Utilizing national patient-register data to control for comorbidity in prognostic studies

Yngvar Nilssen¹
Trond-Eirik Strand¹
Robert Wiik²
Inger Johanne Bakken³
Xue Qin Yu⁴,⁵
Dianne L O’Connell⁶–⁷
Bjørn Møller¹

¹Department of Registration, Cancer Registry of Norway, Oslo, ²Norwegian Patient Register Department, Norwegian Directorate of Health, Trondheim, ³Norwegian Institute of Public Health, Oslo, Norway; ⁴Cancer Research Division, Cancer Council NSW, ⁵School of Public Health, University of Sydney, ⁶School of Public Health and Community Medicine, University of New South Wales, Sydney, ⁷School of Medicine and Public Health, University of Newcastle, Newcastle, NSW, Australia

Objective: To construct an updated comorbidity index (Patient Register Index [PRI]) using national data collections from Norway and compare its predictive ability of 1-year mortality with the Charlson Comorbidity Index (CCI).

Materials and methods: Data regarding over 1.11 million patients registered in the Norwegian Patient Register in 2010 and 2011 were used to construct the PRI. The PRI was evaluated by comparing its model fit and discrimination with the CCI.

Results: Compared with the CCI, the PRI weights decreased for six, increased for four, and were unchanged for seven diseases. When the PRI was added to the model including age and sex, the age effects were reduced by up to 38% for patients older than 50 years. All measures of model fit improved for the PRI model.

Conclusion: Adjustment for comorbidity is especially important for patients 50 years of age or older, and its effect on 1-year mortality is almost comparable to the age effect. The PRI is based on more recent data than the CCI, and is more representative of the general population due to its construction.

Keywords: comorbidity, National Patient Register, Charlson Comorbidity Index, predictive ability, patient-register index, prognostic value

Introduction

Comorbidities are defined as diseases that are concomitant to the disease under study.¹ When survival in medical and epidemiological research is studied, it is important to consider the effect of comorbidity, as it could be a potential confounder or an effect modifier for other prognostic factors.²–⁶

The Charlson Comorbidity Index (CCI) and the Elixhauser method are the most commonly used methods to assess comorbidity.⁷ The CCI is used more often, is less complex than the Elixhauser method, and was therefore chosen as the reference index in the present study. The CCI was originally developed based on data from all patients admitted to the New York Hospital–Cornell Medical Center during a 1-month period in 1984, with the objective of predicting 1-year mortality. The risk of death associated with each of the 19 predefined diseases included in the CCI was expressed as weights with values of 1, 2, 3, or 6. Summing the weights for all contributing diseases gives the CCI score for each patient. The CCI was validated using an external cohort consisting of 685 breast cancer patients receiving their first treatment in 1962–1969 at Yale New Haven Hospital.⁷

The original CCI has been modified and evaluated by many authors.⁹–¹³ Quan et al developed coding algorithms for constructing the CCI based on the codes of the
Tenth Revision of the International Classification of Diseases (ICD-10). Sundararajan et al assessed that the version of the CCI created by Quan et al outperformed all the other versions they considered, though all of the ICD-10 versions of the Charlson algorithm performed satisfactorily. However, a persistent challenge for researchers who wish to use the CCI is that the data set of interest may not contain sufficient medical information to assess all the diseases included in the index. One source of such information in many countries is national patient registers. These are based on hospital administrative databases, from which data are often readily available.

Due to better treatment and technological improvements, a patient’s risk of death has reduced since the CCI was established in the 1980s, and the index has been updated accordingly. Bottle and Aylin and Quan et al both updated the CCI. Bottle and Aylin used national data from the UK and studied in-hospital mortality, while Quan et al used regional data from the Calgary Health Region of Canada to study 1-year case fatality. However, to our knowledge, no previous study has combined the use of national patient-register data and death within 1 year as the study end point.

