available at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejconline.com

Evaluation of alternative prognostic stratifications by prediction accuracy measures on individual survival with application to childhood leukaemia

Paola De Lorenzo[a,b,c,*], *Laura Antolini*[a,c], *Maria Grazia Valsecchi*[a,c]

[a] *Biostatistics Centre, Department of Clinical Medicine and Prevention, University of Milano-Bicocca, Via Cadore 48, 20052 Monza, MI, Italy*
[b] *Statistical Unit, Paediatric Clinic, University of Milano-Bicocca, Via Pergolesi 33, 20052 Monza, MI, Italy*

ARTICLE INFO

Article history:
Received 16 October 2008
Received in revised form 4 December 2008
Accepted 13 January 2009
Available online 14 February 2009

Keywords:
Cancer outcome
Prognostic stratifications
Prediction accuracy
Individual survival

ABSTRACT

A common aim of clinical research is the identification of patients at different prognosis, so that future treatment protocols may be tailored to patients' risk profiles. Establishing a prognostic classification is a difficult process, especially in rare cancers: the availability of a growing number of candidate prognostic factors may lead to competitive stratifications, whose validation may not be feasible due to small numbers and the need for a long follow-up. Our goal is to illustrate a strategy to compare stratifications, based on the performance in clinically relevant subgroups of patients. We investigated different statistical measures and recommend a strategy based on the Brier Score, a measure of prediction inaccuracy on individual survival. Results on an infant leukaemia study show that this method is flexible and easily applied with common statistical software. The method, however, does not overcome the problem of lack of validation on external data.

1. Introduction

A common aim of clinical research is the identification of patients at different prognosis, so that future treatment protocols may be tailored to patients' risk profiles. These profiles are defined on the basis of suitable prognostic factors, i.e. patients' features able to explain the heterogeneity in outcome. From a practical point of view, few risk groups are desirable, as each of them is assigned a different treatment protocol. Moreover, it is important that the prognostic classification provides the greatest discrimination in outcome and that it is validated and possibly calibrated on new data, independent of those used in its development.

Various techniques are available to assess prognostic classifications, for example, measures of discrimination or inaccuracy.[1–3] Methodology has been intensively studied, especially to address validation problems. In practice, a case-by-case approach is desirable because each method has different properties and drawbacks. We focus here on a peculiar problem that arises in rare cancers, where typically the small number of patients involved needs to be followed-up for a long period before outcome can be evaluated. In this context, the prognostic analysis is run under two major constraints: (1) the limited information and (2) the (possibly) large number of candidate prognostic factors, originated by expanding biological knowledge (e.g. when biomarkers and gene expression profiling are added to established prognostic factors). The joint analysis of potential factors may therefore result in alternative stratifications with similar prognostic discrimination ability. Since in this context a proper external

[*] *Corresponding author*: Address: Statistical Unit, Paediatric Clinic, University of Milano-Bicocca, Via Pergolesi 33, 20052 Monza, MI, Italy. Tel.: +39 039 233 3074; fax: +39 039 233 2314/230 1646.
E-mail address: paola.delorenzo@unimib.it (P. De Lorenzo).

[c] These authors contributed equally to the paper.

0959-8049/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ejca.2009.01.011

validation step is unfeasible in most of the cases, we consider the assessment of competitive stratifications by means of the same data on which they were derived. We take into account diverse methods. From the clinical point of view, the most valuable assessment is based on the prediction accuracy at the patient level, evaluated by contrasting the observed individual survival with the predicted one. We therefore adopt this approach and investigate the performance of prognostic classifications by using appropriate decompositions of the Brier Score, a well-known inaccuracy measure that has previously been applied to other settings.[4]

