**Chapter 5**

**Risk Assessment: Prognosis of Patients with Multiple Morbidities[[1]](#endnote-1)**

# [H1] Learning Objectives

**[INSERT NL]**

1. Define prognosis and risk
2. Define the difference between comprehensive and selective approaches to prognosis
3. Calculate the likelihood ratio associated with a large number of predictors
4. Adjust likelihood ratios for rare diseases
5. Adjust likelihood ratios for repeated diseases
6. Calculate likelihood ratios for combinations of diseases
7. Define the difference between detection and prediction
8. Cross-validate findings
9. Use Bayes to aggregate the risk of mortality across a large number of predictors
10. Calculate the accuracy of predictions
11. Identify obvious, rare, and informative predictors

**[END NL]**

# [H1] Key Concepts

**[INSERT BL]**

* Risk factor, prognosis, and severity
* Multimorbidity
* Likelihood ratio for rare diseases
* Likelihood ratio for repeated diseases
* Likelihood ratio for combinations of diseases
* Comorbidity versus complication
* Independent predictors
* Overlapping and redundant predictors
* Receiver operating curve
* Cross-validation
* Area under receiver operating curve

**[END BL]**

# [H1] Chapter at a Glance

Chapter 5 explains how to measure the risk of mortality. Prognostic information has many meaningful uses, and measuring the risk of mortality can help patients and clinicians plan for end-of-life decisions such as setting treatment priorities. Policy analysts can use prognostic tools to evaluate the comparative effectiveness of various treatment options. Administrators can also use prognostic information to anticipate patients’ acuity and nursing needs.

All of these uses presuppose that an accurate measure of prognosis exists. This chapter shows one way to create such a prognostic index. Historically, regression analysis has been used to create predictive models. When thousands of independent variables are involved, structured query language (SQL) provides an easier approach. Chapter 5 provides SQL code for predicting patients’ prognoses.

# [H1] Introduction

We use the terms *risk of mortality*, *severity of illness*, and *prognosis* interchangeably throughout this chapter, though some authors have distinguished among them. The information in this chapter is needed throughout this book—risk scores are used to adjust for differences among patients in analysis of control charts and in many other topics in this book. The chapter focuses on the design and use of the multimorbidity (MM) index (one approach to quantifying patients’ prognoses). Farrokh Alemi and colleagues (Alemi et al. 1999; Alemi and Prudius 2004; Alemi and Uriyo 2011; Kheirbek, Alemi, and Fletcher 2015;Kheirbek, Alemi, and Zargoush 2013; Levy et al. 2015) ,[[2]](#endnote-2),[[3]](#endnote-3),[[4]](#endnote-4),[[5]](#endnote-5),developed this tool to account for the prognosis of patients with multiple diagnoses. The index can be easily constructed inside electronic health records (EHRs) using SQL. It calculates the risk associated with any diagnosis through using *likelihood ratios*, a concept that will be further clarified in this chapter. The MM index uses Bayesian probability models to aggregate the impact of multiple predictors; therefore, this chapter also introduces the reader to probability models and Bayesian calculations.

# [H1] Alternatives to the Multimorbidity Index

Several other investigators have also proposed methods for predicting a prognosis from a patient’s history. Charlson and colleagues were among the first investigators to do so (Charlson et al. 1987). These scholars developed an index that predicted mortality from 22 broad disease categories, including one category for all heart diseases, another for AIDS, and still another for all cancers.Deyo, Cherkin, and Ciol (1992); Romano, Roos, and Jollis (1993); Roos and colleagues (1996); and D’Hoore, Sicotte, and Tilquin (1993) attempted to improve on the initial Charlson index by modifying the broad categories and dropping or adding new categories. Elixhauser and colleagues (1998) continued these modifications by creating a list of 30 broad categories of comorbidities, and van Walraven et al. (2009) organized these categories into an index. It may be helpful to examine how the van Walraven version of the Elixhauser index works.

The Elixhauser comorbidity index predicts the probability of mortality for patients based on their hospital diagnoses using the International Classification of Diseases (ICD) codes. Each time a patient is hospitalized, 5–15 diagnoses are used to report what the patient was treated for. The Elixhauser index uses a select set of these diagnoses to classify the patient into 32 categories. ICD-9 classification was used from approximately 1979 to 2015, when ICD-10 became mainstream in the United States. Exhibit 5.1 shows how some of these categories are organized.

**[INSERT EXHIBIT]**

**Exhibit 5.1** Selected Coding Algorithms for Elixhauser Comorbidities

|  |  |  |  |
| --- | --- | --- | --- |
| *Comorbidities (van Walraven Score)* | *Elixhauser’s Original*  *ICD-9-CM* | *Elixhauser’s AHRQ-Web ICD-9-CM* | *ICD-10* |
| Congestive heart failure, score = 7 | 398.91, 402.11, 402.91, 404.11, 404.13, 404.91, 404.93, 428.x | 398.91, 402.01, 402.11, 402.91, 404.01, 404.03, 404.11, 404.13, 404.91, 404.93, 428.x | I09.9, I11.0, I13.0, I13.2, I25.5, I42.0, 142.5–I42.9, I43.x, I50.x, P29.0 |
| Paralysis, score = 7 | 342.0. 342.1, 342.9–344.x | 342.x–344.x, 438.2–438.5 | G04.1, G11.4, G80.1, G80.2, G81.x, G82.x, G83.0–G83.4, G83.9 |
| Lymphoma, score = 9 | 200.x–202.3x, 202.5–203.0, 203.8, 238.6, 273.3, V10.71, V10.72, V10.79 | 200.x–202.3, 202.5–203.0, 203.8, 238.6, 273.3 | C81.x–C85.x, C88.x, C96.x, C90.0, C90.2 |
| Metastatic cancer, score = 12 | 196.x–199.x | 196.x–199.x | C77.x–C80.x |
| Obesity, score = −4 | 278.0 | 278.0 | E66.x |
| Fluid and electrolyte disorders, score = 5 | 276.x | 276.x | E22.2, E86.x, E87.x |
| Depression, score = –3 | 300.4, 301.12, 309.0, 309.1, 311 | 300.4, 301.12, 309.0, 309.1, 311 | F20.4, F31.3–F31.5, F32.x, F33.x, F34.1, F41.2, F43.2 |
| *Source*: Adapted from Quanet al. (2005) and van Walraven et al. (2009). | | | |

**[END EXHIBIT]**

Walraven and colleagues (2009) provide a scoring for the Elixhauser categories. These scores were derived from predicting mortality of the patients after hospitalization. The authors suggest the rounded scores for each category (for some of these scores, see exhibit 5.1). The score is for any diagnosis that falls into the category. Thus, all types of cancers (e.g., brain and skin) are scored the same. All psychoses are scored the same, independent of their severity. Different categories add different points to the overall score. For example, metastatic cancer adds 12 points. Paralysis adds 7 points. Obesity and depression remove 4 points, as they have negative scores. Keep in mind that negative scores do not make clinical sense, as a disease almost never improves prognosis. These negative scores may make statistical sense. The negative scores typically reflect the impact of confounding among comorbidities (Alemi et al. 2016). Many diseases are not scored, or they are scored as 0. A score of 0 means that the disease is not one of the severe illnesses that typically increase the risk of mortality. Among diseases that are not scored are many diseases that clinicians consider serious illnesses, such as diabetes or arthritis. These unusually lax scoring procedures were developed for ease of use before the availability of massive databases that allow the assessment of the contribution of each disease.

Some investigators (e.g., Quan et al. 2005) have sidestepped differential point systems by scoring each diagnosis category as 1 point and adding the scores. The overall Quan score ranges from 31 to 0 and indicates the number of Elixhauser categories that are present. Diseases such as metastatic cancer and obesity are scored as if they have the same risk of mortality, which again does not make clinical sense but may be sufficient for some analyses.

In contrast to other approaches, the MM index does not rely on broad disease categories. For example, it does not score the 32 categories in the Elixhauser index—instead, it measures each of the underlying diagnoses in these categories. In addition, it scores diseases not included in any categories. In essence, it measures each diagnosis in the patient’s medical history. In Elixhauser and other similar selective methods, diagnoses that fall into broad categories are scored as if they have the same risk. For example, consider the variation in mortality among the 28 diagnoses in the “secondary malignancies” category that are used in variants of the Elixhauser index. In a recent study of the prognosis of heart failure patients (Kheirbek, Alemi, and Fletcher 2015), patients who also had a secondary malignant neoplasm of the brain and spinal cord had an odds ratio of mortality equal to 17.28. In comparison, those who had another variant of a secondary malignancy (i.e., a secondary neuroendocrine tumor of distant lymph nodes) had an odds ratio of 2.43. The same category, secondary malignancies, includes diagnoses that have nearly a ninefold difference in mortality risks. Grouping all secondary malignancies into one category oversimplifies the situation. The MM index does not do so. Because it does not do so, it is designed to be more accurate. The key feature of the MM index is that it is built from thousands of diagnoses, without classifying these diagnoses into categories.

# [H1] The Theory Behind Multimorbidity Index

To effectively model the relationship between thousands of diagnoses and mortality, the MM Index uses the *Bayes data mining model*. For predicting that the patient will be alive, shown by A, after diagnosis D, the Bayes formula is:

**[INSERT EQUATION]**

.

**[END EQUATION]**

The formula states that the probability of being alive given a diagnosis, known as *posterior probability*, can be calculated from , which is the likelihood of observing the diagnosis among living patients. If we show death as Aʹ, then the odds of being alive can be calculated from the ratios

**[INSERT EQUATION]**

.

**[END EQUATION]**

This formula, known as the *odds form* of the Bayes theorem, states that the odds of being alive are the product of , the likelihood ratio of being alive, times , the prior odds of being alive. Under the assumption of independence, the likelihood ratio associated with medical history (i.e., a collection of diseases) is calculated as the product of the likelihood ratio of each of the diseases. In this approach, one assumes that the impact of each disease on mortality is independent from other diseases. This assumption is also made in traditional statistical approaches that use linear logistic regression. Even though the assumption is obviously false, numerous studies have shown that the Bayes formula produces predictions that are as accurate as more complicated models that assume interactions among diseases (de Dombal et al. 1992; Gammerman and Thatcher 1991; Hand and Yu 2001; Monti and Cooper 1999; Todd and Stamper 1994). [[6]](#endnote-6),[[7]](#endnote-7),[[8]](#endnote-8),[[9]](#endnote-9),[[10]](#endnote-10),[[11]](#endnote-11),[[12]](#endnote-12),[[13]](#endnote-13),[[14]](#endnote-14),[[15]](#endnote-15), Under the assumption of independence, the Bayes formula tells us how the odds change once we know the patient's medical history:

**[INSERT EQUATION]**

**[END EQUATION]**

In the above equation, indicates the likelihood ratio associated with the diagnostic in the training data set.

**[H1] Estimating Parameters of the MM Index**

To facilitate the use of the MM index in different EHR systems, appendix A contains an SQL code that will estimate the likelihood ratio associated with diagnoses of patients. It relies on ICD‑9 codes; a similar SQL can be run for ICD-10, thereby allowing the EHR to adjust for changes in the coding procedures. In addition, a similar SQL code can be run for procedures, exposure to medications, and categorized physiological measures, as well as age ranges, thereby allowing the prognostic index to reflect more than diagnoses.

## [H1] Calculation of Likelihood Ratios

The likelihood ratio of mortality associated with each diagnosis (Dx) is calculated using the following formula from the portion of the data set aside for training of the model:

**[INSERT EQUATION]**

, or

.

**[END EQUATION]**

The interpretation of the likelihood ratio is relatively simple. The diagnoses with likelihood ratio above 1 increase the odds of mortality. The higher the number, the higher the risk of mortality. Diagnoses with a likelihood ratio less than 1 decrease the odds of mortality. The lower the number, the more likely they are to do so.

A number of studies have already estimated the likelihood ratios associated with different diagnoses. These likelihood ratios were estimated using data from US Department of Veterans Affairs (VA) medical records for heart failure and for nursing home patients and are publicly available (Kheirbek, Alemi, and Fletcher 2015; Levy et al. 2015). To facilitate the estimation of these ratios for other populations, the SQL code is provided in appendix A.

## [H2] Adjustments for Repeated Diagnoses

If a diagnosis repeats itself (i.e., when treatment is not effective and the patient is repeatedly hospitalized to try different treatments for the same diagnosis), the patient’s prognosis changes. It is important to calculate separate likelihood ratios for each repeated diagnosis, as we see here:

**[INSERT EQUATION]**

.

**[END EQUATION]**

## [H2] Adjustment for Combination of Diagnoses

It is important to note that diagnoses are not independent of each other; the joint likelihood ratio of a pair of diagnoses may be different than the product of the likelihood ratios of each diagnosis. Bayes classifiers assume independence even when this assumption is clearly wrong. In sparse large data with thousands of redundant and overlapping predictors—each of which has a similar impact on prognosis—despite wrong assumptions, Bayes classifiers arrive at correct conclusions. Nevertheless, by scoring the combination of diagnoses, one could create a MM index that has more face validity, following clinicians’ perspective that these combinations matter.

## [H2] Adjustments for Diseases of Which No One or Everyone Dies

Many common diseases are associated with no patient mortality, and there are also rare diseases of which every patient dies. In both of these situations, a likelihood ratio cannot be calculated. In these circumstances, Alemi and Prudius (2002) propose the formulas

**[INSERT EQUATION]**

**[END EQUATION]**

In this equation, indicates the number of patients with the diagnosis.

The calculation of the likelihood ratio occasionally leads to situations in which we are dividing by zero. These occur in diagnoses that result in 100 percent survival or 100 percent mortality. In these situations, the likelihood ratio is estimated from the total number of patients with the diagnosis. The following snippet of SQL code shows how the likelihood ratio is calculated:

**[LIST FORMAT]**

CASE

WHEN [Pts with Dx Alive in 6 Months] is null

THEN [Pts with Dx Dead in 6 Months] + 1

WHEN [Pts with Dx Dead in 6 Months] is null

THEN 1/([Pts with Dx Alive in 6 Months] +1)

ELSE

([Pts with Dx Dead in 6 Months]/[Pts Dead])/

([Pts with Dx Alive in 6 Months]/[Pts Alive])

END  
**[END LIST]**

The code calculates the likelihood ratio as the prevalence of the diagnosis among dead patients divided by the same prevalence among alive patients. In calculating likelihood ratios there are two exceptions. In diagnoses of which everyone dies, the code calculates the likelihood ratio as 1 plus the number of cases. In diagnoses after which everyone lives, the code calculates the likelihood ratio as 1 divided by the sum of the number of cases plus 1. Other methods for adjusting likelihood ratios have been reported in the literature, including adding a fraction of a case to either the denominator or the numerator to avoid division by zero. The adjustment used here has the advantage that it is proportional to the number of patients with the diagnosis. For example, if all 100 patients with a disease died, then the assigned likelihood ratio is 101. If there were only 1 patient with the diagnosis and he died, then the assigned likelihood ratio is 2. In this manner, the assigned likelihood ratio is larger in diagnoses that occur often but everyone dies.

## [H1] Adjustment for Rare Diseases

Although the MM index is derived from a large data repository, there are several diagnoses that are rare and have insufficient observations to estimate a likelihood ratio. In a minority of cases (e.g., when a patient presented with a diagnosis that was not seen in at least 29 cases in the training set), then the likelihood ratio associated with a broader diagnostic category is used to score the patient.

A typical ICD-9 diagnosis is represented by a five-digit number consisting of three initial digits, a period, and two additional digits. The first three digits represent a disease category. Each additional digit after the period represents further refinements. If the patient’s diagnosis is rare, then one could use the likelihood ratio for a broader category of the diagnosis that repeats more often (e.g., by dropping the last digit in the diagnosis code). Appendix A includes the SQL code that can be used to estimate the likelihood ratios associated with three-, four-, and five-digit codes from data in EHRs.

## [H1] Adjustment for Revision 10

To date, the MM index has been evaluated using diagnoses coded with ICD-9. In the ICD-10, a sixth digit was added to further clarify the disease categories. The procedures described in this paper and the computer code provided in appendix A can be used to estimate the prognosis of each code in ICD-10. Because ICD-10 has ten times more codes than ICD-9, reliable estimates for this version cannot be made until data sets ten times are available. Even when the data are available, many disease codes in ICD-10 are unlikely to occur with a frequency sufficient for the prognosis for these codes to be estimated reliably. When ICD-10 codes cannot be estimated reliably, investigators should combine data and rely on higher order codes in ICD-9, using the procedures explained earlier for estimating rare diseases. If ICD-10 codes can be estimated reliably, then these codes should be used instead of ICD-9. By using this method, the best description of the patient should be used. When the estimate is not available, a less precise description should be used.

## [H1] Sample Size Needed to Construct the MM Index

To derive the MM index, it is important to recognize that a large number of parameters are estimated. There are more than 14,000 ICD codes, and in most populations 3,000–5,000 unique diagnoses occur. This means that approximately 3,000–5,000 parameters must be estimated. There are a number of ways to estimate the sample size that would be needed for such a large number of determinations. Some investigators have suggested that the power of the investigation depends on the ratio of the number of subjects to the number of variables, using heuristics such as 10 times (Garson 2008; Hutcheson and Sofroniou 1999; MacCallum et al. 1999)or 20 times (Hogarty et al. 2005) the number of subjects compared to the number of variables in the model (e.g., to estimate 5,000 parameters, 100,000 subjects would be needed). In large data sets and in most of the analyses reported here, the total sample size exceeds 30 times the number of diagnoses, suggesting the estimated model has sufficient power to detect the needed parameters. Other statisticians suggest alternative ways of determining the minimum sample size for estimating likelihood ratios (Hsieh et al. 2003; Hsieh, Bloch, and Larsen 1998; McDonald and Krane 1979).

# [H1] Cross-Validation

The likelihood ratios are estimated from the training set; the predictions are made in a different data set. Statisticians randomly set aside data for the purpose of checking the validity of the model. This process is called *cross-validation*. Typically, fivefold cross-validation is done—the analysis is done five times, each time randomly setting aside one-fifth of the data for validation. The reported accuracy is the average across these five sets. Cross-validation protects against modeling random noise in the training set as if it is real change in the data.As the number of predictors increases, the chance of modeling noise in the training set increases. Because we have thousands of predictors, the chance of modeling noise is large; it is important to cross validate the predictions.

In SQL, random numbers have seed values. (The term *random* *seed values* refers to the starting point of the random number generator.) If the seed value does not change, the same random digit will be generated. One way to randomly select patients to be included in the training and validation set is to use their IDs. One must first convert the ID to a number. If patients are unique, then row numbers can be used as a seed for the random number generator.

## [H2] Prediction versus Detection

Multimorbidity indexes rely on diagnoses to predict outcomes. In EHRs, the statistician has access to diagnoses before and after observing the outcome. Some diagnoses occur before and some after. This is not of concern if our outcome is mortality—no diagnoses (with the exception of autopsy reports) occur after the patient has died. A critical question in predictive modeling is whether the predictors (the patients’ diagnoses) should be limited to the period before the observation of the outcome. Many statisticians feel that, in multivariate analysis, independent variables should occur before the outcome of interest. This is not the case in MM indexes.

A likelihood ratio measures the impact of a diagnosis on the outcome; these ratios do not distinguish whether the diagnosis has occurred before or after the outcome. In this sense, likelihood ratios are measures of associations. They show the association between the diagnosis and outcome. A strong likelihood ratio does not imply that the disease causes the outcome. It simply is a measure of association.

Likelihood ratios can be used in two ways. In one approach, one tries to detect an event that has already occurred but may not have been reported. For example, one might want to detect whether the patient has undiagnosed diabetes or an unreported substance abuse disorder. In another approach, one tries to predict an event that has not yet occurred. For example, one might want to predict whether certain patients will, in the future, abuse pain medications. These two approaches differ the variables they use for predictors. The detection approach can rely on consequences of the outcome. For example, it can rely on repeated skin infections to detect injection of drugs (skin infections are one consequence of drug injection). Here the diagnoses are occurring after substance abuse; a particular pattern in these diagnoses points to the existence of earlier substance abuse.

The prediction approach is different. In forecasting, the statistician must only use the information that is already available. The mechanism by which the outcome occurs can suggest a reasonable set of predictors. Repeated prescription of opioids for surgical pain, for example, may be a predictor for abuse in the future; it describes the mechanism by which increased opioid use occurs. Borderline A1c levels may be a good predictor for future diabetes, as they describe how diabetes comes about. In prediction, one should rely on events that precede the outcome; consequences of the outcome are no longer reasonable predictors.

In deciding which variables to include as predictors in a multimorbidity model, the first task is to decide whether one is predicting or detecting. Of course, if the outcome of interest is mortality, the choice is simple: we are predicting mortality; one rarely wants to detect whether the patient has died if they are not aware of it. For diabetes, anemia, or other conditions, one set of predictors is useful for detection, and another set is useful for prediction. Careful thought is needed so that the right set of predictors is used.

## [H2] Predicting Odds of Mortality

The odds form of Bayes formula guides how the likelihood ratios of various diagnoses are used to predict mortality. First one selects the medical history of the patient. For each diagnosis one looks up the equivalent code, if none exists then a likelihood ratio of 1 to 1 is assigned. Second, these likelihood ratios are multiplied to get the change in posterior odds of mortality. The following snippet of code shows how the multiplication is done in SQL using log and exponential functions (a visual display of how this code works is provided in appendix B):

**[LIST FORMAT]**

Calculate Probability of Mortality Assuming Equal Priors

SELECT Id

, EXP(SUM(LOG(ABS(IIF(LR is null, 1, LR))))) as Odds

, EXP(SUM(LOG(ABS(IIF(LR is null, 1, LR))))) / (1+EXP(SUM(LOG(ABS(IIF(LR is null, 1, LR)))))) AS Prob

INTO #Predict

FROM #history

WHERE LR<>0

GROUP BY id

**[END LIST]**

Several problems arise because of this code. First, the product of the likelihood ratios may exceed the largest number allowed in the computer. Some patients have hundreds of diagnoses in their records, and the product of these likelihood ratios could be a number too large for any computer.

Second, note that no predictor is allowed to have a likelihood ratio of zero. Log of 0 is infinity and not defined in the computer. A likelihood ratio of zero means that the outcome never happens—it has not even a minute chance of happening. The statistician controls for this situation when she creates likelihood ratios. She modifies all likelihood ratios of zero to be a number close to zero but not zero, allowing minute chances.

Third, the likelihood ratio contains a great deal of confounding. If a relatively benign condition such as hypertension tends to occur with a relatively deadly condition such as heart attack, then the likelihood ratio for hypertension will be overstated. Procedures to remove confounding are available in several published papers (Rosenbaum and Rubin 1983).[[16]](#endnote-16)[[17]](#endnote-17)[[18]](#endnote-18) I recommend the use of stratified covariate balancing (search on the web for *R Package StratifiedBalancing*) to remove confounding. Applying methods of removing confounding to the estimation of survival of patients with co-occurring diagnoses remains an active area of research.

Fourth, because likelihood ratios are averages and not causal, bizarre situations may occur in which a disease could have a likelihood ratio less than 1, suggesting that it reduces the risk of mortality. In reality, diseases do not help a patient but may reduce the risk compared to prior odds of mortality, which reflects what will happen on average.

# [H2] Estimating Probability from Odds of Mortality

Many investigators and clinicians may wish to estimate the probability of mortality at different time intervals (e.g., 6 months, 12 months, or 5 years). The odds of mortality before adjusting for the patient’s medical history are called prior odds. The odds after adjusting for a patient’s medical history are called posterior odds. By using the prior odds of mortality and the likelihood ratios to combine the diagnoses, one can transform the MM index to estimate the probability of mortality in a specific period. According to the Bayes formula, the posterior odds of mortality are calculated as follows:

**[INSERT EQUATION]**

.

**[END EQUATION]**

# [H1] Checking the Accuracy of Predictions

## [H2] Methods of Measuring Accuracy

To report the performance of an index, we need to consider the concepts of *sensitivity* and *specificity*. A sensitive index has a high probability of detecting the condition. In exhibit 5.2, we have an index that predicts mortality. Sensitivity of this index refers to our ability to correctly predict patients who are dead. A specific index has a high probability of detecting the opposite condition—patients who are alive.

Exhibit 5.2 shows how sensitive and specific measures might have different types of errors. In this exhibit, we assume that we have the posterior probability of mortality, as predicted by the MM index. One must assume that when the probability is above a fixed cutoff number, the prediction is that the patient dies. Exhibit 5.2 shows how, at a particular cutoff, the sensitivity and specificity of the predictions can be calculated using the following code:

**[LIST FORMAT]**

Decide on a sample of cutoff levels to try ---

DROP TABLE #OrderedData, #Cutoffs

SELECT Row\_Number()Over(ORDER BY Predicted) as Row

, [Predicted] AS [Prob], Actual

INTO #OrderedData FROM [ROC].[dbo].[Data] order By [Predicted]

SELECT (b.[Prob]+a.Prob)/2 as Cutoff

INTO #Cutoffs

FROM #OrderedData b inner join #OrderedData a

ON a.Row = b.Row+1

INSERT INTO #Cutoffs (Cutoff) VALUES (0.0), (1.0);

SELECT top 5 \* FROM #Cutoffs;

-- Classify predicted scores by comparison to cutoff ----

DROP TABLE #temp1

SELECT cutoff

, IIF(a.[Prob] > b.[Cutoff],1.,0.) AS Predicted

, Actual

INTO #Temp1

FROM #Data a Cross Join #Cutoff b

-- Calculate sensitivity and Specificity ---

SELECT Cutoff

, SUM(CAST(Actual AS FLOAT)\*CAST(Predicted AS FLOAT))/

Sum(CAST(Actual AS FLOAT)) AS Sensitivity

, SUM((1-Predicted)\*(1-Actual))/SUM(1-Actual) AS Specificity

, ROW\_NUMBER() OVER(ORDER BY cutoff DESC) AS rnum

INTO #sensspec

FROM #Temp1

GROUP BY Cutoff

ORDER BY Cutoff

**[END LIST]**

**[INSERT EXHIBIT]**

**Exhibit 5.2** Definition of Sensitivity and Specificity at a Cutoff Value

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | *Probability of Mortality* | |  |
| *Alive (Less Than or Equal to Cutoff)* | *Dead (Greater Than Cutoff)* |
| True condition | Alive | True negative | False negative |
| Dead | False positive | True positive |

**[END EXHIBIT]**

There is a natural trade-off between sensitivity and specificity. A cutoff can be chosen to increase sensitivity but decrease specificity. Different cutoff points produce different levels of sensitivity and specificity. One can predict that all patients will die, in which case we would have a sensitive index, but not a very specific one. Similarly, one can predict that all patients will live, in which case we have a specific prediction but not a sensitive one. To illustrate this point, consider the data in exhibit 5.3. In it, we report the number of cases that fall in different ranges of probability of death, as well as whether the patient lived or died. For example, three patients who fell into the probability range from 0 to 0.2 actually died; 33 patients who fell in this range lived.

**[INSERT EXHIBIT]**

**Exhibit 5.3** Probability of Death and Actual Death

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | *Probability of Death* | | | | | *Total* |
| *0–0.2* | *0.2–0.4* | *0 .4–0.6* | *0 .6–0.8* | *0.8–1* |
| True condition | Alive | 33 | 6 | 6 | 11 | 2 | 58 |
| Dead | 3 | 2 | 2 | 11 | 33 | 51 |
| Total | 36 | 8 | 8 | 22 | 35 | 109 |

**[END EXHIBIT]**

From exhibit 5.3, we can calculate the sensitivity and specificity of the predictions at different cutoff levels. Exhibit 5.4 shows the result. For example, at a cutoff level greater than 0.4, there are 33 + 6 = 39 persons predicted to live who actually live. The specificity (i.e., the portion of living patients correctly predicted) at this cutoff level is 39 ÷ 58 = 0.67. There are 2 + 11 + 33 = 46 people who are predicted to die who actually die. The sensitivity (i.e., the portion of dead patients correctly predicted) at this cutoff level was 46 ÷ 51 = 0.90. Exhibit 5.4 provides the sensitivity and specificity at various cutoff levels.

**[INSERT EXHIBIT]**

**Exhibit 5.4** Sensitivity and Specificity at Different Cutoff Levels

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Predicted Dead if > Cutoff* | *≥ 0* | *>0.2* | *>0.4* | *>0.6* | *>0.8* | *>= 1* |
| Correct alive predictions | 0 | 33 | 39 | 45 | 56 | 58 |
| Correct death predictions | 51 | 48 | 46 | 44 | 33 | 0 |
| Specificity | 0 | 0.57 | 0.67 | 0.78 | 0.97 | 1.00 |
| 1-Specificity | 1 | 0.43 | 0.33 | 0.22 | 0.03 | 0.00 |
| Sensitivity | 1 | 0.94 | 0.90 | 0.86 | 0.65 | 0.00 |
| Random | 1 | 0.94 | 0.90 | 0.86 | 0.65 | 0.00 |

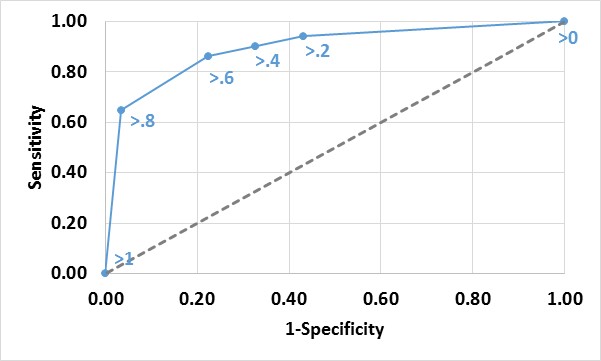
**[END EXHIBIT]**

The receiver operating curve plots sensitivity and specificity of the predictions at different cutoff levels. The convention is to plot 1 minus specificity and not specificity itself. Exhibit 5.5 shows the receiver operating curve for the sensitivity and specificity data in exhibit 5.4. Exhibit 5.5 shows cutoff points so that the reader can see how changes in cutoff points lead to different levels of sensitivity and specificity.

The *area under receiver operating curve* (AROC) estimates the accuracy of the index. An AROC of 1 is perfect prediction; an AROC of .5 is random prediction. The dashed line in exhibit 5.5 shows the random prediction with an area of 0.5. The AROC values range from 0.5 to 1 if the predictions are better than random guesses. The farther the receiver operating curve is from the dashed line, the more accurate the predictions.

**[INSERT EXHIBIT; render in gray scale; make blue line solid black]**

**Exhibit 5.5** Receiver Operating Curve for Data in Exhibit 5.4

****

**[END EXHIBIT]**

To calculate the AROC, we approximate the curve as a series of adjacent trapezoids. The base of the trapezoid is the *x*-axis. The length of the base is the difference of the values of the two points on the *x*-axis and is referred to as *run*. The height of the trapezoid is uneven and corresponds to the height of each point on the *y*-axis. The difference in the two heights is referred to as *rise*. The area for the trapezoid consists of two elements, a triangle, Rise × Run ÷ 2, plus a rectangle, Run × minimum (height). Exhibit 5.6 shows the calculation of AROC for the data in exhibit 5.4. For example, between the two cutoff points >0.2 and >0.4, the triangle has the rise in sensitivity of 0.04 and a run of 0.10. The rectangle below the triangle has run of 0.10 and height of 0.90 (the minimum of sensitivity at these two points). The net triangle and rectangle area are 0.004 and 0.09, for a total area of 0.10. Across all cutoff points the AROC is 0.89, which is relatively large and close to the maximum AROC of 1. These calculations can be easily done in SQL using the following code:

**[LIST FORMAT]**

-- Calculate the area of each section

DROP TABLE #Areas

SELECT

CASE WHEN b.sensitivity> a.sensitivity THEN b.sensitivity ELSE a.sensitivity END \* Abs(b.specificity-a.specificity) + Abs(b.sensitivity - a.sensitivity) \* Abs(b.specificity-a.specificity)/2 AS area

INTO #areas

FROM #sensspec a inner join #sensspec b ON b.rnum-1 = a.rnum

SELECT \* FROM #Areas

-- Calculate the total area under the curve

SELECT SUM (area) AS area FROM #areas

**[END LIST]**

**[INSERT EXHIBIT]**

**Exhibit 5.6** Calculation of Area under Receiver Operating Curve

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | *>0 to >0.2* | *>0.2 to >0.4* | *>0.4 to >0.6* | *>0.6 to >0.8* | *>0.8– >1* | *Total* |
| Height of triangle | 0.06 | 0.04 | 0.04 | 0.22 | 0.65 |  |
| Width of triangle or rectangle | 0.57 | 0.10 | 0.10 | 0.19 | 0.03 |
| Area of triangle | 0.02 | 0.00 | 0.00 | 0.02 | 0.01 |  |
| Height of rectangle | 0.94 | 0.90 | 0.86 | 0.65 | 0.00 |  |
| Area of rectangle | 0.54 | 0.09 | 0.09 | 0.12 | 0.00 |  |
| Total area under curve | 0.55 | 0.10 | 0.09 | 0.14 | 0.01 | 0.89 |

## [END EXHIBIT]

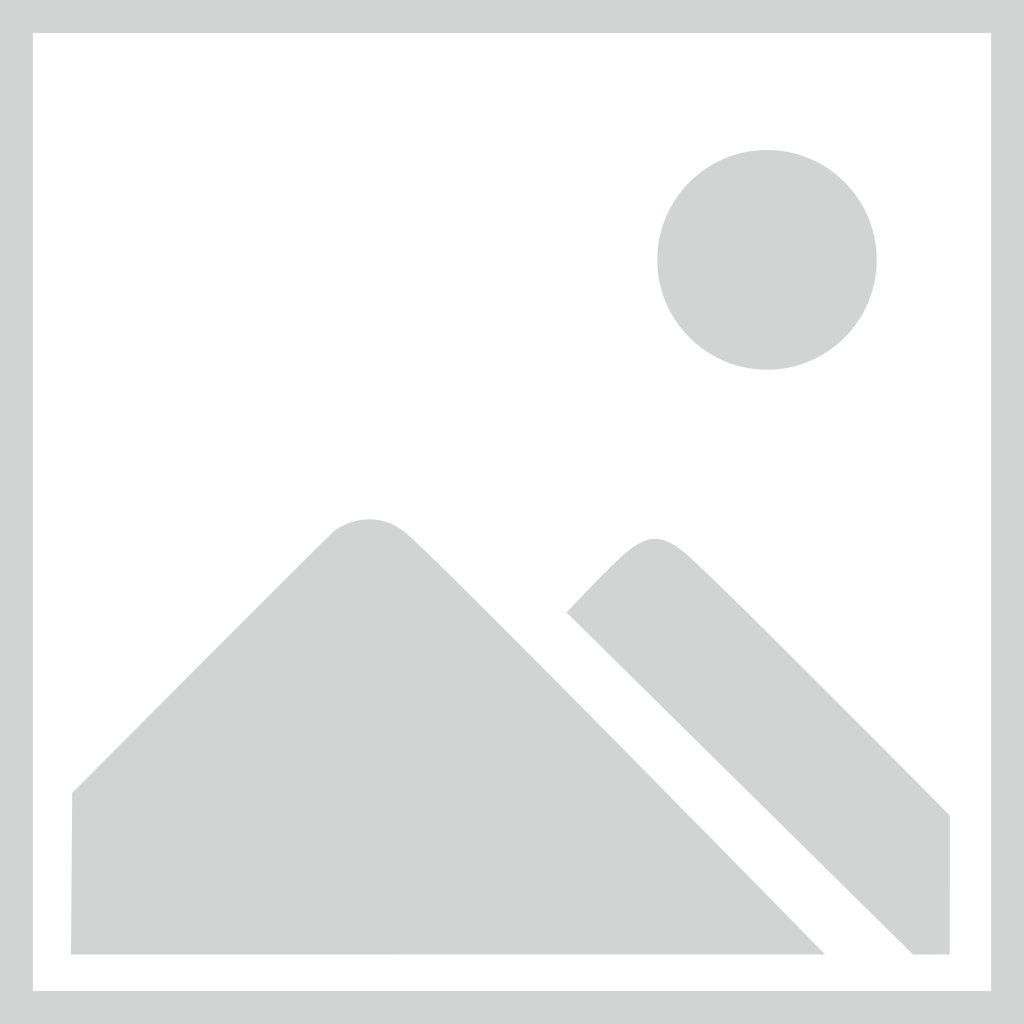
## [H1] MM Index Compared to Physiological Markers

Clinicians are often skeptical about predicting patient outcomes from diagnoses. They believe, and their day-to-day experiences show them, that physiological markers such as blood pressure are most predictive of outcomes. Diagnoses and comorbidities seem too general to identify who is at risk of dying. In addition, clinicians may be concerned with diagnoses being coded in error. Because the healthcare literature is replete with studies that report the use of physiological markers as prognostic indicators, clinicians assume that the use of such markers is the gold standard for prognostication. However, these concerns are not accurate, and as we will review shortly, patients’ diagnoses are highly predictive of mortality in six months.

Alemi and colleagues (1999) examined the prognosis of patients with HIV or AIDS by using an MM index and found that the index was more predictive of patients’ survival than an index developed from an average of physiological indicators, such as CD4 T lymphocyte counts. In as-yet unpublished data (Alemi et al. 2015), the performance of the diabetes MM index was compared to the accuracy of hemoglobin A1c levels for 468,867 diabetic patients. Exhibit 5.7 shows that the sensitivity and specificity of the MM index were superior to prognostic indicators based on HbA1c levels alone. The AROC curve for predicting six-month mortality for HbA1c levels was 0.652; in contrast, AROC for the MM index was 0.812. The MM index was therefore 1.25 times more accurate than the widely used HbA1c levels.

**[INSERT EXHIBIT]**

**Exhibit 5.7.** Comparison of MM Index and HbA1C in Predicting 12-Month Mortality



**[END EXHIBIT]**

In a study of patients in intensive care, we compared the accuracy of the MM index to 13 physiological markers (Min et al. 2016).These markers included sodium, blood urea nitrogen, creatinine, glucose, albumin, bilirubin, white blood cell count, hematocrit, PaO2, PaCO2, pH, eGFR, and lactic acid. We examined the six-month and 12-month mortality of 442,692 unique patients seen in 87 intensive care units of VA medical centers between 2003 and 2013. The MM index, relying solely on diagnostic codes, yielded an AROC of 0.84. In contrast, the logistic regression based on the combined impact of 13 physiological markers yielded an AROC of 0.65. These studies show that the MM index is more accurate in predicting mortality for many existing physiological markers. What matters in predicting prognosis seems to be the history of the patient’s diagnoses, as opposed to any particular physiological marker.

## [H1] MM Indexes Compared to Other Diagnoses-Based Indexes

The MM index has been repeatedly shown to predict mortality better than a wide variety of other tools developed for the purpose. For example, in a study of intensive care patients, we compared the performance of the MM index to comorbidity categories in the Elixhauser list, immunosuppressant medication use, and age. The study reported the six-month and 12-month mortality of 442,692 unique intensive care patients. The MM Index relied on 5,695 diagnoses codes. The cross-validated AROC for the MM index was 0.84. In contrast, the AROC for immunosuppressant medication use was 0.59; for age it was 0.60; for Elixhauser comorbidities it was 0.69; and for all combined variables (including physiological markers), it was 0.80. As in other studies, these differences were all cross-validated and statistically significant. The fact that the MM index, which scores each diagnosis, was more accurate than the Elixhauser categories of comorbidities suggests that grouping diagnoses into broad diagnostic categories reduces the accuracy of predictions.

# [H1] Use of the MM Index for Predicting One Case Example

To illustrate the application of the MM Index to specific cases, we use a case from a recent analysis of the prognosis of nursing home residents. The resident was 81 years old and had ten diagnoses during the last hospital admission (exhibit 5.8). The likelihood ratio of each diagnosis was drawn from the [George Mason University Dataverse](http://arc.irss.unc.edu/dvn/dv/gmu;jsessionid=115a49b698762bb2002983efb8e6) (Levy et al. 2014). For one diagnosis, “chronic airway obstruction, not COPD, and not elsewhere classified,” the Dataverse does not provide any information, and therefore this diagnosis was ignored and scored with a likelihood ratio of 1. The MM score, the product of all likelihood ratios, was calculated as 45.07. In this database, the prior odds of mortality for this population were 0.16. The prior odds were multiplied by the product of the likelihood ratios to obtain the posterior odds. The posterior odds can be expressed as a probability by dividing the posterior odds by one plus the odds. This patient’s diagnoses have resulted in a probability of 0.88 for dying in the next six months. The likelihood ratios in exhibit 5.8 also can be used to explain the prediction. Likelihood ratios above 1 indicate diagnoses that increased the odds of mortality. Based on these data, the main reason for the high estimate of mortality pertained to the patient’s lung cancer, anorexia, and cachexia. Each more than doubled the risk of mortality. Kidney disease also contributed to the high probability of mortality, but to a lesser extent.

**[INSERT EXHIBIT]**

**Exhibit 5.8 Calculation of the MM Index from Diagnoses of an 81-Year-Old Resident**

|  |  |
| --- | --- |
| *Description of Diagnosis* | *Likelihood Ratio* |
| 1. Malignant neoplasm of upper lobe, bronchus, or lobe | 3.18 |
| 1. Other specified chronic ischemic heart disease | 1.41 |
| 1. Abdominal aneurysm without mention of rupture | 1.04 |
| 1. Peripheral vascular disease, unspecified | 0.96 |
| 1. Chronic airway obstruction, not COPD, and not elsewhere classified | Not found |
| 1. Chronic kidney disease, stage IV (severe) | 1.45 |
| 1. Secondary hyperparathyroidism of renal origin | 1.03 |
| 1. Anorexia | 2.16 |
| 1. Nausea and vomiting | 1.02 |
| 1. Cachexia | 3.06 |

|  |  |
| --- | --- |
| *Steps in Calculation* | *Results* |
| 1. MM index (product of all likelihood ratios) | 45.07 |
| 2. Prior odds for all residents | 0.16 |
| 3. Posterior odds for this case (prior odds times MM index) | 7.21 |
| 4. Probability of mortality (posterior odds divided by 1 plus posterior odds) | 0.88 |

**[END EXHIBIT]**

# [H1] Summary

This chapter reviewed the ideas behind and the accuracy of the MM index. MM indexes have had a higher AROC than various physiological measures of prognosis, including ejection fraction for heart failure, HbA1c levels for diabetic patients, and 13 physiological measures for patients in intensive care units.

The MM index is more accurate than existing diagnosis-based indexes such as variants of the Charlson and Elixhauser indexes. To the best of our knowledge, the MM index is the first that scores each disease separately, rather than grouping similar diagnostic codes into broad categories. The improved accuracy of the MM index may be a feature of its scoring each diagnosis without grouping them.

The comprehensive inclusion of thousands of comorbidities in the MM index makes its use in clinical settings difficult, but use of the MM index is more practical now that many clinics employ EHRs. These organizations have access to the patients’ diagnostic history, can score the patient’s prognosis, and can explain the top two or three reasons for the predicted prognosis. The use of the MM index in a clinical setting will be akin to the use of any laboratory test, in which the results are available, but details of how the results were obtained are masked.

The MM index can be further improved in a number of ways, including the examination of interaction among diseases and removal of confounding in estimates of likelihood ratios. In addition, the MM index may be improved if medications, physiological markers, or procedures are used to predict prognosis.

The use of the MM index in informing patients and their family members about chances of death is fraught with difficulties. The MM index reflects average probabilities associated with a disease. No patient is the average patient, given that patients are likely to experience a combination of comorbidities that may radically differ from the average patient. MM index scores may not be appropriate for patients who do not want to deal with averages, thinking that they will beat the odds and have lower than average risk of mortality. Furthermore, a discussion of mortality with patients may be understood differently if it is framed in terms of survival as opposed to mortality. Probabilities may be misunderstood, and patients may prefer to know expected survival days as opposed to the probability of mortality. Obviously, any discussion of prognosis with patients requires empathetic communication on the part of clinicians. These and other limitations continue to frustrate efforts to make data on prognostic information available to patients and their families. Additional research is needed to clarify how best to communicate prognostic information to individual patients.

The MM index can be used easily in policy analysis, decision support, and program evaluation. In these uses, the index enables assessment of comparative effectiveness of treatment. Because the MM index is more accurate than existing comprehensive diagnosis-based indexes, we use it throughout this book.

# [H1] Supplemental Resources

A problem set, solutions to problems, multimedia presentations, SQL code, and other related material are on the course website.

**[H1] References**

Alemi, F., S. Avramovic, D. Aron, and M. Hua. 2015. “Prognosis of Diabetic Patients.” Unpublished paper, George Mason University.

Alemi, F., C. Levy, B. A. Citron, A. R. Williams, E. Pracht, and A. Williams. 2016. “Improving Prognostic Web Calculators: Violation of Preferential Risk Independence.” *Journal of Palliative Medicine* 19 (12): 1325–30.

Alemi, F., and V. Prudius. 2004. A Mathematical Theory for Identifying and Measuring Severity of Episodes of Care. US Patent 10,054,706, filed on January 21, and issued April 20, 2010.

Alemi, F., and M. Uriyo. 2011. “Accuracy of Claims-Based Measures of Severity of Childhood Illnesses.” *Health Outcomes Research in Medicine* 2: e71-e78. http://openonlinecourses.com/ehr/Accuracy%20of%20claims%20based%20measures%20of%20childhood%20severity%20of%20illness.pdf.

Alemi, F., L. Walker, J. Carey, and J. Leggett. 1999. “Validity of Three Measures of Severity of AIDS for Use in Health Services Research Studies.” *Health Services Management Research* 12 (1): 45–50.

Charlson, M. E., P. Pompei, K. L. Ales, and C. R. MacKenzie. 1987. “A New Method of Classifying Prognostic Comorbidity in Longitudinal Studies: Development and Validation.” *Journal of Chronic Diseases* 40 (5): 373–83.

Deyo, R., D. C. Cherkin, and M. A. Ciol. 1992. “Adapting a Clinical Comorbidity Index for Use with ICD-9-CM Administrative Databases.” *Journal of Clinical Epidemiology* 45 (6): 613–9.

de Dombal, F. T., D. J. Leaper, J. R. Staniland, A. McCann, and J. Horrock. 1972. “Computer-Aided Diagnosis of Acute Abdominal Pain.” *British Medical Journal* 2 (5804): 9–13.

D'Hoore, W., C. Sicotte, and C. Tilquin. 1993. “Risk Adjustment in Outcome Assessment: The Charlson Comorbidity Index.” *Methods of Information in Medicine*. 32 (5): 382–87.

Elixhauser, A., C. Steiner, D. R. Harris, and R. M. Coffey. 1998. “Comorbidity Measures for Use with Administrative Data.” *Medical Care*. 36 (1): 8–27.

Gammerman, A., and A. R. Thatcher. 1991. “Bayesian Diagnostic Probabilities Without Assuming Independence of Symptoms.” *Methods of Information in Medicine* 30: 15–22.

Garson, D. G. 2008. “Factor Analysis: Statnotes.” North Carolina State University Public Administration Program. Accessed March 22. www2.chass.ncsu.edu/garson/pa765/factor.htm.

Hand, D. J., and K. Yu. 2001. “Idiot’s Bayes—Not So Stupid After All?” *International Statistical Review* 69 (3): 385–98.

Hogarty, K. Y., C. V. Hines, J. D. Kromrey, J. M. Ferron, and K. R. Mumford. 2005. “The Quality of Factor Solutions in Exploratory Factor Analysis: The Influence of Sample Size, Communality, and Overdetermination.” *Educational and Psychological Measurement* 65: 202–226.

Hutcheson, G., and N. Sofroniou. 1999. “The Multivariate Social Scientist: Introductory Statistics Using Generalized Linear Models.” Thousand Oaks, CA: SAGE Publications.

Hsieh, F. Y., D. A. Bloch, and M. D. Larsen. 1998. “A Simple Method of Sample Size Calculation for Linear and Logistic Regression.” *Statistics in Medicine* 17 (14):1623–634.

Hsieh, F. Y., P. W. Lavori, H. J. Cohen, and J. R. Feussner. 2003. “An Overview of Variance Inflation Factors for Sample-Size Calculation.” *Evaluation and the Health Professions*. 26 (3): 239–57.

Kheirbek, R. E., F. Alemi, and R. Fletcher. 2015. “Heart Failure Prognosis: Comorbidities Matter.” *Journal of Palliative Medicine* 18 (5): 447–52.

Kheirbek, R. E., F. Alemi, and M. Zargoush. 2013. “Comparative Effectiveness of Hypoglycemic Medications Among Veterans.” *Journal of Managed Care and Specialty Pharmacy* 19 (9): 740–44.

Levy, C., R. E. Kheirbek, F. Alemi, J. Wojtusiak, B. Sutton, A. R. Williams, and A. Williams. 2015. “Predictors of Six-Month Mortality Among Nursing Home Residents: Diagnoses May Be More Predictive Than Functional Disability.” *Journal of Palliative Medicine* 18 (2): 100–106.

MacCallum, R. C., K. F. Widaman, S. Zhang, and S. Hong. 1999. “Sample Size in Factor Analysis.” *Psychological Methods* 4: 84–99.

McDonald, R. P, and W. R. Krane. 1979. “A Monte Carlo Study of Local Identifiability and Degrees of Freedom in the Asymptotic Likelihood Ratio Test.” *British Journal of Mathematical and Statistical Psychology* 32: 121–32.

Min, H., S. Avramovic, J. Wojtusiak, R. Khosla, R. D. Fletcher, F. Alemi, and R. E. Kheirbek. 2017. “Comprehensive Multimorbidity Index for Predicting Mortality in Intensive Care Patients.” *Journal of Palliative Care* 20 (10): 35–41.

Monti, S., and G. F. Cooper. 1999. “A Bayesian Network Classifier That Combines a Finite Mixture Model and a Naive Bayes Model.” Presented at the Fifteenth Conference on Uncertainty in Artificial Intelligence, Stockholm, Sweden.

Quan, H., V. Sundararajan, P. Halfon, A. Fong, B. Burnand, J. C. Luthi, L. D. Saunders, C. A. Beck, T. E. Feasby, and W. A. Ghali. 2005. “Coding Algorithms for Defining Comorbidities in ICD-9-CM and ICD-10 Administrative Data.” *Medical Care* 43 (11): 1130–39.

Romano, P. S., L. L. Roos, and J. G. Jollis. 1993. “Adapting a Clinical Comorbidity Index for Use with ICD-9-CM Administrative Data: Differing Perspectives.” *Journal of Clinical Epidemiology* 46 (10): 1075–79.

Roos, L. L., R. K. Walld, P. S. Romano, and S. Roberecki. 1996. “Short-Term Mortality After Repair of Hip Fracture: Do Manitoba Elderly Do Worse?” *Medical Care* 34 (4): 310–26.

Rosenbaum, P. R., and D. B. Rubin. 1983. “The Central Role of the Propensity Score in Observational Studies for Causal Effects.” *Biometrika* 70: 41–55.

Todd, B. S., and R. Stamper. 1994. “The Relative Accuracy of a Variety of Medical Diagnostic Programmes.” *Methods of Information in Medicine* 33: 402–416.

van Walraven, C., P. C. Austin, A. Jennings, H. Quan, and A. J. Forster. 2009. “A Modification of the Elixhauser Comorbidity Measures into a Point System for Hospital Death Using Administrative Data.” *Medical Care* 47 (6): 626–33.

**[H1] Appendix A**

The structured query language (SQL) code for the derivation of parameters of the multimorbidity (MM) index is provided in the next paragraph. Similar SQL code is also available for measurement of episodes of illness (Alemi and Walters 2006).Because the source of data may include millions of records, the code is written in steps; each step generates a temporary file that is used in subsequent steps. In this fashion, if for some reason the server operations are interrupted, intermediary results are still available, and users can start the analysis from the point of the interruption and not from the beginning.

In the first step, we select 80 percent of the cases for calculations of the likelihood ratios and set aside the remaining cases for validation purposes:

**[LIST FORMAT]**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

--> 1. Make data for training cases from 602,050 patients

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

/\* Split data into training and validation sets. Randomly select 80% of cases for training. This is needed for cross-validating study findings. \*/

Print 'Generate a random number'

Drop table #tmp1

SELECT DISTINCT [ScrSSN], Rand(cast(newid() as varbinary)) AS RR

INTO #tmp1

FROM [Src[CohortScrSSN]

Go -- (948,236 row(s) affected)

Print 'Select 80% of cases'

DROP TABLE #tmp2

SELECT ScrSSN

INTO #tmp2

FROM #tmp1

WHERE RR<.8

Go -- (759107 row(s) affected)

Print 'Get date of birth and death'

DROP TABLE #tmp3

SELECT b.[ScrSSN] as ssnID

, Max(DateOfDeath) AS DeathDate

, Max(DateOfbirth) AS BirthDate

INTO #tmp3

FROM #tmp2 a inner join [Src[CohortCrosswalk] b on a.scrssn=b.scrSSN

inner join [Src[SPatient\_SPatient] c on b.SCRSSN = c.SCRSSN and b.sta3n=c.sta3n

WHERE DateOfBirth is not null

GROUP BY b.scrSSN

Go -- (716272 row(s) affected)

Print 'Store as training cases'

DROP TABLE dflt.tcases

SELECT \*

INTO dflt.tCases

FROM #tmp3

Go -- (716272 row(s) affected)

/\*

**[END LIST]**

Make data for training cases from 602,050 patients. In the second set of steps, we calculate the length of service to a patient. This information is used to exclude patients who are not deceased but have no visit after a certain date (at least one year). These patients are likely to have changed providers and may be receiving care outside of the VA.

**[LIST FORMAT]**

Use [Database]

>\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

--> Calculate last encounter.

USE [Database]

Print 'Get last outpatient visit dates for training cases'

DROP TABLE #LastVisit

SELECT t.ssnID, Max(VisitDateTime) as LastVisit, Min(visitdatetime) as FirstVisit

INTO #LastVisit

FROM [dflt[tcases] t left join [Src[CohortCrosswalk] c on t.ssnID = c.scrssn

left join [Src[Outpat\_Visit] o on c.SCRSSN=o.SCRSSN And c.Sta3n=o.sta3n

WHERE VisitDateTime is not null

GROUP BY t.ssnID

Go -- (714820 row(s) affected) 39 minutes

Print 'Get last hospital admission date for training cases'

DROP TABLE #LastAdm

SELECT t.ssnID

, max([AdmitDateTime]) as LastAdmit

, min(admitdatetime) as FirstAdmit

INTO #LastAdm

FROM [dflt[tcases] t left join [Src[CohortCrosswalk] c on t.ssnID = c.scrssn

left join [Src[Inpat\_InpatientDiagnosis] i on c.SCRSSN = i.SCRSSN And c.Sta3n=i.sta3n

WHERE admitdatetime is not null

GROUP BY t.ssnID

Go -- (661407 row(s) affected)

Print 'Combine last visit and last admit'

DROP TABLE #t1

SELECT ssnID, lastadmit, Firstadmit into #t1 from #lastAdm

UNION all

SELECT ssnID, LastVisit, FirstVisit from #lastvisit

Go -- 1376227 row(s) affected)

Print 'Select last encounter'

DROP TABLE dflt.LastEnc

SELECT ssnid, max(lastadmit) as lastEnc, min(firstadmit) as FirstEnc

INTO dflt.lastEnc

FROM #t1

GROUP BY ssnid

Go -- (714854 row(s) affected)

Print 'Remove cases with less than 1 year of follow up'

DELETE FROM Dflt.LastEnc WHERE datediff(dd,firstEnc,LastEnc)<365

Go -- (19928 row(s) affected) Deleted

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

--> \*\*\*\*\*\*\*\*\*\*\*\* Calculate Likelihood Ratios from training cases

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

USE [Datebase]

Print 'Delete cases with less than 1 year follow up'

DROP TABLE #tcases

SELECT t.\*, lastEnc

INTO #tcases

FROM dflt.tcases t inner join dflt.LastEnc l on t.ssnID=l.ssnID

Go --(694926 row(s) affected)

Print 'Get inpatient diagnoses for training cases'

DROP TABLE #Dx0

SELECT ssnID, i.SCRSSN, i.sta3n

, admitdatetime, lastEnc, deathdate

, icd9sid

, iif(deathdate is null

, iif(datediff("dd", admitdatetime, LastEnc)<182, 1,0)

, iif(datediff("dd", admitdatetime, deathdate)<182,1,0)) As Dead182

, datediff("dd", birthdate, [AdmitDateTime]) AS AdmAgeInDays

, datediff("yy", birthdate, [AdmitDateTime]) As AdmAgeInYears

INTO #Dx0

FROM #tcases t left join [Src[CohortCrosswalk] c on t.ssnID = c.scrssn

left join [Src[Inpat\_InpatientDiagnosis] i on c.SCRSSN=i.SCRSSN And c.Sta3n=i.sta3n

WHERE [AdmitDateTime] is not null Go -- (35889752 row(s) affected) Diagnoses

Print 'Identify patients with encounters after death'

DROP TABLE #BadDate

SELECT ssnID, admitdatetime, lastenc

INTO #BadDate

FROM #Dx0

WHERE datediff(dd, admitdatetime, lastenc)<-1

Go -- 0 (0 row(s) affected)

Print 'Get ICD9 codes and descriptions, different stations have same codes'

DROP TABLE #ICDCode

SELECT icd9sid, max(icd9code) as ICD9Code, Max([ICD9Description]) as Long, Max([DiagnosisText]) as Short

INTO #ICDCode

FROM [CDWWork[Dim[ICD9]

GROUP BY ICD9SID

Go -- (2,025,871 row(s) affected)

Print 'Rank order repeated diagnosis, select good date of admissions'

DROP TABLE dflt.tDx

SELECT DISTINCT

d.ssnID as id1, d.admitdatetime

, icd9code, left(icd9code, 6) as ICD6, left(icd9code, 5) as icd5, left(icd9code,4) as icd4, left(icd9code,3) as icd3

, Dead182, Long, Short

, rank() Over (Partition by d.ssnid, icd9code Order by d.admitdatetime) as Repeated

INTO dflt.tDx

FROM #dx0 d left join #Baddate b on d.ssnID=b.ssnID

left join #icdcode i on d.icd9sid=i.ICD9SID

WHERE b.ssnID is null and icd9code not like '%unkn%'

Go -- (30800108 row(s) affected)

>\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

--> 3. Caclulate Likelihood Ratios for ICD-9 codes

>\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Print 'Calculate number of occurences for diagnoses with 6 digits'

DROP TABLE #dx1

SELECT icd6

, Repeated

, count(distinct id1) as n6Dx

, sum(dead182) as n6DeadAndDx

, sum(1-dead182) as n6AliveAndDx

, 'HospDx6' as IndType

, Min(Long) as Long

, Min(Short) as Short

INTO #dx1

FROM dflt.tdx

WHERE len(icd6)=6

GROUP BY icd6, Repeated

HAVING count(distinct id1)>29

Go -- (10928 row(s) affected)

Print 'Calculate number of occurences for diagnoses with 5 digits'

DROP TABLE #dx2

SELECT icd5

, Repeated

, count(distinct id1) as n5Dx

, sum(dead182) as n5DeadAndDx

, sum(1-dead182) as n5AliveAndDx

, 'HospDx5' as IndType

, iif(Min(Long)=max(Long),Min(Long),'') as Long

, iif(Min(Short)=max(short),Min(Short),'') as Short

INTO #dx2

FROM dflt.tdx

WHERE len(icd5)=5

GROUP BY icd5, Repeated

HAVING count(distinct id1)>29

Go --

Print 'Calculate number of occurences for diagnoses using 4 digits'

DROP TABLE #dx3

SELECT icd4

, Repeated

, count(distinct id1) as n4Dx

, sum(dead182) as n4DeadAndDx

, sum(1-dead182) as n4AliveAndDx

, 'HospDx4' as IndType

, iif(Min(Long)=max(Long),Min(Long),'') as Long

, iif(Min(Short)=max(short),Min(Short),'') as Short

INTO #dx3

FROM dflt.tdx

WHERE len(icd4)=4

GROUP BY icd4, Repeated

HAVING count(distinct id1)>29

Go -- (7211 row(s) affected)

Print 'Add 6, 5 and 4 digit diagnoses'

DROP TABLE #dx4

SELECT icd6 as Indicator, n6Dx as nDx, Repeated

, n6DeadAndDx as nDeadAndDx, n6AliveAndDx as nAliveAndDx

, IndType as [Type], Long, Short into #dx4 from #dx1

UNION all

SELECT icd5, n5Dx, Repeated

, n5DeadAndDx, n5AliveAndDx, IndType, Long, Short from #dx2

UNION all

SELECT icd4, n4Dx, Repeated, n4DeadAndDx, n4AliveAndDx

, IndType, Long, Short

FROM #dx3

Go -- (32118 row(s) affected)

Print 'Calculate Likelihood Ratio for ICD Code'

DECLARE @nDead int, @nAlive Int

SELECT @nDead=sum(nDeadAndDx), @nAlive=sum(nAliveAndDx) from #dx4

DROP TABLE Dflt.LR

SELECT [Type], Indicator, Repeated, concat([type]

,' ', Indicator, ' ', Repeated) as Code

, @nDead as 'Admissions Dead'

, @nAlive as 'Admissions Alive'

, nDeadAndDx '# Dead w', nAliveAndDx '# Alive w'

, nDx as 'Cases w'

, iif(nDeadAndDx=0, 1.0/cast((nDx+1) as float)

, iif(nAliveAndDx=0, nDx+1, (cast(nDeadandDx as float)/cast(@nDead as float))/(cast(nAliveandDx as float)/cast(@nAlive as float)))) as LR

, Long, Short

INTO Dflt.LR

FROM #dx4

Go -- (32118 row(s) affected)

**[END LIST]**

**[H1] Reference**

Alemi, F., and S. R. Walters. 2006. “A Mathematical Theory for Identifying and Measuring Severity of Episodes of Care.” *Quality Management in Healthcare* 15 (2): 72–82.

**Appendix B**

This appendix describes how a logarithm function can be used to calculate the product of values in a different row of a single column.

|  |  |
| --- | --- |
| *Graphic* | *Text* |
| https://lh5.googleusercontent.com/kYiAmfjEEJqkJjuZMI2Kia4zhey8vfioxO8rNKy-Wt0m_XcXGQ1v-pdiQY39r2WN7F6Qgx1S8bvWC57uxoyQcNpzde3_Dd5IaI4e0nFed2D8E09FEW7U37SpY3ONAzd58nFx1Uei | In SQL, there are commands for summing values in a column but no commands for finding the product of the values in the same column. However, users can employ use Sum for this task. I demonstrate this by calculating the product of 2 and 4, when 2 is in one row and 4 is in another row of the same column. The first step is to calculate the log of these values—the log of 2 is 0.69 and the log of 4 is 1.39. In the graph we start at the *x*-axis and move to the *y*-axis to get the log. |
| https://lh6.googleusercontent.com/gcKKKw797-Tuy4vEV_qCw39SMv4eNToIURnzADQ7ihJq8i2TFcGm_axsCDBVx0oZjTqx2Z7TuTi9ezz1LKeMVPdjouC6XdetLUAccCs7Smjse61QtdYhsXqBvi8GsctX5RKg_5OS | Next we sum the log of 2 and log of 4 and obtain 2.08. This step is occurring in the margins of the *y*-axis, which may be difficult to see. Better to see this in a cutout. |
| https://lh6.googleusercontent.com/Zdf3HBrjHvsbWE-hPg7y9PID1i67Z-beOUHeSqyGvbRKY7bqN6DKpfIOvM2V-WY8pcnw23PkHEPoaxJZMUEh5z6t0N34HL8bJ-OWP9E1NejgiQqHxYZgWEPYWNkzJgtPdomU7KZX |  |
| https://lh3.googleusercontent.com/0Uj-dceiHRe7HTFuzCyv4uCqDIHQaJwqxqoSXvmORIlV0Axr7SKU9khqK6xaKrF3QzEUGDU9sRG32aAJpEl6yyHfS4DHAoq7GZMCfLBGFoAmduGCxyDRK-cigHLonO52E2FVpZVa | The final step is to take an antilog, which is working backward from the *y*-axis to the *x*‑axis. We see now that 2.08 is the log of 8. Therefore, we conclude that 2 times 4 is 8. Note that we got to 8 by summing the two logs and without use of SQL’s multiplication function. |

**[type: please place endnote before appendices]**

1. This work was supported by appropriation #3620160 from the VA Office of Geriatrics and Extended Care to Dr. Cari Levy at Denver Veterans Affairs in addition to resources from the District of Columbia Veterans Affairs Medical Center. The chapter is based on F. Alemi, C. R. Levy, and R. E. Kheirbek. 2016. “The Multimorbidity Index: A Tool for Assessing the Prognosis of Patients from Their History of Illness,” *eGEMS* 4 (1): 1235. [↑](#endnote-ref-1)
2. [↑](#endnote-ref-2)
3. [↑](#endnote-ref-3)
4. [↑](#endnote-ref-4)
5. [↑](#endnote-ref-5)
6. [↑](#endnote-ref-6)
7. [↑](#endnote-ref-7)
8. [↑](#endnote-ref-8)
9. [↑](#endnote-ref-9)
10. [↑](#endnote-ref-10)
11. [↑](#endnote-ref-11)
12. [↑](#endnote-ref-12)
13. [↑](#endnote-ref-13)
14. [↑](#endnote-ref-14)
15. [↑](#endnote-ref-15)
16. [↑](#endnote-ref-16)
17. [↑](#endnote-ref-17)
18. [↑](#endnote-ref-18)