The aim of the present study was to evaluate the importance of including a comorbidity measure into analyses when predicting 1-year mortality using data from the Norwegian Patient Register (NPR). A second aim was to construct a modified version of the CCI that can be used with patient-register data (the Patient Register Index, PRI), and compare the predictive ability of the CCI and the PRI, as well as the confounding effect of each index on age. An additional aim was to describe the pattern of diseases included in the CCI by age and sex, based on data from all hospitals in Norway registered in the NPR.

### Materials and methods

#### The NPR

The NPR is a national health register covering all sectors of the specialized health care services. Reporting to NPR is mandatory, and the register includes data on all patients treated in Norwegian government-funded hospitals. Personal identification numbers have been reported to the NPR from 2008 onwards. This enables researchers and policy makers to follow the disease trajectory of patients between sectors and hospitals. Also, alignment of data and validation with other national health registries is made feasible.

The study sample was selected from the NPR data reported from all hospitals, which consisted of three main data sources for statistics. The first source was visits for medical treatment for in- and outpatients at publicly financed hospitals. The two other sources were private hospitals and private specialist practices. The government purchases medical treatment from private hospitals and specialists practices as a supplement to services at the public hospitals. The NPR does not include data on privately financed hospital treatments.

The basic data unit in the NPR is hospital visits. However, when a patient is transferred between wards at the same hospital, the record is aggregated. Each episode of national hospital data contains one or more diagnoses, coded according to the ICD-10.

#### Study sample

To provide information on the total disease history of the patient, the personal identification number was used to link episodes of treatment registered at different hospitals, sectors, and years. Initially, all patients in the NPR registered with a hospital visit in 2010 or 2011 were identified, constituting 15,214,796 visits. Figure 1 shows an overview of exclusions of patient visits in the study. Three subgroups of the initial data suffered from incomplete reporting of personal identification numbers. The first group was outpatients at private specialist practices in 2010 and 2011 (3,359,618 visits). The reason for incomplete reporting in this group was mainly related to technical limitations in their information-technology systems. To avoid any selection bias, we then also excluded 10,131,791 outpatient visits registered at publicly financed hospitals. However, we decided to keep...
outpatient visits with surgical procedure codes, because this activity mainly is performed at publicly financed hospitals or publicly financed private hospitals. The second group of visits excluded due to incomplete reporting were data on newborn babies (139,196 visits), since a large proportion had not yet received a personal identification number at the time of registration. The third group was any other patient visits that were lacking a correctly reported personal identification number (49,249 visits). After these exclusions, the data were aggregated, resulting in 1,534,942 eligible visits, corresponding to 1,113,341 unique patients.

The first hospital visit during the study period was defined as the patient’s index visit. The end of follow-up was defined as the date of death if the patient died within 1 year of the index visit, or the date 1 year after the index visit (censoring date). Sex, age category, and ICD-10 codes (truncated to three digits for the index visit) were obtained from the NPR. To avoid the possibility of tracing back to individual patients, we received age information in 5-year age-group categories. Duration in days from the index visit to end of follow-up, vital status, and history of the diseases (dichotomous variables) included in the CCI that were listed up to 1 year prior to the index visit were obtained from the NPR. The NPR routinely obtains information on vital status and date of death from the National Population Register. The data file obtained from the NPR was anonymized, and thus no requirement for regulatory ethical approval in Norway was needed.

The original CCI included 19 diseases, but for the purposes of the present study, leukemia and lymphoma were included in the disease category “any malignancy”, as done by other authors. These 17 CCI diseases were identified by the ICD-10 coding defined by Quan et al in 2005. A primary diagnostic code is the code for the main medical condition causing the admission. Secondary diagnostic codes are codes for diseases that exist at the same time as the primary disease, or diseases that develop and need examination or treatment during the admission. Information from both primary and secondary diagnostic codes was obtained for each patient.

**Statistical analysis**

The PRI weights were estimated using a Cox regression model of time from index visit to death with follow-up censored at 1 year, adjusted for sex, 5-year age categories (0–4 years … 95+ years) and 17 CCI disease indicators. An algebraically correct method was used to define new weights for the CCI diseases by summing the regression coefficients, not the hazard ratios (HRs). The PRI was constructed by multiplying the regression coefficients by a scaling constant – k, then rounding it to the nearest integer, and finally summing it over all the CCI diseases. The scaling constant k was chosen such that the maximum weight for a specific disease in the PRI became the same integer as the maximum weight in the CCI, and thus k was set to 2.3.

Both the construction and comparison of the PRI and the CCI were done through an internal tenfold cross-validation procedure. Ten subsets of the data were randomly defined, and nine of them were used to develop the PRI, with the last subset used to run the comparison. This was done for all ten subsets, and the results are given as the mean values from the ten runs. To predict 1-year mortality, three logistic regression models were fitted, one including sex and age (base model), one including sex, age, and the CCI (CCI model), and one model including sex, age, and the PRI (PRI model).

Model fit was compared using several measures of global fit. The Bayesian information criterion (BIC) is a likelihood-based measure, and for model fit, the lower the BIC-value, the better the fit. Likelihood ratio chi-square statistics were also used to test model fit. McFadden’s $R^2$ is a measure of improvement in fit over the intercept model. The Brier score was used to directly compare the observed outcomes with the predicted probabilities. The C-statistic, which is a summary of a model’s ability to discriminate between those who do and those who do not experience the outcome, was calculated for each model. This measure is the most commonly used in the medical and epidemiological literature when comparing different comorbidity indices. The C-statistic varies from 0.5, which indicates that the discrimination is due to chance alone, to 1, which indicates perfect discrimination. The general classification of discrimination is “acceptable” (c ∈ [0.7, 0.8]), “excellent” (c ∈ [0.8, 0.9]), and “outstanding” (c ∈ [0.9, 1]). To assess which index changed the predicted value most when added to the base model, movement from the estimated probabilities of the outcome were compared. This was done using the net reclassification improvement (NRI), which is defined as the difference in proportions moving up or down in risk among patients who died and those who survived. We also used the integrated discrimination improvement (IDI), which measures the difference in the mean predicted probabilities between those who died and those who did not in the CCI and PRI models compared to the base model. We also calculated how much the probabilities changed on average when a comorbidity index was included in the model:
\[
\Delta \hat{p} = \frac{1}{N} \sum_{i=1}^{N} \left| \hat{p}_{\text{age,sex},i} \right| - \left| \hat{p}_{\text{age,sex,comorbidity},i} \right| \quad (1)
\]

This quantifies the effect of introducing a comorbidity index into a model. To compare the importance of including age, sex, and a comorbidity index in the model, \( \Delta \hat{p} \) was calculated with and without each of the variables in a subanalysis with patients aged over 50 years who were not admitted to hospital due to any of the 17 CCI diseases at the index visit.

In the description of the disease pattern, the patient group was age-standardized according to the Norwegian standard population in 2011.\(^{27}\) All analyses were done using Stata 13.\(^{28}\)

**Results**

The majority of the 1,113,341 patients in the study were women (57.0%). For 68.1% of the patients, the index visit was a hospitalization. Figure 2 presents the proportion of the Norwegian population with one or more of the CCI diseases in 2010–2011 and the total proportion admitted to Norwegian hospitals in that same period. Twenty-two percent of the patients were registered with at least one CCI disease, but among men and women below 50 years of age, the proportion was less than 3%. For patients 50 years of age or older, the proportion with CCI diseases increased with age, and the increase was more prominent among men. For both sexes, there was a strikingly high proportion of hospital admission in the youngest age-group (0–4 years). This was mostly due to diseases of the respiratory system (data not shown). For women, there was a peak around 30–34 years (Figure 2) as a result of birth-related admissions (data not shown). There was a small peak for men aged 20–24 years as well, which was mostly attributed to ICD-10 codes within S00-T98 (“Injury, poisoning, and certain other consequences of external causes”, data not shown).

The age-standardized proportions of the Norwegian population with each of the CCI diseases are shown in Figure 3. The five most common CCI diseases for men and women were myocardial infarction, cerebrovascular disease, chronic pulmonary disease, diabetes without chronic complications, and malignancies, but the order of importance differed.

![Figure 2](https://www.dovepress.com/)

**Figure 2** Total proportion of the Norwegian population admitted to hospitals in 2010–2011 in Norway or with a condition included in the Charlson Comorbidity Index (CCI) present within the previous year, by sex and age-group.

**Notes:** Data shown are for (A) women and (B) men.
The overall proportion of deceased patients within 1 year was 4.8% (52,938 patients), and the 1-year risk of death was significantly higher for patients with at least one CCI disease compared to those with none (HR 4.5, confidence interval 4.4–4.6) (Table 1). The proportion of patients with each CCI disease ranged between 0.05% (acquired immunodeficiency syndrome [AIDS]/human immunodeficiency virus [HIV]) and 4.9% (chronic obstructive pulmonary disease [COPD]). Metastatic tumor was associated with the highest hazard, while rheumatic diseases and diabetes without chronic complications were associated with the lowest hazard for 1-year mortality. Across all CCI diseases, the mean CCI score was 0.4, the median number of diseases present in the patients was 0, and the range was 0–8.

When the weights for the CCI diseases in the PRI were compared with those in the CCI, the weights for seven diseases remained unchanged, four increased in magnitude, and six decreased in magnitude. Except for two CCI diseases – moderate-to-severe liver disease and AIDS/HIV – the weights in the PRI only deviated by one if at all compared to those in the CCI. The maximum observed CCI and PRI values among the patients were 12 and 15, respectively. The overall mean CCI was approximately equal to the mean PRI (0.43 versus 0.42).

Table 2 presents the measures of model fit obtained from the internal cross-validation for the three prediction models. The C-statistic was marginally better for the PRI model than for the CCI model (91.5 versus 91.3). All measures showed improvement in fit when the PRI was included in the model instead of the CCI (Table 2).

Among patients aged 50 years or older who were not admitted to hospital due to any of the 17 CCI diseases, the ten most common categories for admission are listed in Table 3. ICD-10 code H25, “Senile cataract” (4.2%), and R07, “Pain in throat and chest” (3.0%), were the two most common conditions. The rest contributed with less than 3% each.

Figure 4 shows the change in the effect of age when either of the two comorbidity indices was added to the base model for these patients. Including the PRI in the model led to a reduction in the effect of age for all age-groups, ranging from 3.5% in the 55–59-year group to 38% in the 95+ year group. This reduction was greater than the reduction observed for the CCI model.

Using backward elimination from the PRI model for patients 50 years of age or older with an index visit that was not due to a CCI disease, the change in the estimated probability (Δp) was 2.8%, 4.3%, and 0.7% when the PRI, age, and sex were removed, respectively. Similarly, for the CCI
model, $\Delta p$ was 2.6%, 4.7%, and 0.7% when the CCI, age, and sex were removed, respectively.

**Discussion**

We observed that for patients 50 years of age or older, comorbidity based on NPR data was almost equally important as age for predicting 1-year mortality. Measured by the proportion of the Norwegian population with a CCI disease registered in the NPR, the disease burden increased with age, reaching 52% and 38% in the oldest age-group for men and women, respectively. The weights for four CCI diseases were higher in the PRI compared to the corresponding weights in the CCI, while the weights for six diseases were lower in the PRI than in the CCI.

The PRI was marginally better than the CCI in predicting 1-year mortality. The C-statistics showed that the performance of both indices was rated as “outstanding” in discriminating, according to the standard C-statistic classification, between those who died and those who did not. The base model produced a C-statistic of 0.869, while the C-statistic increased to 0.913 and 0.915 when the CCI and the PRI were included in the model, respectively. Therefore, the majority of the ability to discriminate came from sex and age, but both of the indices increased the discrimination ability further, which is in agreement with the findings of Gabbe et al. For patients over 50 years of age, we observed that the PRI was almost as important as age, while the CCI was relatively less important than the PRI. In addition, a reduction in the effect of age for those over 50 years of age increased with age toward 38% when the PRI was added to the base model, showing the importance of including information regarding comorbidity when predicting the probability of death, especially in elderly patients.

The lack of data from privately financed hospitals could cause concern of possible introduced bias. However, in 2008 the total use of privately financed hospitals in Norway was approximately 0.5% of the total health care service. Therefore, excluding these patients would probably not have induced any major bias. Further, we only included inpatients and those outpatients with a surgical procedure code. The exclusion of all the outpatient visits with only medical procedures or no procedures at all probably led to an underestimation of the prevalence of the least severe diseases (those who did not need a hospitalization). The main aim in this study, however, was not to estimate the prevalence of a disease, but to construct a valid modification of the CCI. The private specialist practices are not situated evenly throughout the country. If we had included all outpatient

**Table 1** Number and proportion of patients (present and dead) with Charlson Comorbidity Index (CCI) diseases (not mutually exclusive) registered in the Norwegian Patient Register between January 1, 2010 and December 31, 2011

<table>
<thead>
<tr>
<th>Comorbidity</th>
<th>Number of patients</th>
<th>Proportion with comorbidity (%)</th>
<th>Number dead</th>
<th>Proportion dead (%)</th>
<th>HR</th>
<th>95% CI</th>
<th>CCI</th>
<th>PRI</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>861,553</td>
<td>77.4</td>
<td>12,296</td>
<td>1.4</td>
<td>1.0</td>
<td>–</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Yes</td>
<td>251,788</td>
<td>22.6</td>
<td>40,642</td>
<td>16.1</td>
<td>4.46</td>
<td>4.36–4.55</td>
<td>&gt;0</td>
<td>&gt;0</td>
</tr>
<tr>
<td>Total</td>
<td>1,113,341</td>
<td>100.0</td>
<td>52,938</td>
<td>4.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Type of comorbidity**

- No chr dis
- Myocardial infarction
- Congestive heart failure
- Peripheral vascular dis
- Cerebrovascular dis
- Dementia
- Chronic pulmonary dis
- Rheumatic dis
- Peptic ulcer dis
- Mild liver dis
- Diabetes without chr compl
- Diabetes with chr compl
- Hemiplegia or paraplegia
- Renal dis
- Any malignancy
- Mod or severe liver dis
- Metastatic solid tumor
- AIDS/HIV

**Notes:** Hazard ratios (HRs) for 1-year mortality, and 95% confidence intervals (CIs), adjusted for age and sex, together with weights for both the CCI and Patient Register Index (PRI), are shown.

**Abbreviations:** AIDS, acquired immunodeficiency syndrome; HIV, human immunodeficiency virus; dis, disease; chr, chronic; compl, complications; mod, moderate.
visits, this would have introduced a possible regional bias using the PRI.

The Norwegian Prescription Database reported that approximately 156,000 (3.1%) of Norwegians used medication for diabetes in 2011. The total prevalence of diabetes in Norway is uncertain. In addition to those using medications, there is an unknown number of patients with type 2 diabetes who are managed through lifestyle changes only. However, a Norwegian study reported that 4.3% of the population suffered from diabetes in the period 2006–2008. In our data, we identified approximately 47,000 cases of diabetes treated in hospitals during the study period (1% of the Norwegian population). The Norwegian Institute of Public Health has reported that the prevalence of COPD in Norway is around 200,000 (4%). In our study, 54,410 patients admitted to hospital in 2010–2011 were either treated for COPD or had it registered in their disease history. Lastly, the Cancer Registry of Norway reported a cancer prevalence of 215,000 in 2011. The observed prevalence from the NPR is measured using a combination of prevalence and incidence, since it includes patients with either a newly diagnosed cancer or a diagnosis in the past in continuous need of treatment. However, considering any malignancy and metastatic tumor as one group identified in the NPR had more severe disease and thus poorer prognosis than those suffering from the same diseases, but not admitted to hospital.

The uniqueness of this study is that we used two alternative methods (NRI and IDI) for comparing the quality of the reclassification of patients. None of the previous papers describing modifications of a comorbidity index has to our knowledge used these predictive probability measures. The NRI was 74.5% and 84.0% for the CCI and PRI, respectively, showing that inclusion of a comorbidity index increased the credibility of a model’s predictive ability, and the model including the PRI correctly determined certain other consequences of external causes.
index in a model, as opposed to using the base model, in predicting the 1-year mortality for patients 50 years of age or older.

The present study is closely related to the work done by Quan et al and Bottle and Aylin.11,12 Quan et al developed an updated version of the CCI based on the population of Calgary, Canada, and validated it using data from six different nations. A major advantage of the study of Quan et al was that they validated their updated index externally, in addition to being population-based. The advantage of Bottle and Aylin’s study was that they developed new empirical weights based on English administrative data in 2007–2008 (with over 5 million records). However, they studied in-hospital mortality only, as individual follow-up was not possible. Our proposed PRI is based on much larger patient numbers than Quan et al’s, and more complete follow-up after discharge from hospital compared with that of Bottle and Aylin, which is very important, especially considering that only 35% of the deceased in Norway in 2011 died in hospitals.37

When constructing the PRI, diagnoses of the CCI diseases, recorded both in primary and secondary fields from the index visit as well as the history, and recorded up to 1 year prior to the index visit were included. The rationale behind this was that such an index should be based on the most recent disease history available. Pine et al and Ghali et al both showed improved discrimination when conditions present upon admission were included.38,39 In all subanalyses for patients 50 years of age or older, we excluded data for the patients who were admitted to hospital due to any of the 17 CCI diseases. The rationale behind this was that patients under 50 years of age have a low presence of comorbidities. In addition, we wanted to study comorbidity – diseases in addition to the one causing the index visit – and not morbidity. This study design is similar to the situation researchers meet when they want to add comorbidity information to their study cohort.

One limitation of our study was that we were not able to validate the PRI using external data. Another limitation regards the inclusion criteria of patients: we excluded outpatients registered with only medical procedures or no procedures at all. This excluded many patients and limits the generalizability of the results, but it was necessary due to the large variation regarding the degree of reported personal identification numbers from private specialist practices. Outpatient

Figure 4 The age effects (hazard ratios) relative to the age-group 50–54 years (log scale), for different models based on patients not admitted for any of the 17 Charlson Comorbidity Index (CCI) diseases registered in the Norwegian Patient Register in 2010–2011.

Abbreviation: PRI, Patient Register Index.
consultations are usually scheduled for patients with chronic and non-life-threatening diseases, such as dementia, rheumatism, and diabetes; hence, there is a considerable risk of underestimating the number of patients suffering from these diseases in our study. On the other hand, the strengths of the study are that in contrast to the CCI, the PRI was developed using data from a national register. Secondly, the algebraically correct method of the CCI was used when developing the PRI. Finally, the large sample size and access to complete mortality data are obvious strengths.

**Conclusion**

It is of high importance to include a comorbidity index in observational studies of disease prognosis, especially in elderly patients. Both the CCI and the PRI showed a high degree of discrimination, indicating that both have good predictive ability. However, the PRI explained a larger proportion of the observed effect of age, and the PRI weights reflect the patterns in the data from the NPR. We have shown that weighting the CCI for a specific population slightly improves the performance of the CCI. The PRI is by its construction more representative of the general population, and can be generalized to other countries in situations where data from a national patient register are used.

**Disclosure**

The authors report no conflicts of interest in this work.

**References**


