Some issues in applied statistics in clinical restorative dental research
Tobi, H.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Chapter 8

General discussion
Chapter 8

Introduction

Aim of this chapter is to discuss the three themes of this thesis in the context of methodology of clinical dental research. To reach this, each theme as represented by two chapters, is discussed separately. First, the significance and limitations of the way we dealt with the issues at hand are discussed. Secondly, recent developments are described for each theme. Thirdly, conclusions are drawn and recommendations are made. Finally, in the section "Further research", more general developments are described and priorities for the collaboration between methodology, statistics and dentistry are suggested.

Observer Variation

Significance and limitations

As we pointed out in Chapter 2 "Evaluating the evaluation: Observer variation in the assessment of marginal adaptation", clinical assessments in dental research are usually done by two independent observers. In the literature one may find reports on how observer disagreement is dealt with, but little -if any at all- information is given on the amount and size of observer variation. Both chapters in this thesis on observer variation regard observer variation in the clinical assessment of marginal adaptation. Marginal adaptation is widely accepted and used as an important descriptor of restoration quality. It is shown that for marginal adaptation the percentage of agreement can no longer be regarded sufficient for the description of observer variation. A kappa statistic point estimate of chance-corrected agreement with confidence interval and McNemar’s test is to be preferred. This may be supplemented with log-linear modelling, but in general log-linear modelling adds little when (pairs of) two observers score a series of restorations. Log-linear modelling is very informative though, in the situation where pairs of observers score restorations of different materials. Log-linear modelling allows to take material performance into account and to check for differential observer variation per material (Chapter 3).

We studied the assessment of marginal adaptation but the results can be generalized to other clinical variables that are also assessed on an ordinal or nominal scale with a highly skewed distribution, such as marginal discoloring. Hopefully, our two extensive examples of observer variation in the field of dentistry will stimulate other researchers into giving more information on observer disagreement in their own clinical studies. The limitations of our studies are numerous and stem from the scales, the scoring and the restorations used. With respect to scales: we limited our
General Discussion

studies to marginal adaptation using two different adaptations of the Ryge/USPHS criteria (Ryge, 1980). The standard USPHS criteria were not used so we do not know exactly how our observers would have (dis)agreed on that scale. We note that this seems no longer feasible as the standard USPHS has been altered by various investigators (for example: van Dijken 1996; Geurtsen and Schoeler, 1997) and is no longer used in its original form. A peculiarity of the scoring procedures in our studies is the division of the margin in sections. The study design did not allow for the comparison of assessment outcomes and agreement, between a.) scoring per section taking the worst section as characteristic for the restoration and b.) the scoring of one complete restoration. It would be interesting to investigate whether this procedure of scoring sections is worth the extra effort by increasing agreement compared to the scoring of restorations. In our studies the presence of multiple units within patients is more or less ignored. In Chapter 2 only one restoration at random per patient is included. In Chapter 3 “Observer variation in the assessment of composite resins” it is assumed that the balanced design prevents bias and that with respect to observer agreement the restorations within one patient are independent units.

Recent developments

There is little attention for observer disagreement in clinical dental studies (Chapter 2 of this thesis). This is in contrast to some other clinical fields where the importance of this issue is no longer debatable. Developments in the field of biostatistics are numerous. They can best be categorized in: methods based on agreement beyond-chance (like Cohen’s kappa), log-linear models and latent class models. The discussion on percentage agreement seems resolved: one is discouraged from using percentage agreement as index of agreement. Kappa still undergoes a lot of doctoring. Recent developments involve the proposal of a symmetric matrix with kappa-type coefficients for nominal scales (Kraemer, 1992; Roberts and McNamee, 1998). The diagonal elements are intraclass kappa coefficients of each category relative to all other categories combined (between-subject variance divided by the total variance). The off-diagonal elements investigate confusion between categories. Confusion between categories and the effect of combining categories are very important to analyze when constructing a scale for descriptive and diagnostic purposes because of the validity threat it poses. Log-linear models are especially suited for the analysis of (dis)agreement between more than two observers. Nonetheless, attention is usually limited to pairwise agreement. Rogel, Boëlle and Mary (1998) present a global and partial modelling approach in which higher-order interactions are addressed.
In the early nineties, latent class models were regarded as very promising (Uebersax and Grove, 1990; Agresti, 1992; Guggenmoos-Holzmann, 1993). Recent research on latent class models for the analysis of accuracy of diagnostic tests is extensive. Nonetheless, the textword "latent class" in Medline identifies no application in a dental journal (October 1998). Latent class models have been used for the study of incidence and diagnosis of dental caries but these studies were published in statistical journals (Espeland, Murphy et al., 1988; Formann, 1994). Apparently, the diffusion of latent class models is problematic. There is a reluctance to let go of kappa in favor of log-linear models and, specifically, latent class models.

