CAN WE INTERPRET A HAZARD RATIO AS A RATIO OF MEDIAN SURVIVALS?

Jordi Cortés[1], MSc

José A González[1], PhD

Michael J Campbell[2], PhD

Erik Cobo[1], PhD

[1]Department of statistics and operations research, Barcelona Tech, Spain

[2] School of Health and Related Research, The University of Sheffield, UK.

Corresponding autor (and guarantor): Jordi Cortés jordi.cortes-martinez@upc.edu tel. +34 934015868 – fax +34 934015855

Universitat Politècnica de Catalunya, BarcelonaTech

Jordi Girona street, 1-3

C5 – 2nd floor

08034 Barcelona

Spain

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Running Head*:* Median Ratio vs. Hazard Ratio

Word count*: 1500 (without boxes)*

Number of figures: 5

Number of tables: 0

Abstract

Objective: To evaluate the empirical concordance between the Hazard Ratio (HR) and the Median Ratio (MR) in survival cancer studies.

Study Design and Setting: We selected all cancer survival papers from the New England Journal of Medicine published between 2000 and 2010. The qualitative concordance was estimated by the proportion of measured pairs where the treatment effects for the MR and the HR are in the same direction. The quantitative concordance was assessed through: (1) the mean difference between the logarithms of the measures; (2) the Lin coefficient; and (3) the Bland-Altman plot.

Results: We retrieved 106 measure pairs (HR-MR) corresponding to 54 papers. Concordance was high, both at the qualitative (99/106, 93.4%) and quantitative level (mean MR/HR ratio 1.01, 95% confidence interval: 0.97 to 1.04). However the 95% Bland Altman discordance limits indicate that the MR can be up to 50% higher or 50% lower than the HR.

Conclusion: The average concordance allows trialists to approximate HR from MR to determine sample size. However, the discordance limits are too great to consider that both measures are interchangeable. The actual policy to report only HR is not enough. Our results emphasize the need to attach descriptive survival measures to the HR.

Key words: Hazard Ratio; Median Ratio; Concordance; Survival; Clinical Trial.

Running title: Median Ratio versus Hazard Ratio.

Word count (abstract): 201 (taking into account the headings).

Box 1: What is new?

Key findings:

- ✓ In cancer survival studies, the mean HR/MR was 1.01 with a 95% confidence interval from 0.97 to 1.04.
- ✓ In cancer survival studies, the discordance limits indicates that the median ratio can be up to 50% higher or lower than the hazard ratio.

What this adds to what is known:

- ✓ There is no interchangeability between HR and MR in individual cancer studies.
- ✓ The inverse of the Hazard ratio cannot be interpreted as a measure of the expected life gained.
- ✓ The average concordance provides empirical evidence to approximate HR from MR in sample size calculations.

What is the implication, what should change now:

- ✓ The discordance at individual trial results implies the need to consider additional descriptive measures to help interpret the benefit from HR in survival papers.
- ✓ The actual policy to report only HR is not enough to convey such important information.

Box 2: Measures explanation

Example: a clinician desires to explain to a patient the life-extending effect of an additional treatment (T) versus a clinical guide reference (R).

Median: this can be estimated as the time on the x-axis of the Kaplan Meier curve corresponding to the 50% point on the y-axis (probability of survival). For low mortality rates the time axis might be too short to observe the 50% point. However, depending on the censoring, it may be observed even when fewer than 50% of the patients die, and it does not require that all events are observed.

Median Ratio (MR): the quotient of medians in the Treatment (T) compared to the Control (C). If the median is 8 months in the T group and 4 in C, then the median ratio is 2. The clinician may say "treatment will double your survival time".

Hazard function: the risk[1] of dying at a point in time among patients who survived up to that point[2].

Hazard rate: The frequency of death over a period of time[3]. For example, in group T, 1 out of 12 patients dies every month; while in R, 1 out of 6.

Hazard (rate) Ratio (HR): the quotient of the hazards from both groups[4]. The HR is (1/12) / (1/6) = 1/2. The clinician can say "treatment will halve your probability of death at any moment in time".

