EUROPEAN COMMISSION

Directorate-General for Health & Consumers

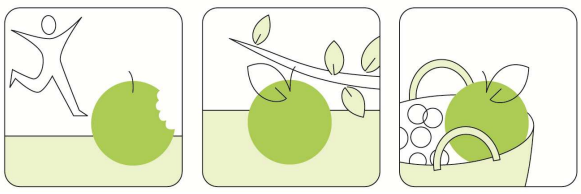
EU Public Health Outcome Research and Indicators Collection
EUPHORIC Project
Grant Agreement n° 2003134

A project funded by the European Commission,
Directorate General for "Health and Consumers"

Deliverable N. 12.7

Statistical procedures for comparative evaluation of outcomes

January 2009 First release



MAIN BENEFICIARY

Istituto Superiore di Sanità, *Italy*

ASSOCIATED BENEFICIARIES

EFORT/EAR Verein zur Unterstützung der Tätigkeit von nationalen Endoprothesenregistern, *Austria*

Sosiaali- ja terveysalan tutkimus- ja kehittämiskeskus, *Finland*

National and Kapodistrian University of Athens, *Greece*

ASL RM E Department of Epidemiology, *Italy*

Institut Municipal d'Assistència Sanitària, *Spain*

Karolinska Institutet, *Sweden*

COLLABORATING PARTNERS

National Center of Public Health Protection, *Bulgaria*

Catalan Agency for Health Technology Assessment and Research, *Spain*

Slovak Arthroplasty Register, *Slovak Republic*

Arthroplasty Register Tyrol, *Austria*

Ludwig Boltzmann Institut Health Technology Assessment, *Austria*

French Society of Orthopaedic and Trauma Surgery, *France*

BQS Bundesgeschäftsstelle Qualitätssicherung gGmbH, *Germany*

Israel Society for the Prevention of Heart Attacks at NCRI, *Israel*

This report was prepared by:

Danilo Fusco, Anna Patrizia Barone, Mariangela D'Ovidio (partner DEASL[1])

[1] Department of Epidemiology ASL Roma E, Italy

Acknowledgment: The authors would like to thank Marina Torre (partner ISS), Rino Bellocco (partner KAR) for their comments.

Introduction

Risk adjustment has shown to be an essential tool to enable comparisons between groups, services, facilities, providers or treatments that are not directly comparable because of differences in confounding factors. For example, comparisons between outcomes from different hospitals are often confounded by the patients' severity, and it is recommended that outcomes should be compared only when differences in "*a priori*" risk are adjusted for.

The objective of risk adjustment is to identify a model which can accurately predict the outcome while controlling for an array of patient risk factors. Applying the risk adjustment models to a hospital's data helps 'level the playing field', so that a hospital can compare its indicator rates to other hospitals more fairly.

The risk factors considered for assessing health care quality include patient clinical and demographic variables such as gender, age, co-morbidities, scores, and a wide range of other preexisting conditions and associated diagnoses. In this report we do not consider hospital level information for risk adjustment because of the possibility that these factors may be the very factors explaining variance in performance between organizations (Multilevel model).

The following describes the steps to develop risk adjustment models for assessing health care quality.

Risk Factor identification

After the selection of factors potentially associated with the outcome under study based on review of the available evidence and a clinical audit (without relevant information on these factors, it is advisable to consider surrogate variables (proxies)), each potential risk factor is recoded to appropriate values for the computerized modelling process. There are four basic types of potential risk factors:

Binary variable

Binary variables include gender and the risk factors associated with the ICD-9-CM/ICD-10 diagnoses.

Binary variables are coded to a value of '1' or '0'.

The missing values for binary variables should be recoded into a separate category in order to be included in predicting the outcome.

Ordinal variable

Ordinal variables are coded as either continuous or design (dummy) variables based on their actual effect on the outcome. The scatter plot or contingency table between the outcome and the ordinal variable can help to determine which coding is appropriate. The missing values for ordinal

variables should be recoded into a separate category in order to be included in predicting the outcome.

Continuous variable

Continuous variables, such as patient age, are handled based on the actual effect on the outcome. Initially, the relationship between age and the indicator rate is examined via scatter plot and bivariate regression analysis. The data are aggregated by age or age group to produce the scatter plot.

