ORIGINAL RESEARCH

Validity of data in the Danish Colorectal Cancer Screening Database

Mette Kielsholm Thomsen[1]
Sisse Helle Njor[1]
Morten Rasmussen[2]
Dorte Linnemann[3]
Berit Andersen[4]
Gunnar Baatrup[5,6]
Lennart Jan Friis-Hansen[7]
Jens Christian Riis Jørgensen[8]
Ellen Margrethe Mikkelsen[1]

[1]Department of Clinical Epidemiology, Aarhus University Hospital, Aarhus, [2]Department of Digestive Diseases K, Bispebjerg Hospital, Copenhagen, [3]Department of Pathology, Herlev and Gentofte Hospital, Herlev, [4]Department of Public Health Programs, Randers Regional Hospital, Randers, [5]Department of Surgery, Odense University Hospital, [6]Department of Clinical Science, University of Southern Denmark, Odense, [7]Center for Genomic Medicine, Rigshospitalet, University of Copenhagen, Copenhagen, [8]Department of Colorectal Cancer Surgery, Vejle Hospital, Vejle, Denmark

Correspondence: Mette Kielsholm Thomsen; Ellen Margrethe Mikkelsen
Department of Clinical Epidemiology, Aarhus University Hospital, Olof Palmes Allé 43-45, DK-8200 Aarhus N, Aarhus, Denmark
Tel +45 8716 7212; +45 8716 8207
Email mko@clin.au.dk; em@clin.au.dk

Background: In Denmark, a nationwide screening program for colorectal cancer was implemented in March 2014. Along with this, a clinical database for program monitoring and research purposes was established.

Objective: The aim of this study was to estimate the agreement and validity of diagnosis and procedure codes in the Danish Colorectal Cancer Screening Database (DCCSD).

Methods: All individuals with a positive immunochemical fecal occult blood test (iFOBT) result who were invited to screening in the first 3 months since program initiation were identified. From these, a sample of 150 individuals was selected using stratified random sampling by age, gender and region of residence. Data from the DCCSD were compared with data from hospital records, which were used as the reference. Agreement, sensitivity, specificity and positive and negative predictive values were estimated for categories of codes "clean colon", "colonoscopy performed", "overall completeness of colonoscopy", "incomplete colonoscopy", "polypectomy", "tumor tissue left behind", "number of polyps", "lost polyps", "risk group of polyps" and "colorectal cancer and polyps/benign tumor".

Results: Hospital records were available for 136 individuals. Agreement was highest for "colorectal cancer" (97.1%) and lowest for "lost polyps" (88.2%). Sensitivity varied between moderate and high, with 60.0% for "incomplete colonoscopy" and 98.5% for "colonoscopy performed". Specificity was 92.7% or above, except for the categories "colonoscopy performed" and "overall completeness of colonoscopy", where the specificity was low; however, the estimates were imprecise.

Conclusion: A high level of agreement between categories of codes in DCCSD and hospital records indicates that DCCSD reflects the hospital records well. Further, the validity of the categories of codes varied from moderate to high. Thus, the DCCSD may be a valuable data source for future research on colorectal cancer screening.

Keywords: colorectal cancer screening, clinical database, data validity

Introduction

In Denmark, a nationwide screening program for colorectal cancer was implemented in March 2014. The program is administered by the five Danish regions, which are administrative units responsible for health care. Screening is offered biennial and free of charge to all citizens aged 50–74 years.[1] The screening procedure is based on a single-sample immunochemical fecal occult blood test (iFOBT), which has been documented to detect invisible amounts of blood in stool samples, associated with bleeding lesions related to precancerous adenomas or bowel cancer, at early

stages of the disease.[2,3] In the case of a positive test result (>100 ng/mL), a colonoscopy is performed,[4] and if relevant, further treatment is provided.

In addition to the screening program, a clinical quality database, the Danish Colorectal Cancer Screening Database (DCCSD), was established to monitor the quality of the screening program. The database comprises data from existing registries: the Invitation and Administration Module (IAM) for the screening program, the Danish National Patient Registry (DNPR) and the National Pathology Registry. Thus, all data are provided electronically and no data are entered manually in the DCCSD. With the screening program, a number of new codes were introduced for diagnoses and procedures performed within the screening program, mainly for data reported to the DNPR.

