Simultaneous and selective inference: Current successes and future challenges

Yoav Benjamini*

Department of Statistics and Operations Research, Sackler School of Mathematical Sciences, Tel Aviv University, Tel Aviv, Israel

Received 1 December 2009, revised 20 June 2010, accepted 12 August 2010

The previous decade can be viewed as a second golden for era Multiple Comparisons research. I argue that much of the success stems from our being able to address real current needs. At the same time, this success generated a plethora of concepts for error rate and power, as well as multiplicity of methods for addressing them. These confuse the users of our methodology and pose a threat. To avoid the threat, it is our responsibility to match our theoretical goals to the goals of the users of statistics. Only then should we match the methods to the theoretical goals. Considerations related to such needs are discussed: simultaneous inference or selective inference, testing or estimation, decision making or scientific reporting. I then further argue that the vitality of our field in the future – as a research area – depends upon our ability to continue and address the real needs of statistical analyses in current problems. Two application areas offering new challenges have received less attention in our community to date are discussed. Safety analysis in clinical trials, where I offer an aggregated safety assessment methodology and functional Magnetic Resonance Imaging.

Key words: Aggregated safety analysis; False discovery rates; Familywise error rate; Functional magnetic resonance imaging; Multiple comparisons procedures.

1 An illustrated concise history of the area

The beginning was in the 1950's. Following the work of Tukey and Scheffe there was a burst of interest in the problem of Multiple Comparisons, an interest reflected in the increase of the number of methodological papers that appeared in the four leading journals (*Biometrika*, *Journal of the American Statistical Association*, *Journal of the Royal Statistical Society B* and *The Annals of Statistics*). Figure 1 displays the time series of this proportion as estimated from a computerized literature search, from 1965 till 2008. This increase lasted for about two decades and terminated with a decline. In 1979, Prof. John Tukey was asked, after giving a talk on Data Analysis and Robustness, what is his opinion about the future of Multiple Comparisons, his blunt response was that ''the field is dead'' (as communicated by Prof. Fuchs).

In spite of the decline the field was not dead. In fact what kept the field alive is a steady stream of new ideas generated by a committed community of researchers that came in response to the needs of medical statistics: multiple looks, multiple arms, two stage designs, bioequivalence and dose response studies, to name a few application needs; the closure principle (Marcus *et al.*, 1976), followed by stepwise procedures such as Holm's (1979) to name just a few principles.

Then, from the deep of 1986 the trend changed upwards. The buildup of ideas described above, and the books by Hochberg and Tamhane (1987), Westfall and Young (1993) and Hsu (1996) had their strong impact, I believe. The work on the False Discovery Rate (FDR) and related ideas kept the

*Corresponding author: e-mail: ybenja@tau.ac.il, Phone: +972547975311, Fax: +97236409357

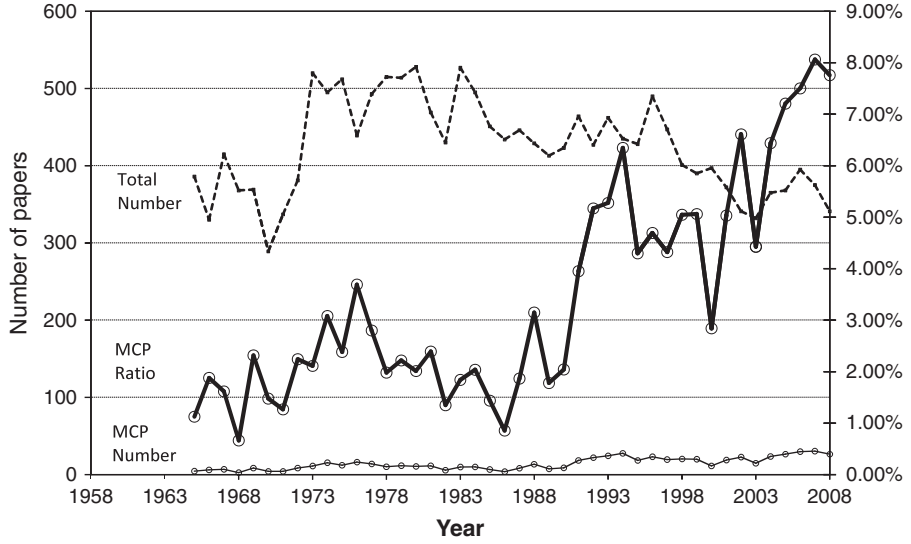

Figure 1 The proportion of papers in the leading four methodological journals that were devoted to simultaneous and selective inference. The total number of papers (left scale) in the *Annals of Statistics*, *Biometrika*, *JASA* and *JRSS B* (dashed line), the number of papers devoted to MCP (slim line) and the proportion (bold lines with empty circles, right scale). Based on computerized search by Johnatan Rosenblat.

momentum. This change brought some of us, in the beginning of 1995, to think about organizing a meeting devoted to Multiple Comparisons. As the call for papers read:

> "The field of multiple comparisons has progressed tremendously over the last 40 years… The International Conference on Multiple Comparisons was planned in order to take stock of how the field had progressed, what some of the current problems of active interest are, and where the field should be going."