The scatter plot clearly shows the relationship between two variables. If a linear relationship exists, age is operationalized in the model as a continuous variable, generally, truncation is necessary to improve model performance. In the event of a nonlinear relationship, age may need an algebraic transformation or be recoded as a design variable based on its actual effect on the indicator rate.

Categorical variable

Categorical variables such as race and hospital type are recoded as design (dummy) variables for inclusion in the risk adjustment models. The categories should represent distinct types. Their effects on the indicator rate can be evaluated by analysis of a contingency table or examination of the univariate logistic model. In some cases, the categorical risk factor may need to be redefined several times during analysis until the categorical risk factor fits the selected regression model well. The missing values should be recoded into a missing category in order to be included in predicting the outcome.

Risk Factor Screening

There is considerable redundancy of information since several diagnostic findings report a common underlying physiologic process. Therefore, the process of model development includes further screening to identify risk factors that are able to predict the event of interest with statistical significance.

All risk factors are checked for collinearity. Two simple methods are used. One method is to use the Variance inflation factor (VIF). This statistic measures the impact of collinearity among the X's in a regression model on the precision of estimation and expresses the degree to which collinearity among the predictors degrades the precision of an estimate. Typically a VIF value greater than 10 is of concern. A second method is to examine results of the pairwise correlation matrix. Neither method is very sensitive for detecting collinearity but can be used for reference. Collinearity among risk factors is also checked further during the modelling process.

Model Building

The model building process is accomplished using the following steps:

1. a full risk adjustment model is developed with all screened risk factors. The estimate for each risk factor is reviewed carefully to compare its crude effect in the univariate logistic regression model to its effect in the full risk adjustment model. If the coefficient for the risk factor changes dramatically, in particular, if it changes direction, then collinearity exists and it is necessary to investigate the problematic risk factors in detail.

2. important biological factors or categorical variables (a priori risk factors) are retained in the model independently of their association with the outcome, according to the information available in the literature, and therefore without any statistical verification.

3. statistically nonsignificant risk factors for the indicator outcome are dropped by the stepwise selection method. Stepwise selection with the 'SLS=0.1' option will retain only those risk factors whose regression coefficient estimates are significant at level 0.1 in the model. The less significant risk factors ($p > 0.1$) are dropped from the model. In many statistical programs there are options to retain in the model the "a priori" risk factors while the stepwise selection is processing (ie. in SAS program the option INCLUDE).

Stepwise selection deals directly with redundancy and identifies the risk factors that strongly and independently affect the outcome. If collinearity between risk factors still exists in the model, interpretation of the effects of such factors on the outcome is complex. In such cases, it is determined whether to avoid simultaneous inclusion in the model.

4. the model resulting from the stepwise selection is reviewed for both clinical and statistical validity. Very large estimated coefficients or standard errors, as well as unreasonable estimates should be investigated further. After all statistically significant risk factors are reviewed and identified, the final model is run for assessing model fit.

Model Fit Assessment

One central problem with the current use of goodness-of-fit measures is that there are no formal standards for their selection and use. In some research areas, there are a number of conventions for the selection of particular methods. However, these conventions are typically more sociological and historical than logical in origin. Moreover, many of these conventions have fundamental shortcomings resulting in goodness-of-fit arguments that often range from uninformative to somewhat misleading to just plain wrong.

In general the goodness-of-fit of a model to data is evaluated in two different ways: 1) through the use of visual presentations methods which allow for visual comparison of similarities and differences between model predictions and observed data; and 2) through the use of numerical measures which provide summary measures of the overall accuracy of the predictions.

1) Visual presentations methods

Overlay Scatter Plots and Overlay Line Graphs

The best method for assessing the accuracy of point predictions is to use overlay scatter plots and overlay line graphs. In these graphical forms, the model and data are overlaid on the same graph. Because the relatively small size of point types used in scatterplots do not typically create strong Gestalt connections by themselves, scatterplots do not emphasize the qualitative trends in the data or model particularly well, whereas line graphs strongly emphasize trends. In overlay graphs, it is important for the model and data to be on the same scales. If the model's performance is measured in arbitrary units, then overlay graphs are inappropriate. Line graphs are not appropriate for categorical x-axis variables. Both overlay and line graphs are not optimal for displays of complex data patterns not fit particularly well by the model because the graphs become very difficult to read.

As an additional consideration, overlay graphs may become too cluttered if error bars are added to them. Also, for line graphs, the usual convention is to have data indicated with closed icons and solid lines and to have model performance indicated in open icons and dotted lines.

2) Numerical measures

Mean Squared Deviation (MSD) or Root Mean Squared Deviation (RMSD)

The most popular measures of goodness-of-fit to exact location are the Mean Squared Deviation and its square root (Root Mean Squared Deviation). That is, one computes the mean of the squared deviation between each model prediction and the corresponding data point:

$$MSD = \frac{\sum_{i=1}^{k} (m_i - d_i)^2}{k} \qquad \text{and} \quad RMSD = \sqrt{MSD} = \sqrt{\frac{\sum_{i=1}^{k} (m_i - d_i)^2}{k}}$$

where, m_i is the model mean for each point i, d_i is the data mean for each point i, and k is the number of points i being compared. The consequence of squaring the deviations is that more emphasis is placed on points that do not fit well than on points that do fit well (i.e., a prediction that is two units off the data produces a penalty four times as large as a point that is one unit off the data). Because of the noise found in almost all behavioral data, this result is desirable because it reduces the tendency to overfit the data. The applicability of RMSD to a broad range of situations and familiarity to the general research community makes it one of the measures of choice for measuring deviation from exact location.

Mean Absolute Deviation (MAD)

A conceptually easier to understand measure of goodness-of-fit to exact location is Mean Absolute Deviation. The MAD is the mean of the absolute value of the deviation between each model prediction and its corresponding data point:

$$MAD = \frac{\sum_{i=1}^{k} |m_i - d_i|}{k}$$

where, m_i is the model mean for each point i, d_i is the data mean for each point i, and k is the number points i being compared. One advantage of this measure is that is provides a value that is very easy to understand, much like r^2. For example, a model fit with a MAD of 1.5 seconds means that the model's predictions were off from the data on average by 1.5 seconds. Unlike MSE and RMS, MAD places equal weighting on all deviations. When the data is not relatively noise-free (as is the case in most behavioral data), then this is a disadvantage, both because of overfitting issues and because one is being penalized for deviations that are often not real.

Note that because MAD involves the absolute value of the deviation, this measure does not differentiate between noisy fits and systematically biased fits (i.e., off above and below the data versus only below the data). Measures of systematic bias are presented in a later section.

Mean Scaled Absolute Deviation (MSAD) and Root Mean Squared Scaled Deviation (RMSSD)

They are very similar to MAD and RMSD except that each deviation is scaled by the standard error of the mean of the data. For example, for MSAD, the absolute value of each model-to-data deviation is divided by the standard error for each data mean (i.e., the standard deviation of each mean divided by the square root of the number of data values contributing to each mean). This type of scaling is similar to the scaling done with standardized residuals in statistical regression and structural equation modeling. There, deviations between the statistical model and the data (i.e., the residuals) are divided by the standard error of the data. The rationale and base computation for MSAD and RMSSD is the same, but MSAD takes the mean absolute value of these standardized residuals and RMSSD takes the square root of the mean squared value of these standardized residuals. MSAD is defined as follows:

$$MSAD = \sum_{i=1}^{k} \frac{|m_i - d_i|}{k\frac{s_i}{\sqrt{n_i}}} = \sum_{i=1}^{k} \frac{|m_i - d_i|\sqrt{n_i}}{k s_i}$$

where m_i is the model mean for each point i, d_i is the data mean for each point i, s_i is the standard deviation for each data mean i, n_i is the number of data values contributing to each data mean d_i, and k is the number points i. By contrast, RMSSD is defined as:

$$RMSSD = \sqrt{\sum_{i=1}^{k} \frac{\left(\frac{m_i - d_i}{s_i/\sqrt{n_i}}\right)^2}{k}} = \sqrt{\sum_{i=1}^{k} \frac{(m_i - d_i)^2 n_i}{k s_i^2}}$$

where m_i is the model mean for each point i, d_i is the data mean for each point i, s_i is the standard deviation for each data mean i, n_i is the number of data values contributing to each data mean d_i, and k is the number points i.

The standard error of the mean ($s_i/\sqrt{n_i}$) combines variability information with amount of data contributing to each point and is used in common statistical tests (t-tests, confidence intervals, etc). This scaling has three advantages. First, it shows how close the models are getting to the true resolution of the data. The reader need not always be given both SE_c and RMSD to evaluate the quality of a fit—these measures combine both pieces of information. Second, this scaling more heavily penalizes misfits to data points that are precisely known than to data points whose locations are imprecisely known—the weighting function is directly the degree of imprecision. Third, like $\underline{r}^2$, MSAD and RMSSD are scale invariant; that is, regardless of whether the data is measured in terms of errors, reaction time, or percent of a certain choice, a MSAD value of 1.5 has the same meaning: on average, the model is 1.5 standard errors off from the data. This scale invariance makes overall evaluation of the quality of the fit to the data quite easy. Of course, in order to compute these values, one needs to have access to variance information about the data, which may not be available for fits to older data sets.

One potential problem with MSAD and RMSSD is that a lower $\underline{n}$ will produce a lower MSAD and RMSSD, thus apparently rewarding fits to low $\underline{n}$ studies. However, including 95%CI bars in the graphical presentations of the data will make clear the low fidelity of the data, thereby reducing the persuasiveness of the model fits. Moreover, low $\underline{n}$ studies also produce highly noisy data, which is usually much harder to fit, especially in fits to a series of points on a single dimension or in fits to multiple data sets. Thus, MSAD and RMSSD do not clearly reward low $\underline{n}$ studies in the same way that badness-of-fit measures do.

Another potential problem with MSAD and RMSSD is the case of data points with zero variance (e.g., zero errors in a condition when fitting error data) because the denominator in the scaling becomes zero and the deviation is then difficult to include in an average. The solution is to use pooled standard errors (from the data set as a whole) for the zero variance data points, under the assumption that the true population variance for those cells is not actually zero.

As with MAD and RMSD, MSAD and RMSSD have the same relative advantages and disadvantages. That is, MSAD does not overweight large deviations but can lead to overfitting problems, whereas RMSSD does overweight large deviations but reduces overfitting problems. Thus we generally recommend RMSSD for most behavioral data. In choosing between MSAD and RMSSD, we suggest the following rule-of-thumb: when MSAD drops below 2, then RMSSD values should be used to avoid overfitting. If the MSAD or RMSSD drop below 1, then the model has reached the fidelity of data.

Bayesian information criterion (BIC)

The BIC is sometimes also named the **Schwarz Criterion**, or **Schwarz Information Criterion** (**SIC**). It is so named because Gideon E. Schwarz (1978) gave a Bayesian argument for adopting it. The BIC is an asymptotic result derived under the assumptions that the data distribution is in the exponential family. Let:

- x = the observed data;
- n = the number of data points in x, the number of observations, or equivalently, the sample size;
- k = the number of free parameters to be estimated. If the estimated model is a linear regression, k is the number of regressors, including the constant;
- $p(x|k)$ = the likelihood of the observed data given the number of parameters;
- L = the maximized value of the likelihood function for the estimated model.

The formula for the BIC is:

$$-2 \cdot \ln p(x/k) \approx BIC = -2 \cdot \ln L + k \cdot \ln p(n)$$

Under the assumption that the model errors or disturbances are normally distributed, this becomes (up to an additive constant, which depends only on n and not on the model):

$$BIC = n \cdot \ln (RSS/n) + k \cdot \ln p(n)$$

where RSS is the residual sum of squares from the estimated model. Note that the term for used in this specialization is equal to the rescaled normal loglikelihood up to an additive constant that depends only on n.

Given any two estimated models, the model with the lower value of BIC is the one to be preferred. The BIC is an increasing function of RSS and an increasing function of k. That is, unexplained variation in the dependent variable and the number of explanatory variables increase the value of BIC. Hence, lower BIC implies either fewer explanatory variables, better fit, or both. The BIC penalizes free parameters more strongly than does the Akaike information criterion.

It is important to keep in mind that the BIC can be used to compare estimated models only when the numerical values of the dependent variable are identical for all estimates being compared. The models being compared need not be nested.