2. Patients and methods

2.1. *The motivating example*

We illustrate the problem with data from Interfant-99, an international multi-centre study in infant Acute Lymphoblastic Leukaemia (ALL), described elsewhere.[5] Four hundred and eighty two infants aged 0–12 months with newly diagnosed ALL were enrolled in the study between 1999 and 2005 and followed-up until December 2006. Median follow-up was about 4 years, ranging between one month and 7.5 years. The cumulative proportion of censored patients at 6 months, 1, 1.5 and 2 years is 1%, 1%, 5% and 8%, respectively. One of the aims of this study was the joint assessment of many candidate prognostic factors, i.e. status of the *MLL* gene (rearranged or not), age at diagnosis, white blood cell (WBC) count at presentation, immunophenotype and early response to Prednisone. The primary end-point was the event-free survival (EFS) defined as the time from diagnosis till one of the following events: early failure (resistance or death), relapse, death in complete remission, second malignancy. The joint analysis of all five candidate prognostic factors was based on 374 patients with complete data and aimed at the identification of three risk groups. Two extreme and ideally small groups, the Low Risk (LR) and High Risk (HR) groups with very good and very bad prognosis, respectively, and a large, remaining group, the Intermediate Risk (IR) group. This choice was dictated by the clinical purpose of defining a rule for treatment allocation of future patients. The analysis led to two alternative stratifications: investigation on possible model refinements did not result in a clear indication as to which classification to recommend for a new clinical protocol.

Both stratifications were identified as a three-variable rule, comprising MLL gene status, age at diagnosis and either WBC count at diagnosis or early response to Prednisone (see Table 1). Both these latter two variables actually convey information on the disease burden. WBC measures it at diagnosis, while the early response to Prednisone represents the residual disease after 1 week of treatment with Prednisone, a consolidated pre-chemotherapy course in most treatment protocols in childhood leukaemia. In detail, while the LR stratum was common to both stratifications, HR and IR patients had conflicting definitions. Nonetheless, the distributions of subjects and events among strata were very similar and as a consequence, only in the HR group the Kaplan–Meier EFS estimates showed some marked difference (Fig. 1). However, by cross-tabulating the stratifications we could clearly identify two subsets of patients which were classified at a different risk depending on the stratification (see Table 1). If, for instance, we focus on the subgroup of 30 patients at HR for Stratification 1 and at IR for Stratification 2, the predicted EFS assigned at the individual level is represented by the HR solid line and by the IR dashed line, respectively (Fig. 1). As HR patients qualify for a more intensive (and possibly more toxic) treatment than IR patients do, this makes quite a difference in practice and deserves deep investigation into the stratifications' performance within subgroups.

2.2. *Statistical methods*

A number of measures of predictive accuracy are available in the literature, including the Harrell's C-index,[1] indexes of prognostic separation, such as SEP, and, more recently, the D-index[3] and measures of inaccuracy at the individual level, for instance, the Brier Score.[2] In what follows, we focus on the application of the C-index and of the Brier Score which, as discussed later, seems more appropriate to our aim.

We assume that the following data have been recorded for each subject: (1) the time-to-event or survival time (if no event is observed, patient is censored at last follow-up) and (2) the risk stratum according to a given classification rule. For example, if the rule is based on the Cox regression model, the risk strata are derived from the linear combination of the prognostic factors (the linear predictor), while in other contexts, such as the classification and regression tree (CART) analysis, strata are simply defined by indicator functions based on the selected covariates.

We consider a scenario in which the classification rule can only be evaluated by using the same dataset that was used to construct it. The resulting over-optimism that affects the estimates of the C-index and the Brier Score can be corrected by applying bootstrap methods. Bootstrap is also used here for the computation of the confidence intervals for the Brier Score.[1,2]

2.2.1. *The Harrell's C-index*

The idea behind this measure is the comparison between ranks of predicted and observed survival times. Given a pair of subjects (i, l), such that i fails while l is still event-free and assuming separation between predicted survival curves ('one-to-one correspondence' between predicted times and predicted survival functions, see[6]), the C-index may be defined as the conditional probability that i, who failed before l, is assigned a lower predicted survival than l is, for any time-point. The estimation of this quantity is commonly computed as the ratio of the number of pairs in which the rank of predicted survival functions equals that of survival times, over the number of pairs for which the underlying survival times can be ranked, i.e. for which the shortest time is an event time. Couples of subjects who are assigned the same predicted survival (tied pairs) cannot be ordered and are therefore assigned a conventional 0.5 weight.

2.2.2. *The Brier Score*

The Brier Score evaluates the prediction error at the individual level. For any subject i, $i = 1, \ldots, n$ assigned to risk stratum j, the observed survival status at a given time-point t (say 0 if failure, 1 if survivor) is compared to the predicted survival at t.

