

BMJ Open Accuracy of colorectal cancer ICD-9-CM codes in Italian administrative healthcare databases: a cross-sectional diagnostic study

Francesco Cozzolino,¹ Ettore Bidoli,² Iosief Abraha,^{1,3} Mario Fusco,⁴ Gianni Giovannini,¹ Paola Casucci,⁵ Massimiliano Orso,¹ Annalisa Granata,⁴ Marcello De Giorgi,⁵ Paolo Collarile,⁶ Valerio Ciullo,⁴ Maria Francesca Vitale,⁴ Roberto Cirocchi,⁷ Walter Orlandi,⁸ Diego Serraino,² Alessandro Montedori,¹ for the D.I.V.O. Group

To cite: Cozzolino F, Bidoli E, Abraha I, *et al.* Accuracy of colorectal cancer ICD-9-CM codes in Italian administrative healthcare databases: a cross-sectional diagnostic study. *BMJ Open* 2018;**8**:e020630. doi:10.1136/bmjopen-2017-020630

► Prepublication history for this paper is available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2017-020630>).

Received 14 November 2017
Revised 30 April 2018
Accepted 1 June 2018

ABSTRACT

Objectives To assess the accuracy of *International Classification of Diseases, Ninth Revision – Clinical Modification* (ICD-9-CM) codes in identifying subjects with colorectal cancer.

Design A diagnostic accuracy study comparing ICD-9-CM codes (index test) for colorectal cancers with medical chart (as a reference standard). Case ascertainment based on neoplastic lesion(s) within the colon/rectum and histological documentation from a primary or metastatic site positive for colorectal cancer.

Setting Administrative databases from the Umbria region, Azienda Sanitaria Locale (ASL) Napoli 3 Sud (NA) region and Friuli Venezia Giulia (FVG) region.

Participants We randomly selected 130 incident patients from each hospital discharge database, admitted between 2012 and 2014, having colorectal cancer ICD-9 codes located in primary position, and 94 non-cases, that is, patients having a diagnosis of cancer (ICD-9 140–239) other than colorectal cancer in primary position.

Outcome measures Sensitivity, specificity and predictive values for 153.x code (colon cancer) and for 154.x code (rectal cancer).

Results The positive predictive value (PPV) for colon cancer diagnoses was 80% for Umbria (95% CI 73% to 87%), 81% for NA (95% CI 73% to 88%) and 80% for FVG (95% CI 72% to 87%). The sensitivity ranged from 98% to 99%, while the specificity ranged from 78% to 80% in the three units. For rectal cancer, the PPV was 84% for Umbria (95% CI 77% to 90%), 80% for NA (95% CI 72% to 87%) and 81% for FVG (95% CI 73% to 87%). The sensitivities ranged from 98% to 100%, while the specificity estimates from 79% to 82%.

Conclusions Administrative databases in Italy can be a valuable tool for cancer surveillance as well as monitoring geographical and temporal variation of cancer practice.

INTRODUCTION

Large-scale population-based studies have relied on administrative databases of patients with specific diseases. Generally,

Strengths and limitations of this study

- This is the first study that has evaluated the accuracy of *International Classification of Diseases, Ninth Revision – Clinical Modification* (ICD-9-CM) codes for colorectal cancer in three large computerised Italian administrative databases using the same cancer case definition.
- The strength of this study is that it used medical chart review as a reference standard to ascertain cases of colorectal cancer.
- The validity assessment for ICD-9-CM codes could be limited since colorectal cancer diagnoses in secondary position were not evaluated.
- Results from the present validation assessment cannot be generalised into other settings.

administrative databases comprise hospital discharge data, prescription data and laboratory data.^{1 2} These data have the advantage of being readily available, and it is less costly to assess long-term outcomes in large cohort populations. Usually, the diagnosis of diseases stored in administrative databases is associated with the *International Classification of Diseases, Ninth Revision* (ICD-9) or *Tenth Revision* (ICD-10) code edition. The ICD is designed to map health conditions to corresponding generic categories together with specific variations.^{3–6} As administrative databases are not generated for research or quality assessment purposes, it is imperative to assess their validity to avoid misclassification and disseminate inaccurate information. The process of validation consists in evaluating the consistency of information within the administrative databases and the information contained in the clinical charts, which are generally considered the gold standard.⁷



For numbered affiliations see end of article.

Correspondence to

Dr Iosief Abraha;
iosief_a@yahoo.it

In Italy, despite the wide availability of administrative databases, only a few regional databases have been validated for a limited number of ICD-9 codes of diseases. These validated databases were able to exploit their capability as documented by a systematic review.²

Colorectal cancer is the third most common cancer worldwide and almost 55% of cases occur in more developed geographical regions with rectal cancer accounting for ~30% of cases.⁸ It is estimated to be the fourth most common cancer cause of death around the world, accounting for approximately 1.2 million new cases and 600 000 deaths per year.^{8,9} As colorectal cancer generates interest in the public and scientific community^{10,11} it is an important concern for the public due to its economic burden.^{12,13} Epidemiology of colorectal cancer and treatment patterns,^{14,15} as well as potential clinical and economic outcomes,^{16–18} can be evaluated using validated administrative databases.

As reported in our published protocol,¹⁹ the objective of the present study was to evaluate the accuracy of the ICD-9-CM codes related to colorectal cancers in three large Italian administrative healthcare databases.

METHODS

Setting and data source

Administrative databases

The administrative database target for the present study was two regional databases and one local database, represented by the Umbria region (890 000 residents), the Local Health Unit 3 of Napoli (NA) (1170 000 residents) and the Friuli Venezia Giulia (FVG) region (1227 000 residents). The corresponding operative units, the Regional Health Authority of Umbria (for the Umbria Region), the Registro Tumori Regione Campania (for the Local Health Unit 3 of NA) and the Centro di Riferimento Oncologico Aviano (for the FVG region), conducted the same validation process.

