

Accepted Manuscript

Diagnostic test guidelines based on high-quality evidence had greater rates of adherence: A meta-epidemiological study

J.W. O'Sullivan, Clinical Researcher, A. Albasri, Clinical Researcher, C. Koshiairis, Statistician, J.K. Aronson, Reader in Evidence-Based Medicine, C. Heneghan, Professor of Evidence-Based Medicine, R. Perera, Professor of Medical Statistics

PII: S0895-4356(18)30189-6

DOI: [10.1016/j.jclinepi.2018.06.013](https://doi.org/10.1016/j.jclinepi.2018.06.013)

Reference: JCE 9690

To appear in: *Journal of Clinical Epidemiology*

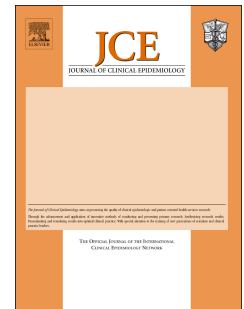
Received Date: 3 March 2018

Revised Date: 14 June 2018

Accepted Date: 28 June 2018

Please cite this article as: O'Sullivan J, Albasri A, Koshiairis C, Aronson J, Heneghan C, Perera R, Diagnostic test guidelines based on high-quality evidence had greater rates of adherence: A meta-epidemiological study, *Journal of Clinical Epidemiology* (2018), doi: 10.1016/j.jclinepi.2018.06.013.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.



Diagnostic test guidelines based on high-quality evidence had greater rates of adherence: A meta-epidemiological study

O'Sullivan JW¹, Albasri A¹, Koshiairis C¹, Aronson JK¹, Heneghan C¹, Perera R¹.

¹ Centre for Evidence-Based Medicine, Nuffield Department of Primary Care Health Sciences, University of Oxford, UK

Jack W O'Sullivan, Clinical Researcher, jack.osullivan@phc.ox.ac.uk

Ali Albasri, Clinical Researcher, ali.albasri@phc.ox.ac.uk

Constantinos Koshiairis, Statistician, constantinos.koshiairis@phc.ox.ac.uk

Jeffrey K Aronson, Reader in Evidence-Based Medicine, jeffrey.aronson@phc.ox.ac.uk

Carl Heneghan, Professor of Evidence-Based Medicine, carl.heneghan@phc.ox.ac.uk

Rafael Perera, Professor of Medical Statistics, rafael.perera@phc.ox.ac.uk

Correspondence to: Dr Jack W O'Sullivan

Centre for Evidence-Based Medicine

Nuffield Department of Primary Care Health Sciences

Radcliffe Observatory Quarter, Oxford, OX2 6GG

Phone number: +441865289300

Manuscript word count: 4,787

Abstract**Objective**

To determine the association between the quality of guidelines for diagnostic tests (both the quality and reporting and the quality of the evidence underpinning recommendations) and non-adherence.

Study design and setting

We conducted a meta-epidemiological study. We previously published a systematic review that quantified the percentage of test use that was non-adherent with guidelines. For the current study, we assessed these guidelines using the Appraisal of Guidelines for Research & Evaluation (AGREE) II tool. We then assessed the quality of evidence underpinning recommendations within these guidelines using Grading of Recommendations Assessment, Development, and Evaluation (GRADE). Linear models were then constructed to determine the association between guideline non-adherence and (a) AGREE II score and (b) GRADE score.

Results

There was no significant association between AGREE II score and non-adherent testing ($p = 0.09$). There was a significant association between GRADE score and non-adherence: recommendations based on low and very low-quality evidence had 38% ($p < 0.01$) and 24% ($p = 0.02$) more non-adherent testing, compared with recommendations based on high quality evidence.

Conclusion

Diagnostic test guideline recommendations based on high-quality evidence are adhered to more frequently.

Abstract word limit: 200

Keywords: GRADE, guidelines, AGREE II, primary care, diagnostic tests, and meta-research.

What is new?*Key Findings*

- There is much heterogeneity in the quality and reporting of diagnostic test guidelines.
- Of the guidelines we examined, most are based on poor-quality evidence.
- Guideline recommendations based on high-quality evidence have a significantly lower rate of non-adherent testing.
- There is no significant association between the quality and reporting of diagnostic test guideline recommendations and the rate of non-adherence.

What this adds to what is known

- Previous studies have shown that the quality and reporting of guidelines pertaining to treatment recommendations is varied and generally poor (using the AGREE tools) [1]. Previous studies have also shown that the quality of evidence underpinning guideline recommendations is poor [2]. We present the first assessment of the quality and reporting of diagnostic test guidelines and are the first to explore the association between guideline quality (both quality and reporting of guidelines and also quality of the evidence underpinning recommendations) and non-adherence.

Implications

- We have highlighted gaps in the literature pertaining to diagnostic tests. Future research should endeavour to produce high-quality randomised controlled trials (RCT) and diagnostic accuracy studies (DTA) to fill these gaps.
- If the evidence is available, policy makers should endeavour to support their guideline recommendations with high-quality research (such as systematic reviews, RCTs and/or DTA). Guideline recommendations based on high-quality evidence are adhered to more frequently. Guidelines developed via expert opinion or consensus, in the absence of evidence, were adhered to much less frequently.

1. Introduction

In the last 15 years, clinical practice guidelines have become increasingly common. Guidelines emerged, in the era of evidence-based medicine (EBM), to try and ensure medical decisions were based on the best available evidence.

In many countries, guidelines serve as the foundation of many performance and quality indicators [3–5]. Although, in some regard, they serve as a framework for the standard of expected medical practice [4], it is important to acknowledge that guidelines inform clinical practice, rather than dictate it. Medical decisions are complex; clinical expertise and patient values should be considered alongside guideline recommendations [6,7]. Guideline recommendations are not applicable to all patients in all clinical situations; it is likely that there will be times when doctors should depart from guidelines.

In many countries, guidelines have important medicolegal implications. Doctors can depart from guidelines in their patients' best interest, however, medical defence companies have issued explicit advice that "doctors must be prepared to explain and justify their decisions and actions, especially if they depart from guidelines issued by a nationally recognised body" [7].

Despite their importance, guidelines have been criticized for their varying quality and reporting [8,9], authors' conflicts of interests [10] and poor-quality evidence supporting their recommendations [11]. Previous research has suggested there is marked variation in how often guidelines are followed [12], but there is a paucity of research exploring the association between guideline quality and adherence, particularly for diagnostic tests. No study has examined the quality of diagnostic test guidelines and no study has looked at the association between guideline quality (in terms guideline quality and reporting, and quality of evidence underpinning recommendations) and guideline non-adherence. It is unclear if adherence to high-quality guidelines is greater than to poor-quality guidelines.

We used a recently published a systematic review that quantified the non-adherence of primary care diagnostic test use with relevant national or international guidelines. Using data from this systematic review we set out to determine if the guidelines used to measure non-adherence were of sufficient quality and whether there is an association between guideline quality and adherence.

