UKMED Supporting Sector Wide Analysis - Governance and Guidelines

UKMED
UK Medical Education Database

Introduction

1. UKMED brings together a substantial wealth of administrative data from across the UK to understand how doctors progress through their medical career, from application to medical school to employment in the health service.

2. Organisations working in the medical education sector can greatly benefit from using UKMED data. While most of these organisations will apply to access UKMED data through the research applications process, there are some sector wide analyses and reviews that can be enhanced by using UKMED data. These types of analyses and reviews do not fit the criteria for access through the:

   2.1. UKMED Research Projects. Research projects outputs are required to be of a standard for publication in an academic journal.

   2.2. UKMED Training Pathways analyses. The limited dataset available through this route only enables descriptive analysis of students and doctors' movement through training pathways, for the purpose of monitoring or planning education and training of doctors.

3. The UKMED Advisory Board has explored how UKMED can support the sector wide reviews and recognised the value that UKMED data can offer.

Legal framework

4. UKMED was established to enable the General Medical Council to carry out its regulatory functions, by supporting research into areas where the GMC has a statutory responsibility. The GMC's functions are set out in the Medical Act 1983 and include:

   4.1. Promoting high standards of medical education and co-ordinating all stages of medical education

   4.2. Establishing and maintaining standards of postgraduate medical education and training

   4.3. Developing and promoting postgraduate medical education

5. The GMC also has statutory functions relating to the maintenance and publication of the medical register, providing guidance on standards of professional conduct.

6. When carrying out its functions, the GMC is able to co-operate with other public bodies concerned with the training, employment and regulation of medical professionals. Those bodies will also have their own statutory responsibilities which may allow them to use data provided by

the GMC. Extracts can only be shared under this protocol if the GMC is satisfied there is a legal basis for the data sharing.

Data excluded from extracts

7. Some data fields may be excluded from the extract. The data fields to be excluded will be determined upon review of the application to ensure that the amount of data released is adequate, relevant and limited.

Proposed governance and access arrangements

8. These extracts will not be used for testing specific hypotheses and some uses of the population extract will not be published. The UKMED Research Projects process is therefore not appropriate for these extracts.

9. To ensure the Advisory Board has sight of all data being shared through UKMED and can see public uses of these data we are proposing arrangements very similar to those currently used for research extracts. The same contractual constraints and requirement to use the Safe Haven would apply.

10. While the output from sector wide reviews and analyses may not be published in academic journals, we understand that these types of output have the potential to influence policy and decision making in the sector. Only after specific criteria has been fulfilled will access be given to organisations seeking to use UKMED data through this route:

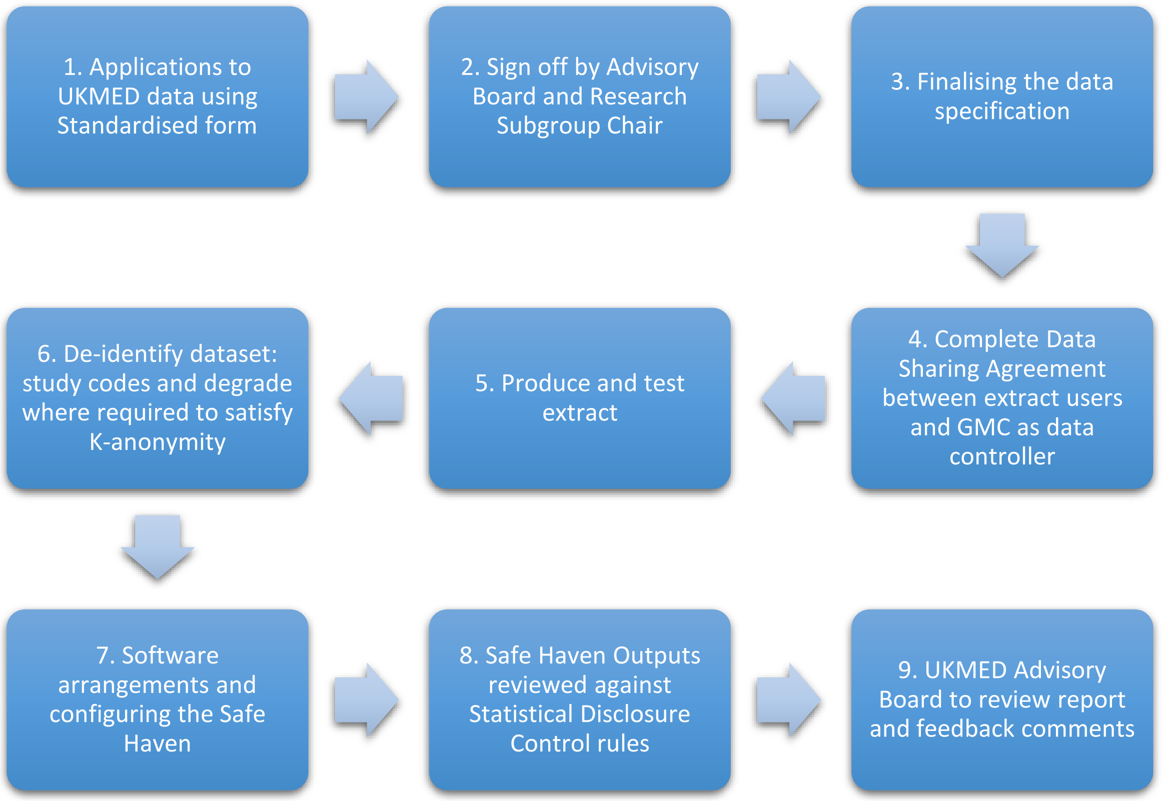
    10.1. Organisations or the associated Review Committee must have a responsibility for medical education, training and workforce development in the UK.

    10.2. Requests to use UKMED data must be supported by senior members of staff at the organisation or a formal request by the associated Review Committee.

    10.3. The purpose of the review/analysis must have a clear output and implication for policy development and demonstrates that UKMED data can help provide the evidence base.

    10.4. Access to UKMED data will cease once the review/analysis has been completed.

Figure 1 – Access arrangements


1. Applications to UKMED data using Standardised form
2. Sign off by Advisory Board and Research Subgroup Chair
3. Finalising the data specification
4. Complete Data Sharing Agreement between extract users and GMC as data controller
5. Produce and test extract
6. De-identify dataset: study codes and degrade where required to satisfy K-anonymity
7. Software arrangements and configuring the Safe Haven
8. Safe Haven Outputs reviewed against Statistical Disclosure Control rules
9. UKMED Advisory Board to review report and feedback comments

Stage 1 - Applications to UKMED data using Standardised form

11. Applications to use UKMED data through this route will be required to fill out an application form under the domains in Table 1.

Table 1 – Domains for applications

Domain	Description
Purpose of the review/analysis	**Requirement:** The review has significant implications for policy or practice in medical education, training and workforce development in the UK.
Data fields required	Data requested is contained within UKMED and is suitable for the review/analysis.
Proposed methodology	Outline of the methodology and approach to using the data appropriately to achieve the purpose of the review/analysis.

Analysis	Outline of the analysis to be undertaken.
Planned output and use	If the work is to support a particular review or working group, then a formal request from Committee chair and Terms of Reference must be included.
Team	Description and details of the individuals requiring access to the UKMED.
Evidence of support	**Requirement:** The review/analysis must be sponsored by senior members of staff.

Stage 2 – Sign off by Advisory Board and Research Subgroup Chair

12. The UKMED Advisory Board and Research Subgroup Chair will review the application and approve the release of data. The Chairs may seek advice, where necessary and appropriate, from members of Research Subgroup or members of the Advisory Board.

Stage 3 – Finalising the data specification

13. The UKMED analyst will work with the approved researcher to complete a final specification of the dataset to be used. The specification will be included in the Data Sharing Agreement.

Stage 4 – Complete Data Sharing Agreement between extract users and GMC as data controller

14. Once the specification is finalised the GMC as Data Controller will issue a Data Sharing Agreement. This will contractually restrict the extract user's use of the data to the agreed purposes. It is important to note that the data cannot be used to support measures or decisions with respect to particular individuals, and cannot be processed in such a way that substantial damage or substantial distress is, or might be, caused to any data subject.

15. If the data set includes data not in the Standard Workforce planning extract - https://www.ukmed.ac.uk/documents/UKMED_Training_pathway_analysis_extracts_reports.pdf The GMC will check with the data contributor that they are content for their data to be released via this route.

Stage 5 – Produce and test extract

16. The GMC will produce the extract to the agreed specification, ensuring that the methodology of production is documented.

Stage 6 – De-identify dataset: study codes and degrade where required to satisfy K-anonymity

17. When providing row by row data, we will pseudonymise individual doctors. Each GMC Reference number contained within the dataset will be replaced by a unique study code. If the dataset

contains multiple records with the same GMC number, these records will have the same unique study code. The unique study code will consist of a concatenation of the project code assigned on approval and a consecutive number. The GMC will hold a table that maps GMC numbers to study codes (STUDY_ID) to allow re-identification in the event of the data being queried. This table will only be accessible to analysts working on the UKMED project. The IDs will change if a new extract is issued. Old extracts will be archived and will not be available to Safe Haven users.

18. The GMC will ensure that individuals cannot be identified using a combination of demographic variables, specialty registration or employment details using data minimisation technique by applying the concept of K-anonymity. This is satisfied if K > 1 for each combination of quasi-identifiers – gender, age, medical school and so forth[1]. To achieve this, it may be the case that some values will be recoded into broader categorisations. We will minimise any reduction in utility by recoding the variables least relevant to the main purpose of the report. If other techniques are used these will be outlined.

19. Data minimisation will have to consider the risks of re-identification that arise from including data in the extracts that are also publicly available, in particular the data on the List of Medical Practitioners and data on employment location.

