Editorials

Problems of multiplicity

Penelope M.A. Brasher PhD,* Rollin F. Brant PhD†

THE use of many significance tests in a single study, e.g., analyses of multiple endpoints, multiple treatment groups, repeated analyses on accumulating data, repeated measurements over time, and subgroup analyses, can increase the probability of obtaining false positive results. Many articles have appeared in medical journals over the years cautioning against multiple testing; Lang and Secic in their book "How to Report Statistics in Medicine", devote a chapter to the problem. However, this sage advice continues to go unheeded by many – a quick review of the research articles published in the *Journal* in 2007 yields several examples of "*P*-ing all over the paper"[1] including an article with twice as many *P*-values as subjects! In this editorial, we briefly describe the problem, and provide some guidance to investigators when reporting research studies.

The problem

If one significance test at level α is performed, the probability of rejecting the individual null hypothesis, although it is in fact true, is $\alpha \times 100\%$. However, if k *independent* tests are performed, each at α the probability of rejecting at least one of the k null hypotheses, *when in fact all are true*, is $1-(1-\alpha)^k$. This later probability is known as the *experimentwise error rate* (EER), or *familywise error rate* (FWER). Thus, for example, if five tests are performed, each at $\alpha = 0.01$, the EER is $1-(1-.01)^5 = 4.9\%$.

Definitions:
type 1 error (α): declaring that a difference exists when it does not.

type 2 error (β): declaring that a difference does not exist when in fact it does.

The preceding paragraph is instructive, as it touches on two of the reasons for the controversy around whether or not to adjust for multiple testing in some situations – the tests are seldom independent, and the global null hypothesis is rarely of interest. However, these do not argue against adjustments for multiple testing, but rather that either appropriate adjustment is made, or appropriate care is taken in the interpretation of unadjusted analyses.

Multiple primary endpoints

In the Utopian world of health research, there would be a single primary endpoint that captures all of the clinically relevant benefits of an intervention in a particular patient population. The demonstration of clinically important changes on this endpoint would lead to changes in practice. However, in some cases, there will be no single endpoint, and the assessment of multiple endpoints will be unavoidable. The appropriate method to adjust for multiplicity will depend on the relevant clinical question.

1) If statistical significance needs to be demonstrated on all of the endpoints, then no adjustment for multiplicity is required, as the EER will be at most equal to the nominal α-level. However, investigators should be aware that the type 2 error rate will be inflated (i.e., power reduced), and one needs to ensure an adequate sample size at the planning stage of the study.

2) Where a therapeutic effect will be declared if one or more of the endpoints is statistically significant at the nominal α-level, then adjustment will need to be made to control the EER. A simple solution is to employ the well-known Bonferroni method.[2] Each test is conducted at α/k, where k is the number of individual tests. However, this strategy can be highly

CAN J ANESTH 2008 / 55: 5 / pp 259–264

From the Centre for Clinical Epidemiology and Evaluation,* Vancouver Coastal Health Research Institute; and the Department of Statistics,† University of British Columbia and Centre for Community Child Health Research Child and Family Research Institute, Vancouver, British Columbia, Canada.

Address correspondence to: Dr. Penelope M.A. Brasher, Centre for Clinical Epidemiology and Evaluation, Vancouver Coastal Health Research Institute, 828 West 10th Avenue, Vancouver, British Columbia V5Z 1L8, Canada. Phone: 604-875-4111, ext 68167; Fax: 604-875-5179 ; E-mail: brasher@interchange.ubc.ca

conservative, i.e., it overcorrects the type 1 error inflation in many situations. Bender and Lange[3] state that the Bonferroni method is a reasonable approach when the number of tests is small (< 5) and the correlations among the test statistics are low. There are a number of modified Bonferroni methods that can be used that are less conservative (more powerful), but continue to control the EER. These include the procedures developed by Holm, Hochberg and Hommel (see reference 3 for details). Although more powerful than the simple Bonferroni method, if the correlations between the endpoints are moderately large (> 0.5) these methods are also conservative. Adjusted *P*-values from re-sampling procedures, which take into account the correlation structure of the test statistics, will be more powerful.[4]

3) If what is of interest is to demonstrate an overall therapeutic benefit based on a number of correlated endpoints, then a method employing a global test statistic such as O'Brien's GLS test[5] may be appropriate. However, one must be cautious, as global tests are best used in situations where the endpoints provide alternative measures of the same fundamental quantity, and one expects the size of the effect to be consistent across endpoints, AND where the individual endpoints are not of interest.[6] Such was the case in the first trials of tissue plaminogen activator (t-PA) in stroke, that considered the simultaneous effect of t-PA on four stroke disability scales.[7]