2. Methods

This study was conducted and is reported in line with the STROBE checklist. No ethics approval was required.

2.1 Study design

We conducted a meta-epidemiological study [13]. Five steps defined the conduct of this study

1. *Measures of non-adherence:* The measures of guideline non-adherence were extracted from a systematic review we previously published [12].
2. *Assessment of guideline quality and reporting:* For each of the measures of non-adherence, we identified the respective guidelines against which adherence was measured. We then used the Appraisal of Guidelines for Research & Evaluation (AGREE) II tool to determine the quality and reporting of these guidelines.
3. *Assessment of evidence quality:* For each identified guideline (from Step 2), we identified the evidence supporting the guideline recommendations. We then assessed the quality of the evidence using Grading of Recommendations Assessment, Development, and Evaluation (GRADE).
4. *Association between guideline non-adherence and guideline quality and reporting (AGREE II):* We constructed linear models to examine the relationship between guideline quality and reporting (Step 2) and measures of guideline non-adherence (Step 1).
5. *Association between guideline non-adherence and evidence quality (GRADE):* We constructed linear models to examine the relationship between evidence quality (Step 3) and measures of guideline non-adherence (Step 1).

Terminology

- *Measures of non-adherence:* the percentage of tests that were either overused or underused, measured against their respective guideline. These measures are taken from primary studies in the previously published systematic review.
- *Quality and reporting of guidelines:* Methods used to construct a guideline and the reporting standards of a guidelines, measured using the AGREE II tool.
- *Evidence quality:* Assessment of the evidence underpinning guideline recommendations using GRADE

2.2 Guideline non-adherence

Our previously published systematic review [12] determined the non-adherence of diagnostic test ordering against their respective guidelines. The full methods of this systematic review are reported in the paper [12]. Briefly, we included primary observational studies that measured the non-adherence of diagnostic test ordering against a national or international guideline. We extracted each primary study's measure of non-adherence: a measure of non-adherence could be either a) Overtesting: the percentage of tests ordered when a specific guideline recommended to not order it or b) Undertesting: the percentage of tests not ordered when a guideline recommended to order it. These measures of non-adherence, along with a description of the type of diagnostic test, the respective guideline, and the relevant recommendation, are listed in Tables 1 and 2 of the systematic review [12].

2.3 Assessment of the quality and reporting of guidelines using the AGREE II tool

From the primary studies included in our systematic review [12], we identified the guidelines that were used to measure non-adherence. The search technique is described in Appendix A.

We then assessed each guideline using the AGREE II tool. The AGREE II is a tool that 'assesses the methodological rigour and transparency in which a guideline is developed' [14]. The tool itself consists of 23 questions organised into six domains (Scope and Purpose, Stakeholder involvement, Rigour of Development, Clarity of Presentation, Applicability and Editorial Independence). Each domain is assigned a score of 0% to 100%; where 100% indicates that the guideline scored perfectly for that domain. Aside from the 23 questions, the AGREE II tool also prompts users to generate an overall score for the guideline (Appendix B). The 'overall assessment' is a quality score (1-7) for the entire guideline, where 7 represents 'Highest possible quality' and 1 'Lowest possible quality'.

Two reviewers (JOS, AA) independently assessed all included guidelines using the AGREE II tool, both reviewers generated a score for each of its six domains and an overall score for the guideline. We then used the formula in the AGREE II guidance to generate an aggregate score for each domain (Appendix C). The score is

expressed as a percentage of the maximum possible score. The exception to this is the scoring for the 'Overall Assessment', where an average of the two reviewer's scores was used. For the overall assessment score, we performed a weighted kappa test to compare agreement between reviewers.

2.4 Assessment of evidence quality (GRADE)

From the identified guidelines, we extracted the reference(s) quoted to support specific diagnostic test recommendations. An example of this process is described in Appendix A.

We searched PubMed and the Bodleian Library of the University of Oxford for the full texts of studies supporting guideline recommendations. If the recommendation was not referenced, the professional societies/organisation or contact author of the guideline was directly emailed. In the absence of clarification from the guideline organisation, we assumed recommendations were based on expert opinion. Some recommendations explicitly stated they were based on 'expert opinion', these were also classified as expert opinion.

Two reviewers assessed the quality of evidence using GRADE [15]. The first reviewer (JOS) made an independent assessment, which was then verified by the second author (AA). Because of uncertainty in assessing publication bias in diagnostic accuracy studies [16], we did not assess other biases. As is recommended by GRADE [15], expert opinion was considered 'very low quality' evidence.

The identified studies underpinning guideline recommendations were of varying designs. When the underpinning studies were diagnostic accuracy studies or systematic reviews of diagnostic accuracy studies, we used GRADE for diagnostic tests and strategies [17]. Otherwise, we used GRADE for interventions [15].

Appendix D describes the differences between these two GRADE systems.

Risk of bias is one of the five categories in GRADE. We used different risk of bias tools to assess primary studies with different study designs, as recommended by GRADE [18]. For systematic reviews, we used the AMSTAR checklist [19], and for randomised controlled trials (RCTs) we used the Cochrane risk of bias tool [20]. We used the Quality Assessment of Diagnostic Accuracy Studies (QUADAS) 2 tool [21] and the Cochrane risk of bias in cohort studies tool [22] for assessing diagnostic accuracy studies and cohort studies respectively.

In some instances, a guideline included multiple recommendations pertaining to one specific diagnostic test (an example of this is listed in Appendix A). This was problematic because we had only one measure of non-adherent diagnostic test ordering, which may have been covered by multiple recommendations or multiple parts

of a recommendation (e.g. different indications for ordering the test) that had different quality of evidence ratings. In these instances, we specified that we would do the following:

1. If GRADE assessments were consistent across all recommendations (e.g. 'moderate quality'), we would use this GRADE assessment.
2. If the GRADE assessments were different across recommendations, we would select the most frequent assessment (e.g. if there were five recommendations or indications for one test, and four were based on 'low quality' evidence, we would choose 'low quality').
3. If the frequency of GRADE assessments were the same, we would, as recommended by GRADE, select the patient-oriented outcome, rather than diagnostic accuracy or another surrogate [17].

Similarly, the GRADE system for diagnostic tests produces a GRADE score for both sensitivity and specificity. This was similarly problematic. If the GRADE score differed between sensitivity and specificity, we chose the more clinically relevant measure. We determined clinical relevance by consulting the guideline to determine if the test was recommended to rule in or rule out disease. If it was used to rule in, we chose specificity; if it was used to rule out, we chose sensitivity.

2.5 Statistical analysis

To test the association between guideline non-adherence and (a) guideline quality and reporting (AGREE II) and (b) evidence quality (GRADE) we constructed the following linear models:

Percentage guideline non-adherence (0-100%) ~ f(guideline quality and reporting (Overall AGREE II Score, 1-7))

Percentage guideline non-adherence (0-100%) ~ f(evidence quality (GRADE: Very low, low, moderate or high))

2.5.1 Association between quality and reporting of guidelines (AGREE II) and guideline non-adherence

We constructed six further models to assess the relationship between guideline non-adherence and the domains of AGREE II. These models are simple linear regression models, in which the percentage of non-adherence to the guideline was modelled against the scores of each AGREE II domain (0-100%).

2.5.2 Association between Evidence quality (GRADE) and Guideline adherence

GRADE categories were converted into numeric categories (Very low = 1, Low = 2, Moderate = 3, and High = 4). This model was a linear regression model with a categorical predictor.

2.5.3 Sensitivity analyses

We performed sensitivity analyses based on the quality of the primary studies from our previously published systematic review. As previously described, these primary studies provided the measures of non-adherence. In the previous systematic review these studies were rated as low, moderate or high risk of bias, in accordance to the risk of bias tool outlined by Hoy et al [23]. More detailed methods and the results of these sensitivity analyses are presenting in Appendix E.

2.5.4 Subgroup analyses

Furthermore, we performed an additional 12 subgroup analyses where the measures of guideline non-adherence (“percentage guideline non-adherence”) were stratified into overuse (the ordering of a test when it is not recommended) or underuse (the failure to order a test when it is recommended). More detailed methods and the results are presented in Appendix F.

We also conducted an additional subgroup analysis exploring the association between guideline non-adherence and the type of recommendation. Guideline recommendations were stratified into “always use”, “use in specific situations” or “never use”. “Always use” recommendations refer to guidelines that encourage the use of a test in certain clinical situations, for instance use of a chest x-ray to confirm or refute the diagnosis of heart failure. Guideline recommendations that were classified as “use in specific situations” include recommendations that have a series of indications for use, for instance The American Society for Gastrointestinal Endoscopy Appropriate use of Gastrointestinal Endoscopy guideline (Appendix G). Lastly, guidelines that were classified as “never use” include recommendations that discourage the use of a test in a certain situation, i.e. do not use imaging for non-red flag low back pain. The detailed methods and results are described in Appendix J.

Lastly, we also explored the relationship between the number of years between study and guideline publication and non-adherence. The methods and results are reported in Appendix K.

For all models, the distributions of the dependent and independent variables were determined separately. Model assumptions were checked and in the case of violations, appropriate transformations were applied.

For all models, coefficients were determined and are presented with corresponding 95% confidence intervals and their P values. All analyses were performed in R version 3.4.1.

3. Results

3.1 Identifying and locating guidelines

We extracted 103 measures of non-adherent testing from the published systematic review [12]. Of these 103 measures of non-adherent testing, we included 62. Measures of non-adherent testing were excluded for several reasons (Figure 1).

>>insert Figure 1 Inclusion and exclusion of measures of non-adherent testing<<

3.2 Assessment of quality and reporting of guideline using the AGREE II tool

Appendix H lists the AGREE II scores for each of the guidelines used for the 62 measures of non-adherence. The median score for the overall assessment was 4.5 (IQR: 3.5 to 5). The weighted kappa between the two reviewers for the 'Overall Assessment' score was 0.77.

The first AGREE II domain (*Scope and Purpose of the guideline*) was the domain with the highest score across guidelines (median score: 92%), whereas Domain 5 (*Applicability of the guideline*) had the lowest (median score: 33%). The median percentage scores for each domain were: domain 1 'Scope and purpose' 92% (IQR: 81% to 97%), domain 2: 'Stakeholder involvement' 44% (IQR: 37% to 66%), domain 3: 'Rigour of development': 60% (IQR: 29% to 84%), domain 4: 'Clarity of presentation' 78% (IQR: 75% to 92%), domain 5: 'Applicability' 33% (IQR: 13% to 44%) and domain 6: 'Editorial Independence' 50% (IQR: 13% to 67%).

3.3 Assessment of evidence quality (GRADE)

Appendix H lists the GRADE scores for the evidence supporting the guideline recommendations. Most (n=45, 73%) were graded as very low. The remaining were graded as low (n=7, 11%) or high (n= 10, 16%) (none were graded as moderate). Of the 45 measures graded as very low-quality evidence, 37 were expert opinion. The GRADE assessments for the 25 guideline recommendations based on evidence (rather than expert opinion) are

displayed in Appendix I. The median number of primary studies within each GRADE assessment was 2 (IQR 1 to 2). For nine of these 25 GRADE assessments, we used GRADE for diagnostic tests and, for the other 16 we used GRADE for interventions.

3.4 Association between guideline non-adherence and guideline quality and reporting (AGREE II)

Percentage guideline non-adherence (0-100%) ~ f(guideline quality and reporting (Overall AGREE II Score, 1-7))

The linear regression model showed no significant association between guideline quality and reporting (AGREE II score) and the percentage of non-adherent testing. As quality and reporting of a guideline improved, there was a non-significant fall in non-adherent testing (coefficient: -4.2% (95%CI: -9.0% to 0.7%, P = 0.09). Figure 2 shows the median percentage of non-adherent testing for each AGREE II Score. The relationship between AGREE II score and non-adherence did become significant when only including measures of non-adherence from primary studies considered at low risk of bias (coefficient: -6.0% (95%CI: -11.4% to -0.6%), P = 0.03, Appendix E).

>>insert Figure 2 Scatter plot and median of guideline quality and reporting (Overall AGREE II Score) and % guideline non-adherence<<

Association between AGREE II domains and guideline non-adherence

We explored the associations between the scores in the six domains of the AGREE II tool and the percentages of non-adherent testing. High scores in three domains of the AGREE II tool were significantly associated with reductions in non-adherent testing (Domain 2: 'Stakeholder involvement', Domain 3: 'Rigour of development' and Domain 5: 'Applicability'). However, these reductions in non-adherent testing were very small (0.5%, 0.4%, and 0.3% respectively) and are therefore unlikely to have implications for policy or clinical practice. Table 1 reports the model coefficients with corresponding 95%CI and P values. Figure 3 presents the scatterplots for each model. When considering measures of non-adherence extracted from primary studies rated at low risk of bias, increases in scores for the same three domains were associated with a decrease in non-adherent testing (Domain 2, 3, and 4). Additionally, increases in scores in domain 5 of the AGREE II tool was

also associated with a reduction in non-adherent testing (Appendix E). These significant findings were also very small (range: 0.4% to 0.5%) and therefore similarly unlikely to have clinical or policy implications.

>>insert Table 1 Results from Models: Domains of AGREE II and non- adherent testing<<

>>insert Figure 3 Distribution of scores for each Domain of the AGREE II tool<<

3.5 Association between guideline adherence and evidence quality (GRADE)

The model showed a significant association between the quality of evidence and the percentage of non-adherent testing (Table 2). Guideline recommendations based on low and very low-quality evidence had 38% (95%CI: 10% to 65%) and 24% (95%CI: 4.5% to 43%) more non-adherent testing, compared with guideline recommendations based on high-quality evidence. Figure 4 presents the median percentage of non-adherent testing for each GRADE category. This relationship was consistent when only considering measures of non-adherence from studies considered at low risk of bias (Appendix E).

>>insert Table 2 Results from Model <<

>>insert Figure 4 Scatter plot and median of each GRADE category and % guideline non-adherent test ordering (1 = Very low, 2 = Low, 3 = Moderate, 4 = High)<<

The additional subgroup analyses are presented in Appendix F, J and K. Appendix F presents the subgroup analyses where the measures of non-adherence are stratified into underuse and overuse. Only one of these results was statistically significant: the association between domain 6 (“Editorial Independence”) of the AGREE

II tool and guideline non-adherence concerning overuse. However, the effect size was very small (-0.3%, (95%CI: -0.5% to -0.1%)) that it is unlikely to have clinical or policy implications.

Appendix J presents the subgroup analysis where guidelines are stratified into never use, always use or use in specific situations. There was a significant association between guideline type and the percentage of non-adherence: compared with “always use” recommendations, recommendations classed as “use in specific situations” had around 30% less non-adherence (coefficient: -29.1% (95%CI: -48.1% to -10.0%, $P = 0.004$). Similarly, recommendations classed as “never use” had almost 40% less non-adherence (coefficient: -39.1% (95%CI: -51.5% to -26.7%, $P < 0.001$).

Appendix K reports the results of the model exploring the relationship between the number of years between study and guideline publication and non-adherence. There was no significant association.

4. Discussion

We present the first analysis exploring the association between the quality and reporting of guidelines, quality of evidence, and non-adherence. There was an inconsistent relationship between the guideline quality and reporting (AGREE II) and guideline non-adherence; there was no association between overall AGREE II guideline score and the percentage of non-adherent testing. Although, this relationship became significant when only considering measures of non-adherence from low risk of bias studies. Furthermore, three of the domains of AGREE II were significantly associated with a decrease in non-adherent testing, these small statistically significant results are unlikely to be important to clinical practice or policy. Of the guidelines we studied, we found that most were based on poor-quality evidence (very low or low), and that guideline recommendations based on high-quality evidence were significantly associated with lower percentages of non-adherent testing. We also found that guideline recommendations that discourage use of a test are adhered to at a much greater rate than those that encourage test use.

4.1 Strengths and weaknesses in relation to other literature

Previous studies have assessed the quality and reporting of guidelines. A 2012 systematic review identified all studies that assessed the quality of guidelines using the AGREE tool (which preceded the AGREE II tool but has the same domains) [1]. The authors reported the mean score for each AGREE domain and found that the ‘Scope and purpose’ and ‘Clarity and presentation’ domains had the highest scores. They also reported that the

‘Applicability’ domain had the lowest score. Our findings are consistent with these results. The use of the AGREE II tool to assess the quality and reporting of guidelines is a strength of our study. A 2013 systematic review concluded that of all the available guideline appraisal tools, the AGREE II was the most comprehensive and appropriate English language tool to systematically assess guideline quality [24].

Our use of GRADE to assess the quality of the underlying evidence is also a strength. GRADE is used by over 100 organisations from 19 countries, including the WHO, the European Commission, the American College of Physicians, the UK’s National Institute for Health and Care Excellence (NICE), and UpToDate. Few studies have audited the quality of the evidence supporting guideline recommendations. However, it is important to note that it was not possible for us to examine the association between the strength of a guideline recommendation and its adherence. GRADE has two classifications for the strength of a recommendation: “strong”: where a guideline panel is “highly confident of the balance between desirable and undesirable consequences” and “weak” (also known as “conditional”): where a guideline panel is “less confident of the balance between desirable and undesirable consequences” [25]. Four factors determine the strength of a recommendation: the balance between desirable and undesirable consequences, quality of the underlying evidence, uncertainty in the values and preferences and resource use [26]. It is plausible that some of the variation in adherence we note may be explained by a varying strength of guideline recommendation. It is more likely that recommendations based on high-quality evidence would be classed as strong recommendations (but not necessarily). Whereas, it is less common for recommendations based on very low or low-quality evidence to be classified as strong recommendations. As such, it may not be unreasonable that recommendations based on very low or low-quality evidence are adhered to less. Future research addressing this question – the association between adherence and strength of guideline recommendation – would be useful. Nevertheless, our study has shown that, of the diagnostic test guidelines we examined, most were not based on high-quality evidence, highlighting the difficult most guideline panels and clinicians face: there is a paucity of high-quality evidence supporting or refuting the use of diagnostic tests in practice. Our study has several other limitations. We found less than two-thirds of the guidelines from our systematic review. It is plausible that these missing data would change our results quantitatively, however, with the full dataset it is less likely that our conclusion would change: the reporting and quality of a guideline has an unclear effect on adherence, but the quality of evidence supporting guideline recommendations is associated with adherence. Further, our inability to locate all the eligible guidelines begs a wider question: if full-time researchers cannot locate a full guideline, how can busy clinicians?

For some of the guideline recommendations, it was unclear if evidence was cited to support the recommendation. It was not uncommon for guidelines to report 'research suggests', but then not provide a reference. In these uncertain cases, we emailed the professional bodies, seeking clarification. If the guideline organisation did not respond or could not provide the primary reference, we considered the recommendation to be 'Expert Opinion'. It is possible that we misclassified some recommendations as 'Expert Opinion'.

Further, as mentioned in the methods and Appendix A, in some instances, we had multiple indications for test use within one guideline recommendation. This was problematic because we had only one measure of non-adherence diagnostic test ordering. In these situations, it would be ideal to have a measure of non-adherence for each of the specific of the indications within the guideline recommendation, but we were restricted by the primary studies within our systematic review, which only provided one measure of non-adherence. This situation was only relevant for 8 of the 62 (13%) measures of non-adherence.

It is also important to note that we only assessed guidelines and diagnostic tests that had been used in the primary studies in our systematic review [12]. The previously published systematic review captured all primary studies that measured the adherence of *any* primary care diagnostic test against a relevant national or international guideline. However, not all diagnostic tests have been studied in the primary literature. Therefore, it was not possible to include all diagnostic tests available to clinicians. Similarly, we have not examined all diagnostic test guidelines available. It is plausible that if the data were available for *all* tests (and their respective guidelines) our results may differ.

Furthermore, the primary studies included in the previously published systematic review were of varying quality. To address this potential limitation, we conducted sensitivity analyses stratified by risk of bias (of primary studies) – presented in appendix E. Although we have tried to address this, it is still plausible that the variation in quality of primary studies may have affected the results.

