UKMED Standard Extracts

Daniel Smith

Kirsty White

V2 – September 2020

Change control

1 V2 includes UCAS data

Introduction

2 One of the key rationales for bringing UKMED data together is that understanding individuals' performance at different points during their study and medical career is helpful to understand the factors that make doctors more or less likely to progress through the training pathways.

3 In Phase 1 we explored whether we could develop a model that would identify the 'value-add' between entry to medical school and graduation but concluded that only limited capability existed within the current database and that a medical licensing assessment may offer greater potential to understand the learning trajectories of cohorts of students within and across medical schools.

4 Having discussed next steps with Advisory Board members, we have identified that there would be value in producing standard datasets that would enable descriptive analysis of students and doctors' movement through training pathways.

5 Two types of extract and accompanying reports have been identified:

a Medical school entry profiles describing the demographics of medical school cohorts.

b A workforce planning extract for those with responsibility for monitoring or planning education and training of doctors.

c A summary of both extracts is described in Table 1 below and a list of fields included in **Appendix A.**

Medical school applicant and entry profiles

6 This dataset would contain demographic information from HESA on applicants and entrants to medical school including information linked on postcode such as the young participation classification (POLAR3) and UKCAT bursary information to assist track widening participation initiatives.

7 This would be used by the Medical Schools Council Selection Alliance as part of their selection monitoring work stream and by the GMC for its work on differential attainment and to enhance information provided through its *State of Medical Education and Practice* publications.

Workforce planning extract

8 This extract would contain wide-ranging data: applicant profiles and applications from UCAS, entry profiles (from HESA and UKCAT Bursary), GMC survey census data describing doctors' specialties and training grades, Annual Review outcomes identifying doctors who do not progress and new data on where doctors work upon completion of training.

9 NHS Education for Scotland (NES) is a special NHS Board with national responsibility for education, training and workforce development for those who work in and with NHS Scotland.[1] Since 1 April 2015 Health Education England has been a Non-Departmental Public Body (NDPB) under the provisions of the Care Act 2014. Under Section 97 of this act HEE has responsibility for workforce planning[2]. Initial conversations with NES suggest that their primary interest is in understanding cross-nation flows by comparing students' country of domicile on entry to medical school to the country they go on to train in and finally work in.

10 The Department of Health (DH)'s role is to set overall policy and strategy direction for the health and social care system, developing evidence-based policies in partnership with arm's length bodies. The Workforce Directorate in DH aims to ensure that we have the right number and mix of staff, in the right place at the right time to deliver patient care. Within that the Workforce capacity and Analytics team currently accesses rich data on an individual anonymised basis for the HCHS (Hospital and Community Health Services) workforce.

11 The workforce capacity and analytics team propose to use UKMED data to better understand the career paths of doctors and the impact of policy and strategy, both backwards looking monitoring of the impact of policy changes and forwards looking for policy development. The current model for accessing UKMED data (research applications) cannot be used by the team because they could not always commit to publishing outputs due to the confidential nature of some policy development and on occasion the need to work to very short-term deadlines.

The legal framework

12 In addition to the GMC's statutory duties to set and secure standards for medical education and training, paragraph 9A(1)(b) of Part II of Schedule 1 of the 1983 Medical Act gives the GMC a statutory duty to co-operate, in so far as is appropriate and reasonably practicable, with public bodies or other persons concerned with:

a "(i) the employment (whether or not under a contract of service) of provisionally or fully registered medical practitioners,"

b "(ii) the education or training of medical practitioners or other health care professionals"

The users of the population extract would meet this definition of public bodies we are obliged to co-operate with and the extract would be designed to allow analysis for the purposes of workforce planning.

