the batch was part of the study.

One coder decided not to participate in this part of the study, and so was not included at this stage. The modal site that was allocated by the four coders was used to classify each record as "difficult" or "not difficult". In those instances where there were two classifications as a "difficult" site and two as a "not difficult" site, one reading from the four was randomly selected using random numbers between 1 and 4, generated using Stata® version 11, and the selected site was used to stratify the tumour site.

Data management and analysis

Data for the analyses were entered in duplicate, in order to identify and correct errors. Data were entered directly from the coded and recoded reports. Stata® version 11 was used for the data analysis. Percentages of observations in agreement, as well as unweighted kappa statistics,[6] were calculated for both strata separately. These calculations were performed for all the observations taken as a whole within each stratum, and also for all pairwise comparisons, for example reading 1 versus reading 2 for each coder for the intra-coder datasets; and coder 1 versus coder 2 for the inter-coder datasets. Byrts' criteria[7,8] were used to interpret the kappa scores.

Codes assigned during the first coding process were compared to those assigned during the recoding process or intra-coder agreement. To assess inter-coder agreement, kappa statistics were calculated for the six possible pairwise combinations of the four coders.

Ethics committee approval

Ethics committee approval to conduct this study was obtained from the University of Pretoria's Faculty of Health Sciences Research Ethics Committee (Protocol Number 204/2010).

Results

Delphi technique

All three coding experts participated in the Delphi technique rounds. After four rounds of the Delphi technique over a period of five months, the panel of experts reached agreement on 44/48 of the tumour sites, and stratified them into 22 "not difficult" and 22 "difficult" sites. The remaining four sites, classified according to the modal opinions, were bladder, lung, myeloma and ovary. The final tumour site stratifications are presented in Table I.

Comparisons of proportions of cancer tumour sites: study sample versus the national dataset

The NCR 2003 report on the frequencies of tumours in South Africa[1] was the most recent dataset that had been coded at the NCR at the time of the study. The frequency distribution by site of tumours reported in 2010 (intra- and inter-coder datasets) were reweighted to take into account the overall national distribution of tumours reported in the 2003 NCR report. The differences between the NCR report and the study datasets were statistically significant (chi-square test, p-value < 0.01).

Table I: Final Delphi technique classification of cancer sites into "difficult" or "not difficult" tumours: National Cancer Registry, 2010

"Difficult" tumour sites	"Not difficult" tumour sites
Anus	Basal cell carcinoma of the skin
Bladder	Breast
Bone	Cervix
Brain and central nervous system	Colorectal
Burkitt's lymphoma	Gum
Connective tissue	Intestine
Endocrine	Kaposi's sarcoma
Eye	Kidney
Haematology other	Larynx
Hodgkin lymphoma	Lung
Ill-defined	Melanoma
Leukaemia	Mesothelioma
Lip	Oesophagus
Liver and bile duct	Ovary
Mouth	Pancreas
Myeloma	Penis
Naso-oropharynx	Prostate
Non-Hodgkin's lymphoma	Squamous cell carcinoma of the skin
Other specified	Stomach
Placenta	Testis
Primary site unknown	Thyroid
Salivary	Tongue
Skin other	Vagina
Uterus	Vulva

Those tumour sites that were under- or over-represented in the two study samples are listed in Table II.

Reports of non-malignant tumours, or those pertaining to foreign residents [242 (10.52%)], and those with only one code recorded [203 (8.82%)] constituted 19.34% of the 2 301 reports in the study, and were excluded from the analysis.

Intra-coder agreement

There were 130 disagreements, spread between 31 tumour sites, when assessing intra-coder agreement (Figure 1 and Table III). Eighty-two of 455 possible pairings for these 31 sites, or 18.02%, were for sites identified as "difficult" to code, while 48 of a possible 1 167 pairings, or 4.11%, were for sites identified as being "not difficult" to code.