Table 1 – Number of patients (number of events) by two competitive prognostic stratifications and corresponding risk groups.

Stratification 1		Stratification 2			
		LR	IR	HR	Total
		MLL negative	Otherwise	MLL positive *and* Age < 6 months *and* Poor PDN response	
LR	MLL negative	78 (18)	0	0	78 (18)
IR	otherwise	0	200 (105)	***24 (21)***	224 (126)
HR	MLL positive *and* age < 6 months *and* WBC ⩾ 300,000	0	***30 (23)***	42 (33)	72 (56)
Total		78 (18)	230 (128)	66 (54)	374 (200)

Abbreviations: LR = Low Risk, IR = Intermediate Risk, HR = High Risk, WBC = White Blood Cell count (cells per μl), and PDN = Prednisone. MLL negative means MLL gene not rearranged, whereas positive means rearranged. Poor Prednisone response is defined as a blast count of 1000 cells per μl or more, after 7 days of therapy.

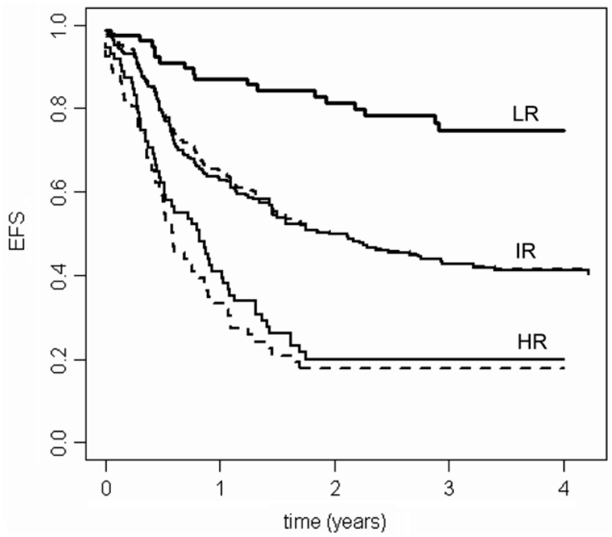

Fig. 1 – Stratum-specific Kaplan–Meier EFS predictions under Stratification 1 (solid-line curves) and Stratification 2 (dashed-line curves). Prediction in LR is represented by one curve only, because LR is the same for both stratifications. Abbreviations: LR = Low Risk, IR = Intermediate Risk, and HR = High Risk.

In practice, the Brier Score is estimated by computing the quadratic difference between observed and predicted survival (the individual error, ER_i) and taking the average over all patients

$$BS(t) = \frac{1}{n}\sum_{i=1}^{n} ER_i. \tag{1}$$

Computation of ER_i for patients censored before t is not possible, because their survival status at t is unknown by definition. To account for censoring in (1), the ER_i components are therefore weighted by a multiplicative factor corresponding to the inverse probability of censoring. Consider, for instance, ER_i for patients who failed before t. The weighting is such that, the lower the probability of being under observation at least up to t, the bigger the weight assigned to ER_i, so that this also accounts for patients who could have been observed to fail within t, if they had not been censored before.

As the Brier Score is estimated with the sample mean error (1), the prediction inaccuracy due to subgroups can be easily isolated by writing the score as a weighted mean of mean errors in subgroups. For example, if g risk strata are defined, and the Brier Score for patients assigned to each stratum j is denoted by $BS_j(t)$, $j = 1,\ldots,g$, we can write $BS(t) = \frac{1}{n}\sum_{j=1}^{g} n_j BS_j(t)$, with $n = \sum_{j=1}^{g} n_j$. The decomposition isolating groups of patients classified in different risk strata depending on the stratification adopted can be used to compare the performance of different prognostic models on the same data.

The Brier Score results can be further illustrated by means of a measure of explained residual variation, which contrasts the Brier Score obtained when the indistinct prediction $BS^0(t)$ is assumed for all strata (e.g. the overall estimated Kaplan–Meier survival), with the Brier Score resulting from the model

$$R^2(t) = \frac{BS^0(t) - BS(t)}{BS^0(t)}. \tag{2}$$

This can be interpreted as the gain in accuracy which is achieved when the indistinct prediction is replaced by the stratum-specific, covariate-driven predicted survival.

