**Please see the below email and attachments sent from Dr. Rod Handy regarding the upcoming Department Journal Club.*

Hello Faculty/Staff:

The Education Committee of the Department of Family and Preventive Medicine is pleased to announce **DFPM Journal Club**, another opportunity for inter-Divisional collaboration and a first time inter-professional opportunity for all of our learners. The selected article (below) offers systematic review and meta-analysis for discussion. We will follow the below worksheet to guide discussion. Perspectives from all disciplines are relevant and encouraged. Please share this opportunity with your students.

The first date is **Wednesday, March 27th from 7:30 am to 8:30 am, and breakfast will be available**. This time and date coincides with FM Grand Rounds: CME is offered.

Participants should check in at 375 Chipeta in the downstairs common area (kitchen) on time to grab a bite and be assigned to one of the following breakout rooms, 12 (or so) participants each:

- Classroom 203
- Classroom 104
- Classroom 124
- Classroom 125
- Conference Room 109 (IVC to St. George campus)
- Classroom @ 421 Wakara (IVC to St. George campus)

Lead faculty facilitators are:

- Clarivette Bosch (FM Division)
- Jennifer Coombs (PA Division)
- Kara Frame (FM Division)
- Karen Schliep (PH Division)
- Virginia Valentin (PA Division)
- Sarang Kim Yoon (OEH Division)

Again, we encourage participation from all students, staff, and faculty of the Department of Family and Preventive Medicine. For future Journal Club dates (or other possible events), we plan to engage faculty from each Division to lead and facilitate such inter-professional learning opportunities.

Thanks, Rod

Rod Handy, MBA, PhD, CIH
Professor and Interim Vice Chair of Education and Research
Director of IH, TRT, and OEH Graduate Studies
Chair of DFPM Education Mission Committee
Dept. of Family and Preventive Medicine (DFPM)
Rocky Mountain Center for Occupational & Environmental Health

BMJ Open

Overtesting and undertesting in primary care: a systematic review and meta-analysis

Jack W O'Sullivan,[1] Ali Albasri,[1] Brian D Nicholson,[1] Rafael Perera,[1] Jeffrey K Aronson,[1] Nia Roberts,[2] Carl Heneghan[1]

To cite: O'Sullivan JW, Albasri A, Nicholson BD, *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

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2017-018557).

Received 6 July 2017
Revised 12 December 2017
Accepted 13 December 2017

[1]Centre for Evidence-Based Medicine, Nuffield Department of Primary Care Health Science, University of Oxford, Oxford, UK
[2]Bodleian Health Care Libraries, University of Oxford, Oxford, UK

Correspondence to
Dr Jack W O'Sullivan;
jack.osullivan@phc.ox.ac.uk

ABSTRACT

Background Health systems are currently subject to unprecedented financial strains. Inappropriate test use wastes finite health resources (overuse) and delays diagnoses and treatment (underuse). As most patient care is provided in primary care, it represents an ideal setting to mitigate waste.

Objective To identify overuse and underuse of diagnostic tests in primary care.

Design Systematic review and meta-analysis.

Data sources and eligibility criteria We searched MEDLINE and Embase from January 1999 to October 2017 for studies that measured the inappropriateness of any diagnostic test (measured against a national or international guideline) ordered for adult patients in primary care.

Results We included 357 171 patients from 63 studies in 15 countries. We extracted 103 measures of inappropriateness (41 underuse and 62 overuse) from included studies for 47 different diagnostic tests. The overall rate of inappropriate diagnostic test ordering varied substantially (0.2%–100%)%). 17 tests were underused >50% of the time. Of these, echocardiography (n=4 measures) was consistently underused (between 54% and 89%, n=4). There was large variation in the rate of inappropriate underuse of pulmonary function tests (38%–78%, n=8). Eleven tests were inappropriately overused >50% of the time. Echocardiography was consistently overused (77%–92%), whereas inappropriate overuse of urinary cultures, upper endoscopy and colonoscopy varied widely, from 36% to 77% (n=3), 10%–54% (n=10) and 8%–52% (n=2), respectively.

