UNIVERSITY OF PLYMOUTH
Faculty of Medicine and Dentistry

UKMED Project P41:
Development of a UKMED multidimensional measure of widening participation status.

Final Report

Paul Lambe, Martin Roberts, Tom Gale, David Bristow.

October 2018

Contents

1 Executive Summary

1.1 Background

- Use of contextual background information to widen participation among students from lower social class backgrounds is common in the selection and admissions processes of UK medical schools and higher education generally.
- However, there is concern about the validity of contextualised admissions decision making because contextual indicators produce conflicting information on disadvantage.
- The hypothesis, that the 'use of multiple, different types of contextual indicators mitigates the risk of false positive socioeconomic classification' was tested using the resources of the UK Medical Education Database (UKMED).
- This study aimed to evaluate available indicators and to combine the most reliable into a single multidimensional measure, an Index of Widening Participation Status (IWPS), which UK medical schools could use in their selection process.

1.2 Methods

- Three samples of non-graduate students aged 20 years and under were analysed, (1) UK domiciled entrants to Standard Entry Programmes at UK medical schools 2008-2015, (2) England domiciled entrants to Standard Entry Programmes at UK medical schools 2008-2014 and, (3) UK domiciled entrants to Medicine With a Gateway Year (MWGY) Programmes 2008-2015.
- Complete case analysis (cases with missing values dropped) using Spearman's correlations (rho), multivariable linear and logistic regression to identify and weight a set of contextual indicators to inform development of a multidimensional measure of widening participation status, termed Index of Widening Participation Status.
- Receiver Operator Characteristic (ROC) analysis to assess the accuracy of the Index of Widening Participation Status as a screening device and to inform cut-score threshold decision making.

- Latent Class Analysis of each sample (complete case analysis) to identify and describe 'typologies' of widening participation status based on students' pattern of response to the contextual indicators and social class position.
- Multiple imputation of missing values on contextual indicators (UK sample only) to enable comparison between the parameter estimates produced by the complete case analysis (model with missing data values) and those produced by the imputed data set, to obtain a more accurate picture of the relationship between the outcome (NS-SEC 3-7 versus NS-SEC 1-2) and the contextual indicators of disadvantage.
- The association between missing NS-SEC data values and a range of sociodemographic variables was examined to determine if particular sub-groups were more likely than others to not self-declare their social class when applying to study medicine.

1.3 Results

- Area-level contextual indicators returned conflicting information on individual's social circumstances and correlated weakly with socioeconomic class. School and individual-level indicators also correlated weakly with socioeconomic class.
- An IWPS derived from weighted scores on multiple types of contextual indicator identified students from lower socioeconomic class backgrounds with a high level of accuracy.
- Findings supported the hypothesis that the 'use of multiple, different types of contextual indicators mitigates the risk of false positive socioeconomic classification'.
- Latent Class Analysis identified three distinct typologies of student.
- Across typologies, the mean IWPS score of students in the 'WP students' latent class was greater than the mean IWPS scores of students in the other latent classes.
- Results of the Latent Class Analysis supported the hypothesis that the 'use of multiple, different types of contextual indicators mitigates the risk of false positive socioeconomic classification'.
- Results of the Latent Class Analysis supported the view that area level contextual indicators can produce conflicting information on disadvantage.

- An IWPS derived from imputed missing values identified students from lower socioeconomic class backgrounds with a level of accuracy comparable to that of the complete case analysis.
- The multiple imputation results indicated that missing data did not heavily bias the parameter estimates of the complete case analysis and thereby supported the inference of the complete case analysis that the 'use of multiple, different types of contextual indicators mitigates the risk of false positive socioeconomic classification'.
- The probability of not self-declaring social class on entry to medical school for BME students from areas of most deprivation whose parents had no higher education qualifications was ten-fold the probability for white students from areas of least deprivation whose parent(s) had higher education qualifications.

1.4 Conclusions

- The findings of this study support the hypothesis, that the 'use of multiple, different types of contextual indicators mitigates the risk of false positive socioeconomic classification'.
- An Index of Widening Participation Status derived from weighted scores on multiple types of contextual indicator of disadvantage identified students from lower socioeconomic class backgrounds with a high level of accuracy.
- Area-level contextual indicators returned conflicting information on individual's social circumstances and correlated weakly with socioeconomic class. School and individual-level indicators correlated weakly with socioeconomic class reflecting the far from straightforward link between contextual indicators, social circumstances and social class.

2 Introduction

Internationally, governments are driving an agenda of widening access to higher education to students from traditionally under-represented groups. [1] In 2012 the UK government's Independent Reviewer on Social Mobility and Child Poverty reported that, "*medicine has a long way to go when it comes to making access fairer, diversifying its workforce and raising*

social mobility". [2] Indeed, research indicates that around 84% of applicants to UK medical schools come from professional parental backgrounds. [3]

However, applicants to UK medical schools are not obliged to disclose their socioeconomic class and, for the approximately 70% that do, data is only available to admissions tutors some weeks after those selected for entry have enrolled. [4] Additionally, the UK's occupation-based socioeconomic class schema is not strictly hierarchical, and lacks occupational within-class homogeneity. [5, 6] For instance, large-scale employers find themselves in the same class as rank-and-file service workers, and Supreme Court judges the same class as fast-food shift supervisors. [6] Furthermore, if applicants were aware that self-reporting particular social class backgrounds could lead to preferred offers there would inevitably be some 'gaming' of the system. [7]

Against this backdrop, the Medical Schools Council (MSC) commissioned a programme of work on widening participation to the study of medicine in the UK. The subsequent MSC report concluded that 'contextual admissions' (admissions processes that adjust entry criteria to take applicants' socioeconomic and educational backgrounds into account) was a powerful tool that medical schools could use to widen participation among students from lower socio-economic classes.[8] Contextual admissions data comprise disparate measures of disadvantage, which, it is assumed, 'can be used to build a picture that more accurately determines socioeconomic background'. [9]

Use of contextual background information is now common in the selection and admissions processes of UK medical schools and higher education (HE) generally, and a range of indicators are used in a wide variety of ways. [10 - 14] Its use enables institutions to identify individuals with academic potential among applicants from lower socioeconomic class backgrounds and inform decisions on 'whom to shortlist, interview, to make standard or reduced offers to, or accept at confirmation or clearing'. [12]

Three types of contextual indicator are typically used: individual-level, area-level and school-level. [9, 10, 12] Individual-level indicators refer specifically to individuals' characteristics or circumstances or to those of their household (e.g. care status, family history of participation in HE, household income). Area-level indicators are proxies for individuals' circumstances derived from administrative and survey data (e.g. neighbourhood measures of average socioeconomic disadvantage and of participation in HE). School-level indicators are proxies

for individuals' circumstances based on the type of school attended and average educational outcomes at the school (e.g. examination performance and HE progression relative to national average).

There is, however, concern about the validity of contextual admissions decision making because of the conflicting information on disadvantage these indicators return, the veracity of self-reported information and potential for 'gaming', and the extent of missing data. [3, 14-17] In order to address these concerns, the MSC recommend that medical schools use multiple contextual indicators, of different types to inform admissions decisions.[9]

It is therefore important to evaluate the strengths and limitations of contextual indicators, singly and in combination, and their association with socioeconomic status. This study aimed to evaluate the association between available indicators and socioeconomic class and to combine those most strongly associated into a single multidimensional measure, which UK medical schools could use in their selection processes. The UK Medical Education Database (UKMED) includes a range of contextual indicators commonly used in selection to the study of medicine and provides a unique opportunity to achieve this aim, the outcome of which can potentially make contextual admissions to medicine fair, transparent and above all, evidence-based. [18]

This report presents findings from the UKMED P41 project in which we tested the hypothesis, that the 'use of multiple, different types of contextual indicators mitigates the risk of false positive socioeconomic classification'.

3 Methods

3.1 Data, study population and variables

The anonymised data for this study was accessed remotely by the authors via the Health Informatics Centre, Safe Haven at Dundee University, (https://www.dundee.ac.uk/hic/hicsafehaven/). The UKMED Data Dictionary (http://www.ukmed.ac.uk/documents/UKMED_data_dictionary.pdf) provides a full list of data types, descriptions and sources.

In its 2018 iteration, the UKMED holds socio-demographic and educational data on all students entering UK medical schools from 2002 until 2015. We restricted the range of data

for our analyses to the years 2008 to 2015, when additional contextual information on entrants became available from the UK Clinical Aptitude Test (UKCAT) database. [19]

The use of proxy measures, such as area-level postcode-based contextual indicators, to determine the social class of graduate and mature students may be misleading as their current postcode may well be unrelated to their earlier educational and social background. [20] Moreover, contextual indicators are UK-based measures and thus not applicable to international and European Union (EU) students. Hence, we excluded graduate, mature, international and EU students from the sample.

Three study populations were analysed: -

(1) UK domiciled, aged 20 years and under, non-graduate entrants to Standard Entry Programmes (n=40190) at UK medical schools 2008-2015.

(2) England domiciled, aged 20 years and under non-graduate entrants to Standard Entry Programmes (n=32825) at UK medical schools, 2008-2014.

(3) UK domiciled, aged 20 years and under, non-graduate entrants to Medicine With a Gateway Year Programmes (MWGY) (n= 810) at UK medical schools, 2008 to 2015, (see appendix tables for frequencies and missing values on socio-demographic and educational background data).

We decided on these three study samples because UKMED data on the A-level performance of the school (or college) attended are only available for English schools and colleges. Hence the UK and England samples. To gain entry to MWGY courses students must fulfil a variety of widening participation criteria and thus this sample provides a touchstone for testing the accuracy of the putative multidimensional measure of widening participation. [10]

The UKMED socio-economic class variable SEC (Office for National Statistics SEC (NS-SEC), 8-class version) included 8 cases reporting 'Long-term unemployed/never worked' which were subsumed into the SEC category 'Routine occupations', thus reducing SEC from its original 8 class to a 7 class version. We created the dichotomous variable LOWERSOC (coded 1 = SEC 3-7, 0 = SEC 1-2) because of concerns about the within-class homogeneity of the UK's occupational–based social class measure NS-SEC, and the schema's assumption that within each class category individuals share similar employment relationships, income security, life chances, life choices, and positions within social hierarchies. [5, 6, 21 - 23]

The contextual admissions indicators held in the UKMED dataset are described in the appendix along with sample frequencies.

We tested the null-hypothesis that 'use of multiple, different types of contextual indicators does not mitigate the risk of false positive socioeconomic classification' on the three study populations.

There were eleven contextual admissions indicators held in the UKMED dataset which could be applied to a UK level analysis and two measures of relative A-level performance which applied only to English schools and colleges. (See Table 2 for description, frequencies and missing values).

3.2 Main statistical analysis

We calculated Spearman's correlations (rho) between contextual indicators and socio-economic class (coded NS-SEC 1 through NS-SEC 7). Using multivariable regression we developed an Index of Widening Participation Status (IWPS) and conducted post-hoc linear regression diagnostics. [24-26]

The accuracy of the IWPS as a screening device (Sensitivity and Specificity of predicted outcomes) was assessed using Receiver Operator Characteristic (ROC) curve analysis. An area under the curve (AUC) statistic 0.9-1 was considered excellent discrimination, 0.8-0.9 good, 0.7 – 0.8 fair, 0.6 – 0.7 poor and 0.5 – 0.6 no better than chance.[27] Thus, the greater the AUC the better the global performance of the index. Generation of a ROC curve for scores on the IWPS enabled determination of an optimum threshold score that maximises true positive and minimises false negative classification.

The likelihood ratio (LR+) for a true positive classification was calculated as follows: Sensitivity/ (1-Specificity). LR+ is the ratio of the chance of a positive classification (NS-SEC 3-7) if the subject is in NS-SEC 3-7 to the chance of a positive classification if the subject is not in NS-SEC 3-7. The likelihood ratio for a true negative classification, LR-, was calculated as follows: (1-Sensitivity)/Specificity. In screening tests for presence or absence of a disease, a high LR+ (e.g.>10) provides evidence to support a diagnosis and a low LR- (e.g. <0.01) provides evidence against a positive diagnosis. [28]

3.3 Latent class analysis

Latent Class Analysis (LCA) is a statistical method that classifies individuals into groups based on conditional probabilities; within each group individuals will have a similar pattern of response to categorical variables. [29, 30] In this case, we sought to identify groups of students with a similar pattern of response to the contextual indicators and social class position (LOWERSOC).

Posterior membership probabilities (maximum likelihood estimates based on patterns of scores on the contextual indicators) assign cases to homogeneous latent classes. Two parameters are produced: latent class probabilities and conditional probabilities. Latent class probabilities indicate the relative size of each class. Within each class there is a set of conditional probabilities relating to the indicators. The conditional probabilities represent the probabilities of being at a particular response level for a particular indicator and thereby enable characterisation of the nature of the types defined by each of the classes (typologies). The underlying assumption of LCA is local independence, that is, within each class all measures are independent as all correlations between the variables are explained through class structure.

Model selection in LCA can involve both absolute fit of a particular model and relative fit of two or more competing models. A common measure of absolute fit in categorical models is the G^2 (aka L^2) likelihood-ratio chi-squared statistic, which tests the hypothesis that the specified LCA model fits the data. However, for the more complex mixture models analysed by this study model selection typically occurs by comparing models with different numbers of latent classes using the Bayesian Information Criterion (BIC) and the Aikake Information Criterion (AIC). BIC is a measure of the goodness-of-fit of a model that considers the number of parameters and the number of observations. AIC is a measure of the goodness-of-fit of a model that considers the number of parameters. When selecting which LCA model fits the data best the model with the lowest BIC and AIC statistics is preferable. Thus, the BIC and AIC information criteria can be used to compare relative fit of models with different numbers of classes (e.g., three versus four classes), with a lower value indicating a more optimal balance between model fit and parsimony.

Whilst BIC and AIC are based on good statistical theory, neither is a ‘gold standard’ for assessing which model should be chosen. In selecting each final model, we also took into

consideration how well the solution could be interpreted, that is, whether the subgroups showed logical patterns, were distinct from other subgroups and could be readily labelled. We did this because LCA is an exploratory approach and should be considered a heuristic for describing population heterogeneity. [31] For each LCA model multiple sets of random starting values were specified to confirm the solution.

LCA study populations, variables and coding:-

LCA Model 1: UK sample (n= 30,595), non-missing on all indicators.

LCA Model 2: England sample (n= 20,690), non-missing on all indicators.

LCA Model 3: Medicine With a Gateway Year sample (n= 630), non-missing on all indicators.

Indicators:-

POLAR quintiles 1 - 5 (ordinal)

IMD quintiles 1 - 5 (ordinal)

SCHOOOL TYPE (binary, 1 = state funded, 0 = privately funded)

PARED (binary, 1= parent has no HE qualifications, 0 = parent has HE qualifications)

BURSARY (binary, 1 = in receipt of bursary/Educational Maintenance Allowance, 0 = not in receipt)

LOWERSOC (binary, 1= NS-SEC 3-7, 0 = NS-SEC 1-2)

The following school performance indicators were included as variables in LCA Model 2, the England sample:-

APFSTE ALEVA = Average point score per A-level entry at school attended by the student in year A –level taken converted to quintiles and coded 1 through 5 lowest to highest score (ordinal).

TALLPPE = Average point score per A-level student at school attended by the student in year A –level taken converted to quintiles and coded 1 through 5 lowest to highest score(ordinal).

3.4 Multiple imputation

Given the missing values on the contextual indicators (Bursary apart) and on our outcome of interest the variable lowersoc (NS-SEC 3-7 versus NS-SEC 1-2), and the potential for biased estimates we created a synthetic data set by replacing missing values by a method of multiple imputation using chained equations (MICE). [32] The aim of this secondary analysis was to examine the impact of missing data on the conclusions reached by our primary complete case analysis. We acknowledge that opinions are divided in the research community between complete case analysis (excluding cases with missing values, list-wise deletion) and imputation. The former considered to lead to biased estimates and the latter,

based on assumptions about data which are often violated, considered to lead to biased estimates of unpredictable direction. [33, 34]

Study population, variables and missing data

We explored the use of multiple imputation in a single subset of the data: UK domiciled, non-graduate entrants to Standard Entry Programmes at UK medical schools 2008-2015, aged under 21 years (n=40190).