Local and regional Italian healthcare administrative databases regularly collect data about patient medical records from public and private hospitals including demographics, hospital admission and discharge dates, vital statistics, the admitting hospital department, the principal diagnosis and a maximum of five secondary discharge diagnoses, as well as surgical and diagnostic procedures. Additionally, these databases record all information regarding drug prescriptions listed in the National Drug Formulary and the basic characteristics of patients' physicians. The unique national identification code of the residents permits linking the different types of information within the database, and since the healthcare is covered almost entirely by the Italian National Health System most residents' significant healthcare information can be traced within the healthcare databases.

The records in the healthcare databases are provided with a code with which it is possible to identify the corresponding medical charts of the patients that are located in secured archives. The code that identifies

a medical chart is generated using several basic codes that take into account the region, the local health unit, the department of admission and other chronologically progressive codes that provide a unique identity to the medical chart event at the national level and avoid duplicate cases.

Source population

The source population was represented by permanent residents aged 18 years or older in the three local or regional areas. Eligible subjects were residents that have been discharged from hospital with a diagnosis of colorectal cancer. Residents that were admitted outside the regional territory of competence were excluded from analysis due to difficulty in obtaining the medical charts.

Case selection and sampling method

In each administrative database, patients with occurrence of diagnosis of colorectal cancer between 1 January 2012 and 31 December 2014 were identified using the ICD-9-CM codes located in primary position of the hospital discharge: (A) 153.x for colon cancer and (B) 154.0, 154.1 and 154.8 for rectal cancer.

To obtain a cohort of first cases in primary position, records subsequent to the index date were deleted. Subsequently, prevalent cases, that is, those with the same diagnosis (ICD-9-CM codes 153.x or 154.0, 154.1 and 154.8 in any position) in the 5 years (2007–2011) before the period of interest, were excluded. This cohort represented our target population from which a sample of cases was obtained using a random sampling method.

For controls (non-cases), first subjects aged 18 or higher with diagnosis of cancer disease, that is, patients having in primary position a diagnosis of cancer (ICD-9 140–239), were identified. Subsequently, from this cohort subjects with colorectal cancer (153.x or 154.0, 154.1 and 154.8 in any position) were excluded obtaining a target population for our controls. From this population we obtained a sample of controls using a random sampling method.

Chart abstraction and case ascertainment

The corresponding medical charts of the randomly selected samples of cases and non-cases were obtained from hospitals for validation purposes. Where available the following information was retrieved from the medical charts: clinical chart number, hospital and ward, date of birth, sex, dates of hospital admission and discharge, signs and symptoms, any diagnostic procedures that contributed to the diagnosis of the cancer, any pharmacological or surgical therapy that was provided for treatment of the cancer.

An initial consensus chart review was performed by trained medical chart reviewers independently examining the same number of medical charts (n=20). The inter-rater agreement regarding the presence or absence of colorectal cancer among the pairs of reviewers within each unit was calculated. Discrepancies were resolved through the involvement of an oncologist (RC).

Case ascertainment of cancer within medical charts was based on (A) The presence of a primary lesion in the colon-rectum, documented by imaging or endoscopy and (b) The histological documentation of cancer from a primary or metastatic site.¹⁹ Following the consensus review, data abstraction was completed independently by the same reviewers. To ensure consistency among all the reviewers, cases with uncertainty were discussed and resolved through third party involvement (RiCh).

Validation criteria

For colon cancer, we considered the ICD-9-CM codes 153.x valid when there is evidence of a neoplastic lesion within the colon documented by endoscopy (eg, colonoscopy) or imaging (eg, abdominal ultrasound or CT scans), and a histological diagnosis from a primary or metastatic site positive for adenocarcinoma, squamous cell carcinoma or neuroendocrine carcinoma.

For rectal cancer, we considered the ICD-9-CM codes 154.0, 154.1 and 154.8 valid when there is evidence of a neoplastic lesion, in the rectosigmoid junction or in the rectum, documented by endoscopy or imaging, and a histological diagnosis from a primary or metastatic site positive for adenocarcinoma or squamous cell carcinoma.

Statistical analysis

We calculated that a sample of 130 charts of cases was necessary to obtain an expected sensitivity of 80% with a precision of 10% and a power of 80%. For specificity calculations, we randomly selected non-cases, that is, records without the ICD-9-codes of interest from hospital discharges. We calculated that a sample of 94 charts of non-cases was sufficient to obtain an expected specificity of 90% with a precision of 10% and a power of 80%.¹⁹ The corresponding medical charts were retrieved and evaluated.

Sensitivity, specificity and predictive values with corresponding 95% CIs were analysed separately for colon and rectum cancer ICD-9-CM codes by constructing 2×2 tables.

In case of missing medical charts we performed a formal sensitivity analysis based on a worst-case scenario in which the missing cases were considered as false positive.

To ensure the quality of any reporting of the results from the present study, the recommended guidelines based on the criteria published by the Standards for Reporting of Diagnostic accuracy (STARD) initiative for the accurate reporting of investigations of diagnostic studies were followed.^{20–22}

Patient and public involvement

Neither patients nor the public were directly involved with the development or design of this study. This was a cross-sectional diagnostic study based on the consultation of medical charts.

RESULTS

Colon cancer

The κ statistics between evaluators was higher than 0.90 across the three operative units.

The exclusion of the estimated prevalent cases of invasive colon cancer in primary position of the hospital discharges allowed the identification of a cohort of 1725 new cases in Umbria, 1414 in NA and 1307 in FVG. From these cases, each unit randomly selected 130 cases of which the corresponding medical charts were requested for evaluation. Two (1.5%) and nine (6.9%) medical charts were not available from Umbria and NA, respectively. **Figure 1** displays the identification of cases from the three operative units. For the non-cases, each unit randomly selected 94 medical charts. Two medical charts of non-cases from Umbria were missing.

