Project Development phase

Code-Layout, Readability And Reusability

DATE	06 May 2023
Team ID	NM2023TMID18418
Project Name	CancerVision: Advanced Breast Cancer Prediction with Deep Learning

Objectives

To assess quality and reusability of coded cancer diagnoses in routine primary care data. To identify factors that influence data quality and areas for improvement.

Methods

A dynamic cohort study in a Dutch network database containing 250,000 anonymized electronic medical records (EMRs) from 52 general practices was performed. Coded data from 2000 to 2011 for the three most common cancer types (breast, colon and prostate cancer) was compared to the Netherlands Cancer Registry.

Measurements

Data quality is expressed in Standard Incidence Ratios (SIRs): the ratio between the number of coded cases observed in the primary care network database and the expected number of cases based on the Netherlands Cancer Registry. Ratios were multiplied by 100% for readability.

Results

The overall SIR was 91.5% (95%CI 88.5–94.5) and showed improvement over the years. SIRs differ between cancer types: from 71.5% for colon cancer in males to 103.9% for breast cancer. There are differences in data quality (SIRs 76.2% − 99.7%) depending on the EMR system used, with SIRs up to 232.9% for breast cancer. Frequently observed errors in routine healthcare data can be classified as: lack of integrity checks, inaccurate use and/or lack of codes, and lack of EMR system functionality.

Conclusions

Re-users of coded routine primary care Electronic Medical Record data should be aware that 30% of cancer cases can be missed. Up to 130% of cancer cases found in the EMR data can be false-positive. The type of EMR system and the type of cancer influence the quality of coded diagnosis registry. While data quality can be improved (e.g. through improving system design and by training EMR system users), re-use should only be taken care of by appropriately trained experts.

Design

We performed a dynamic cohort study in a Dutch network database containing 250,000 anonymized electronic medical records (EMRs) from 52 general practices. We used a 4 step study approach, as described in Fig. 2, to determine Standardized Incidence Rate Ratios (SIRs) between January 1st 2000 and December 31st 2011.

First, we determined our reference standard: the expected incidence rates based on the Netherlands Cancer Registry (NCR) [19] and Statistics Netherlands [20].

Second, observed incidence

Results

The combined SIR for breast, colon, and prostate cancer between 2000 and 2011 was 91.5%, (95%CI 88.5–94.5). This means there is a significant difference between the observed number of cases in the EMR and the expected number according to the NCR .

The SIRs varied over time: from 2000 to 2003 the combined SIR was 66.3%, (95%CI 61.3–71.3), from 2004 to 2007 it was 95.7% (95%CI 90.3–100.9), and from 2008 to 2011 it was 103.8% (95%CI 98.8–108.6). For colon cancer in males the SIR was 71.5%