Cohen’s kappa, log-linear models and latent class models are compared with respect to the concept agreement beyond chance (Guggenmoos-Holzmann and Vonk, 1998). The researchers’ conclusion is that the concept of reliability of assessment has an agreement and an error part to it. These can be represented in one latent class model with a class of systematically consistent and a class of fortuitous ratings. Although attractive from a conceptual point of view, it is doubtful whether this study can lower the revulsion for latent class models.

Conclusions and recommendations

The conclusion is that observer agreement in the assessment of marginal adaptation has to be improved, with respect to the actual agreement as well as the report of observer variation. The latter is easily achieved by including information on the number of observers and their disagreement. In most instances a kappa coefficient with confidence interval and McNemar’s test statistic will suffice. When the research question contains some sort of condition of agreement on other variables, such as dental material used, log-linear modelling can be very informative.

Often researchers report almost 100% rating alpha for marginal adaptation or they use their own alteration of the USPHS scale. Hence, the question arises whether the USPHS criteria need improvement to better suit the need of clinical researchers. In this discussion the predictive value of marginal adaptation needs to be addressed in such a way that the criteria developed have a predictive value in clinical practice too.

In clinical studies, information on observer agreement is indispensable to ensure the reader of a successful calibration of observers. It would be inappropriate and not feasible to expect the clinical researcher and consultant statistician to go to such lengths as would be defensible in a reliability study. After all it is the research question that should steer the data analysis and not the range of available statistical methods. Hence, the importance of straightforward procedures with interpretable results should not be underrated.
Survival analysis

Significance and limitations

In Chapter 4, on the analysis of restoration survival data in split-mouth designs, we described and compared different ways of handling data obtained in such a design. Logistic regression with a random component and Friedman's statistic were compared with the classic Kaplan-Meier estimator which ignores the multiple units within patient structure. It was shown that restorations within patients cannot be regarded as independent units. Since Friedman's statistic suffers from relatively low power and lack of an interpretable risk estimate, the logistic regression with random component approach was preferred and this latter method was used for the analysis of several controlled clinical trials comparing amalgam treatment modalities (Chapter 5).

The limitations of this approach are evident: The uni-dimensional use of the data and the exclusion of patients without complete follow-up. Survival data are two-dimensional: the status (the restoration being a true failure, false failure or censored) and the life span as minimally (in the presence of censored data) observed. In a logistic regression approach the outcomes are reduced to status after a fixed period of time. This method can neither handle patients lost to follow-up nor false failures.

The patients for whom no complete follow-up data are available are excluded from the analysis. When many patients drop out during follow-up the logistic regression with random component approach becomes unusable in its current format. After all, patient drop-out is an important source of bias and if all information available of these patients is excluded from the analysis, the bias may be substantial. In part, this could be dealt with by estimating the odds for replacement for a range of time intervals. In the study design described here, it might be worth considering the odds for replacement within five years, the odds for replacement between five and ten years, and the odds for replacement between ten and fifteen years separately. Some sort of conditioning on surviving the former time-interval could then result in the odds for replacement within 15 years using the available information on people who did not complete the study.

Recent developments

Longevity studies require the follow-up of numerous patients over a substantial number of years. Even after a follow-up of ten years or more, it often is impossible to estimate the median survival time of dental restorations reliably because of the few replacements (Letzel, van't Hof, et al., 1997; Chapter 5).
Hence it is interesting to watch the development of the concepts ‘surrogate endpoint’ and ‘auxiliary endpoint’. A surrogate endpoint has been defined as a response variable that can substitute the ‘true’ endpoint for the purpose of comparing specific treatments in a clinical trial (Prentice, 1989). An auxiliary endpoint is a response variable that can strengthen the ‘true’ endpoint analysis (Fleming, Prentice et al., 1994). The idea is that by establishing the relation between the ‘true’ endpoint (re-restoration or removal of the restoration) and the surrogate or auxiliary endpoint, the size and length of a clinical trial can be shortened. Restoration quality characteristics could perhaps be combined into such an auxiliary variable. Much longitudinal research is needed before this is within reach.