The primary aim of the DCCSD is to monitor the quality of the screening program and the secondary aim is to provide data for research purposes. To fulfill both aims, data completeness and validity must be high. As the screening program is newly established and the medical staffs have to report novel codes for procedures and diagnoses to the DNPR, the validity of data in the DCCSD is unknown. Therefore, the aim of this study was to evaluate the agreement and validity of selected categories of procedure codes "clean colon", "colonoscopy performed", "overall completeness of colonoscopy", "incomplete colonoscopy", "polypectomy", "tumor tissue left behind", "number of polyps", "lost polyps" and "risk group of polyps" and diagnosis codes "colorectal cancer and polyps/benign tumor", comparing DCCSD data with hospital records.

Methods
Study population and setting

The screening program is implemented over a period of 4 years (2014–2017). Home sampling kits are delivered by mail along with instructions and an invitation letter. The target population is invited according to a randomly assigned sequence of birth months, although citizens, who turn 50 or 74 years old within the initial 4-year screening round, have to receive their first invitation no later than 1 month before that particular birthday.

This study was based on 111,810 citizens invited to colorectal cancer screening, via the IAM, between March 3, and May 31, 2014. In this period, a pilot report of the DCCSD estimated that 58% participated by returning for the iFOBT within 3 months and 6.9% of the analyzed samples were positive.[5]

Individuals with negative or faulty screening tests were not registered in the DNPR with any of the diagnosis and

procedure codes we aimed to validate, and therefore we only included participants with a positive iFOBT result. Thus, 4,704 individuals with a positive test result were eligible for the study. From these, we selected a random sample of 150 individuals using simple random sampling within strata of gender, age and region of residence. The sample consisted of 15 men and 15 women from each of the five Danish Regions, while also ensuring a distribution of 10 individuals from the youngest age group (50–59 years) and 20 from the oldest (60–74 years) in each region.

Data

The validated data include a large number of separate codes constituting in total 11 categories of procedures and diagnoses related to colorectal cancer screening (Table 1). Data are linked by a personal registration number (Central Person Registry [CPR] number), which is a unique 10-digit number assigned to each Danish resident at birth or upon immigration.[6] The DCCSD was accepted as a clinical quality database in October 2014 by the Danish Health Data Authority (14/23440) and approved by the Danish Data Protection Agency (2007-58-0014); thus, the study complied with Danish regulations.

The Danish regions' IAM

The IAM was established in addition to the screening program and handles invitations as well as response letters to all residents included in the program. Each region can adjust the invitation rate in order to accommodate regional capacity (especially during the initial implementation).[7] The IAM is

Table 1 Categories of codes for diagnoses and procedures included in the validation of the DCCSD

Category	Code
Clean colon	AFX02C, AFX02D
Colorectal cancer	DC180, DC182, DC183, DC184, DC185, DC186, DC187, DC189, DC209
Polyps/benign tumors	DD120, DD122, DD123A, DD123, DD123B, DD124, DD125, DD126C, DD128
Colonoscopy performed	KUJF32, KUJF35
Overall completeness of colonoscopy	ZPY1A0, ZPY1A1, ZPY1A10, ZPY1A11, ZPY1A12, ZPY1A13, ZPY1A14, ZPY1A15, ZPY1A18
Incomplete colonoscopy	ZPY1A1, ZPY1A10, ZPY1A11, ZPY1A12, ZPY1A13, ZPY1A14, ZPY1A15, ZPY1A18
Polypectomy	KJFA15, KJFA55A, KJFA55B, KJGA05, KJGA52A, KJGA52B
Tumor tissue left behind	ZPY1B01, ZPY1B02
Number of polyps seen	ZPY1Cnn, ZPY1C99
Lost polyps	ZPY1D00, ZPY1Dnn, ZPY1D99
Risk group of polyps	ZPY1E01, ZPY1E02, ZPY1E03

Abbreviation: DCCSD, Danish Colorectal Cancer Screening Database.

updated daily with information on vital status, addresses, etc from the Civil Registration System.[8,9]

From the IAM, we retrieved the CPR numbers and data of all patients eligible for the study population, by November 2014.

The DNPR

The DNPR covers all somatic admissions to Danish hospitals since 1977 and outpatient contacts since 1995. Information in the DNPR includes CPR number, dates of admission and discharge, as well as codes for diagnoses (the International Classification of Diseases, tenth edition [ICD-10]) and procedures. Data are collected for administrative purposes, as well as for research and quality assurance.[10,11] DNPR data for this study were retrieved on February 10, 2015.