Calculation in medical studies with censored times: Median survival is usually calculated as the smallest time for which the Kaplan-Meier survival function is less than or equal to 0.5. The Cox proportional hazards model is the usual way to estimate the HR when some patients are still alive at the end of their follow-up.

Graphical analogy: Looking at the Kaplan-Meier curve, MR assesses the horizontal-axis expansion, that is, life extension or time gained; but HR looks at the vertical-axis decrement, survival descent or death speed [1].

HR advantage: as the same ratio applies to any survival time when the Cox assumption holds, any kind of patient (i.e., moderate, mild, severe) has a common treatment effect measure[5].

Footnotes:

[1] Over time, this risk can be constant, increasing, decreasing, or have any form.

[2] Mathematically: $\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t+\Delta t | T \geq t)}{\Delta t}$

[3] Hazard rate refers to a constant hazard function over the target period.

[4] Cox proportionality assumption states that, independently of the form of the hazard function, the ratio of risks is just a single common value, the HR, for any follow-up time point.

[5] MR and the differences of medians will equal other percentile ratios or differences only under particular parametric assumptions (such as exponential and normal distributions of the events, respectively)

Box 3: Example of suggested wording in the result description

The International Adjuvant Lung Cancer Trial Collaborative Group [2] compared the overall survival effect of Cisplatin in patients with non-small-cell lung cancer. Cisplatin (see figure 1) reduced the death rate by 14% [HR was 0.86, CI 95%, from 0.76 to 0.98]. Through visual inspection of the KM curve, the median increased from 45 to 52 months. The survival time of people in the worst 25% for survival moved from 17 to 19 months, approximately, implying 2 extra months to patients with poorer prognosis. Length of followup was insufficient to provide an estimate of treatment effect for patients with best outcome (best 25%).

Introduction

In cancer survival studies, the Hazard Ratio (HR) is commonly employed to summarize treatment effect size –mainly because it can be easily obtained from the Cox proportional model with good statistical properties, such as efficiency (smaller sample size, more precise confidence intervals, more power) [3]. Unfortunately, the HR is frequently misunderstood [4-6]; and there is empirical evidence that risk ratio measures poorly influence a layperson's decisions [7]. Median survival is a simple summary statistic, especially when mortality is high, and we can use the ratio of medians (MR) to compare two treatment groups. We believe that the MR is a much easier concept to grasp and to communicate to patients. Theoretically, HR and MR will be equivalent if the event rate is constant over time (an exponential model) but, in general, they will disagree if the rate increases or decreases [8].

In some cases, HR and MR may concur, as in the TARGET study [9]: HR survival (Sorafenib vs. placebo) was 0.51 (figure 2), meaning that the hazard rate was around double in the control group. As the median survival for placebo was 2.8 months and for Sorafenib 5.5 months, MR was also 0.51 (swapping the treatments accordingly to make both ratios comparable), showing a perfect concordance. Thus our belief is that, from a patient point of view, in this example it would be more informative to know that median survival is doubled on Sorafenib (or that it is increased by almost 3 months), than to know that the instant probability of death is 49% lower.

However in many other cases, the HR and MR don't always concur.

An important question is whether or not a clinician would be able to recommend an intervention from the results of a clinical trial by deriving the MR from the reported HR.

Furthermore, when planning a study, it may be easier for a trialist to predict an increase in survival, (e.g., 6 months) than to give the HR; and yet the HR is required for sample size determination [10].

Thus, our objective is to evaluate the empirical numerical concordance between the MR and the HR in order to validate HR as an MR indicator in clinical interpretation. A further objective is to validate the MR as an HR estimate for sample size calculations.

Methods

As we generally need a high event rate to observe a 50% survival, we defined our target population as cancer clinical trials. We used the New England Journal of Medicine (NEJM) search engine for *Research* papers published between 2000 and 2010 with the following search terms: the word *cancer* and at least one of the words, *cox* or *hazard.* We further selected Randomized Clinical Trials reporting the HR and the median time of Overall Survival (OS) and/or Progression Free Survival (PFS) outcomes.

