2020-01-01

Evaluating the recruitment process into UK anaesthesia core training: a national data linkage study of doctors' performance at selection and subsequent postgraduate training

Aslet, M

10.1136/postgradmedj-2019-136390
Postgraduate Medical Journal
BMJ Publishing Group

Postgraduate Medical Journal

Evaluating the recruitment process into UK anaesthesia core training: doctors' performance at selection and subsequent postgraduate training- a national data linkage study

Journal:	*Postgraduate Medical Journal*
Manuscript ID	postgradmedj-2019-136390 .R1
Article Type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Aslet, Margaret; University of York, Department of Health Sciences Paton, Lewis; University of York, Health Sciences Gale, Thomas; University of Plymouth, Faculty of Medicine and Dentistry Tiffin, Paul; University of York, Department of Health Sciences; Hull York Medical School
Keywords:	ANAESTHETICS, MEDICAL EDUCATION & TRAINING, EDUCATION & TRAINING (see Medical Education & Training)

Evaluating the recruitment process into UK anaesthesia core training: doctors' performance at selection and subsequent postgraduate training- a national data linkage study

Aslet, M. E. D.,[1] Paton, L. W.,[1] Gale, T.,[2] & Tiffin, P. A.[1,3]

1: Department of Health Sciences, University of York. York, United Kingdom

2: Faculty of Medicine and Dentistry, University of Plymouth, Plymouth, United Kingdom

3: Hull York Medical School (HYMS)

Correspondence to: paul.tiffin@york.ac.uk

Short title: **Evaluating the selection process into anaesthesia core training**

Abstract

Purpose of the study: To explore which factors increase the likelihood of being deemed appointable to core anaesthesia training in the UK and whether those factors subsequently predict performance in postgraduate training.

Study design: Observational study linking UK medical specialty recruitment data with postgraduate educational performance, as measured by Annual Review of Competence Progression (ARCP) outcomes. Data were available for 2782 trainee doctors recruited to anaesthesia core training from 2012 to 2016 with at least one subsequent ARCP outcome.

Results: Both higher interview and shortlisting scores were independent and statistically significant ($p \leq 0.001$) predictors of more satisfactory ARCP outcomes, even after controlling for the influence of postgraduate exam failure. It was noted that a number of background variables (e.g. age at application) were independently associated with the odds of being deemed appointable at recruitment. Of these, increasing age and experience were also negative predictors of subsequent ARCP rating. These influences became statistically non-significant once ARCP outcomes associated with exam failure were excluded.

Conclusions: The predictors of 'appointability' largely also predict subsequent performance in postgraduate training, as indicated by ARCP ratings. This provides evidence for the validity of the selection process. Our results also suggest that greater weight could be applied to shortlisting scores within the overall process of ranking applicants for posts.

What is already known on the subject

- Validity evidence in relation to recruitment to other medical specialties suggests that ratings on selection assessments are largely predictive of future performance
- One regionally-based UK study of selection into anaesthetics training reported similar findings, providing some evidence for the validity of the recruitment process, though national studies are lacking

Main messages

- In recruitment to anaesthetics core training shortlisting and interview scores were strongly predictive of postgraduate performance, as measured by Annual Review of Competence Progression (ARCP) outcomes.
- Our findings support the effectiveness of the UK selection process for anaesthesia core training.

Introduction

How medical trainees are selected into training programmes influences the quality of future consultants. Ideally, applicants should be selected on those factors that subsequently predict clinical competency. In the UK, the anaesthetics recruitment process (summarised in Figure 1) is coordinated by the Anaesthetic National Recruitment Office. Successful applicants then embark on their anaesthetics training (Figure 2).

INSERT FIGURE 1 HERE

INSERT FIGURE 2 HERE

Several studies have investigated how selection scores correlate with subsequent performance [1-6]. Several studies relating to General Practice have shown that performance during the selection process predicts later educational attainment [7-11]. One UK, regionally-based, study of selection into anaesthetics training reported a similar picture [12]. However, national studies are currently lacking.

There are a small number of national assessments in UK postgraduate training, including the Annual Review of Competence Progression (ARCP). This process is used to decide how a doctor progresses through postgraduate training and involves a panel of assessors reviewing a portfolio of evidence [13]. This includes supervisor feedback and workplace based assessments. Success also requires trainees to pass the Primary Fellowship of the Royal College of Anaesthetists exams (Primary FRCA). These exams consist of two, separately taken, sections. The first section consists of a multiple choice based knowledge test, and the second is a clinically orientated test involving an Objective Structured Clinical Examination (OSCE) and a Structured Oral Exam (SOE)[14]. Both of these sections must be passed during core training in order to progress to higher specialist training. Failure in these exams can lead to an unsatisfactory ARCP outcome, where extended training time may be required, or the trainee may be released from programme.

If the selection measures are reasonably valid then performance on such metrics should also predict subsequent outcomes in training, taking into account the impact of selecting out unsuccessful candidates. The role of influences on both performance at recruitment and in training should be taken into account where possible. These factors could include age [15], gender [16-19], country of primary qualification [20], and ethnicity [21-24].

This study aimed to assess the validity of the anaesthetics core training recruitment process by:

- Investigating the predictors of a candidate being deemed appointable to anaesthetics core training
- Evaluating the predictors of ARCP outcomes for those who enter postgraduate anaesthetics training
- Comparing any patterns observed above in order to assess the effectiveness of selection into postgraduate training.

Methods

Data sources and preparation

We obtained selection data for applicants to anaesthetics core training during 2012 to 2016 from the Oriel database supplied to the General Medical Council (GMC). Flow of data through the study is depicted in Figure 3. Our outcome measure for recruitment was 'deemed appointable' rather than 'appointed' in order to reduce the effect of differing competition ratios across different deaneries. Shortlisting and interview performance scores were standardised as z-scores by year and selection centre (i.e. transformed to have a mean of zero and standard deviation of one). This controlled for differences across selection centres and for time. Thus, if doctors had applied more than once in different recruitment years, the mean interview and shortlisting scores for that particular cohort were used to standardise the scores.

INSERT FIGURE 3 HERE

Data for ARCP outcomes for trainees across the UK were available from Health Education England. This was linked to a database supplied by the GMC that included socio-demographic information, and recruitment data. To maintain anonymity the linkage was performed, via the doctors' unique registration number, by the GMC. Only 'competency-based' ARCP outcomes were included (for example, those indicating 'out of programme experience' were excluded). Anaesthetic trainees were identified by their specialty, and those in ACCS Anaesthetics and CAT were coded separately to allow for comparison between the schemes.

ARCP ratings were collapsed and recoded to form a four point ordinal (ordered categorical) scale as follows: 4 = satisfactory, 3 = additional evidence required, 2 = targeted training required but no extra time, 1 = extended training time required/left programme. This allowed the ARCP outcomes to be treated as ordinal, rather than binary (satisfactory/unsatisfactory) or nominal (unordered) in nature. This means that the information contained within the ordering of the (recoded) ARCP outcomes is preserved, increasing study power. Note that, it was important to combine the categories 'extended training time' and 'left programme'. This is because previously it has been shown that, within the context of multilevel modelling (i.e. multiple ARCPs nested within doctors), these two categories cannot be distinguished. This is assumed to be because the need to leave a programme is almost always preceded by a period of extended training time. In addition, unlike Pyne and Ben-Shlomo [15] we included 'insufficient evidence provided' as an intermediate category of outcome. This is because it was previously noted that this outcome was associated with other undesirable outcomes (compared to a 'satisfactory' rating) and therefore was assumed to contain, on average, some information on a trainee. Thus, this approach to recoding and modelling ARCP outcomes has previously been found to be valid, and also leads to models that, at least approximately, fulfil the 'parallel odds' assumption that underlies ordinal logistic regression [25].

We included gender, age (on application and at ARCP), years of NHS experience and self-reported ethnicity (dichotomised as 'White' or 'Black and Minority Ethnic' (BME)) in analyses. We also hypothesised that being a national of, or having trained in, a country with fewer resources might also affect performance at selection or in training. Therefore we controlled for the Gross Domestic Product per capita in US Dollars (GDP) of the relevant countries of nationality and qualification, using 2008 World Bank data [26].

Statistical analyses

Stata 14.1 was used for data management and analysis. Logistic regression was used to model the odds of being deemed appointable to anaesthetics specialty training. 'Random effects' (multi-level) models were used to control for multiple applications, which allowed for nesting of application events within doctors.

We also modelled the odds ratios of obtaining a more versus less satisfactory ARCP rating using random effects ordinal logistic regression models. This accounted for dependency of observations within individual doctors. Analyses were conducted both with and without ARCP outcomes associated with postgraduate exam failure. Trainees who fail the fellowship exam would be highly likely to receive a 'sub-optimal' ARCP rating (for example 'extended training time required'). Therefore, there was a risk that postgraduate exam achievement could be conflated with, more general, ARCP performance. Therefore we conducted additional analyses which excluded the ARCP outcomes that were reported to have occurred in conjunction with postgraduate exam failure. This allowed us, to some extent, to disaggregate the effect of the predictors on ARCP outcomes in general from postgraduate exam performance.

For both sets of analyses, univariable and multivariable models (using a backward stepwise approach) were built. Only interaction terms that were statistically significant (at the $p < 0.05$ level) were included in the final multivariable model. Missing data were managed using listwise deletion. Extensive missing data (greater than 5%) were only observed for the shortlisting and interview scores at specialty selection. Therefore, in order to evaluate the potential impact of the missing data on the results a series of analyses using imputed values for these variables were conducted as a sensitivity analysis (see the technical appendix for further details).

Results

Descriptive statistics

For most variables, there were relatively few missing data, except for standardised shortlist and standardised interview score. 608/2782 (22%) of standardised interview scores were missing, and 1,075 /2782 (39%) of standardised shortlist scores were missing. There were a total of 2,782 doctors who had both recruitment and ARCP outcomes available. 52.30% (1,445/2782) were males, 23.79% (661/2779) described themselves as BME ethnicity, and 26.21% (713/2720) were on the ACCS Anaesthesia training scheme compared to CAT. Further descriptive statistics can be seen in Table 1.

INSERT TABLE 1 HERE

Modelling recruitment outcomes

Figure 4 displays the univariable odds ratios and 95% confidence intervals. The full results can be seen in Table S1 of the technical appendix. We can see the strongest predictor of recruitment is interview score (OR 4.70, 4.18 to 5.30, p<.001). Shortlisting score is a positive, though weaker, predictor (OR 1.29, 1.17 to 1.43).

In terms of demographic predictors, older and/or BME trainees had significantly lower odds of being deemed appointable, whilst experience and gender were not significantly predictive (see Figure 2, and Table S1 of technical appendix). Nationals of wealthier countries were more likely to be deemed appointable (OR 1.58, 1.44 to 1.73). That is, for every $10,000 USD GDP, a candidate's odds of success increased by approximately 60%. Similarly those who qualified in more affluent countries had higher odds of success (OR 1.86, 1.68 to 2.07).

INSERT FIGURE 4 HERE

The results from multivariable regression analysis are shown in Table 2. Since appointability of candidates is determined by their interview score, we excluded this variable from the analysis. Age at application (OR 0.94, 0.90 to 0.98), experience at application (OR 1.35, 1.25 to 1.47), BME applicants (OR 0.66, 0.49 to 0.90), and the GDP of place of qualification (OR 1.48, 1.05 to 2.10) were significant independent predictors of being deemed appointable. Experience at application was non-significant at univariable analysis, but significant at multivariable analysis. Shortlisting score was not an independent predictor of application success. No interaction terms were statistically significant.

INSERT TABLE 2 HERE

Some variation in the results from imputed and non-imputed datasets were observed. Notably, when the missing selection measures were imputed, shortlisting score became a significant predictor in the multivariable model, whilst experience at application became non-significant. The full results from analyses of the imputed datasets can be found in the technical appendix.

Modelling ARCP outcomes

Figure 5 displays the results from univariable analyses predicting ARCP outcomes, with and without postgraduate exam failures. Full results can be found in Table S2 in the technical appendix.

The strongest predictor of ARCP performance, when exam failures were included, was training within an ACCS programme. However, this becomes non-statistically significant when the effects of exam failures are eliminated. BME trainees had reduced odds of having a satisfactory ARCP outcome compared to white trainees (OR 0.61, 0.52 to 0.72). Older candidates (OR 0.90, 0.88 to 0.92), and those with more UK clinical experience (OR 0.79, 0.75 to 0.83) also, on average, had significantly less satisfactory ARCP outcomes. Once ARCP outcomes associated with postgraduate exam failure were excluded from analyses, the effect size reduced for all predictor variables. However, all the predictors, except for the training stream group (ACCS vs CAT) remained statistically significant at the $p<0.05$ level.

INSERT TABLE 3 HERE

The results from the multivariable regression analyses are shown in Table 3. Shortlisting (OR 1.42, 1.24 to 1.62) and interview scores (OR 1.37, 1.19 to 1.57) remained independent predictors of a more satisfactory ARCP outcome. For shortlisting score, for every standard deviation above the mean, trainees had approximately, on average, 40% higher odds of receiving a more satisfactory outcome when controlling for the influence of other variables.

Age (OR 0.94, 0.91 to 0.97), UK clinical experience (OR 0.90, 0.83 to 0.97) and BME status (OR 0.69, 0.55 to 0.86) remained independently predictive of less satisfactory ARCP outcomes when exam failures were included in the analysis.

When ARCP outcomes associated with exam failure were excluded, age, experience and training stream became non-significant predictors. Interview and shortlisting scores became slightly stronger predictors after excluding ARCP outcomes associated with postgraduate exam failure. BME trainees remained significantly less likely to obtain a satisfactory ARCP outcome (OR 0.72, 0.55 to 0.93).

Results for the imputed data were similar to those from the non-imputed (original) data. The full results are contained in the technical appendix. However, in summary, using imputed data, the influence of interview scores on 'appointability' were somewhat diminished (though still statistically significant at the $p<0.05$ level) whilst that of the shortlisting scores increased, though did not reach statistical significance ($p=0.11$). The results for the prediction of ARCP varied relatively little between those for the imputed and non-imputed datasets.

Discussion

This is the first national data linkage study to investigate the predictive validity of a national recruitment process in anaesthesia with measures of in-training clinical performance in training through ARCP outcomes. We found that although shortlisting scores were not predictive of appointability, they did predict ARCP success. Interview scores were not included in recruitment analysis, however they were also predictive of ARCP. At multivariate level, shortlisting and interview scores had very similar ability to predict educational success.

We did not detect a significant difference in ARCP outcomes between the two anaesthetic training streams (ACCS and CAT) once the influence of postgraduate exam failures was controlled for. This is likely to reflect the differing structure of training, where CAT trainees are more likely to take exams from the first year.

We observed that most demographic factors were potential confounding factors – only gender and years of NHS experience were not univariable predictors of recruitment. When confounders were accounted for using multivariable analyses, age at application, experience at application, ethnicity, and GDP of place of qualification remained significant independent predictors of appointability. Interestingly, greater experience was not a significant univariable predictor of appointability but emerged as an independent predictor in the multivariable model. This apparent paradox could be explained by greater years of experience also possibly occurring in trainees who had encountered previous difficulty in obtaining places on anaesthetics or other postgraduate programmes. It may also be a mechanism of the missing data present in the study (see strengths and limitations).

Poorer ARCP outcomes could be predicted by age, experience and ethnicity in both univariable and multivariable models. However, age and experience were not predictive once postgraduate exam outcomes were removed. Previous research has shown that junior doctors aged over 29 years old are more likely to have less than satisfactory ARCP outcomes across a range of specialties [15]. However, more recent research observed no significant difference between age and satisfactory ARCP outcomes within general surgery [27]. Postgraduate exam performance appeared more sensitive to increasing age than ARCP outcomes. This could be due to older trainees being more likely to have competing family responsibilities, making exam preparation more challenging, as has been suggested in previous literature relating to the first year of core training[15].

GDP, of both nationality and place of qualification, were predictors of appointability, although only the latter variable was an independent predictor. Both GDP variables were also significant univariable predictors of ARCP outcome, but neither were independent predictors. This is reassuring in an NHS that continues to rely on international doctors to sustain it.

Candidates, self-reporting as BME, had approximately half the odds of being deemed appointable compared to those reporting White ethnicity. This effect appeared independent of the other factors in the multivariable model. Similarly, BME trainees had higher odds of receiving less satisfactory ARCP outcomes, even when the influence of other background variables was controlled for. This effect diminished modestly, though persisted, when ARCP outcomes related to postgraduate exam failures were excluded. Previous research shows consistently lower performance of BME trainees in UK postgraduate medical examinations [19 21 23] and ARCP [25] across a number of specialties including the anaesthetic Primary FRCA [28]. The reasons underlying such differential performance are likely to be complex, subtle and multifactorial and have been much debated [24 28 29].

Our findings from this national study are in line with those reported by a smaller, regional one that observed that performance during recruitment correlated subsequent work place based assessment ratings [12].

Strengths and limitations

Our study used a large national dataset related to the UK anaesthesia training selection process. Moreover, we could link selection data to information relating to subsequent performance and training. A further strength of this study is that many studies investigating the validity of recruitment processes suffer from 'attenuating effects'. That is that low performing candidates at selection do not generate an assessment score because they do not get appointed. This leads to an underestimation of the recruitment process as it fails to detect ability to filter out poorer performing candidates. However, in our study the shortlisting and interview scores were standardized according to the performance of all applicants, not just those accepted on to training schemes. Thus, this approach should correct for any attenuating effects during the statistical modelling.

The data were relatively complete with the exception of the shortlisting and interview scores. It is likely that the missing values were not 'missing completely at random'. That is, the missing values were likely related to the values of variables that could be observed, as an applicant could not receive an interview without a shortlist score. However, we addressed this via multiple imputation as a form of sensitivity analysis. Indeed, our imputed and non-imputed results were relatively similar suggesting that most missing data were randomly missing, with the exception of shortlisting score. This is unsurprising given that the shortlisting score determines the presence of an interview. Thus, caution must be exercised when interpreting results relating to the former predictor.

Another limitation of the study is that 'clinical experience' could only be measured by date of registration with the GMC. Therefore years of practice outside of the UK could not be accounted for. Furthermore, due to a limited numbers of overseas trainees, we were unable to divide graduates by place of qualification (such as EEA graduates). We instead used the GDP of the country of nationality or qualification and acknowledge that this has less practical use than dividing by whether or not trainees were likely to have taken the Professional and Linguistic Assessments Board exam. This study also focused solely on anaesthetic core trainees, and the findings may not generalise to the context of higher specialist training limited. Nevertheless, it is likely that those variables we observed to predict success at core-training are likely to also apply to performance in subsequent higher specialist training. This may be especially true of educational (as opposed to clinical) outcomes and a high degree of continuity in academic performance in medicine is well recognised [30].

In the present study we used ARCP as one of the main outcomes. This metric lacks granularity and is more likely to discriminate between poorer rather than better performance. Data pertaining to Primary FRCA exam results was not available at the time of the study. Consequently a future study using the Primary FRCA as an outcome variable is likely to add more detail to the picture we have sketched here.

Implications for policy and directions for future research

Our finding that both interview and shortlisting scores made similar sized, independent contributions to the predict ARCP suggest that recruitment centres are correct in using both scores when ranking applicants for training posts. There is an extensive literature on the potential for bias of face-to-face interviews, though structuring the process, as at specialty recruitment, may reduce this risk to some extent. [4 12] The ideal weighting for each station should be informed by studies which are able to link data with clinical performance and our results suggest that the portfolio station should make up a significant portion of the overall selection score.

In line with previous research it was clear that certain demographic groups were less likely to be both appointed at application to core specialty training and also more likely to receive less satisfactory outcomes at subsequent ARCP. Certainly, for groups at risk of less satisfactory ARCP outcomes it may be that additional, targeted support would be beneficial. Such support could also be practical in nature, such as increased access to flexible working for those with caring commitments. The most effective approaches to supporting such individuals could be the subject of future, possibly qualitative, research.

Further research could focus on follow-up of anaesthesia trainees into higher specialist training. Performance in higher specialty trainees (STs) could be observed using subsequent ARCP outcomes, as well as performance in the 'Final FRCA'. This exam, as with the Primary

FRCA, consists of two sections, and must be passed by the end of specialty training year 4 (ST4), in order to allow entry to more advanced specialist training (ST5 to ST7).

Conclusions

Our findings support the effectiveness of the UK national selection process for entry to core anaesthesia training, with both shortlisting and interview scores being strongly predictive of clinical performance measured through ARCP outcomes. Demographic variables effecting appointability were also associated with ARCP outcomes.

Acknowledgments We gratefully acknowledge the assistance of Mr Daniel Smith at the General Medical Council for arranging the data extract used for this study

Author contributions MEDA led on writing, data interpretation and contributed to data analysis. LWP led on data analysis, produced the data visualisations, and contributed to data interpretation and drafting and appraising the manuscript. TG contributed to study design, editing and appraising the content of the manuscript. PAT led on study design and supervision of the project and contributed to writing, editing and appraising the content of the manuscript.

Funding PAT is supported in his research by a National Institute for Health Research (NIHR) Career Development Fellowship CDF-2015-08-11. This paper presents independent research part-funded by the National Institute for Health Research (NIHR). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Competing interests PAT has previously received research funding from the ESRC, the EPSRC, the Department of Health for England, the UKCAT Board, and the GMC. TG has been a member of the RCOA Recruitment Committee since 2010 and Chair of the committee since 2015. He has previously been awarded funding from the Department of Health to develop and evaluate selection methods for anaesthesia training.

Patient consent Not required

Ethics approval The study relied on the analysis of de-identified routinely collected data analysed within a 'safe haven' environment. This was confirmed in writing by the chair of the University Of York Department of Health Sciences Ethics Committee.

Data sharing statement The data and associated STATA syntax used to manage and analyse the data may be made available from the GMC on request within a safehaven environment on an individual basis should a sufficient justification be provided.

Figure legends

Figure 1: Summary of the recruitment process into Anaesthetics Core Training.

Figure 2: Assessment pathway from medical school to the end of anaesthetics training. The focus of our study is the Anaesthetics Core Training stage (in blue).

Figure 3: Data flowchart for the study.

Figure 4: Results from univariable logistic regression analyses for an individual being deemed 'appointable' according to each predictor variable.

Figure 5: Univariable ARCP results for each variable showing how including and excluding exam failures affect ARCP outcomes.

Table 1: Descriptive statistics

Variable	Mean	Standard deviation
Age on application	29.32	3.46
Experience in years on application	2.78	1.72
Standardised shortlist score	0.18	0.83
Standardised interview score	0.17	0.81
Number of jobs applied for during study period	2.25	1.31
Number of jobs deemed appointable during study period	1.49	0.93
Age at ARCP	30.13	3.33
Experience in years at ARCP	3.58	1.44
Number of ARCPs per doctor	1.83	0.93

Table 2: Results from a multivariable logistic regression model predicting the odds of a candidate being deemed appointable to core Anaesthetics training.

Variable	Odds ratio (95% CI)	p
Age at application	0.94 (0.90 to 0.98)	.008
Experience at application	1.35 (1.25 to 1.47)	<.001
BME	0.67 (0.49 to 0.90)	.008
GDP of country of PMQ	1.48 (1.05 to 2.10)	.030

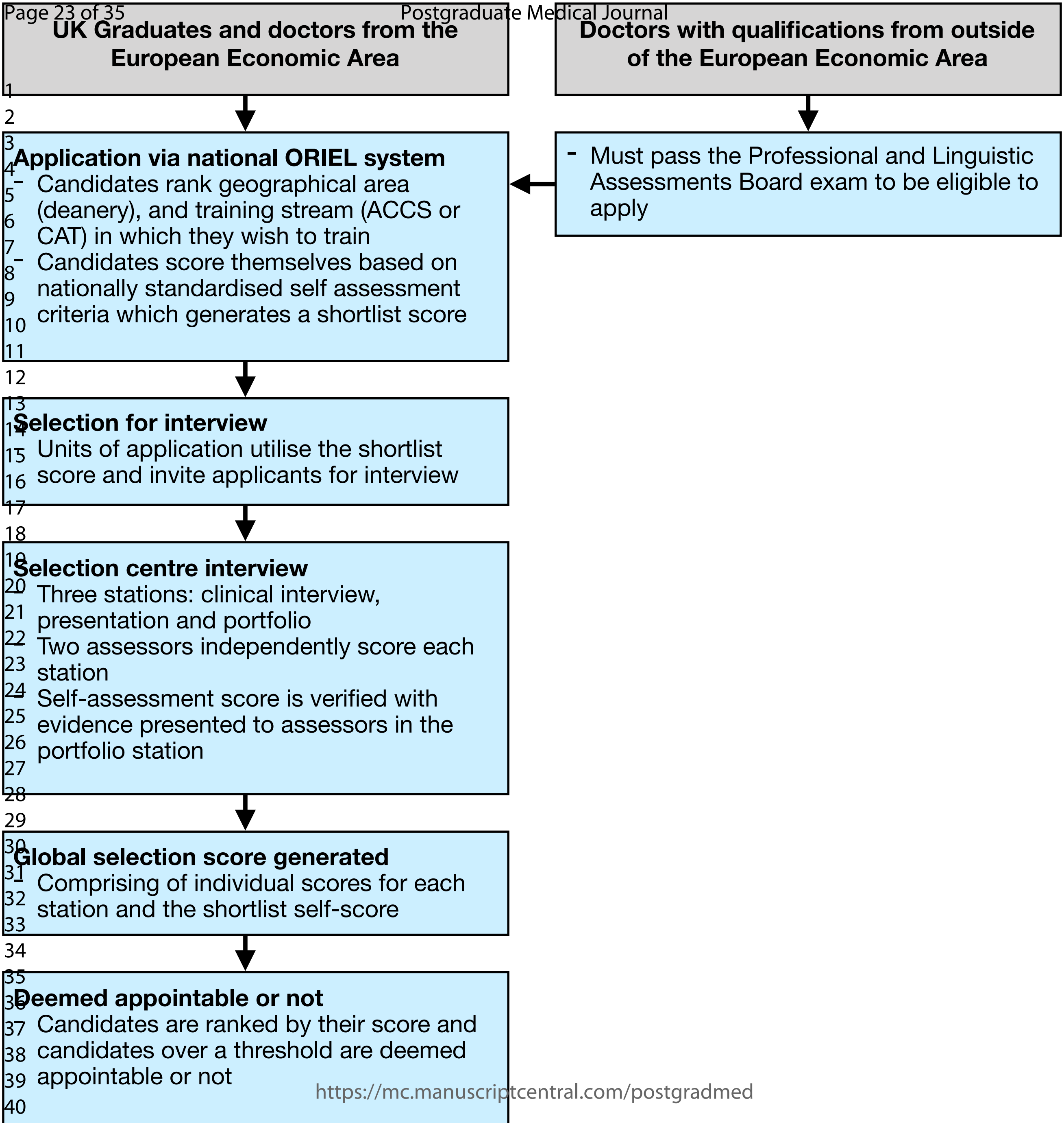
Table 3: Results from multivariable ordinal logistic regression analyses predicting the odds of a satisfactory vs unsatisfactory ARCP outcome amongst anaesthetic trainees. Results excluding ARCP outcomes associated with postgraduate exam failure are in the right column.

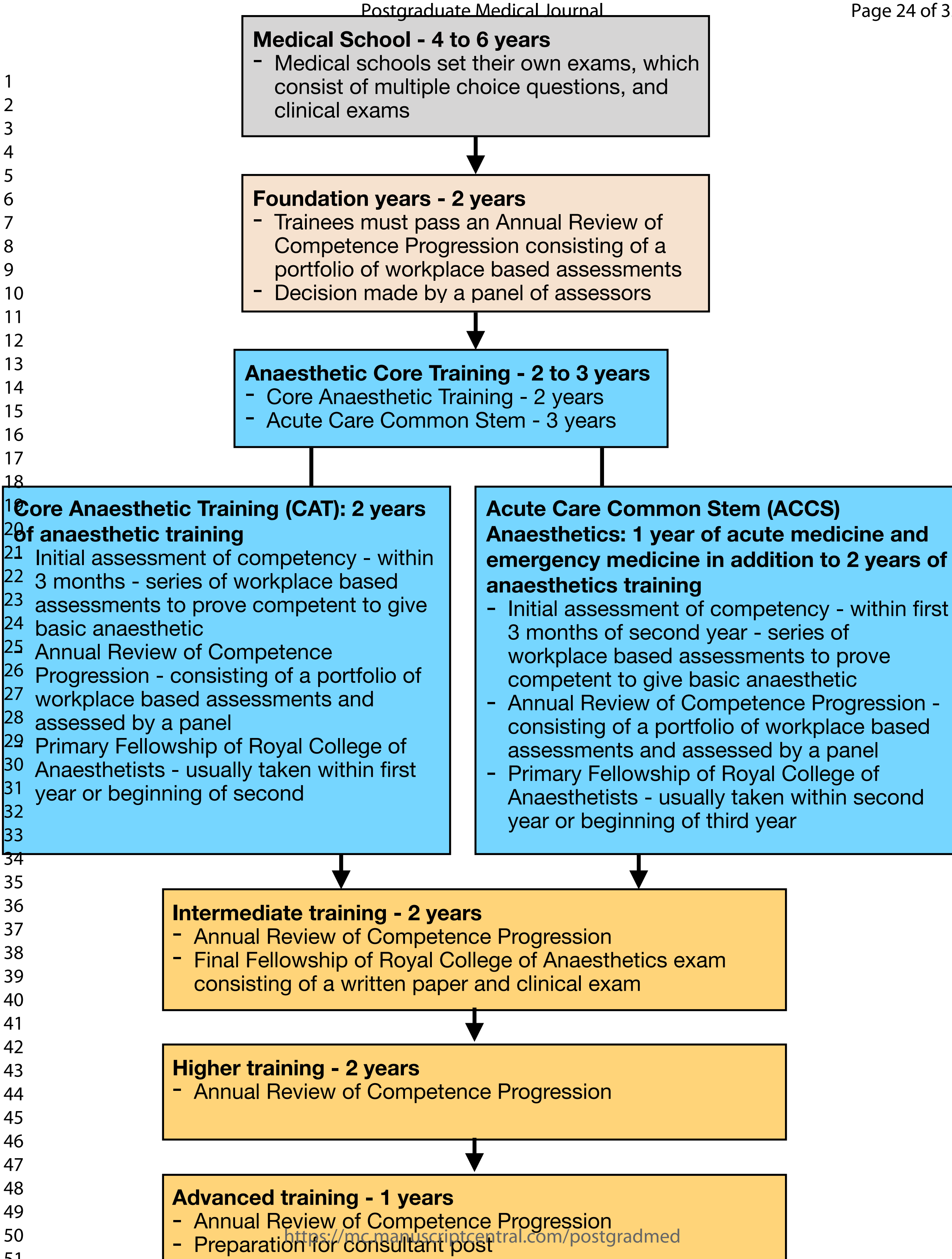
	ARCP results			
	Including exam failure		Excluding exam failure	
Variable	Odds ratio (95% CI)	p	Odds ratio (95% CI)	p
Age at ARCP	0.94 (0.91 to 0.97)	<.001	0.97 (0.94 to 1.01)	.166
Experience at ARCP	0.90 (0.83 to 0.97)	.006	0.97 (0.88 to 1.06)	.489
Standardised interview score	1.37 (1.19 to 1.57)	<.001	1.35 (1.15 to 1.57)	<.001
Standardised shortlist score	1.42 (1.24 to 1.62)	<.001	1.28 (1.10 to 1.48)	.001
ACCS vs CT	1.43 (1.13 to 1.82)	.003	1.01 (0.78 to 1.32)	.934
BME	0.69 (0.55 to 0.86)	.001	0.72 (0.55 to 0.93)	.011

References

1. Chen F, Arora H, Martinelli SM, et al. The predictive value of pre-recruitment achievement on resident performance in anesthesiology. *Journal of Clinical Anesthesia* 2017;39:139-44.
2. Guffey RC, Rusin K, Chidiac EJ, et al. The utility of pre-residency standardized tests for anesthesiology resident selection: the place of United States Medical Licensing Examination scores. *Anesthesia & Analgesia* 2011;112(1):201-06.
3. Pashayan N, Gray S, Duff C, et al. Evaluation of recruitment and selection for specialty training in public health: interim results of a prospective cohort study to measure the predictive validity of the selection process. *Journal of Public Health* 2015;38(2):e194-e200.
4. Reeve P, Vickers M, Horton J. Selecting anaesthetists: the use of psychological tests and structured interviews. *Journal of the Royal Society of Medicine* 1993;86(7):400.
5. Selber JC, Tong W, Koshy J, et al. Correlation between trainee candidate selection criteria and subsequent performance. *Journal of the American College of Surgeons* 2014;219(5):951-57.
6. Yost MJ, Gardner J, Bell RM, et al. Predicting academic performance in surgical training. *Journal of Surgical Education* 2015;72(3):491-99.
7. Ahmed H, Rhydderch M, Matthews P. Do general practice selection scores predict success at MRCGP? An exploratory study. *Education for Primary Care* 2012;23(2):95-100.
8. Davison I, Burke S, Bedward J, et al. Do selection scores for general practice registrars correlate with end of training assessments? *Education for Primary Care* 2006;17(5)
9. Patterson F, Ferguson E, Norfolk T, et al. A new selection system to recruit general practice registrars: preliminary findings from a validation study. *BMJ* 2005;330(7493):711-14.
10. Patterson F, Kerrin M, Baron H, et al. Exploring the relationship between general practice selection scores and MRCGP examination performance. *Final Report to General Medical Council* 2015
11. Patterson F, Tiffin PA, Lopes S, et al. Unpacking the dark variance of differential attainment on examinations in overseas graduates. *Medical Education* 2018;52(7):736-46. doi: doi:10.1111/medu.13605
12. Gale T, Roberts M, Sice P, et al. Predictive validity of a selection centre testing non-technical skills for recruitment to training in anaesthesia. *British Journal of Anaesthesia* 2010;105(5):603-09.
13. Conference of Postgraduate Medical Deans of the United Kingdom. A Reference Guide for Postgraduate Specialty Training in the UK: The Gold Guide, 2018.
14. Royal College of Anaesthetists. The FRCA Examination London: Royal College of Anaesthetists; 2019 [Available from: https://www.rcoa.ac.uk/examinations/overview accessed 15th May 2019 2019.
15. Pyne Y, Ben-Shlomo Y. Older doctors and progression through specialty training in the UK: a cohort analysis of General Medical Council data. *BMJ open* 2015;5(2):e005658.
16. Bowhay AR, Watmough SD. An evaluation of the performance in the UK Royal College of Anaesthetists primary examination by UK medical school and gender. *BMC Medical Education* 2009;9(1):38.
17. Cuddy MM, Swanson DB, Clauser BE. A Multilevel analysis of the relationships between examinee gender and United States Medical Licensing Exam (USMLE) step 2 CK content area performance. *Academic Medicine* 2007;82(10):S89-S93.

18. Richens D, Graham TR, James J, et al. Racial and gender influences on pass rates for the UK and Ireland specialty board examinations. *Journal of Surgical Education* 2016;73(1):143-50.
19. Scrimgeour D, Cleland J, Lee A, et al. Which factors predict success in the mandatory UK postgraduate surgical exam: The Intercollegiate Membership of the Royal College of Surgeons (MRCS)? *the surgeon* 2018;16(4):220-26.
20. Tiffin PA, Orr J, Paton LW, et al. UK nationals who received their medical degrees abroad: selection into, and subsequent performance in postgraduate training: a national data linkage study. *BMJ open* 2018;8(7) doi: 10.1136/bmjopen-2018-023060
21. Esmail A, Roberts C. Academic performance of ethnic minority candidates and discrimination in the MRCGP examinations between 2010 and 2012: analysis of data. *BMJ* 2013;347(347):f5662. doi: 10.1136/bmj.f5662
22. Dewhurst NG, McManus C, Mollon J, et al. Performance in the MRCP (UK) Examination 2003–4: analysis of pass rates of UK graduates in relation to self-declared ethnicity and gender. *BMC Medicine* 2007;5(1):8.
23. Wakeford R. Country of qualification is linked to doctors' General Medical Council performance assessment rate, but is it linked to their clinical competence? *BMC Medicine* 2017;15(1):149.
24. Woolf K, Potts HWW, McManus IC. Ethnicity and academic performance in UK trained doctors and medical students: systematic review and meta-analysis. *BMJ* 2011;342 doi: 10.1136/bmj.d901
25. Tiffin PA, Illing J, Kasim AS, et al. Annual Review of Competence Progression (ARCP) performance of doctors who passed Professional and Linguistic Assessments Board (PLAB) tests compared with UK medical graduates: national data linkage study. *BMJ* 2014;348 doi: 10.1136/bmj.g2622
26. The World Bank. National Gross Domestic Products (GDPs) [20th December 2018]. Available from: http://data.worldbank.org/indicator/ny.gdp.mktp.cd.
27. Scrimgeour D, Brennan P, Griffiths G, et al. Does the Intercollegiate Membership of the Royal College of Surgeons (MRCS) examination predict 'on-the-job'performance during UK higher specialty surgical training? *The Annals of The Royal College of Surgeons of England* 2018;100(8):669-75.
28. Watmough S, Bowhay A. An evaluation of the impact of country of primary medical qualification on performance in the UK Royal College of Anaesthetists' examinations. *Medical Teacher* 2011;33(11):938-40.
29. Woolf K, Rich A, Viney R, et al. Perceived causes of differential attainment in UK postgraduate medical training: a national qualitative study. *BMJ open* 2016;6(11):e013429. doi: 10.1136/bmjopen-2016-013429
30. McManus I, Woolf K, Dacre J, et al. The Academic Backbone: Longitudinal Continuities in Educational Achievement from Secondary School and Medical School to MRCP(UK) and the Specialist Register in UK Medical Students and Doctors. *BMC Med* 2013;11:242.


UK Graduates and doctors from the European Economic Area
Doctors with qualifications from outside of the European Economic Area
- Must pass the Professional and Linguistic Assessments Board exam to be eligible to apply
Application via national ORIEL system
- Candidates rank geographical area (deanery), and training stream (ACCS or CAT) in which they wish to train
- Candidates score themselves based on nationally standardised self assessment criteria which generates a shortlist score
Selection for interview
- Units of application utilise the shortlist score and invite applicants for interview
Selection centre interview
- Three stations: clinical interview, presentation and portfolio
- Two assessors independently score each station
- Self-assessment score is verified with evidence presented to assessors in the portfolio station
Global selection score generated
- Comprising of individual scores for each station and the shortlist self-score
Deemed appointable or not
Candidates are ranked by their score and candidates over a threshold are deemed appointable or not


Medical School - 4 to 6 years
- Medical schools set their own exams, which consist of multiple choice questions, and clinical exams
Foundation years - 2 years
- Trainees must pass an Annual Review of Competence Progression consisting of a portfolio of workplace based assessments
- Decision made by a panel of assessors
Anaesthetic Core Training - 2 to 3 years
- Core Anaesthetic Training - 2 years
- Acute Care Common Stem - 3 years
Core Anaesthetic Training (CAT): 2 years of anaesthetic training
- Initial assessment of competency - within 3 months - series of workplace based assessments to prove competent to give basic anaesthetic
- Annual Review of Competence Progression - consisting of a portfolio of workplace based assessments and assessed by a panel
- Primary Fellowship of Royal College of Anaesthetists - usually taken within first year or beginning of second
Acute Care Common Stem (ACCS) Anaesthetics: 1 year of acute medicine and emergency medicine in addition to 2 years of anaesthetics training
- Initial assessment of competency - within first 3 months of second year - series of workplace based assessments to prove competent to give basic anaesthetic
- Annual Review of Competence Progression - consisting of a portfolio of workplace based assessments and assessed by a panel
- Primary Fellowship of Royal College of Anaesthetists - usually taken within second year or beginning of third year
Intermediate training - 2 years
- Annual Review of Competence Progression
- Final Fellowship of Royal College of Anaesthetics exam consisting of a written paper and clinical exam
Higher training - 2 years
- Annual Review of Competence Progression
Advanced training - 1 years
- Annual Review of Competence Progression
- Preparation for consultant post

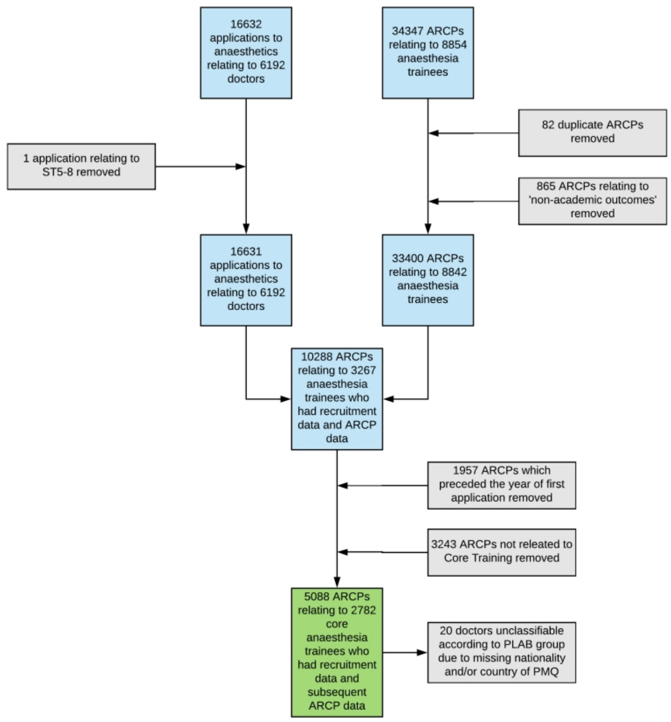


Figure 3: Data flowchart for the study.

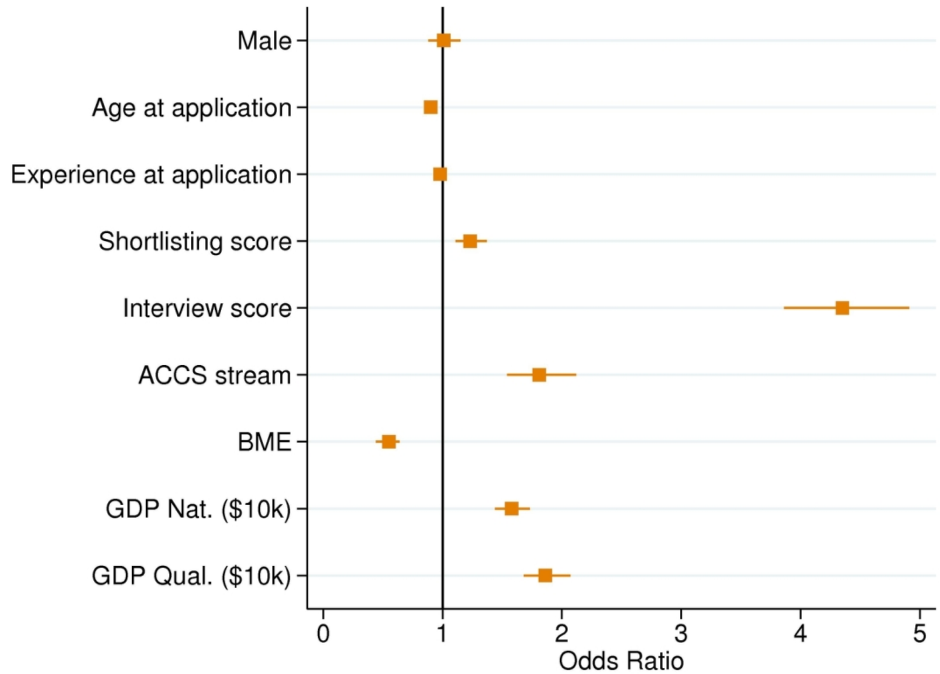


Figure 4: Results from univariable logistic regression analyses for an individual being deemed ‘appointable’ according to each predictor variable.

139x101mm (300 x 300 DPI)

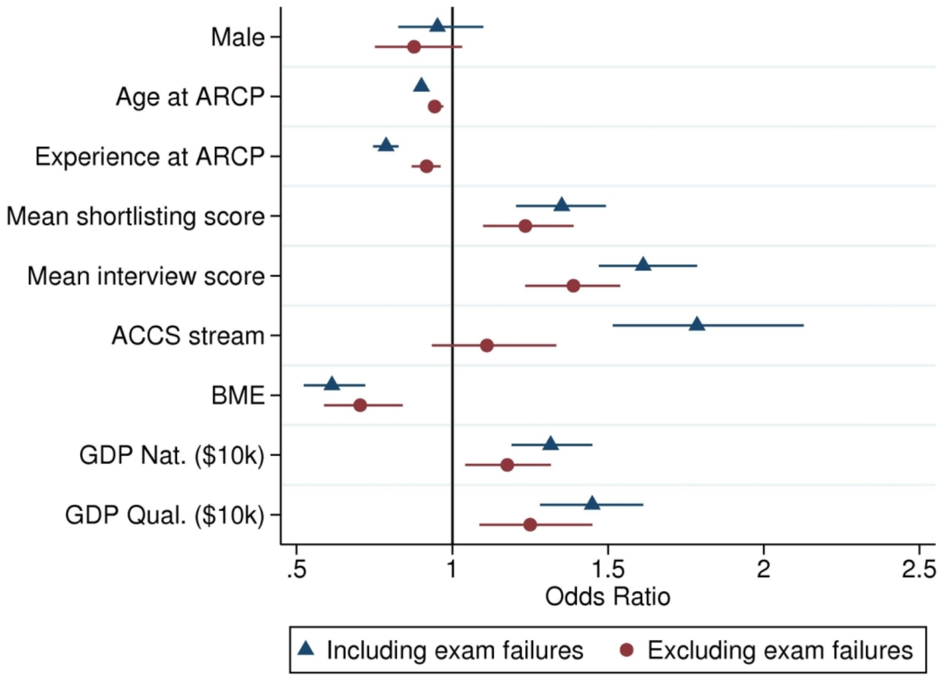


Figure 5: Univariable ARCP results for each variable showing how including and excluding exam failures affect ARCP outcomes.

139x101mm (300 x 300 DPI)

Technical Appendix

Evaluating the recruitment process into UK anaesthesia core training: doctors' performance at selection and subsequent postgraduate training– a national data linkage study: technical appendix

Aslet, M., Paton, L.W., Gale, T., & Tiffin, P.A.

A: Univariable results

Table S1 and Table S2 detail the univariable results for being 'deemed appointable' and ARCP outcome. These results are depicted in Figures 2 and 3 in the main paper.

	Deemed appointable	
Variable	Odds ratio (95% CI)	p
Male gender	1.01 (0.88 to 1.15)	.92
Age at application	0.96 (0.94 to 0.97)	<.001
Experience at application	0.98 (0.95 to 1.02)	.35
Standardised shortlist score	1.23 (1.11 to 1.37)	<.001
Standardised interview score	4.35 (3.86 to 4.91)	<.001
ACCS vs CAT	1.81 (1.54 to 2.12)	<.001
BME	0.55 (0.48 to 0.64)	<.001
GDP of country of nationality ($10k per person)	1.58 (1.44 to 1.73)	<.001
GDP of country of qualification ($10k per person)	1.86 (1.68 to 2.07)	<.001

Table S1: Results from univariable logistic regression predicting the odds of being deemed appointable to anaesthetics training at core level.

	ARCP results			
	Including exam failure		Excluding exam failure	
Variable	Odds ratio (95% CI)	p	Odds ratio (95% CI)	p
Male gender	0.95 (0.82 to 1.09)	.47	0.88 (0.75 to 1.03)	.11
Age at ARCP	0.91 (0.88 to 0.92)	<.001	0.94 (0.92 to 0.97)	<.001
Experience at ARCP	0.79 (0.75 to 0.83)	<.001	0.91 (0.87 to 0.96)	<.001
Standardised shortlist score	1.35 (1.21 to 1.50)	<.001	1.23 (1.10 to 1.39)	<.001
Standardised interview score	1.61 (1.46 to 1.78)	<.001	1.38 (1.23 to 1.54)	<.001
ACCS vs CAT	1.79 (1.52 to 2.12)	<.001	1.11 (0.93 to 1.33)	.23
BME	0.61 (0.52 to 0.72)	<.001	0.70 (0.59 to 0.83)	<.001
GDP of country of nationality ($10k per person)	1.32 (1.19 to 1.46)	<.001	1.18 (1.05 to 1.32)	.01
GDP of country of qualification ($10k per person)	1.44 (1.28 to 1.62)	<.001	1.25 (1.09 to 1.45)	.002

Table S2: Results from univariable analyses for a more satisfactory ARCP outcome. Analyses including exam failure are shown in the left column, and analyses excluding ARCP outcomes linked to postgraduate exam failure are shown in the right column.

B: Multiple imputation

Method

As detailed in the main paper, missing data were relatively uncommon other than for shortlisting score (39% missing) and interview score (22% missing). As such, in addition to the univariable analyses and backwards stepwise multivariable regression using listwise deletion detailed in the main body of the paper, we also performed stepwise multivariable regression on multiply imputed data. We used chained equations, creating 20 imputed data sets.

This portion of the analysis can be thought of as a form of 'sensitivity analysis' for shortlisting score and interview score. That is, if the results between the imputed and non-imputed datasets vary, then this would be evidence that the absent values are 'missing not at random' (MNAR) (i.e. the missing values are neither associated with the observed data nor due to chance). If there is evidence that the absent values are MNAR then results in relation to the affected variables must be interpreted more cautiously. All analyses were performed in Stata version 14.1.

Note that in this portion of the analyses, ARCP outcome was dichotomised, with satisfactory outcomes being coded as 1, and less than satisfactory coded as 0, rather than the four-point ordinal scale used in the main text. This is due to a technical point: Stata does not support ordinal logistic regression for imputed data sets. Thus, we used the binomial outcome as discussed above. This means that the non-imputed results presented in this supplementary appendix will differ slightly from the non-imputed results presented in the main text.

Furthermore, as can be seen in Table S3 shortlisting score was not an independent predictor of being 'deemed appointable' (i.e. it was not in the final multivariable model). However, for the purposes of a sensitivity analysis, we forced inclusion in the final model of both selection variables.

Results

Modelling of recruitment outcomes

	Deemed appointable			
	Original data		Imputed data	
Variable	Odds ratio (95% CI)	p	Odds ratio (95% CI)	p
Age at application	0.94 (0.90 to 0.98)	.01	0.96 (0.94 to 0.98)	.001
Experience at application	1.35 (1.25 to 1.47)	<.001	1.03 (0.99 to 1.07)	.11
BME ethnicity	0.67 (0.49 to 0.90)	.01	0.66 (0.57 to 0.77)	<.001
GDP of country of qualification ($10k per person)	1.43 (1.01 to 2.04)	.05	1.62 (1.44 to 1.82)	<.001
Standardised shortlisting score	1.16 (0.98 to 1.38)	.09	1.11 (1.01 to 1.21)	.02

Table S3: A comparison of the multivariable results for recruitment using non-imputed and imputed data. The left column shows results from the original data.

Table S3 shows the multivariable models for being deemed appointable on both non-imputed and imputed data sets. As can be seen, standardised shortlisting score became a statistically significant predictor (OR 1.11, 1.01 to 1.21) in the imputed dataset. That is, a candidate had 1.11 the odds of being appointed for each standard deviation above the mean a candidate scored. Additionally, in the imputed data, experience at application becomes non-significant.

Modelling ARCP outcomes

The results from the imputed data show slightly reduced effect sizes than the non-imputed results(Table S4). The most notable difference between the two sets of results is a reduction in the odds ratio for mean shortlisting score when using imputed data (Imputed result: OR 1.14, 1.03 to 1.26). A similar trend is seen when excluding ARCP outcomes related to postgraduate exam failure (Table S5). Indeed, mean shortlisting score becomes non-significant when using imputed data.

	ARCP results including exam failure			
	Original data		Imputed data	
Variable	Odds ratio (95% CI)	p	Odds ratio (95% CI)	p
Age at ARCP	0.95 (0.92 to 0.98)	.001	0.94 (0.92 to 0.97)	<.001
Experience	0.91 (0.85 to 0.98)	.016	0.88 (0.84 to 0.93)	<.001
Standardised interview score	1.35 (1.18 to 1.54)	<.001	1.35 (1.22 to 1.48)	<.001
Standardised shortlist score	1.33 (1.18 to 1.51)	<.001	1.14 (1.03 to 1.26)	.015
ACCS vs CT	1.32 (1.05 to 1.66)	.019	1.26 (1.07 to 1.48)	.006
BME	0.71 (0.57 to 0.87)	.001	0.71 (0.61 to 0.83)	<.001

Table S4: A comparison of the multivariable results for ARCP outcome using non-imputed and imputed data, including ARCP outcomes related to postgraduate exam failure. The left column shows results from the original data.

	ARCP results excluding exam failure			
	Original data		Imputed results	
Variable	Odds ratio (95% CI)	p	Odds ratio (95% CI)	p
Age at ARCP	0.98 (0.94 to 1.01)	.29	0.96 (0.94 to 0.99)	.002
Experience	0.97 (0.88 to 1.06)	.47	0.96 (0.91 to 1.02)	.24
Standardised interview score	1.33 (1.14 to 1.54)	<.001	1.25 (1.12 to 1.39)	<.001
Standardised shortlist score	1.23 (1.06 to 1.41)	.01	1.09 (0.96 to 1.22)	.18
ACCS v CT	0.97 (0.75 to 1.25)	.80	0.92 (0.77 to 1.09)	.34
BME	0.73 (0.57 to 0.94)	.01	0.73 (0.62 to 0.87)	<.001

Table S5: A comparison of the multivariable results for ARCP outcome using non-imputed and imputed data, excluding ARCP outcomes related to postgraduate exam failure. The left column shows results from the original data.

Discussion

The imputed analyses generally show similar but slightly weaker trends for all variables for both recruitment and ARCP outcomes. The exception is shortlisting score, which becomes significant when using imputed data to predict appointability. It also shows a large reduction in effect size for ARCP outcomes, and becomes non-significant when postgraduate exam failures are excluded from the analysis. Only candidates who were successfully appointed are included in the dataset. This implies they had a relatively high shortlisting score in order to be offered an interview. This would thus reduce the ability of shortlisting scores to predict less satisfactory ARCP outcomes.

A similar result can be observed when considering interview scores. Only those candidates deemed of a high enough standard to be interviewed would subsequently receive an interview score. Therefore, this limits the ability of this variable to predict trainees who may perform at the lower end of the scale.

The imputed analyses for shortlisting scores show a large reduction in effect size across analyses. This implies that the data missing for this variable are not randomly missing but systematically missing. It may be that those deaneries who return shortlist and interview scores are generally more efficient, which may also improve the quality of the training. This may positively skew shortlisting scores, with those receiving higher scores also more likely have their scores returned. This would explain why, when imputing the data with the assumption of random missingness, the effect size observed is often reduced. Due to these differences, analyses involving shortlisting score should be interpreted with caution.