Of the sites where discrepancies were found in more than 10% of the sample size for the site (Figure 1), the first 18 sites listed in Figure 1 were responsible for 65.4% of all the disagreements. The leading four sites were classified as more

Table II: Under- and over-represented tumour sites in the intra- and inter-coder datasets in 2010 (compared to the prevalence of sites reported by the National Cancer Registry for South Africa in 2003)

Intra-coder dataset*			
"Difficult" tumour sites		"Not difficult" tumour sites	
Over-represented	**Under-represented**	**Over-represented**	**Under-represented**
Anus	Bladder	Breast	BCC
Connective tissue	Leukaemia	Cervix	Lung
Eye	Myeloma	Kaposi's sarcoma	SCC
Haematology other	PSU	Melanoma	
Hodgkin's lymphoma		Penis	
Mouth		Vulva	
Non-Hodgkin's lymphoma			
Uterus			
Inter-coder dataset*			
"Difficult" tumour sites		"Not difficult" tumour sites	
Over-represented	**Under-represented**	**Over-represented**	**Under-represented**
Anus	Bladder	Breast	BCC
Bone	Leukaemia	Cervix	Lung
Burkitt lymphoma	PSU	Kaposi's sarcoma	
Non-Hodgkin's lymphoma	Skin other	Melanoma	
		Penis	

BCC: basal cell carcinoma of the skin, PSU: primary site unknown, SCC: squamous cell carcinoma of the skin

*: The reason why some sites do not appear in both datasets is because the samples used for the two datasets were different

"difficult" to code. Seven of the remaining 14 sites in this list had also been classified as "difficult" to code.

Individual pairwise percentage agreement values for the five intra-observer comparisons ranged from 95.63-97.84% for the "not difficult" sites; and from 74.2-93.46% for the "difficult" sites. The kappa statistics ranged from 0.95-0.97 for the "not difficult" sites, and from 0.72-0.93 for the "difficult" sites. Summaries of the percentage agreement, kappa statistics and Byrts' interpretation of the kappa statistics, for intra-coder agreement, are presented in Table IV.

Inter-coder agreement

Data for this part of the study were available for four coders. Eighty-three of the codings, split between 25 sites, resulted in coding that differed from the consensus site. This is illustrated in Figure 2 and Table V. Fifty of these 83 codings of 499 possible codings, or 10.02%, were for sites classified as "difficult" to code, while 33 of a possible 653 codings, or 5.05%, were for sites classified as "not difficult" to code.

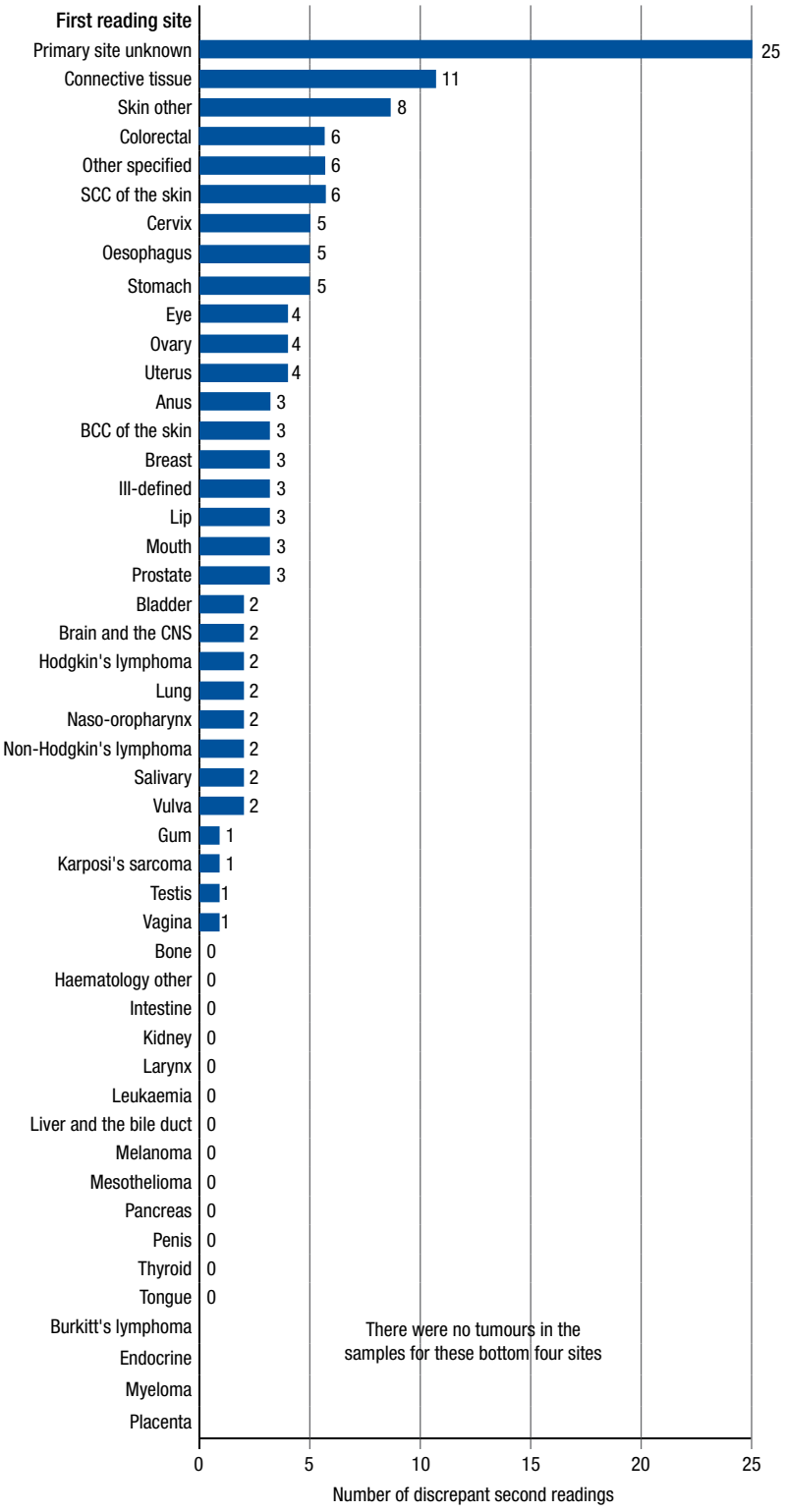
After concentrating on the sites where the discrepant codings made up more than 10% of the possible codings for that

