Appendix 1: Case identification and data management [posted as supplied by author]

Assistance provided by the GMC

Until November 2004, when the Fitness to Practise panel was introduced, the GMC had a standard 3-stage procedure for the investigation of complaints. Each was first reviewed by a medical and a lay screener, who decided whether they were satisfied that the issue did not raise questions over the doctor's fitness to practise (in which case it was dismissed), or required further investigation. Complaints in the latter group were referred confidentially to the Preliminary Procedures Committee (PPC). From there, five potential outcomes were possible:

- no action
- a letter to the doctor
- referral for management of health problems
- referral to the Professional Conduct Committee (PCC)
- voluntary erasure from the medical register

Our interest was in those doctors referred to the PCC and subsequently found to have demonstrated serious professional misconduct, at which point their names, and details of their hearings, are publicly available.

The GMC agreed to:

- Identify 'cases' from the participating medical schools. These doctors would
  - Have 'Finding in Fact' proved against them within the date range 1999-2003, or been admonished/reprimanded or erased in 2004 following earlier PCC hearings
  - Not have an associated medical problem
  - Not have requested voluntary erasure from the Medical Register
- Identify four controls from the same medical school and graduation year, chosen by systematic sampling (see below), who were currently fully registered to practise and had not been investigated by the GMC at any time.
- Liaise confidentially with the medical schools, providing the names of all cases and controls (but not their case/control status), together with a unique ID number.
- Fund the collaborating Medical Schools for clerical time and consumables.
- After completion of data entry at Nottingham, provide the case/control status of the doctors, linked only to the unique ID number, and their current status within the GP or Consultant registers.
- Indicate the type of alleged professional misconduct for each case.

The systematic sampling procedure devised by the GMC for the selection of controls was as follows:
1. For each case, all eligible 'control' doctors matching the above criteria were listed alphabetically by surname
2. The number of 'eligible controls' listed was divided by the number of controls required (ie 4 x number of cases for that medical school graduation cohort)
3. This exercise resulted in blocks of eligible controls, four blocks for every case
4. The first named doctor in each block was chosen as the formal control for that case
5. All cases and matching controls from any one university were combined, sorted by surname, and allocated an ID code

Assistance provided by partner medical schools

Medical schools in which student records were retained for 30 or more years were contacted in late 2007 and asked if they would be willing to participate in this research. They agreed to:

- Obtain reciprocal ethical approval for the study
- Provide a named lead to liaise with the GMC
- Locate, on receipt of the names of cases and controls, the relevant student records, and copy each one so that the identity of the person was fully concealed
- Place each anonymised copy in an envelope marked with the unique ID number and send the copies securely to Nottingham

Data collection and management

A customised database was created in MS Access, designed to receive data on:

- The personal characteristics of each student (ID code, university, age, sex, UK or overseas domicile, and parental occupation if known). Age was categorised as under 21 or 21 and above.
- Overall progress during the medical course, (entered as free text, exam grades if available, and a note of any exam failures)
- Year of exit from the course and whether any years had been repeated. If a student had taken an intercalated degree (requiring an additional year's study) this was also noted.
- Any significant remarks or comments made during the student's years on the course, typed verbatim into text boxes, then flagged as 'adverse comments'.
- Information about satisfactory completion of the pre-registration year, if this was available in the student's file.

The academic references included in UCCA or UCAS forms were reviewed for 'negative comments' and these data were collected in a separate database. See Supplement 3 for full details.

Estimation of Social Class

Parental occupation of the father, as supplied by students, was typed into the database verbatim and used to estimate social class, using the traditional 5-category Registrar General's Scheme.[12] [ref to SC scheme At the end of data collection these data were used to categorise each student's background according to the traditional 5-category Registrar General's Social Class system, which is based on occupations, ie

Class I:	Professional occupations (traditional)
Class II:	Managerial & Technical (new professionals, managerial, employers, higher and intermediate technical, own account agricultural)
Class IIIN:	Skilled non-manual (higher supervisory, intermediate, and semi-routine clerical)
Class IIIM:	Skilled manual (lower supervisory, lower and routine technical, non-professional own account workers
ClassIV:	Semi-skilled (lower technical, semi-routine service and operative, routine sales and service, and agricultural)
Class V:	Unskilled (routine operative)

For a detailed breakdown see
http://www.ons.gov.uk/about-statistics/classifications/current/ns-sec/continuity-issues/index.html

These categories were allocated by printing a report showing only the unique study ID and the occupation, then coding on paper before adding the code back to the database. This was done independently by three people (DJ, JY and DY) and the final coding obtained by triangulation, ie the 'majority view' if two or more codings were the same, or average if all three were different. Because the categories are somewhat pragmatic and have been changed over time, they were collapsed further into Higher (I, II or III) or Lower (IV or V). If only the mother's occupation was shown, this was coded as 'not known'.

Examination & Course Progress Data

A major potential problem was that of handling course examination data. We knew that it would not be possible to collate and analyse numerical data from such varied sources. Course length varied between 5 and 6 years and the progress data were infinitely variable. To overcome this, we compared all the examination and progress data for each group of five students (case + 4 controls, although we did not know which was the 'case'), and flagged any that were obviously poorer than the rest as 'poor performance'. We did this for the 'early/preclinical' and the 'later/clinical' parts of the course; these divisions were somewhat arbitrary but in general corresponded to years 1-2 and 3-5, respectively, for 5-year courses, and years 1-3 and 4-6 for 6-year courses. We took particular care to note whether students had spent longer than usual on the course, either because of the completion of an intercalated degree or because of poor progress and the need to repeat areas of study.

We did attempt to extract School examination results and other pre-admission data from the students' files. However, many pre-dated the University & College Admission Service (UCAS), so there was a lack of consistent information and we are not reporting from this source.