The organizing committee consisted of Charles Dunnett, Yosi Hochberg, Ludwig Hothorn, Sture Holm, J. C. Lemarie, Ruth Marcus, Julie Shaffer, Ajit Tamhane and I. This first conference was held in Israel in 1996, not long after the assassination of the Prime Minister Rabin and the following series of suicide bombers. In spite of the difficulties at the background, the conference attracted some 60 people. It was a success, which have been followed by a series of MCP conferences around the world attracting an ever-growing crowd of researchers and workers in the field: Berlin at 2000; Bethesda at 2002; Shanghai at 2005; Vienna at 2007, and Tokyo at 2009, where this paper was presented.

The large numbers of participants at the recent conferences reflect the increase in the methodological output, which accelerated in the last decade (see again the last part of Fig. 1). This increase is mostly a result of the introduction of microarray technology in Genomics and the multiplicity difficulties that their analysis raised. Other high-throughput technologies emerged soon after, with their impact on MCP, all resulting in an increasing flow of statistical methodology and technology.

2 Current successes and failures

This short overview of the history of MCP makes it evident that the field is currently at a peak in terms of methodological output. About 8% of the publications in the top four general methodo-

logical statistical journals deal with issues of, or related to, simultaneous and selective inference. Many more methodological papers appear in other top journals, and in particular in publications devoted to medical and pharmaceutical statistics as well as bioinformatics. The community of researchers in statistics for whom this area of research is a main one is growing, as is evident from the growing attendance in the MCP conferences. Many regulatory agencies require the use of sophisticated MCP procedures, and in some areas it is clear to potential users that they are in need of newly developed methods. In some areas, notably in the statistical analysis of clinical trials and in psychological research, the use of MCP is the prevailing approach. We have even made it to the regular news. In July 21, 2008, science column in Newsweek, Sharon Begley describes how "In the wild west that is genome research statisticians are the new sheriffs in town. Thanks to the new tools such as the false discovery rate technique they have repeatedly shot down claims" (based on an interview with Professor Brad Efron).

But we should not overlook our failures. Take for example the area of brain research, and in particular the efforts to study the brain using functional Magnetic Resonance Imaging (fMRI). A typical study involves screening through 10–30 thousand hypotheses, which clearly calls for the use of one or another method to address the effect of multiple testing. Still, a meta-analysis of fMRI research conducted by Saxe *et al.* (2006), revealed that out of the 1705 articles in Brede's data base, *p*-values that were not corrected for multiplicities were reported in 49% of the articles.

So, even though the multiplicity problem is well recognized in fMRI research, and all software packages in brain imaging include multiplicity adjustment tools (either random field-based FWER or FDR,) brain researchers do not use the MCP tools almost as often as they do.

Similarly, consider the replication study of genome-wide association signals in UK samples, looking for risk loci for type 2 Diabetes, as reported by Zeggini *et al.* (2007). The study involved a new scan of associations with type 2 diabetes in the UK and three other scans. The main findings of the paper are summarized in a single table, containing information about the selected top ten loci. The information includes details of the locations, and for each location the estimates of odds-ratio, their 95% confidence intervals and *p*-values for each study separately as well as for the combined analysis. The description of the selection process reveals a very complex process:

> "The first wave of validated SNPs sent for replication was selected from the 30 SNPs, … which had, in the WTCCC scan alone, attained the most extreme ($p < 10^{-5}$) significance values.
>
> A heuristic approach for the second wave of priorization of signals is described, where *p*-values between 10^{-2} and 10^{-5} in the primary genome-wide association scan, were chosen if they further exhibited. Corroborating evidence for association with T2D in the companion … scans, as well as biological candidacy and multiple associations within the same locus."

What can be said about the uncertainty involved? There were about 400 000 SNPs tested in the four studies, and we are presented with results about a selected set. There is strong evidence for some of the associations: but we cannot quantify the uncertainty.

What does a *p*-value $\leq 10^{-5}$ mean? Does a 95% confidence interval carry its usual coverage meaning? What does a combined *p*-value of 10^{-6} mean, after such a selection process?

At the heart of our concern in this example is inference following selection, or as it can be called – *selective inference*. Once applied only to the selected few, the interpretation of the usual measures of uncertainty do not remain intact directly, unless properly adjusted. No adjustment of any sort was used in this study except for the statement about lowering the threshold for significance in the combined analysis to 10^{-6}. So while it is quite common to encounter the use of Multiple Comparisons methods in genomic research – some of what is being done is not rigorous enough and is difficult to justify.

The third example of our limitations is in the field of medical research, a field we feel we have made a real impact on. I would like to turn your attention to the paper by Ioannidis (2005) "Why most research findings are false". This paper has generated a wave of responses, both in the scientific press and in public debates (see Boston Globe op-ed of July 27, 2006, at http://www.boston.com/news/science/articles/2006/07/27/science_and_shams/). To quote just a few claims:

"There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field…"

"Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true…"

"…In this essay, I discuss the implications of these problems for the conduct and interpretation of research. It can be proven that most claimed research findings are false."

To a reader of the paper who is familiar with Multiple Comparisons it is clear that the source of the problems he is discussing is the use of nominal hypothesis testing even though many hypotheses are being tested, both within and across studies, and manifested *via* publication bias. In fact Ioannidis is repeating the very same arguments that were brought up by Soriç (1989), where he calculates the expected number of false claims as a proportion of the claims made. Neither the work of Soriç, nor the FDR concept and the large body of methodology it generated is recognized. In fact multiplicity is discussed in passing by Ioannidis, as an irrelevant issue to his concerns.