Conclusions There is marked variation in the appropriate use of diagnostic tests in primary care. Specifically, the use of echocardiography (both underuse and overuse) is consistently poor. There is substantial variation in the rate of inappropriate underuse of pulmonary function tests and the overuse of upper endoscopy, urinary cultures and colonoscopy.

PROSPERO registration number CRD42016048832.

Strengths and limitations of this study

- Generates rate of undertesting and overtesting for specific diagnostic tests against national or international guidelines.
- Only includes data from real clinical encounters rather than surveys or hypothetical clinical vignettes.
- Quantified inappropriate ordering of all types of diagnostic tests rather than just laboratory.
- Systematic reviews are restricted to published literature; thus, rates of inappropriate ordering are not available for all tests available to primary care physicians.
- Included studies measure appropriateness of testing in a particular healthcare setting against a particular guideline, thus reflect test ordering in a specific healthcare setting.

INTRODUCTION

Reaching a diagnosis in primary care is exceedingly complex. The combination of undifferentiated symptoms, a low prevalence of serious disease, a high degree of symptom overlap between serious and benign conditions, patients with multiple complaints and psychological or social distress manifesting somatically all complicate reaching a diagnosis.[1] In around 40% of primary care consultations, a diagnosis cannot be established from the history and physical examination alone,[2] and tests are therefore often needed.[1 3]

Primary care consultations make up most of the care provided in healthcare systems (90% of consultations in the UK,[4] 55% of consultations in the USA[5]) and inappropriate diagnostic testing in primary care therefore has enormous resource implications. Given the calls for £22 billion in efficiency savings from the UK's National Health Service[6] and the $660 billion US Medicare deficit predicted by 2023,[7] ensuring the appropriateness of primary care diagnostic testing is crucial to the sustainability of healthcare systems.[8]

Inappropriate diagnostic tests in primary care can be both inappropriately underused and overused. Underuse of tests, failure to order a test when indicated, can lead to diagnostic errors and delays in diagnosis and the delivery of effective treatment, leading to adverse patient outcomes and further

healthcare costs.[9 10] Overuse of tests, the delivery of tests with no clear benefit or when potential harms outweigh potential benefits, subjects patients to direct harms, such as radiation exposure, as well as potential adverse outcomes (eg, contrast nephropathy),[11] incidental findings[12] and overdiagnosis.[13] Overuse is also a waste of finite healthcare expenditure, diverting resources from beneficial tests and treatments.[14–16]

Many drivers encourage inappropriate underuse and overuse of diagnostic tests in primary care. Greater access to tests,[17] the medicolegal consequences of undertesting,[18] few if any disincentives to overinvestigate[14] and clinical performance measures[19] may all contribute to overuse. Increasing primary care workload,[4] time constraints[19] and difficulty keeping up-to-date with rapidly increasingly evidence[20] may contribute to both inappropriate underuse and overuse.

Guidelines set the standard of care across most healthcare settings.[21 22] Furthermore, they provide a medicolegal framework,[23] inform healthcare policy and improve both care outcomes and processes of care.[24] Despite some recognised limitations, including varying quality of guidelines,[25–27] guidelines are often used as markers of healthcare appropriateness.[28–31] Zhi *et al*,[29] for instance, used guidelines as a measure of appropriateness to estimate underuse and overuse of laboratory testing. They estimated that 45% (95% CI 34% to 56%) of secondary care laboratory testing is underused and 21% (95% CI 16% to 25%) is overused.

Despite the increasing use of healthcare resources,[32] rising healthcare expenditure,[6–8] increasing demands placed on primary care[4] and the apparent drivers of inappropriate testing,[1 4 14 17–20] it is not clear how often diagnostic tests are inappropriately overused or underused in primary care. We therefore conducted a systematic review to quantify the frequency of inappropriate ordering of all types of diagnostic tests from primary care in relation to their respective guidelines and identify tests that are frequently overused and underused.

METHODS

This study was conducted and is reported in line with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses[33] and Meta-analysis of Observational Studies in Epidemiology statements.[34]

Protocol and registration

The protocol has been published and is available online (open access) via the International Prospective Register for Systematic Reviews database (registration ID: CRD42016048832).