We examined the number, proportions and patterns of missing values on the contextual indicators; POLAR quintile, IMD quintile, SCHOOL TYPE (state funded school or college = 1, privately funded school or college = 0), PARED (parent has no HE qualifications = 1, parent has HE qualifications = 0), and BURSARY (in receipt of a bursary or EMA = 1, not in receipt of bursary or EMA = 0), and on the variable LOWERSOC (NS-SEC 3-7 =1, NS-SEC 1-2 =0).

Multiple Imputation by Chained Equations (MICE) is a statistical technique for handling missing data. MICE uses the distribution of observed data to estimate a set of plausible values for missing data. MICE is an iterative process that makes repeated draws from a model of the distribution of variables that have missing observations, to generate multiple complete data sets.

However, an understanding of the underlying processes believed to have generated missing values in the data set is important because different types of missing data require different treatments. Missing data mechanisms comprise three main categories: missing completely at random (MCAR), missing not at random (MNAR) and missing at random (MAR). Each mechanism describes one possible relationship between the propensity of data to be missing and values of the data, both missing and observed.

Missing values on a variable are said to be MCAR if the missingness is independent of both unobserved and observed data, that is, neither the variable nor the unobserved value of the variable predict whether a value will be missing. [35]There is no relationship between the missingness of the data and any values, observed or missing and the missing data points are a random subset of the data. If the observed values are essentially a random sample of the full data set, complete case analysis (listwise deletion) would return the same results as the full data set would have. Thus MCAR is ignorable, however, MCAR is very rare.

A variable is MNAR if unobserved values of the variable predict 'missingness'. MNAR is non-ignorable because the missing data mechanism is related to the missing values. For example, high earners are less likely to reveal income on a survey than respondents with lower incomes. Complete case analysis would then give biased results. MNAR means there is a relationship between the propensity of a value to be missing and its values.

A variable is deemed MAR if the cause of the missing data is unrelated to the missing values, but is related to the observed values of other variables in the data set. MAR means that there is a systematic relationship between the propensity of missing values and the

observed data. For instance, if males respondents are more likely to self-report their weight than female respondents, missing values on weight are MAR.

Multiple imputation assumes that the data are MAR. In order to establish if the missing values on the contextual indicators that comprise the IWPS and the variable lowersoc (NS-SEC 3-7 versus NS-SEC 1-2) are MAR we ran chi-square tests between each of them and other variables in the data set. Importantly, variables which predict missingness on variables to be imputed can be added to the imputation model to increase power and accuracy.

In an iterative process, MICE replicates the incomplete data set multiple times and replaces the missing data in each replication with plausible values drawn from the imputation model. The statistical analysis of interest, in this case logistic regression of the binary outcome NS-SEC 3-7 versus NS-SEC 1-2, was then performed on each completed data set separately. Finally, a single MICE estimate was calculated by working through the estimates (coefficients and standard errors) obtained from each completed data set. Thus, MICE takes into account the uncertainty associated with the imputed values. The estimated variance of the final set of imputed values allows for within-imputation (estimates within each completed data set) and between imputation (between the estimates across completed data sets) variability.

In line with guidance we included all variables with missing values to be imputed in the MICE model, including the outcome variable LOWERSOC (NS-SEC 3-7 versus NS-SEC 1-2) along with the auxiliary variables BURSARY and BME. We specified separate conditional univariate imputations for each variable to be imputed: POLAR quintile and IMD quintile (ordered logistic regression), LOWERSOC, SCHOOL TYPE, PARED (binary logistic). Moreover, we specified the number of iterations at 15, thus exceeding the percentage of missing data in the analysis. [36, 37]

Lastly, we examined the association between missing NS-SEC data values and a range of sociodemographic variables to determine if particular sub-groups were more likely than others to not self-declare their social class when applying to study medicine. A variable _m was created (coded 1 = missing on socioeconomic class and 0 = non-missing), and _m used as the dependent variable in univariate logistic regression models to examine the association between missing on NS-SEC and a range of sociodemographic variables.

We used Stata version 15 for all analyses.

4 Results

The contextual indicators PARENT DEGREE, INCOME SUPPORT and FREE SCHOOL MEALS were excluded from both models due to high proportions of missing data (84%, 85% and

84% respectively). The high correlation between QAHE and POLAR (r_s = 0.75, p<0.001) indicated that both are operationalisations of the same underlying concept. [38] Both IDACI RANK and IDAOPI were highly correlated with IMD (r_s = 0.80, p<0.001, and r_s = 0.79, p<0.001, respectively). Additionally, QAHE, IDACI and IDAOPI are not easily available to admissions decision makers and so were excluded from further analyses.

The over-representation of medical students with NS-SEC 1-2 backgrounds (73.71%) and under-representation of medical students with NS-SEC 3-5 backgrounds (15.82%) or NS-SEC 6-7 backgrounds (7.02%) was confirmed by comparison with the respective proportions in the UK population: 31%, 29% and 25%. [39, 40]

4.1 Results: UK sample

The remaining five contextual indicators correlated significantly (p<0.001) with socioeconomic class: POLAR quintile (r_s = -0.1753), IMD quintile (r_s = 0.2156), SCHOOL TYPE (r_s = 0.1648), PARED (r_s = 0.4329), and BURSARY (r_s = 0.2182). POLAR correlated significantly with IMD (r_s = -0.4300, p<0.001) (Table 4). Cross-tabulations revealed that:

- 61% of subjects from areas of lowest young persons' participation in higher education (POLAR quintile 1) had NS-SEC 1-2 parental backgrounds (Figure 1).
- 49% of subjects from areas of highest deprivation (IMD quintile 5) hadNS- SEC 1-2 parental backgrounds (Figure 2).
- 73% of subjects who attended state school/college had NS-SEC 1-2 parental backgrounds, 5% of those who attended a privately funded school/college had parents in the NS-SEC 6-7 and 16% had parents in the NS-SEC 3, 4 and 5 (Figure 3).
- 41% of subjects who had parents without higher education qualifications had NS-SEC 1-2 parental backgrounds (Figure 4).
- 37% of subjects in receipt of a bursary had NS-SEC 1-2 parental backgrounds (Figure 5).
- 45% of subjects with IMD quintile 5 postcodes, areas of highest deprivation, did not live in POLAR quintile 1 areas of lowest young persons' participation in higher education (Figure 6)

78% (31500/40190) of subjects had values on all five contextual indicators. Subjects were scored: POLAR quintile 1 = 1 point, IMD quintile 5 = 1 point, attended a state funded school

= 1 point, parents had no higher education qualifications = 1 point, in receipt of a bursary = 1 point and zero otherwise. Thus, subjects could score between 0 and 5 points.

Scores were then included as the dependent variable in a linear regression model with POLAR quintile (coded 1-5), IMD quintile (coded 1-5), SCHOOL TYPE (coded 1-0), PARED (coded 1-0) and BURSARY (coded 1-0) as independent predictors.

The regression model's beta coefficients indicated that a unit increase in POLAR quintile was associated with a 0.10 decrease in score, that a unit increase in IMD quintile was associated with a 0.10 increase in score (Table 5).

An IWPS score was created by recoding and summing the five predictor variables.

- POLAR quintile 1 = 1, quintile 2 = 0.9, quintile 3 = 0.8, quintile 4 = 0.7 and quintile 5= 0.6
- IMD quintile 1 = 0.6, quintile 2 = 0.7, quintile 3 = 0.8, quintile 4 = 0.9 and quintile 5 = 1
- SCHOOL TYPE, PARED and BURSARY were weighted 1= state funded school, 1= parent no HE qualifications, and 1= in receipt of a bursary, and zero otherwise.

The IWPS score (mean= 2.38, Std. Deviation =0.82, minimum = 1.2, maximum = 5.0), was used as the sole predictor in a binary logistic regression model with LOWERSOC (NS-SEC 3-7 versus NS-SEC 1-2) as the binary outcome.

A Hosmer-Lemeshow test ($p>0.05$) confirmed adequate model fit and a Wald test ($Chi2(1) = 3907$, $p<0.001$, $n=30,595$) that IWPS score had a significant effect on the outcome LOWERSOC.

The predicted probability of the outcome, when plotted against scores, graphically illustrated that as score increased so the probability of a subject being in NS-SEC 3-7 increased (Figure 7). An AUC of 0.74 (95% confidence interval 0.73 to 0.75) (Figure 8) indicated that the IWPS discriminative ability to correctly classify subjects into NS-SEC 3-7 and NS-SEC 1-2 groups was fair to good.

Cross-tabulation of IWPS scores and the outcome LOWERSOC revealed that 85% (530/620) of those with scores between 4.5 and 5.0 points were from NS-SEC 3-7 backgrounds (Table 6). Moreover, an IWPS score from 4.5 to 5.0 (Table 7) was associated with an LR+ meeting

the criterion of a useful diagnostic test (L=>10). However, the associated LR- statistics indicated a proportion of misclassification.

4.2 Results: England sample

The above analyses were repeated for the sample of non-graduate entrants to Standard Entry Programmes at UK medical schools in academic entry years 2008 to 2014 (school performance indicators not yet available for 2015), aged 20 years and under, whose region of domicile was England at time of entry, and who had attended an English school between 11 and 16 years of age (Table 1).

The seven contextual indicators; POLAR, IMD, SCHOOL TYPE, PARED, BURSARY, TALLPPE_ALEVA and APSFTE_ALEVA, correlated significantly with SEC (Table 8). Cross-tabulation with socio-economic class revealed that:

- 54% of subjects with POLAR quintile 1 postcodes, areas of the lowest rate of persons' participation in higher education qualifications, had NS-SEC 1-2 parental backgrounds.
- 49% of subjects with IMD quintile 5 postcodes, areas of highest deprivation, had NS-SEC 1-2 parental backgrounds.
- 5% of those who attended a privately funded school/college had parents in NS-SEC 6-7 (Semi-routine and routine occupations) and 11% had parents in the NS-SEC 3-5, whilst 73% of subjects who attended state school/college had NS-SEC 1-2 parental backgrounds.
- 41% of subjects whose parents had no higher education qualifications had NS-SEC 1-2 parental backgrounds.
- 37% of subjects in receipt of a bursary had NS-SEC 1-2 parental backgrounds.
- 54% of subjects who had attended APSFTE_ALEVA quintile 1 (lowest average point score per A level student) schools had NS-SEC 1 -2 parental backgrounds (Figure 9).
- 54% of subjects who had attended TALLPPE_ALEVA quintile 1 (lowest average point score per A level entry) schools had NS-SEC 1-2 parental backgrounds (Figure 10).

64% (21,100/32,825) of subjects domiciled in England had non-missing values on the seven contextual indicators; POLAR, IMD, SCHOOL TYPE, PARED, BURSARY, APSFTE_ALEVA and TALLPPE_ALEVA. Weighted scores were calculated using the method described above

(section 4.2). The regression model's beta coefficients weighted for APSFTE and TALLPPE scores as follows.

- APSFTE quintile 1 = 1, quintile 2 = 0.9, quintile 3 = 0.8, quintile 4 = 0.7 and quintile 5= 0.6.
- TALLPPE quintile 1 = 1, quintile 2 = 0.9, quintile 3 = 0.8, quintile 4 = 0.7 and quintile 5= 0.6.

The weighted England WP status scores (mean= 3.74, Std. Deviation =0.95, minimum = 2.4, maximum= 6.9), were included as predictors in a binary logistic regression model with LOWERSOC (1 = NS-SEC 3-7 versus 0 = NS-SEC 1-2) as the binary outcome.

A Hosmer-Lemeshow test (>0.05) confirmed adequate model fit and a Wald test (Chi2(1)= 2541.49, p<0.001, n=20,690) that weighted England WP status score had a significant effect on the outcome NS-SEC 3-7 versus NS-SEC 1-2.

The predicted probability of the outcome, when plotted against scores, graphically illustrated that as score increased so the probability of a subject being in NS-SEC 3-7 increased (Figure 11). An AUC of 0.73 (95% confidence interval 0.72 to 0.74) indicated that the England WP Index's discriminative ability to correctly classify subjects into NS-SEC 3-7 and NS-SEC 1-2 groups was fair to good (Figure 12).

An England WP Index threshold of 6.5 points (Table 9) was associated with an LR+ = 21.56, meeting the criterion of a useful diagnostic test (L=>10). However, the associated LR- was 0.98, an indication of false negative classification. Nevertheless, subjects with England WP status scores between 6.5 and 7 were highly likely to be in NS-SEC 3-7. Indeed, 87% (130/145) of those with scores at or above 6.5 points were from NS-SEC 3-7 backgrounds.

Given these results, we were able to reject the null-hypothesis that 'use of multiple, different types of contextual indicators does not mitigate the risk of false positive socioeconomic classification'.

The addition of the two contextual indicators of school (college) A-level performance, APSFTE_ALEVA and TALLPPE_ALEVA, did not improve the global accuracy of the proposed multidimensional measure of widening participation status or, reduce the proportion of false negative classification.

4.3 Results: MWGY sample

87% (705/810) of subjects had non-missing values on the five contextual indicators; POLAR quintile, IMD quintile, SCHOOL TYPE, PARED and BURSARY. Weighted scores were calculated using the method described above (section 4.2)

The weighted MWGY WP status scores (mean= 3.57, Std. Deviation = 0.79, minimum = 1.2, maximum= 5), were included as predictors in a binary logistic regression model with LOWERSOC (1 = NS-SEC 3-7 versus 0 = NS-SEC 1-2) as the binary outcome.

A Hosmer-Lemeshow test (>0.05) confirmed adequate model fit and a Wald test (Chi2(1)= 98.44, p<0.001, n=630) that weighted MWGY WP status score had a significant effect on the outcome NS-SEC 3-7 versus NS-SEC 1-2. The predicted probability of the outcome, when plotted against scores, graphically illustrated that as score increased so the probability of a subject being in NS-SEC 3-7 increased (Figure 13). An AUC of 0.79 (95% confidence interval 0.76 to 0.82) indicated that the MWGY WP Index's discriminative ability to correctly classify subjects into NS-SEC 3-7 and NS-SEC 1-2 groups was fair to good (Figure 14).

A UK WP Index threshold of 4.8 points (Table 10) was associated with an LR+ = 10.41, meeting the criterion of a useful diagnostic test (L=>10). However, the associated LR- was 0.90, an indication of false negative classification. Nevertheless, subjects with MWGY WP status scores between 4.8 and 5 were highly likely to be in NS-SEC 3-7. Indeed, 91% (50/55) of those with scores above at or above 4.8 points were from NS-SEC 3-7 backgrounds.

4.4 Summary of Results

- Area-level contextual indicators returned conflicting information on individual's social circumstances and correlated weakly with socioeconomic class. School and individual-level indicators correlated weakly with socioeconomic class.
- An IWPS derived from weighted scores on multiple types of contextual indicator identified students from lower socioeconomic class backgrounds with a high level of accuracy.

- Findings supported the hypothesis that the 'use of multiple, different types of contextual indicators mitigates the risk of false positive socioeconomic classification'.

4.5 Results: NS-SEC missing data values

Analysis of the UK sample revealed BME (chi2 (1)=105.41, p<0.001), IMD quintile (chi2 (4)=44.60, p<0.001) and PARED (chi2 (1)=146.21, p<0.001) to be significant predictors of missing data values on NS-SEC (n=31,820). The probability of missing on NS-SEC for BME students from areas of most deprivation whose parents had no higher education qualifications was 0.11 (95% confidence interval 0.10 to 0.13), compared to 0.01 (0.007 to 0.03) for white students from areas of least deprivation whose parent(s) had higher education qualifications. The mean predicted probability of missing data on NS-SEC for the analytic sample was 0.028 (95% confidence interval, 0.026 – 0.03).

4.6 Results: Latent class analysis

4.6.1 Model 1: UK sample (n= 30,595)

Models with one through five classes were compared and selection of the best-fit model was conducted using the associated BIC and AIC statistics (Table 11). Both BIC and AIC statistics were minimised at three classes (BIC = 265598.41, AIC = 265914.92) indicating that a three-class model provided the best fit to the data. Thus, each of the three classes contains a homogeneous group of students who share common characteristics with respect to the contextual indicators and social class, as measured by LOWERSOC (NS-SEC 3-7 versus NS-SEC 1-2).

