The discussion concerning the timing of surgery highlights, among other things, the political aspects of ensuring surgery before the infant’s first birthday (at least during the period from 1984 to March 1994). The ‘suparegional’ status of the Bristol Royal Infirmary qualified it for central funding. However, ensuring that surgery takes place before the age of 1 year may influence some cases to be brought forward as well as others allowed to wait: the authors address the latter concern but not the former. Brining the date of surgery forward for some infants on the grounds of funding rather than clinical need may be just as dangerous as delaying surgery for other children and so the system (the National Health Service) may still be open to criticism. Superficially, the authors’ Table 8 (‘Age at which operations took place’) seems to suggest that the ‘bringing forward’ phenomenon does not exist but such a simple interpretation lacks persuasion. It seems inconceivable that there would be no such scheduling of operations. Can the authors comment on the number of operations in the months immediately after infants’ first birthdays and whether there is any unexpected shortfall?

David Draper (University of California, Santa Cruz)
The topic of this interesting paper is quality assessment: the authors are trying to identify ‘bad’ institutions, whatever ‘bad’ might mean. The approach is largely what might be termed input-output analysis (e.g. Draper (1996)): rather than attempting principally to measure the quality of the processes of care in the 12 surgical centres (because data of this kind are expensive to collect and therefore sparse; Section 6.3 provides some process evidence, but this is limited in scope), the authors were asked to compare the outputs or outcomes of the centres (here, 30-day mortality) after adjusting for the inputs, which in this context would involve measuring patients' sickness before surgery. Detailed information of this kind was not available; the authors had to content themselves mainly with stratifying on patients' ages and type of surgical procedure.

The data on which conclusions must be based in this case are clearly of a lower standard than would be desirable, but the UK Government was placed—by the egregious unadjusted mortality rates at the Bristol Royal Infirmary—in a situation in which ‘something must be done’, and the authors have been admirably careful, with their walter of sensitivity analyses and comparisons between data sources, not to overstep the bounds of what may reasonably be concluded. For future work of this kind, however, consideration of two points may be worthwhile.

1. Simulation of null models, and non-null models featuring different kinds of biases and different sizes of underlying effects, can be a valuable tool, both to calibrate procedures that identify ‘good’ and ‘bad’ institutions and to convince sceptics that differences, when found, are real. In Section 8.3 the authors raise the possibility of excessive false positive or false negative results in performance monitoring based on estimates derived from hierarchical models; these error rates may be estimated effectively with simulation methods, and it is not unusual to find with small numbers of institutions that the actual performance of 95% intervals and (say) 1% P-values differs noticeably from nominal levels in null situations (see Draper and Gittoes (2001) for calibration work along these lines, on behalf of the Higher Education Funding Council of England, in the construction of performance indicators in higher education).

2. Where should the lines be drawn in identifying ‘good’ and ‘bad’ institutions in a way that sensibly trades off false positive and negative findings? It would be salutary to see a careful answer to this question—e.g. via maximization of expected utility in a thorough application of Bayesian decision theory—that made a serious attempt at quantifying the real costs of failing to identify bad institutions and incorrectly blowing the whistle on good ones.

Steve Kendrick and Graham Mitchell (National Health Service Scotland, Edinburgh)
The primary purpose of the paper by Spiegelhalter and his colleagues is to describe the experience of statisticians in the Bristol inquiry. The irony of course is that we must hope that this precise experience is not required again in the context of 'another Bristol'.

We must hope rather that a subsidiary focus, highlighting the potential statistical role in future monitoring of clinical performance, as discussed at greater length in Spiegelhalter et al. (2000), will be of more practical relevance.

There is a fundamental tension here and it is concerned with context. The Bristol inquiry as a context for statistical analysis is entirely different from prospective monitoring of outcomes aimed at 'early warning'. The Bristol inquiry took place in the media spotlight on a quasi-legal adversarial basis (despite attempts to avoid this). In these circumstances, it was imperative that the most rigorous