STATISTICS AND BIOMEDICAL RESEARCH

Article Review by Orgah Adikwu Emmanuel, Nigeria
(M.Sc., Ph.D., in Clinical Research Student of Texila American University)
Email: - emmaorgah@yahoo.com

INTRODUCTION

Biostatistics is the application of statistical techniques to biological data obtained prospectively and/or retrospectively. Statistics plays critical analytical role in biomedical research. It is the bases for building clear inference from the data collected in a biomedical evaluation and without which it would be impossible to declare an outcome from any clinical trial. This critical role of biostatistics in biomedical research was noted by Cadarso-Suárez, and González-Manteiga, (2007), who stated that "the discipline of biostatistics is nowadays a fundamental scientific component of biomedical, public health and health services research" and pointed out traditional and emerging areas of application as "clinical trials research, observational studies, physiology, imaging, and genomics".

At the same time, misuse of biostatistics has resulted in several misleading outcomes and several workers have progressively noted the many statistical errors and shortcomings found in a large number of biomedical publications (Porter, 1999; Cooper, et al., 2002; García-Berthou and Alcaraz 2004; Strasak, et al., 2007; Ercan, et al., 2007; Thiese, et al., 2015). Ercan, et al., (2007) specifically notes that this observations cuts across "every stage of a medical research related to data analysis; design of the experiment, data collection and pre-processing, analysis method and implementation, and interpretation". Similarly, Thiese, et al., (2015), points to data abuses such as "incorrect application of statistical tests, lack of transparency and disclosure about decisions that are made, incomplete or incorrect multivariate model building, or exclusion of outliers".

The role of statistics in medical research starts at the planning stage of a clinical trial or laboratory experiment to establish the design and size of an experiment that will ensure a good prospect of detecting effects of clinical or scientific interest. Statistics is again used during the analysis of data (sample data) to make inferences valid in a wider population. Specifically, statistics has two roles in laboratory experiments and clinical trials.

1. It ensures sound experimental design and optimal use of resources at the planning stage. Adequate statistical analysis of experimental data is usually only possible if the design is statistically sound.

2. It is essential for proper analysis of results or research outcomes. Assertions based on experimental data need to be backed by a relevant statistical analysis. Even for pilot or for retrospective studies some statistical analysis is usually needed.

Statistics therefore plays key roles in all phases of the research project starting from the design stage and continuing through the monitoring, data collection, data analysis and interpretation of the results. Sprent, (2003); Mandrekar and Mandrekar, (2009), highlighted these important roles that statistics plays in the field of biomedical research. A clear understanding of the statistical approach as it relates to the study hypothesis, reported results and interpretation is vital for the scientific integrity and interpretation of the study findings in the general medical community.

In simple situations computation of simple quantities such as P-values, confidence intervals, standard deviations, standard errors or application of some standard parametric or nonparametric tests may suffice. Despite their wide use even these simple notions are sometimes misunderstood or misinterpreted by research workers in other disciplines who have only a limited knowledge of statistics. More sophisticated research projects often need advanced statistical methods including the formulation and testing of mathematical models to make relevant inferences from observed data. Such advanced methods should only be applied with a clear understanding both of their purposes and the implication of any conclusions based upon their use. Statistics provides rational measures to reflect on the degree of uncertainty associated with assertions drawn from a database. At a more sophisticated level it provides indicators for how well data conform to some specified mathematical model, eg, tests goodness of fit to the model, and when appropriate provides estimates of certain constants or parameters in a model.

Overall, Statistics plays an essential part at all stages of the clinical trial, from planning, through conduct and interim analysis, to final analysis and reporting. Statistical methods will be typically utilised in devising the randomization schedules, advise on sample size, specify criteria for measuring treatment differences, and analyze response rates. Close collaboration between statisticians, whether professionals in that field or medical research workers with a sound statistical background, and other members of a research team is needed to ensure a seamless integration of the statistical elements into the reporting and discussion of research outcomes. At the end of the study, the Independent Data Monitoring Committee requires statistically defined data for final processing and interpretation.

