Living to 100 and Beyond: Survival at Advanced Ages
Session 6: Risk Factors

Discussant: Esther Portnoy

One of the advantages of being an academic is that we get to play first with the new toys. The software that Professors Guo and Wang have used is brand new (2001), and even the general area of data mining is less than 20 years old. Now, I don't mean the word "play" in a negative sense. When new tools come out, it really is important to have many different people try them out in different applications, including ones for which they were not designed. That's how we begin to learn how these things really work (not always as advertised), and when they don't; we figure out how to combine them with other tools and suggest ways they might be improved. It may be too great a risk for commercial practicioner to try one of these new things, not being quite sure whether it will yield the required output in a reasonable time. So I hope that you will continue to indulge the academics as we "play" with these toys, trusting that enough of them will eventually become reliable tools to repay your investment.

So, what does this data-mining thing do? It claims to automate the selection of the more important factors affecting an outcome (in this case the mortality rate), and to separate a possibly large amount of data into classes in each of which we can then estimate the parameters of some model for the response as a function of the factors. The hope is that we will be able to steer a reasonable course between two dangers: on the one hand forcing all the data into a single model that obscures some important features, and on the other hand fragmenting the data so much that it is difficult to see patterns.

Unfortunately, I confess that I am almost totally in the dark as to how this selection works. Table 1 lists "variable importance measures" and presumably the higher the number, the more important the variable, but beyond that the numbers mean nothing to me. Having determined that "participant status" is the single most important variable (other than age, which is handled differently), we first divide the data into five groups (status = employee, retiree, disabled, beneficiary or "combined"). I assume that it was theoretically possible to have combined two or more of those groups, but they were somehow found all to be different enough. The retiree group was then further subdivided according to sex. Does that mean that in the retiree group (but not in the other groups) there is not just a difference between males and females, but a difference that is not well modelled by logistic regression? Or might it be that only in the retiree group are there enough individuals of each sex to give statistically significant results? The paper does not tell us.

Once the segments have been determined, the regression is, I assume, standard. What remains is to interpret the results, and here again I find myself asking more questions. There are some inconsistencies in presentation, which probably do not represent any error in the analysis but can leave the reader somewhat confused. For example, random variables are sometimes 0/1 and enter the formula with a nontrivial coefficient; at other

times the coefficient is absorbed into the variable itself. There is too little detail on some of the interaction terms. Most importantly, some of the numerical results are counterintuitive, and I don't see any discussion of possible explanations. Consider, for example, equation (1), describing the model for beneficiaries. I would have thought that, for pay type and union status, the "combined" setting would give a contribution between the two specific settings, but it doesn't. Also, apparently the larger the annuity, the greater the mortality rate, and that seems backwards. (This happens in most of the segments.)

I'm not sure this paper offers much new insight into mortality patterns among the very old, but as the authors say, it demonstrates a new tool that might prove quite useful once a bit more data is available. Papers like this, and the discussions they prompt, help us to be ready to use the tool when the time comes.

If we compare data mining to this year's must-have toy, then Cox regression is like a Barbie doll: new models come out from time to time, but the basic product is well-tested, reliable and popular. The paper by Zhu, Hoag, Julien and Cui relies on SAS software to carry out the Cox regression, producing estimates of parameters and plenty of diagnostics. However, as Gua and Wang point out, if the model is not appropriate then the parameter estimates don't really mean anything, and I am afraid that is what has happened here. Remember that we are particularly interested in mortality at high ages, and want to avoid the tendency merely to extrapolate from the more extensive data that we have at earlier ages. But the Cox model (at least in this relatively simple form) forces us to look only at the average effect of each factor, and obscures any systematic change over time. Thus we read that male mortality is about 30% higher than female, whereas in fact it is quite well established that the discrepancy is considerably higher at younger ages (eg male rates roughly double female rates from the mid-teens to about age 50) and gradually decreases, possibly disappearing or reversing somewhere around age 100. Because almost all of the data here is at ages under 80, this "one-model-fits-all" approach leaves us with very little insight into the very subject we were hoping to address in this seminar. The Cox model does indeed have some advantages (and availability of software is not the only one), but it is too blunt an instrument for the task at hand.

A comment from Brad Efron in *Statistical Science* (August 2001, pp 218-219) warns us of the dangers of "black-boxism", the temptation to rely on the output of software packages without careful examination of the assumptions and results. Efron is hardly a person to be reluctant to use technological tools, and his remarks should be required reading for anyone relying on automatic methods to analyze large data sets.