Class 1 comprised 15% (4590/30595) of the sample, Class 2 19% (5810/30595) and Class 3 66% (20195/30595) (Table 12).

Across the three classes students in Class 1, labelled 'WP students'(Table 12) had the highest probability of being from a lower social class background (0.87 versus 0.12 versus 0.12), the highest probability of having attended a state funded school or college (0.90 versus 0.77 versus 0.64) the highest probability of having parents without HE qualifications (0.83 versus 0.13 versus 0.08), and the highest probability of being in receipt of a bursary or EMA (0.22 versus 0.04 versus 0.01). They also had the highest probability (0.21 versus 0.14 versus 0.01), of being from an area of highest deprivation (IMD quintile 5) and the second highest

probability of being from an area of the lowest rate of HE participation (0.11 versus 0.13 versus 0.01).

The sample mean score on the Index of Widening Participation Status (IWPS) was 2.7 (n=30595, s.d. = 0.80, range 1.2 – 5). The mean IWPS, for students in Class 1, 'WP students', was 3.64, compared to 2.66 for students in Class 2 and 2.03 for students in Class 3 (Table 12, bottom row).

However, there was little difference between students in classes 1 and 2. For instance, 38% of students in Class 2 (conditional probabilities 0.13 + 0.25) lived in areas of the lowest rates of HE participation, POLAR quintiles 1 and 2 (Table 12). Moreover, 39% of students in Class 2 lived in areas of higher deprivation (IMD quintiles 4 and 5). Nevertheless, students in Class 2 were unlikely to have lower social class parental backgrounds (0.12), unlikely to be have parents without HE qualifications or be in receipt of a bursary/EMA.

Students in Class 3, labelled 'Non-WP students' (Table 12) appeared to reflect traditional entrants to the study of medicine with 88% from top two social class parental backgrounds, 36% having attended privately funded schools or colleges, 1% in receipt of a bursary or EMA, 8% having parents without HE qualifications, 2% from areas of lower rates of participation in HE (POLAR quintiles 1 and 2), and 4% from areas of higher deprivation (IMD quintile 4 and 5).

For the three-class model, lower social class (NS-SEC 3–7 versus NS-SEC 1-2) and latent class assignment were significantly and positively correlated (Spearman's rho = 0.39, $p<0.001$, n=30595) (latent classes recoded with 'Non-WP students' =1, and 'WP students'=3). Moreover, lower social class (NS-SEC 3– 7 versus NS-SEC 1 -2) and IWPS score were significantly and positively correlated (Spearman's rho = 0.36, $p<0.001$, n=30595).

4.6.2 Model 2: England sample (n= 20,692)

Models with one through five classes were compared and selection of the best-fit model was conducted using the associated BIC and AIC statistics (Table 13). Both BIC and AIC statistics were minimised at three classes (BIC = 265555.45.87, AIC = 265071.26) indicating that a three-class model provided the best fit to the data.

Class 1 comprised 51% (10555/20690) of the sample, Class 2, 31% (6415/20690) and Class 3, 18% (3725/20690) (Table 14).

Across the three classes, students in Class 3, labelled 'WP students', (Table 14), had the highest probabilities of having attended a state funded school or college (0.98), of having parents without HE qualifications (0.64), of being in receipt of a bursary or EMA (0.26) and being from a lower social class background (0.64).They also had the highest probabilities (0.28), of being from an area of highest deprivation (IMD quintile 5) and of being from an area of the lowest rate of HE participation (0.18). Moreover, across the three classes students in Class 3 had the highest probabilities of having attended schools in the bottom quintile of both school A-level performance indicators (Table 14).

The sample mean score on the Index of Widening Participation Status (IWPS) (see Section 1 for weighting) was 3.72 (n=20690, s.d. = 0.93, range 2.4 – 6.90). The mean score on the IWPS for students in Class 3, 'WP students', was 5.12, compared to 3.14 for students in Class 1 and 3.95 for students in Class 2 (Table 14, bottom row).

Like students in Class 3, students in Class 2 were also highly likely to have attended a state school or college (0.93) but, in stark contrast, had low probabilities of being from a lower social class background (0.15), having parents without HE qualifications (0.13) being in receipt of a bursary or EMA (0.02), living in an area of lowest he participation (0.01) and an area of most deprivation (0.01). Nevertheless, a small minority attended schools in the bottom quintile of both school A-level performance indicators (0.03 and 0.01 respectively) (Table 14).

Students in Class 1, labelled 'Non-WP students' (Table 14) appeared to reflect traditional entrants to the study of medicine with 85% from top two social class(NS-SEC 1-2) parental backgrounds, 58% having attended privately funded schools or colleges, 2% in receipt of a bursary or EMA, 10% having parents without HE qualifications, 7% from areas of lower rates of participation in HE (POLAR quintiles 1 and 2), and 10% from areas of higher deprivation (IMD quintile 4 and 5). In stark contrast to the schools attended by students in the other two classes, nine out of ten students attended schools in the top quintile of both A-level school performance indicators (Table 14).

For the three-class model lower social class (NS-SEC 3-7 versus NS-SEC 1-2) and latent class assignment were significantly and positively correlated (Spearman's rho = 0.32, $p<0.001$, n=20690). Moreover, lower social class (NS-SEC 3 – 7 versus NS-SEC 1 -2) and IWPS score were significantly and positively correlated (Spearman's rho = 0.34, $p<0.001$, n=20690).

4.6.3 Model 3: Medicine With a Gateway Year sample (n= 630)

Models with one through five classes were compared and selection of the best-fit model was conducted using the associated BIC and AIC statistics (Table 15). Both BIC and AIC statistics were minimised at two classes (BIC =6368.01, AIC =6256.95) indicating that a two class model provided the best fit to the data. Class 1 comprised 42% (265/630) of the sample and Class 2 58% (365/630) (Table 16).

Students in Class 2, labelled 'WP students', (Table 16), were marginally more likely to have attended a state funded school or college (0.87 versus 0.86), more likely to have parents without HE qualifications (0.91 versus 0.35), more likely in receipt of a bursary or EMA (0.38 versus 0.14) and more likely from a lower social class background (0.92 versus 0.15). Moreover, 26% of Class 2 students lived areas of lower participation in HE (POLAR quintiles 1 and 2) and 67% in areas of higher derivation (IMD quintiles 4 and 5) indicating a disadvantaged contextual background.

Although students in Class 1 were highly likely to be from top two social class(NS-SEC 1-2) backgrounds (0.85), 19% were from areas of lower participation in HE and 48% from areas of higher deprivation again reflecting the far from straightforward link between area-level contextual indicators and social class. Moreover, 74% of students in Class 2, 'WP students' lived in areas with higher rates of HE participation (POLAR quintiles 3-5) and 33% lived in areas of lesser deprivation (IMD quintiles 1-3).

The sample mean score on the Index of Widening Participation Status (IWPS) (see Section 1 for weighting) was 3.54 (n=630, s.d. = 0.78, range 1.2 – 5). The mean score on the IWPS for students in Class 2, 'WP students', was 3.93, compared to 2.99 for students in Class 1 (Table 16, bottom row).

For the two-class model lower social class (NS-SEC 3-7 versus NS-SEC 1-2) and latent class assignment were significantly and positively correlated (Spearman's rho = 0.89, $p<0.001$, n=630). Moreover, lower social class (NS-SEC 3-7 versus NS-SEC 1-2) and IWPS score were significantly and positively correlated (Spearman's rho = 0.43, $p<0.001$, n=630).

4.7 Summary of results: LCA

- For each of the samples examined LCA identified a subgroup of nominally 'WP students' characterised by high probabilities of being from a lower social class

background, having attended a state funded school or college, of being in receipt of a bursary or EMA, and having parents without HE qualifications.

- And in respect of the England sample, the highest probability of having attended schools in the bottom quintile of both indicators of A-level performance.
- Within each model reported the mean IWPS score of students in the 'WP students' latent class was greater than the mean IWPS scores of students in the other latent classes within the same model.
- Students' IWPS score and latent class assignment were significantly correlated with lower social class position.
- For each of the samples examined LCA identified a subgroup of nominally 'Non-WP students' who appeared to reflect traditional entrants to the study of medicine. These students were characterised by a lower probability of having attended a state funded school or college, and very low probabilities of being from a lower social class background, being in receipt of a bursary or EMA, having parents without HE qualifications, and living in areas of lower HE participation and higher deprivation. And in respect of the England sample, the highest probability out of having attended schools in the top quintile of both indicators of A-level performance.
- However, a very small proportion classified as 'Non-WP' lived in areas of lower participation in HE and or areas of higher deprivation. In further contrast, in models 1 and 2 LCA identified a third subgroup of students characterised by a low probability of being from a lower social class background, a high probability of having attended a state funded school or college, and included a proportion who lived in areas of lowest HE participation and highest deprivation.
- A third typology characterised by a low probability of being from a lower social class background, a high probability of having attended a state funded school or college, low probability of having parents without HE qualifications, low probability of being in receipt of a bursary /EMA, included a proportion who lived in areas of lowest HE participation and highest deprivation.
- These findings reflect the far from straightforward link between contextual indicators, social circumstances and social class as evidenced in Section 1 of this

study. Particularly, that the average characteristics of a neighbourhood are misleading indicators of individuals' social circumstances and socioeconomic class.

- Nevertheless, they support the conclusion reached above, that use of multiple contextual indicators of disadvantage reduces the risk of false positive lower social class position.

4.8 Results: Multiple imputation

The largest percentage of missing values on any of the variables to be included in the MICE model was 12.34% (Table 17, missing data patterns are described in Table 18) and thus in line with guidance the number of model iterations was set at 15.

Analysis of missing data values indicated that the missing values on the contextual indicators and the outcome variable LOWERSOC were MAR because BME and Bursary were significantly associated with missing values on LOWERSOC and many of the contextual variables to be imputed (Table 19). Hence, both were included as auxiliary variables in the MICE model.

Sensitivity analysis was conducted by comparing the parameter estimates from (1) a logistic regression of the complete case contextual indicators and the complete case outcome LOWERSOC with (2) the parameter estimates from a logistic regression of the final imputed values produced by the MICE programme on the contextual indicators and the outcome LOWERSOC (Table 20). Comparison of the coefficients, and their associated significance and standard errors revealed little difference across the models and Wald tests (Table 20) that in both models each contextual indicator had a significant independent effect on the outcome LOWERSOC (NS-SEC 3-7 versus NS-SEC 1-2). Thus, estimates of both models (complete case analysis, n = 30595) and imputed values (n = 40190) provided similar pictures of the relationship between POLAR quintile(1-5), IMD quintile(1-5), SCHOOL TYPE(1-0), PARED (1-0), BURSARY (1-0) and the outcome LOWERSOC (NS-SEC 1-2 versus NS-SEC 3-7), indicating that the parameter estimates of the complete case analysis were not heavily biased by missing data values.

An imputed IWPS score was calculated for all cases (n=40190) by recoding and summing the five imputed predictor variables across the sample:-

- POLAR quintile 1=1, quintile 2= 0.9, quintile 3= 0.8, quintile 4 = 0.7 and quintile 5 = 0.6.

- IMD quintile 1= 0.6, quintile 2= 0.7, quintile 3= 0.8, quintile 4 = 0.9 and quintile 5 = 1.
- SCHOOL TYPE, PARED and BURSARY were weighted 1= state funded school, 1=parent no HE qualifications, and 1= in receipt of a bursary/EMA, and zero otherwise.

The imputation based IWPS score (mean = 2.36, Std. Deviation = 0.82, minimum = 1.2, maximum = 5.0, n = 40190) was used as a sole predictor in a binary logistic regression with imputed values on LOWERSOC (1= NS-SEC 3-7, 0=NS-SEC 1-2) as the binary outcome.

A Hosmer-Lemeshow test ($p>0.05$) confirmed adequate model fit and a Wald test ($Chi2(1) = 4675.30$, $P<0.001$), that IWPS had a significant effect on the outcome imputed LOWERSOC (NS-SEC 3-7 versus NS-SEC 1-2).

The predicted probability of the outcome, when plotted against scores, graphically illustrated that as score increased so the probability of a subject being in NS-SEC 3-7 increased (Figure 15). An AUC of 0.73 (95% confidence interval 0.72 to 0.73) (Figure 16) indicated that the IWPS discriminative ability to correctly classify subjects into NS-SEC 3-7 backgrounds was fair to good.

Cross-tabulation of IWPS scores and the outcome LOWERSOC (NS-SEC 3-7 versus NS-SEC 1-2) revealed that 83% (150/714) of students with scores between 4.5 and 5.0 points were from NS-SEC 3-7 backgrounds (Table 21). Moreover, an IWPS score from 4.5 to 5.0 (Table 22) was associated with an LR+ meeting the criterion of a useful diagnostic test (L+ >+10). However, the associated LR- statistic indicated a proportion of misclassification.

4.9 Summary of results: Multiple imputation

- An IWPS derived from imputed missing values identified students from lower socioeconomic class backgrounds with a level of accuracy approximating that of the complete case analysis.
- The multiple imputation results indicated that missing data did not heavily bias the parameter estimates of the complete case analysis and thereby supports the inference of the complete case analysis that the ‘use of multiple, different types of contextual indicators mitigates the risk of false positive socioeconomic classification’.

5 Conclusions

- The evidence of this study enabled acceptance of the hypothesis that the 'use of multiple, different types of contextual indicators mitigates the risk of false positive socioeconomic classification'.
- The evidence of this study indicates that an Index of Widening Participation Status if used to inform admissions decision-making, based on applicants' weighted scores on the contextual indicators; POLAR, IMD, SCHOOL TYPE, PARED and BURSARY, would facilitate widening participation among applicants from lower socio-economic classes.
- Area-level contextual indicators returned conflicting information on individual's social circumstances and correlated weakly with socioeconomic class. School and individual-level indicators correlated weakly with socioeconomic class.
- However, an IWPS derived from weighted scores on multiple types of contextual indicator identified students from lower socioeconomic class backgrounds with a high level of accuracy.
- The findings of this study reflect the far from straightforward link between contextual indicators, social circumstances and social class. Particularly, that the average characteristics of a neighbourhood are misleading indicators of individuals' social circumstances and socioeconomic class.

6 Limitations and further study

We acknowledge the limitations imposed on the findings of this study by the level of missing data values on the contextual indicators held in the UKMED database.

We also acknowledge that the utility of contextual indicators may be undermined not only by missing data values but also, in respect of aggregate, area based measures such as POLAR, IMD, IDACI, and IDAOPI, by ecological fallacy. Indeed, the majority of disadvantaged families in the UK do not live in areas of low Higher Education participation whereas a substantial minority of relatively wealthy residents do live in areas of low Higher Education participation and in areas of high deprivation. [25] [26]

We also acknowledge that binary contextual indicator type of school/college attended, state funded versus privately funded, and the former a flag for widening participation, may do injustice to the students from disadvantaged backgrounds attending privately funded schools. As Boliver points out, 'some very small private schools have only nominal fees and serve some of the poorest communities in the UK' and a significant minority of pupils at private schools are in receipt of bursaries and scholarships. [10] Additionally, the information on NS-SEC and PARED is self-declared and is open to abuse. [27]

We further acknowledge, and are mindful for our ongoing research, that contextualised admissions are a high stakes assessment for entry to the study of medicine and the validity of any assessment is 'the degree to which evidence and theory support or refute the interpretation of test scores entailed by proposed use of the test'. [28]

7 Acknowledgment

Source - UK Medical Education Database ("UKMED") UKMEDP041 extract generated on 03/04/2018 and refreshed on 16/05/2019. Approved for publication on 07/12/2018. We are grateful to UKMED for the use of these data. However, UKMED bears no responsibility for their analysis or interpretation. The data includes information derived from that collected by the Higher Education Statistics Agency Limited ("HESA") and provided to the GMC ("HESA Data"). Source: HESA Student Record 2002/03 to 2015/16 Copyright Higher Education Statistics Agency Limited. The Higher Education Statistics Agency Limited makes no warranty as to the accuracy of the HESA Data, cannot accept responsibility for any inferences or conclusions derived by third parties from data or other information supplied by it.

