ORAL DISEASES

Oral Diseases (2010) 16, 233–241. doi:10.1111/j.1601-0825.2009.01634.x

http://www.blackwellmunksgaard.com

INVITED MEDICAL REVIEW

Design and statistical analysis of oral medicine studies: common pitfalls

L Baccaglini[1], JJ Shuster[2,3], J Cheng[2], DW Theriaque[4], VJ Schoenbach[5], SL Tomar[1], C Poole[5]

[1]*Department of Community Dentistry & Behavioral Science, College of Dentistry, University of Florida, Gainesville, FL;* [2]*Division of Biostatistics, Department of Epidemiology and Health Policy Research, College of Medicine, University of Florida, Gainesville, FL;* [3]*Shands Clinical Research Unit, Clinical and Translational Science Institute, University of Florida, Gainesville, FL;* [4]*Regulatory Knowledge and Research Support, Clinical and Translational Science Institute, University of Florida, Gainesville, FL;* [5]*Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC, USA*

A growing number of articles are emerging in the medical and statistics literature that describe epidemiologic and statistical flaws of research studies. Many examples of these deficiencies are encountered in the oral, craniofacial, and dental literature. However, only a handful of methodologic articles have been published in the oral literature warning investigators of potential errors that may arise early in the study and that can irreparably bias the final results. In this study, we briefly review some of the most common pitfalls that our team of epidemiologists and statisticians has identified during the review of submitted or published manuscripts and research grant applications. We use practical examples from the oral medicine and dental literature to illustrate potential shortcomings in the design and analysis of research studies, and how these deficiencies may affect the results and their interpretation. A good study design is essential, because errors in the analysis can be corrected if the design was sound, but flaws in study design can lead to data that are not salvageable. We recommend consultation with an epidemiologist or a statistician during the planning phase of a research study to optimize study efficiency, minimize potential sources of bias, and document the analytic plan.
Oral Diseases (2010) **16**, 233–241

Keywords: public health; bias; epidemiology; guideline; methods; statistics

Correspondence: Lorena Baccaglini, DDS, DDS, MS, PhD, Department Epidemiology and Biostatistics, College of Public Health and Health Professions, Department of Community Dentistry & Behavioral Science, University of Florida College of Dentistry, 1329 SW 16th Street, Suite 5182, PO Box 103628, Gainesville, FL 32610 3628, USA. Tel: (352) 273 5962, Fax: (352) 273 5985, E-mail: lbaccaglini@dental.ufl.edu
Received 9 September 2009; accepted 15 September 2009

Introduction

In 2006, oral medicine experts from across the globe gathered in San Juan, Puerto Rico, for one of the largest meetings in oral medicine history, the Fourth World Workshop on Oral Medicine. After 2 years of preparation and systematic reviews of hundreds of published oral medicine articles, the review teams met to reach a final consensus and develop management recommendations for 10 selected clinical conditions (Baccaglini *et al*, 2007a,b). At that time, the review teams also reached another consensus: that the overall poor quality of published oral medicine studies had hindered the successful development of strong evidence-based clinical recommendations.

Since that time, there has been an outpouring of review articles and commentaries in medical, epidemiologic, and statistical journals outlining the urgent need for careful execution and reporting of medical research studies. General guidelines for the reporting of observational studies (strengthening the reporting of observational studies in epidemiology; http://www.strobe-statement.org/) and clinical trials (Consolidated Standards of Reporting Trials; http://www.consort-statement.org/) have been published in and adopted by scientific journals (Vandenbroucke *et al*, 2007) and disseminated through a global network (Enhancing the QUAlity and Transparency Of health Research Network; Moher *et al*, 2008). These guidelines can contribute to improve the reporting of medical research studies (Kane *et al*, 2007; Vandenbroucke *et al*, 2007), although they should not be viewed as rigid prescriptions (Rothman and Poole, 2007).

Oral medicine clinicians and researchers may not have been exposed to such guidelines, due in part to the lack of methodologic articles appearing in journals most frequently read by dentists and oral medicine practitioners and investigators. Furthermore, guidelines for reporting studies may be misconstrued as applying only to the publication phase rather than to the planning

phase, which is problematic because many weaknesses cannot be rectified if the data collection is already underway or has been completed.

To raise awareness about the intricacy of oral medicine research methodology, we have gathered a small team of epidemiologists, statisticians, informaticians, dentists, and oral medicine experts to present some of the most commonly encountered methodologic and reporting issues in oral medicine studies. To avoid drawing rigid chronologic timelines and to stress the need for investigators to have a global understanding of the study from its onset, we have chosen to address the conduct and reporting of research studies simultaneously.

