RESEARCH ARTICLE

Assessment of Statistical Methodologies and Pitfalls of Dissertations Carried Out at National Cancer Institute, Cairo University

Rasha M Allam, Maissa K Noaman, Manar M Moneer*, Inas A Elattar

Abstract

Purpose: To identify statistical errors and pitfalls in dissertations performed as part of the requirements for the Medical Doctorate (MD) degree at the National Cancer Institute (NCI), Cairo University (CU) to improve the quality of medical research. **Methods:** A critical assessment of 62 MD dissertations conducted in 3 departments at NCI, CU, between 2009 and 2013 was carried out regarding statistical methodology and presentation of the results. To detect differences in study characteristics over time, grouping was into two periods; 2009-2010 and 2011-2013. **Results:** Statistical methods were appropriate in only 13 studies (24.5%). The most common statistical tests applied were chi-square, log-rank, and Mann-Whitney tests. Four studies estimated sample size and/or power. Only 37.1% and 38.7% of dissertation results supported aims and answered the research questions, respectively. Most of results were misinterpreted (82.3%) with misuse of statistical terminology (77.4%). Tabular and graphical data display was independently informative in only 36 dissertations (58.1%) with accurate titles and labels in only 17 (27.4%). Statistical tests fulfilled the assumptions only in 29 studies; with evident misuse in 33. Ten dissertations reported non-significance regarding their primary outcome measure; the median power of the test was 35.5% (range: 6-60%). There was no significant change in the characteristics between the time periods. **Conclusion:** MD dissertations at NCI have many epidemiological and statistical defects that may compromise the external validity of the results. It is recommended to involve a biostatistician from the very start to improve study design, sample size calculation, end points estimation and measures.

Keywords: MD theses- National Cancer Institute- study design- statistical methodology

Asian Pac J Cancer Prev, **18 (1)**, 231-237

Introduction

Statistics plays a fundamental role throughout the course of research as it is the science of designing studies and collecting and analyzing data aiming at decision making and scientific discovery if the available evidence is insufficient and/or variable. Thus, statistics is the science of learning from data (Ott and Longnecker, 2010).

In modern medical research projects, statistics is a fundamental constituent. Medicine and statistics reached a stage of development where the number of people with expertise in both areas is declining. Statistics has two roles in medical research. First, during planning, statistics is needed to ensure sound experimental design and best usage of available resources. Sound statistical design is the only possible way for adequate statistical analysis of data which is the second role of statistics. Conclusions based on experimental data should be supported by a relevant statistical analysis. The analytical stage involves two steps; summarization of the data and statistical testing (du Prel et al., 2010).

Currently, almost all researchers have an access to computer software for statistical testing. However, make decisions on what test to do and when and what are the prerequisites is not supported by the software. It cannot tell which correct statistical test to use for which situation and data set. The software offers a large array of statistical tests disregarding its relevance to data the researcher needs to analyze. Hence, knowledge on choosing the correct test is essential for the researcher (Gunawardena, 2011).

Despite the great increase in the use of statistical methods in the field of medical research over the past four decades, there is wide consensus that standards are generally low. A large proportion of published medical research contains statistical errors and flaws (García-Berthou and Alcaraz, 2004). Statisticians are needed from the early stages of any research for proper design to avoid mistakes at this point that may disturb all subsequent stages of medical research. This problem necessitates proper management because inappropriate statistical analysis may yield incorrect conclusions, false results and a waste of resources (Strasak et al., 2007a).

This study comprehensively reviewed dissertations presented during the period from 2009 to 2013 as a part of the requirements for the Medical Doctorate (MD) degree in 3 departments at the National Cancer Institute

*Department of Biostatistics and Cancer Epidemiology, National Cancer Institute, Cairo University, Egypt. *For Correspondence: mailbox_mcs@yahoo.com*

(NCI), Cairo University (CU) to identify errors, flaws and pitfalls in design, statistical analysis and presentation. The ultimate goal was to help medical researchers produce statistically sound output in their future investigations to improve the quality of medical research in the institute.