20. The GMC will maintain an archive of the extracts issued. To avoid additional complexity in satisfying K-anonymity, archived files will not be available in the Safe Haven. The archive is only maintained for any queries regarding outputs.

Stage 7 – Software arrangements and configuring the Safe Haven

21. Extract users will be completing their analysis in the University of Dundee's Health Informatics Centre (HIC) Safe Haven. Users will complete a HIC/GMC Data User Agreement, which the GMC will countersign.

22. Users will need to complete a short course on Data Protection before accessing the Safe Haven and provide evidence of completion to HIC. The course "Research Data and Confidentiality" can be found at: http://byglearning.co.uk/mrcrsclms/course/category.php?id=1.

23. Users will be remotely logging onto a secure server located within HIC to access data and perform analysis, without being able to copy or remove the data from the secure central server.

24. The remote-access Safe Haven utilises a VMware secure environment. In this model data are no longer released externally to researchers for analysis on their own computers but placed on a server at HIC by the GMC, within a secure IT environment, where the user is given secure remote access to analyse it. Researchers will need to install the VMware client on their machine or access via http to use the Safe Haven.

25. The GMC supply the data to HIC and GMC will be responsible for all queries regarding the data. Users will have a named point of contact at the GMC for this purpose. The GMC will transfer files to HIC via a secure file transfer. Within 48 hours HIC will transfer these files to the Safe Haven environment (except during the 2-week Christmas/New Year period when there will be no Safe Haven support available).

[1] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 571-588.

26. HIC are responsible for managing access to the Safe Haven and working with the users to ensure the required software is available. The GMC are responsible for answering any queries on the data supplied.

27. All software within the Safe Haven is licenced for academic research only, unless connected to an academic institution, it is likely that there will be an additional cost to population extract users.

28. For software that is not included as standard and where HIC Safe Haven can support it, extract users must buy the necessary licence along with the software media (to allow installation) (see table 3 for exceptions for SAS, SPSS and STATA) and pay HIC a £250 installation fee per install.

Stage 8 – Safe Haven Outputs reviewed against Statistical Disclosure Control rules

29. Extract users will be completing their analysis in the University of Dundee's Health Informatics Centre (HIC) Safe Haven. Users will complete a HIC/GMC Data User Agreement, which the GMC will countersign.

30. Tabular data: When UKMED users have completed their analysis, outputs intended for the public domain, for example a table of results, will be reviewed by the GMC using the following statistical disclosure controls:

    30.1. 0, 1, 2 are rounded to 0;

    30.2. All other numbers are rounded to the nearest multiple of 5;

    30.3. Percentages based on fewer than 22.5 individuals are suppressed;

    30.4. Averages based on 7 or fewer individuals are suppressed.

    30.5. The requirements relate to headcounts, Full-Person Equivalent (FPE) and Full-Time Equivalent (FTE) data. Financial data are not rounded.

31. Charts and visualisations: Charts and visualisations should also apply statistical disclosure controls to reduce the risk of reidentification. Consider:

    31.1. Histograms

        31.1.1. Aggregating categories to distributions with long tails and small numbers;

        31.1.2. Masking the scale of the X and Y axis by omitting all values and using the maximum or minimum values;

    31.2. Box plots and scatter plots

        31.2.1. Grouping together or averaging the minimum and maximum values as outliers may be attributable to a single observation.

        31.2.2. Grouping together or averaging the original data to generate scatter plots.

32. Data output requests are processed once per day, between the hours of 9:30 and 11:30 on work-days (except during the 2-week Christmas/New Year period when there will be no Safe Haven support available). All requests made in the previous 24hrs will be processed during this period and shared with the GMC. GMC will review the files in line with statistical disclosure controls and if approved, share the output analysis files with researchers via GMC Connect within 2 working

days. Users are strongly encouraged to leave sufficient time in their plans for their output to be reviewed before being passed to them.

33. Output intended for external consumption (for example published on an organisation's website) must be reviewed by email prior to publication. Review will be undertaken by email by persons nominated by the Advisory Board with a four-week turnaround time. All external outputs must contain a clear statement on methodology, in particular criteria for inclusion in the cohorts and the details of the derivation of any variables used. Users will be expected to share derived variables with other extract users.

34. External publications will need to acknowledge UKMED and HESA as the data source using the following statement:

    34.1. "This report uses data from UKMED (www.ukmed.ac.uk). UKMED uses data from the Higher Education Statistics Agency Limited Source: HESA Student Record 2002/03 to 20XX/XX Copyright Higher Education Statistics Agency Limited. Neither the GMC (the data controller for UKMED) or The Higher Education Statistics Agency Limited can accept responsibility for any inferences or conclusions derived by third parties from data or other information supplied by it."

Stage 9 – UKMED Advisory Board to review report and feedback comments

35. Reports and any outcomes using UKMED data will be sent to the UKMED Advisory Board for comments. The Advisory Board will have the opportunity to comment on the use of UKMED data in the review and analyses.