For convenience of exposition, our commentary is divided into three broad areas, corresponding to the primary sections of a manuscript: methods (with considerations related to design, sample size, data entry and management, and statistical analyses), results, and discussion. This division is not meant to imply that the sections are independent of one another. In an actual study, methods, results, and discussion should be highly integrated.

Our commentary is not intended as a recipe for how to conduct and report oral medicine studies, but as a distillation of our collective experience as journal, abstract and grant application reviewers, researchers, and instructors.

Methods

Pitfalls in the implementation and reporting of research methodology may occur during study design, sample size calculations, data management, and statistical analyses.

Study design

In a well-planned study, investigators typically first formulate a clearly defined research question and then carefully choose the design and analyses that are most suitable to answer their question. Even when a specific research question is not explicitly formulated before data collection, such as in large national surveys, the research questions likely to be of interest should be considered so that data collected will enable investigators to answer these and similar research questions through secondary statistical analyses. During the planning stage, consultation with an epidemiologist or statistician is desirable, because study design affects subsequent statistical analyses and defines what inferences can be made. Failure to recognize this link can lead to disappointment in the analysis and publication phases.

Threats to study validity include selection bias, information bias, and confounding. Within these three large and partially overlapping categories, different types of specific biases can be recognized (Jacob, 2002; Delgado-Rodriguez and Llorca, 2004). Investigators should be fully familiar with potential biases that may affect their study and should identify strategies to minimize these biases during the planning stage and before beginning data collection. Two common causes of selection bias in oral medicine studies are improper selection of a comparison group and loss of data.

When designing (and publishing) their studies, investigators should always specify the characteristics of the source population from which the samples are drawn and the recruitment methods used, so that potential sources of bias during participants' selection may be identified. In general, eligibility criteria should be applied evenly to the groups being compared. For example, in a study of oral bacteria in which 100% of cases and only 58% of controls are females (Goodson *et al*, 2009), the gender imbalance can have profound effects on the study results, because gender is linked to a variety of factors, including health habits, which may affect the oral micro-environment and microbial composition.

Investigators are often interested in the effect of a treatment or an exposure. For this objective, the strongest studies are generally those that assess outcomes in participants randomly assigned to either a treatment or a control group. The least desirable study designs are case reports and case series with no comparison group. Studies lacking a contemporaneous, randomly selected comparison group cannot exclude the possibility that observed changes were because of natural disease progression or regression over time rather than because of the treatment or other exposure of interest. This possibility is of particular concern if the more severe cases of an inherently varying condition were chosen for the study, because improvement may be related to the well-known regression to the mean phenomenon.

In some cases, selection bias is introduced by the investigator intentionally. For example, in a case–control study, the investigator may select a control group matched to cases by 5-year age groups. Although matching of treatment and control subjects in a trial improves comparability and reduces the occurrence of bias in the comparison of outcomes, matching of controls to cases actually introduces bias unless the analysis adjusts for the matching factors (Rothman and Greenland, 1998). Selection bias introduced by matching is a frequently overlooked source of bias. Thus, investigators should always report how matching was performed and how the matching factors were treated in the analysis. Of note, caliper matching, e.g., selecting the comparison group by ± 5 years of age or by a fixed number of standard deviations, is inferior to matching within fixed categories (Rothman and Greenland, 1998).

Missing data can be a source of error. In dental research, a frequent cause of missing data is missing teeth (Slade and Caplan, 1999). Depending on the pattern of missing data and on how they are handled, missing data can introduce both random and systematic error. Both parameter estimates (e.g., relative risk, odds ratios, or mean values) and statistical inference [confidence intervals (CI), *P*-values] may be affected.

Additional sources of missing data are attrition (when participants withdraw from a study or become unavailable for other reasons) and skipping of questionnaire

items. Bias can also be introduced when potential participants are lost through self-selection (a more subtle type of participants' loss that occurs during recruitment) or are deleted from an analysis data set because data are missing for some variables (i.e., during complete case analyses). Missing data can sometimes be avoided by checking for missing questionnaire items while the participant is still in the clinic or by collecting the minimum amount of data necessary to answer the study questions, thereby reducing participants' fatigue. Collecting too many data fields will lower the data quality of the key data fields, and hence every data item collected should undergo rigorous review to confirm that it is needed. A large proportion of missing data, unless the data are missing completely at random (i.e., the pattern of missingness is not related to observed or unobserved values of the variables of interest), can threaten the validity of the study. Depending on the missing data pattern assumptions, special analytical techniques, such as multiple imputation, can be used to partially reduce the bias after the data have already been collected, although even the most educated guess of a missing value contains some degree of uncertainty. Multiple imputation is preferable over single imputation, because it takes some of the uncertainty into account by increasing the final standard error estimates.