Ioannidis is not an isolated voice that ignores the multiplicity problem in medical research. An editorial discussing the Hormone Therapy study of Women Health Organization (Fletcher and Colditz, 2002) states:

"The authors present both nominal and rarely used adjusted CIs to take into account multiple testing, thus widening the CIs. Whether such adjustment should be used has been questioned, …"

I was surprised by these claims. With the help of Rami Cohen we examined a sample of 60 papers from NEJM (2000–2004). Our examination revealed that 47/60 studies had no multiplicity adjustment at all, even though all needed it in some form or the other.

In summary, in spite of the very large multiplicity problems encountered in these fields, the use of MCPs has not become the unquestionable practice. The field of simultaneous and selective inference has seen periods of decline in the past. The question that I would like to answer in the rest of this paper is "what can be done to avoid a decline from the current peak?" My answer, detailed below, is to see our mission as providing aid to researchers, to clarify the needs for well tailored MCPs, to simplify their use, to look for new challenging problems in old application areas, and to look for new challenging application areas.

3 Provide aid to researchers

Scientists are very ambitious about getting discoveries, an ambition that may hurt the replicability of the results. Lander and Kruglyak (1995) argue that the concern about replicability should be as central as the ambition to get positive results. In the paper where they discuss guidelines for interpreting genetic dissection of complex traits, they write:

"Adopting too lax a standard guarantees a burgeoning literature of false positive linkage claims, each with its own symbol… Scientific disciplines erode their credibility when substantial proportion of claims cannot be replicated…"

As I mentioned before in the Newsweek item statisticians are described as the new sheriffs in the wild west, shooting down false claims. I fully agree we should shoot down false claims, and these are

abundant in large inference problems. Even though the image of sheriffs is attractive in the eyes of the general public, I argue that in order to be effective with scientists we should take the attitude that we provide aid to researchers combating the danger of producing nonreplicable results and the embarrassment that follows. We should make efforts to convince potential users, in our personal interactions and in our writing, that the control of multiplicity-induced error is in their best interest in large inference problems.

4 Clarify the need to use well-tailored MCPs to practitioners

4.1 The problem

Since the 1950's the concern with the "don't worry be happy" unadjusted approach was formulated as the demand to have almost always no error at all – the FWER approach. Once the FDR managed to break the dichotomy, many other concepts of error rates have been offered, error rates that try to take some middle way between the two extreme approaches. Defining $R_i = 1$ if the i-th hypotheses is rejected, $V_i = 1$ if it is rejected in error, and otherwise setting both at 0, and further denoting $V = \sum_{i=1}^{m} V_i$ and $R = \sum_{i=1}^{m} R_i$, the list of error rates is given in Table 1.

As the reader may feel at this point that the list raises confusion, even for statisticians. Are the differences between these error rates of importance, and what do we need so many error rates for? To increase this sense of bewilderment let me leave the error rates unexplained until Section 4.4 where I shall discuss them and argue that each of them might be appropriate and useful for some inferential situation. Setting the ground for the discussion I shall first introduce two important distinctions.

4.2 Selective *versus* simultaneous inference

We were always aware that "Multiple Comparisons" is concerned with the effect of simultaneous and selective inference on the properties of the usual inferences if unadjusted for multiplicities. But

Table 1 Multiplicity-related error rates.

Unadjusted inference	$\mathrm{E}(V/m) \le \alpha$	
Weak control of FWER	$\Pr(V \ge 1) \le \alpha$	*Only* when *all* $\mathrm{H_i}$ are true
k-FDR	$\mathrm{E}((V-k)_+/\mathrm{R}) \le q$	
False Exceedance Rate	$\Pr(V/R > \mathrm{q}) \le \alpha$	
Weighted FDR (wFDR)	$\mathrm{E}(\Sigma\omega_i R_i/(\Sigma\omega_i V_i)) \le q$	where $\omega_i \ge 0$
False Discovery Rate (FDR)	$\mathrm{E}(V/R) \le q$	
Positive FDR (pFDR)	$\mathrm{E}(V/R\|\mathrm{R} > 0) \le q$	
Fdr	$\mathrm{E}(V)/E(\mathrm{R}) \le q$	
Fdr(z)	$\mathrm{Fdr(z)} = p_0 F_0(z)/F(z)$	
local FDR	$\mathrm{fdr}(z) = p_0 f_0(z)/f(z)$	
k-FEWR	$\Pr(V \ge \underline{\mathrm{k}}) \le \alpha$	
FamilyWise Error Rate (FWER)	$\Pr(V \ge 1) \le \alpha$	
Per-Family Error Rate (PFE)	$\mathrm{E}(V) \le \alpha$	

The references for the origin of the different error rates, as well as further clarification of the notations are given in the text. Emphasized in bold are the error rates at the extreme and the first error rate to be offered in-between. Note that three of these error rates share the same word (FDR) but vary in capitalization. This may be unfortunate, but reflect the commonly used notations by different authors.

only after coming up with the False Coverage-statement Rate criterion in Benjamini and Yekutieli (2005), it became clear to us that selective inference and simultaneous inference are two distinct goals. The confusion arises because methods that assure simultaneous inference also assure selective inference, and within the FWER framework all methods offer simultaneous inference, and therefore answer both concerns.