3. Results

Findings illustrated in this section are obtained applying the `sbrier` and `rcoor.cens` functions available in *R* (`ipred` and `Hmisc` package, respectively) and some ad-hoc written routines to the competitive stratifications described in Section 2.1.[7]

3.1. *The Harrell's C-index*

For both candidate prognostic models in infant ALL, we calculated the Harrell's C-index and its 95% bootstrap confidence interval by taking $B = 1000$ samples from the original dataset. C was equal to 0.679 (95% CI 0.624–0.735) for Stratification 1 and equal to 0.676 (95% CI 0.618–0.734) for Stratification 2.

Results were almost overlapping, indicating that the two classification schemes are nearly indistinguishable in their prognostic discrimination ability when this is measured by the C-index.

3.2. The Brier Score

Brier Score estimates, calculated according to (1) at clinically relevant time-points t, are shown in Fig. 2 for the two prognostic models in infant ALL and for the overall Kaplan–Meier EFS. Bootstrap confidence intervals are again obtained by taking $B = 1000$ samples from the original dataset. Each model actually shows a more accurate prediction than the overall EFS, but Stratification 2 seems to perform consistently better than Stratification 1 across all time-points. Moreover, prediction is more accurate at early time-points, even when the indistinct prediction is used. As a result, the explained variation $R^2(t)$ increases with time and is consistently higher under Stratification 2 as compared to Stratification 1. $R^2(t)$ varies from 4.5% (95% CI 0.2–9.1) at $t = 0.5$ years to 15.5% (95% CI 9.2–21.7) at $t = 2$ under Stratification 1, and from 5.7% (95% CI 0.6–10.8) at $t = 0.5$ to 16.1% (95% CI 9.4–22.8) at $t = 2$ under Stratification 2.

These findings provide only a partial answer to the initial questions, as they give no insight into the performance of the competitive stratifications in the two subsets in which they conflict (see Table 1). In order to clarify this aspect, we calculate a decomposition of the overall Brier Score. We consider three major components: $BS_{CON}(t)$, the error contributed by 320 patients who are classified in the same strata irrespective of the prognostic model; $BS_{30DIS}(t)$, the error due to the 30 patients who are classified at HR under Stratification 1 but at IR under Stratification 2 and finally $BS_{24DIS}(t)$, the error due to the remaining 24 patients discordantly classified at IR under Stratification 1 and at HR under Stratification 2. The Brier Score on the overall 374 patients may then be written as follows:

$$\mathrm{BS}(t) = \frac{1}{374}[320\mathrm{BS}_{\mathrm{CON}}(t) + 30\mathrm{BS}_{30DIS}(t) + 24\mathrm{BS}_{24DIS}(t)].$$

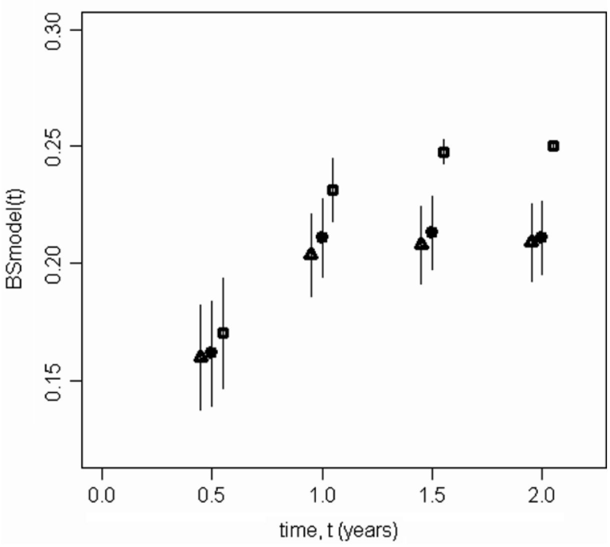

Fig. 2 – 95% Bootstrap confidence interval estimates $BS(t)$ at time-points t = 0.5, 1, 1.5 and 2 years from diagnosis, for Stratification 1 (circles), Stratification 2 (triangles) and for the indistinct prediction with the overall Kaplan–Meier EFS estimate ($BS^0(t)$, represented by squares).