Data that will not be included in the standard extracts

13 The following types of data that will not be included in the standard extracts:

13.1 Measures of attainment on selection tests used by medical schools (UKCAT, GAMSAT, BMAT),

13.2 Measures of attainment at school (A-levels),

13.3 Medical royal college membership exam results,

13.4 National Training Survey responses

13.5 Fitness to Practise data.

Table 1 Proposed annual extracts

Extract name	Tableau report	Description	Data sources Appendix A for field lists	Users
Medical school applicant and entry profiles		Extract provides demographic information on entrants to medical school including information linked on postcode such as the young participation classification (POLAR3).	UCAS, HESA and UKCAT Bursary field	Data monitoring work stream for the MSC selection alliance board includes MSC data analyst and nominated medical school users. GMC analysts (beyond UKMED analysts) will have access to the de-identified dataset in the database.
Workforce Planning	To be developed. This will report on employment outcomes, for instance geographical location, area of work and type (e.g. locum) for those who have left training as defined by CCT or ARCP outcome 4.	The extract will be provided in two tables due to the differences in refresh frequency. It will be possible to link between the tables using STUDY_ID. Annual extract: VW_UKMED_HESAPROGRESSION VW_UKMED_PERSON NTS_TRAINEE This gives Demographic data from HESA on entry to medical school, including the out code[3] part of student's postcode on application to medical school. Details of where the	UCAS_APPLICATIONS VW_UKMED_PERSON_APPLICANT VW_UKMED_PERSON HESA and UKCAT Bursary NTS_TRAINEE GMC NTS census Data collected from LETBs and Deaneries ARCP_OUTCOMES ARCP outcomes collected from LETBs and Deaneries. PRACTICEHISTORY contains GMC data collected for revalidation purposes originally provided from the following payroll systems: by the four Departments of Health ESR – Electronic Staffing Records PCIS – Primary Care Information System	Workforce Capacity and Analytics Quarry House Department of Health Dr Louise Plewes Scottish Government - Health Workforce & Strategic Change Directorate Dr Emma Watson Planning (Medical) – Directorate of Strategy & Planning Health Education England - John Stock NHS Education for Scotland - Analysis, Intelligence and Modelling Steering Group members[6] including Dr Colin Tilley, Programme Director, Workforce Analysis Intelligence Dr Stewart Irvine, Director of Medicine