The concern in simultaneous inference is that all inferences made be simultaneously valid. Hence, if a single confidence interval covers the parameter in 95% of the realizations on the average, simultaneous confidence intervals cover all parameters in 95% of the realizations on the average. In that work we demonstrated that confidence intervals constructed for selected parameters only, where the selection depends on the observed values, fail to ensure nominal coverage, even if we do not require coverage for all parameters at the same time, but just care about the average property over the selected ones. We suggested the False Coverage-statement Rate (FCR) as the appropriate criterion for capturing the error made by constructing confidence intervals for selected parameters. The definition of the FCR parallels that of the FDR, where a discovery is replaced by "a confidence statement on a parameter is made" and a false discovery is replaced by "a confidence statement on a parameter is made but it fails to cover the parameter". A general procedure was offered, that adjusts the confidence level using the factor $|S|/m$ where $|S|$ is the (data dependent) number of the selected, out of the potential m parameters. Instead of $1-q$ confidence intervals, construct marginal $1-q|S|/m$ confidence intervals for the $|S|$ selected parameters. Indeed, in many large problems we do not care about simultaneous coverage but do care about the effect of selection on the average marginal properties over the selected ones, because only the findings about the selected few are of importance. In retrospect we understand that this type of concern in the realm of testing is the one addressed by the FDR criterion and some of its variations. Note that selection need not be direct but can result from partial reporting, or from highlighting, or from attending only to the significant findings.

4.3 Testing *versus* interval and point estimation

There is a distinction between the three in the realm of MCP as to the state of our knowledge. While the FDR offers a well-developed set of methods to address selective inference in testing, the situation is different for estimation. Selective inference for confidence intervals has been offered, as discussed above, but it might well be that the confidence intervals offered are too wide. Estimation after selection is beginning to attract attention: Efron (2008) discusses the issue, and the associated difficulties, from the empirical Bayes point of view. Yekutieli (2009) discusses it from both the Bayesian and empirical Bayes points of view, and tries to understand formally the different effects of selection in both approaches. Soft thresholding used in the context of signal denoising may be viewed as another direction.

Unfortunately adjusting for selection effect in testing does not address estimation following selection. A recent dispute in Social Neuroscience demonstrates vividly this point. Take, for example, a study discussed by Vul *et al.* (2009). The experimenters created a game exposing individuals to social rejection. Then, they measured the brain activity in 13 individuals at the same time as the actual rejection took place. They also obtained a self-report measure of how much distress the subject had experienced during the game. Finally, they searched for the brain region whose correlation with the reported distress was the highest. The reported correlation with activity in anterior cingulate cortex was 0.88 – much beyond what is normally encountered in social studies, and very likely beyond the capability of the self-reported measure. Vul *et al.* (2009) questioned such high estimates, calling them "Voodoo correlations". Jabbi *et al.* protected the practice in a web posted article (www.bcn-nic.nl/replyVul.pdf), arguing that,

"The authors misunderstand the critical role of multiple comparison corrections and conflate issues pertaining to null hypothesis testing and effect size estimates, respectively."

But the fact is that their response fails to recognize that selection may affect estimation, and that using the multiplicity correction for testing does not diminish the selection effect on the estimation. If instead of point estimates confidence intervals were used, the confidence intervals would have to be much wider because of the selection effect, making much lower correlations still plausible.

4.4 Matching error rates to inference needs

After preparing the ground in the previous two subsections, I can now address the issue of recommending the use of a specific error rate according to the purpose of the inference: personal decision-making (e.g. what gene should I further study or what route should I take), policy decision-making (e.g. comparing educational policies), monitoring (*e.g.* a breakout of disease), scientific communication at exploratory levels (e.g. candidate genes) and confirmatory or licensing analysis (e.g. of drugs and medical devices). Otherwise, users will again avoid using control of any error rate in a meaningful way, and rely instead on *ad hoc* solutions, such as the unclearly justified threshold of p-value $\leq 10^{-5}$ in the replicability analysis of Type 2 Diabetes studies discussed in Section 2 (see Benjamini *et al*., 2009 for a discussion of this example).

Unadjusted inference is appropriate where no selection and or simultaneity are at play. I do not care about multiplicity when working on completely different projects, because no selection takes place, and the users do not care about the validity of the simultaneous inference in their project and the projects of others.

Weak control of FWER is rarely of interest as an end result, and more often plays as a building block for other procedures. Testing a main effect in ANOVA is a relevant example, where one is generally not satisfied with weak control and seeks at least a limited set of pairwise comparisons. Donoho and Jin (2004) do give some important examples, from surveillance to astrophysics, where this limited concern is justified. Often such a study will be followed up with an effort to point at the source of the violation of the intersection null hypothesis.

At the other extreme lies the strong control of FWER, that addresses the concern about simultaneous inference, and therefore about the effect of selection as well. It is the most developed concept and offers a solution for any purpose and under a wide variety of statistical scenarios. It is therefore appropriate to serve all needs in regulation, policy making, scientific reporting, *etc*. The only problem, of course, is that this error rate so badly limits the power of the statistical methods that offer such protection really that one should not use it unless simultaneous inference is needed.

The control of k-FWER (Lehmann and Romano, 2005) offers simultaneous inference, and at this stage the error rate is relevant only for testing. I view this error rate to be appropriate for personal decision-making. Suppose I control the FWER and have three discoveries. If I allow myself two errors and now find five discoveries instead of the previous three – I know they are worthless; if I find ten more – I know I have some substance in the newly discovered pool, and the use of k-FWER has offered a way to overcome a hurdle. This is not the situation if all that is being reported is that the hypothesis was rejected at the 3-FWER level. I have to know that this hypothesis is not the single member of the rejected pool.