Table III: Table showing total number of intra-coder disagreements by tumour site, National Cancer Registry, 2010

Site	Disagreement	
	n/N	%
Primary site unknown	25/64	39.06
Connective tissue	11/34	32.35
Skin other	8/19	42.11
Colorectal	6/78	7.69
Other specified	6/14	42.86
SCC of the skin	6/70	8.57
Cervix	5/222	2.25
Oesophagus	5/50	10
Stomach	5/38	13.16
Eye	4/26	15.38
Ovary	4/12	33.33
Uterus	4/50	8
Anus	3/12	25
BCC of the skin	3/137	2.19
Breast	3/346	0.87
Ill-defined	3/4	75
Lip	3/7	42.86
Mouth	3/30	10
Prostate	3/124	2.42
Bladder	2/25	8
Brain and the CNS	2/15	13.33
Hodgkin's lymphoma	2/24	8.33
Lung	2/6	33.33
Naso-oropharynx	2/13	15.38
Non-Hodgkin's lymphoma	2/110	1.82
Salivary	2/8	25
Vulva	2/15	13.33
Gum	1/1	100
Kaposi's sarcoma	1/161	0.62
Testis	1/5	20
Vagina	1/2	50
Bone	0/7	0
Haematology other	0/12	0
Intestine	0/3	0
Kidney	0/9	0
Larynx	0/25	0
Leukaemia	0/5	0
Liver and bile duct	0/7	0
Melanoma	0/23	0
Mesothelioma	0/2	0.
Pancreas	0/2	0
Penis	0/9	0
Thyroid	0/13	0
Tongue	0/17	0
Burkitt's lymphoma	-	-
Endocrine	-	-
Myeloma	-	-
Placenta	-	-

SCC: squamous cell carcinoma, BCC: basal cell carcinoma, CNS: central nervous system

There were 130 disagreements in total out of 1 856 reports, for which two codings were available



SCC: squamous cell carcinoma, BCC: basal cell carcinoma, CNS: central nervous system

There were 130 disagreements in total out of 1 856 reports, for which two codings were available

Figure 1: Graph showing total number of intra-coder disagreements by tumour site, National Cancer Registry, 2010

Table IV: Priority cancer tumour sites for strengthened training and quality control, National Cancer Registry, 2010

Intra-coder samples		Inter-coder samples	
Proportion of expected misclassifications for the 2003 dataset*		Proportion of expected misclassifications for the 2003 dataset*	
Site		**Site**	
Skin other	0.31	BCC of the skin	0.31
PSU	0.21	PSU	0.16
Lung	0.11	Lung	0.12
BCC of the skin	0.04	Colorectal	0.11
Connective tissue	0.03	Naso-oropharynx	0.04
Oesophagus	0.03	Connective tissue	0.03
Colorectal	0.03	Prostate	0.03
Other specified	0.03	Other specified	0.02
Total	**0.78**	**Total**	**0.80**