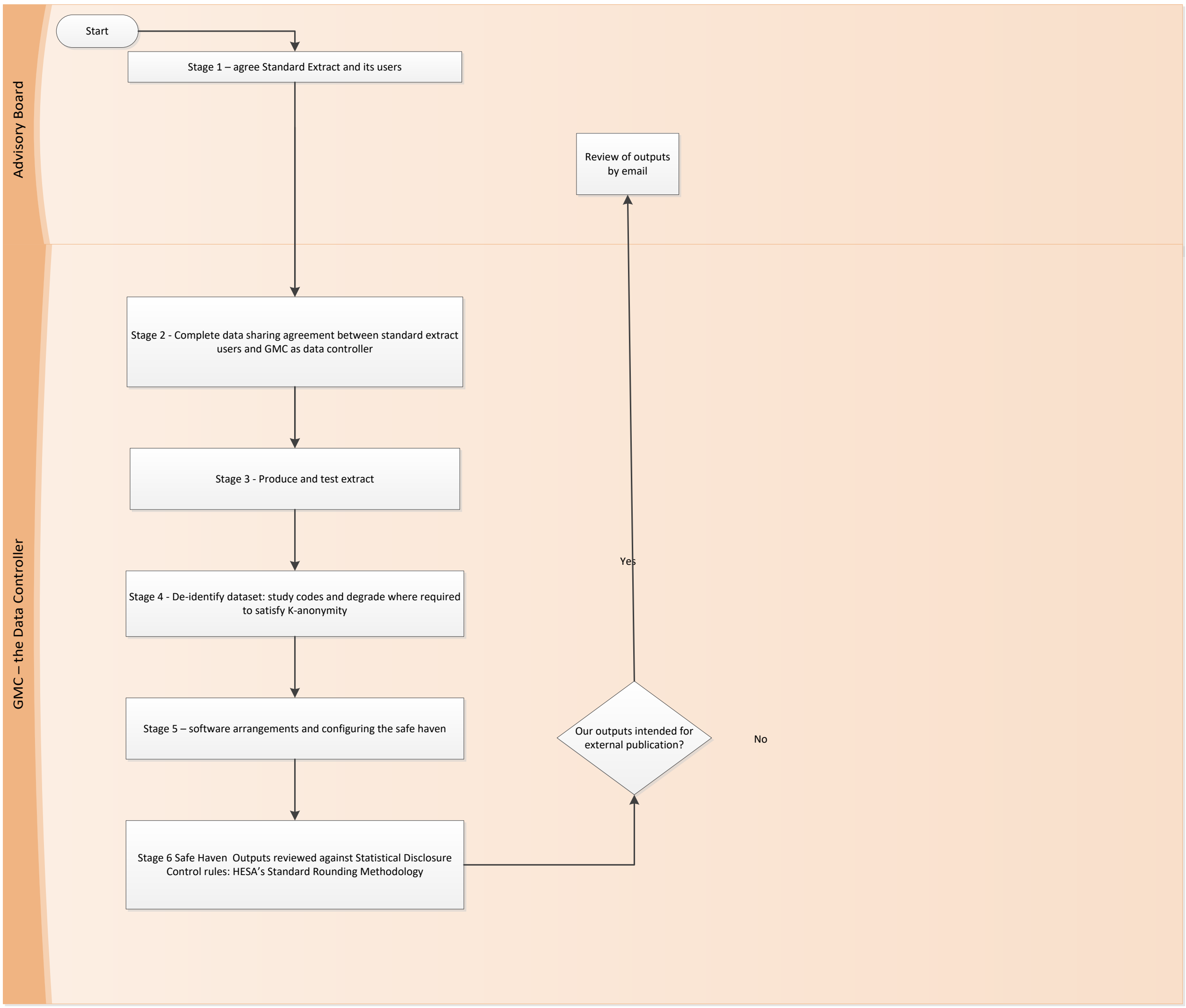
Extract name	Tableau report	Description	Data sources Appendix A for field lists	Users
		trainee was during their postgraduate training including the following: specialty, level and deanery as captured annually on the NTS Census.[4] Quarterly updates: VW_UKMED_PRACTISE_HISTORY Details of where the doctor has worked and the job role from PRACTICEHISTORY and ORGANISATION which contains data from NHS payroll systems. VW_UKMED_SPECIALTYREG Details of specialty and GP registration entries. Contains one row per instance of practice history.	SWISS – Scottish Workforce Information Standard System ISD Scotland's GP Contractor Database will be included Northern Ireland Business Services Organisation – this has not been refreshed recently. A process to refresh is under discussion. ORGANISATION contains the details of each ORGANISATION in PRACTICEHISTORY CR_ONS_POSTCODE contains data from National Statistics Postcode Lookup (NSPL)[5] used to assign Government Office Region and CCGs using the organisation's postcode. The view contains one row per employment assignment. VW_UKMED_SPECIALTYREG Contains data from SPECIALITIES - entries to the specialist register. PERSON which contains information on GP register entries View contains one row per specialist register entry	GMC analysts in the GMC Intelligence Unit (beyond UKMED analysts) will have access to the de-identified dataset in the database. Medical Schools Council analyst

Proposed Governance and access arrangements

14 These extracts will not be used for testing specific hypotheses and some uses of the population extract such as some governmental workforce analysis will not be published. The research process is therefore not appropriate for these extracts.

15 However, to ensure the Advisory Board has sight of all data being shared through UKMED and is able to see public uses of these data we are proposing arrangements very similar to those currently used for research extracts[7]. The same contractual constraints and requirement to use the safe haven would apply. These are summarised in figure 1 and below.

Figure 1 – Access arrangements



Advisory Board
GMC – the Data Controller
Start
Stage 1 – agree Standard Extract and its users
Review of outputs by email
Stage 2 - Complete data sharing agreement between standard extract users and GMC as data controller
Stage 3 - Produce and test extract
Yes
Stage 4 - De-identify dataset: study codes and degrade where required to satisfy K-anonymity
Stage 5 – software arrangements and configuring the safe haven
Our outputs intended for external publication?
No
Stage 6 Safe Haven Outputs reviewed against Statistical Disclosure Control rules: HESA's Standard Rounding Methodology

Stage 1 – Agree Standard Extract and its users

16 The UKMED Advisory board will have sight of the fields and definition to be included in each extract. The specification of each extract will be reviewed annually. The extracts as currently defined are in Appendix A.