The control of the FDR (Benjamini and Hochberg, 1995) offers protection against the effect of selection, but does not offer simultaneous inference. Together with the FCR we can offer advanced methods for testing and reasonable methods for confidence intervals but none for point estimation. As such, it is the adequate goal for policy making and scientific reporting.

The control of k-FDR (Sarkar and GUO, 2008) allows personal assessment of the implications of relaxing the requirements of FDR, and is related to FDR just as the k-FWER to the FWER, with the advantages for personal decision making, and drawbacks for the other purposes.

The weighted FDR (Benjamini and Hochberg, 1997) has not received much attention yet, but it is a very promising approach, with weights capturing importance or even monetary value. Offering selective inference, it can be used to compare the utility of policy making options, allowing the

flexibility to include costs, and allows incorporating previous knowledge in scientific reporting. It can even serve for regulatory purposes, for example to control for selection effect of secondary endpoints in clinical trials.

Both the Fdr = E(V)/E(R) and the pFDR are variations on the FDR. Both were discussed in Benjamini and Hochberg (1995), but they were seriously treated only later: the pFDR by Storey (2002), and the Fdr by Efron *et al.* (2001). In the mixture model, nowadays associated with microarray analysis, where each hypothesis is assumed to come from the null ($H_i = 0$) or from the alternative with $Pr(H_i = 0) = p_0 < 1$, and when the number of hypotheses tested is very large, these are the same. So are the methods associated with these error rates and their properties.

Outside the mixture model, I would avoid the pFDR for policy making because it offers no control when all hypotheses are true. I would avoid Fdr because of the separation between V and R in the same experiment, which is again of no importance in the mixture model.

Within the mixture model the pFDR and the Fdr carry a Bayesian interpretation. If the observed z-value is distributed F, and under the null distributed F_0, then $\mathrm{Fdr}(z) = p_0 F_0(z)/F(z)$ is the posterior probability that the hypothesis is a true null if it is in the set selected by having observed z-value more extreme than z. Its companion is the local FDR, where F and F_0 are replaced by their densities.. The local Fdr at z turns out to be the posterior odds for the hypothesis being a true null given its z-value, calculated under the assumption that all hypotheses have similar priors. Therefore both error rates are appropriate for selective inference for personal decision-making, such, as what leads to follow within a selected set. Both error rates were offered in Efron *et al.* (2001), and Efron further developed an empirical Bayes approach to their estimation (see his review in Efron, 2008). Within this approach it may be possible to offer estimation methods that address selection effects as well.

The False Exceedance Rate was offered by Genovese and Wasserman (2004) as a stronger form of FDR. Using this error rate one can assure that the observed proportion of false discoveries will not be above q up to a desired level of uncertainty. In this sense, it better supports the naïve interpretation of the FDR, but it is not clear yet to what extent one has to give up power in practice.

Note that I have discussed only error rates, and not procedures. The use of the procedure should match the error rate, and not the other way around. There was a period of intense debate as to the estimation of the error rate, rather then its control, as a different and better approach. In my view this is an artificial distinction, and while estimators (say of FDR) are used as working tools one should worry about the properties of a procedure that thresholds according to the value of the estimator in order to focus on discoveries.

In summary, this multiplicity of multiplicity error rates should be welcomed. Still, we have done too little to offer practitioners, nonstatistician in particular, advice as to what situation requires what error rate, the same way we advice them about what test is needed for what situation. Bringing my views to the discussion table, however personal and debatable they seem now, will help, I believe, the buildup of a consensus about most such matters.

5 Simplify the use of the newly developed methodologies

The transfer of new developments in MCP methodologies into easily usable software cannot be overemphasized. Making software available in a variety of forms should be the concern of developers, and should be recognized as important within our community, the same way it is recognized and rewarded in the Bioinformatics or Machine Learning. We should continue with a project as that of Westfall *et al.* (1999) in SAS updating the software. Available methodologies in SPSS are not quite updated. According to my knowledge Matlab does not enjoy a coherent set of MCP procedures. R has more of these, but they are quite scattered. In fact the Bioconductor depository has an extensive set of programs attending multiplicity. A similar depository project for Multiple Comparisons in general – not only those tuned to the analysis of large genomic problems – will

benefit much the MCP field. In fact, a few months after the presentation of this work Dr. Thorsten Dickhaus and Dr. Gilles Blanchard have initiated an effort along these lines, and the current status of the project named μTOSS can be checked at https://r-forge.r-project.org/projects/mutoss/. I was equally happy to learn that the book by Bretz *et al.* (2010) "Multiple Comparisons using R" has been just published, transferring into R realm the work and attitude of Westfall *et al.* (1999).

Finally, we should not forget Excel. It may come as a surprise to us that a large proportion of the people conducting statistical analysis use Excel for this purpose. Economists, managers, biologists, chemists and others, start looking for a solution by a search for an Excel solution to their problem. The success of SAM by Storey and Tibshirani (2003) is an evidence of this phenomenon. These users of statistical analyses within Excel should have available tools that address multiplicity in modern ways.

6 Look for new challenges in old application areas: Safety analysis

As is clear from the concise history of MCP, innovative applications have been the driving force behind developments in the area. I believe this will continue to be the case in the future, if we open our minds to emerging challenges. I shall not try to outline all such opportunities, but rather give an example of a new challenge in an old application area, and an example in a new one. Let me first address the challenge in old application areas: the study of safety in clinical trials.