BCC: basal cell carcinoma, PSU: primary site unknown

*: Expected misclassifications in the entire annual dataset, after applying the proportions misclassified in the samples in 2010 to the total actual cases at each site reported in 2003

site (Figure 2), it was found that 14 sites were responsible for 65.4% of all the disagreements. All these sites, except colorectal, squamous cell carcinoma of the skin, mouth and Burkitt's lymphoma, also appeared in the list of leading sites for misclassification after intra-coder agreement testing.

The inter-coder agreement percentages for the six pairwise comparisons of readings varied from 88.60-96.41% ("not difficult" sites), and from 83.85-95.12% ("difficult" sites). The kappa statistics varied between 0.87-0.96 ("not difficult" sites), and between 0.82-0.94 ("difficult" sites). A summary of the percentage agreement, kappa statistics and Byrts' interpretation of the kappa statistics is presented in Table IV.

The proportions of the samples that were misclassified at each site did not take into account the relative contribution of a particular site to the national dataset for a year. In order to take the relative frequencies of different sites in the full national dataset into account, the misclassified proportions were multiplied by the number of reports recorded for each site in 2003. This procedure provided the number of expected misclassifications by site for the 2003 data.

These expected misclassifications were then expressed as percentages of the total expected misclassifications, and ranked in descending order. Using both the intra- and inter-coder data in this way, eight sites accounted for the top 80% of expected misclassifications. These results are presented in Table VI.

Discussion

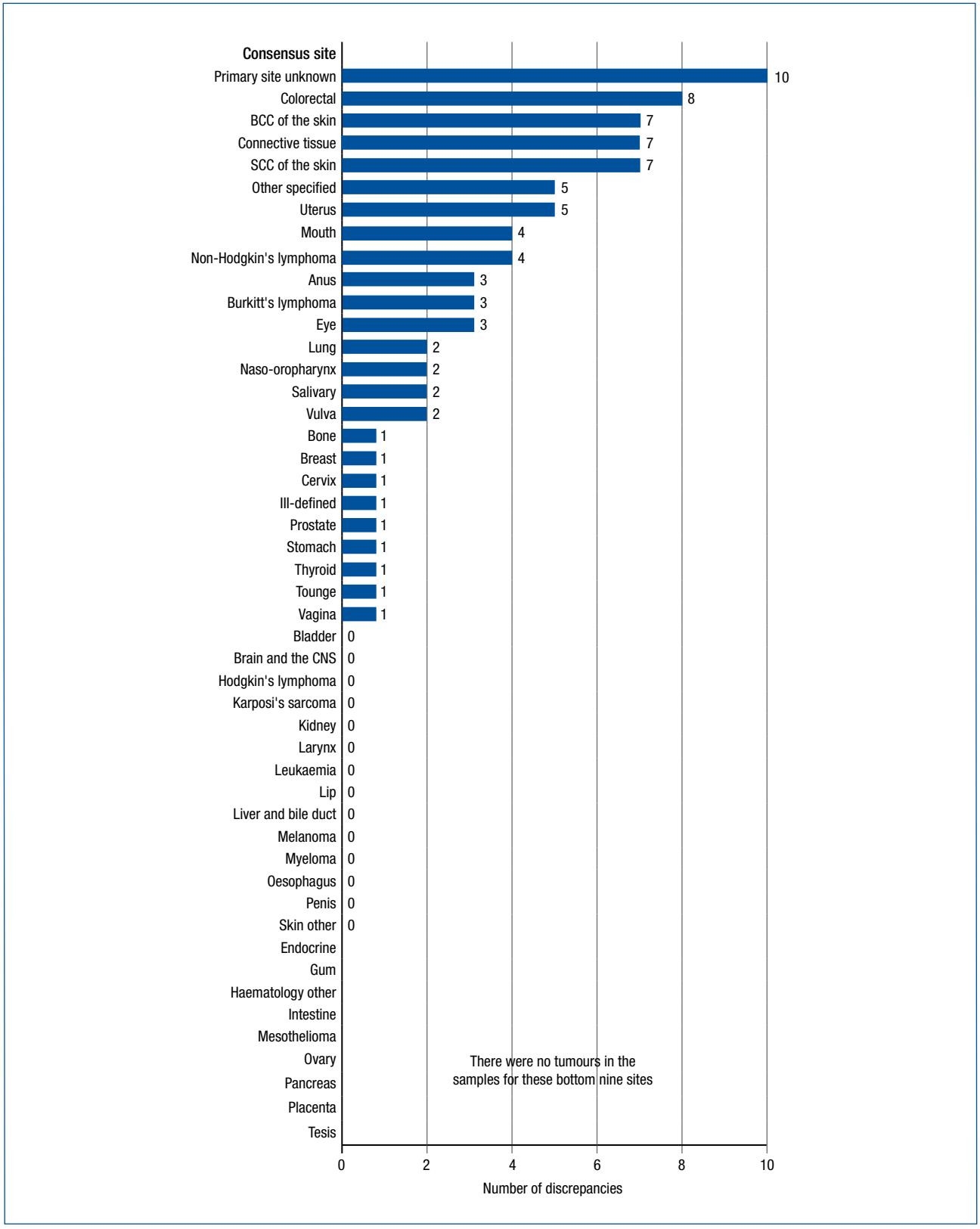
This was the first study to be conducted to assess the repeatability of manual coding of cancer reports at the NCR. Expert opinion on "difficult" and "not difficult" cancer tumour sites was combined with the manual cancer coding of