Stage 2 - Complete data sharing agreement between standard extract users and GMC as data controller

17 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[*].

Stage 3 - Produce and test extract

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

19 Quarterly update files for VW_UKMED_PRACTISE and VW_UKMED_SPECIALTYREG will contain the entire dataset not just the delta.

Stage 4 - De-identify dataset: study codes and degrade where required to satisfy K-anonymity

20 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. Study codes will only be used for one annual cycle of extracts, the same study codes will be used for tables that are refreshed quarterly during the year. This table will only be accessible to analysts working on the UKMED project. The same STUDY_ID will be used for the year and will be constant across the quarterly updates. The IDs will change when a

[*] See section 33 of the Data Protection Act (1998) here: http://www.legislation.gov.uk/ukpga/1998/29/section/33

new annual extract is issued. Old extracts will be archived and will not be available to safe haven users.

21 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[*]. 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.

22 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. [†]

23 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 5 – software arrangements and configuring the safe haven

24 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.

25 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/mrcrsc-lms/course/category.php?id=1.

26 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.

27 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

[*] See 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. http://dataprivacylab.org/dataprivacy/projects/kanonymity/kanonymity2.html

[†] http://www.gmc-uk.org/doctors/register/LRMP.asp

[‡] https://medicine.dundee.ac.uk/sites/medicine.dundee.ac.uk/files/Safe%20haven%20User%20Guide.pdf

computers but placed on a server at HIC by the GMC, within a secure IT environment, where the researcher 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[*].

28 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).

29 Previously written customised code/syntax, libraries of reference data and so forth can be imported once approved by the GMC.

30 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.

31 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.

32 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 6 Safe Haven Outputs reviewed against Statistical Disclosure Control rules: HESA's Standard Rounding Methodology

33 When the user has 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[†]:

a 0, 1, 2 are rounded to 0

b All other numbers are rounded to the nearest multiple of 5

c Percentages based on fewer than 22.5 individuals are suppressed

d Averages based on 7 or fewer individuals are suppressed

[*] https://medicine.dundee.ac.uk/sites/medicine.dundee.ac.uk/files/Safe%20haven%20User%20Guide.pdf

[†] https://www.hesa.ac.uk/content/view/146

- **e** The above requirements relate to headcounts, Full-Person Equivalent (FPE) and Full-Time Equivalent (FTE) data Financial data is not rounded.

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

35 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. This part of the process will be reviewed each year to ensure it is possible to resource and proportionate.

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

> "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 2014/15 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."

V2 of the workforce planning extract will be available from September 2020 and thereafter refresh every quarter.

Appendix A – Extract specification

For details of the tables please see https://www.ukmed.ac.uk/documents/UKMED_data_dictionary.pdf

UCAS_APPLICATIONS

VW_UKMED_PERSON_APPLICANT [without UKCAT data]

NTS_TRAINEE

VW_UKMED_PERSON

VW_UKMED_HESAPROGRESSION

VW_UKMED_PRACTICEHISTORYV

W_UKMED_SPECIALTYREGARCP_OUTCOMES

ARCP_OUTCOMES

End notes

[1] NHS Education for Scotland (2014) *A refreshed strategic framework for 2014-19* Available at:

[2] http://www.legislation.gov.uk/ukpga/2014/23/section/97/enacted

[3] http://www.bph-postcodes.co.uk/guidetopc.cgi

[4] Example from 2016- http://www.gmc-uk.org/NTS_2016___Briefing_Note_2_FINAL_V3.pdf_63386727.pdf there is a similar collection notice for each year.

[5] Available from http://geoportal.statistics.gov.uk/
Office for National Statistics (Edition:February 2017) *National Statistics Postcode Lookup User Guide* Available from:
https://www.arcgis.com/sharing/rest/content/items/8e409cfe4d0a4971986343f3919021e3/data

[6] http://www.nes.scot.nhs.uk/education-and-training/by-theme-initiative/analysis,-intelligence-and-modelling/aim-steering-group.aspx

[7] UK Medical Education Database (UKMED). *UKMED Process for completing UKMED Research v2 December 2016* Available at: http://www.ukmed.ac.uk/documents/UKMED_research_process.pdf [Accessed 28 December 2016]