In terms of ICD-9-CM subgroups, the most common were 153.6 (34%) (ie, ascending colon cancer) followed by 153.3 (20%) (ie, sigmoid colon cancer) in Umbria; 153.3 (33%) followed by 153.6 (21%) in NA; 153.6 (28%) followed by 153.3 (27%) in FVG. The mean age ranged between 68 years and 73 years. The majority of the cases were identified in surgical departments with a percentage higher than 77%. The types of surgical intervention were similar across the three territorial units with hemicolectomy being the most performed surgical intervention. **Table 1** reports the basic characteristics of the incident colon cancer cases in each unit.

Across the three operative units, from the medical charts of 379 cases, 448 lesions were identified. Of 305 positive cases for malignant carcinoma (true positives) 55 also had a benign tumour; whereas of the 74 false positives 14 had a second lesion that resulted negative or with a benign tumour.

Accuracy estimates results were similar across the three units. The positive predictive value (PPV) was 80% for Umbria (95% CI 73% to 87%), 81% for NA (95% CI 73% to 88%) and 80% for FVG (95% CI 72% to 87%); the negative predictive value (NPV) was 99% for Umbria (95% CI 94% to 100%), 98% for NA (95% CI 93% to 100%) and 98% for FVG (95% CI 93% to 100%). The sensitivity of colon cancer cases confirmed by instrumental and histological examinations was 99% (95% CI 95% to 100%) for Umbria, 98% (95% CI 93% to 100%) for NA and 98% (95% CI 93% to 100%) for FVG. The specificity estimates were 78% (95% CI 70% to 86%) for Umbria, 80% (95% CI 72% to 87%) for NA and 78% (95% CI 69% to 85%) for FVG.

Table 2 provides cross tabulation of the ICD-9-CM code results from the results of the medical charts.

Description of misclassifications

The number of false positives that may cause misclassification were 74 across the three operative units and were categorised as cases with missing histological documentation and cases with possible negative histology for colon cancer. Cases with missing histological documentation had evidence of metastases (n=8) and chemotherapy or

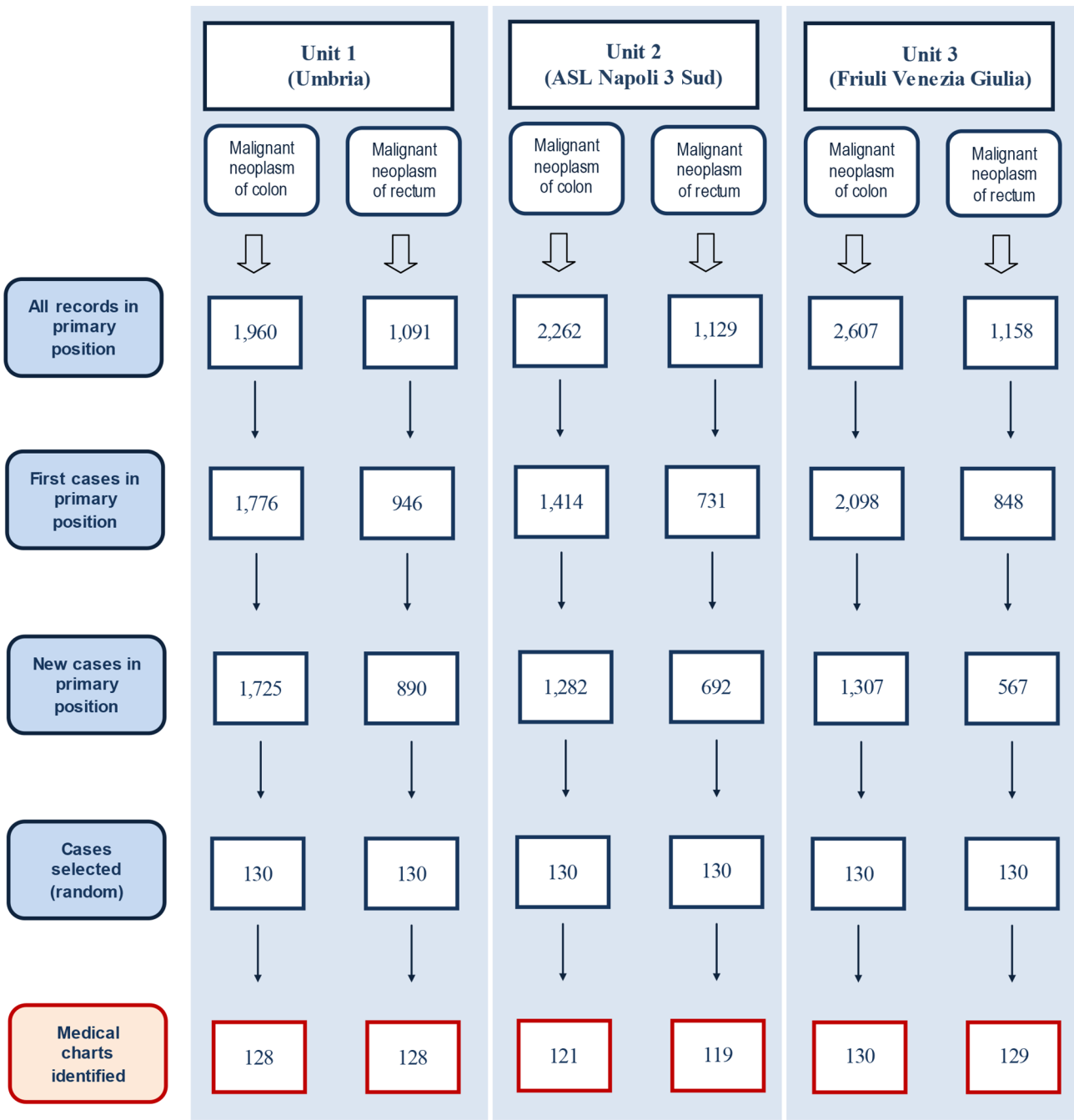


Figure 1 Flow chart of incident colorectal cancer cases identification in primary position from the three administrative databases and the corresponding charts examined.

radiotherapy (n=1), or biopsy was not performed due to cachexia or death (n=2) within the same medical chart; in the subsequent medical chart review 13 cases had evidence of colon cancer and one case had evidence of rectal cancer on histological documentation (table 3).