Multiple treatment groups

When three or more groups are being compared, the general procedures based upon adjusted *P*-values described above (1 & 2) can be employed. These general multiple test procedures can be used for any type of data (binary, categorical, continuous, time to event) and any test statistic (Fisher, χ^2, *t*, etc.). However, if the endpoint is continuous, the usual analysis of variance (ANOVA) and the associated simultaneous test procedures may be preferred. Selecting which of the many available *post hoc* procedures to use will depend on the comparisons of interest.[8] For example, Dunnett's test will be the most powerful if several groups are being compared to a single control. Tukey's test can be used if all pair wise comparisons are of interest.

One cautionary note, the omnibus *F*-test (is there *any* difference between the groups?) is seldom of clinical interest. Thus, in the planning stage of the study, it is important to ensure that the sample size is sufficient to ensure adequate power for the comparisons between treatment groups that are of interest, including accommodation of any inflation in the type 1 error rate.

Repeated analyses on accumulating data (interim analyses)

Repeated testing on accumulating data inflates the type 1 error and multiplicity adjustments are *required* to develop adequate early stopping rules. The reader is referred to the excellent *Lancet* article by Schulz and Grimes[9] for an introduction to interim analyses, including methods that should be employed to protect against an inflated type 1 error rate.

Repeated measures over time

When repeated measurements of the same endpoint are made over time, a common strategy is for investigators to compare the groups at each time point, by using *t* tests. Similar to the problem of multiple endpoints, this approach will only be useful if the number of time points is limited (≤ 3), the intervals between them are large, and each of the time points is of interest in its own right. In such cases, one could control the type 1 error rate with one of the general multiple test procedures, although the adjustments will be too conservative since the repeated measures on an individual are likely to be highly correlated.

Two alternative approaches are the summary measures approach,[10] and regression modeling.[11] In the summary measures approach, the sequence of repeated measures for an individual is reduced to a single summary measure [e.g., rate of change, area under the curve, etc.] then standard univariate statistical methods can be applied to the derived measures. In many cases, however, the number of repeated measures on each subject may not be the same, or they may not be taken at the same interval, or observations may be missing. In such cases, the summary measures approach may not be valid, and the regression modeling approach will be preferred.[11]

Subgroup analyses

Subgroup analyses are often conducted to assess whether or not there are differences in the effect of interest across subgroups of patients. The issue of subgroup analyses was the topic of a "Special Report" in the November 22, 2007 edition of the *New England Journal of Medicine*.[12] The report summarizes common problems with the conduct of subgroup analyses, including problems of multiplicity, and provides guidelines for reporting subgroup analysis. With regard to multiplicity, the authors suggest:

"Indicate the potential effect on type 1 errors (false positives) due to multiple subgroup analyses and how this effect is addressed. If formal adjustments for multiplicity were used, describe them; if no formal adjustment was made, indicate the magnitude of the problem informally".

The authors provide no guidance as to when formal adjustment is necessary. Our view is similar to other authors,[9,12,13] in that subgroup analyses should be undertaken with extreme caution, and for the most part, should be considered exploratory in nature. In such cases, an informal indication of the magnitude of the problem will be sufficient.

Safety analyses

Protection against inflation in the type 1 error rate is less of a concern when considering the analysis of adverse events, as the focus is on not committing a type 2 error, particularly for serious adverse events. A strategy that seems reasonable is to perform individual unadjusted analyses for serious adverse events, and to apply a multiple testing procedure to the remainder.

Summary

We have discussed a number of procedures that mitigate against the "malady" of multiple comparisons. They often do so, however, at the cost of decreased power or increased sample sizes. Prevention at the stage of study design, when feasible, is the best policy. Primary and secondary objectives should be carefully delineated. For outcomes measured over time, the most relevant features of response trajectories (for example, area under the curve) should be clearly identified. In statistical analysis plans, more emphasis should be placed on describing the key features of the data in clinically relevant terms, and researchers should resist the practice of stamping every possible comparison with the label of significance.

