Disturbing examples with multiple testing adjustments

Example 1: Response of cancer pts to IL 2: effect of immunologic HLA type
Rubin et al, 1995.

Data = 2x2, row= HLA type DQ1 present or absent, col= IL2 response = yes or no,
Fisher "exact" test P= 0.01.

Data = three 2x2 tables, for HLA types DQ1…DQ3.
Minimum P-value = 0.01 for DQ1. Times 3 ➔ 0.03

Data = 3 groups of 2x2 tables, including 3 DQ types, 5 DP types, and 7 DR types.
Minimum P-value = 0.01 for DQ1. Times 15 ➔ 0.15

Data = 2 groups of groups of 2x2 tables,
including MHC1 (A, B, C) and MHC2 (DP,DQ,DR). Total # tables = 120.
Minimum P-value = 0.01 for DQ1. Times 120 ➔ 1.2. Sidak: 0.70

What is the "proper" collection of tests to control the Type I error over?
Just the DQ1 test? All DQ tests? All MHC2 tests? All HLA tests?

The same data was reported in an earlier paper. It had:
more prestigious authors
uncorrected data errors
the weakest multiple comparisons adjustment possible (2nd method above)
Our paper used the exact tests to identify a hypothesis (DQ1) of many, then did an *independent* verification test, using survival data .

Example 2: "Comparisons of a priori interest" Cohen Anwar, Day 1983.

Testing 6 methods of measuring echocardiograms—are they equivalent?

6x5/2 = 15 comparisons
Nominal P for method B versus method C = 0.005.
But not "of *a priori* interest", so P = 15 * 0.005 = 0.075, "not significant".

But now investigator states "the comparisons of *a priori* interest were:

A versus D, A versus E, A versus F, D versus E, D versus F, E versus F

Now the adjusted P values for B versus C is

P = (15-6) * 0.005 = 0.045, "significant".

So the inference on B versus C changed depending on how many <u>others</u> were "of *a priori* interest".

<u>Example 3: ECOG 5592 cooperative group clinical trial</u>

Arms: (A) etoposide + cisplatin, (B) taxol+cisplatin+G-CSF, (C) taxol+cisplatin. Should multiple comparisons adjustments be made? Which ones? How many?

The mystery answer: "There are four comparisons:

B > A, C > A, B > C, C > B

so the required significance level will be 0.05 / 4 = 0.0125".

<u>Issues with multiple comparisons methods</u>

- This is sometimes too cautious ("conservative"), if the tests are positively correlated.
- Often it's difficult to decide what collection of tests to throw together into one "bag". Bag = this data set? This article we're writing? All today's analyses??
- For huge numbers of tests (for example, high-thoughput biological data; degrees of freedom is negative, $n < K$), the "family-wise error rate" (FWER) may be far too conservative. One would allow a high probability of **at least one "false positive" (Type I error)** in exchange for **making some true discoveries.**