Considering evidence for ethnicity bias using assessment case scenarios and medical student correctness and certainty

Mike Tweed, Gordon Purdie, Cameron Lacey

ABSTRACT

AIMS: There is inequitable distribution of health risks, exposures, resources and outcomes by ethnicity. This may be contributed to by health professional bias. The aim of this study was to investigate the relationship between ethnicity of patients, within written assessment case scenarios, and medical students’ response correctness and certainty.

METHODS: Otago Medical School students sit a 150 MCQ progress test with item level response certainty. Patient ethnicity for 60 MCQ case scenarios was varied between two ethnic groups (New Zealand European, Māori) and none specified. Analysis of responses by patient ethnicity was undertaken to compare: odds of correctness; level of certainty; correctness for level of certainty and also by year groups and ability.

RESULTS: One thousand one hundred and three students sat the test. There was no significant difference in odds of correctness or level of certainty by the ethnicity of the patient case scenario. These did not differ significantly by year groups or ability groups, or for correctness by level of certainty.

CONCLUSIONS: No systematic differences in correctness or certainty of student responses to case scenarios by patient ethnicity were detected. Further exploration is warranted, including incorporating more ethnicity descriptors, analysis of incorrect answers, analyses for patterns responses over time by individual students, and selecting questions where varying patient ethnicity is expected to alter the correct response or difficulty.

When assessing medical students, assessment content should maximise authenticity, with realistic clinical problems. To provide better validation evidence for results, written assessments of healthcare students are often based on clinical scenarios. Threats to validity include construct under-representation. Therefore, patients’ case scenarios/vignettes should include demographics, with assessment patient demographics matching the healthcare population. When ethnicity, or other demographics, are only included in the case scenario when the correct response or item difficulty is dependent on the demographic, this may serve to cue test-wise students, so reducing validation evidence.

In a research context, New Zealand medical students have demonstrated implicit and explicit ethnic bias favouring New Zealand Europeans. Medical students may demonstrate preferences for some patient demographic groups, and simply including ethnicity in a case description can influence how final-year medical students talked about patients. This variation in response, related to case scenario ethnicity, might indicate a bias related to the assessment or a bias related to the candidates and warrants further exploration.

Doctors’ practice may be affected by a variety of biases, with patient demographics influencing certainty of diagnosis.
Therefore, there is the possibility that the correctness of a clinical decision, and also the certainty in that decision, might be influenced by perception of the patient demographics. Self-monitoring, as reflection-in-action of decision certainty, is authentic to daily practice.\(^\text{17,18}\) Therefore, when assessing medical students there is potential value in considering their demonstration of certainty in responses, stratified by self-monitoring descriptors, to questions related to decision-making situations, and this has proved possible in a variety of assessments.\(^\text{19–22}\) Building on this, it is possible to explore how the ethnicity of a patient, as part of an assessment case scenario, will affect the odds of medical students making correct responses, their certainty in those responses and the odds of correctness for certainty. The aim of this study is to investigate the relationship between ethnicity of patients, within written assessment scenarios, and the responses of medical students.

**Method**

The Retained Knowledge Test (RKT) is a progress test sat by all year 2–5 medical students at Otago Medical School. The RKT is computer-delivered twice per year and consists of 150 MCQ questions in random order. These tests are not answered under examination conditions, with the students having a two-week window to respond but are encouraged to answer in a single sitting if possible. Also the RKT includes student certainty in their responses, with certainty being categorised as no, low, moderate or high, based on the students’ perception of the need for additional resources, should this question be an actual clinical situation.\(^\text{21}\)

The current standard policy is that all case scenarios that relate to an individual patient should include their ethnicity and that the ethnicities across the test as a whole should approximate to the New Zealand census population.

To investigate the relationship between the patient ethnicity of the scenario and correctness, certainty and the relationship between them, six versions of the RKT were created and students were randomised to which version they received (Table 1).

The questions chosen to have variable patient ethnicity were reviewed by clinicians (CL, MT) to ensure that the change in ethnicity did not affect the correct answer and/or the perceived difficulty of the question.

The standard questions were questions where case scenario ethnicity was not included as the question was not about an individual patient, was vital to the question and could not be varied or was not vital but included as per current policy.

**Analysis**

Analysis explored whether correctness and certainty differed by case scenario ethnicity, student year or student total number responses correct.

To estimate the odds ratios of a correct answer between levels of certainty, year cohort and meeting the standard, among all 150 questions a mixed model logistic regression was used with terms for level of certainty, student year, meeting the standard for the year cohort, question group, question and a random term for student. Level of certainty interaction terms with student year and standard were added to the model.