Table V: Table showing cumulative numbers and percentage of inter-coder disagreements by tumour site, National Cancer Registry, 2010

Site	Disagreements	
	n/N	%
Primary site unknown	10/72	13.89
Colorectal	8/56	14.29
BCC of the skin	7/88	7.95
Connective tissue	7/51	13.73
SCC of the skin	7/66	10.61
Other specified	5/26	19.23
Uterus	5/76	6.58
Mouth	4/38	10.53
Non-Hodgkin's lymphoma	4/186	2.15
Anus	3/15	20
Burkitt's lymphoma	3/7	42.86
Eye	3/28	10.71
Lung	2/12	16.67
Naso-oropharynx	2/7	28.57
Salivary	2/20	10
Vulva	2/8	25
Bone	1/19	5.26
Breast	1/175	0.57
Cervix	1/136	0.74
Ill-defined	1/4	25
Prostate	1/56	1.79
Stomach	1/28	3.57
Thyroid	1/12	8.33
Tongue	1/12	8.33
Vagina	1/4	25
Bladder	0/34	0
Brain and the CNS	0/19	0
Hodgkin's lymphoma	0/28	0
Kaposi's sarcoma	0/54	0
Kidney	0/4	0
Larynx	0/12	0
Leukaemia	0/7	0
Lip	0/16	0
Liver and bile duct	0/4	0
Melanoma	0/11	0
Myeloma	0/8	0
Oesophagus	0/31	0
Penis	0/12	0
Skin other	0/4	0
Endocrine	-	-
Gum	-	-
Haematology other	-	-
Intestine	-	-
Mesothelioma	-	-
Ovary	-	-
Pancreas	-	-
Placenta	-	-
Testis	-	-
Testis	-	-

SCC: squamous cell carcinoma, BCC: basal cell carcinoma, CNS: central nervous system

There were 83 disagreements in total out of a possible 1 446. (There were 1 446 paired comparisons with available consensus report codings, after allowing for non-responses)



SCC: squamous cell carcinoma, BCC: basal cell carcinoma, CNS: central nervous system

There were 83 disagreements in total out of a possible 1 446. (There were 1 446 paired comparisons with available consensus report codings, after allowing for non-responses)

Figure 2: Graph showing cumulative numbers and percentage of inter-coder disagreements by tumour site, National Cancer Registry, 2010

Table V: Table showing cumulative numbers and percentage of inter-coder disagreements by tumour site, National Cancer Registry, 2010