Hospital record review

A review of the hospital record of each study participant was performed according to a standardized protocol by an appointed medical doctor in each region. The reviewers extracted the information on all relevant diagnoses and procedures from the medical files and entered them in a standardized spreadsheet.

Statistics

We used data from the hospital records as the reference and compared these data for each patient with DCCSD data. Agreement was calculated as the percent of cases with the same coding for diagnoses and procedures in the DCCSD as in the hospital records. In addition, we calculated sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV).

Sensitivity is the ability of the DCCSD to identify all true positives, ie, participants with a diagnosis or procedure code for the phenomenon of interest registered in the DCCSD and in the hospital records. Sensitivity was calculated as the proportion of true positives in the DCCSD out of all positives in the hospital records.

Specificity is the ability of the DCCSD to identify all true negatives, ie, participants who are not registered with a diagnosis or procedure code for the phenomenon of interest neither in the DCCSD nor in the hospital records. Specificity was calculated as the proportion of true negatives in the DCCSD out of all negatives in the hospital records.

The PPV, which is the probability that a code in the DCCSD is correctly registered compared to the hospital records, was calculated as the proportion of true positives out of all positives in the DCCSD. The NPV, which is the probability that a code absent in the DCCSD is correctly absent, was calculated as the proportion of true negatives out of all negatives in the DCCSD.[12] Exact 95% confidence intervals (CIs) for the binomial distribution were calculated for each estimate.

First, results are presented for 11 categories of codes (Table 2). As multiple codes can be used, eg, to indicate cancer at several locations in the colon, we examined whether a code included in the specific category was recorded – not whether it was the correct code. Second, we estimated the validity of each code separately (Table 3). Thus, the validity of the codes in the DCCSD, eg, cancer, was evaluated as both 1) overall colorectal cancer, not distinguishing between the nine separate diagnosis codes (Table 2), and 2) subtypes of cancer, eg, cancer in cecum (DC180; Table 3). For the separate procedure and diagnosis codes, we estimated only the agreement, specificity and NPV as the numbers were too sparse to estimate meaningful sensitivity and PPV. For completeness of colonoscopy, two categories are presented in Table 2, one for all codes on completeness of colonoscopy "overall completeness" and one for the codes that indicate an incomplete procedure "incomplete colonoscopy". For "number of polyps seen" and "lost polyps", "nn" in the codes ZPY1Cnn and ZPY1Dnn were replaced by numbers representing the number of polyps seen and the size of polyps lost, respectively. If the number in the DCCSD was incorrect, the code was treated as missing in the calculations of validity of the specific codes.

Results

Of the 150 individuals selected for this validation study, hospital records were available for review for 136 (91%). Records for seven patients were not relevant as no colonoscopy was performed, and seven records were not accessible (Figure 1). For the overall category "clean colon" at both colonoscopy and computed tomography (CT) colonography, we found an agreement between the DCCSD and the hospital records of 82.4% (95% CI: 74.9–88.4). Sensitivity was 69.0% (95% CI: 56.9–79.5) due to 22 false negatives. PPV was 96.1% (95% CI: 86.5–99.5), so indicating that this code was rarely recorded in the DCCSD if the statement of "clean colon" was not found in the hospital records. NPV was 74.1% (95% CI: 63.5–83.0; Table 2). For the category "colorectal cancer", the agreement between the DCCSD and the hospital records was 97.1% (95% CI: 92.6–99.2), but three false negatives resulted in a sensitivity of 72.7% (95% CI: 39.0–94.0). Specificity, NPV and PPV were 99.2% (95% CI: 95.6–100.0), 97.6% (95% CI: 93.3–99.5) and 88.9% (95% CI: 51.8–99.7), respectively (Table 2). The agreement in registration of individual types of

Table 2 Validity of categories of colorectal cancer screening codes in the DCCSD compared with hospital records, n=136