8 Tables

Table 1: Sociodemographic characteristics of the samples.

		UK sample n = 40,190		England sample n = 32825		Medicine With a Gateway Year sample n = 810	
Factor	Category	N	%	N	%	N	%
Gender	Male	18185	45.24	14935	45.00	360	44.38
	Female	22010	54.76	17890	55.00	450	55.62
	All	40190	100.00	32825	100.00	810	100.00
Age at entry to medical school	17 years	315	00.78	45	0.14	-	-
	18 years	19410	48.30	15280	46.54	350	43.39
	19 years	16735	41.64	14290	43.52	360	44.38
	20 years	3730	9.29	3215	9.80	100	12.24
	All	40190	100.00	32825	100.00	810	100.00
UK region of domicile (prior to commencement of course)	England	32825	81.68	32825	100.00	790	97.40
	Northern Ireland	2430	6.05			7	SUPPR
	Scotland	3080	7.66			SUPPR	SUPPR
	Wales	1850	4.60			10	SUPPR
	Missing	5	SUPPR				
	All	40190	100.00	32825	100.00	810	100.00
SEC (NS-SEC 1-7) (socio-economic classification of the student's parent if under 21 years of age) with missing values infilled if non-missing on SEC COMBINED	1. Higher managerial /professional	17330	43.12	14200	43.26	75	10.61
	2. Lower managerial /professional	12295	30.59	9890	30.13	210	29.61
	3. Intermediate occupations	3650	9.08	2920	8.90	85	11.87
	4. Small employer own account workers	1870	4.65	1500	4.57	70	8.78
	5. Lower supervisory /technical	840	2.09	650	1.99	30	3.96
	6. Semi-routine occupations	2130	5.30	1850	5.64	155	19.16
	7. Routine occupations	695	11.72	590	1.79	85	10.51
	Missing	1385	3.45	1225	3.73	95	11.50
	All	40190	100.00	32825	100.00	810	100.00
SEC COMBINED (HESA & UKCAT socioeconomic data combined	Managerial /professional occupations	29625	73.71	24090	73.39	290	35.97
	Intermediate occupations	3650	9.08	2920	8.90	85	10.51
	Lower supervisory/technical occupations	840	2.09	650	1.99	35	4.08
	Small employer own account worker	1870	4.65	1500	4.57	75	9.52
	Semi-routine/ routine occupations	2815	7.00	2430	7.41	220	27.19
	Missing	1390	3.46	1300	3.74	105	12.73
	All	40190	100.00	32825	100.00	810	100.00
GOLDTHORPE'S SEC	Salariat (SEC 1 & 2)	29625	73.71	24090	73.38	290	35.60

(not a UKMED variable)	Intermediate (SEC 3, 4 and 5)	6360	15.82	5075	15.45	190	23.24
	Working class (SEC 6 and 7)	2820	7.02	2440	7.44	240	29.67
	Missing	1385	3.45	1225	3.73	95	11.50
	All	40190	100.00	32825	100.00	810	100.00
LOWER CLASSES	Yes = SEC 3, 4, 5, 6, and 7	9180	22.84	7510	22.89	430	52.90
	No = SEC 1 & 2	29625	73.71	24090	73.38	290	35.60
	Missing	1385	3.45	1225	3.73	90	11.50
	All	40190	100.00	32825	100.00	810	100.00
Ethnicity	White	27275	67.87	20825	63.44	245	30.41
	Asian/Asian British	9110	22.67	8320	25.95	300	37.33
	Black/Black British	1025	2.53	980	2.99	170	20.64
	Mixed	1700	4.23	1510	4.61	40	4.70
	Other	960	2.38	880	2.68	55	6.55
	Missing	125	0.31	110	0.33	SUPPR	0.37
	All	40190	100.00	32825	100.00	810	100.00
BME	Yes	12790	31.83	11895	36.33	560	69.22
	No	27275	67.97	20825	63.44	245	30.41
	Missing	120	0.31	110	0.33	SUPPR	0.37
	All	40190	100.00	32825	100.00	810	100.00
Disabled	Yes	100	00.24	80	0.24	95	11.62
	No	10955	27.26	8770	26.27	-	-
	Missing	29140	72.50	23975	73.04	715	88.38
	All	40190	100.00	32825	100.00	810	100.00

Table 2: Contextual indicator frequencies for the samples.

		UK sample n = 40,190		England sample n = 32825		Medicine With a Gateway Year sample n = 810	
Factor	Category	n	%	n	%	n	%
POLAR 3 Postcode based, quintile classification of areas for young participation rates in higher education, where 1 = lowest to 5 = highest.	1	1580	3.94	1390	4.24	65	8.26
	2	3370	8.39	2890	8.80	110	13.47
	3	5570	13.85	4790	14.59	235	28.80
	4	9400	23.38	7755	23.63	200	24.97
	5	20190	50.23	15940	48.55	195	24.10
	Missing	85	0.21	65	0.20	SUPPR	SUPPR
	All	40190	100.00	32825	100.00	810	100.00
IMD Postcode based, quintile classification of areas for index of multiple deprivation where 1= least deprived to 5 = most deprived.	1	14525	36.14	12575	38.30	65	8.16
	2	9260	23.04	8340	25.40	100	12.48
	3	6565	16.33	5915	18.01	145	17.92
	4	3965	9.86	3650	11.12	210	26.21
	5	2360	5.87	2240	6.83	275	33.99
	Missing	3520	8.76	110	0.34	10	1.24
	All	40190	100.00	32825	100.00	810	100.00
QAHE	1	2150	5.35	1805	5.49	90	11.37
	2	4210	10.47	3305	10.07	120	14.46

Postcode based, quintile classification of areas for the proportion of people aged 16 -74 with higher education qualifications where 1= lowest to 5 = highest.	3	6700	16.66	5505	16.77	145	17.68
	4	10760	26.77	9405	28.65	195	24.23
	5	16300	40.54	12745	38.83	260	31.89
	Missing	85	1.21	65	0.20	SUPPR	SUPPR
	All	40190	100.00	32825	100.00	810	100.00
SCHOOL TYPE Attended between ages of 11 and 16.	State funded	27490	68.40	21675	32.82	775	95.55
	Privately funded	12180	30.31	10775	66.03	30	3.34
	Missing	520	1.29	375	1.15	10	SUPPR
	All	40190	100.00	32825	100.00	810	100.00
PARED Parent had higher education qualifications	Yes	27895	69.63	22415	68.28	230	28.43
	No	7245	18.03	6190	18.86	495	61.19
	Missing	4960	12.34	3225	12.86	85	10.38
	All	40190	1000.00	32825	100.00	810	100.00
PARENT DEGREE Parent completed degree course or equivalent.	Yes	4595	11.43	3395	10.95	5	0.62
	No	1810	4.50	1450	4.42	20	2.47
	Missing	33785	84.07	27780	84.63	785	96.91
	All	40190	100.00	32825	100.00	810	100.00
BURSARY Student in receipt/ received UKCAT Bursary or Educational Maintenance Grant (EMA)	Yes	2140	5.32	1850	5.63	240	29.54
	No	38055	94.68	30980	94.37	570	70.46
	Missing	-	-	-	-	-	-
	All	40190	100.00	32835	100.00	810	100.00
INCOME SUPPORT Student's household received income support during school years	Yes	810	2.01	645	1.97	15	1.85
	No	5120	12.74	4020	12.24	5	SUPPR
	Missing	34265	85.25	28160	85.79	790	97.40
	All	40190	100.00	32825	100.00	810	100.00
FREE SCHOOL MEALS Student had free school meals	Yes	420	1.05	360	1.10	10	1.11
	No	5790	14.41	4525	13.78	15	1.61
	Missing	33975	84.54	27940	85.12	790	97.28
	All	40190	100.00	32825	100.00	810	100.00

Continuous indicators	N non-missing	Min	Max	Mean	SD
IDACI RANK (postcode based) Income Deprivation Affecting Children Index where 1 = most deprived to 32482 least deprived	32715	SUPPR	32480	21281.16	8797.68
IDAOPI RANK (postcode based) Income Deprivation Affecting Older People Index where 1 = most deprived to 32482 least deprived	32715	SUPPR	32482	21108.54	8911.68
TALLPPE ALEVA Average point score per A-level entry for year taken,	27550	93.2	281.3	232.13	21.10
APSFTE ALEVA Average point score per A-level student for year taken,	275	258.3	1650	913.69	149.34

Table 3: Correlation matrix of UK level contextual indicators held in the UKMED using Spearman's rank correlation, where -1 represents a perfect negative correlation, +1 a perfect positive correlation, and 0 no association (n=28,190).All coefficients were statistically significantly different from zero at the p<0.001 level.

Indicator	POLAR quintile	IMD quintile	AHE quintile	SCHOOL TYPE	PARED	BURSARY	IDACI decile	IDAOPI decile
POLAR quintile	1.00							
IMD quintile	-0.43	1.00						
QAHE quintile	0.75	-0.35	1.00					
SCHOOL TYPE	-0.16	0.10	-0.21	1.00				
PARED	-0.20	0.21	-0.25	0.16	1.00			
BURSARY	-0.14	0.18	-0.22	0.09	0.23	1.00		
IDACI decile	0.42	-0.80	0.32	--0.14	-0.21	-0.18	1.00	
IDAOPI decile	0.39	-0.7903	00.31	-0.12	-0.21	-0.18	0.76	1.00

Table 4: Correlation matrix of UK level contextual indicators and SEC using Spearman's rank correlation, where -1 represents a perfect negative correlation, +1 a perfect positive correlation, and 0 no association (n=30,595). All coefficients were statistically significantly different from zero at the p<0.001 level.

Indicator	SEC	POLAR quintile	IMD quintile	SCHOOL TYPE	PARETHE	BURSARY
SEC	1.00					
POLAR quintile	- 0.17	1.00				
IMD quintile	0.21	-0.43	1.00			
SCHOOL TYPE	0.16	-0.16	0.08	1.00		
PARENTHE	0.43	--0.20	0.19	0.14	1.00	
BURSARY	0.21	-0.13	0.16	0.08	0.21	1.00

Table 5: Results of linear regression models to determine weighting of scores on contextual indicators.

UK sample (n=31,497)						
Predictor	*Coefficient*	*Std.Error*	*t*	*P>t*	*95% Confidence Interval*	
POLAR	*-1.005*	*0.0015*	*-68.55*	*<0.001*	*-0.1035*	*-0.0978*
IMD	*0.1099*	*0.0014*	*79.36*	*<0.001*	*0.1073*	*0.1127*
SCHOOL TYPE	*0.9905*	*0.0035*	*297.79*	*<0.001*	*0.9839*	*0.9907*
PARED	*1.0205*	*0.0038*	*264.19*	*<0.001*	*0.0290*	*1.0280*
BURSARY	*1.0906*	*0.0068*	*160.05*	*<0.001*	*1.0797*	*1.1064*
England sample (21,102)						
POLAR	*1.03*	*0.007*	*141.01*	*<0.001*	*1.02*	*1.05*
IMD	*1.18*	*0.007*	*176.21*	*<0.001*	*1.17*	*1.19*
SCHOOL TYPE	*1.02*	*0.003*	*294.03*	*<0.001*	*1.01*	*0.9907*
PARED	*1.02*	*0.004*	*264.19*	*<0.001*	*0.03*	*1.04*
BURSARY	*1.18*	*0.006*	*175.01*	*<0.001*	*1.16*	*1.19*
APSTFE	*1.07*	*0.013*	*78.01*	*<0.001*	*1.04*	*1.09*
TALLPPE	*1.01*	*0.011*	*91.28*	*<0.001*	*0.99*	*1.04*
MWGY sample (809)						
POLAR	*-0.12*	*0.010*	*-12.34*	*<0.001*	*-0.15*	*-0.11*
IMD	*0.13*	*0.091*	*21.61*	*<0.001*	*0.11*	*0.15*
SCHOOL TYPE	*1.00*	*0.041*	*24.82*	*<0.001*	*0.92*	*1.08*
PARED	*0.94*	*0.024*	*37.81*	*<0.001*	*0.89*	*0.98*
BURSARY	*1.01*	*0.029*	*34.32*	*<0.001*	*0.95*	*1.06*

Table 6: Cross-tabulation of weighted scores and the outcome NS-SEC 3-7 versus NS-SEC 1&2 at cut-score.

Cut-score	*NS-SEC 1-2*	*NS-SEC 3-7*	*Total*
>=4.5 (range 1.2 to 5.0)	*n=90 (14.63%)*	*n=530 (85.37%)*	*620*

Table 7: Report of sensitivity and specificity for the UK index of Widening Participation.

Cut score	Sensitivity	Specificity	Correctly Classified	LR+	LR-
>= 1.2	100.00%	0.00%	23.71%	1	
>= 1.3	96.31%	10.58%	30.90%	1.077	0.3493
>= 1.4	93.33%	18.09%	35.93%	1.1394	0.3688
>= 1.4	92.75%	19.96%	37.22%	1.1587	0.3633
>= 1.5	91.49%	22.82%	39.10%	1.1855	0.3727
>= 1.6	90.10%	26.22%	41.36%	1.2212	0.3775
>= 1.6	89.72%	26.98%	41.86%	1.2287	0.3811
>= 1.7	89.34%	27.74%	42.34%	1.2364	0.3842
>= 1.8	88.81%	28.93%	43.13%	1.2496	0.3869
>= 1.9	88.52%	29.52%	43.51%	1.256	0.3889
>= 2)	88.27%	29.82%	43.68%	1.2577	0.3934
>= 2.2	88.19%	29.95%	43.76%	1.2589	0.3945
>= 2.3	79.45%	47.80%	55.30%	1.5219	0.43
>= 2.4	71.92%	61.21%	63.75%	1.854	0.4588
>= 2.5	64.65%	71.33%	69.75%	2.255	0.4955
>= 2.6	59.55%	78.72%	74.18%	2.799	0.5138
>= 2.7	55.64%	83.15%	76.63%	3.3018	0.5335
>= 2.8	52.59%	86.13%	78.18%	3.7909	0.5504
>= 2.9	50.22%	88.26%	79.24%	4.2779	0.564
>= 3)	49.05%	89.40%	79.83%	4.6255	0.57
>= 3.2	48.54%	89.93%	80.11%	4.8188	0.5723
>= 3.3	44.09%	91.71%	80.41%	5.3149	0.6097
>= 3.4	39.00%	93.33%	80.45%	5.8461	0.6536
>= 3.5	32.97%	95.03%	80.32%	6.6405	0.7053
>= 3.6	27.56%	96.39%	80.07%	7.6297	0.7516
>= 3.7	22.83%	97.33%	79.67%	8.5526	0.7929
>= 3.8	18.46%	98.11%	79.22%	9.7694	0.8311
>= 3.9	14.01%	98.79%	78.69%	11.5512	0.8705
>= 4)	10.93%	99.24%	78.30%	14.3342	0.8975
>= 4.2	9.51%	99.40%	78.08%	15.7454	0.9104
>= 4.3	8.99%	99.48%	78.03%	17.3374	0.9149
>= 4.4	8.24%	99.54%	77.89%	17.982	0.9218
>= 4.5	7.32%	99.61%	77.73%	18.775	0.9304
>= 4.6	6.22%	99.67%	77.51%	18.8457	0.9409
>= 4.7	4.91%	99.75%	77.26%	19.7493	0.9533
>= 4.8	3.53%	99.81%	76.98%	18.7205	0.9665
>= 4.9	1.93%	99.90%	76.67%	19.5854	0.9817
>= 5	0.57%	99.96%	76.39%	13.1913	0.9948
ROC					
Observations		Area =	Standard Error	95% CI	
30595		0.7412	0.004	0.7317	0.7428

Table 8: England domiciled students who had attended English school or college : Correlation matrix of contextual indicators and SEC using Spearman's rank correlation, where -1 represents a perfect negative correlation, +1 a perfect positive correlation, and 0 no association (n=20,100). All coefficients were statistically significantly different from zero at the p<0.001 level.