Site	Disagreements	
	n/N	%
Primary site unknown	10/72	13.89
Colorectal	8/56	14.29
BCC of the skin	7/88	7.95
Connective tissue	7/51	13.73
SCC of the skin	7/66	10.61
Other specified	5/26	19.23
Uterus	5/76	6.58
Mouth	4/38	10.53
Non-Hodgkin's lymphoma	4/186	2.15
Anus	3/15	20
Burkitt's lymphoma	3/7	42.86
Eye	3/28	10.71
Lung	2/12	16.67
Naso-oropharynx	2/7	28.57
Salivary	2/20	10
Vulva	2/8	25
Bone	1/19	5.26
Breast	1/175	0.57
Cervix	1/136	0.74
Ill-defined	1/4	25
Prostate	1/56	1.79
Stomach	1/28	3.57
Thyroid	1/12	8.33
Tongue	1/12	8.33
Vagina	1/4	25
Bladder	0/34	0
Brain and the CNS	0/19	0
Hodgkin's lymphoma	0/28	0
Kaposi's sarcoma	0/54	0
Kidney	0/4	0
Larynx	0/12	0
Leukaemia	0/7	0
Lip	0/16	0
Liver and bile duct	0/4	0
Melanoma	0/11	0
Myeloma	0/8	0
Oesophagus	0/31	0
Penis	0/12	0
Skin other	0/4	0
Endocrine	-	-
Gum	-	-
Haematology other	-	-
Intestine	-	-
Mesothelioma	-	-
Ovary	-	-
Pancreas	-	-
Placenta	-	-
Testis	-	-
Testis	-	-

SCC: squamous cell carcinoma, BCC: basal cell carcinoma, CNS: central nervous system

There were 83 disagreements in total out of a possible 1 446. (There were 1 446 paired comparisons with available consensus report codings, after allowing for non-responses)

Table IV: Summaries of percentage agreement, kappa statistics and Byrts' interpretation of the kappa statistics for intra- and inter-coder agreement, National Cancer Registry, 2010

	Intra-coder agreement		Inter-coder agreement	
Strata	**"Difficult"**	**"Not difficult"**	**"Difficult"**	**"Not difficult"**
Agreement (%)	83.13	96.50	87.13	96.40
Kappa	0.81	0.96	0.85	0.92
Byrts	Good to very good	Excellent	Very good	Very good
	Combined		**Combined**	
Agreement (%)	92.87		90.12	
Kappa	0.92		0.89	
Byrts*	Excellent		Very good	

*:Byrts' interpretation of the kappa statistic

corresponding cancer reports performed by coders based at the NCR. Quantification of the level of agreement within and between coders was performed. This study did not attempt to investigate the important issue of the factors that may have contributed to misclassifications.

Despite the limited number of Delphi panel participants in our study, we ensured that a panel was selected that varied in professional experience, education and employment, to guarantee a better performance, as well as enrich the results of the Delphi procedure, as advised by Bantel et al.[9]

Our study datasets were not representative of the proportions of different sites that are represented in the national dataset of cancers for 2003. The greatest differences in the tumour site proportions between our dataset and those of 2003 were observed in the intra-coder dataset. The intra-coder dataset over-represented the "difficult" tumour sites that were observed in 2003. Since the "difficult" sites had a lower kappa, our study may actually have underestimated the true kappa that would be expected for all cancers reported in a particular year. As a result, we believe that it is likely that our non-representative sample may have led to an underestimation of the true repeatability. This poor representativeness led us to exclude confidence intervals for our agreement measure estimates from the results.

Increasing the number of individuals making the same decision, i.e. inter-coder assessment, was expected to increase the opportunity for disagreement. Therefore, it was unexpected that inter-coder repeatability would be higher than intra-coder repeatability, as was the case with the "difficult" tumour sites. However, the dataset that was used for the inter-coder kappa estimates was not the same as the one used for the intra-coder estimates. In addition, the intra-coder sample size was larger, and as a result, gave rise to more precise estimates. Hence, the kappas for the inter- and intra-coder estimations should probably not be compared directly. Conclusions should also not be drawn from the observed differences in this case.

As might be expected, the repeatability of manual coding was better for the "not difficult" tumour sites than it was for the "difficult" tumour sites in both the intra- and inter-coder assessments. Care should be taken when interpreting the percentage of results that are discordant for sites where the sample size is small (say less than 10), as the discrepancies could easily have arisen by chance alone.

Although our study revealed high ("very good" to "excellent") kappa statistics for overall intra- and inter-coder agreement, the repeatability of cancer recoding at the NCR was perhaps less than what management would have expected. There is a small, but regular, tendency to misclassify, and this is particularly so for a small subset of sites. Therefore, as a first priority, the NCR should focus its training and quality assurance measures on these sites. This study also provided a baseline against which the success of any future strengthened NCR training and quality programmes can be compared.