THE ROLE OF STATISTICS AT THE PLANNING AND DATA GATHERING STAGE

The first step in a data analysis is the selection criteria used to determine the cases (i.e., data elements) to be included in the analysis to ensure that there is minimal to no selection bias, and that results arising from the analysis of this data is reproducible, i.e., the selection criteria is not ad-hoc. This is obvious if we consider as pointed out by Thiese, et al., (2015) that errors in statistical “data collection and analyses follow directly from study design”. While the selection

of cases is largely driven by the scientific question at hand, it also depends on the subject population, the variables to be explored, adequacy of available follow-up information for the endpoints, and attributes of the missing data values (i.e., missing by design, missing at random etc.). This is a critical step in evaluating the scientific merit of the research study.

Once the analysis data set is identified and set up, the next step is to explore and describe more thoroughly the endpoints and the explanatory (or independent or prognostic) variables. The guidance provided by The STARD Statement, The STROBE Statement and The CONSORT Statement, (Bossuyt, et al., 2003; Von Elm, et al., 2007 and Moher, et al., 2001 respectively), have luckily come to the rescue in guiding researchers to improve the quality of reporting study methods and results of biomedical research. The outcome (i.e., endpoint) variables typically fall into one of the three classes:

1. Categorical, which is further classified into:
    a. Binary or two categories (for example, limited vs. extensive stage disease; male vs. female etc.),
    b. Nominal or multiple categories with no specific order (for example, blood group type: A, B, O, AB etc.), and
    c. Ordinal or multiple ordered categories (for example, performance status: 0 vs. 1 vs. 2; Likert type scales: strongly agree, agree, neutral, disagree), or
2. Continuous (for example, age, white blood cell counts etc.), or
3. Time-to-event (for example, overall survival, time to any recurrence etc.).

The definition of the endpoint, percentage of data completeness, and adequacy of follow-up information are all critical elements that help assess the accuracy and interpretability of the results as it relates to the endpoint. For example, in the case of time to event endpoints, the following issues need careful attention:

- The starting (date of diagnosis, date of randomization etc.) and the ending time point (date of death, date of recurrence, date of last follow-up) used for defining the specific endpoint,
- Information regarding the uniformity (and length) of follow-up (for example, all patients are followed for a minimum of 2 years),
- The loss to follow-up/drop out rates (censors) and the event rate for assessing data completeness and determination of the number of covariates that can be explored, and

- The presence of competing risks (for example, both death and distant recurrence are competing events when the event of interest is local recurrence) and/or recurrent events (for example, recurrent heart attacks in a coronary heart disease study).

For the explanatory variables, running simple descriptive and graphical summaries to identify obvious deviations such as outliers, sparse data within certain groups, and/or questionable data points is usually recommended. At this time, decisions regarding collapsing categories and/or categorization of continuous covariates are explored, which are typically based on

- Distribution of the data,
- Underlying biologic or clinical rationale, and
- Ease of interpretation.

Statistics serves a purpose for identifying optimal cut-point(s) for categorizing continuous covariate as shown by Mazumdar and Glassman, (2000). These techniques which emanates as data and outcome handling procedures include:

- ✓ mean, and median
- ✓ the minimum p-value approach, and two-fold cross validation

Further, statistics is essential in mitigating wrongful categorization of continuous covariate or potential loss of information and incorrect assumption of the distribution of the data post categorization as documented by Altman et al. (1994), Abdolell et al. (2002). Chansky et al., (2009), demonstrated the use of a data-driven approach for categorizing a continuous covariate and discussed the impact of this categorization on the model assumptions and the estimates.

THE ROLE OF STATISTICS AT THE ANALYSIS STAGE

The second and most critical step in biomedical research is the data analysis stage which is useful to draw conclusions from the data. The general analytical approach at this stage of the study typically includes four elements. Namely:

1. Definition of the training and validation data sets, if applicable;
2. Detailed description of all statistical procedures utilized to address the research questions, including the testing and model building framework;
3. Clearly describing the pathway from univariable to multivariable analyses;
4. Setting threshold for declaring statistical/clinical significance a priori for the main effects, interactions and subgroup analyses keeping in mind the multiple comparisons issue.