Categories	Agreement, % (95% CI); a/n	Sensitivity, % (95% CI); tp/(tp+fn)	Specificity, % (95% CI); tn/(tn+fp)	PPV, % (95% CI); tp/(tp+fp)	NPV, % (95% CI); tn/(tn+fn)
Clean colon	82.4 (74.9–88.4); 112/136	69.0 (56.9–79.5); 49/(49+22)	96.9 (89.3–99.6); 63/(63+2)	96.1 (86.5–99.5); 49/(49+2)	74.1 (63.5–83.0); 63/(63+22)
Colorectal cancer	97.1 (92.6–99.2); 132/136	72.7 (39.0–94.0); 8/(8+3)	99.2 (95.6–100.0); 124/(124/1)	88.9 (51.8–99.7); 8/(8+1)	97.6 (93.3–99.5); 124/(124+3)
Polyps/benign tumors	90.4 (84.2–94.8); 123/136	84.7 (73.0–92.8); 50/(50+9)	94.8 (87.2–98.6); 73/(73+4)	92.6 (82.1–97.9); 50/(50+4)	89.0 (80.2–94.9); 73/(73+9)
Colonoscopy performed	96.3 (91.6–98.8); 131/136	98.5 (94.6–99.8); 130/(130+2)	25.0 (0.6–80.6); 1/(1+3)	97.7 (93.5–99.5); 130/(130+3)	33.3 (0.8–90.6); 1/(1+2)
Overall completeness of colonoscopy	92.6 (86.9–96.4); 126/136	93.9 (88.4–97.3); 124/(124+8)	50.0 (6.8–93.2); 2/(2+2)	98.4 (94.4–99.8); 124/(124+2)	20.0 (2.5–55.6); 2/(2+8)
Incomplete colonoscopy	96.3 (91.6–98.8); 131/136	60.0 (26.2–87.8); 6/(6+4)	99.2 (95.7–100.0); 125/(125+1)	85.7 (42.1–99.6); 6/(6+1)	96.9 (92.3–99.1); 125/(125+4)
Polypectomy	95.6 (90.6–98.4); 130/136	92.1 (82.4–97.4); 58/(58+5)	98.6 (92.6–100.0); 72/(72+1)	98.3 (90.9–100.0); 58/(58+1)	93.5 (85.5–97.9); 72/(72+5)
Tumor tissue left behind	87.5 (80.7–92.5); 119/136	75.0 (63.0–84.7); 51/(51+17)	100.0 (94.7–100.0); 68/(68+0)	100.0 (93.0–100.0); 51/(51+0)	80.0 (69.9–87.9); 68/(68+17)
Number of polyps seen	93.4 (87.8–96.9); 127/136	89.4 (79.4–95.6); 59/(59+7)	97.1 (90.1–99.7); 68/(68+2)	96.7 (88.7–99.6); 59/(59+2)	90.7 (81.7–96.2); 68/(68+7)
Lost polyps	88.2 (81.6–93.1); 120/136	81.5 (68.6–90.7); 44/(44+10)	92.7 (84.8–97.3); 76/(76+6)	88.0 (75.7–95.5); 44/(44+6)	88.4 (79.7–94.3); 76/(76+10)
Risk group of polyps	83.8 (76.5–89.6); 114/136	69.6 (55.9–81.2); 39/(39+17)	93.8 (86.0–97.9); 75/(75+5)	88.6 (75.4–96.2); 39/(39+5)	81.5 (72.1–88.9); 75/(75+17)

Notes: a, number in agreement; n, number in total.
Abbreviations: CI, confidence interval; fn, false negatives; fp, false positives; NPV, negative predictive value; PPV, positive predictive value; tn, true negatives; tp, true positives.

cancers ranged between 98.5% and 100%. For the category "colonoscopy performed", the agreement between the DCCSD and the hospital records was 96.3% (95% CI: 91.6–98.8). Sensitivity and PPV were 98.5% (95% CI: 94.6–99.8) and 97.7% (95% CI: 93.5–99.5), respectively (Table 2). For the category "overall completeness of the colonoscopy" (both complete and incomplete), the agreement was 92.6% (95% CI: 86.9–96.4) and sensitivity was 93.9% (95% CI: 88.4–97.3). Only 60% of colonoscopies in the category "incomplete" were reported to the DCCSD (sensitivity, 60%; 95% CI: 26.2–87.8). Specificity was above 99%. Thus, the code for incompleteness was only in one instance reported to the DCCSD, although the colonoscopy was in fact complete (Table 2). For the overall category of "polypectomy", sensitivity and specificity were 92.1% (95% CI: 82.4–97.4) and 98.6% (95% CI: 92.6–100.0), respectively (Table 2). Most polypectomies were of the type "endoscopic polypectomies, large intestine" (KJFA15), which had an agreement of 95.6% (95% CI: 90.6–98.4; Table 3). Overall sensitivity for the category "tumor tissue left behind" (including both "tissue left behind" and "no tissue left behind") was 75% (95% CI: 63.0–84.79; Table 2). For the categories "number of polyps seen" and "lost polyps", sensitivity was 89.4% (95% CI: 79.4–95.6) and 81.5% (95% CI: 68.6–90.7), respectively. Agreement for the specific "number of polyps

seen" (ZPY1Cnn) was 90.4% (95% CI: 84.2–94.8), and for the "number of lost polyps" (ZPY1D00) it was 87.5% (95% CI: 80.7–92.5; Table 3). Information on "risk group of polyps" had an agreement of 83.8% (95% CI: 76.5–89.6), whereas the sensitivity was 69.9% (95% CI: 55.9–81.2) due to 17 false negatives (Table 2).