Material and methods

Study design

A critical assessment of the MD dissertations discussed during the period from 2009 to 2013 in the Medical oncology (MO), Pediatric oncology (PO) and Clinical pathology (CP) departments of NCI, CU. First, all MD dissertations in these departments included in the MD theses index of the NCI library were registered. Dissertation's titles, serial numbers and year of discussion were recorded and the search for each one in the library archive started. The total number of registered dissertations was 62. A data collection form based on Consolidated Standards of Reporting Trials (CONSORT) and Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) checklist was constructed.

The evaluation checklist included each section of the study from the title up to conclusions. For example, the results and conclusion were examined to find if they answer the research question and supports the aim of the study. Presentation of the results in tables, illustrations and explanatory text was assessed. Statistical tests used were examined for appropriateness to the sample size, data type, types of dependent and independent variables and fulfillment of other assumptions of each test.

Period comparison

The studies were divided based on the defense time into 2 time periods; (2009-2010), and (2011-2013). These 2 intervals were compared to determine differences in the study characteristics with time.

Statistical methods

Data were analyzed using SPSSwin statistical package version 21 (SPSS Inc., Chicago, IL). Qualitative data were expressed as frequencies and percentages. Chi-square or Fisher's exact tests were used to examine the relation between qualitative variables as appropriate. A p-value < 0.05 was considered significant. All tests were 2 tailed.

Table 1. Application and Misapplication of Different Statistical Tests Used - NCI, 2009-2013

Test statistic	Number of misapplication	Cause of misapplication	Correction
Mann-Whitney U test (n=20)	3		
	1	Comparing median for censored data (TTP)	Log-rank test
	1	ND	Student t-tests
	1	Comparing categorical variables	Chi square or Fisher's Exact as appropriate
Student t-test (n=16)	21*		
	14	Not ND	Mann-Whitney U test
	5	Small sample size not tested for normality	Testing for normality then select appropriate test
	1	Testing relation between categorical variables	Chi square or Fisher's Exact as appropriate
	1	Paired data and not ND	Wilcoxon signed- rank test
Chi Square (n=50)	44		
	15	Expected count < 1	Use Fisher exact test if it was 2 by 2 table or combine categories and then use Fisher
	18	Expected counts < 5 in more than 25% of cells	Combine categories then use Chi square test
	3	More than one p value in same table	Only one P value
	5	Calculation of direct estimates with censored data	Survival analysis
	2	Only one proportion tested	Summary statistics with CI
	1	Required but not used	To be used
Log rank (n=36)	11		
	1	No post hoc tests for more than 2 groups	Post hoc test
	1	Comparison between time to relapse and DFS	Omit
	9	Response to treatment in relation to survival in Log rank	Do not put response as a prognostic factor as it is related to outcome

ND, normally distributed; TTP, time to progression; CI, confidence interval; DFS, disease free survival; *More than 1 misapplication can occur in one thesis

Table 2. Characteristics of Results Section and Use of Statistical Analysis of MD Dissertations - NCI, 2009-2010 vs. 2011-2013

	Year of defense		
	(2009-2010)	(2011-2013)	p value
	n=29	n=33	
Results support aims	8 (27.6)	16 (48.5)	0.092
Comparable groups in relevant measures	4 (33.3)	8 (57.1)	0.225
Complementary text with data in tables and illustrations	21 (72.4)	30 (90.9)	0.057
Missing data for each variable stated	7 (24.1)	10 (30.3)	0.587
Misinterpretation of results	24 (82.8)	27 (81.8)	0.923
Misapplication of statistical words	24 (82.8)	24 (72.7)	0.346
Type of analysis			
Multivariate	5 (17.2)	8 (24.2)	0.499
Univariate	24 (82.8)	25 (75.8)	
Proper type of analysis	22 (75.9)	26 (78.8)	0.783
Accurate title and labels of tables and graphs	6 (20.7)	11 (33.3)	0.265
Good organization of tables and graphs	8 (27.6)	10 (30.3)	0.814
Satisfactory presentation of statistical output	8 (27.6)	15 (45.5)	0.146
Discrepancies between text and tables	10 (34.5)	16 (48.5)	0.265
Misuse of statistical tests	14 (48.3)	19 (57.6)	0.464
Statistical tests fulfilling assumptions	14 (48.3)	17 (51.5)	0.799

Data is presented as n (%)

Results

The results section shows first a comprehensive view of the dissertation's characteristics then, it shows a comparison between early and more recent studies, i.e. 2009-2010 vs. 2011 to 2013.