Information bias may occur when the variables of interest are misclassified by the study participant or the investigator. For example, in a study comparing two treatments applied to the right and left side of the tongue in patients with bilateral oral hairy leukoplakia, the examiner who measures the size of the lesion after treatment could be influenced in its evaluation if he or she knew which treatment was used on which side and had a pre-existing belief about which treatment was more efficacious.

Whenever possible, investigators should incorporate techniques in their study design that tend to minimize information bias, such as masking and calibration. Masking (also known as blinding) can reduce undue influences that occur in participants, examiners, laboratory personnel, and statistical analysts when they are aware of certain information. However, masking does not predictably reduce information bias. In typical settings, it directs information toward the null, possibly at the expense of increasing the bias.

Calibration of investigators and equipment reduces systematic differences in measurement that tend to occur across different research sites or over time. Investigators should also explicitly and specifically define the variables of interest (outcome, exposure, and covariates) to reduce misclassification, avoid 'cherry picking', and allow replication of the study by other investigators. In oral medicine drug studies, for example, investigators should record (and report) the concentration of the drug used, number of applications and exact method of application, and should clearly define measures of improvement.

Confounding is the mixing of effects of certain 'extraneous' variables (confounders) with the potential effect of the treatment or other main exposure of interest. For example, if older individuals have more amalgam restorations and also a higher prevalence of chronic neurologic conditions, such as multiple sclerosis (MS) than do younger individuals, age differences could confound a comparison of amalgam restorations among patients with MS to amalgam restorations in controls. If participants of various ages are included in the study and age is ignored in the analysis (e.g., the investigators report only a crude odds ratio), the odds ratio for MS with respect to amalgam restorations will be higher than if the analysis was conditioned on (controlled for) age.

A number of techniques can be used at both the design and analysis stages to reduce the effects of confounders. At the design stage, investigators can use restriction, randomization, and matching. Investigators should carefully consider advantages and disadvantages of each technique when selecting the most appropriate design.

Restriction is simply the exclusion of certain categories of participants from a study. For example, if all participants of a study are females, there cannot be confounding by gender, but also nothing can be learned about males or gender differences. Randomization is the assignment of a treatment by a method (e.g., computer-generated random number) that on average enhances comparability of groups at baseline. In contrast to other methods of dealing with confounding, randomization reduces confounding from unknown factors as well as known ones. Randomization should be used whenever possible, because it greatly strengthens inferences that can be made. The specific methods used for randomization and corresponding features that may be required in the statistical analysis should be reported. Matching is widely used in oral medicine research, especially in case–control studies. One of the most common misconceptions is that the function of matching in case–control studies is to control confounding. While matching can control confounding in some circumstances in cohort studies, simple calculations demonstrate that this does not occur in case–control studies (Rothman and Greenland, 1998). However, matching by confounders to improve precision can be advisable in both kinds of designs. Investigators should carefully consider potential disadvantages of matching. One disadvantage is the increased time and expense required to find an appropriately matched comparison group, especially when using multiple matching factors or exact matching to a specific value of a variable rather than to a range of values for that variable. Frequency matching, accompanied by a stratified analysis, is an excellent way to match subjects and avoid the difficulties inherent in implementing individual matching.

Another disadvantage of matching (in a case–control study) is the inability to study the effects of the matching factors in the analysis (e.g., the effects of age can no longer be analyzed in a case–control study matched by age), unless randomized control recruitment is used or the relative control-sampling probabilities are otherwise known (Weinberg and Sandler, 1991).

In addition to considering the use of the above techniques to minimize confounding, investigators should also ensure that data on remaining important

possible confounders are collected for subsequent analyses. Analysis of covariance (ANCOVA) and multiple regression are examples of techniques that allow an investigator to control for confounders analytically, provided that these covariates are selected in advance and are few in number. An underutilized and useful way to determine if data should be collected on certain variables to control confounding and other biases is to draw graphs (directed acyclic graphs; DAGs) that illustrate the expected biologic relationship among different variables (Merchant and Pitiphat, 2002; Shrier and Platt, 2008). DAGs are particularly helpful when selecting potential risk factors or confounders for a multi-factorial disease or behavior (Chattopadhyay *et al*, 2003; Baccaglini *et al*, 2006). Variables identified through DAGs as advisable conditioning variables are then used in restriction, matching, stratification, and statistical adjustment. However, effect measure modifiers (i.e., variables that modify the strength of association between the exposure of interest and the disease) are not visualized by traditional DAGs and these variables should be identified separately (VanderWeele and Robins, 2007).

It is critically important, as part of a good study design, to fully document the analytic plan and have it peer reviewed before collecting any study data. This practice provides protection against reviewers who ask investigators to re-analyze the study in a different way after the manuscript has been submitted for publication.

Sample size

During study design, power and precision calculations are performed to determine the number of participants needed for the study or to work backward to determine the power for a given study size. Oral medicine studies frequently have missing or insufficient statistical power and precision or use incorrect sample size calculations.

Missing or insufficient statistical power or precision occurs in the following scenarios:

(1) Investigators do not report power or precision calculations when a justification for the sample size used in the study is appropriate, such as in hypothesis testing.
(2) During planning, the investigators have not considered whether the sample size is too small to have made a study worth conducting.
(3) Investigators do not report the number of comparisons when all three of the following conditions are present: (a) the results of all the comparisons are not reported (which they should be); (b) the analysis consists of Neyman–Pearson hypothesis testing; and (c) one wishes to control the studywise type I error probability.

Four scenarios are frequently encountered in oral medicine studies in which power and precision calculations are incorrect:

(1) Sample size analyses do not take into account the potential for missing or misclassified data. This typically occurs in longitudinal studies if sample size analyses are based on the number of participants at the beginning of the study and not on the final sample size after attrition (loss to follow-up). Incomplete data collection, such as unanswered questionnaire items or missing laboratory results, can also reduce the precision of a study. Additionally, misclassified data can lower the power of a study by reducing observed differences between groups, or raise it by increasing those differences. The effects of missing and misclassified data extend beyond sample size calculations, because bias may be introduced simultaneously.
(2) The initial sample size calculations are based on analysis plans that differ from those performed at the end of the study. The power is usually overestimated if it is based on continuous variables and the analysis is categorical (e.g., high/medium/low, positive/negative), and usually underestimated if it is based on categorical variables that are analyzed as continuous.
(3) The power and precision analyses do not correspond to the underlying biologic model. For example, this occurs in power calculations for genetic studies that incorrectly use the minor allele frequency instead of the genotype prevalence as the exposure prevalence (Altshuler *et al*, 2008). As humans carry two alleles (one for each chromosome) at each locus, if 30% of the chromosomes are expected to have a polymorphism (or mutation) at that particular locus, then only 9% of the population will be considered exposed under a recessive inheritance model, and 51% will be considered exposed under a dominant model.
(4) Failure to take into account the multivariate nature of a hypothesis. The analysis and power calculation must be multivariate to control studywise error for the hypothesis if the outcome of a study question is multivariate, if null hypothesis testing is to be performed and if the investigator desired to control the studywise type I error probability.

Data entry and management

Common pitfalls in the conduct and reporting of data entry and data management procedures in oral medicine studies include:

(1) Lack of a basic description of the data entry and data management methodology. The type and frequency of data transfer has an impact on the degree of expected error. For example, double data entry, especially if performed by different operators, detects more data entry errors and is preferable to single data entry, although the former procedure is more costly and time consuming. Whenever possible, investigators should minimize the number of manual data transfers and should report the methods used. Direct data entry using programs that minimize invalid entries is helpful. Data fields should also be calculated directly by computer cross-field calculations wherever possible, rather

than be manually calculated and entered. For example, body mass index should be machine calculated from height and weight by a coded algorithm in the data entry program or statistical software program and not be calculated and entered by hand. Investigators should also always maintain a copy of the original data set and a detailed record of any changes made and who made the entry or change. Ideally, those records should be generated automatically each time the data set is modified.

(2) Selectively removing certain data points 'after the fact', such as excluding patients that develop clinical disease in a study of the effects of treatments on Candida counts, should be avoided. In this scenario, a treatment may appear to be more effective once all the participants that develop clinical Candidiasis are deleted from the data set (Patel *et al*, 2008). Another procedure to avoid is the removal of outliers after the final analyses followed by re-analysis of the data, even though the outliers were identified during data management and deemed to be within the range of plausible values. To ensure that these decisions are not influenced by one's knowledge of the results, the data manager should identify and discuss out-of-range, impossible or implausible values with the scientist before the final statistical analyses. Removal of these values should have a strong justification, because this procedure can create bias, or replace one bias (information bias) with another (selection bias). Investigators often have good pilot data to decide at the design phase how best to treat the data (log transformation *vs* rank methods *vs* ordinary parametric analysis, with or without a Satterthwaite correction for unequal variance; Shuster, 2009).

