subject may introduce bias into the calculations and analysis. Please consider the following hypothetical. In this hypothetical, an experimental drug is ineffective against breast cancer, except for a minority of people in the general population with a rare mutation in the epidermal growth factor gene. Now, please imagine that the health of most of the study subjects deteriorates to the point where they can no longer come to the clinic, and where the investigator decides to censor data from the subjects. In this hypothetical, only a fraction of the subjects – perhaps 5% of the total subjects enrolled in the study – having the rare mutation will feel good enough to come to the clinic for more treatment. The result of the censoring will be that the drug is discovered to be dramatically effective against breast cancer. But this will be a misleading and artifactual result because, in fact, the drug is only effective in 5% of the subjects.

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 genes, 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 hazard ratio 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

[18] Motulsky H. E-mail of May 9, 2011.
[19] Hand DJ. Data mining: statistics and more? *Am. Stat*. 1998;52:112–118.
[20] Duerden M. *What are Hazard Ratios? What is…? Series*. Hayward Medical Communications, Hayward Group, Ltd.; 2009;8.
[21] Dawson B, Trapp RG. *Basic and Clinical Biostatistics*. 4th ed. New York, NY: Lange Medical Books; 2004;407.