EUROPEAN COMMISSION

Health & Consumer Protection
Directorate-General

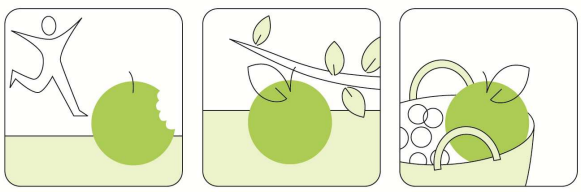
EU Public Health Outcome Research and Indicators Collection
EUPHORIC Project
Grant Agreement n°2003134

A project funded by the European Commission,
Health and Consumer Protection Directorate General

Deliverable N. 10

Risk adjustment methodologies

February 2008



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*

Genetics Research Institute ONLUS, *Italy*

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[1], Anna Patrizia Barone[1], Paola Colais[1], Gloria Tiberi[1]

[1]ASL RM E Department of Epidemiology, Italy

Acknowledgements: The authors would also like to thank Gabriella Badoni, Lucilla Di Pasquale, Mascia Masciocchi and Grazia Rago who were responsible for editing the report.

Preamble

The purpose of this review is to provide a detailed description of different risk adjustment methodologies to compare health care outcomes.

The review describes:

1) the methods for constructing the severity measures;
2) the methods that use the severity measures to obtain "adjusted" outcome measures for valid comparison between groups (stratified analysis, indirect and direct standardization);
3) identification and management of effect modification;
4) the methods to gain the precision of the estimates;
5) the methods used with multiple comparisons;
6) other models (multi-level models) used for risk adjustment.

Introduction

Observational studies that compare groups, services, facilities, providers and treatments must consider the possible differences already existing in the populations under study, above all, the differences in patient characteristics that can determine care outcomes. Therefore, the goal of an investigator is to take into account the possible confounding effect of the different distribution (among groups, services, facilities, providers or treatments) of "a priori" characteristics [1-11], which can influence the occurrence of the outcome.

Empirical evaluation of the confounding effect entails comparing a crude measure of association with the "*adjusted*" one for chosen confounding variables. This is a "*qualitative*" comparison and cannot be based on the result of statistical tests. In other words, the "*acceptable*" amount of confounding depends on the subjective evaluation of the investigator in relation to the features of the study object and the hypothesis under study.

Risk adjustment [11-15] is one of the ways confounding is identified and controlled in observational studies. It 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. The basic idea is to compare exposed to unexposed groups that are homogenous regarding their "*a priori*" risk of occurrence of the outcome under study.

In this approach, also, potential confounders can be considered individually or synthesized in a summarizing score and used for stratified/matched analyses, or for multivariate adjustment.

Two steps in risk adjustment procedures can be distinguished:

- construction of the measure used to define the "*a priori*" risk (patients' severity);
- construction or selection of the methods that use the severity measures to obtain "adjusted" outcome measures for valid comparison [12].

1) Severity measure

The severity measure to use in risk-adjustment procedures should be a good outcome predictor in the population studied, should not be influenced by outcome (e.g., a diagnosis of cardiac arrest is associated with death, but it can be hardly interpreted as an "*a priori*" risk factor), or should not be a link of the causal chain between exposure and outcome (complication).

Severity measures, widely used for risk adjustment, can be classified into three groups: prognostic scores, severity measures based on "*pre-defined*" predictive models or an "*empirical*" approach.

Prognostic Scores

Usually additive, they summarize individual risk of adverse outcomes into scores or groups. They can be based on "*clinical*" data (e.g., medical charts) or "*administrative*" data (e.g., discharge abstracts). The score attribution criteria may derive from clinical evaluations and/or

from multivariate analysis conducted with the empirical approach. The score cannot be directly interpreted as individual odds of the outcome considered. Some prognostic scores based on hospital data are shown in Table 1.

Severity measures based on "*pre-defined* " predictive models

These measures are similar to those described in the previous section, but they are represented by mathematical functions that, when applied to every patient, allow the individual probability of the outcome to be estimated directly. Variable selection and the estimates of their coefficients derive from statistical models applied to an external population, generally very large. Applying such coefficients to the population under study, we obtain an estimate of the outcome expected at the individual level. In this case "*clinical*" or "*administrative*" data may also be used. Some examples of "*pre-defined*" models [11, 16, 17] are shown in Table 2.

The use of these models or prognostic scores assumes a constant effect of predictors on the outcome between the populations.

Severity measures based on an "*empirical*" approach

This approach is based on the necessity to identify confounding factors and to control their effect according to their specific relationship with the outcome in the population under study. Therefore the empirical approach uses a severity measure specific for the time and population under study, calculated analyzing the multivariate relationship between possible predictors and the outcome considered.

The most frequently used instrument for the construction of an empirical severity measure is the multivariate regression analysis.