According to Harrell (2001), training data set (i.e., developmental data set) validation data (i.e., test data set) are used to discover and validate statistical process in order to gauge the predictive accuracy and performance of the original analysis in practice. The common approaches include:

- ✓ Cross validation methods, which refer to repeated partitioning of one large data set into training and test sets, and
- ✓ Use of two independent data sets (with similar data attributes), one utilized only for development, and another used exclusively for validation.

The statistical techniques used to analyze the data are closely tied to the nature of the data and the research hypothesis. Statistical analysis utilises terms and concepts such as normal distribution, binomial distribution, non-parametric methods, analysis of variance, long-tail distribution, exponential distribution, correlation, regression for different data treatment. Tests such as the t-test, chi-squared test, Wilcoxon signed-rank test, Wilcoxon rank-sum test and Mann-Whitney test and others such as the log-rank test are also employed to draw inferences about the overall outcome of the study.

Terms such as standard error, standard deviation, range, interquartile range are also used to describe spread in the data field. The Normal distribution is central to much statistical analysis of measurement data and this is based on sound mathematical modelling and assumptions. The normal distribution is characterised by its mean and standard deviation. Standard error or specifically the standard error of the mean is a measure of how accurately the sample mean estimates the population mean.

A univariate analysis helps to assess the strength of association of the explanatory variable on the outcome of interest by itself, independent of other variables. Summary measures that pertain to a single variable are called univariate statistics and a summary measure is a compact description of the data that conveys information about the distribution of a variable quickly.

Univariate analysis helps to gauge the ability of the covariate to influence outcome on its own. It is well understood in the statistical and clinical literature that in reality the impact (effect size and significance) of an explanatory variable on outcome when explored by itself is susceptible to change when explored in the presence of other explanatory variables. It is thus important to assess the possible independent effect of a covariate on outcome in a multivariate model.

A histogram is a summary measure for a single variable which plots observed values of a variable on the X-axis versus the relative frequency of these values on the Y-axis. An example is the systolic blood pressures of research participants described by a histogram which conveys a sense of how systolic blood pressures are distributed in the entire data set.

According to Mandrekar and Mandrekar, (2009), there are a variety of ways to move from a univariate to a multivariate model. Some of these include:

- Explore only those that are statistically and/or clinically significant in a univariable analysis
- Include all previously known important covariates as the base model and then build a full model adding new covariates in a step wise fashion
- Explore all covariates explored in the univariable analysis in a multivariable analysis regardless of their significance in the univariable setting
- Build a multivariable model using the pool of all covariates through a selection technique

The statistical significance for these analyses is usually determined by the number of models and factors explored, the sample size/number of events and the clinical relevance. A multivariable model also includes the exploration of two-way interaction effects to assess if the effect of a variable on outcome differs depending on the level of another variable. Specifically, two variables are said to interact if a particular combination of the variables leads to results that would not be anticipated on the basis of the main effects of those variables. Also known as bivariate descriptive statistics, it is used to describe the joint relationship between two variables of interest. One important application of multivariate statistics is to identify potential confounding factors by examining the joint distribution of the exposure variable with other variables in the study.

For example, consider an observational study that examines whether aspirin use lowers the risk of myocardial infarction. Aspirin use might be a marker of other health-related factors that also influence myocardial infarction risk, thereby confounding this association. Multivariate statistics could be used to describe the joint distribution of the exposure variable (aspirin use) with potential confounding variables in the study. Since aspirin use is a binary variable, multivariate statistics could simply be the tabulation of means and standard deviations for continuous study variables according to values of aspirin use.

CONCLUSION

Biostatistics is the soul of biomedical research without which the research cannot yield any meaningful outcome. It is equivalently essential for researchers to exploit statistical techniques very early in the planning of research. The estimation of sample size and determination of whether such sample size have the needed power or whether it is adequately representative of the study population can only be achieved by exploiting appropriate statistical model.