The collected outcomes were: title; publication year; outcome category (primary/secondary); outcome type (OS/PFS); HR and its 95% confidence interval (CI); survival median for each group; sample size; and events in each group.

Unreported median survivals were estimated visually from the Kaplan-Meier (KM) curves when they crossed the 50^{th} percentile.

The average concordance was assessed with a paired t-test, using the mean difference between the logarithm of measures and weighted by the inverse standard error (SE) of the log(HR). As some curves came from repeated outcomes obtained in the same paper, the SE of the estimation was corrected by the Intraclass Correlation Coefficient (ICC).

The qualitative concordance was estimated as the proportion of measure pairs where the MR and the HR are in the same direction.

The quantitative concordance was assessed through a Lin coefficient [11] and the Bland-Altman plot [12] of the logarithms of the study measures. Also, subgroup analyses were performed based on outcome type (OS/PFS); publication year (≤2005/>2005); statistically significant HR; and medians reported versus visually estimated.

Results

Paper flow and description

The literature search retrieved 348 studies, but most of the medians were either not achieved or the KM plots were not presented (244, 70.1%). We further excluded 50 papers mainly because HR or HR standard error were unreported (34, 9.8%); or studies were not properly randomized (9, 2.6%); or due to other reasons (7, 2.0%). The remaining 54 articles provided 106 HR-MR pairs. The MR was recovered directly from 21 (19.8%) KM curves.

The median number of papers per year was 4 (interquartile range [IQR]: 2 to 6.5). Papers provided the following pairs of measures: 1 (22%), 2 (67%) and 3 or more (11%). The mean sample size was 571 (range, 81 to 1867)

and the mean number of events was 401 (73 to 1095). From the 106 pairs, both HR and MR geometric means were 0.74.

Qualitative concordance

In 99 of 106 pairs of measures (93.4%, 95% confidence interval [CI]: 86.4 to 97.1), HR and MR are in the same direction. Of the remaining 7 pairs, there are 4 where MR equals to 1; and 3 where the measures point in opposite directions. Figure 3 contains the Kaplan Meier (KM) survival curves for the 3 discordant studies [13-15]. The common feature in them is that the curves cross each other at some point.

Quantitative concordance

The mean MR/HR ratio was 1.007 (95% CI: 0.953 to 1.064). The 95% Bland-Altman limits of discordance were 3/2 and 2/3, and were symmetric; that is, either MR was 50% higher than HR or HR was 50% higher than MR (figure 4). Lin's concordance correlation coefficient was 0.77 (95% CI: 0.69 to 0.85). Trials based on larger sample sizes and interventions with smaller effects seem to provide higher concordance between MR and HR.

Figure 5 reproduces KM curves of 3 studies [15-17] which presented large numerical discrepancies, despite their higher sample size (highlighted in figure 4 with red circles); the one on the right also presented qualitative discordance. The rationale behind providing the HR as a single intervention measure does not hold in these three studies. The hazard ratios are not constant, i.e., the effect cannot be summarized in one single value.

Discussion

Interpretation

In absence of direct estimators of HR target effect, the almost perfect average concordance allows trialists to approximate target HR from MR in sample size calculations. However, the weak individual concordance prevents patients, physicians, meta-analysts, and health managers from obtaining an estimate of the MR from the HR. In order to derive an MR estimate from HR while taking discordance into account, we should consider ranges from 2/3 and 3/2 of the HR—leaving aside the fact that there is additional uncertainty in the HR estimation. For example, if the HR is 3/4 and MR is not available, the expected MR is 4/3, implying an extended median survival of 33%. But, in fact, it can be any number between 0.89 (=4/3*2/3) and 2 (4/3*3/2). In the first case 0.89 median survival decreases by about 10%; in the second (2), median survival doubles. In other words, if median survival in the control arm was 10 months, discordance between HR and MR can imply an effect ranging anywhere from a one-month loss of life up to a 10-month gain in life. This imprecision due to discordance is too high to derive the MR from the HR.