6.1 Safety analysis in clinical trials

When inferring about efficacy of a drug the statistical goals are quite clear: showing success for well-defined endpoints. A large set of quantitative tools is available: hypotheses testing and confidence intervals, dose response analysis, particular analyses for survival data, bioequivalence and longitudinal analysis. The tools for addressing the relevant multiplicity issues are also well developed: primary and secondary endpoints, comparisons with control, simultaneous confidence intervals, dose response methods, bioequivalence, gate-keeping procedures, multiple looks and so on.

In contrast, safety analysis remains at the descriptive stage with no quantitative inference. An evidence for this situation can be found in some of FDA's guidance to industry as appear, for example, in their document "Clinical Data Needed to Support the Licensure of Trivalent Inactivated Influenza Vaccines". A detailed description of guidelines about efficacy analysis is given, but no indication that any analysis should be done regarding safety – only reporting.

Discussing the role of safety in the assessment of clinical trials, Ioannidis *et al.* (2004) noted that in the CONSORT (CONsolidated Standards Of Reporting Trials) statement, only one out of the 22 recommendations addressed safety. They added ten additional recommendations, but when refering to the role of statistical methods to address safety they wrote: "Using only descriptive statistics to report harms is perfectly appropriate in most RCTs because most trials lack power…". So even here safety analysis remains at the descriptive stage.

Table 2 presents part of a larger table taken from Fowler *et al.* (2006), where the results for safety are compared between Daptomycin treated and the standard therapy for Bacteremia and Endocarditis caused by *Staphylococcus*. It is brought here to demonstrate the way safety data are usually reported: number and percentage of some adverse events in each group separately, and *p*-value of the significance of the difference. The larger table looks the same but involves about 30 safety endpoints.

There are inherent difficulties in the way safety is assessed and inferred from such reports.

(i) First and foremost, the test of the simple hypothesis of equal safety is not the relevant one for inference! The conclusion one wants to make in safety analysis is that the new treatment (Daptomycin in this example) is as safe as the standard therapy. That calls for the null

Table 2 Safety results from Table 5 in Fowler *et al.* (2006).

Adverse events	Daptomycin ($N = 120$)	Standard therapy ($N = 116$)	p-Value
	No. of patients (%)		
Any drug-related adverse events	42 (35.0)	49 (42.2)	0.29
Any serious adverse event	62 (51.7)	52 (44.8)	0.30
Any drug-related serious adverse event	3 (2.5)	6 (5.2)	0.33
Death	18 (15.0)	19 (16.4)	0.86
Stopped study drug because of drug-related adverse event	10 (8.3)	13 (11.2)	0.51

The results for safety are compared between Daptomycin treated and the standard therapy for Bacteremia and Endocarditis caused by *Staphylococcus*. The results here reflect only a small part of the results in the original table.

hypothesis to be tested that the proportion of adverse events is higher in the new treatment group than in the standard therapy group. But of course if the safety is actually at the same level in both groups the power to show safety is the same as the level of the test. The situation is therefore similar to that of showing bioequivalence, or noninferiority, when inferring about efficacy (for related approach see Bauer *et al.*, 2001).

(ii) The multiplicity problem is very different in safety analysis than in efficacy analysis. When facing multiple safety endpoints, the concern is not about making even one type I error in the set of tests of equalities of the measures of harm. If it were, such concern about multiplicity makes it more difficult to detect the harm caused by the new drug. Therefore, adding safety endpoints will reduce the sensitivity of the analysis – an unacceptable result. Alas if no adjustment is made, adding safety endpoints will assure that some nonexisting safety problem will always be detected.

(iii) An endpoint of efficacy, whether primary or secondary may fire back and become a safety problem.

(iv) Lack of power because adverse events are few. This is often mentioned as the problem underlying any formal analysis of safety.

(v) Finally, in many cases the analysis should weigh increased benefits against reduced safety. Combined endpoints offer a good theoretical direction that is hard to follow in practice.

In the spirit of this general paper I shall not get into a detailed discussion of the solutions to all of these problems. I do not even have them. But I would like to show that a fresh way to address safety is feasible, and much new research can be done to solve the problems raised *via* tests of conjunction, creating a tool for Aggregated Safety Assessment.

6.2 Aggregated safety assessment

For this purpose, let us turn somewhat technical. Let the mean of the i-th measure of harm under the new treatment be η_{ti}, and under the standard treatment η_{si}. Define, for some prespecified $C_i > 1$,

$$H_{0i}: \eta_{ti} \geq C_i \eta_{si} \quad \text{and} \quad H_{1i}: \eta_{ti} < C_i \eta_{si},$$

the alternative meaning that the new treatment is not worse than the standard by more than the factor C_i. We want to show that all alternatives H_{1i} hold, implying safety in all measures, and this

can be done *via* the special test for the conjunction of alternatives: $H_0 = \cup H_{0i}$, so rejecting H_0 means that all the alternatives hold.

If p_i is the p-value for testing H_{0i}, in order to reject $\cup H_{0i}$

$$\text{all } p_i \leq \alpha.$$

Equivalently, define an adjusted p-value for the test of conjunction

$$p_{adj} = \max(p_i),$$

and if $padj \leq \alpha$ conclude that treatment is (up to some factors) safer in all measures. Note that since being safe at level α requires all p-values to be less than α, the more measures for safety (safety endpoints) are tested the more difficult it becomes to show safety – just as the case is for efficacy.