Indicator	IMD quintile	APSFTE quintile	TALLPPE quintile	POLAR quintile	SCHOOL TYPE	PARED	BURSARY
IMD quintile	1.00						
APSFTE quintile	-0.11	1.00					
TALLPPE quintile	-0.11	0.63	1.00				
POLAR quintile	-0.35	0.13	0.24	1.00			
SCHOOL TYPE	0.11	-0.29	-0.51	0.21	1.00		
PARED	0.21	-0.15	-0.18	-0.24	0.15	1.00	
BURSARY	0.20	-0.11	-0.11	-0.13	0.11	0.26	1.00

Table 9: Report of sensitivity and specificity for the England index of Widening Participation.

Cut score	Sensitivity	Specificity	Correctly Classified	LR+	LR-
>= 2.4	100.00%	0.00%	23.66%	1	
>= 2.5	97.12%	8.50%	29.47%	1.0614	0.3387
>= 2.6	94.12%	16.30%	34.71%	1.1245	0.3608
>= 2.6	93.18%	19.34%	36.81%	1.1552	0.3527
>= 2.7	91.87%	22.43%	38.86%	1.1844	0.3624
>= 2.8	90.03%	26.70%	41.68%	1.2282	0.3734
>= 2.8	89.30%	28.33%	42.76%	1.246	0.3778
>= 2.9	89.03%	29.32%	43.45%	1.2596	0.3741
>= 3)	88.03%	30.69%	44.66%	1.2701	0.39
>= 3.1	87.60%	31.44%	44.73%	1.2777	0.3944
>= 3.2	87.30%	31.94%	45.04%	1.2826	0.3978
>= 3.2	87.28%	31.96%	45.05%	1.2828	0.3981
>= 3.3	87.17%	32.24%	45.23%	1.2864	0.3979
>= 3.4	87.13%	32.29%	45.26%	1.2868	0.3985
>= 3.5	84.56%	37.21%	48.41%	1.3466	0.415
>= 3.6	80.51%	44.67%	53.15%	1.4552	0.4362
>= 3.6	77.25%	50.32%	56.69%	1.555	0.4521
>= 3.7	75.16%	52.95%	58.21%	1.5975	0.4691
>= 3.8	69.38%	62.19%	63.89%	1.8352	0.4923
>= 3.8	65.30%	67.66%	67.10%	2.0193	0.5129
>= 3.9	63.89%	69.18%	67.93%	2.0731	0.522
>= 4)	59.91%	75.59%	71.88%	2.454	0.5304
>= 4.1	56.58%	79.99%	74.45%	2.8272	0.5429
>= 4.2	53.84%	83.57%	76.53%	3.276	0.5524
>= 4.3	51.78%	85.86%	77.80%	3.6626	0.5616
>= 4.4	50.67%	87.40%	78.71%	4.0223	0.5644
>= 4.5	49.02%	88.72%	79.33%	4.3476	0.5746

>= 4.6	46.79%	89.83%	79.64%	4.5995	0.5923
>= 4.7	44.00%	91.32%	80.12%	5.0689	0.6133
>= 4.8	39.75%	92.78%	80.23%	5.5025	0.6494
>= 4.9	35.56%	94.10%	80.25%	6.0268	0.6848
>= 5)	31.68%	95.12%	88.11%	6.4903	0.7183
>= 5.1	27.94%	95.99%	79.89%	6.9615	0.7507
>= 5.2	24.31%	96.94%	79.75%	7.9325	0.7809
>= 5.3	20.81%	97.65%	79.47%	8.8615	0.8109
>= 5.4	17.63%	98.20%	79.14%	9.8039	0.8388
>= 5.5	15.71%	98.68%	79.05%	11.928	0.8542
>= 5.6	13.71%	98.94%	78.77%	12.9632	0.8722
>= 5.7	11.97%	99.13%	78.51%	13.8001	0.888
>= 5.8	10.44%	99.25%	78.24%	13.9716	0.9024
>= 5.9	9.56%	99.40%	78.14%	15.8938	0.9099
>= 6)	8.37%	95.53%	77.96%	17.637	0.9206
>= 6.1	7.11%	99.59%	77.71%	17.543	0.9327
>= 6.2	5.94%	99.71%	77.52%	20.4098	0.9433
>= 6.3	4.55%	99.78%	77.25%	20.5564	0.9566
>= 6.4	3.72%	99.82%	77.08%	20.9708	0.9645
>= 6.5	2.59%	99.88%	76.86%	21.5651	0.9752
>= 6.6	1.61%	99.93%	76.67%	23.1722	0.9845
>= 6.7	0.86%	99.97%	76.52%	27.1003	0.9917
>= 6.8	0.35%	99.98%	76.41%	18.2849	0.9967
>= 6.9	0.02%	99.99%	76.34%	3.2255	0.9999

ROC				
Observations	Area =	Standard Error	95% CI	
20,690	0.7337	0.0044	0.72251	0.7443

Table 10: Report of sensitivity and specificity for the MWGY index of Widening Participation.

Cut score	Sensitivity	Specificity	Correctly Classified	LR+	LR-
>= 1.2	100.00%	0.00%	48.99%	1	
>= 1.3	99.14%	3.94%	50.58%	1.032	0.2192
>= 1.4	98.06%	7.05%	51.64%	1.055	0.2756
>= 1.4	98.06%	7.26%	51.75%	1.0573	0.2677
>= 1.5	98.06%	7.47%	51.85%	1.0597	0.2603
>= 1.6	98.06%	8.51%	52.38%	1.0717	0.2285
>= 1.6	98.06%	8.71%	52.49%	1.0742	0.2231
>= 1.7	98.06%	9.75%	53.02%	1.0865	0.1993
>= 1.8	98.06%	10.37%	53.33%	1.0941	0.1874
>= 2.2	97.84%	10.58%	53.33%	1.0942	0.2041
>= 2.3	95.25%	24.90%	59.37%	1.2682	0.1909
>= 2.4	92.66%	35.89%	63.70%	1.4453	0.2046
>= 2.5	89.42%	43.57%	66.03%	1.5845	0.2429
>= 2.6	87.69%	50.83%	68.89%	1.7834	0.2422
>= 2.7	85.31%	54.56%	69.63%	1.8777	0.2692
>= 2.8	83.37%	61.00%	71.96%	2.1374	0.2727
>= 2.9	80.78%	66.60%	73.54%	2.4183	0.2886
>= 3)	68.46%	74.07%	79.91%	2.5341	0.2934
>= 3.2	78.83%	69.09%	73.86%	2.5502	0.3064
>= 3.3	76.67%	71.99%	74.29%	2.7375	0.324
>= 3.4	73.00%	74.48%	73.76%	2.8607	0.3625
>= 3.5	66.74%	78.22%	72.59%	3.0636	0.4253
>= 3.6	59.40%	81.95%	70.90%	3.2906	0.4955
>= 3.7	52.92%	85.06%	69.31%	3.5424	0.5535
>= 3.8	42.55%	89.00%	66.24%	3.8695	0.6455
>= 3.9	31.97%	92.53%	62.86%	4.2798	0.7353
>= 4)	93.57%	61.69%	28.51%	4.4328	0.764
>= 4.2	26.35%	94.81%	61.27%	5.0803	0.7768
>= 4.3	25.70%	94.81%	60.95%	4.9553	0.7836
>= 4.4	24.19%	95.02%	60.32%	4.8582	0.7978
>= 4.5	22.25%	96.06%	59.89%	5.6435	0.8094
>= 4.6	20.09%	96.89%	59.26%	6.4544	0.8248
>= 4.7	16.85%	97.30%	57.88%	6.2462	0.8546
>= 4.8	10.80%	98.96%	55.77%	10.4104	0.9014
>= 4.9	4.54%	99.38%	52.91%	7.2873	0.9606
>= 5)	1.30%	99.59%	51.43%	3.1231	0.9912
ROC					
Observations		Area =	Standard Error	95% CI	
945		0.7877	0.0149	0.7585	0.8168

Table 11: Model 1 goodness-of-fit statistics for latent class analysis models with one through five classes, UK sample (n=30,595).

Number of Classes	Degrees of Freedom	G^2 (L^2)	p-value	AIC	BIC
1	387	18298.57	<0.001	281763.42	281863.36
2	374	5422.77	<0.001	268913.56	269121.77
3	361	2081.59	<0.001	265598.41	265914.92
4	348	1649.73	<0.001	266192.57	266617.33
5	337	773.80	<0.001	267338.64	267855.01
AIC = Aikake information criterion statistic; BIC = Bayesian information criterion statistic.					

Table 12: Outline of conditional and latent class probabilities for the three-class model, UK sample (n=30,595).

Conditional Probabilities			
Three-class model	Class I 'WP students'	Class II	Class III 'Non-WP students'
Indicators			
State funded school	0.90	0.77	0.64
Parent has no HE	0.83	0.13	0.08
BURSARY	0.22	0.04	0.01
LOWERSOC	0.87	0.12	0.12
POLAR quintile			
1	0.11	0.13	0.01
2	0.19	0.25	0.01
3	0.23	0.33	0.07
4	0.24	0.28	0.22
5	0.23	0.01	0.69
IMD quintile			
1	0.17	0.05	0.56
2	0.19	0.25	0.27
3	0.21	0.31	0.13
4	0.22	0.25	0.03
5	0.21	0.14	0.01
Latent Class Probability 95% Confidence Interval Standard Error	0.15 0.14 – 0.16 0.01	0.19 0.18 – 0.20 0.01	0.66 0.65 – 0.67 0.01
Descriptive statistics by latent class			
Mean Index of Widening Participation Status	3.64 (s.d.= 0.61) min = 1.6 , max = 5	2.66 (s.d.= 0.59) min = 1.6 , max = 4	2.03 (s.d.= 0.58) min = 1.2 , max = 4.2

Table 13: Model 2 goodness-of-fit statistics for latent class analysis models with one through five classes, England sample (n=20,692).

Number of Classes	Degrees of Freedom	G^2 (L^2)	p-value	AIC	BIC
1					
2	9959	18618.93	<0.001	271086.76	271404.26
3	**9938**	**12561.43**	**<0.001**	**265071.26**	**265555.45**
4	9918	10186.12	<0.001	265079.30	265559.21
5	9900	9525.05	<0.001	265081.27	265563.83
AIC = Aikake information criterion statistic; BIC = Bayesian information criterion statistic.					

Table14: Outline of conditional and latent class probabilities for the three-class model England sample (n=20,690).

Conditional Probabilities			
Three-class model	*Class I 'Non-WP students'*	*Class II*	*Class III 'WP students'*
Indicators			
State funded school	0.42	0.93	0.98
Parent has no HE	0.10	0.13	0.64
BURSARY	0.02	0.02	0.26
LOWERSOC	0.15	0.15	0.64
POLAR quintile			
1	0.02	0.01	0.18
2	0.04	0.06	0.25
3	0.10	0.15	0.26
4	0.22	0.29	0.20
5	0.62	0.49	0.11
IMD quintile			
1	0.44	0.47	0.07
2	0.27	0.30	0.15
3	0.19	0.16	0.22
4	0.08	0.06	0.28
5	0.02	0.01	0.28
TALLPPE quintile			
1	<0.001	0.03	0.07
2	<0.001	0.13	0.17
3	<0.001	0.29	0.23
4	0.01	0.54	0.29
5	0.99	0.01	0.24
APSFTE quintile			
1	<0.001	0.01	0.06
2	<0.001	0.08	0.13
3	<0.001	0.23	0.17
4	0.12	0.33	0.23
5	0.87	0.35	0.41
Latent Class Probability 95% Confidence Interval Standard Error	0.51 0.50 – 0.52 0.01	0.31 0.30 – 0.32 0.01	0.18 0.17 – 0.19 0.01
Descriptive statistics by latent class			
Mean Index of Widening Participation Status	3.14 (s.d.= 0.64) min = 2.4 , max = 5.7	3.95 (s.d.= 0.47) min = 2.6, max = 6.1	2.00 (s.d.= 0.63) min = 2.9, max = 6.9

Table 15: Model 3 goodness-of-fit statistics for latent class analysis models with one through five classes, Medicine With a Gateway Year sample (n=630).

Number of Classes	Degrees of Freedom	G^2 (L^2)	p-value	AIC	BIC
1	387	605.74	<0.001	6437.67	6496.98
2	374	399.03	>0.05	6256.95	6368.01
3	361	272.28	>0.05	6276.20	6385.02
4	348	224.56	>0.05	6298.48	6399.72
5	337	207.56	>0.05	6137.49	6408.48
AIC = Aikake information criterion statistic; BIC = Bayesian information criterion statistic.					

Table 16: Outline of conditional and latent class probabilities for the two-class model Medicine With a Gateway Year sample (n=630).

Conditional Probabilities		
Two-class model	Class I *'Non-WP students'*	Class II 'WP students'
State funded school	0.96	0.97
Parent no HE	0.35	0.91
BURSARY	0.14	0.38
LOWERSOC	0.15	0.92
POLAR quintile		
1	0.08	0.10
2	0.11	0.16
3	0.27	0.30
4	0.24	0.23
5	0.30	0.21
IMD quintile		
1	0.15	0.03
2	0.19	0.10
3	0.18	0.20
4	0.24	0.30
5	0.24	0.37
Latent Class Probability 95% Confidence Interval Standard Error	0.42 0.32 – 0.52 0.05	0.58 0.48 – 0.68 0.05
Descriptive statistics by latent class		
Mean Index of Widening Participation Status	2.99 (s.d.= 0.64) min = 1.2, max = 4.6	3.93 (s.d.= 0.64) max = 2.3 - 5

Table 17: Summary of proportions of missing values for multiple imputation (n=40190)

Variable	Non-missing	Missing	% Missing	Unique values	Minimum	Maximum
POLAR quintile	40105	85	2.09	5	1	5
IMD quintile	36670	3520	8.75	5	1	5
SCHOOL TYPE	39670	520	1.29	2	1	0
PARED	35230	4960	12.34	2	1	0
BURSARY	40190	0	0.00	2	1	0
LOWERSOC	38805	1385	3.38	2	1	0

Table 18: Summary of patterns missing values for multiple imputation (n=40190)

Percent	POLAR quintile	IMD quintile	School Type	PARED	Lowersoc
76%	1	1	1	1	1
11	1	1	1	0	1
8	1	0	1	1	1
2	1	1	1	1	0
<1	1	1	1	0	0
<1	1	1	0	1	1
<1	1	0	1	0	1
<1	1	0	1	1	0
<1	1	1	0	0	1
<1	1	0	0	1	1
<1	0	0	1	1	1
<1	0	1	1	1	1
<1	0	0	0	1	1
<1	1	1	0	1	0
<1	1	0	1	0	0
<1	1	1	0	0	0
<1	0	0	1	0	1
<1	0	0	1	1	0
<1	0	0	0	1	0
<1	1	0	0	0	1
<1	0	0	0	0	0
<1	0	0	0	0	1
<1	0	0	1	0	0
<1	0	1	1	0	0
<1	0	1	1	1	0
<1	0	1	1	0	1
<1	1	0	0	1	0
100%					

Table 19: Results of univariate binary logistic regression models of the outcome missing on a contextual indicator and the auxiliary variables Bursary and BME (n=40190).

Auxiliary variable	Contextual Indicator	Chi-square	p-value
BME	POLAR quintile (1= missing, 0= non-missing)	87.20	<0.05
BME	IMD quintile (1= missing, 0= non-missing)	471.20	<0.001
BME	School Type (1= missing, 0= non-missing)	26.16	<0.001
BME	PARED (1= missing, 0= non-missing)	106.12	<0.001
BME	LOWERSOC (1= missing, 0= non-missing)	245.56	<0.001
BURSARY	IMD quintile (1= missing, 0= non-missing)	34.90	<0.001
BURSARY	SCHOOL TYPE (1= missing, 0= non-missing)	14.11	<0.05
BURSARY	LOWERSOC (1= missing, 0= non-missing)	177.00	<0.001

Table 20: Results of the multivariate logistic regression of the outcome lowersoc using (1) complete case data set of the contextual indicators and (2) the final imputed data set of the contextual indicators.