Statistical methods were mentioned explicitly in 53 studies (85.5%), but were complete only in 28 studies (52.8%) and appropriate in only 13 (24.5%). One study used only descriptive statistical methods, and 61 studies used both descriptive and analytical statistical methods (98.4%). Level of significance was mentioned in only 4 studies. Also number of tails was mentioned in 4 studies and all of them were 2-tailed. The most commonly statistical package used was SPSS version 17.

Only 37.1% and 38.7% of dissertation results supported the aim and answered the research question, respectively. Most of the results were organized in

Table 3. Comparison between Strasak et al., (2007b) Review and Present Study Regarding Design of Study, Data Analysis, Documentation and Presentation

Category	WKW	WMW	Present Study n=62 (%)
	n= 15 (%)	n= 7 (%)	
Design of study			
No sample size/power calculation (overall)	73.3	57.1	93.5
Prospective study design	26.7	28.6	43.5
Retrospective study design	26.7	28.6	19.4
Study design not classifiable	20.0	0.0	45.2
Data analysis			
Use of a wrong statistical test	20.0	42.9	53.2
Failure to include a multiple-comparison correction/α level correction	20.0	14.3	0.0
Special errors with $\chi 2$-tests			
No Yates correction if small numbers	13.3	0.0	0.0
Use of $\chi 2$ when expected numbers in a cell < 5	6.7	28.6	36.0
Documentation			
Failure to state number of tails	80.0	85.7	93.5
Failure to specify which test was performed on a given set of data	26.7	14.3	100.0
Presentation			
Giving standard error (SEM) instead of SD for statistical description	6.7	0.0	6.0
p = NS, p < 0.05, p > 0.05 etc. instead of reporting exact p-values	46.7	71.4	8.1

WKW, Wiener Klinische Wochenschrift; WMW, Wiener MedizinischeWochenschrift

Table 4. Comparison between Strasak et al., (2007b) Review and Present Study Regarding Type and Frequencies of Statistical Tests Used

Types and frequencies	WKW n =35 (%)	WMW n =16 (%)	Present study n =62 (%)
No statistical methods	2.9	25	0
Descriptive statistics only	22.9	31.3	1.6
Inferential methods	74.3	43.8	98.4
t-tests	28.6	12.5	61.3
Contingency table analysis (χ2, Fishers exact test)	19.4	31.3	95.2
Non-parametric tests	28.6	25	59.7
One-way ANOVA	5.7	0 0	6.5
Correlation coefficients	22.9	2 12.5	20.1
Regression	25.7	1 6.3	9.7
Survival analysis	11.4	0 0	58.1
Confidence intervals	14.4	212.5	58.3

WKW, Wiener KlinischeWochenschrift; WMW, Wiener MedizinischeWochenschrift

the order of the importance of objectives (79%) with complementary text, data in tables and illustrations (82.3%). Unfortunately, most of the results were misinterpreted (82.3%) with misuse of statistical words (77.4%). Only 24 studies involved multiple groups, the comparability between groups was tested in 12/24 studies. Tabular and graphical data display was independently informative in 36 dissertations (58.1%); only 17 (27.4%) had accurate title and labels. Statistical material was reasonably presented in 37.1%; however discrepancies between text and tables were seen in 41.9%. Statistical tests fulfilled the assumptions in 29 studies; while evident misuse was observed in 33 studies. Most of the studies (n=56) reported the exact p-value (90.3%). Confidence interval use was applicable in 36 dissertations; it was mentioned only in 21 (58.3%).