The origin of the test of conjunction is Berger (1982) (see also Hochberg and Tamhane, 1987). Tests of partial conjunctions have recently become of interest in fMRI studies, as well as in replicability analysis (Benjamini *et al.*, 2009).

Aggregated safety assessment takes this point of view one step further. For a given set of safety measures, the smaller the factors C_i, the stronger the conclusion about safety. Take $C_i = c > 1$ for all i, and define

$$H_{0i}(c) : \eta_{ti} \geq c\, \eta_{si} \quad \text{and} \quad H_{1i}(c) : \eta_{ti} < c\, \eta_{si}.$$

Again, in order to show that all $H_{1i}(c)$ hold, test $H_0(c) = \cup H_{0i}(c)$, using the individual p_i (c) for testing $H_{0i}(c)$. For that purpose define

$$p_{adj}(c) = \max_i(p_i(c)),$$

so $p_{adj}(c)$ can be used to test whether treatment is less harmful than c-times-standard in all measures.

Figure 2 displays $p_{adj}(c)$ as a function of c, the Aggregated Safety Curve. The smaller c is, the larger the adjusted p-value is, which means the stronger the safety requirement is, the weaker the evidence the data offer to support it. As a first step, the curve is a display of quantitative assessment of aggregated safety (see also Bauer and Kieser, 1996). The curve can further be summarized using area under the curve methodology or using the values of some prespecified points on it. For example one may use $p_{adj}(1.25)$ (about 0.25 in Fig. 2), or the c that can be supported at the.05 level (about 1.43 in Fig. 2). Other useful summaries of this curve may also be conceived.

It may be appropriate (i) to consider a subset of the safety endpoints to denote as primary ones, (ii) to make even finer distinctions of importance between the safety endpoints by weighing them differently, (iii) to incorporate structure of safety/adverse events and of (iv) body systems or (v) to use hierarchical testing to gain power. These issues as well as the relative merits of the summaries can wait for a more thorough treatment of the method in the future.

This is just one possible starting point for quantitative safety analysis. I am trying to make the case that safety analysis can be quantitative and inferential, while raising new simultaneous and selective problems. Thus *old areas of applications may offer new challenges – requiring new methodologies*.

7 Look for new challenges in new application areas: fMRI analysis

The scientific scene is changing rapidly. New areas appear and older ones change dramatically as a result of new technologies. Sequencing and microarray technologies have changed the field of genomics, and gave rise to the flourishing field of Bioinformatics. fMRI has changed dramatically brain research. These areas of research emerged fast. In some of them statisticians were not present at the birth bed, and the fields are dominated by other professions. In others, statisticians were present, and both these research areas and the statistical methodology benefited from their presence

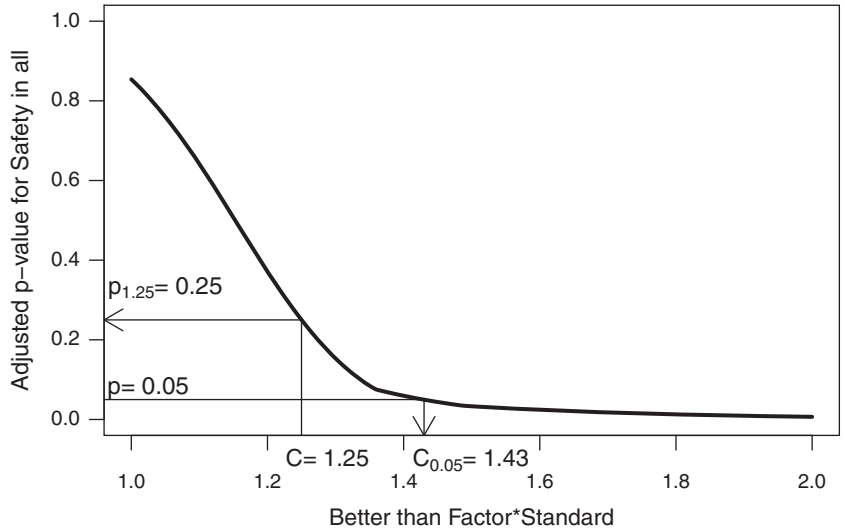

Figure 2 Aggregated safety curve for the five safety endpoints in Table 2. The figure displays the $p_{\text{adj}}(c)$ for showing that all safety endpoints are no more larger than c times their value under the standard treatment, as a function of c. The curve can further be summarized using area under curve methodology or using prespecified points on it. For example, one may use $p_{\text{adj}}(1.25)$ (about 0.25), or the c that can be supported at the 0.05 level (about 1.43).

at that time. In most cases, the problems are further characterized by their large dimension, so in particular they enriched the area of Multiple Comparisons.

The area of research into the functioning brain using MRI is an example of a fast developing area with only a handful of statisticians being involved, and unfortunately their number even decreased this past year with the untimely death of Prof. Keith Worsley, a very dominant statistician in the area. Multiplicity problems are not only abundant in fMRI research; they are central to the area as there are fierce debates as to the goals of analyses. Should the emphasis be on inference on active voxels, those volumetric pixels for which measurements are available, on active regions, or on both? How can we make use of predefined regions (Regions Of Interest, ROI) whether general or matched individually defined (functional ROI)? Should not we be interested in topological features such as peaks and cusps? How need multiplicity be addressed when combining single subject analysis into group analysis? The debates regarding the goals in the fMRI community are in dire need for the active participation of statisticians whose area of interest is Multiple Comparisons. There are clearly new MCP challenges in emerging new application areas.

