Epidemiological and Clinical Data Management

Case 1 - Dr. Wood is the principal investigator for a large, multi-center cohort study of cancer in adults. Over the last year, two postdoctoral fellows, each working with their respective tenure-track mentor, had embarked on studies examining risk factors for finger cancer. Because he had noted a strong north-south gradient in the U.S. Atlas of Cancer Mortality, PD 1 studied the relationship with climate and temperature, while PD 2 examined the associations with occupation, pollution, and genes.

Eager to confirm his hypothesis and impress his mentors, PD 1 started his analyses, identifying his main residential history questions, creating new variables related to annual seasonal temperatures from a NOAA database, and working up other covariates for potential confounding. Meanwhile, based in part on her previous experience implementing field studies, PD 2 was painstakingly reviewing the questionnaires, cataloging the myriad exposures that had been quantified, and drafting a careful and complex analytical plan.

The initial analyses of PD 1 showed substantial variation in the geographic distribution of finger cancer in the cohort, and there was a striking risk-annual ambient temperature gradient such that the persons in warmest regions were at greatly and significantly reduced risk. (A lower temperature threshold effect was also suggested by the data, however.) PD 1 was very excited by these ground-breaking results, which he explained on the basis of hemodynamics, and shared them with his mentor, TT 1. TT 1 agreed that PD 1 should complete the analyses and get internal clearance in time for an upcoming AACR late-breaking session abstract deadline, even though consideration of the entire database was lacking. The abstract was accepted for oral presentation and PD 1 was invited to participate in a press conference at the meeting. PD 1, TT 1, and Dr. Wood were ecstatic, and planned for rapid submission to a high profile journal.

At the same time, PD 2 had begun to produce some very interesting results, including age and sex differences, and had DNA samples from a nested case-control set sent to the genotyping facility. Dr. Wood was not impressed with her progress, however, especially in light of PD 1's AACR acceptance, and she asked PD 2 to present her initial findings at the next Branch meeting.

After going over the data and slides with TT 2, PD 2 presented her results to the group. She had found independent, positive associations for the 45-65 age range (in men only) and showed a RR (Relative Risk) of 10 for the use of argon-infused, sub-zero gloves (included in the Apparel module of the study questionnaire only after two visits to the TEQ (Technical Evaluation of Questionnaires Committee) and at the insistence of a previous fellow). A gasp went around the room, and eyes turned to PD 1 and TT 1. They revealed that they had looked at the glove variable but did not keep it in the final models owing to "some" attenuation of the main finding. Also, they had learned of specific factories in Montana, North Dakota, Wisconsin, Michigan, and New York that could have been explored in the data but were not. Dr. Wood was not looking forward to her next meeting with the Division Director.

Questions

What should the investigators and Branch do with this new information?

What steps could have been taken earlier to avoid the present situation?

What are the implications for the abstract accepted by AACR? How do pressures of meeting submissions and publishing in competitive fields affect decisions regarding which data to include?

What are the steps in evaluating and managing the data before they are analyzed? Where can the most critical errors occur? Who has oversight of data linkages and database integrity?

What responsibility does the PI have for monitoring data-related tasks and knowing which piece of primary data was used in each analysis, which was not, and why?

What are some of the pitfalls regarding a priori and post-hoc hypotheses? Data exploration? Testing for confounders?

What constitutes original data in epidemiology? Is it the primary record, the questionnaire, the lab assays? Is it the electronic entry? Edited data on the servers?

Case 2 - You do an analysis of a risk factor, say body mass index, and multiple outcomes—i.e. diabetes incidence, risk of disability, risk of heart disease, and death. All the data are consistent with the exception of one endpoint.

How should you handle this?

Case 3 - You are involved in a clinical protocol comparing a clinical intervention with usual care. Overall, there is no difference between your intervention and control. However, on careful analysis, you see that there is a clear dichotomy in response—with a large group having a modest response but a small group with a very substantial response.

How do you analyze the data?

Case 4 - You are conducting a multicenter trial and note that all centers but two have results consistent with a positive outcome for the trial. You determine that the intervention was not applied as rigorously at these centers as in others.

Can you exclude these centers from the analysis?