Overall, the most common descriptive measure used were frequencies in all studies, followed by means and standard deviations in 50 studies (80.6%), then the medians, interquartile ranges or ranges in 39 of the studies (62.9%). The frequency of usage and misusage of the different tests is shown in Table 1. The Mann-Whitney U test was the most commonly used test for the comparison of numerical variables in two groups; 20 (32.3%). The

Table 5. Comparison between *Hanif* and *Ajmal* Study and Present Study Regarding Statistical Methodology, Design and Statistical Errors

Statistical Methodology	Hanif and Ajmal study (2011) n = 80 (%)	Present study n = 62 (%)
Design of study not given	52.5	85.5
No Sample size calculation/ power calculation (overall)	92.5	93.5
Sampling Selection criteria not given	75.0	79.0
No statistical methods	26.3	14.5
Data analysis technique defined	48.7	85.5
No statistical package defined with version	70.0	32.3
Descriptive statistics only	28.8	1.6
Inferential methods with descriptive	41.3	98.3
Contingency table analysis	30.0	95.2
t-tests	13.8	61.3
Basic Chi-square, Fisher's Test	30.0	95.2
Non-Parametric tests	3.8	59.7
Analysis of Variance	7.5	6.5
Correlation coefficient	7.5	21.0
Logistic Regression	8.7	9.7
Survival Analysis	0.0	58.1
Confidence interval	15.0	58.1
Use of wrong statistical analysis	28.7	21.0
Incompatibility of statistical test with type of data examined	20.0	53.2
Over all inappropriate interpretation	13.7	82.3

Table 6. Comparison between Leucuța et al. Study and Present Study

Category	Leucuța et al., Study (2013) n = 170 (%)	Present study n=62 (%)
Summarize each variable with descriptive statistics	97.1	100.0
Verify that data conformed to the assumptions	12.4	46.8
Indicate whether and how any allowance or adjustments were made for multiple comparisons	44.0	0.0
Report how any outlying data were treated in the analysis	4.9	0.0
Say whether tests were one- or two-tailed	7.8	6.5
Report the alpha level (e.g. 0.05)	75.5	6.5
Name the statistical package or program used	32.8	79.2
Report total or group sample size for analyses	80	100
95% confidence coefficient to indicate the precision of an estimate	11.1	58.3

Student's t-test was performed in 16 (25.8%). Each of the paired t-test and Wilcoxon signed-rank matched-pairs test was used once (1.6%). The sample size for the use of the paired t-test was small and the assumption of normality was not tested. For the comparison of numerical variables in more than two groups, Kruskal-Wallis test was used more than one way analysis of variance (ANOVA); in 9.7% and 6.5% of the studies, respectively. Kruskal-Wallis test was not followed by a post hoc test in 2/3 of the theses. Use of ANOVA was inappropriate in all of the 4 dissertations; the distribution was asymmetric. The comparison between proportions was done using either Chi-square or Fisher exact test; the former was more commonly used in 50 studies (80.6%). The assumption concerning the sample size was violated in more than half of the theses.

Spearman's rho correlation coefficient was used more than Pearson's correlation coefficient (16.1% versus 4.8%, respectively). Half of times the correlations were inappropriate. The candidate correlated survival time or a variable with itself or categorical variables. The logistic regression was performed 6 times (9.7%), only one candidate did not mention the odds ratio or confidence interval.

The survival estimates were calculated with Kaplan-Meier methods in 36 studies (58.1%), and the survival curves compared using the Log-rank test. Multivariate analysis (Cox proportional hazard model) was done in 9/36 studies (25%) for testing independent prognostic effect of different statistically significant variables on univariate levels. It was misused in 3 dissertations, in one the author did not include the hazard ratio and confidence interval and in 2, the response to treatment was included in the model as a prognostic factor.

The analysis of 49 dissertations was univariate. Of these dissertations, 12 needed further multivariate analysis. Ten dissertations reported non-significance of their primary outcome measure. The power of the test was calculated for these studies; it ranged from 6% to 60% with a median of 35.5%.

The time period of the study was divided into two; 2009-2010 (n=29) and 2011-2013 (n=33) to show if there was any differences in the characteristics of the MD dissertations presented. There was no significant difference between the 2 time periods, Table 2.

Discussion

The ultimate goal of this study was to review past patterns in research to tune the future directions so as to maximize the achievements and minimize the shortcomings. This critical assessment was limited to the number of available MD dissertations in the period from 2009 to 2013 that were archived in the NCI library. Efforts of all researchers and supervisors were acknowledged and appreciated; however, errors were rather common. Statistical methods were inappropriate in 75% of the studies.

Using inappropriate statistical methods can be a waste of time and financial resources, and is detrimental to the scientific concepts and to humanity. Using incorrect statistical methods can produce misleading, suboptimal, incoherent results amenable to be cited by other researchers (Ercan et al., 2007).