Problèmes de multiplicité

L'utilisation de nombreux tests d'hypothèse dans une seule étude, par exemple les analyses de points d'aboutissement multiples, de groupes de traitement multiples, les analyses répétées sur des données accumulées, les mesures répétées dans le temps, ou les analyses de sous-groupes, peuvent augmenter la probabilité d'obtenir des résultats faux positifs. Au fil des années, nombre d'articles ont été publiés dans les revues médicales pour mettre en garde les chercheurs contre les tests multiples ; Lang et Secic, dans leur ouvrage « How to Report Statistics in Medicine », consacrent en effet un chapitre entier à ce problème. Pourtant, ce conseil avisé est resté lettre morte auprès de plusieurs chercheurs, comme le démontre un survol rapide des articles de recherche publiés par le *Journal* en 2007 : on dénombre plusieurs exemples de « *P*-lifération partout dans l'article » (en anglais, « *P*-ing all over the paper »)[1], notamment un article comprenant deux fois plus de valeurs *P* que de sujets ! Dans cet éditorial, nous décrivons brièvement ce problème et proposons quelques recommandations aux chercheurs qui désirent rapporter les résultats de leurs recherches.

Le problème

Si un test d'hypothèse au niveau α est effectué, la probabilité de rejet de l'hypothèse nulle individuelle, bien qu'elle soit en fait vraie, est $\alpha \times 100$ %. Cependant, si des tests *k indépendants* sont effectués, chacun à α la probabilité de rejeter au moins une des hypothèses nulles *k*, *alors qu'en fait toutes sont vraies*, est de $1-(1-\alpha)^k$. Cette probabilité ultérieure est connue sous le nom de *taux d'erreur lié à l'expérience* (EER – 'experimentwise error rate'), ou *taux d'erreur lié à la famille* (FWER – 'familywise error rate'). Ainsi par exemple, si cinq tests sont effectués, chacun à α = 0,01, l'EER est de $1-(1-,01)^5$ = 4,9 %.

Définitions:

Erreur de type 1 (α): affirmation qu'une différence existe alors qu'elle n'existe pas.

Erreur de type 2 (β): affirmation qu'une différence n'existe pas alors qu'en fait elle existe.

Le paragraphe précédent est très instructif étant donné qu'il aborde une des deux raisons de la controverse existant autour de la correction ou non des données lors de tests multiples dans certaines situations ; ces tests sont rarement indépendants, et l'hypothèse nulle totale est rarement digne d'intérêt. Toutefois, ces données ne vont pas à l'encontre des corrections pour effectuer des tests multiples ; il s'agit plutôt de soit faire une correction adéquate, soit faire suffisamment attention lors de l'interprétation de données non corrigées.

Points d'aboutissement primaires multiples

Dans le monde utopique de la recherche en santé, il n'y aurait qu'un seul point d'aboutissement primaire qui permettrait d'appréhender tous les bénéfices pertinents, d'un point de vue clinique, d'une intervention sur une population de patients donnée. La démonstration des modifications importantes au niveau clinique sur ce point d'aboutissement conduirait à des change-

ments de la pratique. Toutefois, dans certains cas, il n'y aurait pas de point d'aboutissement unique, et l'évaluation de points d'aboutissement multiples serait inévitable. La bonne méthode à utiliser afin de corriger la multiplicité dépendra de la question cliniquement pertinente.

1) Si une signification statistique pour tous les points d'aboutissement doit être démontrée, il n'est pas nécessaire de procéder à des corrections pour la multiplicité des points d'aboutissement, étant donné que l'EER sera tout au plus égal au niveau α nominal. Cependant, les chercheurs devraient être conscients que le taux d'erreurs de type 2 sera plus élevé (en d'autres termes, la puissance sera réduite), et il faut s'assurer dès la phase de planification de l'étude qu'il y aura une taille d'échantillon adaptée.

2) Dans les cas où un effet thérapeutique sera confirmé si un ou plusieurs des points d'aboutissement est statistiquement significatif au niveau α nominal, il faudra alors effectuer une correction pour contrôler l'EER. Une solution simple consiste à utiliser la méthode de Bonferroni.[2] Chaque test est effectué à α/k, où k constitue le nombre de tests individuels. Cette stratégie peut toutefois s'avérer très conservatrice, c'est-à-dire qu'elle sur-corrige l'augmentation d'erreurs de type 1 dans de nombreuses situations. Selon Bender et Lange,[3] la méthode de Bonferroni est une approche raisonnable lorsque le nombre de tests est restreint (< 5) et les corrélations entre les tests statistiques sont peu nombreuses. Il existe plusieurs modifications de la méthode de Bonferroni qui sont moins conservatrices (plus puissantes) et qui peuvent être utilisées tout en continuant à contrôler l'EER. Les procédures développées par Holm, Hochberg et Hommel (voir référence 3 pour plus de détails) font partie de telles méthodes. Bien que plus puissantes que la méthode de Bonferroni simple, si les corrélations entre les points d'aboutissement sont modérément nombreuses (> 0,5), ces méthodes sont alors également conservatrices. Les valeurs *P* ajustées tirées de procédures de ré-échantillonnage tenant compte de la structure de corrélation de ces variables auxiliaires seront alors plus puissantes.[4]