Although not strictly a methodological limitation, it is important to consider that guideline recommendations may not be applicable to all patients in all clinical situations. In this study, we do not determine nor state the appropriate rate of guideline adherence, nor do we explore when non-adherence to a recommendation may be appropriate – for example, if a clinician and patient reach a shared decision to not investigate a symptom due to a patient's advanced age. It remains contentious, on a population-level, what rate of guideline adherence is acceptable, and, in a sense, our study provides data to help explore this issue. We do not imply that clinicians and patients should always follow guidelines; there will always be times when it is appropriate to deviate from

recommendations. Medical decisions should consider the best available evidence, clinical guidance, clinician expertise and experience and patient values. We present data that shows that diagnostic test recommendations based on high-quality evidence have greater rates of adherence, compared to those based on low or very-low quality of evidence. We feel this novel finding will be of value to guideline developers.

4.2 Implications

4.2.1 Implications for patients

Domain II of the AGREE II tool ('Stakeholder Involvement') directly addresses the involvement of patients and the public in the development of guidelines. This domain had the second lowest median score; guidelines scored on average 44% of what is expected. Greater involvement of patients in guidelines can improve the quality of guidelines [27]. Further patient involvement can also help guidelines transition from didactic documents to ones that encourage shared decision making [28], which may also lead to greater adherence to recommendations.

4.2.2 Implications for clinicians

Our results have highlighted the caution that clinicians should exert when following clinical practice guidelines. Most recommendations we located were based on no evidence at all and we noted consistent flaws in guidelines, most significantly lack of patient involvement in guideline construction and limited advice about how to implement guidelines in routine clinical practice (Domain 5).

4.2.3 Implications for policymakers

The implications for guideline makers are clear. Guideline recommendations based on high-quality evidence are more likely to be adhered to. When high-quality evidence is available, recommendations should be based on the available evidence. Nevertheless, there remains clinical questions where there is a paucity of high-quality evidence. In these scenarios, it is not possible for guideline developers to base their recommendations on high quality evidence. When this is the case, it is important for the strength of the recommendation to reflect the underlying evidence, i.e. it is less likely to be appropriate for recommendations based on very-low quality evidence to warrant a strong recommendation. Unfortunately, many guideline recommendations with a "strong" strength of recommendation are based on low or very low evidence [2].

There are further implications for policy makers: 1. Guideline developers can highlight evidence gaps. If they develop a guideline and only find very-low or low-quality evidence to support their recommendation, this information should be publicised, especially to academic centres, who can fill these gaps. 2. Policy makers may want to consider targeting performance indicators around guideline recommendations that are supported by high-quality evidence and have a strong strength of recommendation. Conversely, where only very-low or low evidence exists, targeted performance indicators are less likely to be appropriate.

4.2.4 Implications for researchers

Guidelines based on high-quality evidence are adhered to at a significantly greater rate than guidelines based on low or very-low-quality evidence. Future research should be directed to improving the underlying evidence in guidelines based on low or very-low evidence. This should lead to greater adherence and thus mitigate geographical variation [29] and reduce wasteful diagnostic test ordering.

Furthermore, future research investigating the relationship between the strength of a guideline recommendation and its adherence would be advantageous.

5. Conclusions

Of the guidelines that we examined, we have shown that most are based on evidence of low or very low quality. There is heterogeneity in diagnostic test guideline quality and reporting (AGREE II score). Our results show that guideline recommendations based on high-quality evidence are adhered to more frequently.

Author contribution

JOS, CH, and RP conceived the idea for the study. JOS and AA performed the AGREE II and GRADE assessment, JOS and CK performed the statistical analysis. JOS drafted the manuscript with JKA, of which all the authors contributed and approved.

Acknowledgements

The authors have no acknowledgements.

Competing interest statement

All authors declare no competing interests.

Funding

This paper did not receive any dedicated funding.

Appendix file captions

Appendix A: Extended Methods

Appendix B: AGREE II Tool

Appendix C: Formula used to calculate score for AGREE II domains

Appendix D: Differences between GRADE for diagnostic accuracy research compared with GRADE for intervention research

Appendix E: Sensitivity analysis: Risk of Bias

Appendix F: Subgroup analyses: Overuse vs. underuse

Appendix G: Specific guideline recommendations (that were too long for Appendix H)

Appendix H: Guideline recommendations and their corresponding AGREE II and GRADE scores

Appendix I: GRADE profiles

Appendix J: Subgroup analyses: Type of recommendation (always use vs. never use vs. use in specific situations)

Appendix K: Raw data file

References

- [1] Alonso-Coello P, Irfan A, Solà I, Gich I, Delgado-Noguera M, Rigau D, et al. The quality of clinical practice guidelines over the last two decades: A systematic review of guideline appraisal studies. *Qual Saf Heal Care* 2010;19:1–8. doi:10.1136/qshc.2010.042077.
- [2] Alexander PE, Brito JP, Neumann I, Gionfriddo MR, Bero L, Djulbegovic B, et al. World Health Organization strong recommendations based on low-quality evidence (study quality) are frequent and often inconsistent with GRADE guidance. *J Clin Epidemiol* 2016;72:98–106. doi:10.1016/j.jclinepi.2014.10.011.

- [3] Garber AM. Evidence-based guidelines as a foundation for performance incentives. *Health Aff (Millwood)* 2005;24:174–9. doi:10.1377/hlthaff.24.1.174.
- [4] Ransohoff DF, Pignone M, Sox HC, HC S, R G, J K, et al. How to Decide Whether a Clinical Practice Guideline Is Trustworthy. *JAMA* 2013;309:139. doi:10.1001/jama.2012.156703.
- [5] Roland M, Guthrie B. Quality and Outcomes Framework: what have we learnt? *BMJ* 2016;354:i4060. doi:10.1136/bmj.i4060.
- [6] Sackett DL, Rosenberg W, Grey M, Haynes RB, Richardson W. Evidence based medicine: what it is and what it isn't. *BMJ* 1996;312.
- [7] Fryar C. Doctors can depart from guidelines in patients' best interests. *BMJ* 2015;350.
- [8] Shaneyfelt TM, Mayo-Smith MF, Rothwangl J. Are guidelines following guidelines? The methodological quality of clinical practice guidelines in the peer-reviewed medical literature. *JAMA* 1999;281:1900–5.
- [9] Grilli R, Magrini N, Penna A, Mura G, Liberati A, Thomson R, et al. Practice guidelines developed by specialty societies: the need for a critical appraisal. *Lancet (London, England)* 2000;355:103–6. doi:10.1016/S0140-6736(99)02171-6.
- [10] Gale EAM. Conflicts of interest in guideline panel members. *BMJ* 2011;343.
- [11] Lenzer J. Why we can't trust clinical guidelines. *BMJ* 2013;346.
- [12] O'Sullivan JW, Albasri A, Nicholson B, Perera R, Aronson J, Roberts N, et al. Overtesting and undertesting in primary care: a systematic review and meta-analysis. *BMJ Open* 2018;8:e018557. doi:10.1136/bmjopen-2017-018557.
- [13] Murad MH, Wang Z. Guidelines for reporting meta-epidemiological methodology research. *Evid Based Med* 2017;22:139–42. doi:10.1136/ebmed-2017-110713.
- [14] Brouwers MC, Kho ME, Browman GP, Burgers JS, Cluzeau F, Feder G, et al. AGREE II: Advancing guideline development, reporting and evaluation in health care. *J Clin Epidemiol* 2010;63:1308–11. doi:10.1016/j.jclinepi.2010.07.001.
- [15] Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: An

- emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008;9:8–11.
doi:10.1136/bmj.39489.470347.AD.
- [16] Rogozinska E, Khan K. Grading evidence from test accuracy studies: what makes it challenging compared with the grading of effectiveness studies? *Evid Based Med* 2017;22:81–4.
doi:10.1136/ebmed-2017-110717.
- [17] Schünemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. GRADE: grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ* 2008;336:0–b. doi:10.1136/bmj.a139.
- [18] Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, et al. GRADE guidelines: 4. Rating the quality of evidence - Study limitations (risk of bias). *J Clin Epidemiol* 2011;64:407–15.
doi:10.1016/j.jclinepi.2010.07.017.
- [19] Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Med Res Methodol* 2007;7:10. doi:10.1186/1471-2288-7-10.
- [20] Higgins JPT, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *Bmj* 2011;343:d5928–d5928.
doi:10.1136/bmj.d5928.
- [21] Whiting PF, Rutjes AWS, Westwood ME, Mallet S, Deeks JJ, Reitsma JB, et al. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Ann Intern Med* 2011;155:529–36.
- [22] Cochrane Collaboration. Tool to assess risk of bias in cohort studies. *Cochrane Methods* 2013:1–4.
- [23] Hoy D, Brooks P, Woolf A, Blyth F, March L, Bain C, et al. Assessing risk of bias in prevalence studies: modification of an existing tool and evidence of interrater agreement. *J Clin Epidemiol* 2012;65:934–9. doi:10.1016/j.jclinepi.2011.11.014.
- [24] Siering U, Eikermann M, Hausner E, Hoffmann-Eßer W, Neugebauer EA. Appraisal tools for clinical practice guidelines: A systematic review. *PLoS One* 2013;8. doi:10.1371/journal.pone.0082915.
- [25] Andrews J, Guyatt G, Oxman AD, Alderson P, Dahm P, Falck-Ytter Y, et al. GRADE guidelines: 14.

- Going from evidence to recommendations: The significance and presentation of recommendations. *J Clin Epidemiol* 2013;66:719–25. doi:10.1016/j.jclinepi.2012.03.013.
- [26] Andrews JC, Schünemann HJ, Oxman AD, Pottie K, Meerpohl JJ, Coello PA, et al. GRADE guidelines: 15. Going from evidence to recommendation - Determinants of a recommendation's direction and strength. *J Clin Epidemiol* 2013;66:726–35. doi:10.1016/j.jclinepi.2013.02.003.
- [27] Diaz del Campo P, Gracia J, Blasco JA, Andradas E. A strategy for patient involvement in clinical practice guidelines: methodological approaches. *BMJ Qual Saf* 2011;20:779–84. doi:10.1136/bmjqs.2010.049031.
- [28] Berger ZD, Brito JP, Ospina NS, Kannan S, Hinson JS, Hess EP, et al. Patient centred diagnosis: sharing diagnostic decisions with patients in clinical practice. *Bmj* 2017;4218:j4218. doi:10.1136/bmj.j4218.
- [29] O'Sullivan JW, Heneghan C, Perera R, Oke J, Aronson JK, Shine B, et al. Variation in diagnostic test requests and outcomes: a preliminary metric for OpenPathology.net. *Sci Rep* 2018;8:4752. doi:10.1038/s41598-018-23263-z.

Table 1 Results from Models: Domains of AGREE II and non-adherent testing

| AGREE II Domain | Coefficient (95%CI) | P value |
|-----------------|-------------------------|---------|
| Domain 1 | 0.1% (-0.5% to 0.7%) | 0.7 |
| Domain 2 | -0.5% (-0.7% to -0.2%) | <0.01* |
| Domain 3 | -0.4% (-0.6% to -0.1%) | <0.01* |
| Domain 4 | -0.3% (-0.9% to 0.3%) | 0.3 |
| Domain 5 | -0.3% (-0.6% to -0.01%) | 0.04* |
| Domain 6 | -0.2% (-0.4% to 0.1%) | 0.2 |

Table 2 Results from Model

| GRADE category | Coefficient (95%CI) | P value |
|---------------------------|------------------------|---------|
| 'High' Reference standard | | |
| Low | 37.7% (10.4% to 65.0%) | <0.01* |
| Very low | 23.9% (4.5% to 43.3%) | 0.02* |

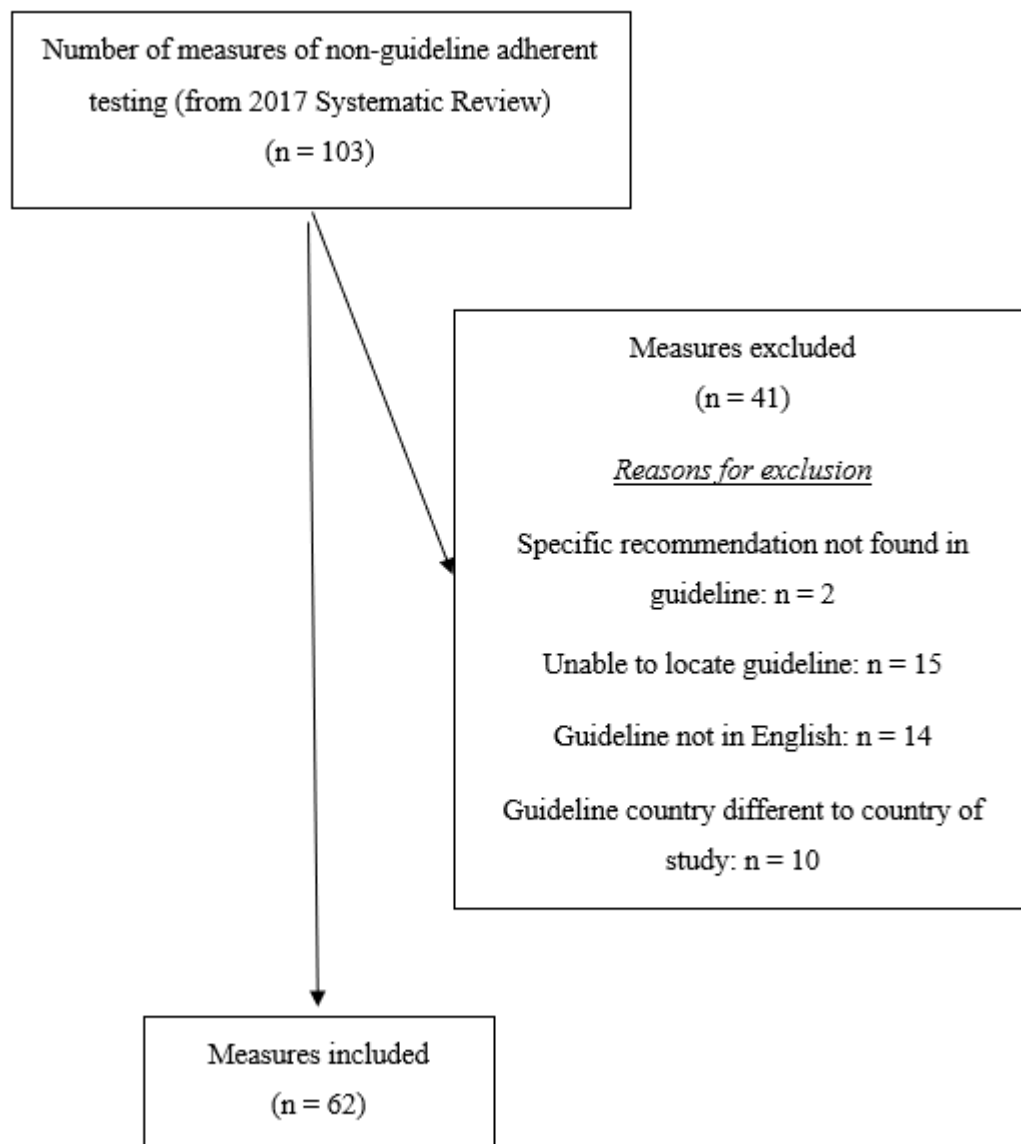


Figure 1 Inclusion and exclusion of measures of non-adherent testing

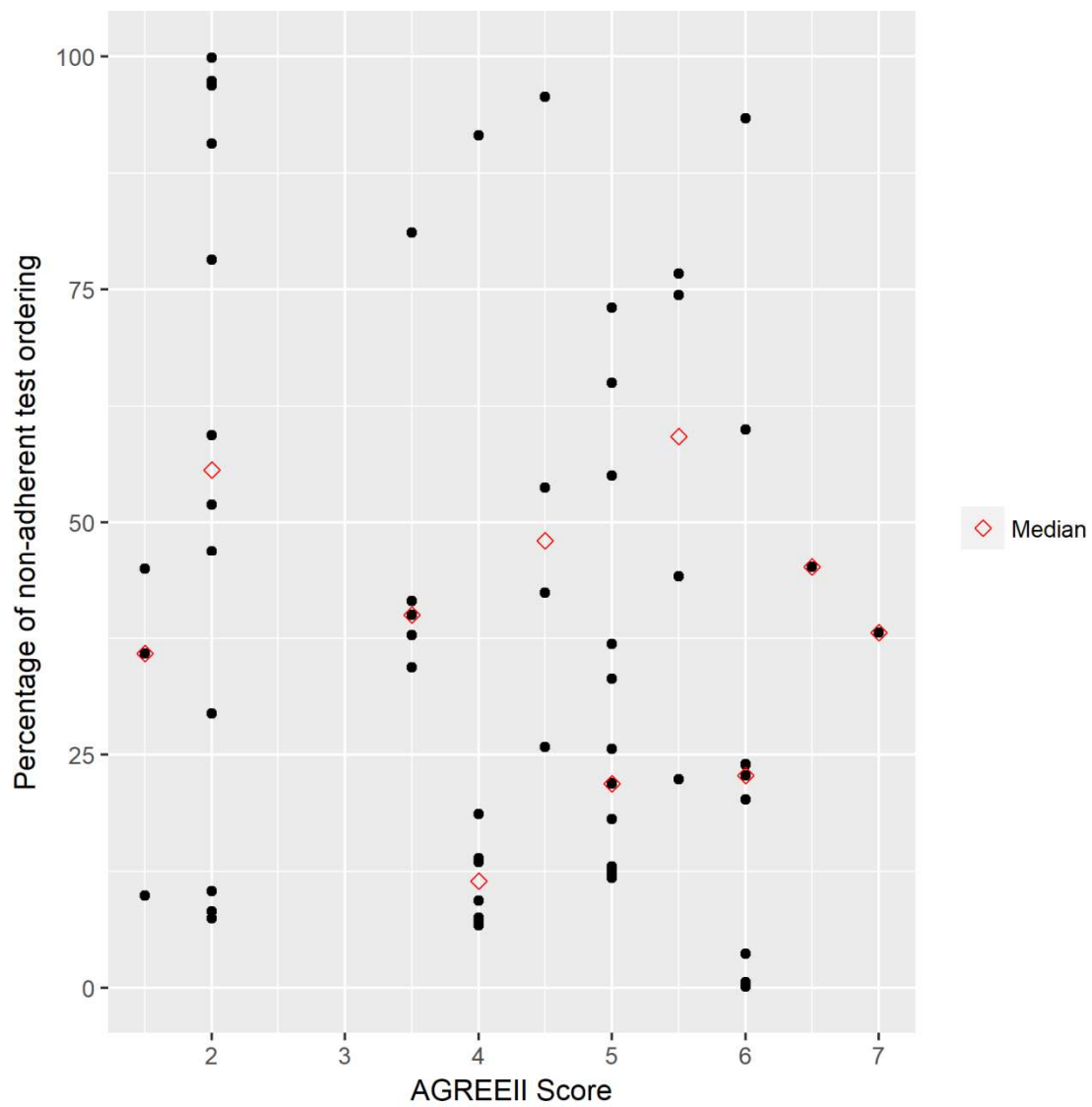


Figure 2 Scatter plot and median of guideline quality and reporting (Overall AGREE II Score) and % guideline non-adherence