(3) Excessive use of categorization, often based on arbitrary cut-off points. A common practice is to transform a continuous variable into a binary variable. Sole reliance on this procedure makes biologic interpretations more difficult (this is especially true when multiple heterogeneous groups are grouped together), it reduces statistical power by throwing away information, it can lead to residual confounding, and it also inhibits the investigator's ability to observe a dose-response relationship (Royston *et al*, 2006). Continuous and categorical analyses should be carried out in concert to take advantage of their dovetailing strengths and limitations.

Statistical analyses

Pitfalls in performing and reporting statistical analyses are very common. Statistical analyses may be missing, excessive or incorrect. The most frequent scenarios related to missing or incomplete statistical analyses include:

(1) No statistical analyses were conducted, even though it would have been possible and informative to perform them. This includes, for example, not reporting intent-to-treat analyses when indicated (e.g., in a randomized trial in which we are studying risk) or stating that an association is present or stronger in one subgroup and absent or weaker in another subgroup without conducting the necessary analysis to measure or test the difference.

(2) Missing basic univariate analyses, such as description of demographic characteristics of the sample, so that it is not clear to whom the results may be generalizable.

(3) Reporting or performing only crude bivariate analyses in the presence of strong confounders, such as in the amalgam and MS example (Young, 2007), or in the presence of effect measure modifiers (interactions).

(4) Reporting adjusted results without specifying the covariates used for adjustment or the method used to select the covariates for inclusion in the model (Groenwold *et al*, 2008).

Excessive use of statistical analyses occurs in the following scenarios:

(1) The investigators perform multiple analyses beyond those originally planned and do not label them as exploratory.

(2) Using multiple pairwise tests instead of an overall test to compare multiple groups when a conclusion or inference is drawn about the ensemble of hypotheses and the investigator chooses to control the studywise type I error probability. For example, this occurs when using multiple individual *t*-tests instead of (a) an overall error controlled multiple comparison (e.g. Tukey's test or Bonferroni correction) or (b) an F test followed by post hoc comparisons if the overall test is significant.

(3) Statistically comparing randomized groups at baseline "to see if the randomization worked". This consideration is also applicable to observational studies.

Descriptive analyses (typically reported in the first table of a trial or observational study report) should not include *P*-values, null hypothesis tests, estimated associations, or CI for those estimates (Vandenbroucke *et al*, 2007). There are a number of reasons for not performing these tests, including: (a) the study was not designed (and may not be powered) to conduct these tests; (b) the variables tested may not have prognostic importance; (c) we are not testing if the variables are important and (d) if multiple tests are performed, a few will likely turn out positive by chance, regardless of the number of participants in the study (Altman, 1984).

The incorrect choice of statistical analyses is one of the most common flaws of oral medicine studies. This primarily occurs when the analyses do not fully relate to the hypothesis or the study design.

Ignoring the underlying study design during statistical analyses is a very serious and frequent problem. An especially common problem is the use of statistical tests that assume data independence (such as unpaired *t*-tests, logistic regression, ANOVA and chi-square) when data

are not independent, such as in study designs that use matching, repeated measures, and clustered sampling. This may occur in analyses of any kind of study, including trials and other cohort studies. Examples of typical analysis errors include:

(1) Using an unpaired *t*-test when comparing two topical treatments applied to the right or left sides of the tongue or two periodontal surgeries performed on the maxillary and mandibular quadrants of the same patient. In these cases, the inappropriate use of the unpaired *t*-test rather than an analysis that takes into account the paired design (Lesaffre *et al*, 2007) makes the conclusions conservative.
(2) Cases and controls are matched by gender and age, and then chi-square tests that ignore matching are used in the analyses. As noted above, matching in case–control studies introduces selection bias that requires adjustment for the matching factor in the analysis phase.
(3) Data from multiple visits are compared by using an ANOVA or multiple unpaired *t*-tests.
(4) Analyzing duplicate or triplicate measurements performed during one experiment as if they were from separate experiments. This often occurs during laboratory analyses, where repeated measurements are often used to improve precision.
(5) Analyzing surveys employing complex sampling designs as if data were from a simple random sample, e.g., incorrectly assuming that all persons in the US had an equal probability of being selected for a national survey conducted with multistage sampling. Nationally representative samples, such as the National Health and Nutrition Examination Survey (NHANES), are typically selected by using a multistage clustered stratified design with oversampling that requires complex statistical analyses, such as generalized linear models (Caplan *et al*, 1999). Failure to account for the complex design, for example by using standard *t*-tests or chi-square tests, leads to incorrect results. If clustering is ignored, standard errors will be underestimated, leading to narrower CI and smaller *P*-values. If sample weights are ignored, both parameter estimates and standard errors will be incorrect.
(6) Analyzing multiple teeth from the same patient as if they were from different patients. Teeth from the same patient are not independent, because they are exposed to similar risk factors (e.g., diet or genetics).
(7) Analyzing multiple histologic sections taken from the same tooth or oral lesion as if they were from different patients (Epstein *et al*, 2003).
(8) Analyzing data collected from multiple members of the same family or from students in the same classroom as if the family members and students were unrelated biologically or environmentally. Investigators should also be aware that multiple levels of clustering may occur. For example, analyses of schoolchildren conducted in different schools should take into account clustering of students within a classroom and clustering of classrooms within a school. In laboratory research, clustering effects may occur when animals live in the same cage.
(9) Analyzing a survey of dental practice-based networks that include a variable number of dentists within each practice as if all dental practitioners worked in a solo practice, i.e., ignoring clustering of dentists within the same office when analyzing the data at the dentist level.
(10) Estimating an odds ratio for a variable used in matching for a case–control study.
(11) Performing multiple-testing adjustments that assume test independence when the tests are not independent.