3) Si l'étude cherche à démontrer un bienfait thérapeutique global sur la base de plusieurs points d'aboutissement corrélés, une méthode employant une variable auxiliaire globale comme le test GLS d'O'Brien[5] pourrait être adéquate. Il faut toutefois être prudent, étant donné que les tests globaux sont utilisés de façon optimale dans les cas où les points d'aboutissement procurent des mesures alternatives de la même quantité fondamentale, et qu'il est prévu que la taille de l'effet coïncide à tous les points d'aboutissement, ET lorsque les points d'aboutissement individuels ne sont pas pertinents en soi.[6] Tel était le cas dans les premières études traitant de l'activateur tissulaire du plasminogène (t-PA) lors d'accidents cérébraux vasculaires qui étudiaient l'effet simultané du t-PA sur quatre échelles d'incapacité d'accidents cérébraux vasculaires.[7]

Groupes de traitement multiples

Lorsque trois groupes ou plus sont comparés, les procédures générales basées sur des valeurs *P* ajustées décrites ci-dessus (1 & 2) peuvent être utilisées. Ces procédures générales de test multiple peuvent être utilisées pour n'importe quel type de données (binaires, catégoriques, continues, temps jusqu'à l'événement) et toute variable auxiliaire (Fisher, χ^2, *t,* etc.). Cependant, si le point d'aboutissement est continu, l'analyse de variance (ANOVA) usuelle et les procédures de test simultanées associées devraient être privilégiées. La sélection de l'une des nombreuses procédures ultérieures dépendra des comparaisons pertinentes à l'étude.[8] Par exemple, le test de Dunnett sera le plus puissant si plusieurs groupes sont comparés à un seul groupe témoin. Le test de Tukey peut être utilisé si toutes les comparaisons par paire sont pertinentes à l'étude.

Un avertissement toutefois : les tests-F dits omnibus (y a-t-il une *quelconque* différence entre les groupes ?) ne représentent que rarement un intérêt clinique. Au cours de la phase de planification de l'étude, il est dès lors important de s'assurer que la taille de l'échantillon est suffisamment grande afin de garantir une puissance adéquate pour effectuer les comparaisons entre les groupes de traitement pertinents, y compris les ajustements à toute augmentation du taux d'erreurs de type 1.

Analyses répétées sur des données accumulées (analyses intermédiaires)

L'analyse répétée effectuée sur des données en cours de collecte augmente les erreurs de type 1 et des corrections de multiplicité sont *requises* afin de développer des règles d'arrêt précoce adaptées. Nous renvoyons le lecteur à l'excellent article de Schulz et Grimes[9] pour une introduction aux analyses intérimaires, comprenant des méthodes à utiliser pour se prémunir contre un taux d'erreurs de type 1 plus élevé.

Mesures répétées dans le temps

Lors de mesures répétées dans le temps du même point d'aboutissement, une stratégie courante consiste à comparer les groupes à chaque point dans le temps à l'aide de tests *t*. Tout comme le problème des points

d'aboutissement multiples, cette approche ne sera utile que si le nombre de points dans le temps est restreint (≤ 3), si les intervalles entre eux sont importants et si chacun des points dans le temps est pertinent en soi. Dans de tels cas, le taux d'erreurs de type 1 pourrait être surveillé à l'aide de l'une des procédures générales de test multiple, malgré le fait que les corrections seront trop conservatrices étant donné que les mesures répétées sur un individu risquent d'être considérablement corrélées.

Deux autres approches possibles sont l'approche par mesures agrégées,[10] et le modèle de régression.[11] Lors d'une approche par mesures agrégées, la séquence de mesures répétées pour un individu donné est réduite à une mesure agrégée unique (par ex., taux de changement, zone inférieure à la courbe, etc.) ; ensuite, des méthodes standard d'analyse univariée peuvent être appliquées aux mesures dérivées. Dans de nombreux cas pourtant, le nombre de mesures répétées sur chaque sujet peut ne pas être le même, ou les mesures peuvent ne pas être prises au même intervalle, ou certaines observations peuvent manquer. Dans de tels cas, l'approche par mesures agrégées pourrait ne pas être valable, et l'analyse par modèle de régression sera alors à privilégier.[11]

