In reviewing data from a clinical trial, the statistician can analyze the data from the total population of study subjects, as well as from specific subgroups. These subgroups typically include subjects between 18 and 65 years of age versus subjects over 65 years, subjects previously treated with chemotherapy versus those who are treatment-naive, and subjects with wild-type gene, for example, epidermal growth factor gene versus those with a mutated gene. If there is reason to suspect that expression of a given gene is relevant to response to a study drug, or that a mutation in the gene is relevant to response, then subgroup analysis can be performed when the study is completed, and when all of the data are collected. Dr Harvey Motulsky (18) has emphasized that good methodology in study design requires the definition of subgroups before initiating the clinical trial, and not after the clinical trial when the data are available, and that defining subgroups after the clinical trial can raise the issue of "data mining." Data mining has been described as, "data dredging or fishing and ... the process of trawling through data in the hope of identifying patterns" (19).

d. Hazard Ratio

The hazard ratio is the ratio of (chance of an event occurring in the treatment arm)/(chance of an event occurring in the control arm) (20). The HR has also been defined as, the ratio of (risk of outcome in one group)/(risk of outcome in another group), occurring at a given interval of time (21). In the situation where the hazard for an outcome is exactly twice in Group A than in Group B, the value of the hazard ratio can be either 2.0 or 0.5. The result of the calculation (whether HR = 2.0 or 0.5) depends on whether the investigator chooses to calculate the ratio of hazards for (Group A)/(Group B) or, alternatively, to calculate the ratio of hazards for (Group B)/(Group A) (22,23).

The term "hazard" refers to the probability that an individual, under observation in a clinical trial at time t, has an event at that time (24). It represents the instantaneous event rate for an individual who has already survived to the time "t."

The two arms of a clinical trial can be compared by way of the hazard ratio and the P value. The following serves as a starting point for defining hazard ratio and P value, as it applies to two curves in a Kaplan–Meier plot. The hazard ratio is a measure of the magnitude of the difference between the two curves in the Kaplan–Meier plot, while the P value measures the statistical significance of this difference. These two definitions serve only as starting points for our present goal in arriving at accurate, correct definitions. The following are the correct definitions. The numerical value of the hazard ratio expresses the relative hazard reduction achieved by the study drug compared to

[18]Motulsky H. E-mail of May 9, 2011.

[19]Hand DJ. Data mining: statistics and more? The American Statistician. 1998;52:112–18.

[20]Duerden M. What are hazard ratios? What is ... ? series. Hayward Medical Communications, Hayward Group, Ltd., 2009. p. 8.

[21]Dawson B, Trapp RG. Basic and clinical biostatistics. 4th ed. New York, NY: Lange Medical Books; 2004, p. 407.

[22]Machin D, Cheung YB. Survival analysis: a practical approach. 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc.; 2006. p. 62.

[23]Crowley J. Handbook of statistics in clinical oncology. New York, NY: Marcel Dekker; 2001. p. 541.

[24]Duerden M. What are hazard ratios? What is ... ? series. Hayward Medical Communications, Hayward Group, Ltd. 2009. p. 8.