Five papers (see figures 3 and 5) showed large disagreements between HR and MR: these discordances are fewer than those found (25/128, 20%) by Michiels et al [18]. The finding highlights that sometimes the proportional assumption does not hold: HR cannot convey a nonexistent constant effect through time, and so it cannot be used to summarize an intervention effect. Furthermore, it emphasizes the importance of complementing HR with supplementary information.

Limitations

Both internal and external validity are compromised; the former, because of the high proportion of papers that either did not provide the medians — despite existing recommendations [19]— or they were not reached; and the latter because, for convenience, our data came from a single, highly esteemed and highly rated medical journal [20]. Furthermore, in order to improve our chances of observing the median in both groups, we also restricted our study to cancer trials. If concordance is not good enough in those narrower conditions, it could be expected to be even worse in more general situations.

Implications

As Schwartz et al [7] summarize, "ratio measures without the underlying absolute risks often exaggerate readers' perceptions of benefit or harm". And so, the actual policy to report only HR is not enough. When studying dichotomy outcomes, reporting risk differences in order to complement relative risks has been accepted for clinical trials [21]. But when studying "time to an event", the HR can often be assumed to summarize the effect size in a unique measure that would apply to any patient (mild, moderate or severe). In order to facilitate decisions made by patients, their physicians and healthcare providers, we advise that authors and editors do the following: along with the report of HR and its 95% CI, include a graphical

interpretation of life extension in the Kaplan-Meier curve (e.g., see the lines indicating the median survival groups in Figure 1).

Acknowledgements: to Matthew Elmore for English editing and to anonymous reviewers for their valuable suggestions and clever proposals.

Conflict of Interest Statement: The Authors declare that there is no conflict of interest.

Contributorship: This paper is mainly based on the JC Master Thesis. EC had the original idea; MC upgraded the objectives; JAG and EC designed the study; JC collected the data; JC and JAG analyzed the data; and all authors interpreted the results, contributed to and approved the paper.

References

1. Lau EW, Ng GA. Visual illusions created by survival curves and the need to avoid potential misinterpretation. Medical Decision Making 2002;22:238–44.
2. The International Adjuvant Lung Cancer Trial Collaborative Group. Cisplatin-Based Adjuvant Chemotherapy in Patients with Completely Resected Non–Small-Cell Lung Cancer. N Engl J Med 2004;350:351-360
3. Kleinbaum DG, Klein M. Survival Analysis: A Self-learning Approach. 2nd ed. New York: Springer, 2005.
4. Spruance SL, Reid JE, Grace M, Samore M. Hazard Ratio in Clinical Trials. Antimicrob Agents Chemother 2004;48:2782-92.
5. Hernán M. The hazards of hazard ratios. Epidemiology 2010;21:13–5.
6. Case LD, Kimmick G, Pasketta ED, Lohmana K, Tuckerb R. Interpreting Measures of Treatment Effect in Cancer Clinical Trials. Oncologist 2002;7:181-7.
7. Schwartz LM, Woloshin S, Dvorin EL, Welch HG. Ratio measures in leading medical journals: structured review of accessibility of underlying absolute risks. BMJ 2006;333: 1248.
8. Carroll KJ. On the use and utility of the Weibull model in the analysis of survival data. Control Clin Trials 2003;24:682–701.
9. Escudier B, Eisen T, Stadler WM, et al. Sorafenib in advanced clear-cell renal-cell carcinoma. N Engl J Med 2007;356:125-34.
10. Machin D, Campbell MJ, Tan SB, Tan SH. *Sample size tables for clinical studies*. 3rd ed. Chichester: Wiley-Blackwell, 2008.
11. King TS, Chinchilli VM. Robust estimators of the concordance correlation coefficient. J Biopharma Statist 2001;11:83–105.

12. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;i:307-10.
13. Karapetis CS, Khambata-Ford S, Jonker DJ, et al. K-ras mutations and benefit from cetuximab in advanced colorectal cancer. N Engl J Med 2008;359:1757-65.
14. Goodwin PJ, Leszcz M, Ennis M, et al. The effect of group psychosocial support on survival in metastatic breast cancer. N Eng J Med 2001;345: 1719–26.
15. Mok TS, Wu YL, Thongprasert S, et al. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. N Engl J Med 2009;361:947-57.
16. Jonker DJ, O'Callaghan CJ, Karapetis CS, et al. Cetuximab for the treatment of colorectal cancer. N Engl J Med 2007;357:2040-48.
17. Löwenberg B, Ossenkoppele GJ, van Putten W, et al. High-dose daunorubicin in older patients with acute myeloid leukemia. N Engl J Med 2009;361:1235–48.
18. Michiels S, Piedbois P, Burdett S, Syz N, Stewart L, Pignon JP. Meta-analysis when only the median survival times are known: A comparison with individual patient results. Int J Technol Assess Health Care 2005;21:119–25.
19. Altman DG, De Stavola BL, Love SB, Stepniewska KA. Review of survival analyses published in cancer journals. Br J Cancer 1995;72:511-18.
20. Saha S, Saint S, Christakis DA. Impact factor: a valid measure of journal quality? J Med Libr Assoc 2003;91:41–6.
21. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trial. BMJ 2010;340:c869.

Figure 1. Figure 1A from paper *Cisplatin-Based Adjuvant Chemotherapy in Patients with Completely Resected Non–Small-Cell Lung Cancer* [2]. Dashed lines are added to the graphic in order to distinguish the different survival percentiles in both groups: 75 (green), 50 (orange) and 25% (red). The latter were not estimable because the time of follow-up was not long enough.

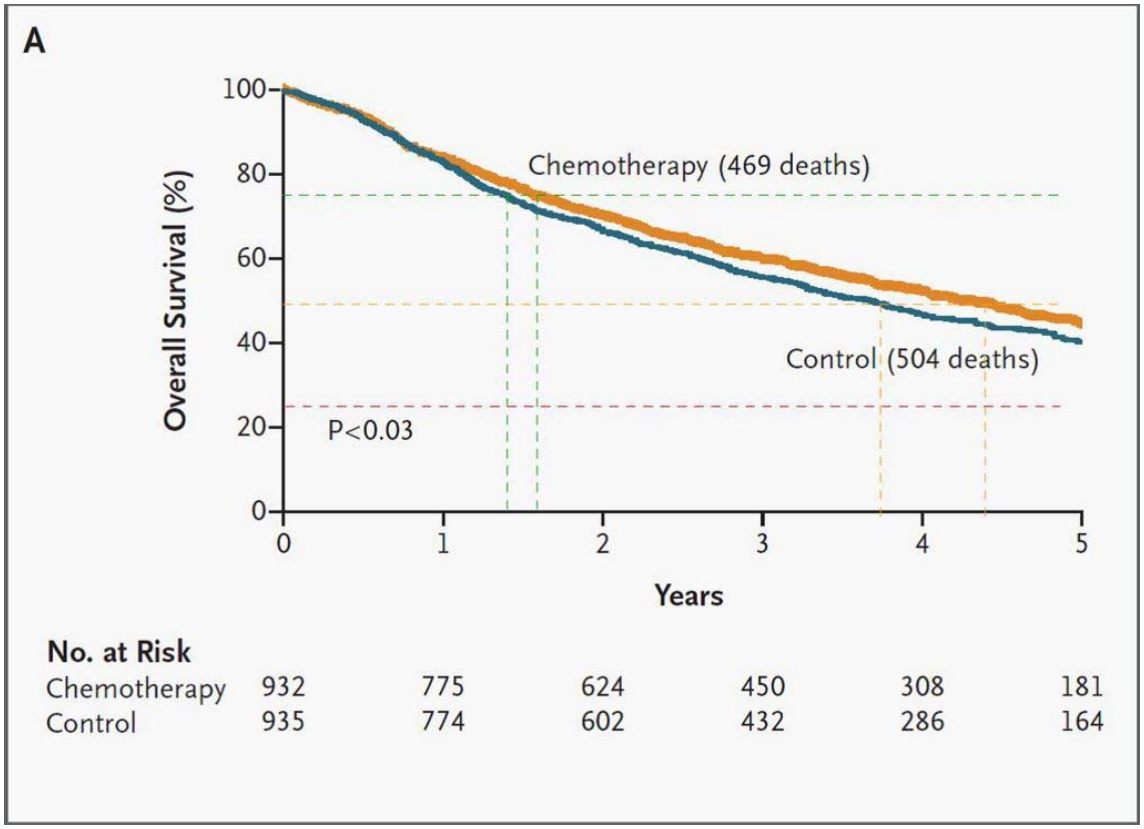

Figure 2. Example of Kaplan-Meier curves with equal HR and MR.