Lastly, there is considerable debate as to whether in hypothesis-driven research one should or should not lock in their statistical methods at the design stage. One argument in favor of choosing a robust method and locking in the statistical methods at the design stage is that, when the statistical analysis plan is changed based on diagnostic testing for assumptions after the data have been collected, both the power and type I error properties of a test become unpredictable because the errors of the diagnostic tests would need to be incorporated into the global errors structure (Shuster, 2005). A frequent reason for changing the statistical plan at the end of the study is the presence of outliers. For example, investigators may change a two-sample *t*-test into a Wilcoxon test when data or residuals are not normally distributed. However, even in this case, certain strategies can be followed to prepare for potential outliers at the design stage, rather than changing the analysis plan (Shuster, 2009).

Results

When preparing the results section of a manuscript, investigators should ensure that results are complete, correct, and consistent. Common examples of missing or incomplete results include:

(1) Not reporting measures of variability of the data, especially standard deviations that accompany mean values or quartiles that accompany medians.
(2) Selectively reporting results based on significance, i.e., omitting results that are not statistically significant (Rifai *et al*, 2008).
(3) Reporting only conclusions or highly condensed summaries of analytic results (e.g., significant/not significant). Reporting only *P*-values also complicates the clinical and biologic interpretation of the study results, because *P*-values depend on sample size. Thus, a small (or large) *P*-value could reflect the large (or small) size of a study rather than a clinically meaningful difference (Young, 2007). For this reason, many journals prefer or require authors to report CI for key comparisons (Altman, 2005).

(4) Not reporting failure to achieve planned accrual. Often, papers simply report the results without acknowledging the low accrual, and its implications. If this happens, the study still needs to be reported. If the reason can be documented as not being based on an interim analysis of the study question, then the bias will be minimal. Conditional power calculations can be made to assess what the results might have been if indeed the study had been run to completion. In any case, the reasons for failure to complete the accrual should be completely disclosed.

Confusion of definitions when interpreting results is also common. The three more frequently misinterpreted statistical terms are risk ratios, correlations, and *P*-value or CI. Examples include:

(1) Interpreting incidence odds, prevalence, prevalence odds, rates, and hazards as if they were equivalent to measuring 'risk' (Slade and Caplan, 1999; Katz, 2006). For example, if investigators measure the prevalence of a very common outcome, such as seropositivity to herpes simplex virus type 1, in older *vs* younger individuals and find that the prevalence odds ratio is equal to two (POR = 2), it would be incorrect to conclude that older individuals have double the risk (or double the prevalence) of seropositivity than younger individuals or that they are twice as likely to be seropositive. First, risk cannot be estimated directly in a cross-sectional study, so investigators could have more appropriately chosen to calculate prevalence ratios (PR), being careful not to interpret them as risk ratios (RR). Secondly, odds and risks are calculated differently: 'risk' is the probability of an event (*P*), whereas 'odds' is the probability of an event divided by the probability of that event not occurring ($P/1 - P$). For this reason, odds ratios will overestimate risk ratios (and PR, which are also ratios of probabilities), especially for common outcomes (i.e., large *P*). Thus, in the example above, POR will overestimate PR, which is a preferable measure less susceptible to misinterpretation.
(2) Incorrectly interpreting the *P*-value as the probability that the hypothesis is true, or interpreting a CI as having a 95% chance or confidence that the true estimate lays within that interval (Poole, 2001).
(3) Using the terms association and correlation as if they were synonymous. Correlation is a specific type of association, but not all associations are correlations.