Search strategy

We searched Embase (OvidSP) and MEDLINE (OvidSP) databases from January 1999 to October 2017 for studies of any design measuring how often diagnostic test guidelines were followed in primary care (see online supplementary file 1: search Strategy). Our search strategy can be summarised as: 'Ambulatory Care AND adherence AND guideline AND diagnostic tests AND inappropriate'. Conference abstracts published after 2015 were also searched for in these databases to capture data not yet published. We also searched the WHO International Clinical Trials Registry Platform (http://apps.who.int/trialsearch/), ClinicalTrials.gov, and the reference lists of included studies.

Eligibility criteria

We included studies of any design if they measured the rate of inappropriate ordering (overuse) or not ordering (underuse) of diagnostic tests ordered from primary care against national or international guidelines. We considered all diagnostic tests ordered in adults. We also included studies that measured diagnostic tests ordered from primary care but performed in secondary care (eg, upper endoscopy). We included the control arms of randomised controlled trials (RCTs) if they offered exclusively usual care and the preintervention periods of studies that used interrupted time series designs (before and after studies).

We excluded studies if they met the following criteria: >20% of participants were children (>20% under 18 years old); diagnostic tests not ordered by general practitioners; and screening or monitoring tests or publication before 1999 (studies after 1999 were considered to ensure that results would more closely reflect current practice). We defined a screening test as a test on an asymptomatic or symptomatic person without signs or symptoms related to that test.[35 36] We defined monitoring tests as 'a test for a patient with an established diagnosis, for which the test is used to measure progression of the disease'.[37] We excluded studies if they did not give a measure of appropriateness or if appropriateness was measured against local guidelines, such as a guideline specific to a hospital or region, rather than international or national guidelines.

Study selection and data extraction

Three reviewers (JWO and AA or BDN) independently screened titles, abstracts and full texts for eligibility. The same reviewers assessed risks of bias and extracted the following data from included studies: patient demographics, eligibility criteria, name and type of diagnostic test, duration of study (days), guideline name and recommendation, total number of tests performed and the number of tests ordered when the specific guideline recommended not ordering (inappropriate overuse) or the number of tests not ordered when the guideline recommended ordering it (inappropriate underuse). The last two data points (overuse and underuse) represent 'measures of inappropriateness'. When studies measured inappropriateness of multiple tests, we extracted data on each test and presented them as individual measures of inappropriateness. When studies measured tests across different periods, we extracted measures for each time point and considered each one as an individual measure of inappropriateness.

We assessed the quality of included studies using a modified version of the Hoy risk of bias tool.[38] This tool has been validated to assess the internal and external validity of prevalence studies.[38] Our modified version of this tool kept the same domains but adjusted the wording of the tool to reflect prevalence of inappropriate testing rather than prevalence of disease. Our tool (and results) is available in online supplementary file 2: risk of bias.

Statistical analysis

The primary outcome was the prevalence of inappropriate diagnostic testing. Inappropriate testing was measured in two ways:

1. Overuse: a diagnostic test was ordered when the relevant guideline recommends not ordering it, for instance, imaging for non-red flag low back pain (LBP).
2. Underuse: a diagnostic test was not ordered when the relevant guideline recommended ordering it, for instance, spirometry to confirm or refute the diagnosis of chronic obstructive pulmonary disease (COPD).

We expressed measures of inappropriateness as percentages (%), where the numerator represents the total number of times a guideline recommendation was not followed, and the denominator represents the total number of times a guideline recommendation could have been followed. For instance, the number of times imaging was inappropriately ordered for non-red flag headache as a percentage of the total number of patients who presented with non-red flag headache. As our included data are percentages, we calculated Clopper-Pearson 95% CIs for each individual measure of appropriateness. We conducted sensitivity analyses with high risk of bias studies excluded.

Where the same guideline and recommendation were used by multiple studies (eg, five studies measured inappropriate underuse of spirometry testing in patients with COPD[39–43] using the Global Initiative for Chronic Obstructive Lung Disease (GOLD) guideline), we pooled the measures and assessed heterogeneity. We combined measures of inappropriateness using a random-effects meta-analysis with 95% CIs(Clopper-Pearson), for the reason that each measure of appropriateness contributed relatively evenly to pooled estimates. We performed double arcsine transformation on prevalence data to stabilise the variance[44] and pooled the data using the inverse variance method.[45] We assessed heterogeneity using the I^2 statistic.[46] We did not combine measures of overuse and underuse, as they have different denominators: overuse involves the total number of tests ordered, whereas underuse involves the total number of times a test should have been ordered. We performed analyses using R V.3.3.2 (R project).

RESULTS

Study selection and characteristics

We included 63 studies from 14 716 references identified from independent searches by two authors (JWO and AA or BDN) (see figure 1). Of the 63 included studies, 55 were observational studies, 6 were before-and-after studies and 2 were RCTs. The two RCTs investigated the effect of implementing an intervention to reduce inappropriate testing. These studies were conducted in 15 countries and included 357 171 patients (see online supplementary file 3: table 1). Online supplementary file 4: table 1 shows the 103 measures of inappropriateness extracted from included studies for 47 different diagnostic tests measured against 77 guideline recommendations (41 measured underuse and 62 measured overuse). Guideline recommendations came from 42 different guideline organisations from 15 countries.

Fourteen studies measured inappropriateness of more than one diagnostic tests for the same condition (eg, chest X-ray, electrocardiography and transthoracic echocardiography to confirm or refute a diagnosis of heart failure). Two studies[47 48] measured inappropriateness across multiple time periods. No studies measured both underuse and overuse of the same test.

Included studies measured inappropriateness in one of three ways:

1. Patients with specific symptoms were assessed (prospectively or retrospectively) to see if they had received an inappropriate diagnostic test (overuse) or had not received the appropriate diagnostic test (underuse) in line with the relevant guideline recommendation (eg, records for patients with non-red flag LBP to see if they received imaging[49]). Eighteen studies used this method.
2. Patients who had undergone a diagnostic test were identified (via hospital or national databases), and an assessment of whether the test was inappropriate (as per the defined guideline recommendations) via individual patient data was made (overuse). For instance, patients who had an upper endoscopy.[50] Twenty-two studies used this method.
3. Patients with a diagnosis were identified via hospital or national databases and assessed to see whether they had received the appropriate diagnostic test (as per the defined guideline) to confirm or refute the diagnosis via individual patient data (underuse). For instance, assessing if patients with a diagnosis of COPD had spirometry to confirm or refute the diagnosis.[39] Twenty-three studies used this method.

Risk of bias

Two-thirds of the studies (n=44) were graded as being at low risk of bias, 15 (24%) at moderate risk and 4 (6%) at high risk (see online supplementary file 2: risk of bias). Moderate or high risk studies were at an increased risk of non-response bias (>20%), non-objective collection of data and/or unclear intervals between symptom onset and diagnostic test use. Supplementary file 2: risk of bias outlines risk of bias scores in detail.