Cases with possible negative histology for colon cancer were 24 adenoma, 9 with negative histology and 2 adenocarcinoma in situ. Of these cases, there was evidence of metastases in three cases within the same medical chart whereas nine cases had a second lesion that missed the histological documentation; evaluation of subsequent medical charts showed that there was evidence of

colon cancer in eight cases and of rectal cancer in two cases based on histological documentation. Detailed descriptions for each of the administrative databases are displayed in table 3.

Sensitivity analysis showed that the nine missing medical charts in the NA administrative database reduced specificity from 80% to 74% (95% CI 66% to 82%), however, with no statistical difference.

Rectal cancer

The κ statistics between evaluators was equal to or higher than 0.90 across the three operative units.

Table 1 Characteristics of patients with colorectal cancer who were identified in the three administrative healthcare databases

Characteristics	Unit 1 (Umbria)	Unit 2 (ASL Napoli 3 Sud)	Unit 3 (Friuli Venezia Giulia)
Invasive colon carcinoma			
Incident cases (N medical chart reviewed)	128	121	130
Overall number of colon lesions	161	143	144
Subjects with more than one lesion (%)	33 (26)	22 (18)	14 (11)
ICD-9 code, N (%)			
153.0	9 (7)	7 (6)	1 (1)
153.1	11 (9)	5 (4)	11 (8)
153.2	14 (11)	19 (16)	19 (15)
153.3	26 (20)	40 (33)	35 (27)
153.4	11 (9)	4 (3)	18 (14)
153.5	1 (1)	1 (1)	–
153.6	44 (34)	25 (21)	36 (28)
153.7	3 (2)	2 (2)	–
153.8	1 (1)	2 (2)	1 (1)
153.9	8 (6)	16 (13)	9 (7)
Admission to department, N (%)			
Medical	28 (22)	27 (22)	24 (18)
Surgical	102 (78)	94 (78)	106 (82)
Sex, N (%)			
Male	71 (55)	69 (57)	62 (48)
Age, years, N (%)			
<40	1 (1)	3 (3)	1 (1)
40–59	14 (11)	5 (4)	21 (16)
≥60	115 (88)	113 (93)	108 (83)
Instrumental diagnosis, N (%)			
Colonoscopy	94 (73)	93 (77)	24 (18)
Abdomen ultrasound	37 (29)	13 (11)	14 (11)
CT scan (including abdomen)	85 (66)	86 (71)	63 (48)
MRI (including abdomen)	6 (5)	3 (2)	–
Surgical procedures, N (%)			
Hemicolectomy	66 (52)	53 (44)	65 (50)
Other surgical excisions	18 (14)	25 (21)	35 (27)
Histological documentation, N (%)			
Biopsy	66 (52)	57 (47)	33 (25)
Resection specimens (after surgical intervention)	89 (70)	80 (66)	102 (78)
Invasive rectal carcinoma			
Incident cases (N medical charts reviewed)	128	119	129
Overall number of rectal lesions	161	137	141
Subjects with more than one lesion (%)	33 (25)	18 (15)	12 (9)
ICD-9 code, N (%)			
154.0	41 (33)	48 (40)	39 (30)
154.1	87 (67)	61 (51)	84 (65)

Continued

Table 1 Continued

Characteristics	Unit 1 (Umbria)	Unit 2 (ASL Napoli 3 Sud)	Unit 3 (Friuli Venezia Giulia)
154.8	2 (2)	10 (8)	6 (5)
Admission to department, N (%)			
Medical	18 (14)	25 (22)	23 (18)
Surgical	112 (86)	94 (78)	106 (82)
Sex			
Male	86 (66)	68 (57)	64 (50)
Age, years, N (%)			
<40	–	–	1 (1)
40–59	15 (12)	26 (22)	19 (15)
≥60	115 (88)	93 (78)	109 (84)
Instrumental diagnosis, N (%)			
Colonoscopy/sigmoidoscopy	97 (75)	105 (88)	32 (25)
Ultrasound	38 (29)	24 (20)	12 (9)
CT scan (including abdomen)	85 (65)	85 (71)	55 (43)
MRI (including abdomen)	15 (12)	18 (15)	12 (9)
Surgical procedures, N (%)			
Anterior resection	54 (42)	27 (23)	22 (17)
Rectal resection	15 (12)	20 (17)	25 (19)
Other	21 (16)	34 (29)	55 (43)
Histological documentation, N (%)			
Needle biopsy	66 (51)	67 (56)	32 (25)
Resection specimens (after surgical intervention)	88 (68)	69 (58)	79 (61)

ICD-9, International Classification of Diseases, Ninth Revision.

After excluding prevalent cases, the incident cases for rectal cancer were 890 for Umbria, 692 for NA and 567 for FVG. From these incident cohorts 130 cases were randomly identified but 2, 11 and 1 medical charts were not available respectively from each operative unit (figure 1). For the non-cases, each unit randomly selected 94 medical charts. Two medical charts of non-cases from Umbria were missing.

The mean age ranged between 69 years (NA) and 72 years (Umbria and FVG). Most of the patients with

a diagnosis of rectal cancer were identified in surgical departments (from 78% to 86%). The most common ICD-9-CM subgroup was rectal cancer (154.1), 67% in Umbria, 51% in NA and 65% in FVG. The most frequent type of surgical intervention was anterior resection (54%) in Umbria, and other types of surgical interventions in NA (34%) and FVG (55%). Complete descriptions of basic characteristics of the cases are displayed in table 1.