The procedure can be separated into the following steps [18-22]:

1. *Selection of the outcome under study.*
2. *Choice of the most appropriate statistical model for data analysis depending on the outcome under study*

   We can consider:
   - linear regression models, used when the outcome is a continuous variable; logistic regression models, applied when the outcome variable is dichotomous (presence vs absence);
   - Poisson's regression models, used when the outcome variable is a number of units or events;
   - survival models, used when the outcome variable is survival time.
3. *"A priori" identification of individual characteristics as possible risk factors for the outcome*:

   Selection of factors potentially associated with the outcome under study is usually based on review of the available evidence about the association of interest. Without relevant information on these factors, it is advisable to consider surrogate variables (proxies).
4. *Descriptive analysis and first screening of risk factors:*

   Occurrence of risk factors selected in the population under study has to be described and those factors present in a small fraction of subjects have to be eliminated (e.g. smaller than 1%).
5. *Selection of "a priori" risk factors:*

   Some factors are included in the model independently of their association with the outcome, according to the information available in the literature, and therefore without any statistical verification.
6. *Second screening of risk factors:*

   The statistical model that predicts the outcome includes:
   - the factors considered in step 5;
   - the remaining factors that "*survived*" the first screening;
   - interactions between factors considered potentially relevant

and, a subsequent selection, in the above two groups, based on the significance of the multivariate associations with outcome. Selection is made through automatic stepwise procedures [23,24].

Since the crude association between each factor and outcome could be biased by confounding effects of other factors, adjusted estimates aim to obtain the best possible unbiased estimates of the "*real*" association between each factor and outcome.

It must be carefully noted that only the interactions defined "*a priori*" as relevant for the specific study could be taken into account. Implicit limits of validity and power prevent any "*screening*" of all candidate interactions.

7. *Estimate of the coefficients of the model*

   A string of coefficients (b_0, b_1,…,b_k), representing the best estimate of the real coefficients (β_0, β_1,…, β_k), are calculated, based on available observations, by iterative numerical methods available in the most statistical analysis programs.

8. *Evaluation of the performance of the model chosen:*

   The predictive capability of the severity measure constructed can be quantified to make external comparisons by using different statistics [24-29]:

• *Coefficient of determination* R^2

The ability of the model to explain data under study is calculated by R^2 statistic. This statistic is the proportion of variance explained by the model over total variance and ranges from 0 to 1. The higher the values, the better the degree of adaptation.

• *Adjusted* R^2

This statistic is interpreted in the same way as the above, but the greater the number of factors the more penalized adjusted R^2 is.

• *Pearson chi-square*

This statistic is calculated by grouping the population under study by patterns of existing factors (values assumed by each factor for each subject) and calculating the number of expected and observed events for each pattern. The latter are compared using a chi-square test to evaluate if the differences between expected and observed events are not statistically significant. If continuous variables are present, it is preferable to use Hosmer-Lemeshow test.

• *Hosmer-Lemeshow test*

This test estimates the ability to predict a number of expected events corresponding to those observed (calibration). This statistic is calculated by grouping the population under study into percentiles (usually deciles) and calculating the number of expected and observed events for every percentile. The latter are compared using a chi-square test to evaluate if the differences between the expected and the observed events are not statistically significant.

• *C-index* (area under the ROC curve)

The ability to distinguish subjects having the outcome from others who do not (capacity to discriminate) is estimated by the C-index, also called *c* statistic or area under the Receiver Operating Characteristic (ROC) curve. This index varies between 0 and 1 and higher values indicate a better ability to discriminate.

• *AIC (Akaike Information Criterion)*

AIC measures the degree of adaptation of the model to data under study, considering the number of factors included in the model. The greater the number of factors the more penalized AIC is. Low values of the AIC indicate good adaptation of the model.

These statistics are only some of the methods used to select the best predictive model.

Table 3 summarizes the regression models for which the previous statistics can be used.

Finally, we must emphasize that it is possible to construct empirical risk measures that include a prognostic score instead of, or in addition to individual factors (e.g, Charlson index or APR-DRG risk class). This choice has some advantages from a practical viewpoint, but has the defect of using neither the informative content of the variables detected nor the empirical evidence about the relationship in the best way.

Table 1. Prognostic scores

(attributing of scores or severity classes)

Measure	Data used	Type of measure	Construction criterion	Outcome measured
	Clinical data			
APACHE III[30-33]	17 physiologic parameters and other clinical information	Integer scores from 0 to 299 measured within 24 hours of ICU admission	Empirical modelling with clinical guidance	In-hospital death for patients in Intensive Care Unit
Canadian[34,35]	Condition-specific clinical variables entered at time of referral for cardiac surgery	Range of scores from 0 to 16 based on odds ratio for 6 key risk factors	Logistic regression model	In-hospital mortality, ICU stay, and postoperative length of stay
EuroSCORE[36]	Condition-specific clinical variables	Range of scores from 0 to 39 based on 17 weighted risk factors	Logistic regression model	Operative mortality
	Administrative data			
AIM[37]	Discharge abstract	Scores from 1 to 5 within DRGs	Empirical modelling	Length of hospital stay within DRGs
APR-DRGs[33,,38,39]	Discharge abstract. DRG-specific variables	Four classes of illness severity Four risk of mortality classes	Empirical modelling with clinical guidance	Use of resources In-hospital mortality
Body Systems Count[40]	Discharge abstract	Number of organ systems involved with disease	Clinical judgment	Number of organ systems involved with disease
Charlson Severity Score[41,42]	Discharge abstract	Integer from additive scale representing number and severity of comorbidities	Clinical judgment with empirical guidance	Risk of death within one year of medical hospitalization
Cleveland[43]	Discharge abstract. Condition-specific variables	Range of scores from 0 to 33 based on odds ratio for each of 13 risk factors	Empirical modelling (univariate analysis)	In-hospital death or death within 30 days of operation