For instance, at $t = 1$ year, $BS_{30DIS}(t)$ was equal to 0.261 and 0.276 under Stratification 1 and 2, respectively, showing a slightly better performance under Stratification 1 that classifies these 30 patients at HR. On the contrary, $BS_{24DIS}(t)$ was 0.328 and 0.208 under Stratification 1 and 2, respectively, indicating that Stratification 2 achieves in this case a more accurate prediction. Notice, however, that Stratification 2 allocates the 24 patients to the HR stratum. In conclusion, for both subsets in which the models give conflicting classifications, the better prediction seems to be ensured by assignment to HR. We calculated this partition at other clinically relevant t, obtaining similar results over time.

Pushing this reasoning further, one may therefore end up with a new stratification proposal in which the HR stratum is defined as a combination of the two alternative HR strata: MLL gene rearrangement *and* age <6 months *and* (poor Prednisone response *and/or* WBC $\geqslant$300,000 cells per µl). LR stratum would remain unchanged, while IR would again be defined as the complementary stratum. If this New Stratification is adopted, the EFS predictions and their accuracy by Brier Score compare to the corresponding quantity for the initial models as shown in Fig. 3 (see grey solid line curve in left-hand graph and grey squares in right-hand graph, respectively). Predictions based on the New Stratification seem to be at least as accurate as those originally available and superior at late time-points.

4. Discussion

In this paper, we illustrated commonly used measures for comparison of prognostic models. Our motivating example on infant ALL introduced a non-standard problem. Existing data led to two competitive, apparently equivalent prognostic classification schemes that could not be validated on new data, but were rather the only evidence which future treatments could be based on. Such situations may frequently arise in clinical research in small populations or in rare diseases. From the statistical point of view, when a prognostic model is fitted and the measure of discrimination or inaccuracy is calculated on the same data, it is expected that this gives an overoptimistic evaluation of the model's performance. To account for this, it is possible to apply a bootstrap resampling technique that results in corrected measures. We recognise, however, that definitive results may only be achieved by external validation.

We first explored the use of measures of separation, in particular the Harrell's C-index that in our application gave inconclusive results. This is not completely unexpected, as both candidate models offer a prognostic classification that is based on few (three) strata and thus induces a high occurrence of ties. In practice, ties between predictions within pairs prevent the required ranking, which can only be approximated with a conventional weight in the formula. Of course, the more frequent the ties, the less meaningful the value attained by the index. Impracticability of these within pair com-

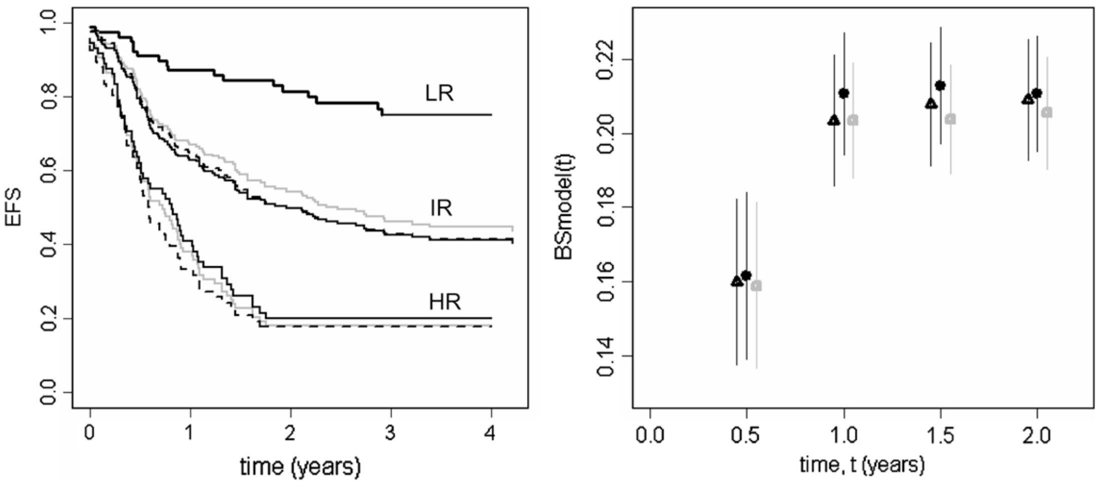

Fig. 3 – Left-hand graph: stratum-specific Kaplan–Meier EFS predictions under Stratification 1 (black solid line), Stratification 2 (black dashed line) and Stratification New (grey solid line). Right-hand graph: 95% bootstrap confidence interval estimates $BS(t)$ at time-points t = 0.5, 1, 1.5 and 2 years from diagnosis, for Stratification 1 (circles), Stratification 2 (triangles) and Stratification New (grey squares). Abbreviations: LR, Low Risk; IR, Intermediate Risk; HR, High Risk.