Inconsistencies in reporting are often due to lack of attention to details. The most common examples include:

(1) Percentages do not add up to 100% after accounting for rounding error.
(2) Counts divided by the total do not correspond to correct percentages. For example: 'Of 20 cases, five (21%) were positive to herpes simplex virus antibodies'.
(3) Results in the text and tables do not match.
(4) Categories are not mutually exclusive (e.g., age categories 0–5, 5–10, and 10–15).
(5) Inadvisable and inconsistent rounding, such as variable, excessive or insufficient use of significant digits.

Discussion

The most common pitfall of the discussion (and conclusion) section is the over-interpretation of the results, such as conclusively claiming that there is an association, that there is no association or that there is or is not a cause–effect relation.

Conclusively claiming that there is no association is virtually never a correct statement and it is particularly inappropriate when sample size is small or when the conclusion is based solely on hypothesis testing from a single study. CI can pin down the set of 'plausible outcomes'. In any case, except for bioequivalence research, a non-significant result is virtually always equated to an inconclusive result, not to a 'no difference' result.

Conclusively claiming that there is an association is particularly inappropriate when based on a single study in which multiple analyses were performed using the same outcome. This may occur in analyses of microarray data and genome-wide studies where the number of analyses can reach several hundred to millions. As the number of analyses increases, the number of false positive associations also increases.

Asserting that there is or is not an effect is stronger than the claim of a positive (or negative) association, because there can be an association without a cause–effect relationship, and there can be an absence of association or an absence of statistical significance when there is a cause–effect relationship. No design is capable of showing unequivocally that there is a cause–effect relationship, although all studies contribute to the scientific basis for assessing that relationship, with the greatest contribution provided by adequately powered, well-designed randomized clinical trials.

Claims of the presence or absence of a cause–effect relation are especially misleading when they are based only on the study at hand and not on the totality of the scientific evidence (Poole *et al*, 2003). Investigators should carefully consider their choice of words in the discussion section, to avoid drawing conclusions that are not supported by the study results. For example, the use of the words 'deterioration' or 'improvement' that imply the observation of an event over time should not be used when progression over time cannot be measured, i.e., if the study is not longitudinal. An 'increase' or a 'decrease' can also only be observed in a longitudinal study, where at least two measurements are made at two different time points.

Lastly, the discussion and conclusion sections should not be used to defend a seriously flawed design or analysis. Minor limitations that could not be completely addressed should be pointed out to the readers, but

major limitations should be addressed directly in the design and analyses rather than acknowledged 'after the fact', when the study has already been completed.

Summary

Oral medicine studies require very careful planning prior to implementation. Design flaws incorporated at the beginning of a study can irreparably damage the validity of the final results. Statistical analyses should be consistent with the research question and design scheme. Results should be presented properly and interpreted cautiously. Investigators with clinical and laboratory expertise but with minimal to no formal or informal statistical background should consider consulting a statistician or an epidemiologist early during the study design phase and before the data are collected. Early discussions among different team members are more likely to identify the most efficient and least biased study design and the most appropriate statistical analyses.

In this article, we have presented a review of some of the potential pitfalls that should be considered during these early discussions and throughout the study. This review is not meant to be used as a comprehensive checklist for all oral medicine research studies, but it represents the consensus of our team at this specific point in time, on what are some of the most commonly encountered pitfalls of research studies that have been proposed are in progress, or have already been completed and submitted for publication (Rothman and Poole, 2007).

Acknowledgements

This work was partially supported by grants M01RR00082 and U54RR025208 from the National Institute of Research Resources, National Institutes of Health, and by grant R21DE018714 from the National Institute of Dental and Craniofacial Research, National Institutes of Health.

Author contributions

L Baccaglini wrote first draft, incorporated changes received from co-authors, reviewed and modified manuscript. JJ Shuster, J Cheng, DW Theriaque, VJ Schoenbach, SL Tomar and C Poole reviewed and modified manuscript.

References

Altman DG (1984). A fair trial? *Brit Med J* **289:** 336–337.

Altman DG (2005). Why we need confidence intervals. *World J Surg* **29:** 554–556.

Altshuler D, Daly MJ, Lander ES (2008). Genetic mapping in human disease. *Science* **322:** 881–888.

Baccaglini L, Schoenbach VJ, Poole C *et al* (2006). Association between herpes simplex virus type 1 and *Helicobacter pylori* in US adolescents. *Oral Surg Oral Med Oral Pathol Oral Radiol Endod* **101:** 63–69.

Baccaglini L, Atkinson JC, Patton LL, Glick M, Ficarra G, Peterson DE (2007a). Management of oral lesions in HIV-positive patients. *Oral Surg Oral Med Oral Pathol Oral Radiol Endod* **103**(Suppl. S50) e1–e23.