The structure of the data in a split-mouth design can be regarded hierarchical. The restorations within patients are a sample (of restorations) within a sample (of patients) of a sample (of dentists). Models especially suited for data resulting from this multistage sampling are called multilevel models. The multilevel approach is making its entrance into medical statistics from educational measurement (Bryke and Raudenbusch, 1992; Goldstein 1995). Extensions of the multilevel approach to survival data are being formulated and further developed, but the presence of censored data causes problems. Consequently, there is yet no multilevel analysis software available that can handle survival data.

A lot of conceptual and mathematical work is currently being done on frailty models. Frailty is a term used for unobserved heterogeneity. A frailty survival model is a random effect model where the random effect (the frailty) has a multiplicative effect on the hazard (Aalen, 1994; Hougaard, 1995). Translated in a restorative dentistry setting: in the one unit per patient situation, the frailty would describe the influence of unobserved risk factors. In the multiple restorations within patient situation, the frailty would reflect the neglected commonalities that account for the dependence of the units. Conditional on the frailty, the multiple units are independent. One of the interesting preliminary questions would be what frailty distribution best reflects the actual within patient correlation.

Conclusions and recommendations

Survival analysis for correlated data obtained in a split-mouth design remains cumbersome. Survival of restorations within patients are not to be regarded as independent units. The performance of these restorations correlate at least with regard to durability. Since so many restorations last beyond the scope of a feasible follow-up in the context of a controlled clinical trial, the use of surrogate or auxiliary endpoints is tempting. These, however, can never fully replace the real event of interest: replacement of the restoration. Hence,
techniques that can handle scarce events need to be developed or existing techniques need to be further refined.
In the mean time, Kaplan-Meier curves can be used for descriptive purposes. To evaluate risk factors and treatment modalities, logistic regression modelling with a random component can be used - be it with caution: it is an easy to use device with clearly identified drawbacks.

Cost-effectiveness analysis

Significance and limitations

The increasing concerns over costs of public health services have resulted in increasing importance of 'value for money' considerations in the allocation of resources. Nonetheless, economic evaluations in dentistry are relatively scarce, and the quality of many of the available studies is rather poor (Elixhauser, Halpern et al., 1998). Hence, we felt the need for a paper like Chapter 6 'An outline of cost-effectiveness analysis in dentistry' that explains and appraises guidelines for economic evaluation proposed by the BMJ Economic Evaluation Working Party (Drummond and Jefferson, 1996). We concluded that a supplement is necessary to ensure due attention for ethical, methodological and statistical issues and suggested the inclusion of a guideline on the statistical analysis. The need for more statistical advice has been recently confirmed by the results of a review of 45 randomized clinical trials by Barber and Thompson (1998) who report a lack of statistical awareness.

In Chapter 7 cost-effectiveness of composite resins and amalgam for the rerestoration of old amalgam Class II restorations is investigated. The rerestoration of such old restorations will remain a common procedure for quite some years. No prior cost-effectiveness analysis has been published on this issue. In our study, no significant difference in effectiveness could be found within five years after placement: a full 100% survival was observed for both amalgam and composite resins restorations. And although the quality of marginal adaptation indicated slightly in favor of composite resins, the need for repair indicated in favor of amalgam. Using treatment time as a proxy of treatment costs, the conclusion was that rerestoration with composite resins more or less doubles the costs compared to rerestoration with amalgam from the perspective of the field of dentistry.

The major limitation of this economic evaluation is the limited perspective. The amalgam versus composite debate raises a lot of awareness and emotions. This makes information on utilities and willingness-to-pay for composite resin restoration instead of amalgam restorations indispensable for decision makers.
Recent developments

As in any other type of research, a certain number of patients is needed to warrant sufficient power and accurate estimates. Briggs and Gray (1998) explore power analysis techniques for deriving desired sample size assuming independence between costs and effects. In practice, however, costs and effectiveness are likely to be associated so a technique based on the joint distribution of costs and effects is more precise. Such a technique has been shown technically possible in a simulation study, but one should question the usefulness because too much prior information is required (AI, van Hout, et al., 1998). A practical approach is to calculate the desired sample size for costs and effects separately, and take the largest of the two. Including more patients in the economic analysis than would be necessary for an efficacy trial can cause an ethical problem (O’Brien, Drummond et al., 1994; Briggs and Gray, 1998).