The tumour sites that were misclassified the most were interpreted as those that the coders had the most difficulty coding. It is noted from Table IV, that for intra- and inter-coder assessment, eight out of the 48 sites were involved in the top 80% of expected misclassifications if the rates of misclassification are to be applied to the complete national dataset. Six sites were common to both analyses and there were two unique sites listed for each stratum. The common sites were basal cell carcinoma of the skin, primary site unknown, lung, colorectal, connective tissue and other specified.

We used the Delphi panel technique to classify the 48 sites into two strata. The Delphi panel technique is well established for this kind of undertaking. However, our application of the technique had some limitations. Firstly, the selection criteria for consensus changed after the last round of the Delphi panel. (The mode of the panel responses was used for the last four sites where consensus had not been reached after four rounds). According to the rules of the Delphi technique, selection criteria should be the same in all rounds.[10] Secondly, no feedback was given to the Delphi panel after each round of the technique regarding panel responses, other than the list of tumour sites they had disagreed on and were required to reclassify. Nevertheless, the fact that the kappa statistics were lower for the "difficult" strata in all cases made us confident that the strata were reasonably correctly identified by the Delphi panel.

This study will be used by NCR management staff to focus on those sites that are most likely to have a greater impact on the registry data and reports. It would also be useful for trainers to know what the coding errors were. For example, if basal cell carcinoma was often miscoded as "skin other", then this would help trainers to understand better what the issues are that need to be addressed in the training programme. Further qualitative research, perhaps in the form of focus groups, may help to identify why errors are being made.

Conclusion

Intra- and inter-coder agreement among experienced coders in manual cancer coding was described. The level of agreement, quantified in terms of kappa statistics, showed good to excellent agreement beyond what would be expected by chance for both intra- and inter-coder repeatability. The estimated measures of agreement probably underestimate the true measures because of bias in our samples.

Important areas that would benefit from further attention are recommended, in particular the focused training of coders and the strengthening of the quality assurance system. Training should focus especially on the coding of the sites: basal cell carcinoma, primary site unknown, lung, colorectal, connective tissue, other specified, skin other, naso-oropharynx, prostate and oesophagus.

Improvements in the repeatability of the coding process will enhance the credibility of the NCR's annual cancer incidence reports. In addition, the impact of training and quality assurance measures may be monitored by repeating this study in the future.

Acknowledgments

This study was carried out by Nomathemba Dube in partial fulfilment of the requirements for her Master's degree in Public Health in the School of Health Systems and Public Health at the University of Pretoria. We would like to express our gratitude to Margaret Urban and Dikeledi Rasoaisi, who formed part of the Delphi panel of experts.

Funding

The study received financial support from the South African Field Epidemiology and Laboratory Training Programme (SAFELTP), funded by the Centers for Disease Control and Prevention (CDC). The contents of this report are solely the responsibility of the authors and do not reflect the views of the SAFELTP or the CDC.

Competing interests

The authors declare no competing interests.

References

1. Mqoqi N, Kellett P, Sitas F, Jula M. Cancer in South Africa, 1998-1999. Johannesburg: National Cancer Registry of South Africa, National Health Laboratory Services; 2004.
2. Bray F, Parkin DM. Evaluation of quality data in the cancer registry: principles and methods. Part 1: Comparability, validity and timeliness. Eur J Cancer. 2009;45(5):9:747-755.
3. Bhurgri Y, Bhurgri A, Hasan SH. Comparability and quality control in cancer registration; Karachi (data monitoring 1995-2001). J Pak Med Assoc. 2002;52(7):7:301-307.
4. Brewster D. Improving the quality of cancer registration data. J R Soc Med. 1995;88(5):268-271.
5. Stata® (version 11). StataCorp, Texas; 2009.
6. Altman D. Practical statistics for medical research. Washington: Chapman and Hall; 1999.
7. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol. 1993;46(5):7:423-429.
8. Girdler-Brown BV. Biostatistics BOS 870 Week 2. Introduction to health measurement: hypothesis testing. Pretoria: University of Pretoria; 2010.
9. Bantel K. Comprehensiveness of strategic planning: the importance of heterogeneity of a top team. Psychol Rep. 1993;73(1):35-49.
10. Campbell SM, Hann M, Roland MO, et al. The effect of panel membership and feedback on ratings in a two-round Delphi survey: results of a randomised controlled trial. Med Care. 1999;37(9):964-968.