Analyses de sous-groupes

Les analyses de sous-groupes sont souvent effectuées pour évaluer s'il existe ou non des différences dans l'effet à l'étude entre différents sous-groupes de patients. La question des analyses de sous-groupes a fait l'objet d'un « Rapport spécial » dans l'édition du 22 novembre 2007 du *New England Journal of Medicine*.[12] Ce rapport résume les problèmes courants rencontrés lors d'analyses de sous-groupes, notamment les problèmes de multiplicité, et fournit des recommandations pour rapporter des analyses de sous-groupe. En ce qui a trait à la multiplicité, l'auteur suggère :

« Indiquez l'effet potentiel sur les erreurs de type 1 (faux positifs) provoqué par des analyses multiples de sous-groupe et la manière dont vous abordez cet effet. Si des corrections formelles pour contrer la multiplicité ont été utilisées, décrivez-les ; si aucune correction formelle n'a été faite, indiquez de manière informelle la magnitude du problème. »

Par contre, les auteurs ne proposent pas de recommandations quant aux cas dans lesquels une correction formelle est requise. Notre position rejoint celle d'autres auteurs,[9,12,13] à savoir que les analyses de sous-groupe devraient être effectuées avec prudence et, dans la plupart des cas, considérées comme étant de nature exploratoire. Dans de telles situations, une indication informelle de la magnitude du problème suffira.

Analyses de sécurité

La protection contre une augmentation du taux d'erreurs de type 1 n'est que peu pertinente lorsqu'on analyse les événements indésirables, étant donné qu'il s'agit surtout de ne pas commettre d'erreurs de type 2 et ce, particulièrement dans le cas d'événements indésirables graves. Une stratégie apparemment raisonnable consiste à effectuer des analyses individuelles non corrigées dans les cas d'événements indésirables graves, et d'appliquer une procédure de test multiple aux autres analyses.

Résumé

Nous avons présenté le pour et le contre de plusieurs procédures qui pourraient soigner la « maladie » des comparaisons multiples. Ces approches y parviennent néanmoins au prix d'une puissance réduite ou d'une taille d'échantillon accrue. La prévention dès la phase de conception d'une étude demeure, lorsqu'elle est possible, la meilleure tactique. Les objectifs primaires et secondaires devraient être soigneusement définis. En ce qui concerne les résultats mesurés à plusieurs points dans le temps, les aspects les plus pertinents des trajectoires de réponses (par exemple, l'aire sous la courbe) devraient être clairement identifiés. Dans les plans d'analyse statistique, il faudrait se concentrer davantage sur la description des caractéristiques clés des données en termes de pertinence clinique. Enfin, les chercheurs devraient résister à la tentation de coller l'étiquette du « statistiquement significatif »à chaque comparaison possible.

References

1 *Lang TA, Secic M.* How To Report Statistics in Medicine: Annotated Guidelines for Authors, Editors, and Reviewers. Philadelphia: American College of Physicians; 2006: 63.

2 *Bland JM, Altman DG.* Multiple significance tests: the Bonferroni method. BMJ 1995; 310: 170.

3 *Bender R, Lange S.* Adjusting for multiple testing – when and how? J Clin Epidemiol 2001; 54: 343–9.

4 *Westfall PH, Young SS.* Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment. NY: Wiley; 1993.

5 *O'Brien PC.* Procedures for comparing samples with multiple endpoints. Biometrics 1984; 40: 1079–87.

6 *Sankoh AJ, Huque MF, Dubey SD.* Some comments on frequently used multiple endpoint adjustment methods in clinical trials. Stat Med 1997; 16: 2529–42.

7 *Tilley BC, Marler J, Geller NL, et al.* Use of a global test for multiple outcomes in stroke trials with application to the National Institute of Neurological Disorders and Stroke t-PA Stroke Trial. Stroke 1996;

27: 2136–42.
8 *Ramsey FL, Schafer DW*. The Statistical Sleuth: a Course in Methods of Data Analysis. CA: Duxbury Press; 2002.
9 *Schulz KF, Grimes DA*. Multiplicity in randomised trials II: subgroup and interim anslyses. Lancet 2005; 365: 1657–61.
10 *Matthews JN, Altman DG, Campbell MJ, Royston P*. Analysis of serial measurements in medical research. BMJ 1990; 300: 230–5.
11 *Fitzmaurice GM, Laird NM, Ware JH*. Applied Longitudinal Analysis. NJ: Wiley; 2004.
12 *Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM*. Statistics in medicine – reporting of subgroup analyses in clinical trials. N Engl J Med 2007; 357: 2189–94.
13 *Assmann SF, Pocock SJ, Enos LE, Kasten LE*. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 2000; 355: 1064–9.