	Complete Case (n=30594)			MICE (n=40190)		
Predictor	Coef.	Std. Err.	P>z	Coef.	Std. Err.	P>z
POLAR quintile coded 1-5	-0.0275	0.0143	<0.05	-0.0314	0.0122	<0.05
IMD quintile coded 1-5	0.1863	0.0139	<0.001	0.1822	0.0119	<0.001
SCHOOL TYPE coded 1 -0	0.3813	0.0363	<0.001	0.3817	0.0304	<0.001
PARED coded 1-0	1.9534	0.0331	<0.001	1.7941	0.0285	<0.001
BURSARY coded 1-0	1.0071	0.0628	<0.001	1.0278	0.0528	<0.001
Wald Tests						
POLAR quintile coded 1-5	Chi2(1) =3.70, p<0.05			Chi2(1) = 6.61, p<0.001		
IMD quintile coded 1-5	Chi2(1) = 177.33, p<0.001			Chi2(1) = 235.00, p<0.001		
SCHOOL TYPE coded 1 -0	Chi2(1) = 110.30, p<0.001			Chi2(1) = 156.72, p<0.001		
PARED coded 1-0	Chi2(1) = 3472.42, p<0.001			Chi2(1) = 3421.25, p<0.001		
BURSARY coded 1-0	Chi2(1) = 257.02, p<0.001			Chi2(1) = 377.80, p<0.001		

Table 21: Results of the cross-tabulation of imputed IWPS scores and imputed values on lowersoc (NS-SEC 3-7 versus NS-SEC 1-2).

Cut-score	NS-SEC 1-2	NS-SEC 3-7	Total
>=4.5 Range 1.2 - 5.0	n=150 17.36%	n=715 82.64%	865

Table 22: Report of sensitivity and specificity for the imputed values derived IWPS score.

Cutpoint	Sensitivity	Specificity	Classified	LR+	LR-
(>= 1.2)	100	0	0.2407	1	
(>= 1.3)	0.9571	0.1128	0.316	1.0787	0.3805
(>= 1.4)	0.9242	0.1906	0.3672	1.1418	0.3976
(>= 1.4)	0.9175	0.2089	0.3795	1.1598	0.3949
(>= 1.5)	0.9031	0.2387	0.3986	1.1863	0.4058
(>= 1.6)	0.8877	0.2732	0.4211	1.2214	0.411
(>= 1.6)	0.8827	0.2809	0.4257	1.2274	0.4177
(>= 1.7)	0.878	0.2886	0.4305	1.2342	0.4227
(>= 1.8)	0.8725	0.3	0.4378	1.2465	0.4249
(>= 1.9)	0.8695	0.3061	0.4417	1.2531	0.4262
(>= 2)	86.73	30.91	44.34	1.2552	0.4295
(>= 2.2)	0.8664	0.3102	0.4441	1.2561	0.4306
(>= 2.3)	0.777	0.4873	0.557	1.5155	0.4576
(>= 2.4)	0.6981	0.6176	0.6369	1.8254	0.4888
(>= 2.5)	0.6229	0.7156	0.6933	2.1904	0.527
(>= 2.6)	0.5705	0.786	0.7341	2.6655	0.5465
(>= 2.7)	0.5309	0.8294	0.7576	3.1118	0.5656
(>= 2.8)	0.5012	0.8584	0.7724	3.5396	0.5811
(>= 2.9)	0.4771	0.8784	0.7818	3.9234	0.5953
(>= 3)	46.49	88.91	78.7	4.1913	0.6019
(>= 3.2)	0.4596	0.8941	0.7895	4.3412	0.6044
(>= 3.3)	0.4186	0.9125	0.7936	4.7825	0.6372
(>= 3.4)	0.3707	0.9291	0.7947	5.2304	0.6773
(>= 3.5)	0.3179	0.9463	0.795	5.919	0.7208
(>= 3.6)	0.2658	0.9604	0.7932	6.709	0.7645
(>= 3.7)	0.222	0.9707	0.7905	7.5681	0.8015
(>= 3.8)	0.1801	0.9788	0.7866	8.5074	0.8376
(>= 3.9)	0.1387	0.986	0.7821	9.9386	0.8735
(>= 4)	10.91	99.05	77.84	11.5169	0.8995
(>= 4.2)	0.096	0.9924	0.7767	12.6877	0.9109
(>= 4.3)	0.0912	0.9935	0.7763	13.9829	0.9148
(>= 4.4)	0.0838	0.9942	0.7751	14.4552	0.9215
(>= 4.5)	0.0738	0.9951	0.7734	15.0171	0.9308
(>= 4.6)	0.0631	0.9959	0.7714	15.3958	0.9408
(>= 4.7)	0.0496	0.9969	0.7689	15.9404	0.9533
(>= 4.8)	0.0347	0.9977	0.766	15.3627	0.9675
(>= 4.9)	0.0187	0.9988	0.7629	15.4339	0.9825
(>= 5)	0.57	99.95	76.03	11.5689	0.9948

9 Figures

Figure 1: POLAR quintile by socioeconomic class, UK domiciled, non-graduate entrants to Standard Entry programmes at UK medical schools, 2008-2015, (n=38,730)

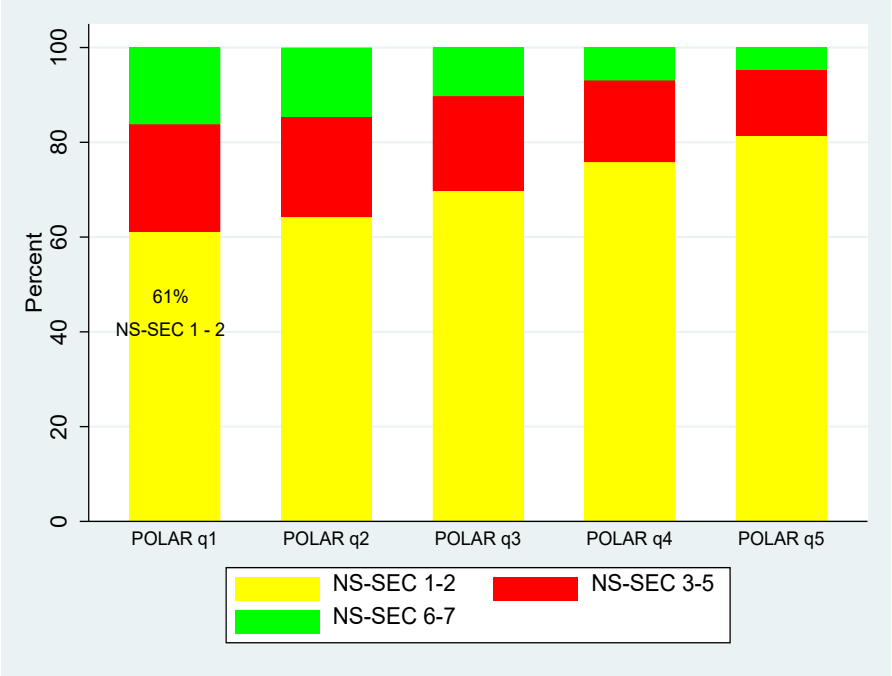


Figure 2: IMD quintile by socioeconomic class, UK domiciled, non-graduate entrants to Standard Entry programmes at UK medical schools, 2008-2015, (n=35,365).

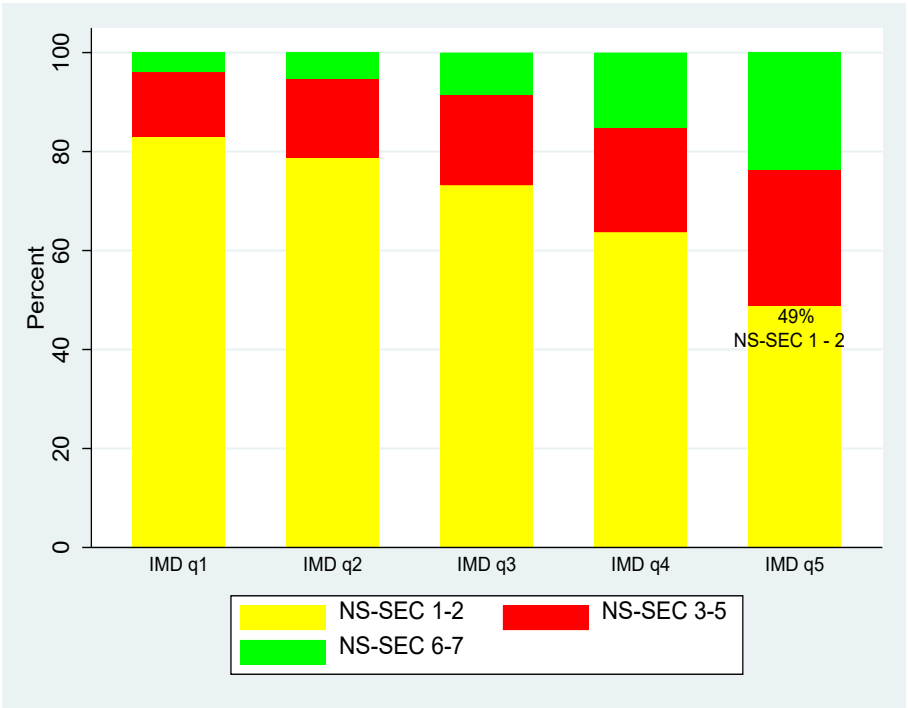


Figure 3: Type of secondary school attended by socioeconomic class, UK domiciled, non-graduate entrants to Standard Entry programmes at UK medical schools, 2008-2015 (n= 38,310).

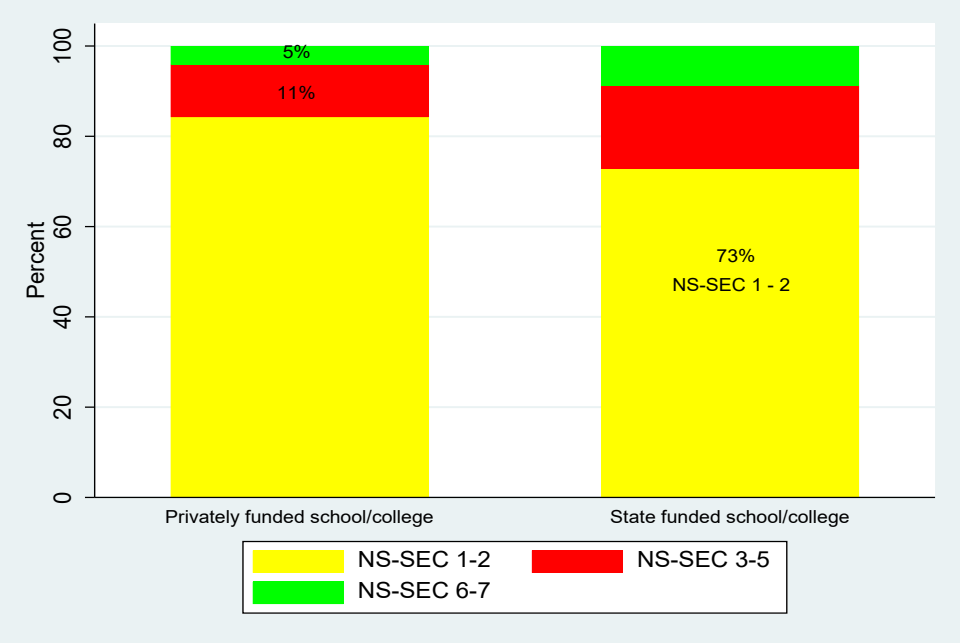


Figure 4: Whether a student's parents had higher education qualifications by socioeconomic class, UK domiciled, non-graduate entrants to Standard Entry programmes at UK medical schools, 2008-2015 (34,245).

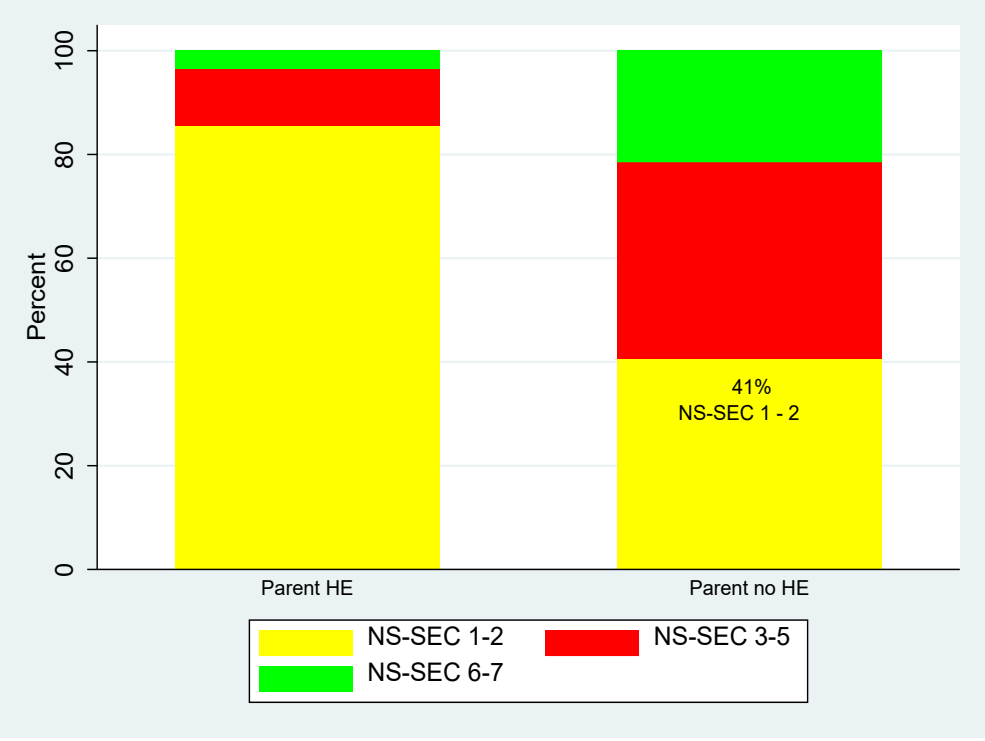


Figure 5: Receipt of a bursary by socioeconomic class, UK domiciled, non-graduate entrants to Standard Entry programmes at UK medical schools, 2008-2015 (38,805).

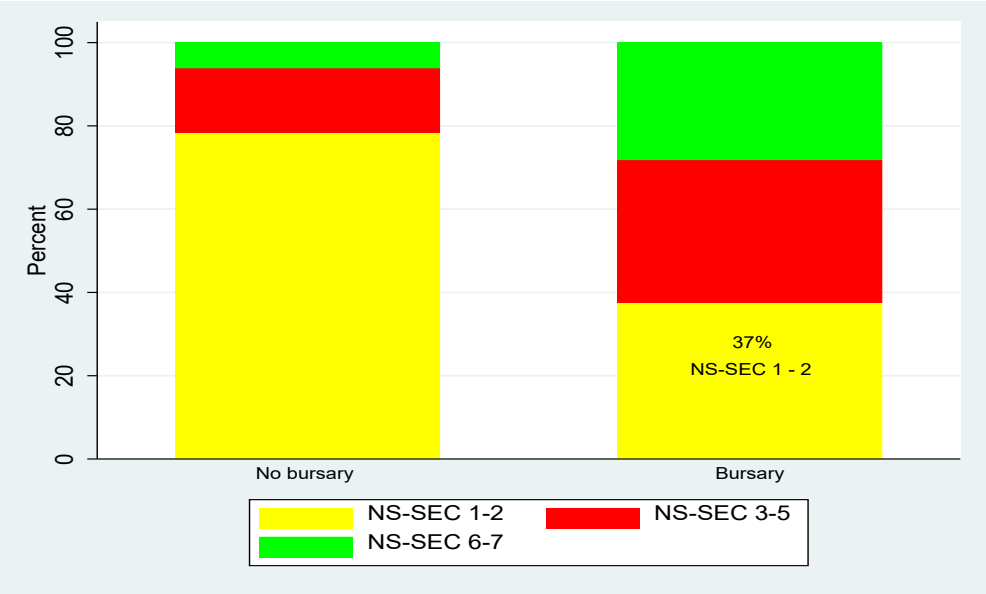


Figure 6: POLAR quintile by IMD quintile, UK domiciled, non-graduate entrants to Standard Entry programmes at UK medical schools, 2008-2015 (n=36,650). Spearman's rho r_s = -0.4300, p<0.001.