**Table 1: Patient ethnicity across six versions of the RKT.**

<table>
<thead>
<tr>
<th>RKT version</th>
<th>1–90</th>
<th>91–110</th>
<th>111–130</th>
<th>131–150</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Standard</td>
<td>Standard</td>
<td>Māori</td>
<td>NZ Euro</td>
<td>None</td>
</tr>
<tr>
<td>2 Standard</td>
<td>Standard</td>
<td>Māori</td>
<td>None</td>
<td>NZ Euro</td>
</tr>
<tr>
<td>3 Standard</td>
<td>Standard</td>
<td>NZ Euro</td>
<td>None</td>
<td>Māori</td>
</tr>
<tr>
<td>4 Standard</td>
<td>Standard</td>
<td>NZ Euro</td>
<td>Māori</td>
<td>None</td>
</tr>
<tr>
<td>5 Standard</td>
<td>Standard</td>
<td>None</td>
<td>Māori</td>
<td>NZ Euro</td>
</tr>
<tr>
<td>6 Standard</td>
<td>Standard</td>
<td>None</td>
<td>NZ Euro</td>
<td>Māori</td>
</tr>
</tbody>
</table>
to test for differences in levels of certainty odds ratios by student year and standard. To estimate the odds ratios of a correct answer between ethnicity of case scenarios, among non-standard questions, a mixed model logistic regression was used with terms for ethnicity of case scenario, student year, question group, question and random term for student. Level of certainty, meeting the standard, and interaction terms of ethnicity of case scenario with level of certainty, student year, standard and question were added to the model to test for differences in case scenario ethnicity odds ratios by level of certainty, student year, standard and question. For the odds ratios of levels of certainty between ethnicity of case scenarios a model was used with the same terms and multinomial distribution and cumulative logit link was used. Meeting the standard and interaction terms of ethnicity of case scenario with student year, standard and question were added to the model to test for differences in case scenario ethnicity odds ratios by level of certainty, student year, standard and question. When there was a significant interaction, tests for differences between ethnicity of case scenarios for each level of the other term in the interaction were adjusted for multiple comparisons with the Holm-Bonferroni method. The glimmix procedure of SAS 9.4 (SAS Institute Inc., Cary, North Carolina, US) was used, with between-within degrees of freedom.

Ethics approval
University of Otago Ethics committee approval was granted (reference number D17/126).

Results
One thousand one hundred and three students took the RKT, with number of students, correct responses and levels of certainty described in Table 2.

For all 150 questions, the odds of being correct were significantly related to level of certainty, year and meeting the standard (all p<0.0001) (Table 3). The odds ratios for level of certainty were significantly different between the year cohorts (p<0.0001) and whether the student met the standard (p<0.0001).

There were no significant differences in the odds of a correct answer between the ethnicity scenarios (p=0.84) (Table 4). These odds ratios did not differ significantly between levels of certainty (p=0.23), the year cohorts (p=0.29), whether the student met the standard (p=0.92) or question (p=0.57).

There were no significant differences in the odds of increasing certainty between the ethnicity of case scenarios (p=0.33). The odds ratio for increasing certainty between Māori and New Zealand European was 0.98 (95%CI 0.94, 1.02; p=0.23) and between no ethnicity and New Zealand European 1.00 (95%CI 0.97, 1.04; p=0.87). These odds ratios

<table>
<thead>
<tr>
<th>Year 2</th>
<th>Year 3</th>
<th>Year 4</th>
<th>Year 5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Took RKT</td>
<td>288</td>
<td>276</td>
<td>262</td>
<td>277</td>
</tr>
<tr>
<td>Above standard</td>
<td>68.8% (198)</td>
<td>78.3% (216)</td>
<td>84.4% (221)</td>
<td>91.7% (254)</td>
</tr>
<tr>
<td>Correct (%)</td>
<td>31 (25–36)</td>
<td>39 (34–44)</td>
<td>50 (43–55)</td>
<td>57 (51–64)</td>
</tr>
<tr>
<td>Certainty</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None (%)</td>
<td>10 (0–74)</td>
<td>47 (1–68)</td>
<td>25 (1–50)</td>
<td>8 (0–36)</td>
</tr>
<tr>
<td>Low (%)</td>
<td>52 (16–85)</td>
<td>24 (13–53)</td>
<td>31 (18–48)</td>
<td>31 (19–48)</td>
</tr>
<tr>
<td>Moderate (%)</td>
<td>9 (5–16)</td>
<td>14 (7–21)</td>
<td>19 (11–27)</td>
<td>23 (15–32)</td>
</tr>
<tr>
<td>High (%)</td>
<td>2 (1–5)</td>
<td>7 (2–13)</td>
<td>11 (4–20)</td>
<td>18 (8–29)</td>
</tr>
</tbody>
</table>

%percentage, median and interquartile range.
*Percent of 150 questions, †Percent of questions answered.
did not differ significantly between the year cohorts (p=0.74) or whether the student met the standard (p=0.56). The odds were significantly differed between questions (p=0.048). There were no individual questions with significant differences between ethnicity of case scenarios after adjusting for multiple comparisons.