Table 1. Prognostic scores
(continued from previous page)

Measure	Data used	Type of measure	Construction criterion	Outcome measured
Disease Staging[33,44,45]	Discharge abstract. Condition-specific variables	Three stages with substages within each stage Number of comorbidities within each of three major stages	Clinical judgment	Risk of death or functional impairment Number of comorbidities within each of three major stages
New England[46,47]	Discharge abstract. Condition-specific variables and comorbidity index	Scoring system based on coefficients used to calculate probability of operative mortality	Logistic regression model	In-hospital mortality
Parsonnet[48-50]	Discharge abstract. Condition-specific variables	Scores between 0 and 158 based on 14 weighted risk factors	Additive multiple regression model	Death within 30 days of operation
PMCs Severity Score[51]	Discharge abstract	Range of scores from 1 to 7	Empirical modelling	In-hospital morbidity and mortality

AIM =Acuity Index Method; Canadian = Ontario Ministry of Health Provincial Adult Cardiac Care Network; Cleveland = Cleveland Clinic Foundation Risk Stratification System; EuroSCORE = European System for Cardiac Operative Risk Evaluation; New England = Northern New England Cardiovascular Disease Study Group; Parsonnet = Parsonnet Risk Stratification Model.

Table 2. Severity measures based on "p*re-defined*" models

(producing a direct estimate of outcome probability)

Measure	Data used	Type of measure	Construction criterion	Outcome measured
	Clinical data			
MedisGroups[33,52]	Clinical variables collected at time of hospital admission	Probability of in-hospital death ranging from 0 to 1	Logistic regression model	In-hospital mortality
VA[53,54]	Condition-specific clinical variables	Risk interval (percent mortality interval) assigned to patient based on variables measured 30 days after operation	Logistic regression model	In-hospital death and morbidity
	Administrative data			
Disease Staging[21,44,45]	Discharge abstract. Condition-specific variables	Probability of in-hospital death ranging from 0 to 1	Empirical modeling	In-hospital death
NY[55,56]	Discharge abstract. Condition-specific variables	Probability of in-hospital death ranging from 0 to 1	Logistic regression model	In-hospital death
STS[57-59]	Discharge abstract. Condition-specific variables	Risk interval (percent mortality interval)	Bayesian algorithm; more recently converted to logistic regression model	In-hospital death and morbidity

NY = New York State Department of Health Cardiac Surgery Reporting System; STS = Society of Thoracic Surgeons Risk Stratification System; VA = Veteran's Administration Cardiac Surgery Risk Assessment Program.

Table 3. Statistics used in regression models

Adaptation measure	Regression model			
	Linear	**Logistic**	**Surviv. Analysis**	**Poisson**
R^2	**x**			
Adjusted R^2	**x**			
Pearson Chi-square		**x**		**x**
Hosmer-Lemeshow		**x**		**x**
ROC		**x**		
AIC	**x**	**x**	**x**	**x**

2) Use of the severity measure for risk adjustment

Once a severity measure has been constructed, the comparison of the outcomes between groups, providers, populations or treatments, can be carried out by three different methods [29, 60-61]:

- Stratified analysis;
- Indirect standardization;
- Direct standardization.

Stratified analysis

If the factors selected and used to define the severity measure are limited and if they are all represented by categorical variables, or if we choose to use a prognostic score (with, by definition, a limited number of modalities), patients can be subdivided into strata by their characteristics (age class, gender, score value, etc.). This subdivision allows each stratum to include patients who are homogeneous for severity. Therefore it is possible to measure the association of interest (between groups and outcome) in each stratum and then calculate the weighted mean of stratum-specific measures of the association (by the Mantel-Haenszel estimate, for example). This mean will be a "*risk adjusted*" estimate of the association considered.

The weighted mean of the stratum-specific measures of association has relevance only if such measures are "*reasonably*" homogenous. If, instead, these measures vary greatly through the strata, the stratum-specific measures should be presented to clearly show the stratum effect rather than being "*hidden*" in a single measure of association (even if "*adjusted*").

Indirect standardization

This method can be adopted when using:

- a severity measure based on a pre-defined model in which variables' selection and the estimates of their coefficients derive from statistical models applied to an external population, generally very large. Applying such coefficients to the population under study, we obtain a direct estimate of the probability of the outcome for every patient.
- a severity measure based on an empirical model in which variables' selection and the estimates of their coefficients derive from statistical models applied to the population under study. Therefore the empirical approach uses a severity measure specific for the time and population under study.

Indirect standardization calculates the expected outcome and compares it with the observed outcome for each group.

The outcome expected in each group is based on the distribution (in that group) of factors included in the severity measure. In other words, it is the outcome we would obtain if patients' risk factors were related to the outcome as those in the reference population, that is, the same population used to construct the severity measure. The coefficients can be either empirical or pre-defined.

Calculating the expected outcome

Calculation techniques are different according to the outcome under study and, therefore, to the model used to construct the severity measure:

1. Logistic model

The logistic model can be selected either from an external population (coefficients of risk factors are predefined) or from the population under study (empirical model). Once it has been applied, the probability of a given outcome for patient i is calculated as:

$$p_i = \frac{\exp(b_0 + X_{1i}\, b_1 + X_{2i}\, b_2 + ... + X_{ki}\, b_k)}{1 + \exp(b_0 + X_{1i}\, b_1 + X_{2i}\, b_2 + ... + X_{ki}\, b_k)}$$

The number of events expected in a group will be obtained as sum of the probabilities pi on the total number of subjects included in that group.

2. Poisson's model

The Poisson's model can be selected either from an external population (coefficients of risk factors are predefined), or from the population under study (empirical model). Once it has been applied, the expected number of events μi for that particular combination of selected risk factors is calculated as:

$$\mu_i = \exp\,(b_0 + X_{1i}\, b_1 + X_{2i}\, b_2 + ... + X_{ki}\, b_k)$$

The expected number of events in a group will be obtained as the sum of the expected events according to the different combinations of predictors present in that group.

3. Survival analysis

Once the function h, describing trend of survival time, has been selected and the corresponding model, that can be derived from an external population (coefficients of risk factors are predefined) or from the population under study (empirical model), has been applied, the expected survival time for that combination of selected risk factors is calculated as:

$$t_i = h^{-1}(b_0 + X_{1i}\, b_1 + X_{2i}\, b_2 + ... + X_{ki}\, b_k)$$

The expected time for one group will be obtained as the sum of expected times for all the subjects of that group.

It should be emphasized that expected time is calculable only for models in which survival function is defined in a parametric form.

Comparing observed with expected outcome

Once the expected outcome has been obtained, for each of the groups compared, the standardized ratio (SR) is calculated as:

SR = observed/expected outcome

Standardized ratios tell us how often the outcome in the group considered is more frequent (SR values > 1) , or less frequent (SR values < 1) than it would be based on:

- the distribution of the severity measure in that group;
- the relationship between such a measure and the outcome in the "*reference*" population.

Therefore, indirect standardization allows a "*risk adjusted*" comparison between the outcomes observed in a group and those observed in the reference population.

Direct standardization

A direct comparison between groups is possible by applying the method of direct standardization.

Theoretically, 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). When comparing n groups, n-1 dummy variables must be added to the selected model. The estimates of the coefficients from the non-reference groups are "*risk adjusted*" measures of association (Odds Ratio, Rate Ratio or Hazard Ratio, depending on the study outcome and design) between the exposure, "*belonging to group X and not to the reference group*", and the outcome under study.

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.

Prognostic additive or categorical scores

If severity is measured with a prognostic score, it must be used instead of or in addition to potential confounders in the explicative model containing the outcome variable and the dummy variables representing the groups. The scores must be treated as categorical variables in the analysis.

Pre-defined models

In this case the coefficients estimated in the external population from which the model has been constructed are applied to each subject under study allowing individual probability of the outcome to be calculated based on the risk factors included in the severity measure.

Individual probability will then be included in the explicative model as an independent quantitative variable, with the dummy variables representing the groups, used for the risk-adjusted comparison in the population under study.

3) Identification and management of effect modification

Effect modification [2,7] is a term used, in epidemiology, to describe when one or more factors modify the relationship between the exposure and the outcome under study. It is important to distinguish effect modification from confounding, defined as the existence of a factor associated with the exposure and the outcome under study that is totally or partially responsible for the association (or for the lack of association) observed between the exposure and the outcome. Effect modification implies, on the contrary, a different effect of exposure on the outcome, as a function of a third factor called an "*effect modifier*". In a simple case of a dichotomous effect modifier, the effect of exposure on the outcome will be different in the absence, or presence, of this factor.

Some authors use the term "*effect measure modification*" more appropriately. In fact, the estimate of the effect modification depends on the measure of association used, and on the reference model used to define the combined effect of the two factors. If the association is measured in terms of the relationship between rates, risks, and odds, the effect modification will be estimated as the excess from a multiplicative model (if RR=2 is estimated for one factor and

RR=5 for another, the effect modification is present when the RR with both factors is other than 2x5=10). If, on the other hand, the association is measured in terms of difference between rates, the effect modification will be estimated as the excess from an additive model (if RD=2 for 1000 person-years is estimated for one factor and RD=5 for 1000 person-years for another, the effect modification is present for any RD different from 5+2=7 in the presence of both factors).

For comparative evaluation of outcomes, it is important to define "*a priori"* the potential effect modifiers, based on available evidence or specific research hypotheses. In stratified analysis, effect modification describes the heterogeneity of measures of association across strata; in statistical multivariate models it is determined by adding the interaction terms between exposure and the factor under study. In both cases, whether effect modification is present or not, has to be formally determined by using appropriate statistical tests. The level of sensitivity chosen (that is the p value considered sufficient to reject the null hypothesis of absence of modification) depends on the subjective judgment on the "*importance*" of the effect modification phenomenon to be studied and is conditioned by the size of the population.

Once a significant effect modifier has been detected, risk adjustment procedures subdivide the population under study into as many groups as the levels of the factor considered. The severity measure in each group is calculated again applying the procedures previously described.

4) Precision of estimates

The measures used to compare outcomes are subject, like all others, to errors and bias.

Confounding is a particular type of error, which occurs frequently in observational studies, due to implicit lack of randomization to the exposure, and that can be "*discovered*" and "*corrected*". Effect modification, on the other hand, is not an error, but an actual phenomenon which occurs in the study population; it does not need to be corrected, but measured. In observational studies the confounding effect has the characteristics of "*systematic error*", and has to do with the estimate's validity, and is not reduced by increasing the size of the population observed.

Other systematic errors, that can considerably influence the validity of results but which are beyond the scope of this paper, can derive from inaccurate measurements of subjects' attribution to groups, from inaccurate assessment of the outcome, from inaccurate measurements of potential confounders, meaning misclassification of exposure, of outcome and confounders respectively.

A different problem concerns the possibility that the measures produced (either "*crude*" or "*adjusted*" to take account of the confounding) are affected by "*random error*".

The precision in epidemiologic measurements corresponds to the reduction of random error.

The random error can be defined as an error which occurs as a result of the process of selecting a specific sample and varies unpredictably in size and direction. It is not a systematic error, and tends to decrease as the sample size increases. The magnitude of its effect can be quantified using inferential statistical methods [12, 23, 29].

This can be made in two ways:

- calculating the estimate, as a single value and as an interval, called the "*confidence interval*", to which the probability of containing the "*true*" value of the measure is attributed;
- calculating the probability (p-value) that the differences observed among the outcomes in the compared groups are entirely due to random error, under the hypothesis that there are no differences ("*null*" hypothesis) between the "*true*" occurrence of outcomes in the compared groups.

The two approaches are based on the same theoretical assumptions and are strongly related. In fact, if the 95% confidence interval of a two group comparison does not include the value assumed by the outcome measures when the 2 groups have the same outcome ("*null*" hypothesis), the p-value of that comparison will be less than 5%.

In general, the measures of association for samples of similar size, adjusted using an empirical severity measure constructed with a large amount of variables tend to be less precise than more "*parsimonious*" measures.

Even if the most strictly "*scientific*" part of the studies on comparative evaluation of outcomes is to produce punctual estimates of risk-adjusted measures of association, their confidence intervals and their p-values, the nature of the comparisons, in most cases, require some qualitative conclusions:

- is it possible to say that treatment X is more effective than treatment Y?
- is it possible to say that the hospitals A, B and C have worse performances than the national average?
- is it possible to say that the residents in the region K have, severity being equal, worse outcomes than the residents in the region H?

Putting aside the problems of data accuracy or hidden and therefore uncontrollable confounding factors, it is possible to answer these questions by defining a traditional level of statistical significance from which we can determine the existence of a "*significant*" difference between groups. It is then necessary to define a threshold for the p-values, below which the observed differences can be considered "*significant*", since the probability of error is "*acceptable*".

Since we think it indispensable that the threshold level is chosen "*a priori*", independently of the results of the evaluation, and made known to decision makers beforehand, such a choice should be based on a careful evaluation of the possible costs and benefits of identifying as "*worse*" a group that is "*truly*" worse than others, compared to the costs and benefits of declaring a group worse, when actually it is equal to the others.

In any case, we should be careful to avoid confusion between precision of estimates, their statistical significance, on one side, and validity on the other. Very precise estimates that are strongly biased or "*statistically non-significant*" estimates that are valid could be obtained. In no instance the statistical significance of an estimate must be interpreted as a judgement on its true or false nature.

Identification of actual confounders

The selection of factors that actually produce confounding should be specific for each comparisons. The inclusion within a risk adjustment model of factors that do not actually induce a relevant bias in the estimate of the measure of association may cause a loss of precision and implies additional costs of collecting the relevant information. Therefore the selection of the "best" risk adjustment models should aim to the maximum parsimony.

The *"Change-in-estimate"* is one of the available methods to select actual confounders [2, 20, 62]. It improves parsimony of the model and gains precision of the estimates, by eliminating variables that are not actual confounders.

According to this method, initially a "*less parsimonious*" model is chosen. This model includes all selected potential confounders and the exposure of interest. Subsequently, all those factors which do not modify, or modify the estimated effect of exposure only by a small amount, are excluded from the model. The variation of the estimates usually considered acceptable, to include the confounder, can vary according to the phenomenon studied (usually from 10% to 20%). However the choice is arbitrary; the relevance of the observed confounding effect must be appraised in relation to the study specific aim, design and utilization of the study results.

This method can only be used to compare one group with the reference.

5) Multiple comparisons

Usually, in the evaluation of actual outcomes more than two groups (e.g. different providers, hospital or areas) are compared [63, 64].

Many analysis methods suit multiple comparisons without particular problems, but in some cases it is necessary to adopt specific precautions.

In the indirect standardization method, each group can be compared with an external population (when using pre-defined models) or, in the case of empirical measures, with all groups under study or with a particular subgroup (benchmark). In the latter case, the groups included in the benchmark must be large enough to assure that the precision of the parameters' estimates, and consequently those of the expected outcomes, is satisfactory. Since comparisons are made between each group and the reference, it is not correct to use standardized ratio values to make comparisons between groups. This could be possible only under the hypothesis, to be verified case by case, that the characteristics used for the adjustment of the outcome under study are homogenously distributed between groups.

In the "*change-in-estimate*" procedure, the actual confounders can be identified only for single comparisons and therefore this procedure must be repeated for each pair of groups considered, defining, as many risk adjustment models as there are comparisons. It would probably increase the precision of estimates, but would make impossible to compare the groups directly, despite the use of direct standardization.

One possible solution to this problem is to use, for all comparisons, a single risk adjustment model, including the true confounders in at least one of the comparisons of interest. This could be "*the best possible compromise*" between the demands of parsimony and making valid multiple comparisons. However, if there were several groups to compare, this solution would be too onerous, in terms of time and in terms of calculation complexity. In this case, therefore, it is better to use for all comparisons the risk adjustment model containing all factors selected based on their multivariate association with the outcome. The latter is the option usually chosen, for example, for comparative evaluations of providers at the regional or national level.

6) Other models used for risk adjustment

The models used for direct standardization are usually criticized for three reasons:

a. they do not expressly consider the "*hierarchical*" nature of data, i.e. aggregation of patients studied in hospitals, geographic area, administration typology, etc.
b. if there are no cases in some of the "*cells*" the models will not converge;
c. they do not allow the analysis of the effect of the variables associated to that group (geographic area, administrative typology).

Hierarchical structure cannot be considered as fortuitous or be ignored during data analysis, because without necessary corrective actions it tends to produce levels of precision that are not justified by the data analyzed and to underestimate the random variability between groups [17, 65].

A well-known study conducted by Bennett in 1976, reported greater progress in elementary school children who had been "*exposed*" to a specific style of teaching. Data were analysed using a multiple regression analysis with a "*classic*" approach [66].

Subsequently, in 1981, Aitkin demonstrated that when the hierarchical nature of the data was considered, such differences disappeared. In conclusion, the children tended to have more similar rates of progress [67].

The hierarchical (or "*multi-level*") models introduced in the early '90s by Goldstein address such problems. They expressly consider that the aggregation of patients of various groups is not due to chance and the differences, adjusted for the different case-mix, need to be specifically modelled. In this way the punctual estimates are concentrated around the general mean, meaning they are more conservative [17, 68-71].

Various statistical approaches exist that specify models like this, but, as indicated by Goldstein and Spiegelhalter, "*statistical preferences between Bayesian, likelihood and quasi-likelihood methods are usually more of philosophical then practical importance*" [72].

Until now the main problem limiting the use of hierarchical models has been the poor diffusion of software for statistical analysis. This limit also is going to be overcome.

Conclusions

This review of epidemiologic methods to compare health care outcomes cannot be considered exhaustive. Its basic message is that empirical techniques should be used whenever appropriate data are available. No sophisticated empirical method is generally enough: a model that contradicts established physiologic principles is not valid, regardless of the statistical rigor used in its derivation. The methods described, like any type of scientific knowledge, cannot measure "*reality*" as it "*truly*" is, but can produce "*images*" of it, defining limits and uncertainties in terms of validity and precision. On these assumptions, everyone, politicians, managers, epidemiologists, and clinicians should make decisions based on the validity and precision of study results in order to promote and improve the health of patients and population, by using the best scientific knowledge available.

References

1. Arcà M, Fusco D, Barone AP, Perucci CA. Risk adjustment and outcome research. Part I. J Cardiovasc Med (Hagerstown). 2006;7(9):682-90
2. Rothman KJ, Greenland, eds. Modern epidemiology. Philadelphia: Lippincott-Raven 2nd ed., 1998
3. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. N Engl J Med. 2000; 342(25):1878-86
4. Vandenbroucke JP. When are observational studies as credible as randomised trials? Lancet 2004; 363(9422):1728-31
5. Deeks JJ, Dinnes J, D'Amico R, Sowden AJ et al. Evaluating non-randomised intervention studies. Health Technology Assessment 2003; 7(27): 1-186
6. Greenland S, Brumback B. An overview of relations among causal modelling methods. International Journal of Epidemiology 2002; 31:1030-1037
7. McNamee R. Confounding and confounders. Occupational and Environmental Medicine 2003; 60: 227-23
8. Sonis J. A Closer Look at Confounding. Family Medicine 1998; 30(8): 584-8
9. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic Research. Principles and quantitative methods. New York: Van Nostrand Reinhold Company Inc., 1982
10. Miettinen OS, Cook EF. Confounding: essence and detection. American Journal of Epidemiology 1981; 114: 593-603
11. Iezzoni LI. Risk Adjustment for measuring healthcare outcomes. Health Administration Press 2nd ed., 1997
12. Pearson ML, Stecher B. Risk Adjustment Methods in Health Care Accountability. In Stecher B, Kirby SN, (eds). Organizational Improvement and Accountability: Lessons for Education from Other Sectors, RAND, MG-136-WFHF, Chapter 7, 2004, pp. 95-105
13. Shaughnessy PW, David FH. Overview of Risk Adjustment and Outcome Measures for Home Health Agency OBQI Reports: Highlights of Current Approaches and Outline of Planned Enhancements. Center for Health Services Research, UCHSC. September 2002 (http://www.cms.hhs.gov/quality/hhqi/RiskAdj1.pdf)
14. Spiegelhalter D, Grigg O, Kinsman R, Treasure T. Risk-adjusted sequential probability ratio tests: applications to Bristol, Shipman and adult cardiac surgery. International Journal for Quality in Health Care 2003; 15:7-13
15. Silva LK. Validity of the risk adjustment approach to compare outcomes. Cad Saude Publica 2003;19(1): 287-95
16. O'Keefe K. Accounting for Severity of Illness in Acutely Hospitalized Patients: a Framework for Clinical Decision Support using DYNAMO. Wipro GE Healthcare. Copyright General Electric Company 1997-2005
17. (http://www.gehealthcare.com/inen/prod_sol/hcare/resources/library/article07.html)

18. Ferraris VA, Ferraris SP Risk Stratification and Comorbidity. In: Cohn LH, Edmunds LH Jr, eds. Cardiac Surgery in the Adult. New York: McGraw-Hill 2003:187224
19. Marshall G, Henderson WG, Moritz TE, Shroyer AL, Grover FL, Hammermeister KE. Statistical methods and strategies for working with large data bases. Medical Care 1995; 33(10 Suppl):OS35-42
20. Robins JM, Greenland S. The role of model selection in causal inference from nonexperimental data. American Journal of Epidemiology 1986;123(3): 392-402
21. Greenland S. Modeling and variable selection in epidemiologic analysis. American Journal of Public Health 1989; 79(3): 340-349
22. Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. J Clin Epidemiol 1996 Aug;49(8): 907-16
23. Kleinbaum DG. Epidemiologic methods: the “art" in the state of the art. J Clin Epidemiol. 2002; 55(12):1196-1200
24. Clayton D, Hills M. Statistical Models in Epidemiology. Oxford University Press, New York, 1993
25. Kleinbaum DG, Kupper LL, Muller KE, Nizam A. Applied Regression Analysis and other multivariable methods. Duxbury Press by Brooks/Cole Publishing Company 3nd ed., 1998
26. Brown H, Prescott R. Applied mixed models in medicine. John Wiley & Sons, Ltd, 2003
27. Hosmer DW, Lemeshow S. Applied Logistic Regression. New York: Wiley,1989
28. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143(1): 29-36
29. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988; 44: 837-845
30. DeLong ER, Peterson ED, DeLong DM, Muhlbaier LH, Hackett S, Mark DB. Comparing risk-adjustment methods for provider profiling. Statistics in Medicine 1997;16(23): 2645-64
31. Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG, et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. Chest 1991;100:1619-36
32. Thomas JW, Ashcraft ML. Measuring severity of illness: six severity systems and their ability to explain cost variations. Inquiry 1991; 28(1): 39-55
33. Knaus WA, Wagner DP, Zimmerman JE., Draper EA. Variations in Mortality and Length of Stay in Intensive Care Units. Annals of Internal Medicine 1993; 118(10): 753-761
34. Iezzoni LI, Ash AS, Shwartz M, Daley J et al. Predicting Who Dies Depends on How Severity Is Measured. Implications for Evaluating Patient Outcomes. Ann Intern Med 1995; 123(10): 763-770
35. Tu JV, Jaglal SB, Naylor CD. Multicenter validation of a risk index for mortality, intensive care unit stay, and overall hospital length of stay after cardiac surgery. Steering Committee of the Provincial Adult Cardiac Care Network of Ontario. Circulation 1995; 91: 677- 684
36. Guru V, Gong Y, Rothwell DM, Tu JV. Report on Cardiac Surgery in Ontario Fiscal Years 2000 & 2001. The Institute for Clinical Evaluative Sciences, Toronto Ontario, Canada in collaboration with the Steering Committee of the Cardiac Care Network of Ontario, 2003
37. Nashef SAM, Roques F, Michel P, Gauducheau E et al. European system for cardiac operative risk evaluation (EuroSCORE). European Journal of Cardio-thoracic Surgery 1999; 16: 9-13
38. Iezzoni LI, Shwartz M, Ash AS, Hughes JS, Daley J, Mackiernan YD, et al. Evaluating severity adjustors for patient outcome studies. Final report. Prepared for the Agency for Health Care Policy and Research under grant no. RO1-HS06742. Boston: Beth Israel Hospital, 1995