Overall, the agreement for the individual diagnosis and procedure codes varied from 84% ("clean colon at colonoscopy" [AF02C]) to 100% ("cancer in cecum" [DC180] and "benign tumor in descending colon" [DD124]; Table 3). For most individual codes, the specificity was above 95%; however, for "colonoscopy" (KUJF32) and "complete colonoscopy" (ZPY1A0), four cases were false negatives; thus, the specificity was 75.0% (95% CI: 47.6–92.7) and 71.4% (95% CI: 41.9–91.6), respectively. The NPVs for the majority of the individual codes were 92% or above; however, for the codes "clean colon at colonoscopy" (AFX02C), "colonoscopy" (KUJF32) and "complete colonoscopy" (ZPY1A0), the NPVs were 77.0% (95% CI: 66.8–85.4), 48.0% (95% CI: 27.8–68.7) and 58.8% (95% CI: 32.9–81.6), respectively.

Discussion

We had access to hospital records for >95% of the 143 sampled individuals, who had had a colonoscopy performed;

Table 3 Validity of specific colorectal cancer screening codes in the DCCSD compared with hospital records, n=136

	DCCSD code	Agreement, % (95% CI); a/n	Specificity, % (95% CI); tn/(tn+fp)	NPV, % (95% CI); tn/(tn+fn)
Clean colon at colonoscopy	AFX02C	83.8 (76.5–89.6); 114/136	97.1 (89.9–99.6); 67/(67+2)	77.0 (66.8–85.4); 67/(67+20)
Clean colon at CT colonography	AFX02D	97.8 (93.7–99.5); 133/136	100.0 (97.2–100.0); 131/(131+0)	97.8 (93.6–99.5); 131/(131+3)
Cancer in cecum	DC180	100.0 (97.3–100.0); 136/136	100.0 (97.3–100.0); 135/(135+0)	100.0 (97.3–100.0); 135/(135+0)
Cancer in ascending colon	DC182	99.3 (96.0–100.0); 135/136	100.0 (97.3–100.0); 135/(135+0)	99.3 (96.0–100.0); 135/(135+1)
Cancer in transverse colon	DC184	99.3 (96.0–100.0); 135/136	100.0 (97.3–100.0); 134/(134+0)	99.3 (95.9–100.0); 134/(134+1)
Cancer in sigmoid colon	DC187	98.5 (94.8–99.8); 134/136	99.2 (95.9–100.0); 131/(131+1)	99.2 (95.9–100.0); 131/(131+1)
Colorectal cancer unspecified	DC189	98.5 (94.8–99.8); 134/136	98.5 (94.8–99.8); 134/(134+2)	100.0 (97.3–100.0); 134/(134+0)
Rectal cancer	DC209	99.3 (96.0–100.0); 135/136	100.0 (97.3–100.0); 133/(133+0)	99.3 (95.9–100.0); 133/(133+1)
Benign tumor in cecum	DD120	96.3 (91.6–98.8); 131/136	99.2 (95.8–100.0); 128/(128+1)	97.0 (92.4–99.2); 128/(128+4)
Benign tumor in ascending colon	DD122	97.1 (92.6–99.2); 132/136	99.2 (95.8–100.0); 130/(130+1)	97.7 (93.5–99.5); 130/(130+3)
Benign tumor in right colic flexure	DD123A	98.5 (94.8–99.8); 134/136	99.2 (95.9–100.0); 132/(132+1)	99.2 (95.9–100.0); 132/(132+1)
Benign tumor in transverse colon	DD123	98.5 (94.8–99.8); 134/136	99.2 (95.8–100.0); 129/(129+1)	99.2 (95.8–100.0); 129/(129+1)
Benign tumor in left colic flexure	DD123B	97.8 (93.7–99.5); 133/136	98.5 (94.7–99.8); 131/(131+2)	99.2 (95.9–100.0); 131/(131+1)
Benign tumor in descending colon	DD124	100.0 (97.3–100.0); 136/136	100.0 (97.3–100.0); 133/(133+0)	100.0 (97.3–100.0); 133/(133+0)
Benign tumor in sigmoid colon	DD125	93.4 (87.8–96.9); 127/136	97.9 (92.5–99.7); 92/(92+2)	92.9 (86.0–97.1); 92/(92+7)
Multiple benign tumors in the colon	DD126C	99.3 (96.0–100.0); 135/136	100.0 (97.3–100.0); 133/(133+0)	99.3 (95.9–100.0); 133/(133+1)