Model Application

Once the final risk adjustment model is determined, the comparison of the outcomes between providers should be estimated by indirect and direct standardization procedures. In this report we will only describe the processes of comparison using risk adjustment models for binary outcome.

1) Indirect standardization

The indirect method first estimates the probability of the outcome (event) for each patient, applying the final risk adjustment model. By summing the probabilities of the outcome for all patients at each hospital, the number of expected events is obtained.

Standardized ratios (SRs) are calculated dividing the observed by the expected number of events at each hospital and multiplying this ratio per 100.

SR is an estimate of what a hospital's observed outcome rate would have been if its patients had the same mean outcome rate and the same risk factor associations as the reference population.

The relationship between the Poisson and χ^2 distributions is employed to calculate the lower and upper limits of confidence intervals for the SRs:

$$LL = \frac{\chi^2_{\frac{\alpha}{2},2\cdot observed}}{2\cdot expected} \qquad UL = \frac{\chi^2_{1-\frac{\alpha}{2},2\cdot(observed+1)}}{2\cdot expected}$$

2) Direct standardization

This method compares the observed outcomes in each group with those of the reference population (with a constant and well-defined distribution of the severity measure). It applies the relationship between risk factors and outcomes in the different groups under study to the reference population.

From an operational viewpoint, the calculation and the comparison of the expected outcomes between groups are made simultaneously.

In fact, direct standardization applies multivariate statistical models (chosen, as always, according to the type of outcome considered) in which, in addition to the risk factors selected for the severity measure, "*dummy*" variables are included. These variables represent the groups compared (a dummy variable assumes a value of 1 for the subjects belonging to the group considered and value 0 for all other subjects).

In this case, however, contrary to indirect standardization, it is possible to derive the adjusted measures of association for any two of the n groups considered from the coefficients of the explicative model.

Reference group

Benchmark

The benchmark could be defined as the group of the best performing hospitals. It should be defined by the following steps:

1. X hospital dummies are added to the model and the corresponding adjusted ORs are estimated. At this step the hospital with the highest number of patients is chosen as reference category.
2. After ranking all hospitals by adjusted ORs, the Y hospitals with the lowest adjusted ORs are selected as the reference group. This group is selected by an iterative procedure that at each step includes one hospital in the reference group. The procedure stops when the hospital to include in the reference group is significantly different (p=0.10) from the benchmark defined in the previous step.

Population under study

The population under study should be used as a reference by applying the ”deviation contrast” method, in which estimated coefficients from the logistic regression model have a sum of zero. One parameter is dropped as redundant during estimation (parameter for the first hospital) and found afterwards using minus the sum of the estimated parameters, or by re-estimating the model using a different omitted category.

References

- Arcà M, Fusco D, Barone AP, Perucci CA. Risk adjustment and outcome research. Part I. J Cardiovasc Med (Hagerstown). 2006;7(9):682-90
- Busemeyer JR, Wang YM. Model comparisons and model selections based on generalization criterion methodology. Journal of Mathematical Psychology 2000; 44(1): 171-189.
- Estes WK. On the communication of information by displays of standard errors and confidence intervals. Psychonomic Bulletin & Review 1997; 4(3): 330-341.
- Forster MR. Key concepts in model selection: Performance and generalizability. Journal of Mathematical Psychology 2000; 44(1): 205-231.
- Myung, J. The importance of complexity in model selection. Journal of Mathematical Psychology 2000; 44(1): 190-204.
- Roberts S, Pashler H. How persuasive is a good fit? A comment on theory testing. Psychological Review 2000; 107(2): 358-367.
- Schwarz G. Estimating the dimension of a model. The Annals of Statistics 1978; 6: 461-464.
- Simon HA. What is an "explanation" of behavior? Psychological Science 1992; 3: 150-161.
- Trafton JG, Trickett SB, Mintz FE. Overlaying images: Spatial transformations of complex visualizations. In Model-Based Rasoning: Scientific Discovery, Technological Innovation, Values. Pavia, Italy, 2001.
- Wasserman L. Bayesian model selection and model averaging. Journal of Mathematical Psychology 2000; 44: 92-107.