The aim of economic evaluation is to help decision makers. Local differences between unit costs, incidence, etceteras are likely and this complicates the use of published studies. After all, these local differences have an impact on the expected effectiveness gain or costs reduction. To improve the use of published economic evaluations, local re-analysis is promoted (Bryan and Brown, 1998). The feasibility of a local re-analysis, however, depends on the authors of the economic evaluation to be explicit. This explains to some extent the interest in improving economic reports (Mason and Drummond, 1995; Nuijten, Pronk et al., 1998; Jefferson, Smith et al., 1998).

Recently, a new framework has been proposed which is called the ‘Net health benefits’ approach (Stinnett and Mullahy, 1998). Here, the index of interest is not a ratio but a difference. The net health benefit (NHB) of an intervention is interpreted as the net effectiveness gained by investing resources in this intervention rather than investing in a marginally cost-effective program. An advantage of this approach over the cost-effectiveness ratio is that a negative difference is easier to interpret than a negative ratio. It is also easier to estimate a confidence interval for a difference than for a ratio.

Conclusions and recommendations

In restorative dentistry one often faces the difficulty of multiple outcome variables. The advice is to use oral health related quality of life measures for summary measures (Slade, Strauss, et al., 1998), although additional work is needed to make them better suited for policy decision making (Weintraub, 1998).

Economic evaluation in dentistry is a developing area where different disciplines meet; dentistry, economics as well as epidemiology and statistics.
General Discussion

The quality of economic evaluations in dentistry can benefit from guidelines such as developed by Drummond and Jefferson (1996) and amended in Chapter 6. By following these guidelines most issues will be satisfactorily addressed. However, the report of a cost-effectiveness or cost-utility ratio with confidence interval is unlikely to become standard due to all sorts of problems. In time, the NHB may demonstrate to meet practical and theoretical demands and replace the ratios.

Further research

As mentioned in the introduction of this thesis, the methodology of good clinical research encompasses a range of issues from the design of the study to the report of results. The quality improvement of clinical research is a partly cyclic process. To evaluate the effectiveness of two types of restorative treatments, for example, one needs information on longevity and valid quality characteristics. To know which quality characteristics to look at, one needs longitudinal research that evaluates the prospective value of clinical criteria. The construction of such evaluation instruments involves issues of validity and reliability. And so on and so forth.

There are three major issues which will effect the process of improving quality of clinical restorative research as well as the research questions that evolve from dental practice.

The introduction of fluoride dentifrices and the improvement of dental hygiene in the general population results in less decayed, missing or filled teeth as demonstrated in most parts of the western world (Pine, 1997). This decreasing prevalence of caries also means that less cavities will become available for research. In general, one will have to deal with the available data much more efficiently: less restorations will have to yield at least the same level of information on functionality, longevity and so on. Furthermore, the subgroups with relatively poor dental health will yield information for treatment effectiveness, but these subgroups may not be representative for the population at large.

The aging population in combination with improved oral health also has implications for research in restorative dentistry. Because of the peculiarity of the dental status in the elderly, many new questions arise. For example, can we consider longevity of the restoration the main determinant of restoration functionality in the elderly as we do in the younger adult population? What are the main predictive factors in this population for the prognosis of restoration functionality?

Last but not least, there is the changing attitude of the recipient of dental care: the patient. In general, the patient has become a critical consumer of care. Consequently, the patient perception of and patient satisfaction with the
restorative treatment is becoming a new dimension of treatment outcome, additional to treatment results as perceived by the dental profession. In combination with the growing importance of efficient resource allocation this calls for research into willingness-to-pay. Concluding, one can think of an overfilled agenda for the further improvement of clinical restorative research which feasts dentists and methodologists/statisticians with numerous interesting issues. The focus should be on how we can answer clinical research questions by obtaining data of as high a quality as possible and deal with these data as efficiently as possible. In particular, the development of study design needs further attention because the study design determines to a great extent whether the data analysis can answer the questions. Consequently, clinical researchers ought to become better aware of the quality gain that results from the early inclusion of a statistician and methodologist in the research team. We - consultant statisticians and methodologists - should refrain from indulgence in the mathematical and numerical work this offers, and aim to fully understand the practical needs of the dentist researcher.

References