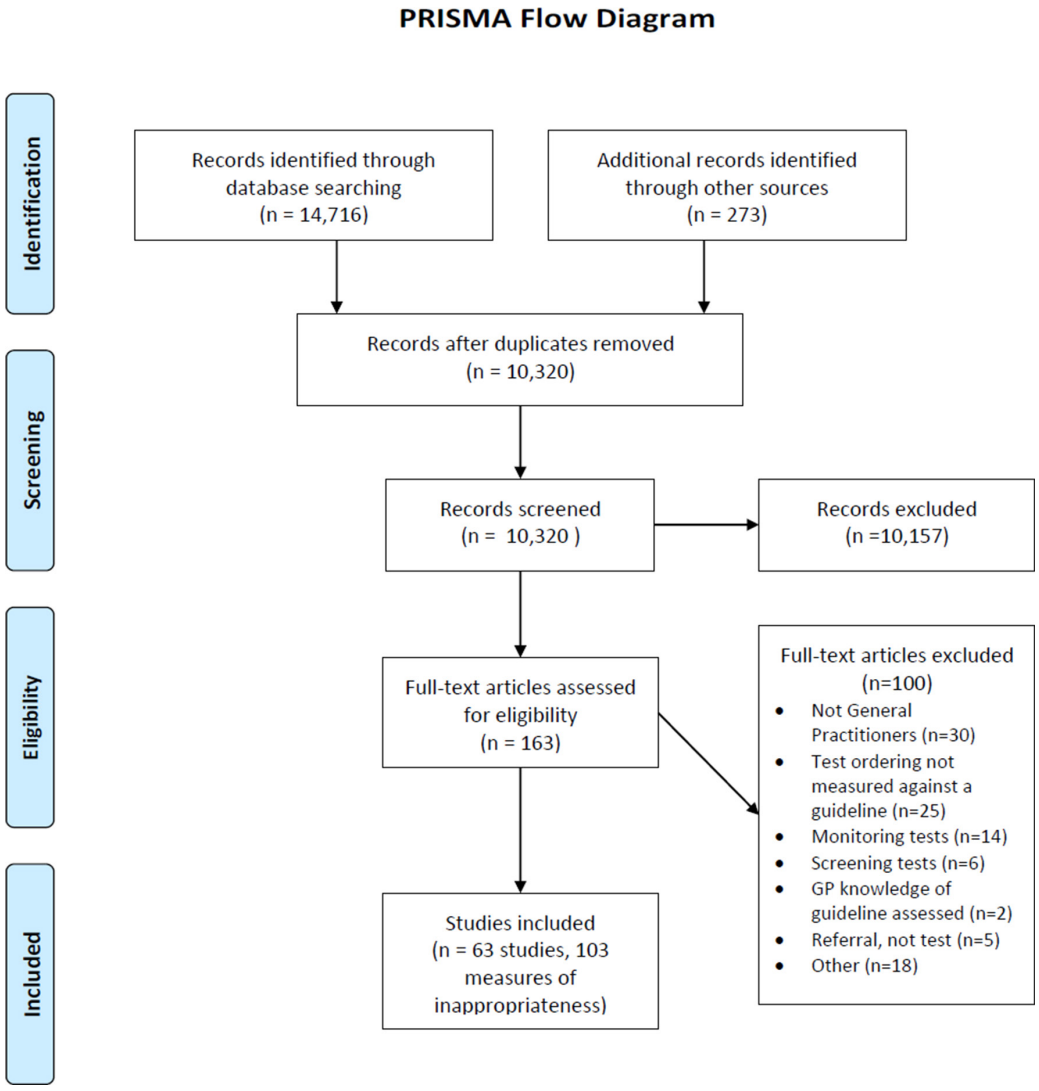

Figure 1 PRISMA flow diagram. GP, general practitioner; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

Percentage of diagnostic tests ordered in line with specific guideline recommendations

There was large variation in the rate of inappropriate diagnostic test ordering. The 103 diagnostic test guideline recommendations were not followed 0.2%–100% of the time (see online supplementary file 4 table 1); wide variation was largely sustained (0.2%–99.94%) when a further analysis was conducted excluding studies judged to be of high risk of bias. The prevalence of underuse varied 8.2%–100%, whereas overuse varied between 0.2% and 94.2%. Similarly, this variation was essentially maintained on exclusion of high risk studies (under use 9.8%–99.9%, overuse 0.2%–94.2%).

Underused tests

Online supplementary file 4 table 1 shows that 17 tests were underused more than 50% of the time. Echocardiography was the most frequently studied (n=4, twice in the UK and once in Poland and Brazil) . In patients with heart failure, echocardiography was underused between 54% and 89% (n=3) of the time and in atrial fibrillation 56% (n=1).

For some tests, there was large variation in the rate of underuse (figure 2). Underuse of pulmonary function tests (PFTs) to confirm or refute COPD, measured against the GOLD, National Institute for Health and Care Excellence (UK) and Danish National Board of Health guidelines, varied from 26% to 78% (n=8). None of the studies that studied echocardiography or PFTs were considered high risk of bias and thus results did not change on further analysis excluding high-risk studies.