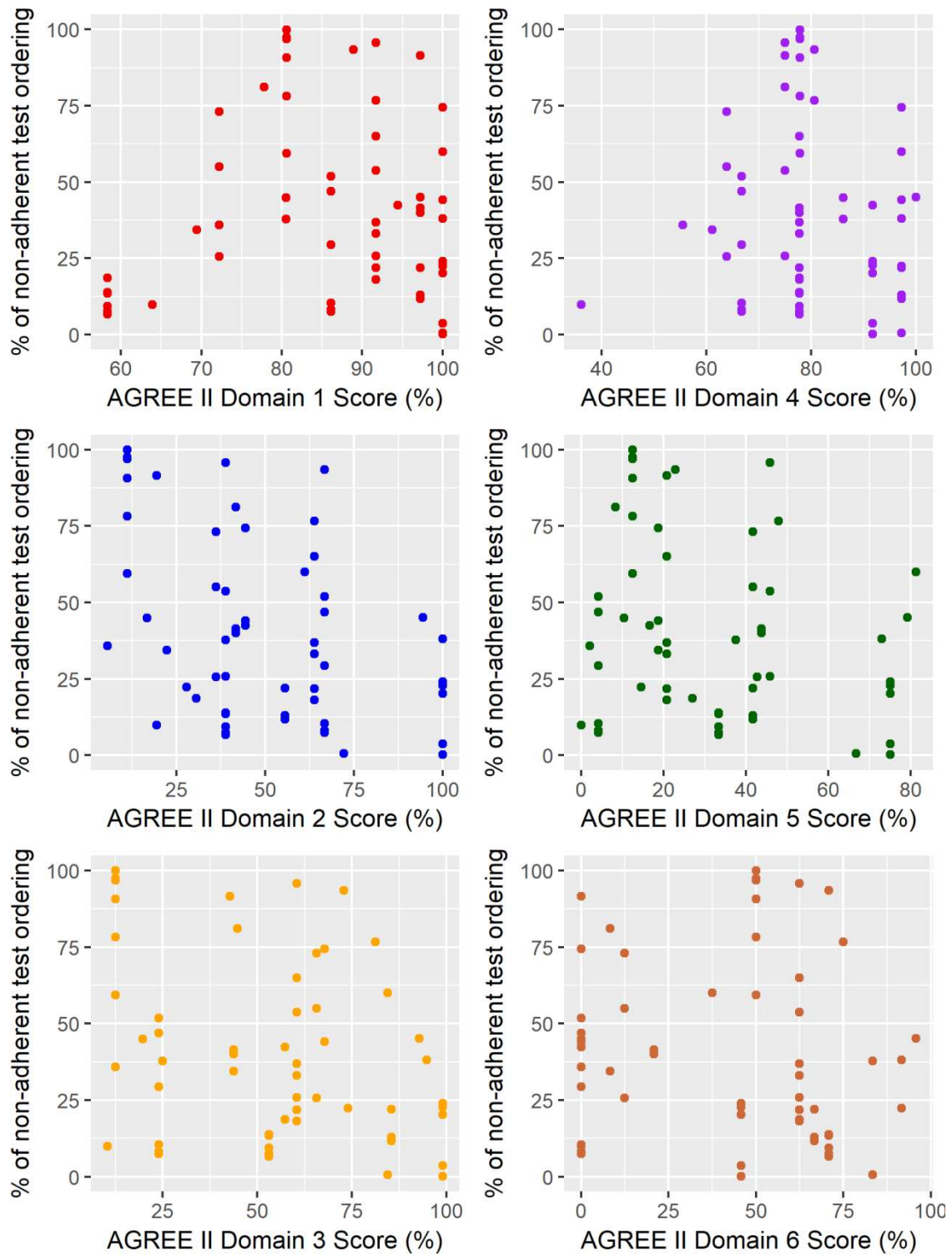


Figure 3 Distribution of scores for each Domain of the AGREE II tool

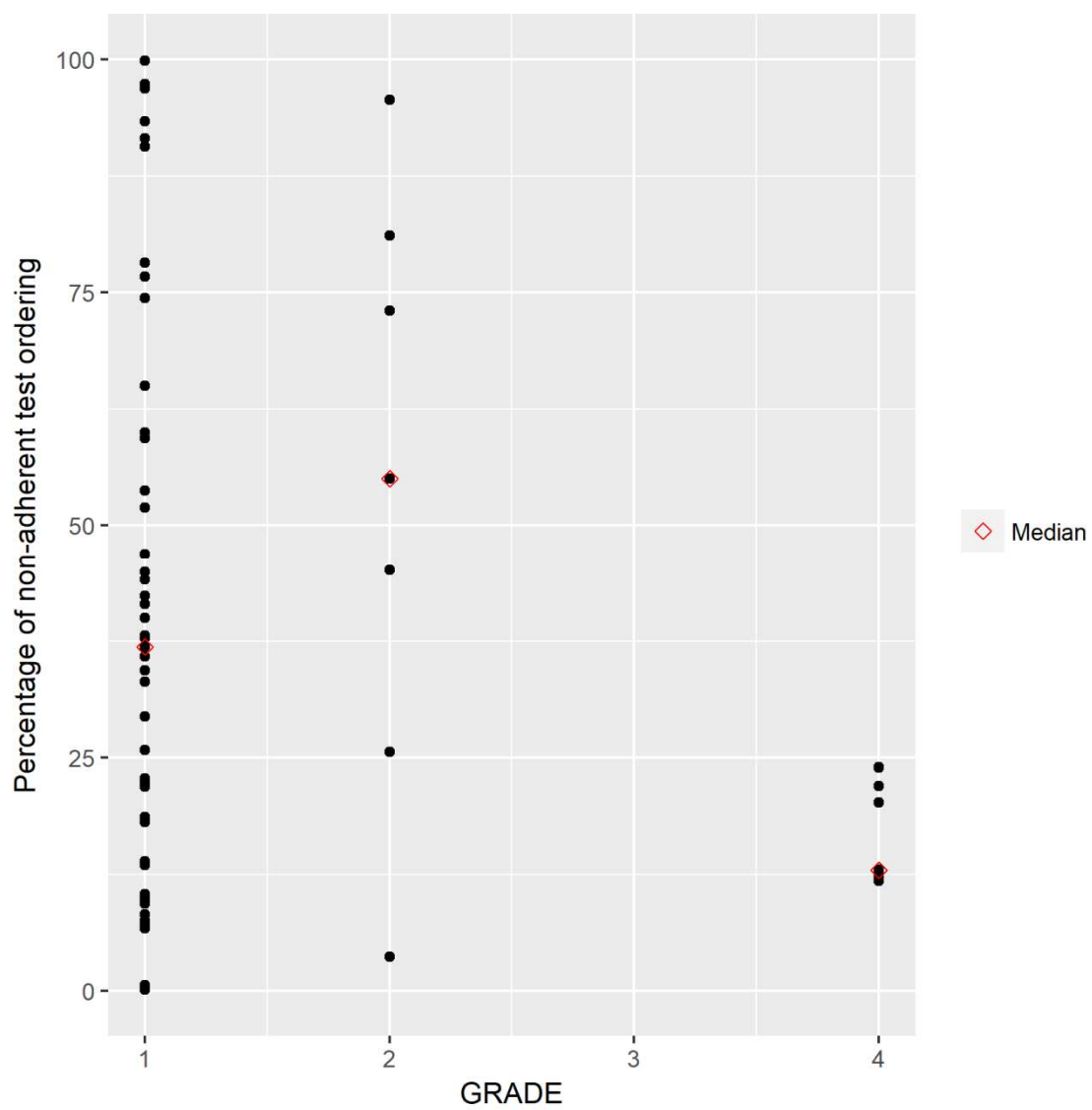


Figure 4 Scatter plot and median of each GRADE category and % guideline non-adherent test ordering (1 = Very low, 2 = Low, 3 = Moderate, 4 = High)