**Discussion**

No systematic differences in correctness or certainty of student responses between case scenario patient ethnicities were detected. Given that there were increased odds of being correct with increased certainty, year group and ability which are similar to results described previously, it is likely that the students approached this RKT similarly to other assessments.

There is considerable literature on the relationships between patient ethnicity and healthcare outcomes, with many factors contributing to this. This is highlighted for outcomes for Māori compared with New Zealand European for many healthcare indicators. Robust processes are required in the development of investigations of racial/ethnic bias in assessment of medical student and healthcare professionals.

This study included three versions of each question with the ethnicity of the patient in the scenario varied, and reviewed by clinicians to ensure that no difference in responses was to be expected. The randomised design of question allocation and randomised order of question delivery, with the responses to 60 questions by 1,103 students add to the robustness of the results, which have narrow confidence intervals.

There are limitations to this study. Several factors could have an effect on responses. All questions that related to individual patients had ethnicity described, apart from where removed for the study, with three

**Table 4:** Odds ratio of a correct answer for ethnicity of case scenarios.

<table>
<thead>
<tr>
<th>Ethnicity of case scenario</th>
<th>Percent correct</th>
<th>Odds ratio (95% confidence interval)</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>NZ European</td>
<td>45.3% (9,889/21,852)</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Māori</td>
<td>45.1% (9,867/21,873)</td>
<td>0.99 (0.95, 1.03)</td>
<td>0.56</td>
</tr>
<tr>
<td>None</td>
<td>45.2% (9,886/21,871)</td>
<td>1.00 (0.95, 1.04)</td>
<td>0.85</td>
</tr>
</tbody>
</table>
versions of each of these study questions, which should minimise any effect due to cueing. All questions were reviewed by practising clinicians to ensure no expected difference in correct answer or difficulty, which should minimise any effect due to complexity. Students are already used to ethnicity descriptors, which may or may not be vital to the correct response and/or difficulty of a question, which should minimise any effect due to unfamiliarity.

Another limitation is that patient ethnicity bias by students may have led to a different incorrect response being given. The analysis here was correct/incorrect only, where all incorrect responses were treated the same. An analysis of different incorrect responses allowing for certainty would add to the analysis, but was not possible as the incorrect responses had not been stratified for safety.

The case scenario ethnicities used as demographic descriptors in this analysis were limited to Māori and New Zealand European. There is greater ethnic diversity in current day New Zealand. Bias across many ethnicities could be investigated in a study with a sufficient number of questions.

It may be that patient ethnicity bias that would be expressed in clinical practice were not necessarily detectable in this research. This was a single low-stakes test with cohort-level analysis, so inferring an absence of bias among all individual students, especially related to high-stakes assessments and subsequent practice, was not possible. It is conceivable that a small number of students may have had response patterns that may indicate a degree of bias, but that this information was lost in the cohort level analysis. Extrapolating to the high cognitive load and stress of medical practice is limited and it is possible that paper-based case scenarios will not reveal bias that is present. Variation in certainty could reflect cognitive load, and there was no difference detected at different levels of certainty or ability in this study. This study set out to investigate for the presence of bias in clinical decision-making. Bias related to the perceived ethnicity of the patient may be primarily due to other factors such as personal interactions, rather than clinical decision-making. In addition, the biases may affect health outcomes related to some diagnostic conditions more than others. However, there were no differences between case scenario ethnicities detected in this study between questions for certainty or for individual question level analysis for certainty.

This study does raise future areas for investigation. One question worth exploring would relate to differences in responses when the patient ethnicity does make a difference to correct response or difficulty, such as sore throat management. Given the important association between patient ethnicity and health both in New Zealand and internationally, there is some risk associated with unintended learning consequences of assessments using only scenarios in which patient ethnicity has no impact on correct response or difficulty. This requires exploration as it may promote students to ignore inclusion of patient ethnicity in decision making, which is not desirable for practice.

In conclusion, analysis of student responses allowing for the ethnicity of the patient in the case scenario can be used to investigate for evidence of variations in response patterns that may indicate bias based on ethnicity. In this study, no difference in correctness or certainty was detected.
Competing interests:
Nil.

Author information:
Mike Tweed, Senior Lecturer, Department of Medicine, University of Otago, Wellington; Gordon Purdie, Biostatistician, University of Otago, Wellington; Cameron Lacey, Senior Lecturer, Māori Indigenous Health Institute (MIHI), University of Otago, Christchurch.

Corresponding author:
Dr Mike Tweed, Senior Lecturer, Department of Medicine, University of Otago, PO Box 7343, Wellington 6242.
mike.tweed@otago.ac.nz

REFERENCES:


