The New Comorbidity Index

A Development and Validation Study

Part II - Data Analysis

Rikke Beck Nielsen
and
Sinna Pilgaard Ulrichsen

AALBORG UNIVERSITY

Dep. of Mathematical Sciences • Master's Thesis • 1. Sep. - 30. May 2010

TITLE:

The New Comorbidity Index
A Development and Validation Study

PROJECT PERIOD:

From 1. September 2009
To 30. May 2010

PROJECT GROUP:

Rikke Beck Nielsen
Sinna Pilgaard Ulrichsen

SUPERVISOR:

Poul Svante Eriksen

COPIES: 10

NUMBER OF PAGES: 110

AALBORG UNIVERSITY

Dep. of Mathematical Sciences • Master's Thesis • 1. Sep. - 30. May 2010

Table of contents

Part I

Data analysis

Methods

1.1 Data sources

The study was based on the Danish medical and administrative registries which included information on the entire Danish population. The Danish National health service provides tax-supported health care with free access to primary and hospital care for the approximately 5.4 million residents [?]. The civil registration number given to every resident since 1968 ensured a direct connection between the Danish registries. [?]

The Civil Registration System
The Civil Registration System contains information on civil registration number, name, address, citizenship and date of death if any, emigration, and immigration for the entire Danish population since 1968. [?]

The Danish National Registry of Patients (DNRP)
This registry receives data from all non-psychiatric hospitals in Denmark with information on 99.5% of all discharges since 1977. From 1995 and onwards the outpatient and emergency room visits were included. This registry contains the civil registration number, information on dates of admission and discharge, surgical procedures performed and up to 20 discharge diagnoses coded by physicians according to the International Classification of Diseases, the 8^{th} revision (ICD-8) until 1994 and the 10^{th} revision (ICD-10) from 1994 and onwards. [?] [?]

1.2 Definitions

Study design
Because we had a large population-based dataset with complete follow-up for mortality the cohort study design was chosen. This study design requires a large sample size and has the advantage that it is possible to compute the risk of the outcome.
A dataset containing all first time inpatient admissions to a nonpsychiatric hospital due to pneumonia in Denmark between 1997 and 2007 was collected from the Danish National Registry of Patients. This dataset was used in cohort studies with 30 day and 1 year mortal-

ity as outcomes. Comorbidities were used as exposure and potential confounders were sex and age.

First time hospital admissions due to pneumonia
All adult (age $\geq$ 15 years) patients with the discharge diagnosis of pneumonia (ICD-10 codes J12 - J18) were found in the Danish National Registry of Patients. All readmissions due to pneumonia were removed and patients with first time admissions before the study period were excluded (ICD-8 codes used from 1977 throughout 1994 are 480 - 486, 073, 471). Patients with legionellosis (ICD-10 code A481) and ornithosis (ICD-10 code A709) were excluded since these diseases are rare and have a higher risk of death than ordinary pneumonia.

Mortality
Outcomes in the studies were death from any cause within 30 days or 1 year following the admission date. The date of death was found in the Danish Civil Registration System.

Comorbidity
Information on all comorbidities diagnosed prior to admission due to pneumonia, was found in the DNRP. The diagnosis codes were listed in table A.1 and A.2 in the appendix.

Potential confounders
The information on the potential confounders, sex and age, were found in the Danish civil registration system. Other potential confounders could be residence and hospital of treatment, but these were not taking into account.

1.3 Statistical analysis

Validation of the Charlson comorbidity index
In the initial investigation of the compliance of CCI Kaplan-Meier curves were made to assess CCI's crude predictor capabilities of 30 day mortality. A logistic regression model was then made with adjustments for age and sex, and with CCI as a predictor for mortality, and the fit of the model to the data was assessed. The choice of the logistic regression model was made because this model handles the binary response well. The choice was possible because there was no general

problem with censured data. Advantages of this model were that all model assumptions concerned the model equation and were verifiable, and that the model parameters were easy to interpret [**?**].

Index developing
A logistic regression with 30 day mortality as an endpoint was made with all the 19 diseases from CCI as predictors. The regression was made both with and without adjustment for sex and age, where age was modeled by a restricted cubic spline. An extended list of diseases was also used to make an adjusted logistic regression.
Furthermore logistic regressions with pairwise interaction was made, in order to investigate the possibility of sex-dependent comorbid effects, of age-dependent comorbid effects and of pairwise comorbid effects. A logistic regression model including time since diagnosis was also made, in order to take the progression of the diseases into account. This was done both with and without first degree interactions. On the basis of the log odds ratios from the regression models, new weights for each disease were found and new comorbidity indexes were made.

As an alternative to the logistic regression model, the naive Bayes method and classification trees were used to develop indexes. Both methods were used on the extended list of diseases. These methods were chosen because of their simplicity, the naive Bayes method because of the simple theory and the classification trees because of the simple resulting index.

Validating the developed comorbidity indexes
To assess the validity of the developed comorbidity indexes, we made a chi-squared test and a logistic regression including the indexes, for both 30 day and 1 year mortality.
The chi-squared test was made to assess the indexes' crude ability to predict death. The corresponding deviance residuals were made to assess the indexes' ability to separate and order the comorbidity groups. The logistic regression was made to assess the indexes' performance when adjusted for sex and age. The performance of the indexes in a logistic regression was evaluated on the basis of the Hosmer-Lemeshow test statistic, the generalized R^2 and the area under the ROC curve.

Data presentation

In this chapter two datasets are introduced. The first dataset was used to develop a number of comorbidity indexes and the second dataset was used to validate these.

2.1 Training data

The dataset in this section consisted of a cohort of Danish pneumonia patients, and was called the training data. The training data contained all first time inpatient admissions to a nonpsychiatric hospital due to pneumonia in Denmark between 1997 and 2006 as described in section 1.2.

	30 day mortality in no.	**30 day mortality in %**	**no. all**	**% all**
Sex				
Females	14585	15	100227	49
Males	16973	16	104620	51
Age				
[15-40]	289	2	16743	8
[41-65]	4310	8	51505	25
[66-80]	11871	15	76932	38
[81-110]	15088	25	59667	29
Total	31558	15	204847	100

Table 2.1 *Basic information about the training data.*

Basic information about the dataset is shown in table 2.1. As seen in the table 51% of the patients were males and percentage-wise a few more males than females died within 30 days of the pneumonia diagnosis. The age distribution was leaning towards the elderly part of the population and as expected the mortality increased with age.
Information about the 19 diseases from the Charlson comorbidity index can be seen in table 2.2, 2.3 and 2.4. Information on three additional diseases: alcohol related disorders, a history of obesity and hypertension are in table 2.4. The ICD-10 and ICD-8 diagnosis codes used to find

all the diseases can be seen in appendix in table A.2. These diseases are known to influence the mortality among people in general [**?**], so including them in an index might improve it.
In general the patients with a previous diagnosis of any of the diseases had a higher mortality than the patients without, except for patients with chronic pulmonary disease, lymphoma, AIDS / HIV or history of obesity. The difference in mortality was especially pronounced among the patients with dementia and the patients with a metastatic solid tumor, here the difference in percentages of the mortalities were 17 and 16 respectively.

	30 day mortality in no.	**30 day mortality in %**	**no. all**	**% all**
Myocardial infarction				
Yes	3421	19	18448	9
No	28137	15	186399	91
Congestive heart failure				
Yes	5315	23	23330	11
No	26243	14	181517	89
Peripheral vascular disease				
Yes	3506	21	16614	8
No	28052	15	188233	92
Cerebrovascular disease				
Yes	7312	23	31573	15
No	24246	14	173274	85
Dementia				
Yes	2285	32	7079	3
No	29273	15	197768	97

Table 2.2 *The distribution of the first five diseases from the Charlson comorbidity index in the training data.*

	30 day mortality in no.	**30 day mortality in %**	**no. all**	**% all**
Chronic pulmonary disease				
Yes	5656	14	39680	19
No	25902	16	165167	81
Connective tissue disease				
Yes	1786	16	10950	5
No	29772	15	193897	95
Ulcer disease				
Yes	3542	20	17311	8
No	28016	15	187536	92
Mild liver disease				
Yes	947	21	4477	2
No	30611	15	200370	98
Diabetes type I and II				
Yes	3076	19	16189	8
No	28482	15	188658	92
Hemiplegia				
Yes	234	18	1284	1
No	31324	15	203563	99
Moderate to severe renal disease				
Yes	1434	20	7088	3
No	30124	15	197759	97
Diabetes with end organ damage				
Yes	1714	20	8575	4
No	29844	15	196272	96
Any tumor				
Yes	7088	24	29880	15
No	24470	14	174967	85

Table 2.3 *The distribution of the next nine diseases from the Charlson comorbidity index in the training data.*

	30 day mortality in no.	**30 day mortality in %**	**no. all**	**% all**
Leukemia				
Yes	418	20	2077	1
No	31140	15	202770	99
Lymphoma				
Yes	560	15	3455	2
No	30998	16	201392	98
Moderate to severe liver disease				
Yes	319	25	1264	1
No	31239	15	203583	99
Metastatic solid tumor				
Yes	1440	31	4588	2
No	30118	15	200259	98
AIDS / HIV				
Yes	31	7	437	0
No	31527	15	204410	100
Alcohol related disorders				
Yes	2122	17	12391	6
No	29436	15	192456	94
History of obesity				
Yes	796	13	6197	3
No	30762	15	198650	97
Hypertension				
Yes	4953	18	28076	14
No	26605	15	176771	86

Table 2.4 *The distribution of the last five diseases from the Charlson comorbidity index and the three additional diseases in the training data.*

2.2 Validation data

The dataset presented in this section was used to validate the developed indexes' ability to predict both 30 day and 1 year mortality, and was called the validation data. The validation data contained all first time inpatient admissions to a nonpsychiatric hospital due to pneumonia in Denmark in 2007 as described in section 1.2.
Basic information about the dataset is shown in table 2.5. The distribution of these variables were similar to the ones in the training data. Comparing 30 day mortality and 1 year mortality, it is seen that the 1 year mortality was approximately twice the size for all variables.

	30 day mortality in %	**1 year mortality in %**	**no. all**	**% all**
Sex				
Females	15	32	11327	49
Males	16	35	11822	51
Age				
[15-39]	1	3	1678	7
[40-64]	9	20	5811	25
[65-79]	14	33	7870	34
[80-110]	25	50	7790	34
Total	16	33	23149	100

Table 2.5 *Basic information about the validation cohort.*

Information about the 19 diseases from the Charlson comorbidity index can be seen in table 2.6, 2.7 and 2.8. Information on the three additional diseases: alcohol related disorders, a history of obesity and hypertension are in table 2.8. The general pattern of the diseases were similar to the ones for the training data, except for connective tissue disease, leukemia and alcohol related disorders. Connective tissue disease showed the opposite tendency for 30 day mortality in this validation data than in the training data. So patients in the validation data with the disease had a lower 30 day mortality than patients without. For leukemia and alcohol related disorders having the diseases made no difference in 30 day mortality in the validation data, but it did in the training data.

	30 day mortality in %	**1 year mortality in %**	**no. all**	**% all**
Myocardial infarction				
Yes	19	39	2252	10
No	15	33	20897	90
Congestive heart failure				
Yes	21	47	2665	12
No	15	31	20484	88
Peripheral vascular disease				
Yes	21	45	2086	9
No	15	32	21063	91
Cerebrovascular disease				
Yes	23	47	4052	18
No	14	30	19097	83
Dementia				
Yes	33	61	1051	5
No	15	32	22098	95
Chronic pulmonary disease				
Yes	13	33	4504	19
No	16	33	18645	81
Connective tissue disease				
Yes	14	34	1326	6
No	16	33	21823	94
Ulcer disease				
Yes	20	44	2045	9
No	15	32	21104	91

Table 2.6 *The distribution of the first eight diseases from the Charlson comorbidity index in the validation data.*

	30 day mortality in %	**1 year mortality in %**	**no. all**	**% all**
Mild liver disease				
Yes	18	39	600	3
No	15	33	22549	97
Diabetes type I and II				
Yes	17	38	2208	10
No	15	33	20941	90
Hemiplegia				
Yes	19	38	167	1
No	16	33	22982	99
Moderate to severe renal disease				
Yes	20	43	1124	5
No	15	33	22025	95
Diabetes with end organ damage				
Yes	17	38	1384	6
No	15	33	21765	94
Any tumor				
Yes	24	55	4211	18
No	14	29	18938	82
Leukemia				
Yes	16	48	269	1
No	16	33	22880	99
Lymphoma				
Yes	13	37	507	2
No	16	33	22642	98
Moderate to severe liver disease				
Yes	22	46	180	1
No	15	33	22969	99

Table 2.7 *The distribution of the next nine diseases from the Charlson comorbidity index in the validation data.*

	30 day mortality in %	**1 year mortality in %**	**no. all**	**% all**
Metastatic solid tumor				
Yes	30	72	723	3
No	15	32	22426	97
AIDS / HIV				
Yes	2	12	51	0
No	16	33	23098	100
Alcohol related disorders				
Yes	16	34	1664	7
No	16	33	21485	93
History of obesity				
Yes	12	26	953	4
No	16	34	22196	96
Hypertension				
Yes	18	39	5330	23
No	15	32	17819	77

Table 2.8 *The distribution of the last two diseases from the Charlson comorbidity index and the three additional diseases in the validation data.*

The compliance of CCI with pneumonia patients

In this chapter the performance of the original Charlson comorbidity index on pneumonia patients is assessed. Since this is basically a validation of the CCI on pneumonia patients, no training data is necessary, and therefore both the training data and the validation data, described in chapter 2 are used.
To test the compliance of the original Charlson comorbidity index with the pneumonia patients, a logistic regression with 30 day mortality as an outcome and the index as a predictor was made. The logistic regression was adjusted for age and sex. Kaplan-Meier curves were also made to see how well the CCI separated the different comorbidity groups.

Calculation of the CCI

The CCI was calculated for every patient by adding the weights of all the comorbid diseases in the Charlson index, the patient has had prior to the pneumonia admission. The 19 diseases and their weights can be seen in table A.1 in the appendix in part I. There were however some exceptions when adding the weights. If a patient had been diagnosed with both mild liver disease and moderate or severe liver disease, the weight from mild liver disease should not be added to the index, since the moderate or severe liver disease diagnosis overrules the mild liver disease diagnosis. The same was in evidence for diabetes versus diabetes with end organ damage and any tumor versus metastatic solid tumor.
After the index had been calculated it was divided into four groups: 0, 1-2, 3-4, ≥ 5, since this is common practice [**?**].

Results

The distribution of CCI is shown in table 3.1. The index values was concentrated on the small values and the mortality increased with the Charlson comorbidity index, as seen in other studies
[**?**], [**?**].

The Charlson comorbidity index	30 day mortality in no.	30 day mortality in %	no. all	% all
0	8008	10	83382	37
1-2	15518	17	93602	41
3-4	7598	22	35362	16
5+	4030	26	15650	7

Table 3.1 *Distribution of CCI.*

The Kaplan-Meier curves in figure 3.1 showed that, as expected the groups were ordered with the lowest value having the highest survival probability, the second lowest value having the second highest and so on. It was also seen that the distance between the groups was similar. This shows that the CCI can separate the different comorbidity groups quit well.

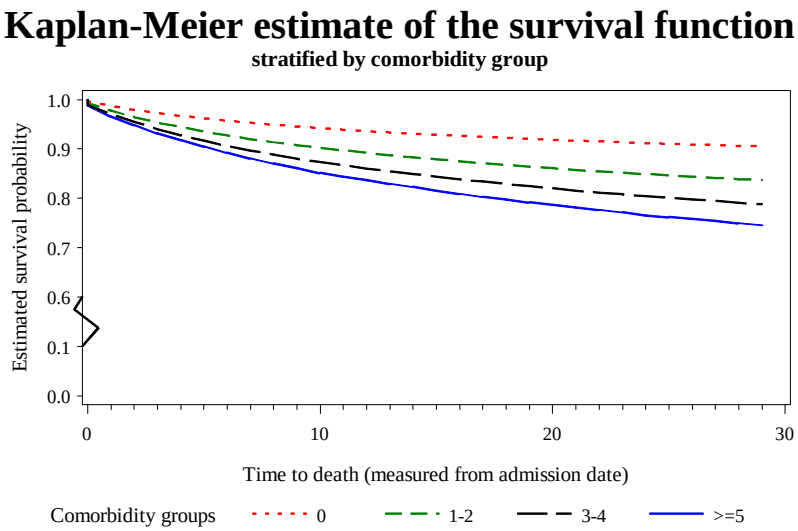


Figure 3.1 *Kaplan-Meier curves for 30 day mortality.*

Next, in order to assess the predictability of CCI the logistic regression adjusted for age and sex was made. The first step was to test the linearity assumptions for age. This was done by making a restricted

cubic spline with five knots on age and plot it against logit of the predicted probability, which can be seen in figure 3.2.

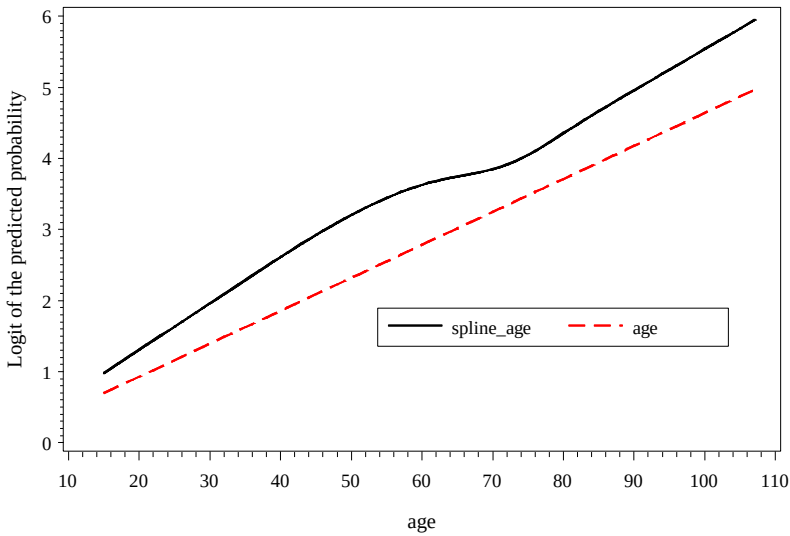


Figure 3.2 *The spline curve and the linear curve for age*

Since figure 3.2 showed that death within 30 days was not linear in age, age was from this point on modeled with a restricted cubic spline.

A new logistic regression adjusted for sex and age modeled by a restricted cubic spline was then made. To assess the fit of the model the Hosmer-Lemeshow χ^2 statistic was calculated and can be seen in table 3.2.

Group	Total	Observed deaths	Expected deaths	Deviance residuals
1	22927	380	496	-4
2	22655	1190	1323	-4
3	22598	1828	1921	-2
4	22769	2466	2463	0
5	22745	3035	2960	1
6	22811	3629	3507	2
7	22685	4384	4099	4
8	22770	5113	4836	4
9	22664	5843	5734	1
10	23372	7286	7816	-6

Table 3.2 *Hosmer-Lemeshow χ^2 test.*

When looking at the probability groups made by the Hosmer-Lemeshow χ^2 test, it seemed that the model overestimated the number of deaths in the highest probability group. The Hosmer-Lemeshow χ^2 value on 154 (p-value$<$0.0001) indicated a lack of fit, however this could be caused by the high number of observations. The χ^2 value also indicated that there is unmodeled information in the data.

In addition to the Hosmer-Lemeshow test, regression diagnostics were calculated. For each of the predictors the $\Delta\hat{\beta}$ was plotted against the predicted probability to see if any of the observations were overinfluential. None of the $\Delta\hat{\beta}$'s exceeded 0.4 which means that there were no overinfluential observations. The value 0.4 was chosen since a change in an estimate of 0.4 changes the odds by 1.5, which normally is the maximal acceptable change.[**?**] The area under the ROC-curve for this model was 0.698, which means that the model discriminated poorly between outcomes.

The significance of the index were also investigated. As seen in table 3.3 all the predictors were statistically significant, so the original CCI had a significant amount of information about mortality.

Effect	**DF**	**Wald Chi-Square**	**Pr** $> \chi^2$
sex	1	170	$<$.0001
$0 < age$	1	245	$<$.0001
$34 < age$	1	45	$<$.0001
$62 < age$	1	47	$<$.0001
$73 < age$	1	26	$<$.0001
CCI	3	1302	$<$.0001

Table 3.3 *Test of statistical significans.*

The odds ratios from the logistic regression adjusted for age and sex, seen in table 3.4 showed that the risk of dying within 30 days increased with the value of the comorbidity index as wanted.

Effect	Odds ratio Estimate	Lower 95% CL	Upper 95% CL
sex 1 vs 0	1.25	1.21	1.29
$0 < age$	1.07	1.06	1.07
$34 < age$	0.95	0.94	0.96
$62 < age$	1.60	1.40	1.83
$73 < age$	0.41	0.29	0.58
CCI 1-2 vs 0	1.49	1.43	1.55
CCI 3-4 vs 0	1.95	1.85	2.05
CCI $\geq$5 vs 0	2.83	2.66	3.01

Table 3.4 *Odds Ratio Estimates.*

To assess the predictability of the CCI, the predicted probabilities were plotted against a smoothed curve of the observed probabilities. Figure 3.3 show that CCI is good at predicting the probability of dying within 30 days in the range with many observations.

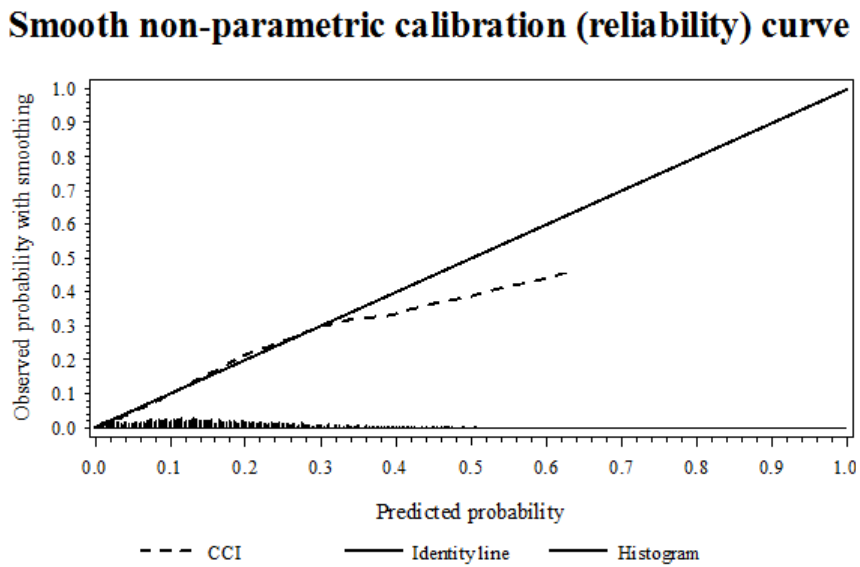


Figure 3.3 *Calibration plot*

Conclusion

The above analysis showed that the Charlson comorbidity index was a significant predictor for death in a logistic regression. The log odds ratios for the comorbidity groups were nicely separated, since the 95% confidence intervals did not overlap. The Kaplan-Meier curves showed that the index gave a good crude discrimination between the comorbidity groups, even though it was poor at discriminating between outcome. When the probability of dying was below 30%, the CCI estimated the probability of dying within 30 days among pneumonia patients well. However there were still room for improvement, as especially the high probability groups overestimated death.

Index development using logistic regression

After having seen that there was room for improvement of the Charlson comorbidity index, we proceeded by developing the index. In this chapter logistic regression models were used to develop indexes. When seeking to improve an index we wished to strengthen the performance and still keep the simplicity. Therefore in the following sections we started with a very simple model and then increased the complexity.

All the following regressions had some similarities. All, except the first regression, were adjusted for sex and age, which was modeled by a restricted cubic spline with five knots as in chapter 3. The weights were then calculated by multiplying the log odds ratio by ten and rounding to the nearest integer. The weights of the registered diseases were added with the same exceptions as in CCI and the values were then grouped into appropriate intervals to make the index. The grouping was done so that each interval had a prevalence of more than 1% in order to ensure the applicability of the index on smaller datasets. We grouped the values because they had a large range and if this was not done, too many degrees of freedom would be used, and it would be better just to include the diseases directly in the model instead of an index.

The index was then included in a logistic regression adjusted for sex and age (modeled by a restricted cubic spline) in order to assess the index's applicability. The Hosmer-Lemeshow χ^2 value, the generalized R^2 value and the area under the ROC curve, were stated for comparison. Kaplan-Meier curves and calibration curves were also made to assess the index's ability to separate the groups and to predict the probability of death.

4.1 An index based on a crude logistic regression

At first a crude logistic regression, containing only the 19 diseases was made. This regression had a Hosmer-Lemeshow χ^2 value of 488 with 6 degrees of freedom.

The Hosmer-Lemeshow test for the regression only subdivided into 8 groups instead of 10. This was caused by the low number of covariate patterns and by the fact that the test arranges observations with the same covariate pattern into the same probability group.

The χ^2 value was high, which indicated a lack of fit, but this may be caused by the high number of observations. The χ^2 value also showed that there still was unmodeled information in the data.

Along with the χ^2 value the $\Delta\hat{\beta}$'s were calculated. All of the $\Delta\hat{\beta}$'s were below 0.4 so none of the observations were overinfluential.

The area under the ROC-curve was 0.636 for this regression, which means that the model discriminated rather poorly. A Wald test showed that four of the diseases (myocardial infarction, connective tissue disease, hemiplegia and lymphoma) were statistically insignificant, but since all of the diseases were of interest none of them were left out.

The new weights were then calculated on the basis of the log odds ratios. The parameter estimates for the diseases and their weights can be seen in table 4.1.

For each observation the appropriate weights were then added, and the values grouped into ≤ 0, 1-3, 4-6, 7-9, 10-12, 13-15 and ≥ 16.

Kaplan-Meier curves were then made and can be seen in figure 4.1. It is seen that after approximately 8 days the different comorbidity groups are separated nicely.

Effect	Parameter estimate	Weight
Myocardial infarction	0.04	0
Congestive heart failure	0.47	5
Peripheral vascular disease	0.24	2
Cerebrovascular disease	0.49	5
Dementia	0.90	9
Chronic pulmonary disease	-0.17	-2
Connective tissue disease	0.02	0
Ulcer disease	0.25	3
Mild liver disease	0.30	3
Diabetes I and II	0.10	1
Hemiplegia	0.06	1
Moderate to severe renal disease	0.14	1
Diabetes with end organ damage	0.08	1
Any tumor	0.57	6
Leukemia	0.41	4
Lymphoma	-0.04	0
Moderate to severe liver disease	0.63	6
Metastatic solid tumor	1.12	11
AIDS/HIV	-0.70	-7

Table 4.1 *Parameter estimates and the assigned weights.*

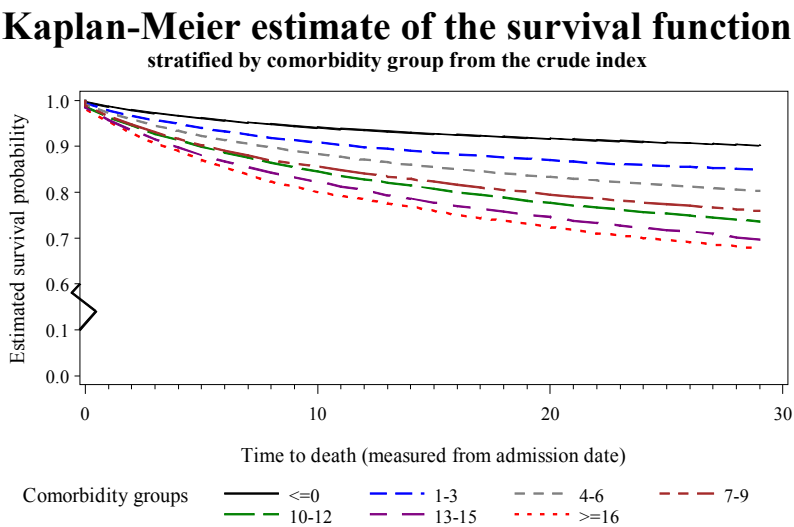


Figure 4.1 *Kaplan-Meier curves for the new index based on a crude logistic regression.*

To assess the applicability of the new index, a logistic regressions with the index as a predictor was made.
The Hosmer-Lemeshow test gave the following:

Group	Total	Observed deaths	Expected deaths	Deviance residuals
1	20296	336	452	-5
2	20584	980	1184	-6
3	20430	1518	1682	-4
4	20491	2133	2112	0
5	20887	2784	2637	3
6	20883	3378	3159	4
7	20537	4003	3676	5
8	20322	4638	4324	5
9	20506	5408	5324	1
10	19911	6380	7008	-8

Table 4.2 *Hosmer-Lemeshow χ^2 test.*

Table 4.2 show that the new index is bad at estimating the number of deaths, especially in the low and highest probability groups.

The statistics for the logistic regression are summarized in table 4.3.

Predictors	χ^2_{HL}	R^2	ROC
Crude index	266 (8 DF)	0.1007	0.705

Table 4.3 *The goodness of fit statistics for the new index.*

To see how well the new index predicts death within 30 days, the observed and the predicted probabilities were plotted against one another.

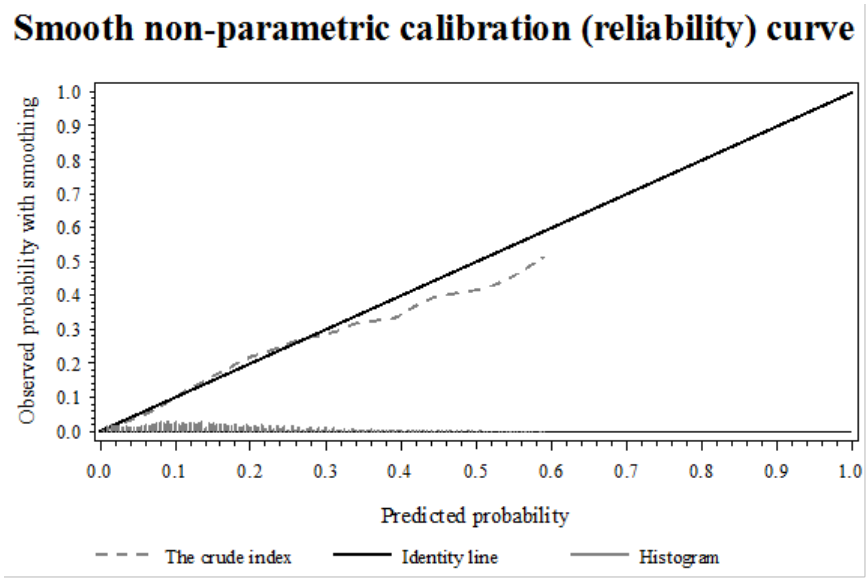


Figure 4.2 *Observed and predicted probabilities for the new index.*

Figure 4.2 show that the new index is good at predicting the probability of dying within 30 days, especially in the range with many observations.

4.2 An index based on an adjusted logistic regression

Age and sex are known to influence mortality in general. So to make the index useful on datasets with different age and sex distributions, we now adjust for age and sex in the regression models used to develop the index.

The Hosmer-Lemeshow test for this regression gave a χ^2 value of 255 with 8 degrees of freedom. The test again indicated a lack of fit, but still it may be caused by the high number of observations.

The Wald test showed that now only three diseases (connective tissue disease, lymphoma and AIDS / HIV) were statistically insignificant. Again the three diseases were of interest so they were kept. The area under the ROC curve was 0.709, so the model discriminated poorly.

The new weights were calculated based on the following parameter estimates. Note that the diseases with '*' have a notably different weight compared to the weights in the crude index.

Effect	**Parameter estimate**	**Weight**
Myocardial infarction	-0.05	-1
Congestive heart failure*	0.23	2
Peripheral vascular disease	0.20	2
Cerebrovascular disease*	0.27	3
Dementia*	0.62	6
Chronic pulmonary disease	-0.12	-1
Connective tissue disease	-0.02	0
Ulcer disease*	0.14	1
Mild liver disease*	0.68	7
Diabetes I and II	0.09	1
Hemiplegia*	0.38	4
Moderate to severe renal disease	0.22	2
Diabetes with end organ damage	0.21	2
Any tumor*	0.43	4
Leukemia	0.44	4
Lymphoma	0.06	1
Moderate to severe liver disease*	1.08	11
Metastatic solid tumor	1.23	12
AIDS /HIV*	0.26	3

Table 4.4 *Parameter estimates and the assigned weights.*

The appropriate weights were added and the values were then grouped into the groups ≤ 0, 1-3, 4-6, 7-9, 10-12, ≥ 13.
The Kaplan-Meier curves for the index are seen in figure 4.3.

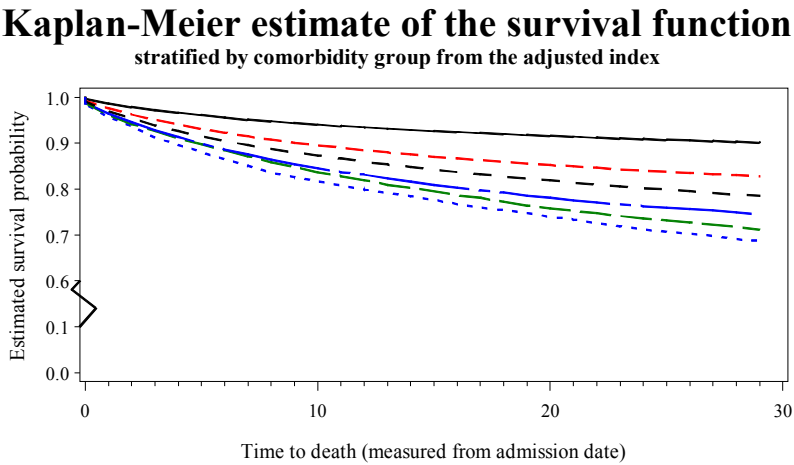


Figure 4.3 *Kaplan-Meier curves for the new index based on an adjusted logistic regression.*

Figure 4.3 shows that the curves are nicely separated.

A logistic regression adjusted for sex and age was then made, including the adjusted index as a predictor. The Hosmer-Lemeshow test is seen in table 4.5.

Group	Total	Observed deaths	Expected deaths	Deviance residuals
1	20717	327	449	-6
2	20432	972	1137	-5
3	20769	1489	1679	-5
4	20308	2107	2094	0
5	20156	2677	2547	3
6	20071	3229	3026	4
7	20543	3978	3666	5
8	20466	4621	4335	4
9	20488	5440	5267	2
10	20897	6718	7359	-7

Table 4.5 *Hosmer-Lemeshow χ^2 test.*

Table 4.5 show that this index also was bad at estimating the number of deaths, in the low and highest probability groups.

A comparison of the overall test statistics of the developed indexes is seen in table 4.6.

Predictors	χ^2_{HL}	R^2	**ROC**
Crude index	266	0.1007	0.705
Adjusted index	257	0.1052	0.709

Table 4.6 *Comparison of the indexes.*

The test statistics showed that the adjusted index fitted the data slightly better than the crude index. The Hosmer-Lemeshow χ^2 value also showed that the adjusted index modeled a little more information than the crude index. A comparison of how well the indexes predicted the probability of dying within 30 days is seen in figure 4.4. The figure shows that there was no major difference in the predictability of the two new indexes, especially when the probability of dying was below 40%.

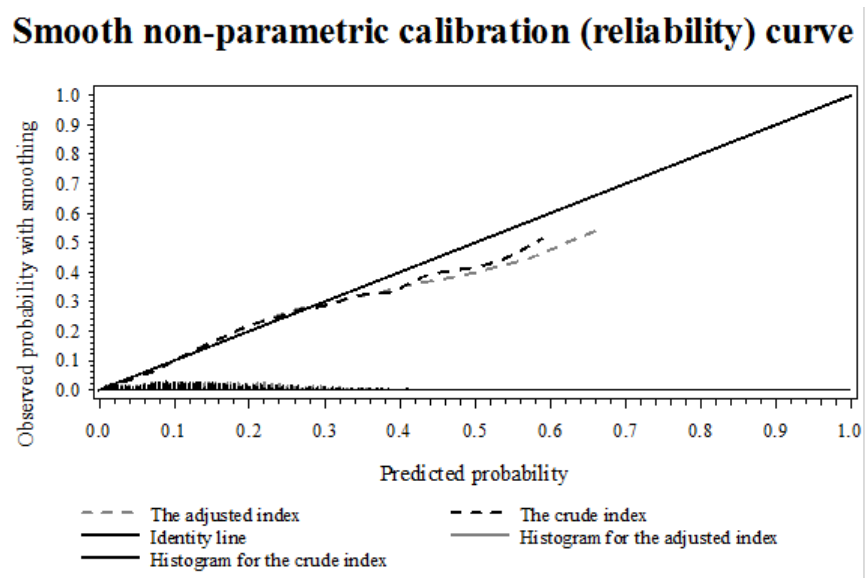


Figure 4.4 *Observed and predicted probabilities for the two new indexes.*

4.3 An index including 22 diseases

In this section three additional diseases were included in the index. The three diseases were alcohol related disorders, a history of obesity and hypertension. These diseases are known to influence the mortality among people in general [**?**], so including them in the index might improve it.

The index was made on the basis of the 19 diseases from the Charlson index and the three additional diseases in the same way as in section 4.2. This means that a logistic regression with 30 day mortality as outcome and the 22 diseases as predictors adjusted for sex and age was made.
The Hosmer-Lemeshow test statistic for this model was 272 with 8 degrees of freedom. The $\Delta\hat{\beta}$'s were also calculated and all of the $\Delta\hat{\beta}$'s were below 0.4 so none of the observations were overinfluential.

The area under the ROC-curve was 0.711 for this regression, which means that the model discriminated better than the model without the three additional diseases but still rather poorly. A Wald test showed that four of the diseases (myocardial infarction, connective tissue disease, lymphoma and AIDS / HIV) were statistically insignificant, but again they were all kept.

The new weights were then calculated from the log odds ratios. The parameter estimates for the diseases and their weights can be seen in table 4.7. Note that mild liver disease and moderate to severe liver disease have notably different weights compared to index with only 19 diseases.

Effect	**Parameter estimate**	**Weight**
Myocardial infarction	-0.04	0
Congestive heart failure	0.23	2
Peripheral vascular disease	0.21	2
Cerebrovascular disease	0.28	3
Dementia	0.59	6
Chronic pulmonary disease	-0.13	-1
Connective tissue disease	-0.01	0
Ulcer disease	0.13	1
Mild liver disease*	0.48	5
Diabetes I and II	0.11	1
Hemiplegia	0.38	4
Moderate to severe renal disease	0.26	3
Diabetes with end organ damage	0.24	2
Any tumor	0.43	4
Leukemia	0.44	4
Lymphoma	0.07	1
Moderate to severe liver disease*	0.88	9
Metastatic solid tumor	1.24	12
AIDS / HIV	0.26	3
Alcohol related disorders	0.51	5
History of obesity	-0.14	-1
Hypertension	-0.12	-1

Table 4.7 *Parameter estimates and the assigned weights. The diseases marked by '*' have notably different weights compared to the index with only 19 diseases.*

For each observation the appropriate weights were then added, and the values was grouped like earlier into the groups ≤ 0, $1-3$, $4-6$, $7-9$, $10-12$, ≥ 13 to make the index. The Kaplan-Meier curves for this index can be seen in figure 4.5.

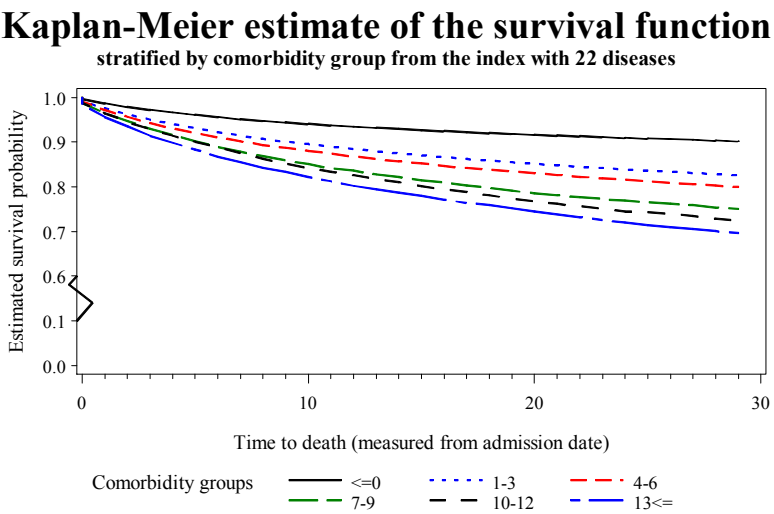


Figure 4.5 *Kaplan-Meier curves for the index with 22 diseases.*

It is seen that after approximately 8 days the different comorbidity groups are nicely separated.

The logistic regression with the index as a predictor was then made. The Hosmer-Lemeshow test for this model can be seen in table 4.8.

Group	Total	Observed deaths	Expected deaths	Deviance residuals
1	20642	282	430	-7
2	20499	902	1111	-6
3	20480	1499	1630	-3
4	20853	2153	2141	0
5	20415	2751	2587	3
6	20637	3356	3137	4
7	20294	3956	3652	5
8	20130	4618	4296	5
9	20369	5425	5279	2
10	20528	6616	7295	-8

Table 4.8 *Hosmer-Lemeshow* χ^2 *test.*

Table 4.8 showed that the regression did not predict death very well in the lower and higher end of the probability scale.
The statistics for this regression are summarized in the bottom row in table 4.9, where the statistics from the indexes found in section 4.1 and 4.2 also are noted for comparison.

Predictors	χ^2_{HL}	R^2	ROC
Crude index (19 diseases)	266	0.1007	0.705
Index (19 diseases)	257	0.1052	0.709
New index (22 diseases)	300	0.1067	0.710

Table 4.9 *The goodness of fit statistics for all the developed indexes.*

Looking at the values in table 4.9, there seems to be no notably difference between the index (19 diseases) and the new index (22 diseases). To see how well the new index predicts death within 30 days, the observed against the predicted probabilities were plotted. The result can be seen in appendix in figure A.1 and is very similar to the calibration plots for the indexes based on the 19 diseases.

4.4 An index including interaction with sex

In the ongoing search for an improved index, we next investigated if including interaction between each disease and sex made a difference. An example of a disease with a known risk difference is myocardial infarction [**?**].

Since only 1 percent of the subjects have hemiplegia we did not allow interaction between sex and this disease. The same goes for leukemia, moderate to severe liver disease and AIDS / HIV. This was done to ensure, that we only included extra terms in the index with enough statistical power and that we did not make the index unnecessarily complicated.

A logistic regression with the 22 diseases and interaction between each of the remaining 18 diseases and sex was made. On the basis of this regression the significant interaction terms were identified.

Effect interacting with sex	**Parameter estimate**	**p-value for int. with sex**
Myocardial infarction	0.17	$<$.0001
Congestive heart failure	0.01	0.8917
Peripheral vascular disease	0.00	0.9269
Cerebrovascular disease	0.05	0.1191
Dementia	-0.01	0.8326
Chronic pulmonary disease	0.06	0.0896
Connective tissue disease	0.10	0.0891
Ulcer disease	-0.04	0.3844
Mild liver disease	-0.01	0.8848
Diabetes I and II	0.06	0.3232
Hemiplegia	0	-
Moderate to severe renal disease	0.01	0.9313
Diabetes with end organ damage	0.06	0.3710
Any tumor	-0.09	0.0129
Leukemia	0	-
Lymphoma	0.08	0.4464
Moderate to severe liver disease	0	-
Metastatic solid tumor	-0.20	0.0040
AIDS / HIV	0	-
Alcohol related disorders	0.05	0.4146
History of obesity	-0.02	0.8279
Hypertension	0.01	0.7142

Table 4.10 *Parameter estimates and the corresponding p-value for the interaction terms in the logistic regression.*

Looking at table 4.10 it is seen that only myocardial infarction, any tumor and metastatic solid tumor had a significant interaction term, so these were the only ones included in the index. The logistic regression used to develop an index therefore had the 22 diseases, interaction between sex and the diseases: myocardial infarction, any tumor and metastatic solid tumor.

The Hosmer-Lemeshow test statistic for this model was 286 with 8 degrees of freedom. The area under the ROC-curve was 0.711 for this regression, which means that this model discriminated slightly better

than the model without interaction with sex but still rather poorly.

The new weights were then calculated from the log odds ratios. In table 4.11 the parameter estimates for the diseases and their weights are shown. Note that the diseases without interaction terms have weights identical to the ones in the index with only the 22 diseases.

Effect	**Parameter estimate**	**Weight**
Myocardial infarction, female	0.08	1
Myocardial infarction, male	-0.11	-1
Congestive heart failure	0.23	2
Peripheral vascular disease	0.21	2
Cerebrovascular disease	0.28	3
Dementia	0.59	6
Chronic pulmonary disease	-0.13	-1
Connective tissue disease	-0.01	0
Ulcer disease	0.13	1
Mild liver disease	0.48	5
Diabetes I and II	0.11	1
Hemiplegia	0.38	4
Moderate to severe renal disease	0.26	3
Diabetes with end organ damage	0.24	2
Any tumor, female	0.38	4
Any tumor, male	0.47	5
Leukemia	0.44	4
Lymphoma	0.07	1
Moderate to severe liver disease	0.88	9
Metastatic solid tumor, female	1.13	11
Metastatic solid tumor, male	1.33	13
AIDS / HIV	0.26	3
Alcohol related disorders	0.51	5
History of obesity	-0.14	-1
Hypertension	-0.12	-1

Table 4.11 *Parameter estimates and the assigned weights.*

For each observation the appropriate weights were then added, and the values were grouped like earlier into the groups ≤ 0, $1-3$, $4-6$, $7-9$, $10-12$, ≥ 13. The Kaplan-Meier curves can be seen in figure 4.6.

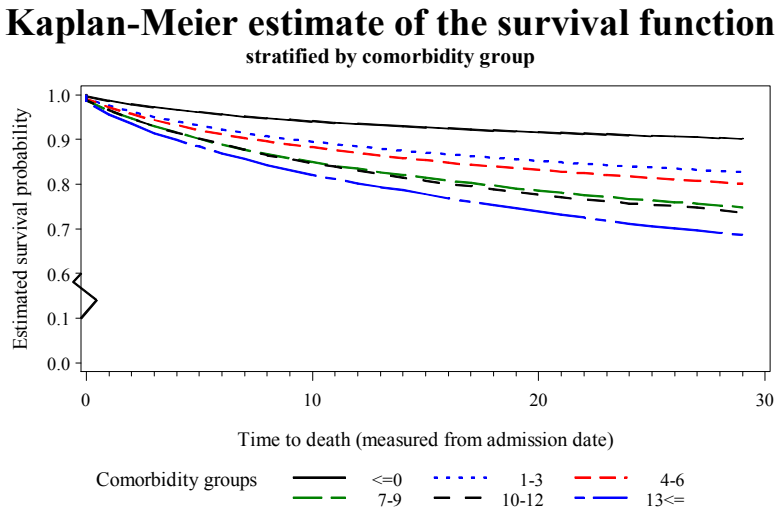


Figure 4.6 *Kaplan-Meier curves for the index with interaction with sex.*

It is seen that after approximately 11 days the different comorbidity groups are separated, though the curves for group 7-9 and group 10-12 lay very close.

To assess the applicability of the index, a logistic regression was made with the index as a predictor. The Hosmer-Lemeshow test for this model can be seen in table 4.12.

Group	Total	Observed deaths	Expected deaths	Deviance residuals
1	20627	281	430	-7
2	20448	895	1107	-6
3	20420	1506	1623	-3
4	20877	2144	2139	0
5	20434	2710	2582	3
6	20517	3373	3119	5
7	20865	4041	3754	5
8	20633	4772	4432	5
9	20504	5530	5378	2
10	19468	6306	6995	-8

Table 4.12 *Hosmer-Lemeshow χ^2 test.*

The table showed that the regression did not predict death well in the lower and higher end of the probability scale.
The statistics for this regression are summarized in the bottom row in table 4.13, where the statistics from the indexes found in the previous sections also are noted for comparison.

Predictors	χ^2_{HL}	R^2	ROC
Crude index (19 diseases)	266	0.1007	0.705
Index (19 diseases)	257	0.1052	0.709
Index (22 diseases)	300	0.1084	0.710
New index with sex int. (22 diseases)	308	0.1065	0.711

Table 4.13 *The goodness of fit statistics for the developed indexes.*

The Hosmer Lemeshow χ^2 value, the R^2 value and the area under the ROC curve, does not seem to be improved with this new index. The calibration plot for this index was very similar to the plot for the indexes with 19 diseases and can be seen in appendix in figure A.2.

4.5 An index including interaction with age

In this section we investigate if the diseases have any interaction with the patients' age. Age is known to have an influence on mortality when the prognosis of diseases is assessed [**?**], [**?**], but this effect may now be the same for different diseases. Therefore we include interactions between age and the 22 diseases in the index in this section.

As was the case with interaction with sex, we only allowed interaction for diseases with more than 1 percent prevalence and only if the interaction term was significant. This meant that interaction between age and the diseases hemiplegia, leukemia, moderate to severe liver disease and AIDS / HIV, was not considered.
In order to keep the index relatively simple, age was not included as a continuous variable, so it was discretized. A clinical relevant discretization was to use the intervals [15-40], [41-65], [66-80], [81-110] [**?**]. A quick look at the mortality rate against age seen in figure 4.7 showed, that these intervals seemed to separate age nicely in intervals with different mortality rates.

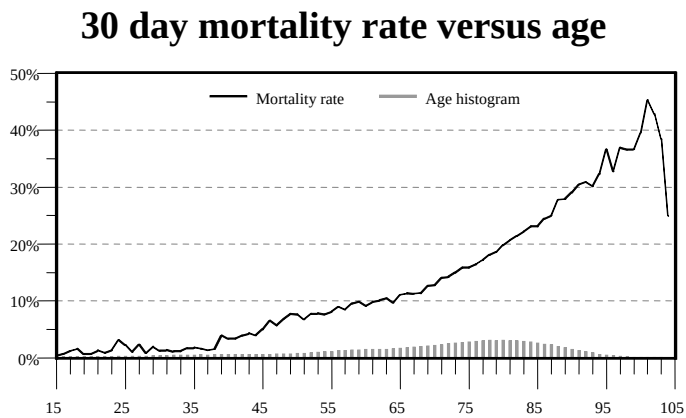


Figure 4.7 *The mortality rate against age in years with an age histogram.*

On the basis of the frequency table 4.14 for each of the diseases versus age in intervals, the first two intervals were collapsed because less than 1% of the patients were under 40 years of age and had one of the diseases.

Diseases	**15-40**	**41-65**	**66-80**	**81-110**
Myocardial infarction	0.0%	1.5%	4.4%	3.1%
Congestive heart failure	0.5%	1.3%	5.0%	5.1%
Peripheral vascular disease	0.1%	1.3%	4.3%	2.5%
Cerebrovascular disease	0.1%	2.0%	6.9%	6.4%
Dementia	0.0%	0.2%	1.3%	1.9%
Chronic pulmonary disease	0.8%	4.3%	9.8%	4.4%
Connective tissue disease	0.2%	1.1%	2.3%	1.8%
Ulcer disease	0.1%	1.8%	3.6%	3.0%
Mild liver disease	0.1%	0.9%	0.6%	0.2%
Diabetes I and II	0.1%	1.0%	2.0%	1.3%
Moderate to severe renal disease	0.2%	0.9%	1.5%	0.9%
Diabetes with end organ damage	0.1%	1.1%	2.0%	1.0%
Any tumor	0.1%	2.4%	5.7%	4.6%
Lymphoma	0.1%	0.6%	0.7%	0.3%
Metastatic solid tumor	0.1%	0.8%	1.0%	0.4%
Alcohol related disorders	0.5%	3.4%	1.7%	0.4%
History of obesity	0.2%	1.1%	1.3%	0.5%
Hypertension	0.2%	2.7%	6.3%	4.6%

Table 4.14 *Frequency table for diseases versus age intervals. Age is given in years.*

To be sure, that we did not divide the disease variable in more categories, than they had statistical power to handle, we looked at the entries in table 4.14 again. All entries in this frequency table were larger than 1% after collapsing the first two age intervals, except for the diseases dementia, mild liver disease, moderate to severe renal disease, diabetes with end organ damage, lymphoma, metastatic solid tumor, alcohol related disorder and history of obesity. For this reason interaction between age and these diseases was not included in the index.
This means that a logistic regression with the 22 diseases and interaction between each of the relevant diseases and discretized age was made. On the basis of this regression the significant interaction terms

were identified as myocardial infarction, congestive heart failure, cerebrovascular disease, chronic pulmonary disease and any tumor.

The logistic regression used to develop a new index had therefore the 22 diseases and interaction between discretized age and the diseases: myocardial infarction, congestive heart failure, cerebrovascular disease, chronic pulmonary disease and any tumor as predictors.
The Hosmer-Lemeshow test statistic for this model was 188 with 8 degrees of freedom. The area under the ROC-curve was 0.712 for this regression, which means that this model discriminated better than the model without interaction with age but still rather poorly.

The new weights were calculated from the log odds ratios like before. The parameter estimates for the diseases and their weights can be seen in table 4.15. Note that the diseases without interaction were given the same weights as in the index without interaction. For all the interaction terms, the diseases has greater influence on mortality, the younger the patient is.

Effect	**Parameter estimate**	**Weight**
Myocardial infarction, 15-65	-0.25	-3
Myocardial infarction, 66-80	-0.07	-1
Myocardial infarction, 81-110	0.03	0
Congestive heart failure, 15-65	0.37	4
Congestive heart failure, 66-80	0.30	3
Congestive heart failure, 81-110	0.17	2
Peripheral vascular disease	0.20	2
Cerebrovascular disease 15-65	0.44	4
Cerebrovascular disease 66-80	0.36	4
Cerebrovascular disease 81-110	0.19	2
Dementia	0.59	6
Chronic pulmonary disease 15-65	-0.26	-3
Chronic pulmonary disease 66-80	-0.10	-1
Chronic pulmonary disease 81-110	-0.12	-1
Connective tissue disease	-0.01	0
Ulcer disease	0.13	1
Mild liver disease	0.48	5
Diabetes I and II	0.11	1
Hemiplegia	0.37	4
Moderate to severe renal disease	0.26	3
Diabetes with end organ damage	0.24	2
Any tumor, 15-65	0.99	10
Any tumor, 66-80	0.48	5
Any tumor, 81-110	0.21	2
Leukemia	0.45	5
Lymphoma	0.07	1
Moderate to severe liver disease	0.88	9
Metastatic solid tumor	1.27	13
AIDS / HIV	0.27	3
Alcohol related disorders	0.51	5
History of obesity	-0.14	-1
Hypertension	-0.12	-1

Table 4.15 *Parameter estimates and the assigned weights.*

For each observation the appropriate weights were then added, and a new comorbidity index was made by grouping the values into the groups ≤ 0, $1-3$, $4-6$, $7-9$, $10-12$, $13-15$, ≥ 16. The Kaplan-Meier curves can be seen in figure 4.8.

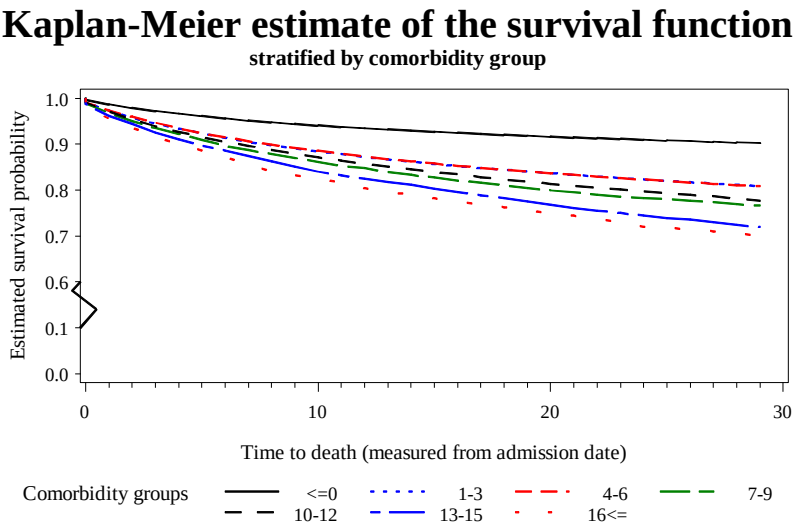


Figure 4.8 *Kaplan-Meier curves for the index with interaction with age.*

It is seen that the curves for $1-3$ and $4-6$ overlap and that the curves for $10-12$ lies above the one for $7-9$.

The logistic regression was made with the grouped index as a predictor. The Hosmer-Lemeshow test for this model can be seen table 4.16.

Group	Total	Observed deaths	Expected deaths	Deviance residuals
1	20621	280	418	-7
2	20492	866	1052	-6
3	20241	1453	1518	-2
4	20 40	2047	1982	1
5	20581	2737	2594	3
6	20727	3390	3202	3
7	20307	3938	3739	3
8	20692	4674	4497	3
9	20540	5476	5323	2
10	20606	6697	7233	-6

Table 4.16 *Hosmer-Lemeshow χ^2 test.*

The table showed that the regression did not predict death well on the lower and higher end of the probability scale.
The statistics for this regression are summarized in the bottom row in table 4.17, where the statistics from the indexes found in the previous sections also are noted for comparison.

Predictors	χ^2_{HL}	R^2	**ROC**
Crude index (19 diseases)	266	0.1007	0.705
Index (19 diseases)	257	0.1052	0.709
Index (22 diseases)	300	0.1084	0.710
Index with sex int. (22 diseases)	286	0.1072	0.711
New index with age int. (22 diseases)	197	0.1124	0.712

Table 4.17 *The goodness of fit statistics for all the developed indexes.*

The Hosmer-Lemeshow χ^2 value and the R^2 value both indicates that the performance of the new index was slightly better than the others. The area under the ROC curve showed that all indexes were poor at discriminating between outcomes.

The calibration plot for this new index was very similar to the ones for the indexes with only 19 diseases and can be seen in the appendix in figure A.3.

4.6 An index including pairwise interaction between diseases

In this section pairwise interaction between the diseases are considered. This is done in order to take into account that the effect of having 2 diseases not necessarily is as the added weights would indicate.

Not all pairwise interactions were considered as possible covariates. Before including any interaction terms a 22×22 frequency table of the diseases was made. Only the combinations of diseases with a prevalence of 1% or more, were included as covariates. There were included 42 interaction terms.

A logistic regression with the 22 diseases and the pairwise interactions as covariates was made.

Not all of the interaction terms were statistically significant, so the list of interaction terms were further shortened, by only including the 16 interaction terms that were statistically significant. The 16 remaining interaction terms can be seen in table 4.19.

A new logistic regression containing the 22 diseases and the remaining 16 interactions terms was then made.

The Hosmer-Lemeshow test for this regression gave a χ^2 value of 269 which indicated a lack of fit.
The area under the ROC curve for this regression was 0.712, which shows that the model was poor at discriminating between the outcomes.

The new weights were then calculated. They can be seen in table 4.18 and 4.19.

Effect	**Parameter estimate**	**Weight**
Myocardial infarction	0.02	0
Congestive heart failure	0.34	3
Peripheral vascular disease	0.23	2
Cerebrovascular disease	0.41	4
Dementia	0.70	7
Chronic pulmonary disease	-0.13	-1
Connective tissue disease	-0.01	0
Ulcer disease	0.19	2
Mild liver disease	0.49	5
Diabetes I and II	0.16	2
Hemiplegia	0.38	4
Moderate to severe renal disease	0.25	3
Diabetes with end organ damage	0.18	2
Any tumor	0.53	5
Leukemia	0.45	5
Lymphoma	0.08	1
Moderate to severe liver disease	0.89	9
Metastatic solid tumor	1.24	12
AIDS/HIV	0.26	3
Alcohol related disorders	0.58	6
History of obesity	-0.14	-1
Hypertension	-0.14	-1

Table 4.18 *The parameter estimates and the assigned weights.*

Effect	**Parameter estimate**	**Weight**
Myocardial infarctions and Congestive heart failure	-0.17	-2
Myocardial infarctions and Any tumor	-0.14	-1
Myocardial infarctions and Hypertension	0.10	1
Congestive heart failures and Cerebrovascular disease	-0.18	-2
Congestive heart failures and Any tumor	-0.20	-2
Peripheral vascular diseases and Cerebrovascular disease	-0.14	-1
Peripheral vascular diseases and Diabetes with end organ damage	0.20	2
Cerebrovascular diseases and Dementia	-0.30	-3
Cerebrovascular diseases and Ulcer disease	-0.13	-1
Cerebrovascular diseases and Any tumor	-0.20	-2
Chronic pulmonary diseases and Alcohol related disorders	-0.16	-2
Chronic pulmonary diseases and Hypertension	0.07	1
Ulcer diseases and Any tumor	-0.14	-1
Ulcer diseases and Alcohol related disorders	-0.21	-2
Ulcer diseases and Hypertension	0.09	1
Diabetes I and II and Hypertension	-0.18	-2

Table 4.19 *The parameter estimates and the assigned weights for the pairwise interaction terms.*

The weights of the appropriate diseases and interactions were added and grouped into ≤ 0, 1-3, 4-6, 7-9, 10-12, ≥ 13. The Kaplan-Meier curves for this index are seen in figure 4.9. The Kaplan-Meier curves show that the index separate the comorbidity groups even though group 7-9 and 10-12 lay very close.

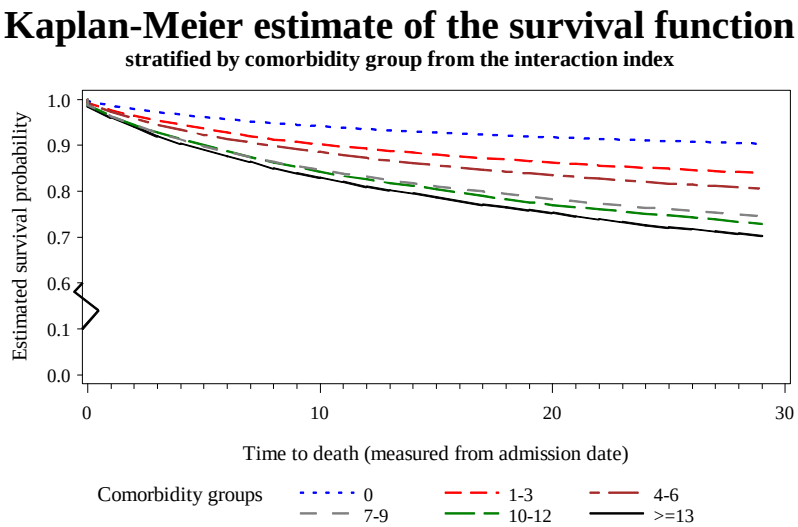


Figure 4.9 *Kaplan-Meier curves for the index with pairwise interaction between the diseases.*

A logistic regression with the new index as a predictor was made and the overall test statistics were calculated.
The Hosmer-Lemeshow test showed that the index including pairwise interaction between the diseases, also was very bad at predicting the number of deaths in the low probability groups as well as in the highest group. The Hosmer-Lemeshow test is seen in table 4.20.

Group	Total	Observed deaths	Expected deaths	Deviance residuals
1	20426	279	423	-7
2	20521	885	1100	-6
3	20266	1472	1598	-3
4	20350	2093	2068	1
5	20894	2776	2626	3
6	20495	3342	3099	4
7	20621	4007	3702	5
8	20532	4640	4392	4
9	20536	5525	5347	2
10	20206	6539	7203	-8

Table 4.20 *Hosmer-Lemeshow χ^2 test.*

A comparison of the overall test statistics for the new index and the indexes from the previous sections is seen in table 4.21.

Predictors	χ^2_{HL}	R^2	ROC
Crude index (19 diseases)	266	0.1007	0.705
Index (19 diseases)	257	0.1052	0.709
Index (22 diseases)	300	0.1084	0.710
Index with sex int. (22 diseases)	286	0.1072	0.711
Index with age int. (22 diseases)	197	0.1124	0.712
Index including int. between diseases	290	0.1072	0.711

Table 4.21 *Comparison of the indexes.*

The Hosmer-Lemeshow χ^2 value and the R^2 value both indicates that there was no difference in the performance of the index with interaction and the index (22 diseases). The area under the ROC curve showed that these indexes were equally poor at discriminating between outcomes.

The calibration plot for this index was similar to the ones for the indexes with 19 diseases, and it can be seen in figure A.4 in the appendix.

4.7 An index including 22 diseases and all first degree interaction terms

In the previous sections different interactions were examined, and in this section a model is made including all the previously found interaction terms.

This means that a logistic regression model containing, the 22 diseases, interactions between sex and myocardial infarction, any tumor and metastatic solid tumor, between age and myocardial infarction, congestive heart failure, cerebrovascular disease, chronic pulmonary disease and any tumor, and containing the 16 pairwise interactions between diseases was made.

The Hosmer-Lemeshow χ^2 value for this regression was 170 with 8 degrees of freedom. The area under the ROC-curve was 0.713 which shows that the model is poor at discriminating between outcomes.

After fitting the model the weights were calculated and the index was made as in the previous sections with the groups ≤ 0, 1-3, 4-6, 7-9, 10-12, 13-15, ≥ 16. The assigned weights can be seen in table 4.22 and 4.23.

Effect	**Parameter estimate Male/Female**	**Weight Male/Female**
Myocardial infarction, 15-65	-0.24/-0.08	-2 /-1
Myocardial infarction, 66-80	-0.06/0.11	-1/1
Myocardial infarction, 81-110	0.03/0.20	0/2
Congestive heart failure, 15-65	0.46	5
Congestive heart failure, 66-80	0.40	4
Congestive heart failure, 81-110	0.27	3
Peripheral vascular disease	0.23	2
Cerebrovascular disease 15-65	0.54	5
Cerebrovascular disease 66-80	0.48	5
Cerebrovascular disease 81-110	0.31	3
Dementia	0.69	7
Chronic pulmonary disease 15-65	-0.24	-2
Chronic pulmonary disease 66-80	-0.10	-1
Chronic pulmonary disease 81-110	-0.12	-1
Connective tissue disease	-0.01	0
Ulcer disease	0.19	2
Mild liver disease	0.49	5
Diabetes I and II	0.16	2
Hemiplegia	0.37	4
Moderate to severe renal disease	0.25	3
Diabetes with end organ damage	0.18	2
Any tumor, 15-65	1.07/1.00	11/10
Any tumor, 66-80	0.59/0.52	6/5
Any tumor, 81-110	0.34/0.27	3/3
Leukemia	0.46	5
Lymphoma	0.07	1
Moderate to severe liver disease	0.89	9
Metastatic solid tumor	1.36/1.17	14/12
AIDS/HIV	0.26	3
Alcohol related disorders	0.57	6
History of obesity	-0.14	-1
Hypertension	-0.14	-1

Table 4.22 *The parameter estimates and the assigned weights.*

Effect	**Parameter estimate**	**Weight**
Myocardial infarction and Congestive heart failure	-0.20	-2
Myocardial infarction and Any tumor	-0.14	-1
Myocardial infarction and Hypertension	0.09	1
Congestive heart failure and Cerebrovascular disease	-0.16	-2
Congestive heart failure and Any tumor	-0.11	-1
Peripheral vascular disease and Cerebrovascular disease	-0.16	-2
Peripheral vascular disease and Diabetes with end organ damage	0.19	2
Cerebrovascular disease and Dementia	-0.28	-3
Cerebrovascular disease and Ulcer disease	-0.13	-1
Cerebrovascular disease and Any tumor	-0.14	-1
Chronic pulmonary disease and Alcohol related disorders	-0.13	-1
Chronic pulmonary disease and Hypertension	0.06	1
Ulcer disease and Any tumor	-0.14	-1
Ulcer disease and Alcohol related disorders	-0.21	-2
Ulcer disease and Hypertension	0.09	1
Diabetes I and II and Hypertension	-0.19	-2

Table 4.23 *The parameter estimates and the assigned weights.*

The Kaplan-Meier curves for this index is seen in figure 4.10.

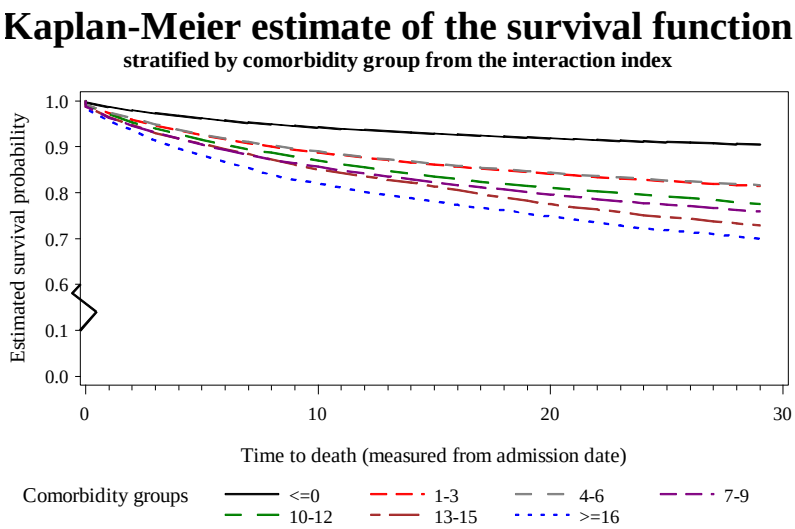


Figure 4.10 *Kaplan-Meier curves for the new index based on a logistic regression including all first degree interaction terms.*

The Kaplan-Meier curves show that this index is rather poor at separating the different comorbidity groups, since the curves for 1-3 and 4-6 are overlapping.

The index was then included in a logistic regression. The Hosmer-Lemeshow test for this regression is seen in table 4.24.

Group	Total	Observed deaths	Expected deaths	Deviance residuals
1	20287	274	406	-7
2	20450	844	1027	-6
3	20479	1452	1520	-2
4	20458	2099	2029	2
5	20415	2752	2586	3
6	20483	3337	3174	3
7	20254	3855	3734	2
8	20530	4741	4443	4
9	20443	5337	5262	1
10	21048	6867	7377	-6

Table 4.24 *Hosmer-Lemeshow χ^2 test.*

It is seen that this index, including all interaction terms also is bad at predicting death in the low probability groups as well as in the highest group. This is supported by the χ^2 value.

In table 4.25 a comparison of the new index and the indexes from previous sections is made.

Predictors	χ^2_{HL}	R^2	ROC
Crude index (19 diseases)	266	0.1007	0.705
Index (19 diseases)	257	0.1052	0.709
Index (22 diseases)	300	0.1084	0.710
Index with sex int. (22 diseases)	286	0.1072	0.711
Index with age int. (22 diseases)	197	0.1124	0.712
Index including int. between diseases	290	0.1072	0.711
Index including all interaction terms	192	0.1139	0.712

Table 4.25 *Comparison of the indexes.*

Table 4.25 show that the index including all interaction terms performs better than the other indexes, since the χ^2 value is smaller and both the R^2 value and the area under the ROC-curve is larger for this index.

Again the calibration plot was similar to the plots for the indexes with 19 diseases, and it can be seen in figure A.5 in the appendix.

4.8 An index including time since diagnosis

In this section we take the time since diagnosis into account in the index. This is done because some diseases, like diabetes, is known to get worse as time passes, so the mortality may increase with the time since diagnosis. Other diseases, like any tumor, may be cured after some time, so the diagnosis no longer have any influence on the mortality. Including the time since diagnosis can be seen as an attempt to include the severity of a disease, which is not registered in the Danish national patient registry.
We defined 'time since diagnosis' as the number of years since the first diagnosis. The first diagnosis was chosen because most of the diseases are chronic diseases, so the time for the first diagnosis reflects the time the patient got the disease. For the diseases that might be cured like leukemia it is not possible to see in the Danish national patient registry if a diagnosis is for a new diseases or a relapse, so using the time for the first diagnosis is a reasonable approximation of the time, the patient got the disease.

The variable 'time since diagnosis' is defined as shown in table 4.26.

Category	**label**
No diagnosis	'No diagnosis'
First diagnosis less than 1 year ago	'<1 year'
First diagnosis 1 or more years ago and less than 5 years ago	'1-4 years'
First diagnosis 5 or more years ago and less then 10 years ago	'5-10 years'
First diagnosis 10 or more years ago	'≥10 years'

Table 4.26 *The definition of the categories for the variable 'time since diagnosis'.*

To test if there is any new information in this variable, a logistic regression with all the normal discrete disease variables and the new 'time since diagnosis' variables was made.
For the eight diseases (dementia, connective tissue disease, diabetes I and II, hemiplegia, diabetes with end organ damage, lymphoma, AIDS / HIV and history of obesity) the 'time since diagnosis' was not signif-

icant, so these were modeled be the normal binary variable. The other 14 diseases had a significant 'time since diagnosis' variable, so these were therefore modeled by this new variable.

In the regression with the 14 diseases modeled by 'time since diagnosis' and the 8 diseases modeled by the normal discrete variable not all the values of the 'time since disease' variable were significant. The variables with one or more insignificant values are shown in table 4.27.

The categories in table 4.27 with a p-value higher than 0.05 were collapsed with the reference category, 'No diseases'. This was done since a high p-value showed, that the estimated parameter for the category was not significantly different from zero.
This meant, that myocardial infarction, moderate to severe renal disease and leukemia was described by a 'time since diagnosis' variable with the categories: 'No diagnosis or $5\geq$ years', '<1 year' and '1-4 years'. For chronic pulmonary disease the categories were 'no diagnosis or <1 year', '1-4 years', '5-10 years' and '≥ 10 years'. For moderate to severe liver disease and hypertension the categories were 'no diagnosis or ≥ 10 years', '<1 year', '1-4 years' and '5-10 years'.

With these modifications the index was made on the basis of a logistic regression containing the 14 'time to diagnosis' variables, the 8 normal discrete variables and with 30 day mortality as outcome.

The Hosmer-Lemeshow test statistic for this model was 263 with 8 degrees of freedom. The area under the ROC-curve was 0.718 for this regression, which means that the model discriminated better than the models without the 'time since diagnosis' variable but still rather poorly.

Effect	**Parameter estimates**	***p*-value for 'time since diagnosis'**
Myocardial infarction		
<1 year	-0.10	0.0174
1-4 years	-0.14	0.0007
5-9 years	-0.03	0.5430
≥10 years	-0.06	0.1010
Chronic pulmonary disease		
<1 year	0.02	0.5327
1-4 years	-0.18	<.0001
5-9 years	-0.10	0.0012
≥10 years	-0.15	<.0001
Moderate to severe renal disease		
<1 year	0.55	<.0001
1-4 years	0.18	0.0008
5-9 years	0.07	0.4000
≥10 years	-0.10	0.2356
Leukemia		
<1 year	0.84	<.0001
1-4 years	0.33	0.0007
5-9 years	0.15	0.2484
≥10 years	0.23	0.2433
Moderate to severe liver disease		
<1 year	1.35	<.0001
1-4 years	0.85	<.0001
5-9 years	0.59	0.0010
≥10 years	0.29	0.0989
Hypertension		
<1 year	-0.18	<.0001
1-4 years	-0.16	<.0001
5-9 years	-0.11	0.0026
≥10 years	0.01	0.6840

Table 4.27 *Parameter estimates and the corresponding p-value for the variable 'time since diagnosis'.*

The new weights were then calculated from the log odds ratios and they can be seen in table 4.28. Note that diabetes I and II and diabetes with end organ damage both had insignificant 'time since diagnosis'. This may be because the division of diabetes already takes the severity of diabetes into account. Note also, that the uneven tendency for cerebrovascular disease and chronic pulmonary disease may be enhanced by the round-off.

Effect	**Weight for the years**			
	<1	1-5	5-10	≥ 10
Myocardial infarction	1	-1	-	-
Congestive heart failure	4	2	2	2
Peripheral vascular disease	4	2	2	1
Cerebrovascular disease	4	2	3	3
Dementia	6			
Chronic pulmonary disease	-	-2	-1	-2
Connective tissue disease	0			
Ulcer disease	3	1	1	1
Mild liver disease	11	5	3	3
Diabetes I and II	1			
Hemiplegia	4			
Moderate to severe renal disease	5	2	-	-
Diabetes with end organ damage	3			
Any tumor	9	4	2	2
Leukemia	8	3	-	-
Lymphoma	1			
Moderate to severe liver disease	13	8	6	-
Metastatic solid tumor	17	10	5	3
AIDS / HIV	2			
Alcohol related disorders	7	6	5	4
History of obesity	-1			
Hypertension	-2	-2	-1	-

Table 4.28 *Parameter estimates and the assigned weights.*
'-' means that the interval is not defined for the disease.

For each observation the appropriate weights were then added and grouped like earlier into the groups ≤ 0, $1-3$, $4-6$, $7-9$, $10-12$, $13-15$, ≥ 16. Kaplan-Meier curves were made and they can be seen in figure 4.11.

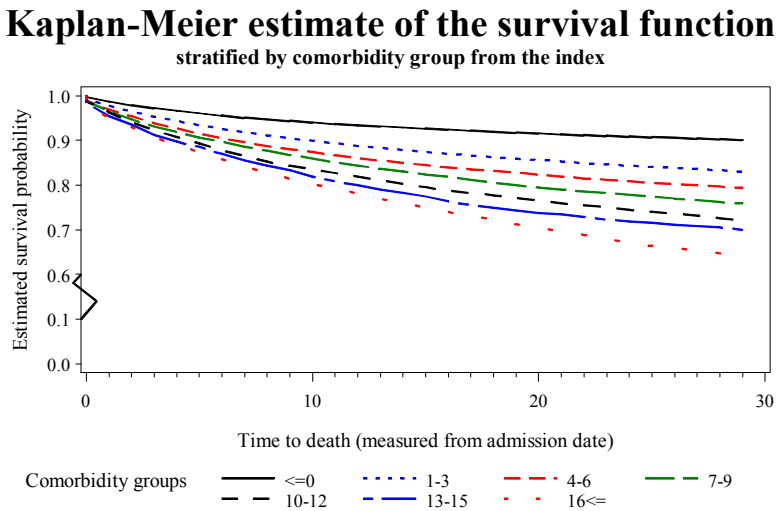


Figure 4.11 *Kaplan-Meier curves for the index with 'time since diagnosis'.*

It is seen that after approximately 2 days the different comorbidity groups are nicely separated.

To assess the applicability of the index, a logistic regression was made with the index as a predictor. The Hosmer-Lemeshow test for this model can be seen in table 4.29.

Group	Total	Observed deaths	Expected deaths	Deviance residuals
1	20421	270	415	-7
2	20310	867	1056	-6
3	20479	1417	1570	-4
4	20404	2032	2020	0
5	20206	2615	2477	3
6	20472	3171	3036	2
7	20871	4028	3704	5
8	20283	4671	4310	5
9	20513	5488	5336	2
10	20888	6999	7634	-7

Table 4.29 *Hosmer-Lemeshow χ^2 test.*

The table showed that the regression did not predict death very well on the lower and higher end of the probability scale.
The statistics for this regression are summarized in the bottom row in table 4.30, where the statistics from all of the indexes found in previous sections also are noted for comparison.

Predictors	χ^2_{HL}	R^2	**ROC**
Crude index (19 diseases)	266	0.1007	0.705
Index (19 diseases)	257	0.1052	0.709
Index (22 diseases)	300	0.1084	0.710
Index with sex int. (22 diseases)	286	0.1072	0.711
Index with age int. (22 diseases)	197	0.1124	0.712
Index including int. between diseases	290	0.1072	0.711
Index including all interaction terms	192	0.1139	0.712
Index with 'time since diagnosis'	281	0.1155	0.717

Table 4.30 *The goodness of fit statistics for some of the developed indexes. The index in the bottom row is the index with the variable 'time since diagnosis'*

From the higher R^2 value and the higher area under the ROC curve, it is seen that the index with 'time since diagnosis' performs better than the other indexes listed in table 4.30.

The calibration plot for this index was also similar to the plots for the indexes with 19 diseases, and it can be seen in figure A.6 in the appendix.

4.9 Index including time since diagnosis and interaction

In this section we extend the index from section 4.8 by also including first degree interaction terms. We do this in an attempt to model all available information in the training data.
Like in section 4.8 we use the 'time since diagnosis' variable on the 14 diseases, where it was significant, and the normal binary variable for the other 8 diseases. The 'time since diagnosis' variable is defined as in section 4.8.

We started this index development with three initial investigations in order to find the diseases that interact with sex, age and other diseases respectively. Then we included the significant interactions in a model and used this model to develop an index.

First we investigated which diseases had significant interaction with sex. This was done by identifying all the diseases where the frequency of females and males with the disease were more than 1%. The identified diseases were congestive heart failure, cerebrovascular disease, dementia, chronic pulmonary disease, connective tissue disease, diabetes I and II, diabetes with end organ damage, any tumor, history of obesity and hypertension. A logistic regression with interaction between these diseases and sex was made. The regression also included the 14 diseases modeled by the 'time since diagnosis' variable and the 8 diseases modeled by the normal binary diseases variable. Only one disease, congestive heart failure had significant interaction with sex.

In the same way as with sex we investigated which diseases had a significant interaction with age. We used the same discretization of age as in section 4.5. The diseases with frequencies larger than 1% for the age intervals 15-65 years, 66-80 years and over 80 years were identified as chronic pulmonary disease, connective tissue disease and diabetes I and II. The interaction terms between these diseases and the dis-

cretized age were included in a regression like before. The disease with significant interaction with age was chronic pulmonary disease.

We then investigated if any interaction among pairs of diseases were significant. None of the pairs of diseases where one or both of the diseases were modeled by the 'time since diagnosis' variable had frequencies over 1%. For this reason we only looked at interaction between the binary variables for the diseases. The 42 pairs of diseases with frequencies over 1% were included in the regression like before. The 14 significant pairs can be seen in table 4.32.

The model used to develop an index contained: 14 'time since diagnosis' variables, 8 normal binary disease variables, interaction between sex and congestive heart failure, interaction between the discretized age variable and chronic pulmonary disease and interaction between the pairs of diseases in table 4.32.

The Hosmer-Lemeshow test statistic for this model was 263 with 8 degrees of freedom. The area under the ROC-curve was 0.719 for this regression, which means that the model discriminated rather poorly like the model with only the 'time since diagnosis' variable. A Wald test showed that two of the diseases (Connective tissue disease and AIDS / HIV) were statistically insignificant, but since all of the diseases were of interest none of them were left out. The interaction terms between chronic pulmonary disease and alcohol related disorder and between ulcer disease and hypertension were also insignificant. They were kept in the model because the initial investigation showed significance.

The new weights were then calculated from the log odds ratios and they can be seen in table 4.31 and 4.32. Note that the uneven tendencies for congestive heart failure, peripheral vascular disease and chronic pulmonary disease may be enhanced by the round-off. Note also that any tumor has remarkably larger weights compared to the weight in the index with only 'time to diagnosis'.

Effect	**Weight for the values**			
	<1	1-5	5-10	≥10
Myocardial infarction	2	-1	-	-
Congestive heart failure, female	4	3	2	3
Congestive heart failure, male	5	2	3	2
Peripheral vascular disease	4	2	3	1
Cerebrovascular disease	5	4	4	4
Dementia	7			
Chronic pulmonary disease and				
15 to 65 years of age	-	-3	0	-3
66 to 80 years of age	-	-1	0	-2
over 80 years of age	-	-2	-1	0
Connective tissue disease	0			
Ulcer disease	4	2	2	1
Mild liver disease	11	5	3	4
Diabetes I and II	2			
Hemiplegia	4			
Moderate to severe renal disease	5	2	-	-
Diabetes with end organ damage	2			
Any tumor	15	10	3	3
Leukemia	8	3	-	-
Lymphoma	1			
Moderate to severe liver disease	13	9	6	-
Metastatic solid tumor	17	11	5	3
AIDS / HIV	2			
Alcohol related disorders	7	7	6	5
History of obesity	-1			
Hypertension	-2	-2	-1	-

Table 4.31 *Parameter estimates and the assigned weights.*
'-' means that the interval is not defined for the disease.

Combinations	**Weight**
Myocardial infarction and Congestive heart failure	-1
Myocardial infarction and Any tumor	-2
Myocardial infarction and Hypertension	1
Congestive heart failure and Cerebrovascular disease	-2
Peripheral vascular disease and Cerebrovascular disease	-1
Peripheral vascular disease and Diabetes with end organ damage	2
Cerebrovascular disease and Dementia	-3
Cerebrovascular disease and Ulcer disease	-1
Cerebrovascular disease and Any tumor	-2
Chronic pulmonary disease and Alcohol related disorders	-1
Ulcer disease Any tumor	-1
Ulcer disease Alcohol related disorders	-2
Ulcer disease Hypertension	1
Diabetes I and II Hypertension	-2

Table 4.32 *The assigned weights for combinations of diseases.*

For each observation the appropriate weights were then added and a new comorbidity index was made by grouping the values like earlier into the groups ≤ 0, $1-3$, $4-6$, $7-9$, $10-12$, $13-15$, ≥ 16. Kaplan-Meier curves were made and they can be seen in figure 4.12.

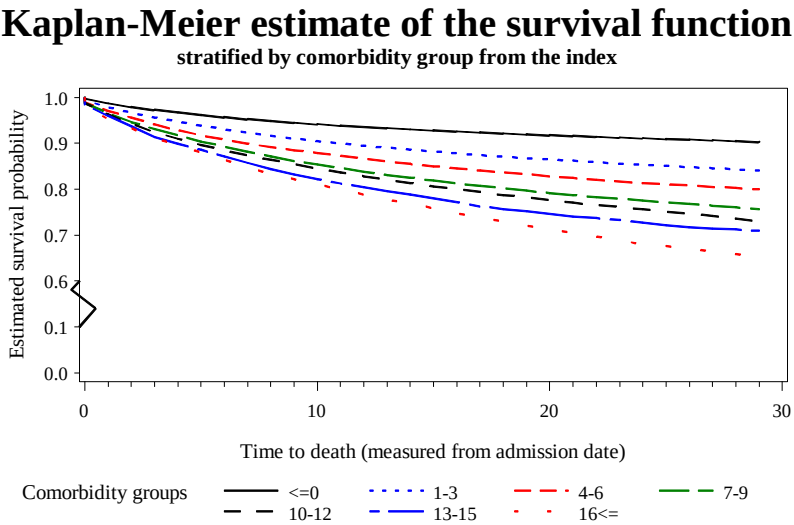


Figure 4.12 *Kaplan-Meier curves for the index with 'time since diagnosis' and first degree interactions.*

It is seen that the curves for the different comorbidity groups are nicely separated.

To assess the applicability of the index, a logistic regression was made with the new index as a predictor. The Hosmer-Lemeshow test for this model can be seen in table 4.33.

Group	Total	Observed deaths	Expected deaths	Deviance residuals
1	20484	275	415	-7
2	20251	863	1047	-6
3	20587	1416	1564	-4
4	21042	2078	2083	0
5	20181	2630	2489	3
6	20443	3236	3056	3
7	20517	3998	3664	6
8	20592	4667	4402	4
9	20475	5569	5372	3
10	20275	6826	7466	-7

Table 4.33 *Hosmer-Lemeshow χ^2 test.*

The table showed that the regression did not predict death very well in the lower and higher end of the probability scale.
The statistics for this regression are summarized in the bottom row in table 4.34, where the statistics from the indexes found in previous sections also are noted for comparison.

Predictors	χ^2_{HL}	R^2	**ROC**
Crude index (19 diseases)	266	0.1007	0.705
Index (19 diseases)	257	0.1052	0.709
Index (22 diseases)	300	0.1084	0.710
Index with sex int. (22 diseases)	286	0.1072	0.711
Index with age int. (22 diseases)	197	0.1124	0.712
Index including int. between diseases	290	0.1072	0.711
Index including all interaction terms	192	0.1139	0.712
Index with 'time since diagnosis'	281	0.1155	0.717
Index with 'time since diagnosis' and int.	273	0.1166	0.718

Table 4.34 *The goodness of fit statistics for all the developed indexes. The index in the bottom row is the index with the variable 'time since diagnosis' and interaction terms.*

From the higher R^2 value and the higher area under the ROC curve, it is seen that the index with 'time since diagnosis' and interactions performs slightly better than the other indexes listed in table 4.34.
Again the calibration curve behaves as the ones for the indexes with 19 diseases and it can be seen in the appendix in figure A.7.

4.10 Summary

In this section a summary of all the results and indexes found in this chapter is given.

Comparing the weights from the crude logistic regression and the weights from the regression adjusted for sex and age, it was clear, that this adjustment was important. Only four out of nineteen diseases got the same weight in the two regressions.
The weights obtained from the different regressions adjusted for sex and age were very similar, but when including the three additional diseases in the index it was seen that the weights for the liver diseases changed. This may be caused by the fact that alcohol related disorders and liver diseases are related.

The weight for myocardial infarction was zero in the indexes were it was modeled only by the discrete disease variable without any interactions. The disease had significant interaction with sex, age and three diseases, when it was modeled by the discrete disease variable, and it had a significant 'time since diagnosis' variable with the intervals '<1' and '1-5 years'. This indicated, that myocardial infarction is a disease with a complex effect and that it is related to both sex, age and other factors.

The Wald tests for the diseases often showed insignificance for myocardial infarction, connective tissue disease, lymphoma and AIDS / HIV. Changing the definition of these diseases or reconsidering their clinical relevance is possibly needed to change this result.

The Kaplan-Meier curves shown in this chapter all had nicely separated curves when interaction between age and the diseases was not included in the index. For the indexes with interaction with age some curves overlapped, and for the index with the 22 diseases and interaction with age the curve for comorbidity group 10-12 lay under the one for 7-8.

The calibration curves and Hosmer-Lemeshow tests illustrated nicely, that the indexes predicted death equally well.All indexes had difficul-

ties in the probability range with few observations.

The test statistics for the indexes can be seen in table 4.35. The goodness-of-fit statistics for the indexes were also similar, however the small differences are of interest since our dataset contains so many observations. The over all tendency in the table is that the indexes perform better, the more complex they become.

Predictors	χ^2_{HL}	R^2	**ROC**
Crude index (19 diseases)	266	0.1007	0.705
Index (19 diseases)	257	0.1052	0.709
Index (22 diseases)	300	0.1084	0.710
Index with sex int. (22 diseases)	286	0.1072	0.711
New index with age int. (22 diseases)	197	0.1124	0.712
Index including int. between diseases	290	0.1072	0.711
Index including all interaction terms	192	0.1139	0.712
Index with 'time since diagnosis'	281	0.1155	0.717
Index with 'time since diagnosis' and int.	273	0.1166	0.718

Table 4.35 *The goodness of fit statistics for all the developed indexes.*

Index development using naive Bayes

In this chapter the process of developing an index with naive Bayes is described, and the performance is assessed. When using the naive Bayes method it is not possible to adjust for sex and age, so instead we included these in the index. Because a simpler index is often to prefer we also made an index without age.

Using naive Bayes to develop a comorbidity index was done by first estimating the conditional probability function

$$\text{logit}(P(Y = 1|X = x)) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i \tag{5.1}$$

as described in chapter 3 in part I. Then weights were found from the log odds ratios like in chapter 4 and the index value was computed by summing the weights where the variable $x_i = 1$.

To estimate function (5.1) the training dataset described in chapter 2 was used. The variables in the training dataset were the binary variables for the 22 diseases, the binary sex variable and the continuous age variable given in years. The continuous age variable was discretized, since this has been shown to improve the index as described in chapter 3 in part I. The clinical relevant discretization used in section 4.5 was also used here. A quick look at the mortality rate against age seen in figure 4.7 showed, that these intervals seemed to separate age nicely in intervals with different mortality rates. To describe the 30 day mortality in the best way, we divided the last interval in the two intervals, [81-90] and [91-110]. This extra division was made because the mortality rate increased notably between 80 and 105 years. When calculating the log odds ratios for age, the interval, [66-80] was chosen as the reference interval, since this is the interval with the largest number of patients.
The weights found by multiplying the log odds ratios with 10 and rounding off to the nearest integer can be seen in table 5.1. Summing the weights where the variable $x_i = 1$ and grouping into the intervals ≤ 0, 1-5, 6-10, 11-15, 16-20, ≥ 21 then gave the index value.

Description	Log odds ratio	Weight
Myocardial infarction	0.24	2
Congestive heart failure	0.56	6
Peripheral vascular disease	0.42	4
Cerebrovascular disease	0.62	6
Dementia	1.01	10
Chronic pulmonary disease	-0.11	-1
Connective tissue disease	0.07	1
Ulcer disease	0.38	4
Mild liver disease	0.30	3
Diabetes I and II	0.21	2
Hemiplegia	0.20	2
Moderate to severe renal disease	0.34	3
Diabetes with end organ damage	0.33	3
Any tumor	0.54	5
Leukemia	0.33	3
Lymphoma	0.06	1
Moderate to severe liver disease	0.62	6
Metastatic solid tumor	0.94	9
AIDS/HIV	-0.87	-9
Alcohol related disorder	0.13	1
History of obesity	-0.22	-2
Hypertension	0.19	2
Sex	0.13	1
Age group 15-40	-2.34	-23
Age group 41-65	-0.69	-7
Age group 81-90	0.54	5
Age group 91-110	0.95	10

Table 5.1 *Weights given to the 22 diseases, sex and age intervals with naive Bayes.*

Note that both dementia, metastatic solid tumor and being older than 90 years were given very high weights. Having chronic pulmonary disease, AIDS/HIV, history of obesity and being between 15 and 65 years resulted in negative weights, so patients with these diseases or this age had a lower mortality than the patients without.

We also made an index with naive Bayes excluding age. This index resulted in the same weights for the 22 diseases and sex because of the strong independence assumption.

Evaluation of the index from naive Bayes

In this section we call the index with the 22 diseases, sex and age, the extended index, and the index without age, the simple index.
To get a visual impression of how well the indexes separated the groups of patients with different mortality rates, Kaplan-Meier curves were made for both indexes. They can be seen in figure 5.1.

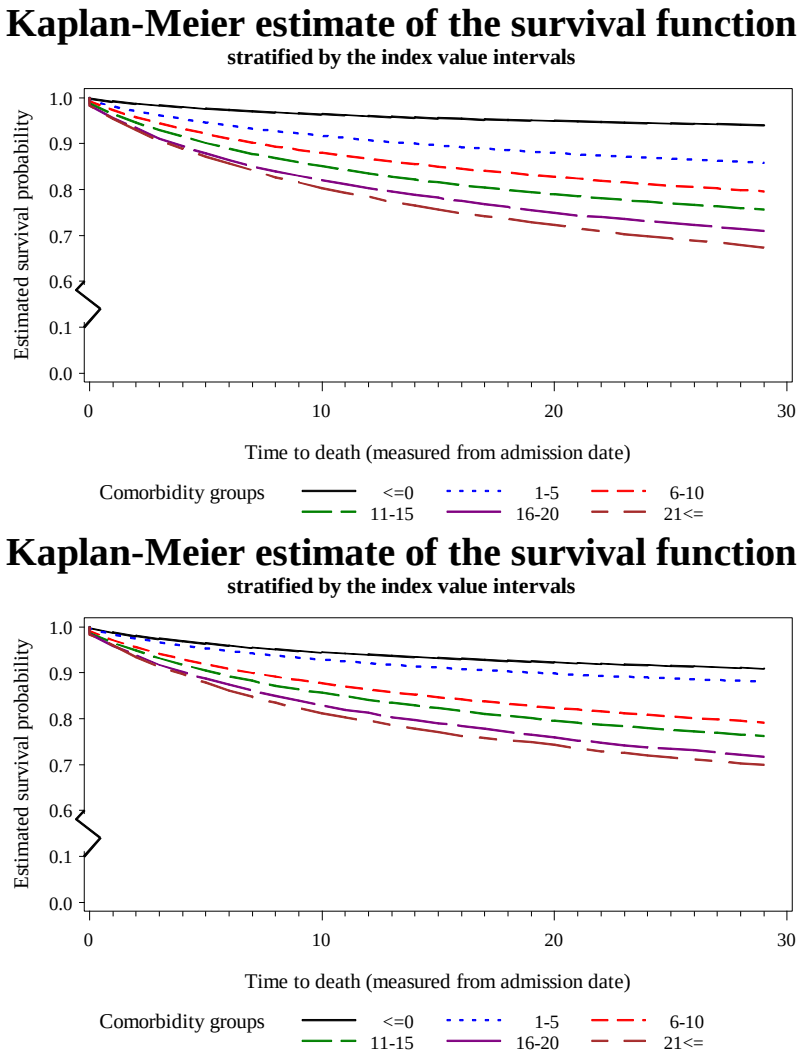


Figure 5.1 *Kaplan-Meier curves for the indexes. The upper plot is for the extended index (the 22 diseases, sex and age) and the lower is for the simple index (the 22 diseases and sex).*

The Kaplan-Meier curves showed, that both indexes separate well, since the curves almost have no overlap except during the first 5 to 6 days. The highest and lowest curves are slightly more extreme for the extended index than for the simple index.

To evaluate the predictive performance of these indexes, a logistic regression model was made for each index with adjustment for sex and for age modeled as a cubic spline was made. These logistic regression models gave the results seen in table 5.2 and 5.3.

Group	Total	Observed deaths	Expected deaths	Deviance residuals
The extended index				
1	20322	406	471	-3
2	20339	1180	1266	-2
3	19998	1526	1544	0
4	20693	2219	1932	7
5	20857	2612	2784	-3
6	20137	3271	3226	1
7	20641	4007	3841	3
8	20263	4374	4469	-1
9	20431	5242	5253	0
10	21166	6721	6773	-1
The simple index				
1	20578	361	467	-5
2	20641	1031	1220	-5
3	20488	1662	1731	-2
4	20629	2120	2170	-1
5	20787	2877	2661	4
6	20131	3295	3088	4
7	20540	3922	3697	4
8	20380	4603	4338	4
9	20503	5323	5289	0
10	20170	6364	6896	-6

Table 5.2 *Hosmer-Lemeshow χ^2 test.*

Table 5.2 showed that both regressions in general predicted death poorly.

Index	χ^2_{HL}	R^2	ROC
The extended index	88	0.0942	0.697
The simple index	197	0.0965	0.700

Table 5.3 *Summary of the logistic regression models with the indexes as predictor adjusted for sex and age.*

Table 5.3 shows that the simple index performed slightly better than the extended index.

To see how well the indexes predict death within 30 days, the observed and the predicted probabilities were plotted against one another.

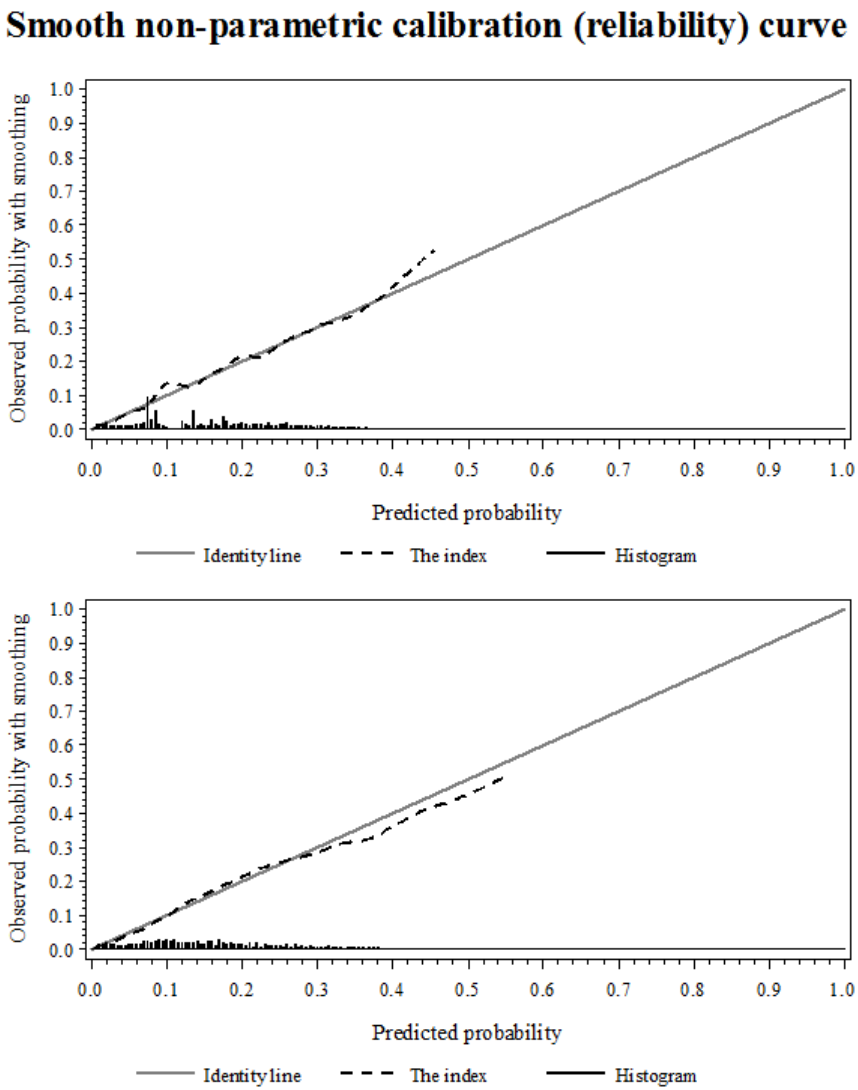


Figure 5.2 *Observed and predicted probabilities for the indexes. The upper plot is for the extended index and the lower plot is for the simple index.*

Figure 5.2 show that both indexes predicts the probability of dying within 30 days fairly. The extended index has a bump near the predicted probability of 0.1, and the simple index differs in the high end of the probability scale. This is confirmed by the Hosmer-Lemeshow test in table 5.2.

Discussion

Because of the unrealistic independence assumptions in naive Bayes, the individual weights should be interpreted with care. A weight does not describe causes, but associations.
The fact that some diseases have a large weight, is not necessary due to the fact, that the disease causes a high mortality risk. It can be due to populational differences between the patients with the disease and the patients without.
A negative weight can be due to differences in clinical procedures and might not be because the disease it-self causes a low mortality risk.

The Kaplan-Meier curves indicates, that the extended index is the better one, since this index has a slightly better separation.
The performance of the indexes in logistic regression models are quite similar. The higher R^2 and area under the ROC curve for the simple index indicates, that this is the better index. The result from the regression indicates therefore, that the simple index is the better choice.

Since the Kaplan-Meier curves and the performance of the logistic regression models give contradictory results, the overall conclusion is that both the indexes are good. Further investigation is needed, to give a final conclusion.

Index development using classification trees

In this chapter the classification tree method is used to develop indexes. In classification trees it is not possible to adjust directly for sex and age, so to take these effects into account we grew a tree including sex and age.
Since age was the most dominating predictor of death, a tree was also grown without it in order to get as much information out of the diseases as possible.

The classification trees was grown using the rpart procedure in the R program. The rpart procedure was written based on [?] and the theory behind it can be seen in chapter 4 in part I.

Classification tree including age

A classification tree including the 22 diseases, sex and age was grown. The tree was restricted so that it would not consider a split if the node contained less than 320 observations. This was done because a classification based on less than 320 observations would be very uncertain and because higher than 320 observations resulted in a tree containing only the root. The complexity parameter α was set to zero since we wanted as large a tree as possible with the above restrictions.
This resulted in the tree seen in figure 6.1.

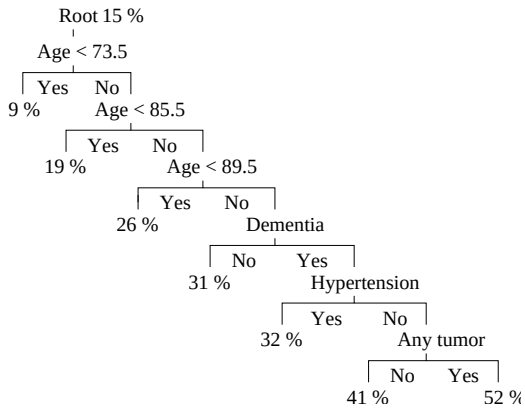


Figure 6.1 *Classification tree with the 22 diseases, sex and age as covariates, with $\alpha = 0$ and $minsplit = 320$. At the root and each terminal node the estimated probability of death is given.*

The log odds ratio for each terminal node was calculated by:

$$\log[\text{Odds for the terminal node/Odds for the root node}]. \qquad (6.1)$$

This log odds ratio compares subjects having the given covariate pattern with all the subjects in the population.

The index was calculated by multiplying the log odds ratio by 10 and then rounding to the nearest integer.
The covariate patterns, their corresponding log odds ratios, the weights and the proportion of observations in each pattern are seen in table 6.1.

Covariate pattern	**Log odds ratio**	**Weights**	**% obs.**
age<73.5	-0.63	-6	50.4
73.5≤age<85.5	0.28	3	35.8
85.5≤age<89.5	0.66	7	8.4
age≥89.5, Dementia=no	0.89	9	6.0
age≥89.5, Dementia=yes, Hypertension=yes	0.94	9	0.1
age≥89.5, Dementia=yes, Hypertension=no, Any tumor=no	1.31	13	0.3
age≥89.5, Dementia=yes, Hypertension=no, Any tumor=yes	1.80	18	0.1

Table 6.1 *The resulting covariate patterns from the classification tree, their weights and the proportion of observations.*

Table 6.1 show that in general, the more diseases a subject has and the older the subject is, the higher weight it gets, just as expected.

The weights were grouped because three patterns had very few observations. The weights were collapsed into the groups ≤ 0, 1-4, 5-8, ≥ 9. This index made a nice separation of the Kaplan-Meier curves, as is seen in figure 6.2.

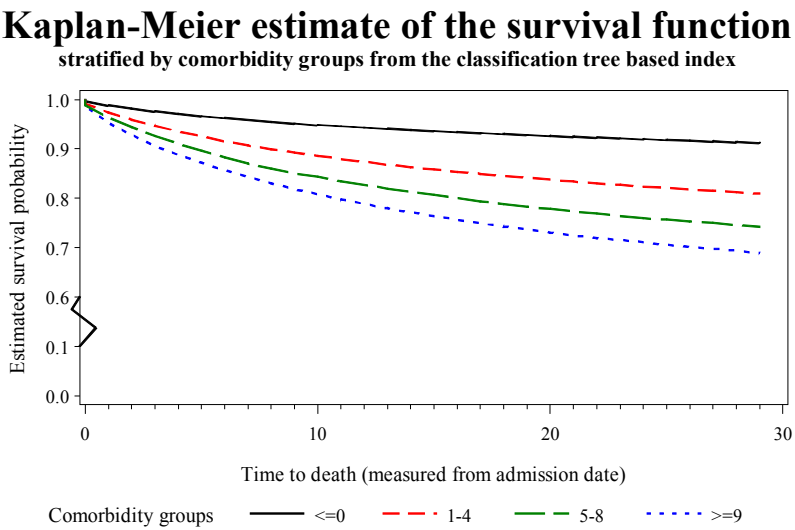


Figure 6.2 *Kaplan-Meier curves for the new index based on a classification tree including the 22 diseases, sex and age as predictors.*

To assess the applicability of the index it was included as a predictor in a logistic regression. The logistic regression was adjusted for age(modeled by a restricted cubic spline) and sex and was, as all previous regressions, made in SAS 9.2.

In table 6.2 the test statistics for this new index can be seen.

Predictors	χ^2_{HL}	R^2	**ROC**
New index	14	0.0841	0.679

Table 6.2 *Test statistics for the index based on a classification tree with age.*

The Hosmer-Lemeshow χ^2 value was very small. The area under the ROC curve indicated that the index was very poor at discriminating between outcomes.
To assess the predictability of the indexes the observed probability was plotted against the predicted probability, as seen in figure 6.3. It shows that the index is good at predicting the probability of dying within 30 days.

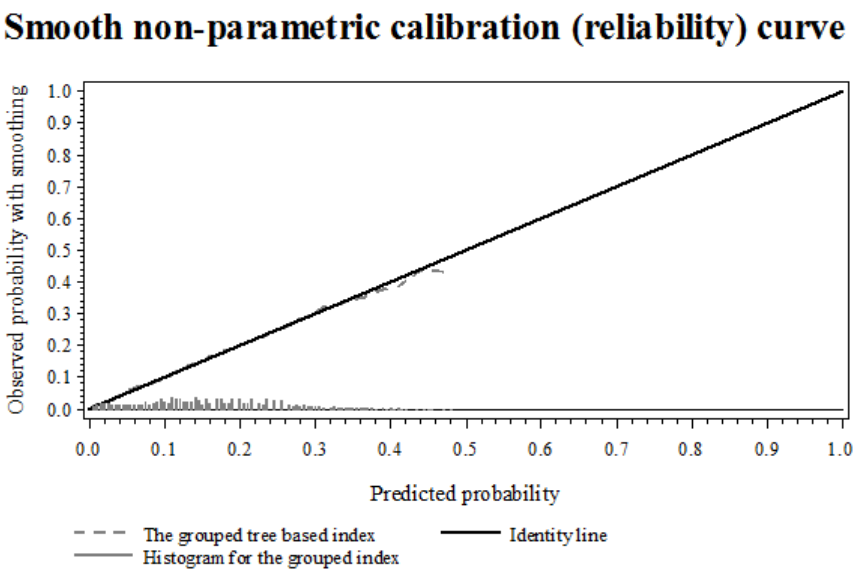


Figure 6.3 *Calibration plot of the tree based index with age.*

Classification tree without age

In figure 6.1 it is seen that age is the most dominant predictor of death, and compared to this predictor some of the others may not contain enough information to be included in the tree. So in order to get as much information out of the 22 diseases, age was excluded as a predictor.
In this tree the minimum number of observations in a node was set to 195. If this parameter was set any higher the tree only contained the root. The complexity parameter was again set to zero. This tree is seen in figure 6.4.

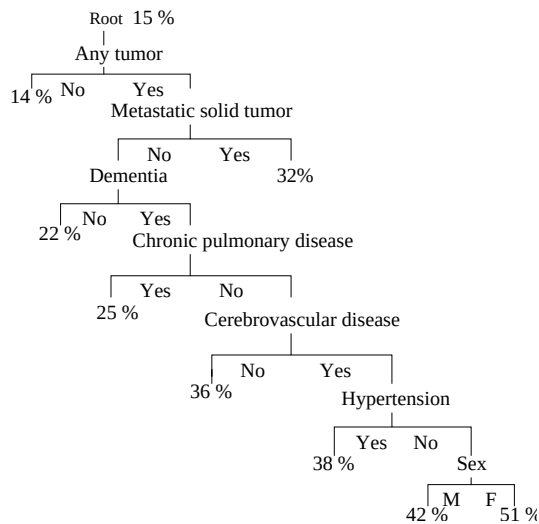


Figure 6.4 *Classification tree with the 22 diseases and sex as covariates, with $\alpha = 0$ and $minsplit = 195$. At the root and each terminal node the estimated probability of death is given.*

The index was then again made by multiplying the log odds ratio, calculated as in equation (6.1), by 10 and rounding to the nearest integer. Table 6.3 shows that again the index value increased with the number of comorbidities, as wanted.

Covariate pattern	Log odds ratio	weights	% obs.
Any tumor=no	-0.11	-1	85.4
Any tumor=yes, Metastatic solid tumor=yes	0.94	9	1.9
Any tumor=yes, Metastatic solid tumor=no, Dementia=no	0.44	4	12.2
Any tumor=yes, Metastatic solid tumor=no, Dementia=yes, Chronic pulmonary disease=yes	0.58	6	0.1
Any tumor=yes, Metastatic solid tumor=no, Dementia=yes, Chronic pulmonary disease=no, Cerebrovascular disease=no	1.14	11	0.3
Any tumor=yes, Metastatic solid tumor=no, Dementia=yes, Chronic pulmonary disease=no, Cerebrovascular disease=yes, Hypertension=yes	1.19	12	0.0
Any tumor=yes, Metastatic solid tumor=no, Dementia=yes, Chronic pulmonary disease=no, Cerebrovascular disease=yes, Hypertension=no, sex=M	1.40	14	0.1
Any tumor=yes, Metastatic solid tumor=no, Dementia=yes, Chronic pulmonary disease=no, Cerebrovascular disease=yes, Hypertension=no, sex=F	1.74	17	0.1

Table 6.3 *The resulting covariate patterns from the classification tree, their weights and the proportion of observations in each pattern.*

To ensure that there were enough observations in each index value the weights were grouped into the groups ≤ 0, 1-4, 5-8, and ≥ 9. The Kaplan-Meier curves, seen in figure 6.5 show that this index separated the comorbidity groups fairly even though the group 5-8 contained few observations.

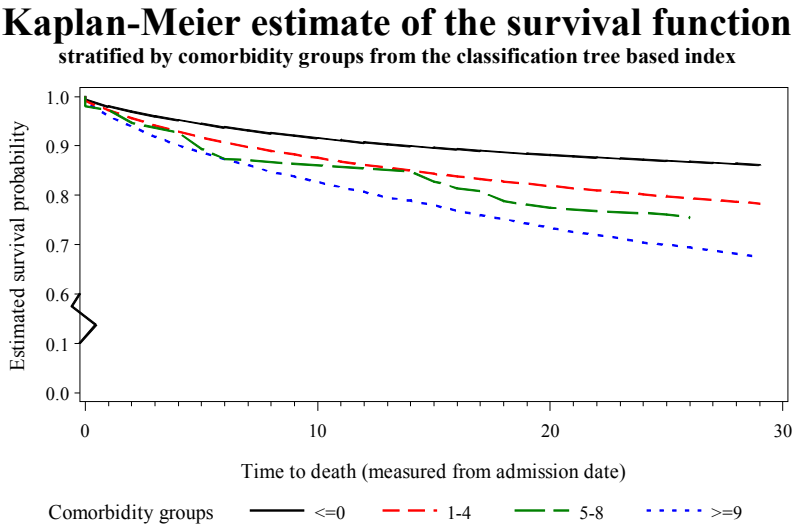


Figure 6.5 *Kaplan-Meier curves for the new index based on a classification tree including the 22 diseases and sex as predictors.*

This index was included as predictors in a logistic regression adjusted for sex and age (modeled by a restricted cubic spline).

The Hosmer-Lemeshow test for the regression, showed that the tree based index was poor at predicting the number of deaths in the low and highest probability groups, as seen in table 6.4 .

Group	Total	Observed deaths	Expected deaths	Deviance residuals
1	20562	397	477	-4
2	20953	1245	1306	-2
3	20381	1706	1856	-3
4	21449	2399	2401	0
5	19384	2623	2565	1
6	20631	3355	3200	3
7	21415	4038	3902	2
8	20471	4455	4367	1
9	20393	5284	5118	2
10	19208	6056	6367	-4

Table 6.4 *Hosmer-Lemeshow χ^2 test.*

The logistic regression gave the following goodness of fit statistics

Predictors	χ^2_{HL}	R^2	ROC
Index with age	14	0.0841	0.679
New index without age	78	0.0871	0.692

Table 6.5 *Comparison of the classification tree based indexes.*

It is seen that both the R^2 value and the area under the ROC curve were higher for the without age. The area under the ROC curve again indicated that the indexes were poor at discriminating between outcome. The χ^2 value showed that a model containing an index based on a classification tree still didn't model all the information in the data.

The predicted and observed probabilities were again plotted against one another to assess the predictability of the index. Figure 6.6 show that the index predict death fairly in the range with many observations. The index including age was however better than the index without age at predicting death in the range with few observations.

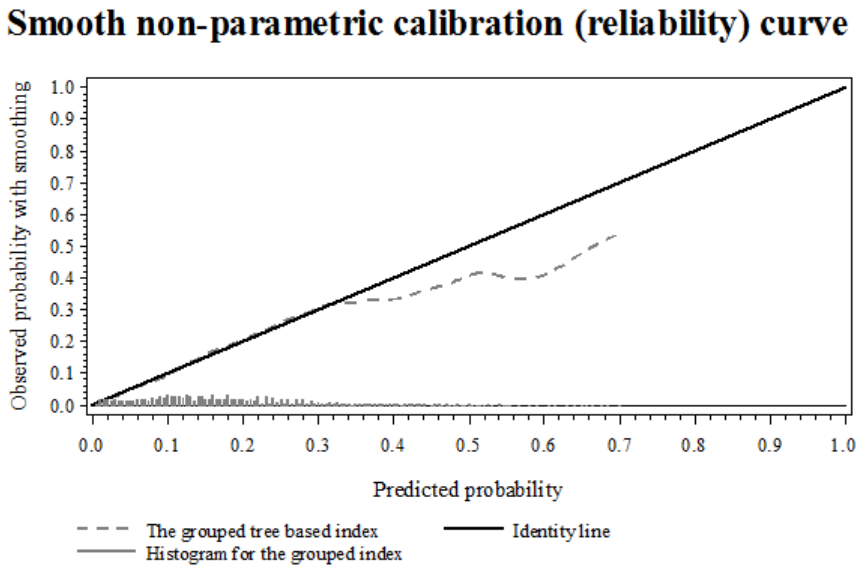


Figure 6.6 *Calibration plot of the classification tree based index.*

Conclusion

TThe index made, whether it included age or not, discriminated between the comorbidity groups and predicted death within 30 days fairly. However it was very bad at discriminating between outcome.

Validating the indexes

In the previous chapters many different indexes have been developed. In this chapter these indexes are validated. This is done in order to see whether the indexes can be applied to other datasets besides the training set and to determine, if possible which of the indexes are best.

The validation was made on the dataset containing pneumonia patients admitted to a hospital in 2007, as described in chapter 2. In the validation both 30 day and 1 year mortality were used as an outcome.

When the indexes were developed, their performance on the training dataset were assessed. This was done in order to give an indication of which of the indexes that might be the best one. The test statistics for these regressions are seen in table 7.1.

Predictors	χ^2_{HL}	R^2	**ROC**
CCI index	145	0.0948	0.698
Crude index (19 diseases)	266	0.1007	0.705
Index (19 diseases)	257	0.1052	0.709
Index (22 diseases)	300	0.1067	0.710
Index (22 diseases) with interaction with sex	308	0.1065	0.711
Index (22 diseases) with interaction with age	197	0.1124	0.712
Index (22 diseases) with interaction between diseases	290	0.1072	0.711
Index (22 diseases) with first degree interactions	192	0.1139	0.712
Index with 'time since diagnosis'	281	0.1155	0.717
Index with 'time since diagnosis' and first degree interactions	273	0.1166	0.718
Simple naive Bayes index	197	0.0965	0.700
Extended naive Bayes index	88	0.0942	0.697
Tree index without age	78	0.0871	0.692
Tree index with age	14	0.0841	0.679

Table 7.1 *Comparison of the test statistics from the regressions made on the training set. The statistics for CCI is for the training data only.*

Table 7.3 show that the indexes with 'time since diagnosis' performed the best. Overall the naive Bayes indexes and the classification tree indexes performed worse than the indexes based on a logistic regression.

7.1 Validation on 30 day mortality

We started the validation by making Pearson's χ^2 test on the contingency table for 30 day mortality versus the index groups, to assess the crude performance of all the indexes. The χ^2 value was standardized by $(\chi^2 - DF)/\sqrt{(2DF)}$, so that a direct comparison of the indexes was possible. The χ^2, and the standardized χ^2 values can be seen in table 7.2.

Index	χ^2 **value**	**DF**	**Standardized** χ^2
CCI	458	3	186
Crude index (19 diseases)*	893	6	256
Index (19 diseases)*	798	5	251
Index (22 diseases)*	717	5	225
Index (22 diseases) with interaction with sex*	725	5	228
Index (22 diseases) with interaction with age	210	6	59
Index (22 diseases) with interaction between diseases*	733	5	230
Index (22 diseases) with first degree interactions	640	6	183
Index with time since diagnosis*	824	6	236
Index with time since diagnosis* and first degree interactions	722	6	207
Simple naive Bayes index*	742	5	233
Extended naive Bayes index*	1341	5	422
Tree index without age	344	3	139
Tree index with age*	1001	3	407

Table 7.2 *The χ^2 test of all the indexes. The star marks the best indexes*

When looking at table 7.2 it is seen that the extended naive Bayes index, the classification tree including age and the crude index were the three best indexes at predicting death. This was however not surprising since these three indexes were the only ones where age was included, and age was the most dominant predictor of death. In the crude index age was not included directly as in the two others, however age was included indirectly since the weights of the diseases were not adjusted for age. This meant that some of the older patients ended up in the higher groups and some of the younger patients in the lower groups and since old people die more often than young people, the high groups have a higher mortality. An example is dementia, which is a disease with a higher prevalence among older people than younger. This disease had a higher weight in the crude index than in the adjusted indexes, because of the general effect age had on mortality.

When looking at the rest of the indexes it is seen that the seven best indexes are: index (19 diseases), index with 'time since diagnosis', simple naive Bayes index, index (22 diseases) with interaction between diseases, index (22 diseases) with interaction with sex, index (22 diseases) and the index with 'time since diagnosis' and first degree interactions. These indexes are marked by a star in table 7.2.

Important features of an index are its ability to separate between the groups and to put them in the correct order. To see how well the indexes did this, the deviance residuals for the comorbidity groups were calculated. If the index separated the groups well and put them in the correct order, the deviance residuals should go from being negative in the low groups to being positive in the high groups. The deviance residuals for the indexes can be seen in figure 7.1.

Index / groups	**0**	**1-2**	**3-4**	**≥ 5**					
CCI	-14	2	9	10					
Index / groups	**≤ 0**	**1-3**	**4-6**	**7-9**	**10-12**	**≥ 13**			
Index (19 diseases)	-16	0	12	10	10	7			
Index (22 diseases)	-16	1	9	9	10	7			
Index (22 diseases) with interaction with sex	-16	1	9	9	10	10			
Index (22 diseases) with interaction between diseases	-16	-1	9	11	10	8			
Index / groups	**≤ 0**	**1-3**	**4-6**	**7-9**	**10-12**	**13-15**	**≥ 16**		
Crude index (19 diseases)	-17	-2	6	10	12	11	8		
Index (22 diseases) with interaction with age	-9	4	6	5	4	10	3		
Index (22 diseases) with first degree interactions	-17	5	4	9	6	8	6		
Index with time since diagnosis	-16	1	10	8	9	6	12		
Index with time since diagnosis and first degree interactions	-16	2	7	8	8	5	12		
Index / groups	**≤ 0**	**1-5**	**6-10**	**11-15**	**16-20**	**≥ 21**			
Simple Naive Bayes	-13	-9	9	11	11	7			
Extended naive Bayes	-23	-3	7	11	14	16			
Index / groups	**≤ 0**	**1-4**	**5-8**	**≥ 9**					
Tree without age	-7	11	-1	10					
Tree with age	-18	8	14	16					

Figure 7.1 *The deviance residuals for all indexes.*

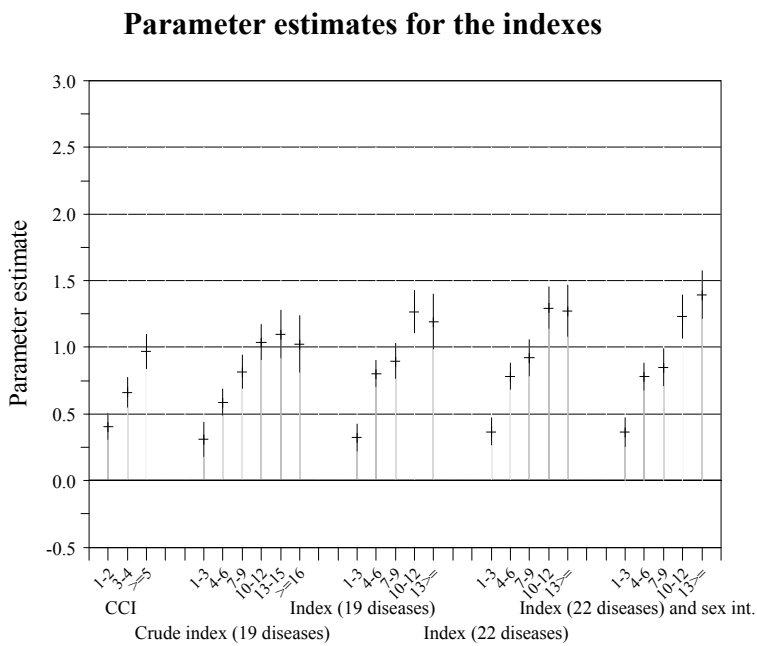
Figure 7.1 shows that all the indexes, except for the tree index without age, arranged the groups as wanted. The tree index without age didn't do quite as well, since the index group 5-8 containing very few observations was negative.

At last the performance of the indexes when included in a logistic regression adjusted for sex and age was assessed. In table 7.3 it is seen that when adjusting for sex and age the extended naive Bayes index no longer was the best. When comparing the indexes, it is seen that the index including 'time since diagnosis' and interaction terms was the best one, however it was only slightly better than the index containing only the 'time since diagnosis'. The next best indexes were the index (22 diseases) and first degree interactions, the index (22 diseases) and the index (19 diseases). It is worth noticing that the indexes including interaction terms did not perform noticeably better than the indexes without.

Predictors	χ^2_{HL}	R^2	**ROC**
CCI	17	0.1153	0.700
Crude index (19 diseases)	48	0.1265	0.713
Index (19 diseases)	39	0.1322	0.717
Index (22 diseases)	37	0.1329	0.718
Index (22 diseases) with interaction with sex	35	0.1320	0.718
Index (22 diseases) with interaction with age	15	0.1160	0.703
Index (22 diseases) with interaction between diseases	38	0.1310	0.716
Index (22 diseases) with first degree interactions	30	0.1340	0.718
Index with time since diagnosis	36	0.1405	0.727
Index with time since diagnosis and first degree interactions	27	0.1478	0.728
Simple naive Bayes index	31	0.1184	0.704
Extended naive Bayes index	13	0.1157	0.703
Tree index without age	40	0.1104	0.703
Tree index with age	2	0.1009	0.682

Table 7.3 *Comparison of the goodness of fit statistics.*

The indexes ability to separate and order the groups when included in a logistic regression are very important features, and therefore the parameter estimates including the confidence limits were plotted. These are seen in figure 7.2 and 7.3.



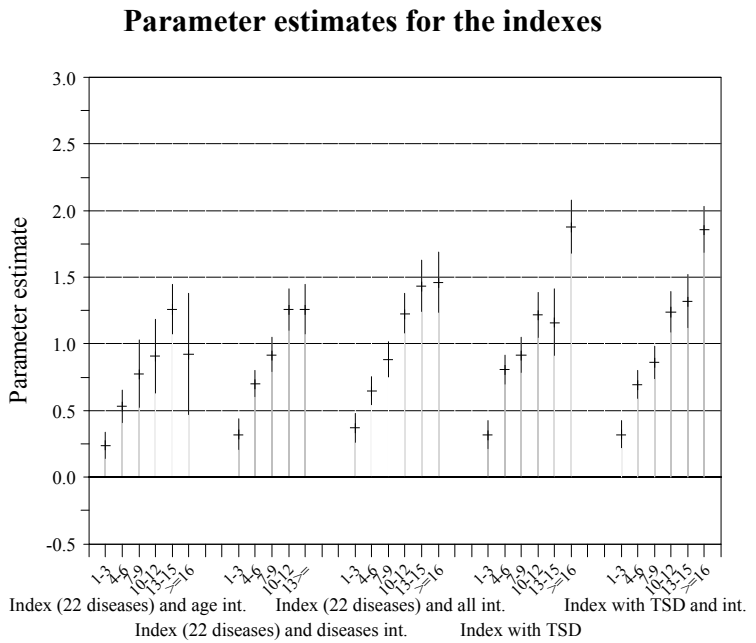


Figure 7.2 *Parameter estimates with 95% confidence interval for 30 day mortality. Note that TSD is 'time since diagnosis'.*

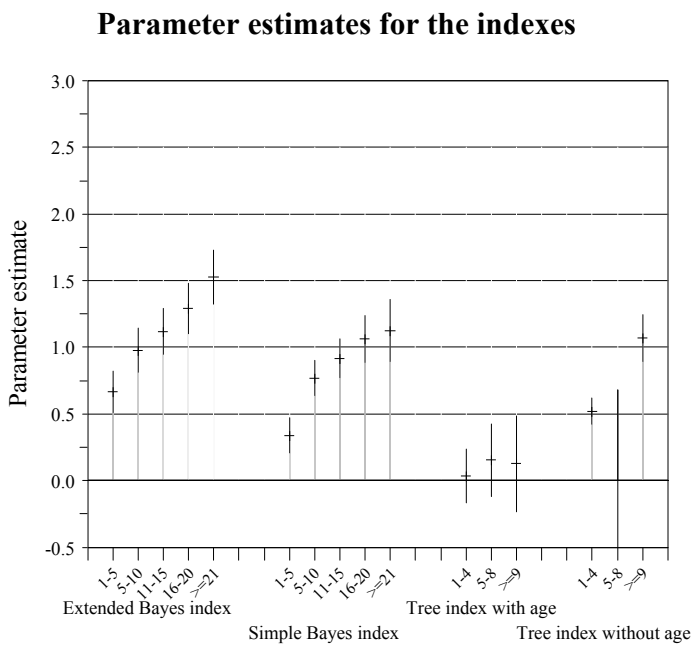


Figure 7.3 *Parameter estimates with 95% confidence interval for 30 day mortality.*

It is seen that the indexes separated the comorbidity groups fairly and that parameter estimates increased nicely with the comorbidity groups, except for the last comorbidity group. The 95% confidence intervals became wider for higher comorbidity groups in general. This was not the case for CCI. The parameter estimates for the developed indexes also have a wider range than the parameter estimates for CCI, except for the tree based indexes. Note that the group for 5-8 in the tree index without age had a parameter estimate on -1.5 [-3.7;0.7].
The index with 'time since diagnosis' and first degree interactions performed the best, since the parameter estimates had the widest range and the comorbidity groups were increasing and nicely separated. The index with 'time since diagnosis' performed almost as good. The index (22 diseases) and all first degree interactions performed slightly better than the rest of the indexes with 19 or 22 diseases. The naive Bayes indexes separated the groups nicely, but the parameter estimate had a smaller range.

7.2 Validation on 1 year mortality

To see if the same tendencies for the indexes are in evidence with 1 year mortality as an outcome, the χ^2 tests were also made for 1 year mortality versus comorbidity groups. These can be seen in table 7.4.

Index	χ^2 **value**	**DF**	**Standardized** χ^2
CCI*	1565	3	638
Crude index (19 diseases)*	2351	6	777
Index (19 diseases)*	2198	5	883
Index (22 diseases)*	1956	5	617
Index (22 diseases) with interaction with sex*	1997	5	630
Index (22 diseases) with interaction with age	684	6	196
Index (22 diseases) with interaction between diseases*	2001	5	631
Index (22 diseases) with first degree interactions	1838	6	579
Index with time since diagnosis*	2164	6	623
Index with time since diagnosis and first degree interactions	1974	6	568
Simple naive Bayes index	1923	5	607
Extended naive Bayes index*	2990	5	944
Tree index without age	1216	3	495
Tree index with age*	1847	3	753

Table 7.4 *A χ^2 test of all the index.*

For the 1 year mortality the extended naive Bayes index was again the best. However since it still included age this was not a surprise. For the rest of the indexes, the best indexes were the index (19 diseases), the crude index (19 diseases), tree index with age, CCI , the index (22 diseases) with interaction between diseases, the index (22 diseases) with interaction with sex, index with 'time since diagnosis', index (22 diseases) and the simple naive Bayes index. They are marked by a star in table 7.4.

The indexes ability to separate and order the groups were also assessed for the 1 year mortality. This can be seen in table 7.4.

Index / groups	**0**	**1-2**	**3-4**	**$\geq$ 5**							
CCI	-22	2	14	19							
Index / groups	**$\leq$ 0**	**1-3**	**4-6**	**7-9**	**10-12**	**$\geq$ 13**					
Index (19 diseases)	-25	2	17	11	16	12					
Index (22 diseases)	-24	4	13	11	16	13					
Index (22 diseases) with interaction with sex	-24	3	13	11	14	16					
Index (22 diseases) with interaction between diseases	-24	-1	13	13	15	13					
Index / groups	**$\leq$ 0**	**1-3**	**4-6**	**7-9**	**10-12**	**13-15**	**$\geq$ 16**				
Crude index (19 diseases)	-25	-3	12	14	18	13	12				
Index (22 diseases) with interaction with age	-12	6	7	6	9	14	7				
Index (22 diseases) with first degree interactions	-24	6	7	11	11	13	11				
Index with time since diagnosis	-24	3	11	14	15	9	16				
Index with time since diagnosis and first degree interactions	-24	3	9	11	13	11	17				
Index / groups	**$\leq$ 0**	**1-5**	**6-10**	**11-15**	**16-20**	**$\geq$ 21**					
Simple Naive Bayes	-20	-13	13	17	14	10					
Extended naive Bayes	-31	-3	10	17	18	18					
Index / groups	**$\leq$ 0**	**1-4**	**5-8**	**$\geq$ 9**							
Tree without age	-11	18	2	19							
Tree with age	-23	10	16	19							

Figure 7.4 *The deviance residuals for all indexes.*

Figure 7.4 shows that all the indexes arranged the groups as wanted. Notice that the tree index without age still has very few observation in the group 5-8.

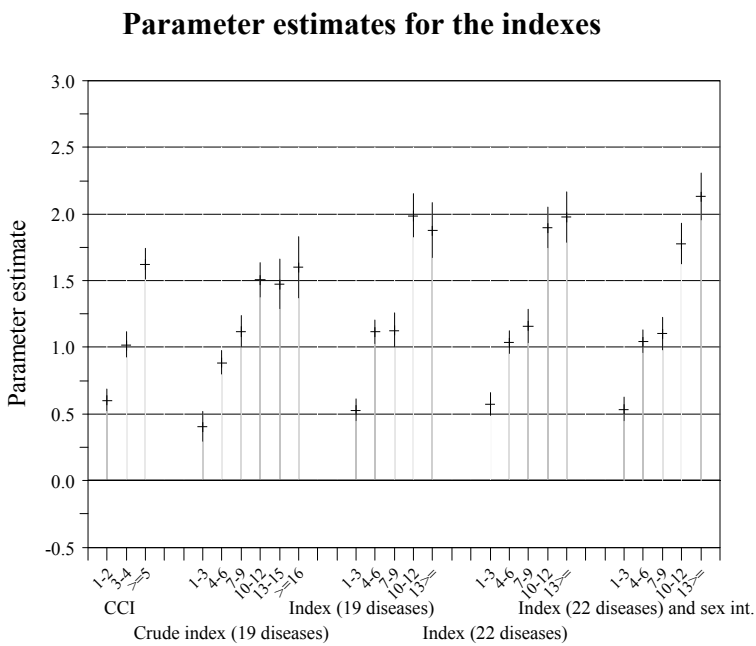
The indexes were again included in a logistic regression adjusted for sex and age. The test statistics can be seen in table 7.5.

Predictors	χ^2_{HL}	R^2	**ROC**
CCI	50	0.1975	0.736
Crude index (19 diseases)	116	0.2016	0.747
Index (19 diseases)	84	0.2177	0.753
Index (22 diseases)	66	0.2181	0.752
Index (22 diseases) with interaction with sex	70	0.2200	0.753
Index (22 diseases) with interaction with age	14	0.1904	0.731
Index (22 diseases) with interaction between diseases	81	0.2168	0.751
Index (22 diseases) with first degree interactions	45	0.2226	0.754
Index with time since diagnosis	62	0.2398	0.761
Index with time since diagnosis and first degree interactions	44	0.2498	0.764
Simple naive Bayes index	62	0.1869	0.733
Extended naive Bayes index	26	0.1776	0.729
Tree index without age	77	0.1743	0.738
Tree index with age	3	0.1525	0.698

Table 7.5 *Comparison of the goodness of fit statistics.*

Table 7.5 show that also this time the extended naive Bayes index and the classification tree did far worse than the rest of the indexes in predicting death. When comparing the rest of the indexes it is seen that the index including 'time since diagnosis' and first degree interactions performs the best followed by the index with 'time since diagnosis'. The index with all first degree interactions, index with interaction with sex, the index with 22 diseases and the index with 19 diseases are the next best indexes, with no major difference in their performance. Again it is seen that including interactions does not increase performance notably.

To see how well the indexes separated and ordered the groups when included in a logistic regression, the parameter estimates and their confidence limits are plotted. In figure 7.5 and 7.6 it is seen that



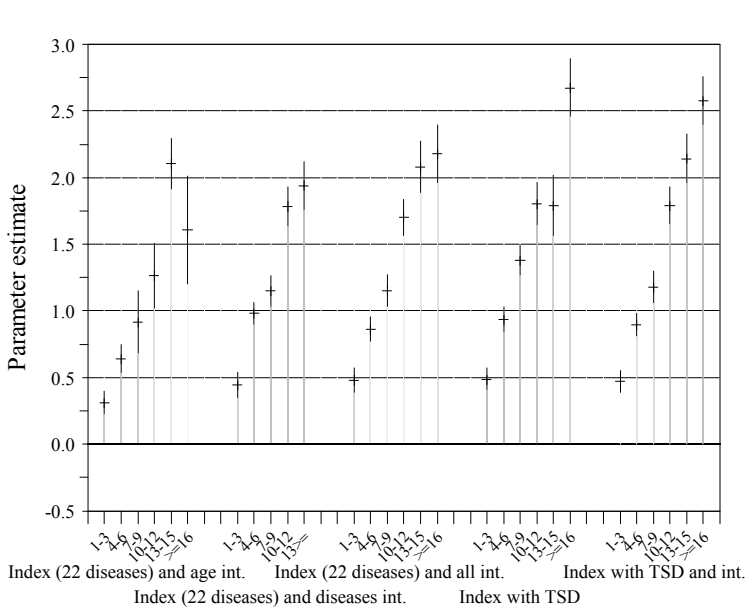


Figure 7.5 *Parameter estimates with 95% confidence interval for 1 year mortality*

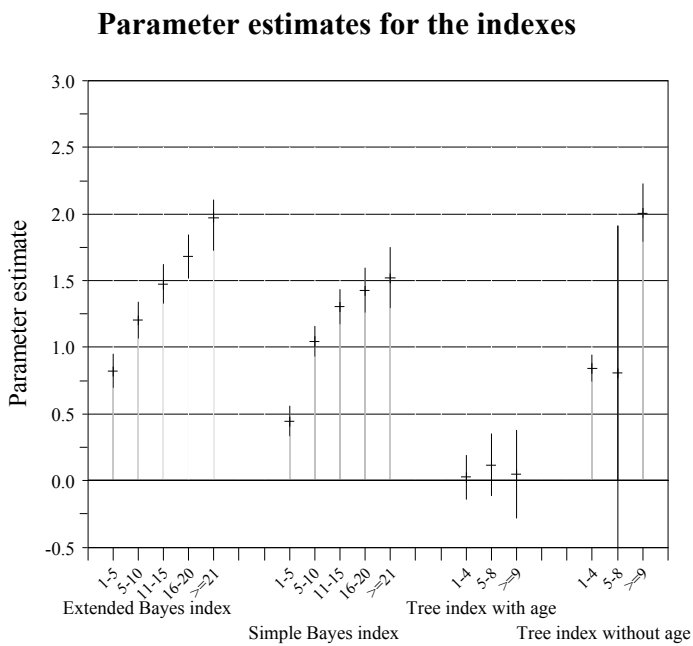


Figure 7.6 *Parameter estimates with 95% confidence interval for 1 year mortality*

Again it is seen that the indexes separated the comorbidity groups fairly and that parameter estimates increased nicely with the comorbidity groups, except for some of the last comorbidity groups. The 95% confidence intervals also became wider for higher comorbidity groups in general. This was still not the case for CCI. The parameter estimates for the developed indexes also have a wider range than the parameter estimates for CCI, except for the tree based indexes. Note that the group for 5-8 in the tree index without age had a parameter estimate on 0.8 [-0.8;1.9].
The index with 'time since diagnosis' and first degree interactions performed again the best, since the parameter estimates had the widest range and the comorbidity groups were increasing and nicely separated. The index with 'time since diagnosis' performed almost as good. The index (22 diseases) and all first degree interactions performed slightly better than the rest of the indexes with 19 or 22 diseases. The naive Bayes indexes separated the groups nicely, but the parameter estimate had a smaller range.

Conclusion

When looking at the overall performance of the indexes it is seen that the index including 'time since diagnosis' and first degree interactions is the best one, since it both had a high χ^2 value when the crude performance was assessed and had high R^2 value and area under the ROC curve, when the adjusted performance was assessed. The next best indexes were the index including ' time since diagnosis', the index including 22 diseases and the index including 22 diseases an first degree interactions.
The validation showed that there were no major difference between the index with 19 diseases and the indexes with 22 diseases (except the index with all interactions). So taking both the complexity and the performance into account we find that the index (22 diseases) is the best out of these.

The indexes including 'time since diagnosis' did perform better than the indexes including the 22 diseases, however one should assess if the small increase in performance is worth the extra complexity.

In the situation were a crude comorbidity index is needed, the extended naive Bayes is the best choice, since it takes both diseases, sex and age into account. However if simplicity also is of some importance the tree index with age should be chosen, since its crude performance is nearly as good and it contains only 6 diseases compared to 22 in the extended naive Bayes index.

7.3 Comparison of CCI and the best indexes

We found that the index with 19 diseases and the index with 'time since diagnosis' were the two best indexes, so in order to see if there were any difference between the two, they were both included in a logistic regression adjusted for sex and age. In order to see whether these two indexes differed from CCI similar logistic regressions were made, and at last a logistic regression with all three indexes was made. The Wald test statistics for these four regressions can be seen in table 7.6.

Index	**Standerdized Wald χ^2**	**p-value**
CCI	1	0.1351
Index (22 diseases)	84	$<$.0001
CCI	2	0.0288
Index (22 diseases) and first degree interactions	83	$<$.0001
Index (22 diseases)	6	0.0001
Index (22 diseases) and first degree interactions	12	$<$.0001
CCI	4	0.0043
Index with 'time since diagnosis'	119	$<$.0001
CCI	2	0.0419
Index with 'time since diagnosis' and first degree interactions	134	$<$.0001
Index with 'time since diagnosis'	5	0.0004
Index with 'time since diagnosis' and first degree interactions	16	$<$.0001
Index (22 diseases)	3	0.0273
Index with 'time since diagnosis'	42	$<$.0001
Index (22 diseases) and first degree interactions	0	0.5219
Index with 'time since diagnosis' and first degree interactions	46	$<$.0001
CCI	4	0.0085
Index (22 diseases)	0.3	0.3071
Index (22 diseases) and first degree interactions	0	0.8853
Index with 'time since diagnosis'	4	0.0022
Index with 'time since diagnosis' and first degree interactions	8	$<$.0001

Table 7.6 *The significance test for the four logistic regressions.*

In table 7.6 it is seen that all the developed indexes contain more information than CCI. Further more the table showed that the indexes with interaction terms were better than the ones without and that including 'time since diagnosis' improved the performance of the indexes.

When comparing the test statistics, seen in table 7.7, from the logistic regressions containing the indexes separately, it is seen that all the indexes perform better than CCI and that the performance increase with the complexity and the inclusion of 'time since diagnosis'.

Predictors	χ^2_{HL}	R^2	**ROC**
CCI	17	0.1153	0.700
Index (22 diseases)	37	0.1329	0.718
Index (22 diseases) with first degree interactions	30	0.1340	0.718
Index with time since diagnosis	36	0.1405	0.727
Index with time since diagnosis and first degree interactions	27	0.1478	0.728

Table 7.7 *Comparison of the goodness of fit statistics.*

Conclusion

The comparison of CCI and the best developed indexes showed that all indexes performed better than CCI and that the index with 'time since diagnosis' and first degree interaction terms perform the best. The index (22 diseases) and first degree interaction terms did not perform better than the index with 'time since diagnosis', but it was more complex, so we do not recommend using this one if the variable 'time since diagnosis' is available.
When choosing between the indexes it depends on the situation at hand and is a balance between complexity and performance.

Discussion

I this chapter we evaluate the strengths and weaknesses of the methods used throughout this thesis. We also discuss the results and finally we present the conclusion of this thesis and ideas for future work.

The data

The dataset used in all the analyses, consists of data from the Danish National Registry of Patients. DNRP consists of 99.5 % of all hospitalizations in Denmark, and the use of this registry ensures that the entire Danish population is represented. The use of DNRP also makes it possible to have a very large dataset, which ensures statistical power. DNRP does, however also contain some error coding, so the comorbidities may not all be correctly registered. We assume that the incorrect registration is random, hence we do not expect the parameter estimates to be biased.

The pneumonia diagnoses, defining our study population, have a positive predictive value of 90 % in DNRP. The high positive predictive value of the pneumonia diagnoses ensures a high validity of our results. The choice of pneumonia patients do however put some restrictions on the generalization of the developed indexes. For instance, our analysis show that patients with chronic pulmonary disease have a lower mortality than patients without. This effect is caused by information bias, since the increased surveillance of patients with chronic pulmonary disease causes the pneumonia to be diagnosed at an earlier stage. Hence we do not expect the indexes to perform well on patient groups other than hospitalized pneumonia patients.

The statistical methods

Generalized linear models are well known models which often are used in statistical analysis. When analyzing binary outcomes logistic regression and Cox regression are to be used.
The logistic regression is the main method used in our analysis. The logistic regression is chosen over the Cox regression, because it handles binary response well, and because there is no general problem with censored data in our dataset. The assumptions of linearity and additivity in the logistic regression and in the Cox regression are relatively

easy to verify, where as the proportional hazard assumptions in the Cox regression can be difficult to meet in practise. The logistic regression is more robust than the Cox regression, which produces a larger variance. However the large dataset reduces this increase in variance, so the advantage of the Cox regression is minimal.

Alternative methods to the generalized linear models are the naive Bayes method and the classification tree.
The naive Bayes method is characterized by strong independence assumptions, which makes this model very simple. These independence assumptions make adjustments for confounding impossible and makes generalizations of results difficult. However the method has been shown to perform surprisingly well in spite of this.
The classification tree method is characterized by its ability to include many parameters and all their interactions in a relatively simple setting. The classification tree has the advantage that it ranks the different covariate patterns by the degrees of association with outcome, so it can also be used to assess the importance of the different patterns. This however makes adjustment for confounders difficult.