The common parlance of ‘garbage in garbage out’ is an axiom that holds very true in the use of statistics in clinical research because what is not statistically planned at the commencement of the study cannot yield any meaningful data neither could be analysed at the data analysis stage and by implication, any useful clinical evidence. It is therefore essential that all biomedical

researchers should not only be conversant with statistics but should be capable of applying and interpreting statistically analyzed results.

REFERENCES

1. Abdolell M., LeBlanc M., Stephens D., et al. Binary partitioning for continuous longitudinal data: categorizing a prognostic variable. Statistics in Medicine 2002;21:3395–3409. [PubMed: 12407680]

2. Altman D. G., Lausen B, Sauerbrei W, et al. Dangers of using “optimal” cutpoints in the evaluation of prognostic factors. Journal of the National Cancer Institute 1994;86:829–835. [PubMed: 8182763]

3. Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., Gatsonis C.A., Glasziou, P.P., Irwig, L.M., et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med 2003;138:W1-12. http://dx.doi.org/10.7326/0003-4819-138-1-200301070-00010.

4. Cadarso-Suárez, C. and González-Manteiga W. (2007). Statistics In Biomedical Research; ARBOR Ciencia, Pensamiento y Cultura CLXXXIII 725 353-361 ISSN: 0210-1963

5. Chansky, K., Sculier, J.P., Crowley, J.J., et al. The International Association for the Study of Lung Cancer Staging Project: prognostic factors and pathologic TNM stage in surgically managed non-small cell lung cancer. J Thorac Oncol 2009;4(7):792–801. [PubMed: 19458556]

6. Cooper, R.J., Schriger D.L., Close R.J.H. Graphical literacy: the quality of graphs in a large-circulation journal. Ann Emerg Med. 2002;40:317–22.

7. Ercan I., Yazıcı B., Yang, Y., Özkaya G., Cangur S., Ediz, B., Kan, I., (2007); Misusage of Statistics in Medical Research. *Eur J Gen Med; 4(3):128-134*

8. García-Berthou E., Alcaraz C., Incongruence between test statistics and P values in medical papers. BMC Med Res Method. 2004;4:13–7.

9. Harrell, F.E., Jr. Regression modelling strategies with applications to linear models, logistic regression, and survival analysis. Springer-Verlag; New York: 2001.

10. Mandrekar, J. N. and Mandrekar, S. J., (2009). Biostatistics: A toolkit for exploration, validation and interpretation of clinical data. *J Thorac Oncol*.; 4(12): 1447–1449. doi:10.1097/JTO.0b013e3181c0a329.

11. Mazumdar, M., Glassman, J. R., Categorizing a prognostic variable: review of methods, code for easy implementation and applications to decision-making about cancer treatments. Statistics in Medicine 2000;19:113–132. [PubMed: 10623917]

12. Moher,. D., Schulz, K.F., Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomized trials. BMC Med Res Methodol 2001;1:2. http://dx.doi.org/10.1186/1471-2288-1-2.

13. Porter, A. M., Misuse of correlation and regression in three medical journals. J Roy Soc Med. 1999;92:123–8.

14. Sprent, P., Statistics in medical research. Swiss Med Wkly 2003;133(3940):522–9. [PubMed: 14655052]

15. Strasak, A. M., Zaman, Q., Pfeiffer, K. P., Göbel, G., Ulmer, H., (2 0 0 7). Statistical errors in medical research –a review of common pitfalls SWISS MED WKLY; 137:44–49 www.smw.ch

16. Thiese, M. S., Arnold, Z. C., Walker, S. D. (2015). The misuse and abuse of statistics in biomedical research: Biochemia Medica 2015;25(1):5–11 http://dx.doi.org/10.11613/BM.2015.001

17. Von Elm E., Altman D.G., Egger M., Pocock, S.J., Gotzsche, P.C., Vandenbroucke, J.P., The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Prev Med 2007;45:247-51. http://dx.doi.org/10.1016/j.ypmed.2007.08.012.