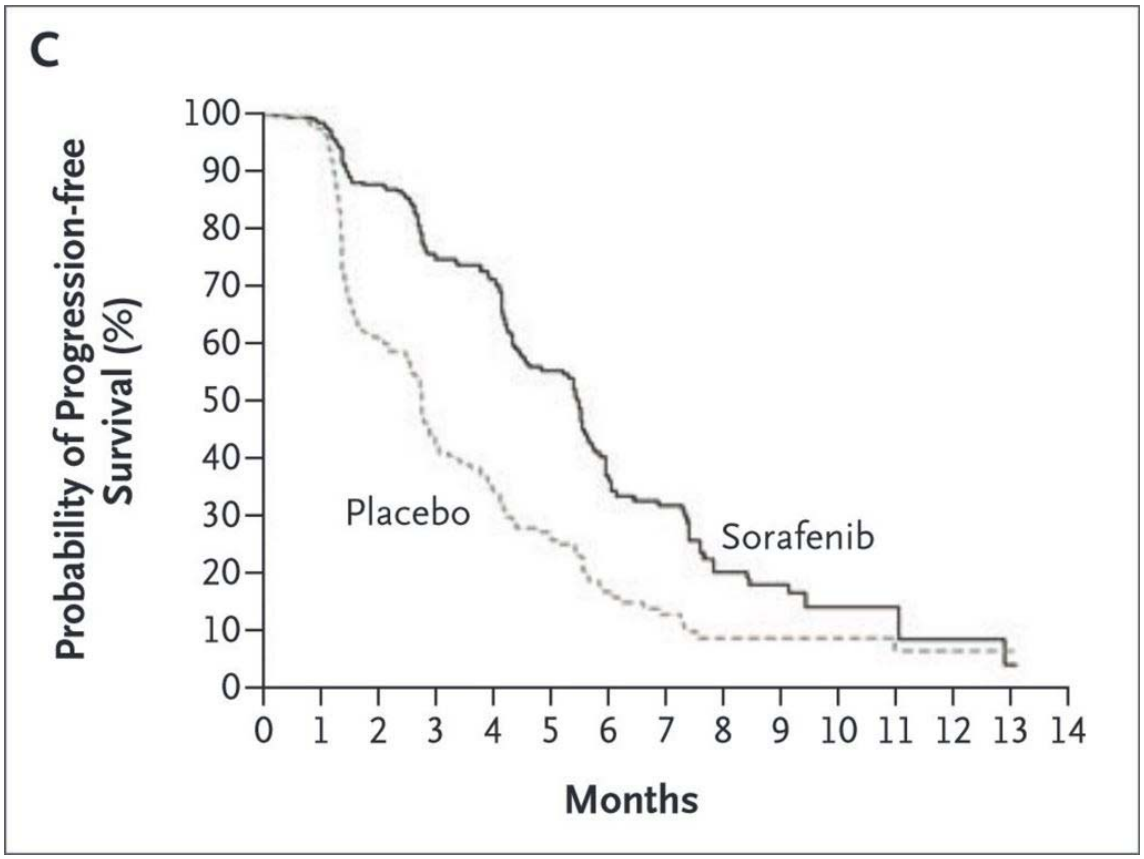

Figure 3. KM survival curves for 3 out of 4 studies with discordant directions in the treatment effect. From left to right: HRs of 0.98, 1.06 and 0.74 favor Cetuximab, control and Getifinib; and MRs of 1.02, 0.98 and 1.02 favor supportive care, psychosocial support and carboplatin-paclitaxel, respectively.