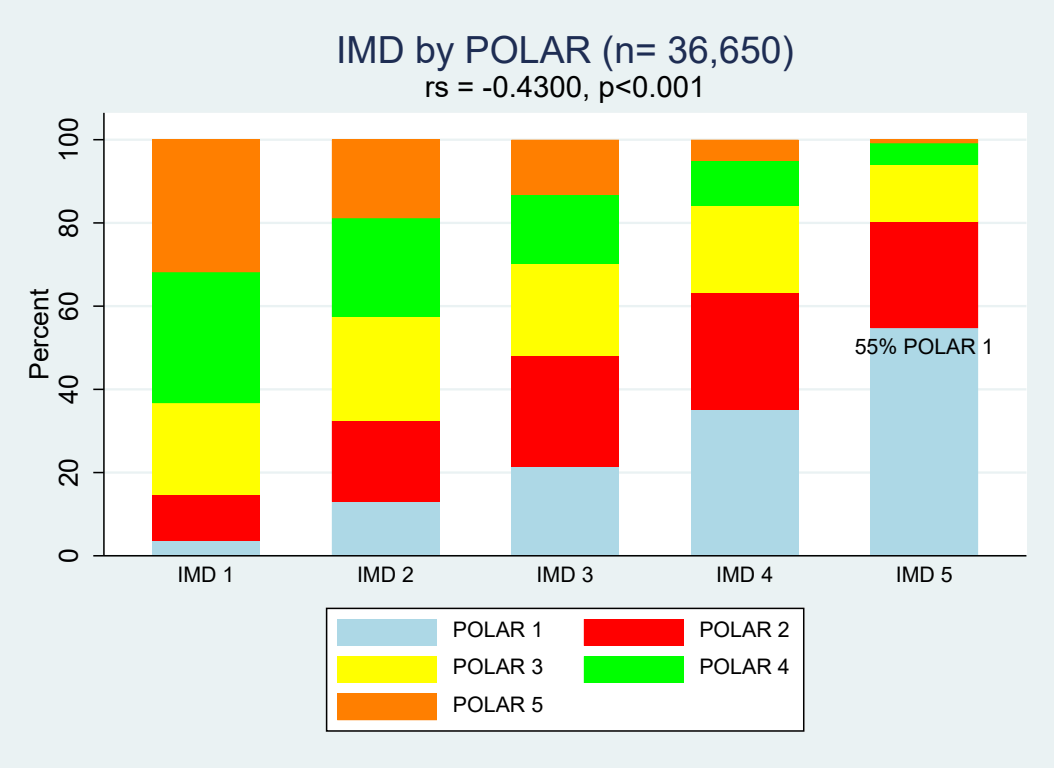


Figure 7: Graphically illustrated predicted probability of the outcome NS-SEC 3-7 versus NS-SEC 1 - 2 adjusted by scores on the UK Widening Participation Index.

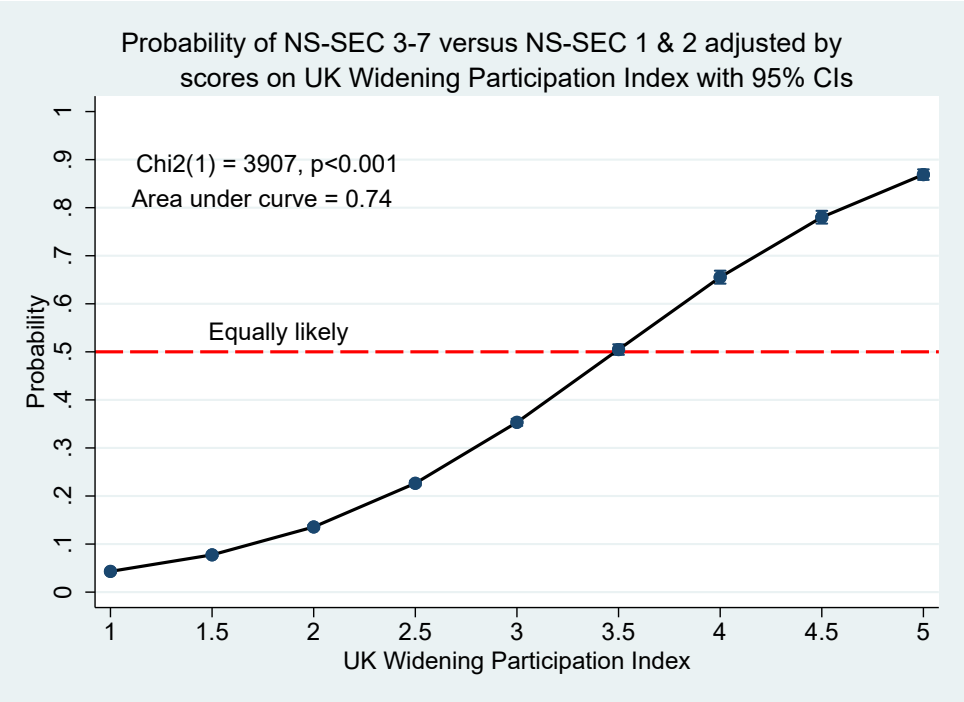


Figure 8: Receiver Operating Characteristic (ROC) curve UK sample.

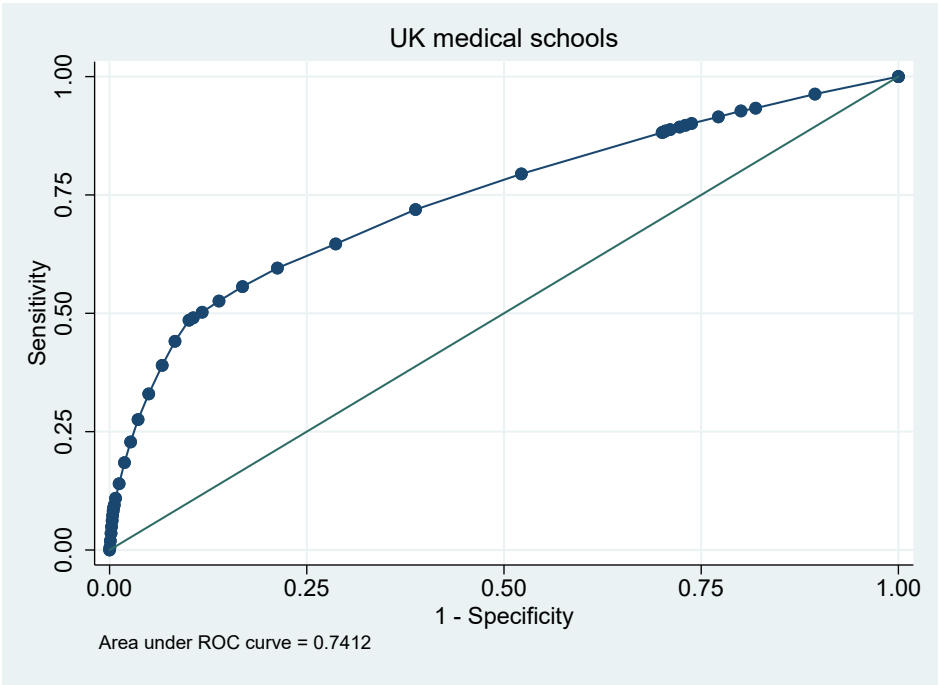


Figure 9: Quintile APSFTE_ALEVA (average point score per A-level student at school attended at time of taking the examination) by socioeconomic class, England domiciled, non-graduate entrants to Standard Entry programmes at UK medical schools, 2008-2014

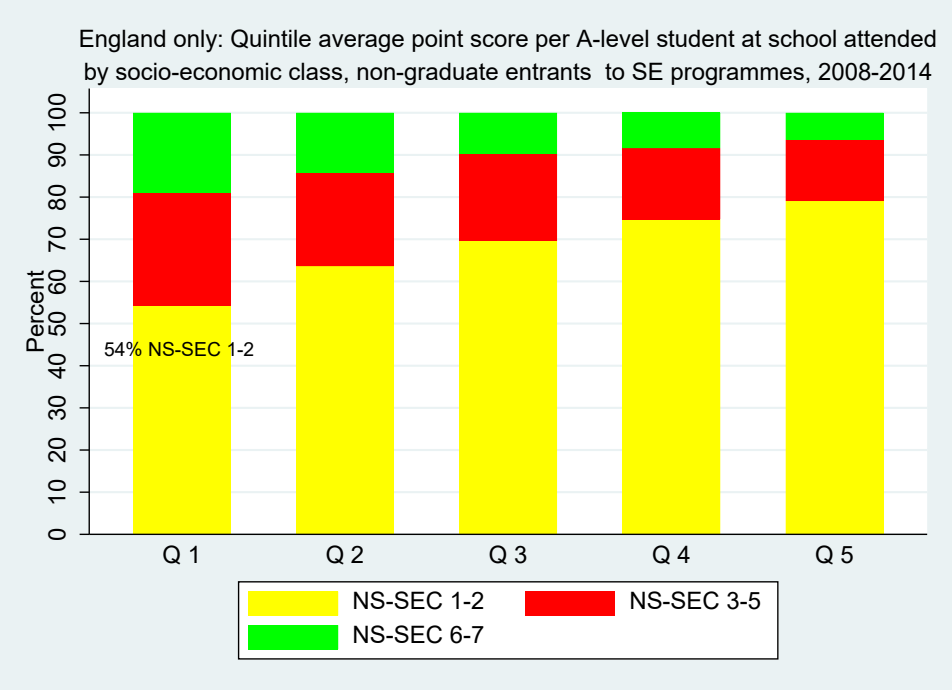


Figure10: Quintile TALLPPE_ALEVA (average point score per A-level entry at school attended at time of taking the examination) by socioeconomic class, England domiciled, non-graduate entrants to Standard Entry programmes at UK medical schools, 2008-2014.

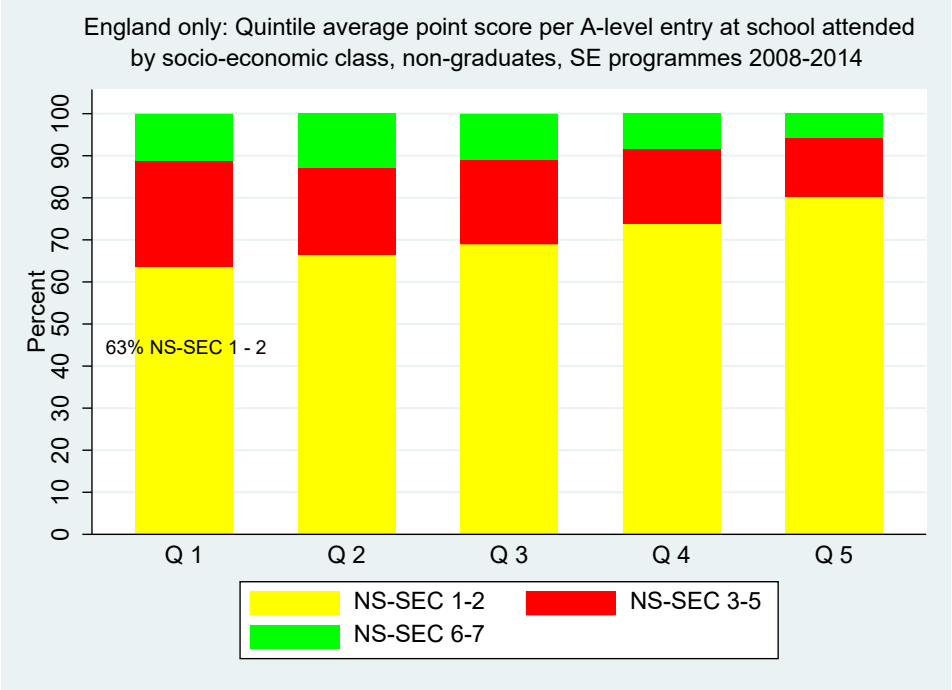


Figure 11: Graphically illustrated predicted probability of the outcome NS-SEC 3-7 versus NS-SEC 1 - 2 adjusted by scores on the England Widening Participation Index.

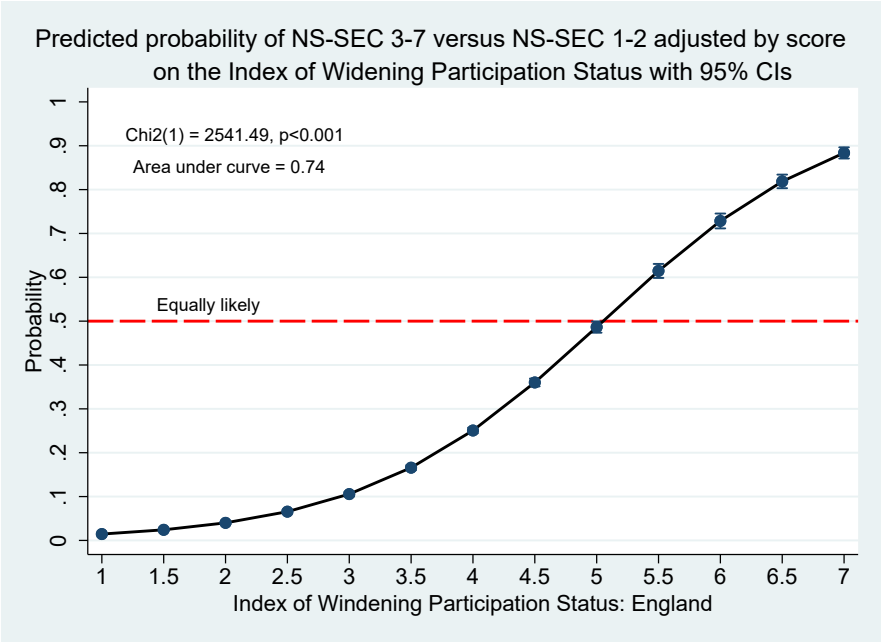


Figure 12: Receiver Operating Characteristic (ROC) curve England sample.

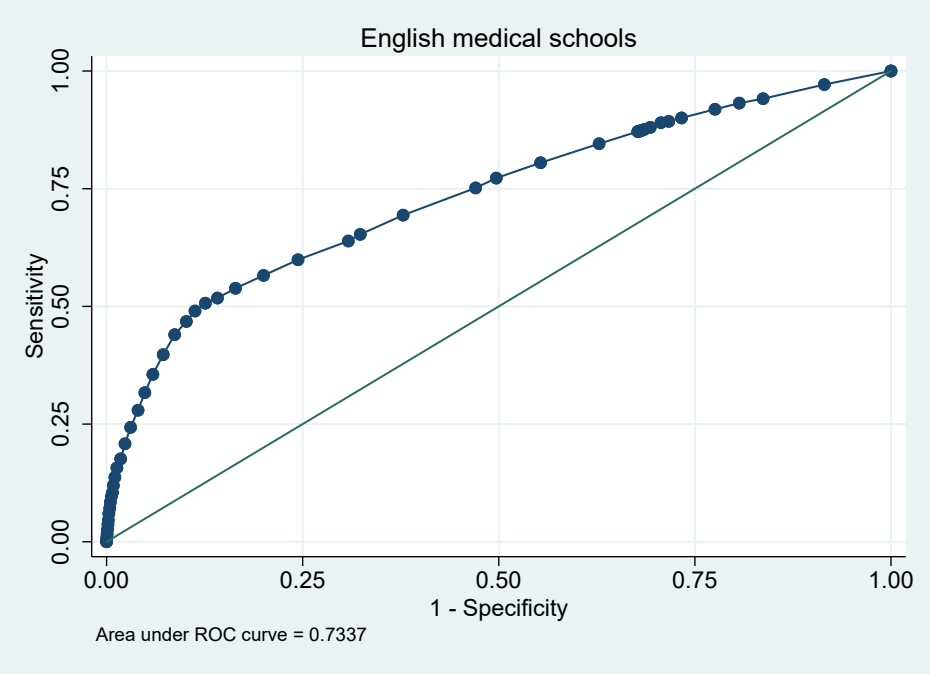


Figure 13: Graphically illustrated predicted probability of the outcome NS-SEC 3-7 versus NS-SEC 1 - 2 adjusted by scores on the Medicine With a Gateway Year Widening Participation Index.

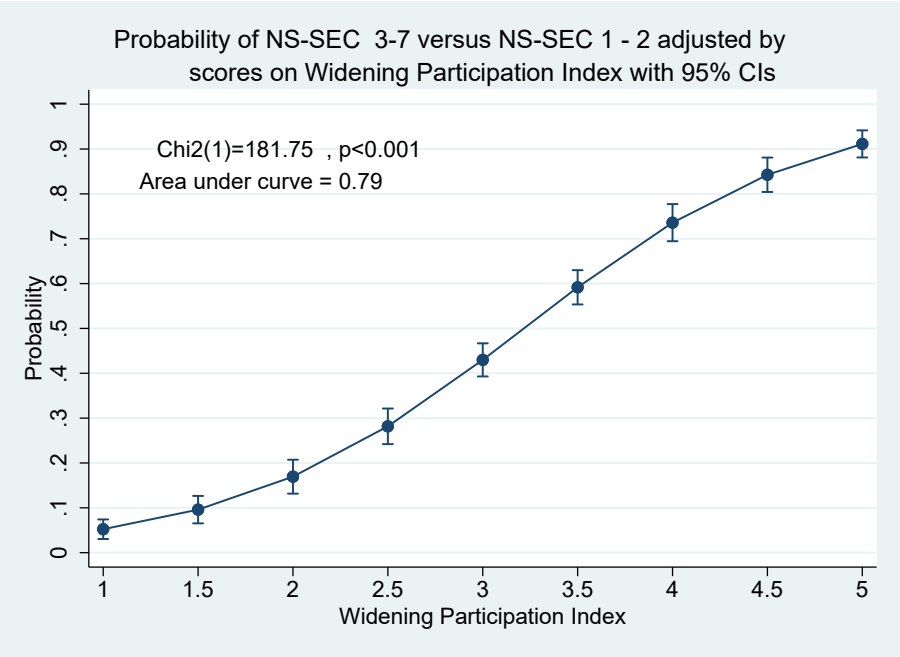


Figure 14: Receiver Operating Characteristic (ROC) curve Medicine With a Gateway Year sample.

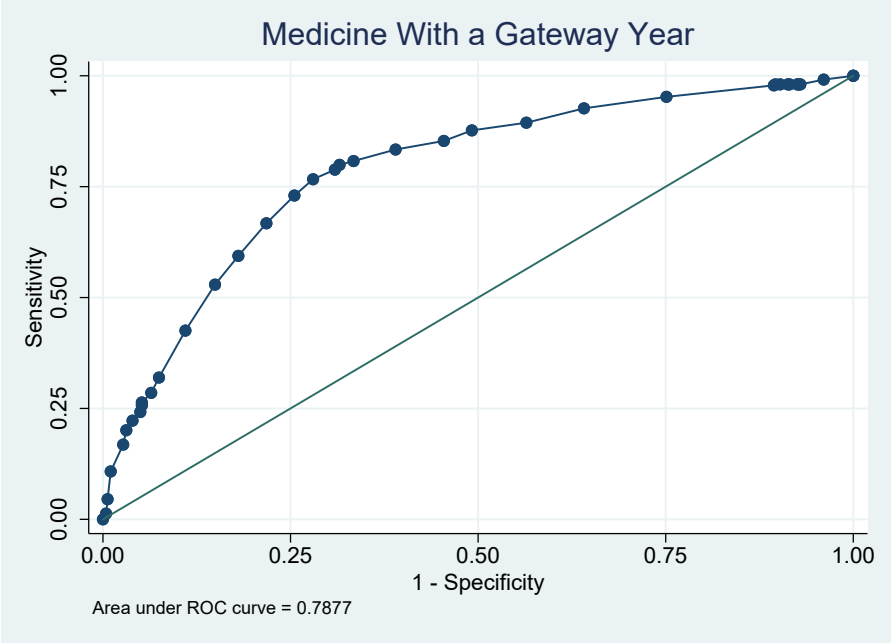


Figure 15: Graphically illustrated predicted probability of the outcome NS-SEC 3-7 versus NS-SEC 1 - 2 adjusted by scores derived from the imputed data set (n=40190).

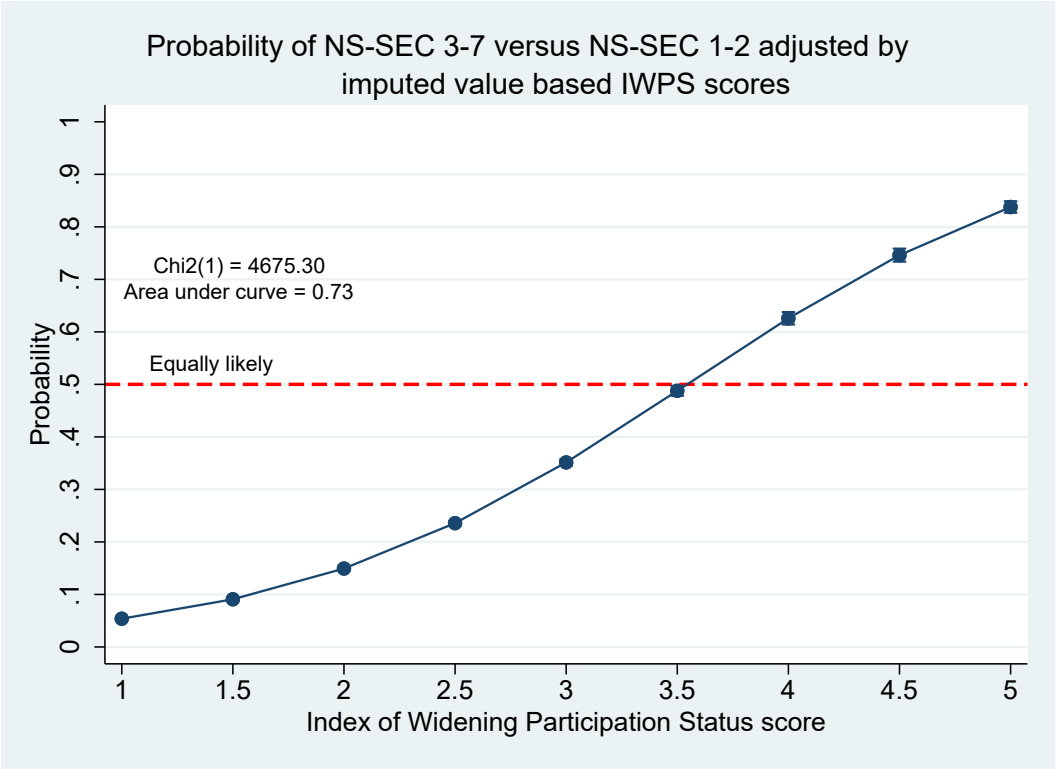
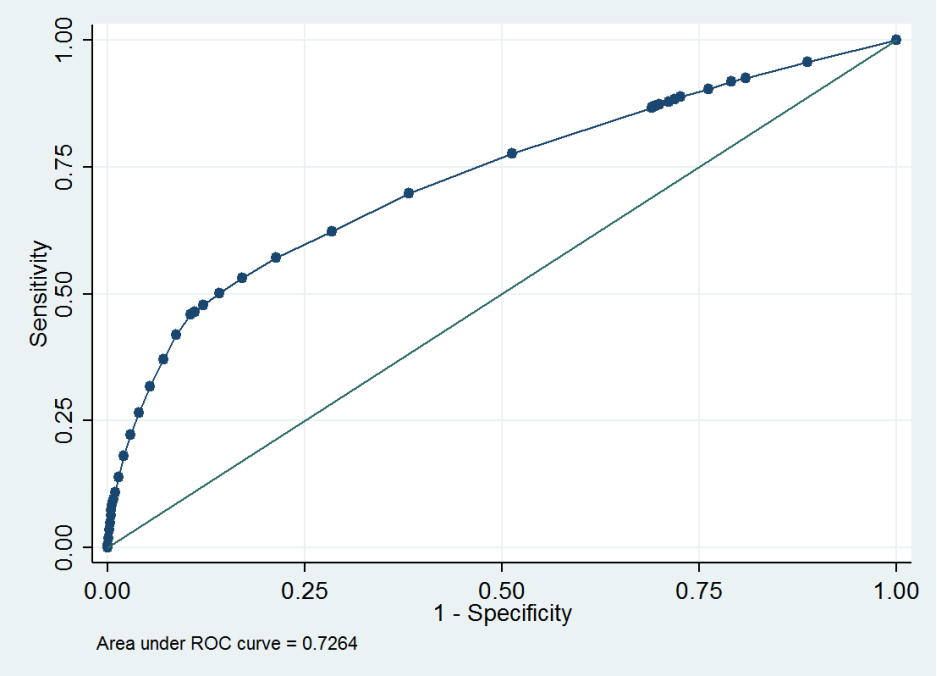


Figure 16: Receiver Operating Characteristic (ROC) curve of predicted probability of the outcome NS-SEC 3-7 versus NS-SEC 1 - 2 adjusted by scores derived from the imputed data set (n=40190).



10 References

1. Shah M, McKay J (eds). *Achieving Equity and Quality in Higher Education: Global Perspectives in an Era of Widening Participation.* Palgrave Macmillan, 2018.

2. Fair Access to Professional Careers: A progress report by the Independent Reviewer on Social Mobility and Child Poverty. Available at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/61090/IR_FairAccess_acc2.pdf

3. Steven K, Dowell J, Jackson C, Guthrie B. Fair access to medicine? Retrospective analysis of UK medical schools application data 2009-2012 using three measures of socioeconomic status. *BMC Med Educ*. 2016; 16(1):11

4. Higher Education Statistics Agency, Student Statistics. Available at www.hesa.ac.uk

5. Connelly R, Gayle V, Lambert P. A review of occupation-based social classification for social survey research. *Methodological Innovations*. 2016; 9:1-14.

6. Bergman M, Joye D. *Comparing Social Stratification Schemata: CAMPSIS, CSP-CH, Goldthorpe, ISCO-88, Treiman, and Wright*. Cambridge Studies in Social Research, No.10. SSRP Publications, 2005.

7. Gorard S, Boliver V, Siddiqui N, Banerjee P. Which are the most suitable contextual indicators for use in widening participation to HE? *Research Papers in Education*. Available at https://doi.org/10.1080/02671522.2017.1402083

8. Medical Schools Council. Selecting for Excellence Final Report. Medical Schools Council. 2014. Available at https://www.medschools.ac.uk/media/1203/selecting-for-excellence-final-report.pdf

9. Medical Schools Council Selection Alliance. Indicators of good practice in contextual admissions. Medical Schools Council. 2018. Available at https://www.medschools.ac.uk/media/2413/good-practice-in-contextual-admissions.pdf

10. Medical Schools Council Selection Alliance. Entry requirements for UK medical schools - 2018 entry. Medical Schools Council 2017. Available at https://www.medschools.ac.uk/media/2357/msc-entry-requirements-for-uk-medical-schools.pdf

11. Cleland J, Nicholson S, Patterson F, Thomas L, Wilde K. The use of contextual data in medical school selection processes: A mixed-method programme of research. 2016. Unpublished Report to the Medical Schools Council. 2016.

12. Boliver V, Crawford C, Powell M, Craige W, *Admissions in Context: The use of contextual information by leading universities*. The Sutton Trust. 2017.

13. Gorard S, Boliver V, Siddiqui N, Banerjee P. Which are the most suitable contextual indicators for use in widening participation to HE? Research Papers in Education, DOI:

10.1080/02671522.2017.1402083. Available at https://www.tandfonline.com/doi/abs/10.1080/02671522.2017.1402083

14. Moore J, Mountford-Zimdars A, Wiggans J. Contextualised Admissions: Examining the Evidence. Cheltenham: Supporting Professionalism in Admissions Programme. The National Audit Office. 2008. Widening Participation in Higher Education in the United Kingdom. NAO Report Available at www.nao.org.uk

15. Garrud P. *Help and hindrance in widening participation: commissioned research report.* Medical Schools Council. 2014. Available at https://www.medschools.ac.uk/media/2446/selecting-for-excellence-research-dr-paul-garrud.pdf

16. Boliver V., Gorard S., Siddiqui N., Will the use of Contextual Indicators Make UK Higher Education Admissions Fairer? *Educational Sciences*. 2015; 5(4):306-22.

17. Bridger K, Shaw J, Moore J. *Fair Admissions to Higher Education: Research to Describe the Use of Contextual Data in Admissions at a Sample of Universities and Colleges in the UK*. Supporting Professionalism in Admissions (SPA). Cheltenham,UK,2015.

18. Dowell J, The UK Medical Education Database: What is it? Why and how might I use it? *BMC Med Educ*. 2018: 18:6.

19. United Kingdom Clinical Aptitude Test (UKCAT). What is the UKCAT? http://www.ukcat.ac.uk/about-the-test/what-is-the-ukcat/

20. Cleland J, Dowell J, McLachlan J, Nicholson S, Patterson F. Identifying best practice in the selection of medical students (literature review and interview survey). General Medical Council. 2013. Available at https://www.gmc-uk.org/about/what-we-do-and-why/data-and-research/research-and-insight-archive/identifying-best-practice-in-the-selection-of-medical-students

21. Goldthorpe J. Class analysis and the reorientation of class theory: the case of persisting differentials in educational attainment. *The British Journal of Sociology* 2010; 61(s1): 311-35.

22. Goldthorpe J. Social class mobility in modern Britain: changing structure, constant process. *Journal of the British Academy*, 2016; 4:89-111.

23. Goldthorpe J, McKnight A. (2006) 'The Economic Basis of Social Class', in S Morgan, D Grusky and G Fields (eds), Mobility and Inequality: Frontiers of Research from Sociology and Economics (Stanford CA, Stanford University Press) 109-36.

24. Hosmer D, Lemeshow S, Sturdivant R. *Applied Logistic Regression*. 3rd ed. Hoboken, New Jersey: Wiley, 2013.

25. Long J, Freese J. *Regression models for categorical dependent variables using Stata*. 3rd ed. College Station, Texas: Stata Corp LP, 2014.

26. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J. Clin. Epidemiol.* 1996;49(12):1373-79.

27. Zhou X, Obuchowski N, McClish D. *Statistical methods in diagnostic medicine*. New York: Wiley, 2002.

28. Petrie A, Sabin C. *Medical Statistics*, (2007) Third Edition. Oxford: Wiley-Blackwell, 115-117.

29. Hagenaars J, McCutcheon A, 2002, (Eds). Applied Latent Class Analysis. Cambridge: Cambridge University Press.

30. McCutcheon A, (1987). Latent Class Analysis. (Sage University Paper series on Quantitative Applications in the Social Sciences, No 07-064). Newbury Park, CA: Sage

31. Lanza S. Cooper B. Latent Class Analysis for Development Research. Child Development Perspectives, Volume 10, Number 1, 2016, 59-64.

32. Rubin D. Multiple Imputation for Nonresponse in Surveys. New York: Wiley 1987.

33. Pampaka M, Hutcheson G, Williams J. Handling missing data: analysis of a challenging data set using multiple imputation. International Journal of Research and Education, 2014, 39,1:19-37

34. Mackinnon A. The use and reporting of multiple imputation in medical research; a review. Journal of Internal Medicine. Volume 268, Issue 6. December 2010, 586-593.

35. Little R, Rubin D. (2002) Statistical analysis with missing data (2nd ed.). New York: John Wiley.

36. White I., Royston P, Wood A. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med, 2011; 30(4):377-99.

37. Rezvan P. Lee K. Simpson J. BMC Medical Research Methodology 2015, 15:30. DOI 10.1186/s12874-015-0022-1.

38. Schuwirth L, Van Der Vleuten C. (2010) How to design a useful test: the principles of assessment. In Swanick T, editor, Understanding Medical Education. Oxford;Wiley, 195-207.

39. Office for National Statistics; National Records of Scotland; Northern Ireland Statistics and Research Agency (2016): 2011 Census aggregate data. UK Data Service (Edition: June 2016). DOI: http://dx.doi.org/10.5257/census/aggregate-2011-1

40. Seyan K, Greenhalg T, Dorling D. The standardised admissions ratio for measuring widening participation in medical schools: analysis of UK medical school admissions by ethnicity, socioeconomic status, and sex. *BMJ*. 2004; 328(7455): 1545-1546.