Benign tumor in the rectum	DD128	94.9 (89.7–97.9); 129/136	97.5 (92.9–99.5); 118/(118+3)	96.7 (91.8–99.1) 118/(118+4)
Colonoscopy	KUJF32	87.5 (80.7–92.5); 119/136	75.0 (47.6–92.7); 12/(12+4)	48.0 (27.8–68.7); 12/(12+13)
Colonoscopy with biopsy	KUJF35	89.7 (83.3–94.3); 122/136	90.3 (83.7–94.9); 112/(112+12)	98.2 (93.8–99.8); 112/(112+2)
Complete colonoscopy	ZPY1A0	91.9 (86.0–95.9); 125/136	71.4 (41.9–91.6); 10/(10+4)	58.8 (32.9–81.6); 10/(10+7)
Not complete colonoscopy	ZPY1A1	98.5 (94.8–99.8); 134/136	100.0 (97.3–100.0); 133/(133+0)	98.5 (94.8–99.8); 133/(133+2)
Endoscopic polypectomy, large intestine	KJFA15	95.6 (90.6–98.4); 130/136	96.3 (89.6–99.2); 78/(78+3)	96.3 (89.6–99.2); 78/(78+3)
Endoscopic mucosal resection, large intestine	KJFA55A	98.5 (94.8–99.8); 134/136	99.3 (95.9–100.0); 133/(133+1)	99.3 (95.9–100.0); 133/(133+1)
Rectoscopic polypectomy, rectum	KJGA05	94.9 (89.7–97.9); 129/136	98.4 (94.3–99.8); 123/(123+2)	96.1 (91.1–98.7); 123/(123+5)
Tumor tissue left behind	ZPY1B01	94.9 (89.7–97.9); 129/136	99.2 (95.8–100.0); 128/(128+1)	95.5 (90.5–98.3); 128/(128+6)
No tumor tissue left behind	ZPY1B02	89.7 (83.3–94.3); 122/136	98.7 (92.8–100.0); 74/(74+1)	87.1 (78.0–93.4); 74/(74+11)
Number (nn) polyps seen	ZPY1Cnn	90.4 (84.2–94.8); 123/136	97.1 (90.1–99.7); 68/(68+2)	86.1 (76.5–92.8); 68/(68+11)
No lost polyps	ZPY1D00	87.5 (80.7–92.5); 119/136	94.3 (87.2–98.1); 83/(83+5)	87.4 (79.0–93.3); 83/(83+12)
Largest polyp lost, size in mm	ZPY1Dnn	96.3 (91.6–98.8); 131/136	96.9 (92.3–99.2); 126/(126+4)	99.2 (95.7–100.0); 126/(126+1)
High-risk polyps removed	ZPY1E01	92.6 (86.9–96.4); 126/136	97.5 (93.0–99.5); 119/(119+3)	94.4 (88.9–97.7); 119/(119+7)
Medium-risk polyps removed	ZPY1E02	91.2 (85.1–95.4); 124/136	97.4 (92.7–99.5); 114/(114+3)	92.7 (86.6–96.6); 114/(114+9)
Low-risk polyps removed	ZPY1E03	88.2 (81.6–93.1); 120/136	93.8 (87.5–97.5); 105/(105+7)	92.1 (85.5–96.3); 105/(105+9)

Notes: a, number in agreement; n, number in total.
Abbreviations: CI, confidence interval; CT, computed tomography; fn, false negatives; fp, false positives; NPV, negative predictive value; tn, true negatives; tp, true positives.

thus, 136 patients were included in the analysis. In general, we found high agreement for categories of diagnosis and procedure codes in the DCCSD, when comparing DCCSD data with hospital records.

Agreement was highest for "colorectal cancer" (97.1%) and lowest for "lost polyps" (88.2%). Sensitivity varied between moderate and high with 60.0% for "incomplete colonoscopy" and 98.5% for "colonoscopy performed". Specificity was 92.7% or above except for the categories "colonoscopy performed" and "overall completeness of colonoscopy", where the specificity was low; however, the

estimates were imprecise. The low number of false positives and the general high specificity indicate that the proportion of missing data in the categories of codes in the DCCSD is limited. Considering the individual codes, the validity is less clear as the estimates are somewhat imprecise.