39. 3M Health Information Systems. All Patient Refined DRGs (APR-DRGs), 1995. (http://www.3mhis.com)
40. Edwards N, Honemann D, Burley D, Navarro M. Refinement of the Medicare diagnosis-related groups to incorporate a measure of severity. Health Care Financing Review 1994; 16(2): 45–64
41. Mendenhall S. DRGs must be changed to take patient's illness severity into account. Modern Healthcare 1984 Nov 15; 14(15): 86-8
42. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. Journal of Chronic Diseases 1987; 40(5): 373-83
43. Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. Journal of Clinical Epidemiology 1992; 45(6): 613-9
44. Higgins TL, Estafanous FG, Loop FD, et al. Stratification of morbidity and mortality outcome by preoperative risk factors in coronary artery bypass patients: a clinical severity score. JAMA 1992; 267(17): 2344-8
45. Gonnella JS, Hornbrook MC, Louis DZ. Staging of disease: A case-mix measurement. JAMA 1984; 251 (5): 637-44
46. Markson LE, Nash DB, Louis DZ, Gonnella JS. Clinical outcomes management and disease staging. Evaluation and the Health Professions 1991; 14(2): 201-27
47. O'Connor GT, Plume SK, Olmstead EM, et al. Multivariate prediction of in-hospital mortality associated with coronary artery bypass graft surgery. Northern New England Cardiovascular Disease Study Group. Circulation 1992; 85:2110-18
48. O'Connor GT, Plume SK, Olmstead EM, et al. A regional intervention to improve the hospital mortality associated with coronary artery bypass graft surgery. The Northern New England Cardiovascular Disease Study Group. JAMA 1996; 275(11): 841-6
49. Parsonnet V, Dean D, Bernstein AD. A method of uniform stratification of risk for evaluating the results of surgery in acquired adult heart disease. Circulation 1989; 79(6 Pt 2): I3-12
50. Parsonnet V, Bernstein AD, Gera M. Clinical usefulness of risk-stratified outcome analysis in cardiac surgery in New Jersey. Ann Thorac Surg 1996; 61(2 Suppl):S8-11; discussion S33-4
51. Martinez-Alario J, Tuesta ID, Plasencia E, et al. Mortality prediction in cardiac surgery patients: comparative performance of Parsonnet and general severity systems. Circulation 1999; 99: 2378-2382
52. Young WW, Kohler S, Kowalski J. PMC Patient Severity Scale: derivation and validation. Health Services Research 1994; 29(3): 367-90
53. Steen PM, Brewster AC, Bradbury RC, Estabrook E, Young JA. Predicted probabilities of hospital death as a measure of admission severity of illness. Inquiry 1993; 30: 128-141
54. Hammermeister KE, Johnson R, Marshall G, Grover FL. Continuous assessment and improvement in quality of care: a model from the Department of Veterans Affairs Cardiac Surgery. Ann Surg 1994; 219(3): 281-90
55. Grover FL, Shroyer AL, Hammermeister KE. Calculating risk and outcome: the Veterans Affairs database. Ann Thorac Surg 1996; 62(5 Suppl): S6-11; discussion S31-2
56. Hannan EL, Kilburn H, O'Donnell JF, et al. Adult open heart surgery in New York State: an analysis of risk factors and hospital mortality rates. JAMA 1990; 264(21): 2768-74
57. Hannan EL, Kumar D, Racz M, Siu AL, Chassin MR. New York State's Cardiac Surgery Reporting System: four years later. Ann Thorac Surg 1994; 58(6):1852-1857
58. Edwards FH, Clark RE, Schwartz M: Coronary artery bypass grafting: the Society of Thoracic Surgeons National Database experience. Ann Thorac Surg 1994; 57:12-19
59. Hattler BG, Madia C, Johnson C, et al: Risk stratification using the Society of Thoracic Surgeons Program. Ann Thorac Surg 1994; 58:1348-52

60. Edwards FH, Grover FL, Shroyer AL, et al: The Society of Thoracic Surgeons National Cardiac Surgery Database: current risk assessment. Ann Thorac Surg 1997; 63: 903-908
61. Nursing Home Quality Measures Resource Manual. Revised Resource Manual, Texas Medical Foundation 2004 (http://www.tmf.org/nursinghomes/manual/)
62. Bailit J, Garrett J. Comparison of risk-adjustment methodologies for cesarean delivery rates. Obstet Gynecol. 2003;102(1): 45-51
63. Mickey RM, Greenland S. The impact of confounder selection criteria on effect estimation. American Journal of Epidemiology 1989; 129(1): 125-137
64. Health Grades. The Healthcare Quality Experts. Copyright 1999-2005 Health Grades, Inc. (http://www.healthgrades.com)
65. Krumholz HM, Rathore SS, Chen J, Wang Y, Radford MJ. Evaluation of a Consumer-Oriented Internet Health Care Report Card: the Risk of Quality Ratings Based on Mortality Data. JAMA 2002; 287:1277-1287
66. Goldstein H. Multilevel Statistical Models, 3rd ed. Hodder Arnold, 2003
67. Bennett N. Teaching Styles and Pupil Progress. Open Books, 1976
68. Aitkin, M., Anderson, D., Hinde, J. Statistical modelling of data on teaching styles. Journal of the Royal Statistical Society 1981; Part A, 144: 148-161
69. Aylin P, Alves B, Best N, Cook A et al. Comparison of UK paediatric cardiac surgical performance by analysis of routinely collected data 1984-96: was Bristol an outlier? Lancet 2001; 358(9277):181-187
70. Shahian DM, Blackstone EH, Edwards FH et al.; STS workforce on evidence-based surgery. Cardiac surgery risk models: a position article. Ann Thorac Surg. 2004 Nov; 78(5):1868-77
71. Normand SL, Glickman M, Gatsonis CA. Statistical methods for profiling providers of medical care: issues and applications. Journal of the American Statistical Association 1997; 92: 803-814
72. Sklo M, Nieto FG. Epidemiology: beyond the basics. Aspen Publishers, Inc. Gaithersburg, Maryland, 2000
73. Goldstein H, Spiegelhalter D. League tables and their limitations: statistical issues in comparisons of institutional performance. Journal of the Royal Statistical Society 1996; 159: 385-443