Results

When validating the Charlson comorbidity index we showed that the index was a significant predictor of death. Including the index in a logistic regression also showed, that the comorbidity groups were nicely separated, since the 95% confidence intervals for the log odds ratios of the groups were not overlapping. This separation was also seen when no adjustments were made, since the Kaplan-Meier curves for the comorbidity groups also were nicely separated.

The weights obtained from the different regressions adjusted for sex and age, were very similar, but when including the three additional diseases in the index it was seen that the weights for the liver diseases changed. This may be caused by the fact that alcohol related disorders and liver diseases are related.

When making an index we emphasize three important features an index should have. Firstly the index must be able to predict death both with and without adjustment for sex and age. Secondly mortality should in-

crease with the comorbidity groups. Thirdly the index must be simple to use and easy to interpret. Another desirable property is that the weights reflect the effect the disease has on the mortality in general.

All indexes including CCI were able to predict death both with and without adjustment for sex and age, some however better than others. The mortality of the different indexes was not strictly increasing in all cases.
When including pairwise interaction terms and the variable 'time since diagnosis' in the indexes they become notably more complex, especially for the indexes with interaction between diseases. Only the indexes with all interaction terms and the indexes including the variable 'times since diagnosis' compensated for the extra complexity by an extra increase in performance.
A high weight from the crude logistic regression and the naive Bayes method can illustrate, that having a disease increases the mortality risk or that patients with the disease generally are older than patients without. For the indexes adjusted for age, a high weight can no longer be explained by differences in age distributions, so the weight illustrates the effect of the disease. This means that weights in the crude index and the naive Bayes indexes do not have the same intuitive interpretation as the indexes from a logistic regression adjusted for sex and age. This is also the case for the classification tree method, since the weights are based directly on covariate patterns and not on individual covariates.

When adjusting for sex and age all the developed indexes performed better than CCI except the tree index with age. With these adjustments the best performing indexes were (in descending order) the index with 'time since diagnosis' and first degree interactions, the index with 'time since diagnosis', the index with 22 diseases and first degree interactions and the index with 22 diseases. Their performance increased with their complexity, so which index to choose depends on the situation at hand.

Without adjustments for sex and age, the best performing index was the extended naive Bayes index. The tree index with age performed almost as good, so if simplicity is of great importance this should be used.

8.1 Conclusion

The Charlson comorbidity index was able to predict death well among the cohort of pneumonia patients, and it was able to separate the different comorbidity groups nicely. For these reasons the index is still usable.

All of our developed indexes performed well and most of them better than CCI. Our analysis showed that the index with 'time since diagnosis' and first degree interactions, the index with 'time since diagnosis', the index with 22 diseases and first degree interactions and the index with 22 diseases were the best indexes at predicting death among the cohort of pneumonia patients. Choosing the best index among these is a balance between performance and simplicity and depends therefore on the situation at hand.

Future work

The severity of diseases is known to be an important factor when predicting death, so to improve the index further this variable could be included. The variable is however not easy to assess since it is not registered directly in DNRP.

We tried adding the diseases alcohol related disorders, history of obesity and hypertension, but this did not effect the performance of the index substantially. This indicates that it might be an idea to investigate if it is possible to remove some of the diseases from the index without effecting its performance.

Before applying the indexes on patient groups other than hospitalized pneumonia patients, it needs to be validated on that group.

The indexes are for now all discrete and made by grouping into intervals, but instead of grouping the index, the index may perform better as a continuous variable.

Appendix

A.1 Calibration plots for indexes developed using logistic regressions

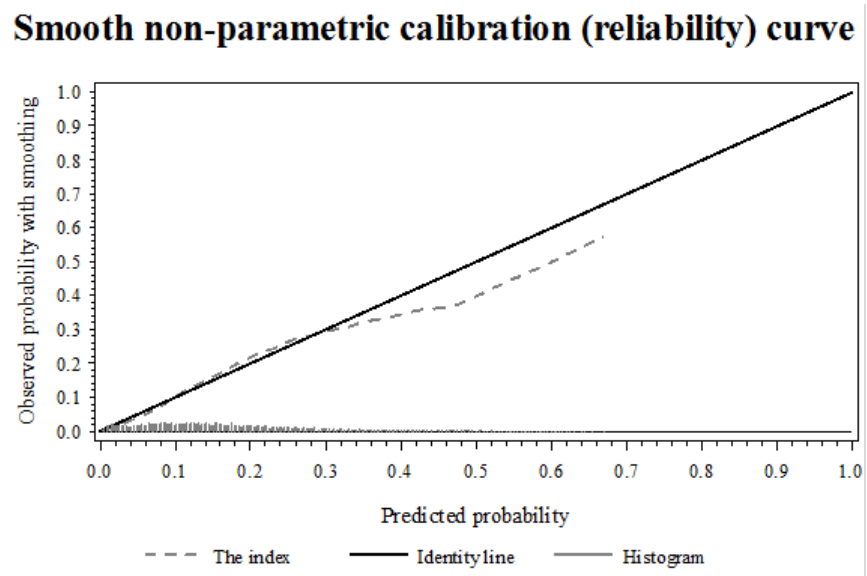


Figure A.1 *Observed and predicted probabilities for the index with 22 diseases.*

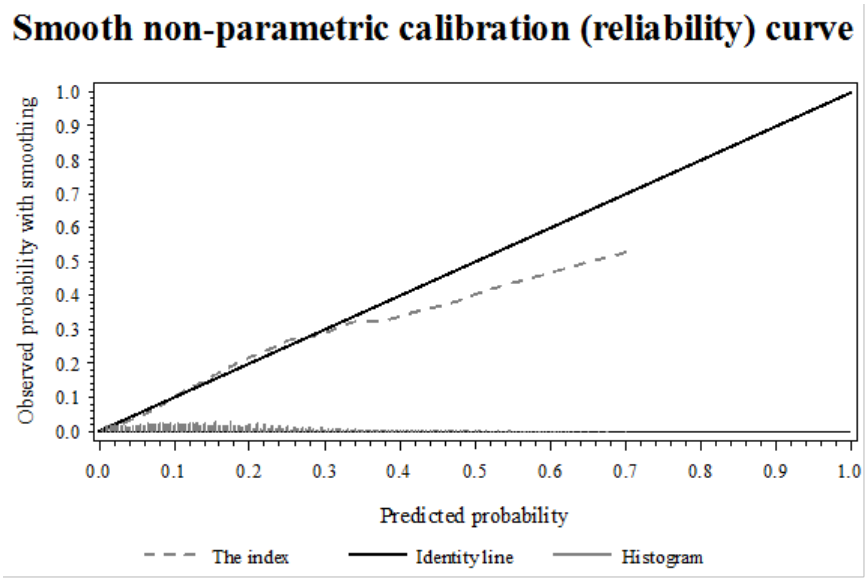


Figure A.2 *Observed and predicted probabilities for the index with 22 diseases and interaction with sex.*

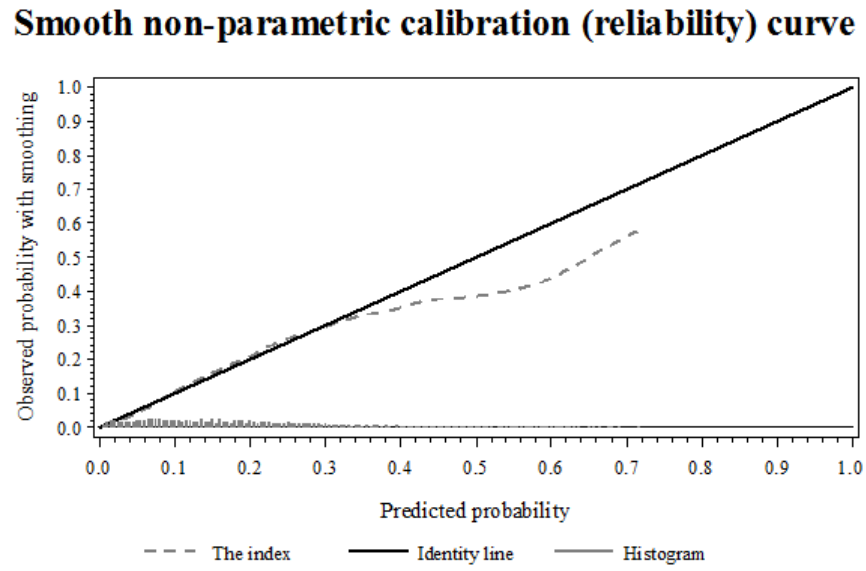


Figure A.3 *Observed and predicted probabilities for the index with 22 diseases and interaction with age.*

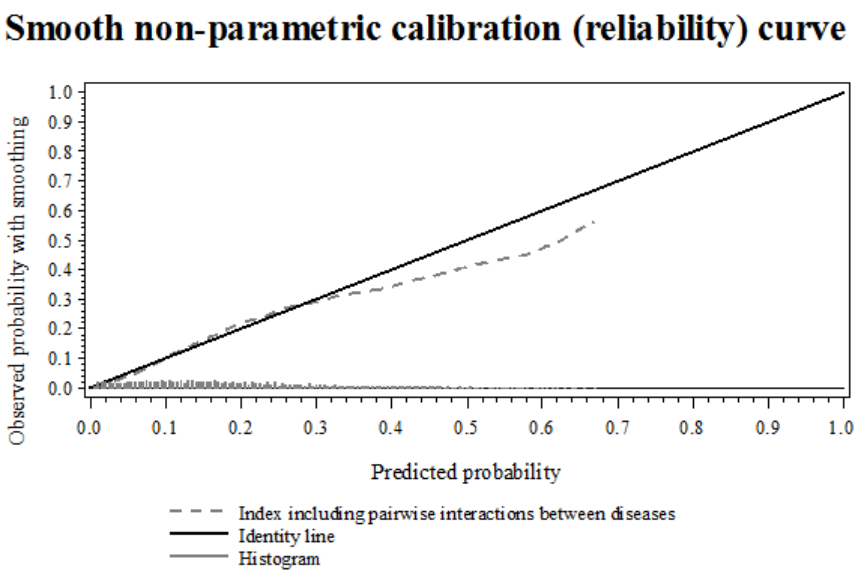


Figure A.4 *Observed and predicted probabilities for the index including 22 diseases and pairwise interaction between the diseases.*

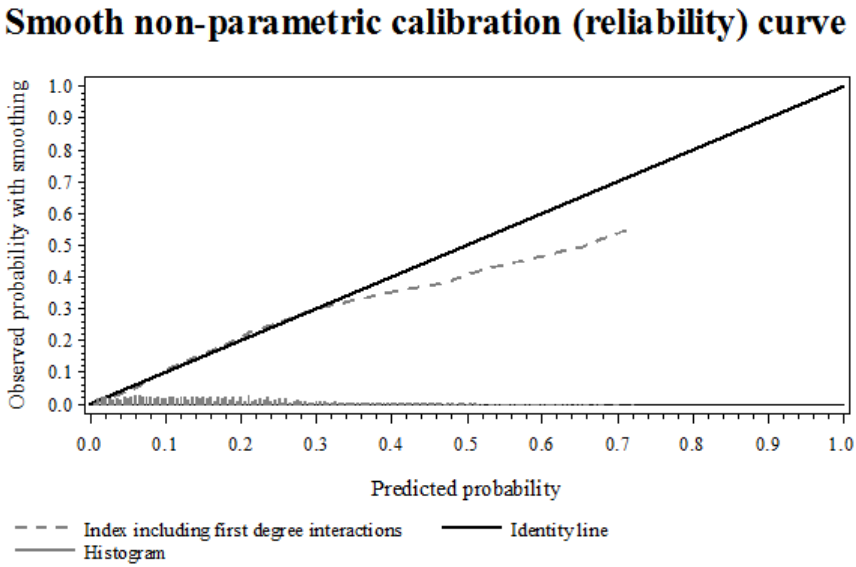


Figure A.5 *Observed and predicted probabilities for the index including 22 diseases and all interaction terms.*

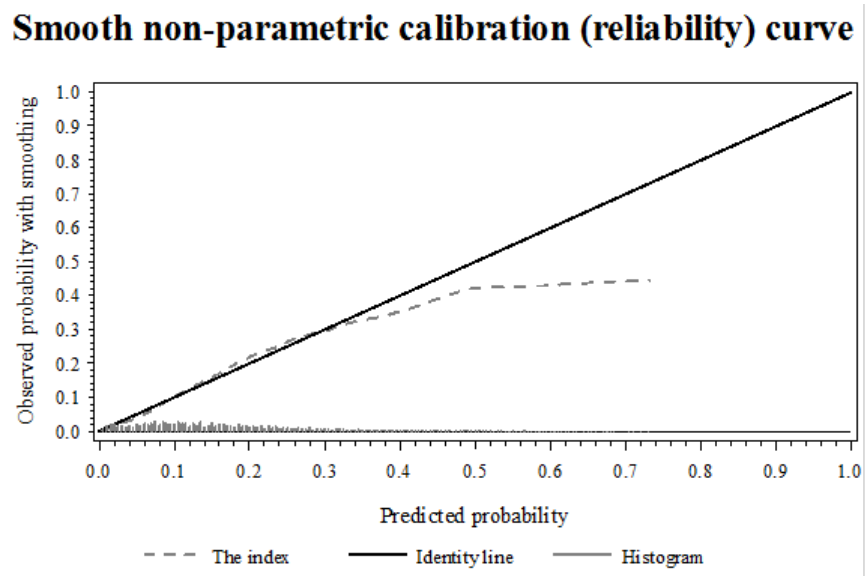


Figure A.6 *Observed and predicted probabilities for the index including 'time since diagnosis'.*

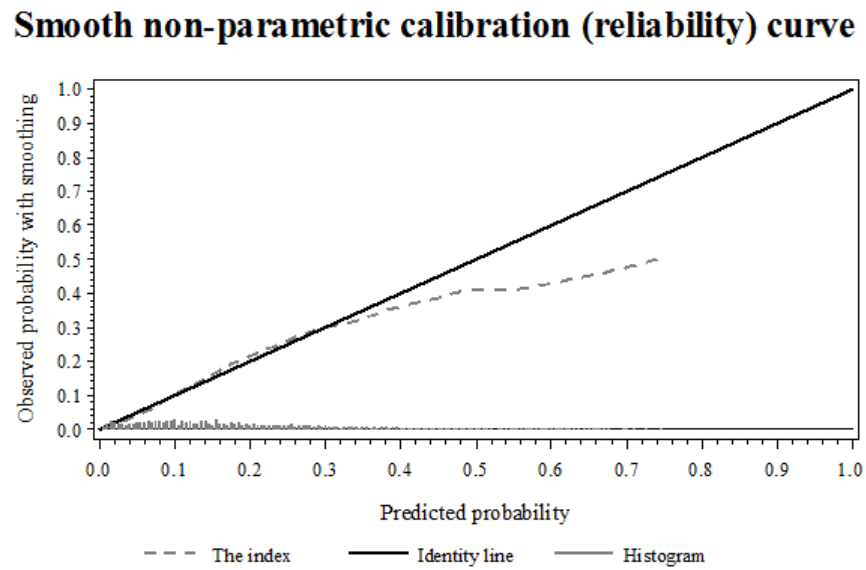


Figure A.7 *Observed and predicted probabilities for the index including 'time since diagnosis' and first degree interaction terms.*

A.2 ICD codes for comorbidities

Disease	ICD-8	ICD-10
Myocardial infarction	410	I21, I22, I23
Congestive heart failure	427.09, 427.10, 427.11, 427.19, 428.99, 782.49	I50, I11.0, I13.0, I13.2
Peripheral vascular disease	440, 441, 442, 443, 444, 445	I70, I71, I72, I73, I74, I77
Cerebrovascular disease	430-438	I60-I69, G45, G46
Dementia	290.09-290.19, 293.09	F00-F03, F05.1, G30
Chronic pulmonary disease	490-493, 515-518	J40-J47, J60-J67, J68.4, J70.1, J70.3, J84.1, J92.0, J96.1, J98.2, J98.3
Connective tissue disease	712, 716, 734, 446, 135.99	M05, M06, M08, M09, M30, M31, M32, M33, M34, M35, M36, D86
Ulcer disease	530.91, 530.98, 531-534	K22.1, K25-K28
Mild liver disease	571, 573.01, 573.04	B18, K70.0-K70.3, K70.9, K71, K73, K74, K76.0
Diabetes I and II	249.00, 249.06, 249.07, 249.09, 250.00, 250.06, 250.07, 250.09	E10.0, E10.1, E10.9, E11.0, E11.1, E11.9
Hemiplegia	344	G81, G82
Moderate to severe renal disease	403, 404, 580-584, 590.09, 593.19, 753.10-753.19, 792	I12, I13, N00-N05, N07, N11, N14, N17-N19, Q61
Diabetes with end organ damage	249.01-249.05, 249.08, 250.01-250.05, 250.08	E10.2-E10.8, E11.2-E11.8
Any tumor	140-194	C00-C75
Leukemia	204-207	C91-C95
Lymphoma	200-203, 275.59	C81-C85, C88, C90, C96
Moderate to severe liver disease	070.00, 070.02, 070.04, 070.06, 070.08, 573.00, 456.00-456.09	B15.0, B16.0, B16.2, B19.0, K70.4, K72, K76.6, I85
Metastatic solid tumor	195-198, 199	C76-C80
AIDS/HIV	079.83	B20-B24, Z21, Z219

Table A.1 *IDC codes for the diseases in the Charlson comorbidity index.*

Disease	ICD-8	ICD-10
Alcohol related disorders	291, 303, 979, 980, 577.10	F10, K860, Z721, R780, T51, K292, G621, G721, G312, I426
History of obesity	277.99	E65, E66
Hypertension	400-404	I10-I15

Table A.2 *IDC codes for the three new diseases.*

References

[Andersen et al., 1999] Andersen, T., Madsen, M., Jørgensen, J., Mellemkjær, L., and Olsen, J. (1999).
The Danish National Hospital Register. A valuable source of data for modern health sciences. *Danish Medical Bulletin*, vol. 46:263–268.

[David W. Hosmer, 2000] David W. Hosmer, S. L. (2000).
Applied Logistic Regression. John Wiley & Sons, second edition.

[Extermann, 2000] Extermann, M. (2000).
Measuring comorbidity in older cancer patients. *European Journal of Cancer*, vol. 36:453–471.

[Ezzati et al., 2002] Ezzati, M., A D Lopez, A. R., Hoorn, S. V., Murray, C. J. L., and Group, T. C. R. A. C. (2002).
Selected major risk factors and global and regional burden of disease. *Lancet Journal 2002*, 360:1347–60.

[Frank E. Harrell, 2001] Frank E. Harrell, J. (2001).
Regression Modeling Strategies. Springer, first edition.

[Gunnersen and Bisgaard, 2007] Gunnersen, S. J. and Bisgaard, M. P. (2007).
Danmark i tal 2007. Danmarks Statistik.

[Gustafsson et al., 2004] Gustafsson, F., Torp-Pedersen, C., Seibæk, M., Burchardt, H., Køber, L., and study group, T. D. (2004).
Effect of age on short and long-term mortality in patient admitted to hospital with congestive heart failure. *European Heart Journal*, 25:1711–1717.

[Kornum et al., 2007] Kornum, J. B., Thomsen, R. W., Riis, A., Lervang, H. H., Schønheyder, H. C., and Sørensen, H. C. (2007).
Type 2 Diabetes and Pneumonia Outcomes. A population-based cohort study. *Diabetes Care*, 30:2251–2257.

[Leo Breiman, 1984] Leo Breiman, Jerome H. Friedman, R. A. O. C. J. S. (1984).
Classification and regression Trees. Wadsworth, first edition.

[Rask-Madsen et al., 1997] Rask-Madsen, C., Jensen, G., Køber, L., Melchior, T., Torp-Pedersen, C., and Hildebrand, P. (1997). Age-related mortality, clinical heart failure, and ventricular fibriliation in 4259 Danish patients after acute myocardial infarction. *European Heart Journal*, 18:1426–1431.

[Sørensen et al., 2008] Sørensen, H. T., Christensen, T., Schlosser, H. K., Pedersen, L., and eds. (2008). *Use of Medical Databases in Clinical Epidemiology*. SUN-TRYK, Aarhus Universitet.

[Thomsen et al., 2006] Thomsen, R., Riis, A., Nørgaard, M., Jacobsen, J., Christensen, S., MCDonald, C., and Sørensen, H. (2006). Rising incidence and persistently high mortality of hospitalized pneumonia: a 10-year population-based study in Denmark. *Journal of International Medicin*, vol. 259:410–417.

[Vaccarino et al., 1999] Vaccarino, V., Parsons, L., Every, N. R., Barron, H. V., Krumholz, H. M., and the National Registry of Myocardial Infarction 2 Participants (1999). Sex-based Differences in early Mortality after Myocardial Infarction. *The New England Journal of Medicine*, 341:217–225.