A major strength of this study is that the study population was extracted from the IAM, a registry independent from the DNPR. This enabled us not only to validate correctness of information of the DNPR data in the DCCSD but also to address the degree of missing data, which can lead to underestimations. A common limitation of validity studies is that
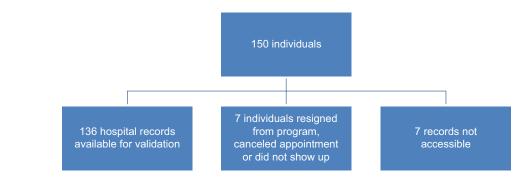
Figure 1 Flowchart of hospital records.

individuals are identified in the registry which is under review and then data are compared with hospital records. When, as in this study, the population can be defined from a third and complete source, clinical registries and hospital records can be compared by calculating specificity and NPVs in addition to sensitivity and PPVs.

One limitation of our study is the relatively small sample size and especially the low numbers of people with cancer, leading to imprecise estimates. Although we used a random sample securing data from all regions, gender and ages, the 136 patients might not be representative of all patients with a positive screening test in the Danish colorectal cancer screening program. It was only possible to have one medical doctor to review the hospital records from each region, and therefore the results may be subject to an unknown level of imprecise extraction of data. The information was delivered in a uniform way via a standardized spreadsheet. This means that interpretation mistakes in the communication between reviewers and researchers were minimized. We used hospital records as the gold standard when evaluating agreement and validity of codes recorded in the DCCSD. This approach is typically used in validity studies,[13] but hospital records do not necessarily represent perfect information as information could be missing or be incorrectly stated. Some of the specific codes had lower validity, mainly colonoscopy procedure codes. These codes (AFX02C, KUJF32, ZPY1A0, ZPY1B02, ZPY1D00 and ZPY1E03) were implemented at the start of the screening program, whereas the remaining codes were already in use for registration of colorectal cancer diagnostic workup and treatment. As time passes, medical staff will become more familiar with the program and the coding procedures and definitions described in the clinical guidelines;[14] thus, the validity of these codes is likely to increase. In addition, implementation of the screening program has put pressure on colonoscopy units, which may have led to some delays in reporting to the DCCSD. If the time period from the end

of the study to data extraction had been longer, it is possible that less data would be missing.

Concerning the specific codes for high-, medium- and low-risk polyps removed, numbers of false negatives varied from 7 to 9, resulting in a low sensitivity for the category "risk group of polyps". Hence, for future studies including data on risk groups, it is recommended also to use the pathology data in the DCCSD, which were not included in this validation study.

Helqvist et al[15] examined the quality of colorectal cancer diagnosis codes (2001–2006) in the DNPR. They compared DNPR data to the data from the Danish Cancer Registry, which has a sensitivity and PPV of >99%.[16] Helqvist et al were able to include >25,000 people with ICD-10 diagnosis codes C18, C19 or C20 for colorectal cancer in their analysis. They found a sensitivity of 93.4% and a PPV of 88.9% for overall colorectal cancer registration. We found a PPV equal to theirs, but a lower sensitivity (72.7%). Because of the small number of cancers in our sample, the CIs were wide for both estimates.

When using DCCSD data for quality assessment and research purposes, some reservations should be noted. First, use of some of the specific codes may result in missing or incorrect data, eg, defining a study population based on a colonoscopy procedure may result in an incomplete study population. A combination of different codes could be used to minimize this problem. Second, patients lacking a code in the DCCSD (false negatives), as well as patients registered with a code in the DCCSD not verified in the hospital records (false positives), might differ from other patients in, eg, disease severity. This may cause misclassification and bias if they differ in relation to the outcome of a specific study.

Conclusion

The high level of agreement between categories of codes in the DCCSD and hospital records indicates that DCCSD reflects the hospital records well. Further, sensitivity and

specificity of the categories appear to vary from moderate to high. Thus, the DCCSD may be a valuable data source for future research on colorectal cancer screening. Considering the specific codes, the validity is less clear and therefore the risk of missing data should be taken into account when using the individual codes for research purposes.

Acknowledgments

Disclosure

The authors report no conflicts of interest in this work.