Across the three operative units, from the medical charts of 378 cases, 439 lesions were identified. Of 307

Table 2 Cross tabulation of the index test (ICD-9-CM code) results from the results of the reference standard (medical chart)

Type of cancer (ICD-9-CM)	Operative unit	TP	FP	TN	FN
Colon cancer (153.x)	Unit 1 (Umbria)	103	25	91	1
	Unit 2 (ASL Napoli 3 Sud)	98	23	92	2
	Unit 3 (Friuli Venezia Giulia)	104	26	92	2
Rectal cancer (154.0, 154.1, 154.8)	Unit 1 (Umbria)	108	20	92	0
	Unit 2 (ASL Napoli 3 Sud)	95	24	94	0
	Unit 3 (Friuli Venezia Giulia)	104	25	92	2

ICD-9-CM, International Classification of Diseases, Ninth Revision – Clinical Modification; ASL, Azienda Sanitaria Locale; TP, True positive; FP, False positive; TN, True negative; FN, False negative.

Table 3 Colon cancer: reason for incorrect identification of cases and non-cases

Invasive colon cancer			
Type of misclassification	Umbria	ASL Napoli 3 Sud	Friuli Venezia Giulia
False positives			
1 Missing histological examination	13*	15†	11‡
2 Possible negative histology	12§	8¶	15**
a) Adenoma (from biopsy specimen)	2	3	7
b) Adenoma (from surgical specimen)	7	1	4
c) Negative	3	3	3
d) Adenocarcinoma in situ	–	1	1
Total	25	23	26
False negatives			
1 Possible colon cancer relapse	1	–	–
2 Metastatic colon cancer	–	2	–
3 Unclear/histological exam missing	–	–	2
Total	1	2	2

*Metastatic lesions from instrumental exam (n=3); positive for colon adenocarcinoma from histological documentation in a subsequent admission (n=3); positive for rectal adenocarcinoma from histological documentation in a subsequent admission (n=1); deceased (n=1).

†Metastatic lesions from instrumental exam (n=4); chemotherapy or radiotherapy (n=1); previous colon cancer diagnosis (n=1); biopsy not performed (patient with cachexia) (n=1); positive for colon adenocarcinoma from histological documentation in a subsequent admission (n=4).

‡Positive for colon adenocarcinoma from histological documentation in a subsequent admission (n=6); metastatic lesions from instrumental exam (n=1).

§Histological documentation missing for the second lesion (n=4).

¶Histological documentation missing for the second lesion (n=5); positive for colon adenocarcinoma from histological documentation in a subsequent admission (n=1); metastatic lesions from instrumental exam (n=2).

**Metastatic lesions from instrumental exam (n=1); positive for colon adenocarcinoma from histological documentation in a subsequent admission (n=7); positive for rectal adenocarcinoma from histological documentation in a subsequent admission (n=2).

positive cases with malignant carcinoma (true positives) 57 also had a benign tumour; whereas of the 69 false positives 6 had a second lesion that resulted negative or with a benign tumour.

The accuracy estimates were similar across the three units. PPV was 84% for Umbria (95% CI 77% to 90%), 80% for NA (95% CI 72% to 87%) and 81% for FVG (95% CI 73% to 87%); NPV was 100% for Umbria (95% CI 96% to 100%), 100% for NA (95% CI 96% to 100%) and 98% for FVG (95% CI 93% to 100%).

The sensitivity of rectal cancer was 100% for Umbria (95% CI 97% to 100%), 100% for NA (95% CI 96% to 100%) and 98% (95% CI 93% to 100%) for FVG. The specificity was 82% (95% CI 74% to 89%) for Umbria, 80% (95% CI 71% to 87%) for NA and 79% (95% CI 70% to 86%) for FVG. Table 2 provides cross tabulation of the ICD-9-CM code results from the results of the medical charts.

Description of misclassifications

The number of false positives that may cause misclassification in the three operative units were 39 cases with missing histological documentation and 30 cases with possible negative histology for rectal cancer. Cases with missing histological documentation had evidence of, within the same medical chart, metastases (n=6), chemotherapy or radiotherapy (n=4), or biopsy not performed

due to inoperability or death (n=2); in the subsequent medical chart review, three cases had evidence of rectal cancer on histological documentation (table 4).

Cases with possible negative histology for rectal cancer had 23 adenomas, 5 were with negative histology and 2 were adenocarcinoma in situ. Of these cases, within the same medical chart, there was evidence of metastases (n=2), metastasis+chemotherapy (n=2), chemotherapy or radiotherapy (n=1) whereas four cases had a second lesion that missed the histological documentation; evaluation of subsequent medical charts showed that there was evidence of rectal cancer in nine cases and of colon cancer in four cases based on histological documentation. Table 4 provides detailed description for false positives and false negatives for each of the administrative databases.

Sensitivity analysis based on the worst-case scenario did not show any statistical difference when missing data were considered false negative or false positive, although in the NA administrative database the specificity was reduced from 80% to 73% (95% CI 64% to 80%) due to the 11 missing medical charts of the cases.

DISCUSSION

Case definition of diseases is important when validating administrative databases since it may influence the sensitivity or PPV.^{19 23 24} In our study the ascertainment of cases

Table 4 Rectal cancer: reason for incorrect identification of cases and non-cases

Invasive rectal cancer			
Type of misclassification	Umbria	ASL Napoli 3 Sud	Friuli Venezia Giulia
False positives			
1 Missing histological examination	14*	17†	8‡
2 Possible negative histology	6§	7¶	17**
a) Adenoma (from biopsy specimen)	4	0	11
b) Adenoma (from surgical specimen)	2	3	3
c) Negative	0	2	3
d) Adenocarcinoma in situ	0	2	0
Total	20	24	25
False negative			
1 Possible rectal cancer relapse	0	0	1
2 Possible rectal cancer metastasis	0	0	0
3 Unclear/histological exam missing	0	0	1
Total	–	–	2

*Metastatic lesions from instrumental exam (n=3); deceased (n=1); chemotherapy or radiotherapy (n=1); inoperable (n=1).

†Metastatic lesions from instrumental exam (n=3); chemotherapy or radiotherapy (n=3); positive for rectal adenocarcinoma from histological documentation in a subsequent admission (n=3).

‡Histological documentation missing for two lesions (one patient).

§Positive for rectal adenocarcinoma from histological documentation in a subsequent admission (n=1); metastasis +chemotherapy (n=2).