The commonness of statistical misuse can be explained by lacking basic statistical knowledge among the medical community in general. Nevertheless, in other cases misuse may be deliberately done to attain a desired result. A systematic review found that 33.7% of surveyed research admitted to questionable practices, including adjusting results to improve the outcome, questionable interpretation of data, concealment of methodological details and dropping observations based on "feeling they were inaccurate" (Fanelli , 2009).

In the current study, statistical tests fulfilled the assumptions in 29 studies; thus 33 studies had evident misuse. If the assumptions of statistical test are not appropriately considered, significant errors and misinterpretation of results are possible. These errors may completely invalidate results and consequently linked conclusions (Jamart, 2008). The study can be appropriately planned and performed, but, incorrect analytical methodology can be grave enough to waste all efforts and costs though incorrect inferences. Actually, the majority of published articles are devoid of discussion of statistical assumptions. One study reported this in nearly 90% of evaluated articles (Williams et al., 1997).

Statistical tests are precisely designed for specific types of data. With the large collection of tests now available in computer programs, comprehensive consideration must be given to their assumptions to guide careful selection.

Many articles fail to report which statistical tests were utilized during data analysis (Strasak et al., 2007a).In the current study, 53.2% of the dissertations used improper statistical tests. Ercan et al., (2012) revised 181 original articles submitted to the TKJMS for detection of statistical errors. An inappropriate statistical test was used in 28.2% of the reviewed manuscripts. Welch and Gabbe (2002) and Hanif and Ajmal (2011) reported comparable rates; 31.7% and 28.8%, respectively.

Strasak et al., (2007b) evaluated the quantity and quality of the use of statistics in two Austrian Medical Journals. All "original research" papers in some articles of two journals; Wiener Klinische Wochenschrift (WKW) and Wiener Medizinische Wochenschrift (WMW) were screened for their statistical content. Their results are compared to this study in Tables 3 and 4.

In the current study, the overall rate of inappropriate interpretation of statistical analysis results was 82.3%. This figure is high and may consequently ruin the deduced conclusions. Drawing conclusions from a study which are insufficiently supported by the data should be avoided. If claiming significance of effects, one has to ensure, that a statistical significance test has been employed. Lack of statistical significance does not invariably mean there was no effect or no difference at all (Strasak et al., 2007a).

The rate of inappropriate interpretation was 10.50% in the study by Ercan et al., (2012), 52.6% in the study by Welch and Gabbe (2002), 4% in the report by Lukiæ and Marušiæ (2001), 13.8% in the study of Hanif and Ajmal (2012) and 17% in the study of McGuigan (1995). The high rate in the current study can be explained by lack of experience of MD candidates compared to the more qualified researchers submitting articles to famous journals. Tables 5 shows more comprehensive comparison of the errors detected in the current study with those reported by Hanif and Ajmal (2011). They reviewed 80 research articles published in indexed and recognized local journals of Pakistan in comparison to the present results. Also, table 6 shows a comparison of the current results with Leucuța et al., (2015) study, who evaluated all pharmaceutical papers published in six Romanian journals, in 2013.

Interpretation related errors were categorized to Harris et al., (2011) into 24% for "lack of understanding the limitations of the analysis, and the need for replication and sensitivity analysis"; 10% for "drawing inferences that go beyond the data such as casual claims for cross-sectional data; 10% for "comparing p-values in separate tests (e.g. in paired t test) to assess group differences; and 5% for the "too much made from "marginally significant" results".

In the current study, ten dissertations reported non-significance of their primary outcome measure. The power of the test ranged from 6% to 60% with a median of 35.5%. A review by Charan and Saxena (2014) was designed to critically evaluate negative studies published in prominent Indian Medical Journals for reporting of statistical and methodological parameters between years 2000 and 2011. Power was reported only in 11.8% studies. Biased negative studies not only reflect poor research effort but also have an impact on 'patient care' as they prevent further research with similar objectives, leading to potential research areas remaining unexplored. Hence, published 'negative studies' should be methodologically strong. All parameters that may help a reader to judge validity of results and conclusions should be reported in published negative studies.