parisons is the crucial weakness of the method in these type of applications, because it does not allow the analysis of the model's performance within subsets of patients with conflicting stratifications. In addition, censored survival times give only a partial contribution to the C-index, that is when censoring does not prevent the times ranking within the couple. However, no formal adjustment for censoring is available in the literature; this deserves further methodological work.

Other measures of discrimination, namely SEP and D-index, were applied but proved unable to detect any difference in the performance of the two stratifications. In our opinion, these unsatisfactory results (not reported) reflect the drawback already found for the Harrell's C-index applied in the context of few risk strata and assessment of individual predictions.

For these reasons, we turned our attention to a measure of inaccuracy of predictions at the individual level, the Brier Score. This is estimated with an overall mean prediction error; therefore, it can be decomposed to gain additional insight into contributions due to individuals or relevant subgroups. Investigation into the model's performance can in fact be customised to the appropriate detail. In our case, this allowed adjustment of the original classification strategy, leading to a final proposal. An important advantage of this methodology is that computations are promptly performed with routines available in widely used statistical packages. As for other measures of inaccuracy, it is difficult to interpret the calculated absolute values of the Brier Score and, as a consequence, of the explained variation R^2, since the 'maximum attainable' level is not easily obtained. However, relative comparisons among stratifications are feasible and, in our data, both R^2 and the Brier Score showed the same tendency, at any time-point. A complete picture over time could be gained by calculating the integrated Brier Score,[2] although in practice interest is frequently confined to clinically relevant time-points.

It could be argued that the problem of selection between candidate models could also have been solved within the model development framework. For example, the analysis with a Cox model, possibly refined with interaction terms, could have clarified the impact of candidate factors on prognosis, eventually leading to a 'final' model. Apart from considering that in applied clinical research the aim is at defining interpretable and applicable classification rules, it is well-known that even very strong prognostic factors may provide a limited predictive ability.[8] In addition, the need to define a parsimonious prognostic classification often produces alternatives that have very similar performance. Model development and evaluation of predictive performance seem therefore to respond to different questions. In our opinion, as establishing a prognostic classification is a complicated multi-step process, both of them should be rigorously investigated, especially when the crucial stage of validation is prevented by the peculiarities of the clinical context.

Conflict of interest statement

None declared.

Acknowledgements

Sources of support: this research was partially supported by AIRC (Associazione Italiana per la Ricerca sul Cancro) and by Fondazione Tettamanti for Research on Childhood Leukaemia.

We also acknowledge the Interfant-99 Study Group and its Chairman, Professor Rob Pieters, for making the data available. We had no writing assistance and no extra funding for this paper. No ethics committee approval was needed.

REFERENCES

1. Harrell Jr FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and

adequacy, and measuring and reducing errors. *Stat Med* 1996;**15**:361–87.

2. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. *Stat Med* 1999;**18**:2529–45.
3. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med* 2004;**23**:723–48.
4. Mackillop WJ, Quirt CF. Measuring the accuracy of prognostic judgments in oncology. *J Clin Epidemiol* 1997;**50**:21–9.
5. Pieters R, Schrappe M, De Lorenzo P, et al. A treatment protocol for infants younger than 1 year with acute lymphoblastic leukaemia (Interfant-99): an observational study and a multicenter randomised trial. *Lancet* 2007;**370**:240–50.
6. Antolini L, Boracchi P, Biganzoli E. A time-dependent discrimination index for survival data. *Stat Med* 2005;**24**:3927–44.
7. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. ISBN: 3-900051-07-0. <http://www.R-project.org>.
8. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic or screening marker. *Am J Epidemiol* 2004;**159**:882–90.