References

1. Sundhedsstyrelsen [The Danish Health Authority] [webpage on the Internet]. Tarmkræftscreening [Colorectal cancer screening]; 2014. Available from: https://sundhedsstyrelsen.dk/da/sygdom-og-behandling/screening/tarmkraeftscreening#. Accessed February 9, 2016. Danish.
2. Hewitson P, Glasziou P, Watson E, Towler B, Irwig L. Cochrane systematic review of colorectal cancer screening using the fecal occult blood test (hemoccult): an update. *Am J Gastroenterol*. 2008;103(6): 1541–1549.
3. Garborg K, Holme Ø, Løberg M, Kalager M, Adami HO, Bretthauer M. Current status of screening for colorectal cancer. *Ann Oncol*. 2013;24(8): 1963–1972.
4. Nielsen KT. Dccg's Nationale Retningslinier for Diagnostik Og Behandling Af Kolorektal Cancer – Screening [DCCG National Guidelines for Diagnostics and Treatment of Colorectal Cancer – Screening]; 2014. Available from: http://dccg.dk/retningslinjer/august2014/2014_screening.pdf. Accessed September 19, 2016. Danish.
5. Styregruppen for Dansk Tarmkræftscreeningsdatabase (DTS), Kompetencecenter for Epidemiologi og Biostatistik Nord (KCEB-Nord). Dansk Tarmkræftscreeningsdatabase Pilotrapport 2014 [Danish Colorectal Cancer Screening Database Pilot Report 2014]; 2015. Available from: https://www.regionh.dk/kliniskedatabaser/rkkp-databaser/Documents/_Pilot_datavalidering_DTS_20150528_endelig_inklbilag.pdf. Accessed September 19, 2016. Danish.
6. Pedersen CB. The Danish Civil Registration System. *Scand J Public Health*. 2011;39(Suppl 7):22–25.
7. Tværregional arbejdsgruppe for implementering og videndeling. Manual for Implementering Og Drift Af Tværregional Tarmkræftscreening [Manual for Implementing and Running of Mutiregional Colorectal Cancer Screening]; 2014. Available from: http://www.regionshospitalet-randers.dk/siteassets/afdelinger/afdeling-for-folkeundersogelser/pdf-episerver/retningslinjer/2014_09_25-manual-for-implementeirng-og-drift-af-tvarregional-tarmkraftscreening---version-1.pdf. Accessed September 19, 2016. Danish.
8. Pedersen CB, Gøtzsche H, Møller JO, Mortensen PB. The Danish Civil Registration System. A cohort of eight million persons. *Dan Med Bull*. 2006;53(4):441–449.
9. Schmidt M, Pedersen L, Sørensen HT. The Danish Civil Registration System as a tool in epidemiology. *Eur J Epidemiol*. 2014;29(8): 541–549.
10. Andersen TF, Madsen M, Jørgensen J, Mellemkjoer L, Olsen JH. The Danish National Hospital Register. A valuable source of data for modern health sciences. *Dan Med Bull*. 1999;46(3):263–268.
11. Lynge E, Sandegaard JL, Rebolj M. The Danish National Patient Register. *Scand J Public Health*. 2011;39(7 suppl):30–33.
12. Loong T. Understanding sensitivity and specificity with the right side of the brain. *BMJ*. 2003;327(7417):716–719.
13. Schmidt M, Schmidt SAJ, Sandegaard JL, Ehrenstein V, Pedersen L, Sørensen HT. The Danish National Patient Registry: a review of content, data quality, and research potential. *Clin Epidemiol*. 2015;7:449–490.
14. Arbejdsgruppen vedr. registrering i screeningsprogrammet for tarmkræft. Screenings- og adenomkontrol program for tyk og endetarmskræft. Guidelines for koloskopi og patolog [Screening and Adenoma Follow Up for Colorectal Cancer. Guideline for Colonocscopy and Pathology]. Available from: http://www.danskpatologi.dk/doc/pdf/CRC-screeningsprogram%20retningslinjer%20v12.pdf. Accessed December 14, 2016. In Danish.
15. Helqvist L, Erichsen R, Gammelager H, Johansen MB, Sørensen HT. Quality of ICD-10 colorectal cancer diagnosis codes in the Danish National Registry of Patients. *Eur J Cancer Care (Engl)*. 2012;21(6): 722–727.
16. Storm HH, Michelsen EV, Clemmensen IH, Pihl J. The Danish Cancer Registry – history, content, quality and use. *Dan Med Bull*. 1997;44:535–539.