8 Concluding remarks

Those working in the area of MCP have had an exciting, wonderful and rewarding decade. We have been very successful but have to make efforts in order to keep it going. Working on simple guidelines as to what to do, when, and why, while offering ways to do so easily, is essential in order to make our methodologies and technologies useful. Getting involved in new application areas will ensure the vitality and relevance of our field as a research area, and our continued success.

Acknowledgements The support of the Israeli Science Foundation Grant is thankfully acknowledged.

Conflict of Interest
The authors have declared no conflict of interest.

References

Bauer, P., Brannath, W. and Posch, M. (2001). Multiple testing for identifying effective and safe treatments. *Biometrical Journal*, **43**, 605–616.

Bauer, P. and Kieser, M. (1996). A unifying approach for confidence intervals and testing of equivalence and difference. *Biometrika* **83**, 934–937.

Benjamini, Y., Heller, R. and Yekutieli, Y. (2009) Selective inference in complex research. *Philosophical Transactions of the Royal Society A* **367**, 4255–4271.

Benjamini Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* **57**, 289–300.

Benjamini, Y. and Hochberg Y. (1997) Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics*, **24**, 407–419.

Benjamini, Y. and Yekutieli, D. (2005). False discovery rate-adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association* **100**, 71–81.

Berger, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometric* **24**, 295–300.

Bretz, F., Hothorn, T. and Westfall, P. (2010). *Multiple Comparisons Using R*. Chapman & Hall/CRC, London, Boca Raton, FL.

Donoho, D. and Jin, J. S. (2004). Higher critisism for detecting sparse heterogeneous mixtures. *Annals of Statistics* **32**, 962–994.

Efron, B. (2008) Microarrays, empirical Bayes and the two-groups model. *Statistical Science* **23**, 1–22.

Efron, B., Tibshirani, R., Storey, J. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96**, 1151–1160.

Fletcher, S. W. and Colditz G. A. (2002). Failure of estrogen plus progestin therapy for prevention. *Journal of the American Medical Association*, **288**, 366–369.

Fowler, V. G., Boucher, H. W., Corey, G. R., Abrutyn, E., Karchmer, A. W., Rupp, M. E., Levine, D. P., Chambers, H. F., Tally, F. P., Vigliani, G. A., Cabell, C. H., Link, A. S., DeMeyer, I., Filler, S. G., Zervos, M., Cook, P., Parsonnet, J., Bernstein, J. M., Price, C. S., Forrest, G. N., Fakenheuer, G., Gareca, M., Rehm, S. J., Brodt, H. R.., Tice, A., Cosgrove, S. E., (2006). Daptomycin *versus* standard therapy for bacteremia and endocarditis caused by *Staphylococcus*. *New England Journal of Medicine* **355**, 653–665.

Genovese, C. and Wasserman, L. (2004) A stochastic process approach to false discovery control. *Annals of Statistics* **32**, 1035–1061.

Hochberg, H. and Tamhane, A. (1987). *Multiple Comparisons Procedures*. Wiley: New York.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65.

Hsu, J. C. (1996). *Multiple Comparisons*. Chapman & Hall: London.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS medicine* **2**, e124.

Ioannidis, J. P. A., Evans, S. J. W., Gotzsche, P. C., O'Neill, R. T., Altman, D. T., Schvlz, K., Moher, D. (2004). Better reporting of harms in randomized trials: an extension of the CONSORT statement. *Annals of Internal Medicine* **141**, 781–788.

Lander, E. S. and Kruglyak L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics*, **11**, 241–247.

Lehmann, E. L. and Romano, J. P. (2005). Generalizations of the familywise error rate. *Annals of Statistics*. **33**, 1138–1154.

Marcus, R., Peritz, E. and Gabriel, K. R. (1976). Closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655–660.

Sarkar, S. K. and Guo, W. G. (2009). On a generalized false discovery rate. *Annals of Statistics*, **37**, 1545–1565.

Saxe, R., Brett, M. and Kanwisher, N. (2006). Divide and conquer: A defense of functional localizers. *NeuroImage*, **30**, 1088–1096.

Soriç, B. (1989) Statistical ''discoveries'' and effect size estimation. *Journal of the American Statistical Association*, **84**, 608–610.

Storey J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, **64**, 479–498.

Storey, J. D. and Tibshirani, R. (2003). SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. In: Parmigiani, G., Garrett, E.S., Irizarry, R.A. and Zeger, S.L. (Eds.), *The Analysis of Gene Expression Data: Methods and Software*, Springer, New York.

Vul, E., Harris, C., Winkielman, P. and Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, **4**, 274–290.

Westfall, P., Tobias, R., Rom, D., Wolfinger, R. and Hochberg, Y. (1999). *Multiple Comparisons and Multiple Tests Using the SAS System*. SAS Publications, North Carolina. ISBN: 978-1-58025-397-0.

Westfall, P. H. and Young, C. (1993). *Resampling Based Multiple Comparison Procedures*. Wiley-Interscience.

Zeggini, E., Weedon, M. N., Lindgren, C. M., Frayling, T. M. *et al*. (2007). Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316**, 1336–1341.