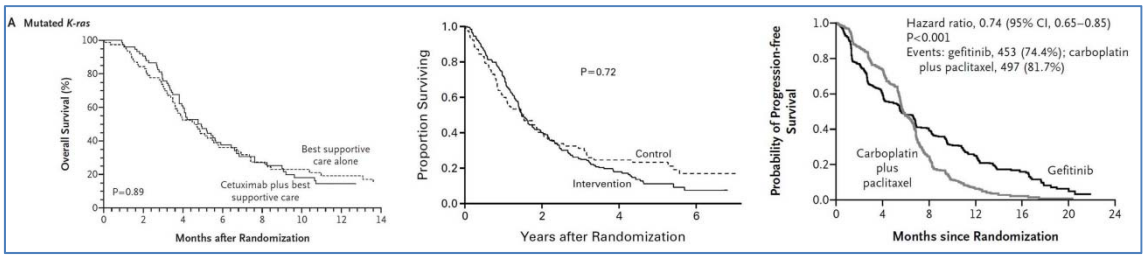

Figure 4. Left: scatter plot of the log(HR) vs log(MR). The size of the dots is proportional to the inverse of the SE of Log (HR). The dashed line represents the line of equality. Right: Bland-Altman concordance plot discrepancies as function of averages. The blue lines represent the weighted mean of all discrepancies and this average, ± 2 weighted SD. Red dots are pairs belonging to large studies and with higher quantitative discordance.

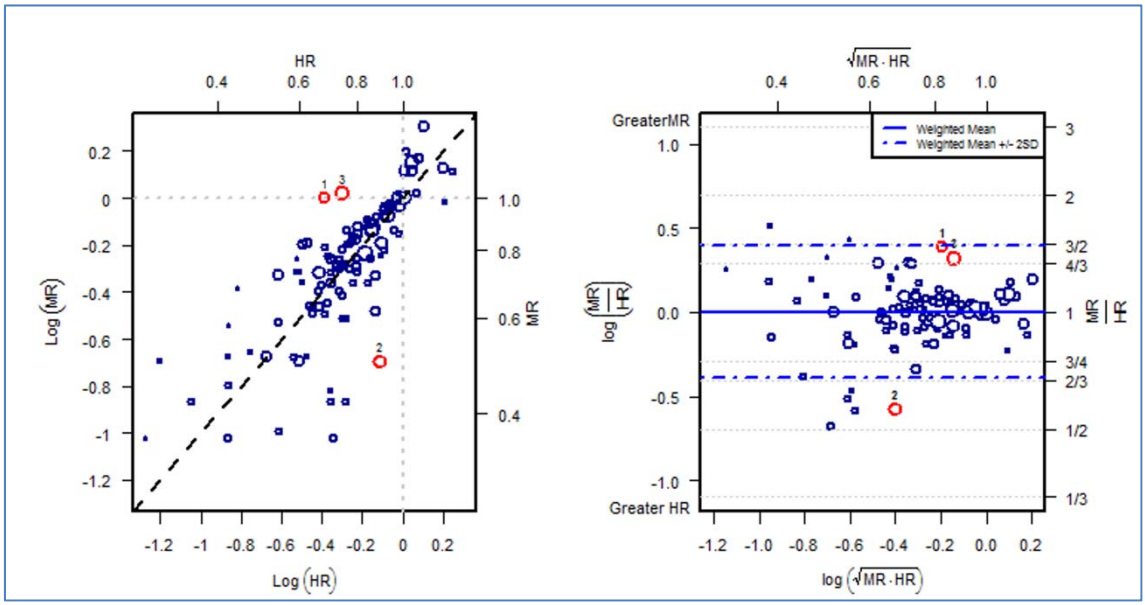

Figure 5. Curves corresponding to measures with great discordance between MR and HR in large studies. Each one has some feature discouraging MR or HR as summary measures: (1) similar behavior up to a certain point and subsequent bifurcation (HR: 0.68, MR: 1.00); (2) large number of patients with event in instant 0 (HR: 0.89, MR: 0.50); and (3) crossed curves (HR: 0.74, MR: 1.02).