We can conclude that the quality of MD dissertations at the NCI has many defects from the epidemiological and statistical points of view. This may compromise the power of the results and their external validity. These will result in studies which lack scientifically sound basis. This can affect the capability of the resulting research articles to be published on an international basis in highly ranked medical journals and eventually influence the international rank of the institute and consequently the university. Poor quality research work constitutes waste of time and money.

To overcome these consequences, education, training, and application of the basics of the epidemiology, biostatistics, and research methodology for all levels of medical researchers are recommended through: 1) addition of courses in the undergraduate curriculum of medical student, 2) application of research methodology in small projects during graduation of undergraduates, 3) refreshing courses for postgraduate students, 4) continuous lectures, demonstrations and workshops for postdoctoral staff members to be linked to advancement of medical research methodology. In addition, research articles and dissertation should be revised by a specialized epidemiologist and biostatistician before discussion or publication. Prior reviewing of the study protocols by a professional epidemiologist can ensure good quality of the research from the start. Following well stated guidelines as CONSORT guidelines for clinical trials or STROBE guidelines for observational study in writing help to reach the high quality articles. Encouraging collaboration with other medical centers increases the sample size and hence, raises the external validity, power, and generalization. This will ensure proper utilization of limited resources (time, effort and money) in performing proper research aiming at significant contribution to medical literature.

References

Charan J, Saxena D (2014). Reporting of various methodological and statistical parameters in negative studies published in prominent Indian medical journals: A systematic review. *J Postgrad Med*, **60**, 362-5.

du Prel J, Röhrig B, Hommel G, Blettner M (2010). Choosing statistical tests. part 12 of a series on evaluation of scientific publications. *Dtsch Arztebl In*t, **107**, 343–8.

Ercan I, Ocakoğlu G, Sigirli D, Özkaya G (2012). Assessment of submitted manuscripts in medical sciences according to statistical errors. *Turkiye Klinikleri J Med Sci*, **32**, 1381-7.

Ercan I, Yazici B, Yang Y, et al (2007). Misusage of statistics in medical research. *Eur J Gen Med*, **4**, 128-4.

Fanelli D (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS One*, **4**, e5738.

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

Gunawardena N (2011). Choosing the correct statistical test in research. *Sri Lanka Journal of Child Health*, **40**, 149-53.

Hanif A, Ajmal T (2011). Statistical errors in medical journals (A

critical appraisal). *Ann King Edward Med Univ*, **17**, 178-82.
Harris A, Reeder R, Hyun J (2011). Survey of editors and reviewers of high-impact psychology journals: statistical and research design problems in submitted manuscripts. *J Psychol*, **145**, 195-9.
Jamart J (1992). Statistical tests in medical research. *Acta Oncol*, **31**, 723-7.
Leucuța DC, Drugan T, Farcaş A, Achimaş A (2015). Statistical reporting in pharmaceutical papers from Romanian journals. *Farmacia*, **63**, 394-401.
Lukić IK, Marusić M (2001). Appointment of statistical editor and quality of statistics in a small medical journal. *Croat Med J*, **42**, 500-3.
McGuigan SM (1995). The use of statistics in the British journal of psychiatry. *Br J Psychiatry*, **167**, 683-8.
Ott RL, Longnecker M (2010). An introduction to statistical methods and data analysis, Sixth Ed. Brooks/Cole, Cengage Learning.
Strasak AM, Zaman Q, Pfeiffer KP, Gobel G, Ulmer H (2007a). Statistical errors in medical research-a review of common pitfalls. *Swiss Med Wkly*, **137**, 44-9.
Strasak AM, Zaman Q, Marinell G, Pfeiffer KP, Ulmer H (2007b). The Use of statistics in medical research: A comparison of Wiener Klinische Wochenschrift and Wiener Medizinische Wochenschrift. *Aust J Stat*, **36**, 141-2.
Welch GE, Gabbe SG (2002). Statistics usage in the American journal of obstetrics and gynecology: has anything changed? *Am J Obstet Gynecol*, **186**, 584-6.
Williams JL, Hathaway CA, Kloster KL, Layne BH (1997). Low power, type II errors, and other statistical problems in recent cardiovascular research. *Am J Physiol*, **273**, 487-3.