Baccaglini L, Brennan MT, Lockhart PB, Patton LL (2007b). World Workshop on Oral Medicine IV: process and methodology for systematic review and developing management recommendations Reference manual for management recommendations writing committees. *Oral Surg Oral Med Oral Pathol Oral Radiol Endod* **103**(Suppl. S3) e1–e19.

Caplan DJ, Slade GD, Gansky SA (1999). Complex sampling: implications for data analysis. *J Public Health Dent* **59:** 52–59.

Chattopadhyay A, Kumar JV, Green EL (2003). The New York State Minority Health Survey: determinants of oral health care utilization. *J Public Health Dent* **63:** 158–165.

Delgado-Rodriguez M, Llorca J (2004). Bias. *J Epidemiol Community Health* **58:** 635–641.

Epstein JB, Zhang L, Poh C, Nakamura H, Berean K, Rosin M (2003). Increased allelic loss in toluidine blue-positive oral premalignant lesions. *Oral Surg Oral Med Oral Pathol Oral Radiol Endod* **95:** 45–50.

Goodson JM, Groppo D, Halem S, Carpino E (2009). Is obesity an oral bacterial disease? *J Dent Res* **88:** 519–523.

Groenwold RH, Van Deursen AM, Hoes AW, Hak E (2008). Poor quality of reporting confounding bias in observational intervention studies: a systematic review. *Ann Epidemiol* **18:** 746–751.

Jacob RF (2002). Bias in dental research can lead to inappropriate treatment selection. *Dent Clin North Am* **46:** 61–78.

Kane RL, Wang J, Garrard J (2007). Reporting in randomized clinical trials improved after adoption of the CONSORT statement. *J Clin Epidemiol* **60:** 241–249.

Katz KA (2006). The (relative) risks of using odds ratios. *Arch Dermatol* **142:** 761–764.

Lesaffre E, Garcia Zattera MJ, Redmond C, Huber H, Needleman I (2007). Reported methodological quality of split-mouth studies. *J Clin Periodontol* **34:** 756–761.

Merchant AT, Pitiphat W (2002). Directed acyclic graphs (DAGs): an aid to assess confounding in dental research. *Community Dent Oral Epidemiol* **30:** 399–404.

Moher D, Simera I, Schulz KF, Hoey J, Altman DG (2008). Helping editors, peer reviewers and authors improve the clarity, completeness and transparency of reporting health research. *BMC Med* **6:** 13.

Patel M, Shackleton JA, Coogan MM, Galpin J (2008). Antifungal effect of mouth rinses on oral Candida counts and salivary flow in treatment-naive HIV-infected patients. *AIDS Patient Care STDS* **22:** 613–618.

Poole C (2001). Low *P*-values or narrow confidence intervals: which are more durable? *Epidemiology* **12:** 291–294.

Poole C, Peters U, Il'yasova D, Arab L (2003). Commentary: this study failed? *Int J Epidemiol* **32:** 534–535.

Rifai N, Altman DG, Bossuyt PM (2008). Reporting bias in diagnostic and prognostic studies: time for action. *Clin Chem* **54:** 1101–1103.

Rothman KJ, Greenland S (1998). Matching. In: Winters R, ed. *Modern epidemiology*. 2nd edn. Lippincott, Williams and Wilkins: Philadelphia, PA, pp. 147–161.

Rothman KJ, Poole C (2007). Some guidelines on guidelines: they should come with expiration dates. *Epidemiology* **18:** 794–796.

Royston P, Altman DG, Sauerbrei W (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med* **25:** 127–141.

Shrier I, Platt RW (2008). Reducing bias through directed acyclic graphs. *BMC Med Res Methodol* **8:** 70.

Shuster JJ (2005). Diagnostics for assumptions in moderate to large simple clinical trials: do they really help? *Stat Med* **24:** 2431–2438.

Shuster JJ (2009). Student t-tests for potentially abnormal data. *Stat Med* **28:** 2170–2184.

Slade GD, Caplan DJ (1999). Methodological issues in longitudinal epidemiologic studies of dental caries. *Community Dent Oral Epidemiol* **27:** 236–248.

Vandenbroucke JP, von Elm E, Altman DG *et al* (2007). Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. *Epidemiology* **18:** 805–835.

VanderWeele TJ, Robins JM (2007). Directed acyclic graphs, sufficient causes, and the properties of conditioning on a common effect. *Am J Epidemiol* **166:** 1096–1104.

Weinberg CR, Sandler DP (1991). Randomized recruitment in case–control studies. *Am J Epidemiol* **134:** 421–432.

Young J (2007). Statistical errors in medical research – a chronic disease? *Swiss Med Wkly* **137:** 41–43.