¶Histological documentation missing for the second lesion (n=3); positive for colon adenocarcinoma from histological documentation in a subsequent admission (n=2).

**Metastatic lesions from instrumental exam (n=2); positive for rectal adenocarcinoma from histological documentation in a subsequent admission (n=8); positive for colon adenocarcinoma from histological documentation in a subsequent admission (n=2); histological documentation missing for the second lesion (n=1); chemotherapy or radiotherapy (n=1).

for the validation of an ICD-9 code within an administrative database was based on the presence of a primary lesion in the colon or rectum confirmed by histological documentation of cancer from a primary or metastatic site. The performance of the ICD-9 codes related to colorectal cancer based on the same case definition was evaluated in the three administrative databases by consulting medical charts that were our reference standard. Results showed that ICD-9-CM codes for colon cancer (153.x) and rectal cancer (154.0, 154.1 and 154.8), based on the 'case definition', performed well in terms of sensitivity across the three databases. False positive rates influenced specificity and PPVs and this may be due to our stringent criteria that the two elements of our case definition had to be present in the first medical chart. We chose to report all diagnostic accuracy though we decided to select non-cases from an oncological population because we aimed to select a population similar to that of the cases, except having the neoplasm of interest (colon and rectal cancer). Although this can be a limitation with regard to the accuracy measures of sensitivity and specificity, our overall results also comprise PPVs which are based exclusively on the cases indicating the ability of the administrative database to identify correctly the subjects with disease according to our case definition that varied between 80% and 84%.

In some of the cases, the false positives could be explained by the absence of histological documentation

in the first medical chart. This does not necessarily mean that patients classified as false positive cases did not have colorectal cancer, since there were several indicators that can prove the presence of malignant cancer. These include the confirmation of the malignant disease in other sources such as subsequent medical charts, the administration of chemotherapy or radiation therapy, the presence of metastases. Apart from 3 cases that did not perform biopsy due to cachexia or death, there were 15/755 (2%) with negative histology and 4 (0.5%) adenocarcinoma in situ that resulted in important misclassifications (tables 3 and 4).

During our data extraction, we found that cases with at least two colorectal lesions were identified, varying between 10% to 25% across the three cohorts of subjects. During the validation process, a subject with two lesions, one benign and the other malignant, was classified as a true positive whereas another with two lesions, one benign and the second with missing histological documentation, was classified as false positive.

Synchronous colorectal neoplasms, that is two or more primary tumours identified in the same patient and at the same time, have been described in the medical literature with a rate of 33%.²⁵ Researchers that aim to validate colorectal cancer ICD-9 codes will need to make a thorough evaluation of the number of lesions and their respective instrumental and histological documentation.

Table 5 Tables and accuracy measures for different case definition algorithms**Colorectal cancer**

Operative unit	Algorithm	TP	FP	TN	FN	Sensitivity	Specificity	PPV
Umbria	Case definition 1: lesion and histology	211	45	183	1	100	80	82
	Case definition 2: case definition 1+chemotherapy or radiotherapy	214	42	183	1	100	81	84
	Case definition 3: case definition 2+metastasis	220	36	183	1	100	84	86
	Case definition 4: case definition 3+subsequent medical chart	225	33	183	1	100	85	87
ASL Napoli 3 Sud	Case definition 1: lesion and histology	193	47	186	2	99	80	80
	Case definition 2: case definition 1+chemotherapy or radiotherapy	197	43	186	2	99	81	82
	Case definition 3: case definition 2+metastasis	206	34	186	2	99	85	86
	Case definition 4: case definition 3+subsequent medical chart	217	23	186	2	99	89	90
Friuli Venezia Giulia	Case definition 1: lesion and histology	208	51	184	4	98	78	80
	Case definition 2: case definition 1+chemotherapy or radiotherapy	209	50	184	4	98	79	81
	Case definition 3: case definition 2+metastasis	213	46	184	4	98	80	82
	Case definition 4: case definition 3+subsequent medical chart	238	21	184	4	98	90	92

PPV, positive predictive value; ASL, Azienda Sanitaria Locale; TP, True positive; FP, False positive ; TN, True negative; FN, False negative.

In a post hoc analysis we re-evaluated different combinations of algorithms by adding other elements such as radiotherapy, chemotherapy, metastasis and histological documentation from subsequent medical charts and found that the PPVs increased across the three operative units as shown in [table 5](#).

Comparison of accuracy results with other settings

Another Italian study evaluated the accuracy of colorectal cancer ICD-9 codes using hospital administrative databases in Piedmont province and found a combined sensitivity for colorectal cancer of 72.4% but with a higher PPV (88%).²⁶ While their PPV was higher than our findings, the sensitivity was much lower than ours. These discrepancies could be due to the methodological approaches

that differed between our study and the Piedmont study. The Piedmont study used the cancer registry as a reference standard, and the population of interest was selected based on an algorithm which was based on a combination of ICD-9 related to malignant neoplasm of the colon, rectum and rectosigmoid junction in the primary position as well as in the secondary position and any ICD-9-CM procedure code leading to surgical diagnosis-related group payment. Another potential reason for discrepancy between the Piedmont study and ours may be due to the fact that we limited our target population to those who were incident cases but in the primary position. In the Piedmont study, the authors also performed a sensitivity analysis by limiting the analysis only to those

in the primary position but they did not report the data and concluded simply that the analysis did not lead to any gain in PPV which seems to be their primary objective. Another potential explanation could be that the authors did not consult any medical chart and no case definition was elaborated against which to test the presence of the disease.

Two other research groups have evaluated the accuracy of colon or rectal cancer diagnosis in administrative databases in other settings.^{15 27} In Denmark, Helqvist *et al*²⁷ evaluated the validity of ICD-10 colorectal cancer (C18 for colon cancer, C19 for cancer in the colorectal junction and C20 for rectal cancer) coding in the Danish National Registry of patients, using the Danish Cancer Registry as a reference standard. The overall accuracy of the colorectal cancer codes was 89% in terms of PPV that was defined as the number of patients with a colorectal diagnosis in the Danish National Registry of Patients (DNRP) and the Danish Cancer Registry (DCR) (numerator), divided by the number of all patients with a colorectal cancer diagnosis registered in the DNRP (denominator). However, the study did not provide any case definition statement. In France, Quantin *et al*¹⁵ developed two algorithms to validate the ICD-10 codes related to colorectal cancer in an administrative database using a cancer registry as a reference standard. The first algorithm, based only on diagnostic and procedure codes, provided good sensitivity and a PPV lower than ours (75%) while the second algorithm, that considered the past history of the patient, overestimated the number of incident cases by almost 50%.¹⁵ Both studies differ from ours in terms of the reference standard used, index test used (medical charts vs cancer registries) and location of the diagnosis (primary or any position).

A systematic review of administrative databases that validated colorectal cancer worldwide is currently being completed and will provide a complete account of validation of administrative databases.¹

Strengths and limitations

The main strength of our study is that we used medical chart review as a gold standard and the presence of histological documentation in addition to an imaging or endoscopic presence of a primary lesion as a requirement for validation. In contrast to other studies, we separately validated codes related to colon cancer from the codes related to rectal cancer. In addition, our study assessment was based on a prepublished protocol and we can state that there was no deviation from protocol. We used detailed and explicit eligibility criteria, a duplicate and independent process for medical chart review and data abstraction following recommended guidelines based on the criteria published by the STARD initiative for the accurate reporting of investigations of diagnostic studies.^{20 21 28}

Our assessment was limited to the diagnosis of colorectal cancers in the primary position and this might underestimate future epidemiological incidence of cancer. We are unsure whether the obtained accuracy results can be

generalised to new cases of cancer in patients who were diagnosed in day hospital or day surgery facilities. Further research is needed to address the validity of ICD-9 codes in outpatient settings.

Another limitation of our study is that there were missing charts with respect to the estimated sample size. However, the number of missing charts was very low for Umbria and FVG (ranged from 0% and 2.1%) and quite low for ASL Napoli 3 Sud (6.9% charts missing for colon cancer and 8.5% for rectal cancer). In general, a study population lower than the estimated sample size leads to the same diagnostic accuracy estimates but with broader CIs. Nevertheless, to be more conservative, we also decided to present a 'worst case' scenario in which the missing charts were considered as false positives.

A potential limitation in our assessment could be the choice of a non-case population that was arbitrary. We chose to select non-cases from an oncological population because we aimed to select a population similar to that of the cases, except having the neoplasm of interest (colon and rectal cancer). In our opinion, using this approach there was a chance of finding false negatives. Choosing the non-cases in other ways, for example, from patients with other types of diseases, the chances of finding false negatives would have been very low, which was not our primary concern.

However, although the sensitivity and specificity are influenced by the choice of non-cases, PPV is based only on the cases, and its value represents the ability of the administrative database to identify correctly the subjects with disease.

A possible limitation related to the implications of our results for future research is that validation studies of administrative databases are related to the context in which they are generated and are not generalisable to other settings.

CONCLUSION

The present study concerns two regional and local areas in Italy and shows that administrative healthcare databases from Umbria, NA and FVG can be used to identify hospitalised subjects with colon and rectal cancers. We proposed a simple case definition for case ascertainment within colorectal cancer and the obtained accuracy is acceptable. The present study will add value to the knowledge of the colorectal cancer diseases given that it covers different areas of Italy and can contribute to improving the cancer treatment patterns, although the presented results may not be generalisable in other settings.

Author affiliations

¹Health Planning Service, Regional Health Authority of Umbria, Perugia, Italy

²Cancer Epidemiology Unit, Centro di Riferimento Oncologico Aviano, Aviano, Italy

³Centro Regionale Sangue, Azienda Ospedaliera di Perugia, Perugia, Italy

⁴Registro Tumori Regione Campania, ASL NA 3 Sud, Brusciano, Italy

⁵Health ICT Service, Regional Health Authority of Umbria, Perugia, Italy

⁶SOC Epidemiologia Oncologica, Centro di Riferimento Oncologico, Aviano, Italy

⁷Department of Digestive Surgery and Liver Unit, University of Perugia, Perugia, Italy

^aDirezione Regionale Salute, Regional Health Authority of Umbria, Perugia, Italy

Collaborators Giuliana Alessandrini, David Franchini, Michele Gobbato, Fabrizio Stracci, Rita Chiari, Chiara Grisci

Contributors AM, IA, MF and DS conceived the original idea of the study. IA, DS, AM, MF, EB, GG, FC, MO and WO designed the study. PCa, MDG, AG, MFV, PCo and VC identified the cohort using administrative database with the supervision of WO, EB, DS, MF, AM and FS. IA, FC, MO, AG, PCo, VC, MFV, undertook the data abstraction with the supervision of AM, GG, WO, FS, MF, EB and DS. IA, AM, DS and RC performed case ascertainment. IA, AM, FC, EB, MF and MO performed the analysis. GG, PCa, MDG, PCo, AG, VC, MFV, RC, WO helped in the interpretation of the data. The initial draft of the manuscript was prepared by IA, AM, FC, MF, EB and RC. GG, PCa, MO, AG, MDG, PCo, VC, MFV and WO critically revised the manuscript for important intellectual content. All the authors read and approved the final manuscript. AM, MF and EB are the guarantors of the data for the respective operative units.

Funding This study was developed within the D.I.V.O. project (Realizzazione di un Database Interregionale Validato per l'Oncologia quale strumento di valutazione di impatto e di appropriatezza delle attività di prevenzione primaria e secondaria in ambito oncologico) supported by funding from the National Centre for Disease Prevention and Control (CCM 2014), Ministry of Health, Italy. The study funder was not involved in the study design or the writing of the protocol.

Competing interests None declared.

Patient consent Not required.

Ethics approval Regional Ethics Committee of Umbria (CEAS), authorisation number: 2656/15 (04/11/2015).

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement No additional data are available.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>

© Article author(s) (or their employer(s) unless otherwise stated in the text of the article) 2018. All rights reserved. No commercial use is permitted unless otherwise expressly granted.

REFERENCES

1. Abraha I, Giovannini G, Serraino D, *et al.* Validity of breast, lung and colorectal cancer diagnoses in administrative databases: a systematic review protocol. *BMJ Open* 2016;6:e010409.
2. Abraha I, Montedori A, Eusebi P, *et al.* The Current State of Validation of Administrative Healthcare Databases in Italy: A Systematic Review. *Pharmacoepidemiology and Drug Safety* 2012;21:400–00.
3. World Health Organization. *International statistical classification of diseases and health related problems*. 10th revision. Geneva: WHO, 1992.
4. Cozzolino F, Abraha I, Orso M, *et al.* Protocol for validating cardiovascular and cerebrovascular ICD-9-CM codes in healthcare administrative databases: the Umbria Data Value Project. *BMJ Open* 2017;7:e013785.
5. Montedori A, Abraha I, Chiatti C, *et al.* Validity of peptic ulcer disease and upper gastrointestinal bleeding diagnoses in administrative databases: a systematic review protocol. *BMJ Open* 2016;6:e011776.
6. Rimland JM, Abraha I, Luchetta ML, *et al.* Validation of chronic obstructive pulmonary disease (COPD) diagnoses in healthcare databases: a systematic review protocol. *BMJ Open* 2016;6:e011777.
7. Rawson NSB, Shatin D. Assessing the validity of diagnostic data in large administrative healthcare utilization databases. In: Hartzema A, Tilson H, Chan K, eds. *Pharmacoepidemiology and Therapeutic Risk Management*: Harvey Whitney Books, 2008.
8. Ferlay J, Soerjomataram I, Dikshit R, *et al.* Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer* 2015;136:E359–E386.
9. Brenner H, Kloor M, Pox CP. Colorectal cancer. *Lancet* 2014;383:1490–502.
10. Pucciarelli S, Gagliardi G, Maretto I, *et al.* Long-term oncologic results and complications after preoperative chemoradiotherapy for rectal cancer: a single-institution experience after a median follow-up of 95 months. *Ann Surg Oncol* 2009;16:893–9.
11. Marmorino F, Salvatore L, Barbara C, *et al.* Serum LDH predicts benefit from bevacizumab beyond progression in metastatic colorectal cancer. *Br J Cancer* 2017;116:318–23.
12. Bousquet PJ, Caillet P, Coeuret-Pellicier M, *et al.* [Using cancer case identification algorithms in medico-administrative databases: Literature review and first results from the REDSIAM Tumors group based on breast, colon, and lung cancer]. *Rev Epidemiol Sante Publique* 2017;65 Suppl 4(Suppl 4):S236–s42.
13. Di Costanzo F, Ravasio R, Sobrero A, *et al.* Capecitabine versus bolus fluorouracil plus leucovorin (folinic acid) as adjuvant chemotherapy for patients with Dukes' C colon cancer: economic evaluation in an Italian NHS setting. *Clin Drug Investig* 2008;28:645–55.
14. Deshpande AD, Schootman M, Mayer A. Development of a claims-based algorithm to identify colorectal cancer recurrence. *Ann Epidemiol* 2015;25:297–300.
15. Quantin C, Benzenine E, Hägi M, *et al.* Estimation of national colorectal-cancer incidence using claims databases. *J Cancer Epidemiol* 2012;2012:1–7.
16. Dehal A, Abbas A, Johna S. Comorbidity and outcomes after surgery among women with breast cancer: analysis of nationwide in-patient sample database. *Breast Cancer Res Treat* 2013;139:469–76.
17. Kinski A. Clinical and economic outcomes analyses of women developing breast cancer in a managed care organization. *Am J Clin Oncol* 2005;28:51–7.
18. Mittmann N, Liu N, Porter J, *et al.* Utilization and costs of home care for patients with colorectal cancer: a population-based study. *CMAJ Open* 2014;2:E11–E17.
19. Abraha I, Serraino D, Giovannini G, *et al.* Validity of ICD-9-CM codes for breast, lung and colorectal cancers in three Italian administrative healthcare databases: a diagnostic accuracy study protocol. *BMJ Open* 2016;6:e010547.
20. Benchimol EI, Manuel DG, To T, *et al.* Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. *J Clin Epidemiol* 2011;64:821–9.
21. De Coster C, Quan H, Finlayson A, *et al.* Identifying priorities in methodological research using ICD-9-CM and ICD-10 administrative data: report from an international consortium. *BMC Health Serv Res* 2006;6:77.
22. Bossuyt PM, Reitsma JB, Bruns DE, *et al.* Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ* 2003;326:41–4.
23. Montedori A, Bidoli E, Serraino D, *et al.* Accuracy of lung cancer ICD-9-CM codes in Umbria, Napoli 3 Sud and Friuli Venezia Giulia administrative healthcare databases: a diagnostic accuracy study. *BMJ Open* 2018;8:e020628.
24. Orso M, Serraino D, Abraha I, *et al.* Validating malignant melanoma ICD-9-CM codes in Umbria, ASL Napoli 3 Sud and Friuli Venezia Giulia administrative healthcare databases: a diagnostic accuracy study. *BMJ Open* 2018;8:e020631.
25. AIRT Working Group. Italian cancer figures--report 2006: 1. Incidence, mortality and estimates. *Epidemiol Prev* 2006;30(1 Suppl 2):8–10.
26. Baldi I, Vicari P, Di Cuonzo D, *et al.* A high positive predictive value algorithm using hospital administrative data identified incident cancer cases. *J Clin Epidemiol* 2008;61:373–9.
27. Helqvist L, Erichsen R, Gammelager H, *et al.* Quality of ICD-10 colorectal cancer diagnosis codes in the Danish National Registry of Patients. *Eur J Cancer Care* 2012;21:722–7.
28. Bossuyt PM, Reitsma JB, Bruns DE, *et al.* Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. *Ann Intern Med* 2003;138:40–4.