Validity of data in the Danish Colorectal Cancer Screening Database

Mette Kielsholm Thomsen¹
Sisse Helle Njor¹
Morten Rasmussen²
Dorte Linnemann³
Berit Andersen⁴
Gunnar Baatrup⁵,⁶
Lennart Jan Friis-Hansen⁷
Jens Christian Riis Jørgensen⁸
Ellen Margrethe Mikkelsen¹

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

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. In the case of a positive test result (>100 ng/mL), a colonoscopy is performed, 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.

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. 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). The IAM is

<table>
<thead>
<tr>
<th>Table 1 Categories of codes for diagnoses and procedures included in the validation of the DCCSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category: Clean colon</td>
</tr>
<tr>
<td>Category: Colorectal cancer</td>
</tr>
<tr>
<td>Category: Polyps/benign tumors</td>
</tr>
<tr>
<td>Category: Colonoscopy performed</td>
</tr>
<tr>
<td>Category: Overall completeness of colonoscopy</td>
</tr>
<tr>
<td>Category: Incomplete colonoscopy</td>
</tr>
<tr>
<td>Category: Polypectomy</td>
</tr>
<tr>
<td>Category: Tumor tissue left behind</td>
</tr>
<tr>
<td>Category: Number of polyps seen</td>
</tr>
<tr>
<td>Category: Lost polyps</td>
</tr>
<tr>
<td>Category: Risk group of polyps</td>
</tr>
</tbody>
</table>

**Abbreviation:** DCCSD, Danish Colorectal Cancer Screening Database.
updated daily with information on vital status, addresses, etc from the Civil Registration System.\textsuperscript{5,9}

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.\textsuperscript{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.\textsuperscript{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
“number of polyps seen” and “lost polyps”, sensitivity was 75\% (95\% CI: 63.0–84.79; Table 2). For the categories (including both “tissue left behind” and “no tissue left behind”) “endoscopic polypectomies, large intestine” (KJFA15), which was 92.1\% (95\% CI: 82.4–97.4) and 98.6\% (95\% CI: 92.6–100.0), respectively. 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” (ZPY1A0), “benign tumor in descending colon” [DD124]; 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;
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
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, 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; 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 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%. 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
The following individuals have contributed to the establishment of the DCCSD and in developing the idea behind this validity study: Mona Skarbye, Department of Gastrointestinal Surgery, Regional Hospital Slagelse, Slagelse, Denmark; Per Gandrup, Department of Surgery A, Aalborg University Hospital, Aalborg, Denmark; Henrik Nørgaard, Danish Radiology Society, Denmark; Per Ejstrup, Danish Society for Gastroenterology and Hepatology, Denmark.

Disclosure
The authors report no conflicts of interest in this work.

References