Overused tests

Eleven tests were overused more than 50% of the time (figure 3). Echocardiography was consistently overused, for instance in 'routine perioperative evaluation of ventricular function with no symptoms or signs of cardiovascular disease', whereas other tests (urinary cultures, upper endoscopy and colonoscopy) were overused at varying rates. The overuse of echocardiography

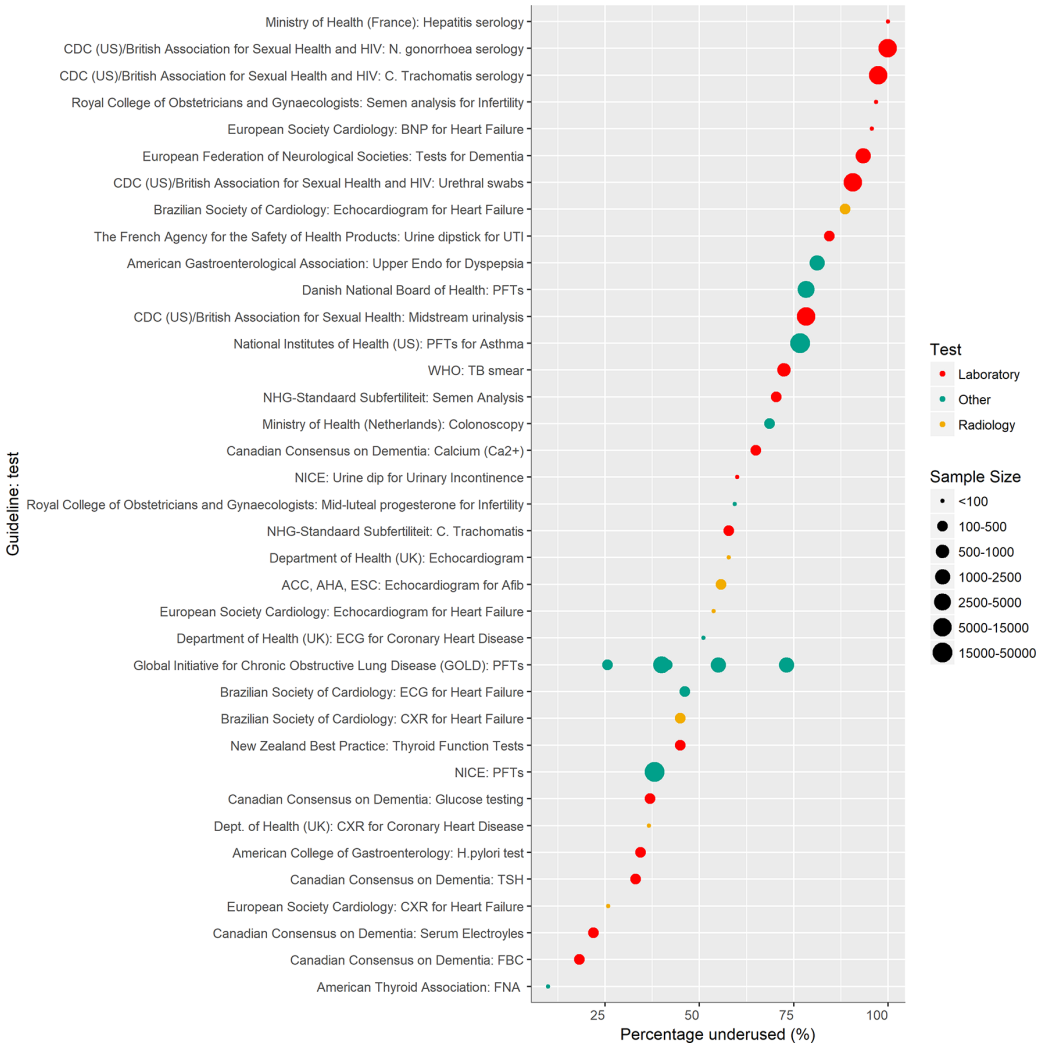

Figure 2 Rates of underuse. ACC, American College of Cardiology; AFib, atrial fibrillation; AHA, American Heart Association; CDC, Centers for Disease Control and Prevention; CXR, chest X-ray; ESC, European Society of Cardiology; FBC, full blood count; FNA, fine needle aspiration; GOLD, Global Initiative for Chronic Obstructive Lung Disease; NICE, National Institute for Health and Care Excellence; PFTs, pulmonary function tests; TB, tuberculosis; TSH, thyroid stimulating hormone; UTI, urinary tract infection.

was studied in the UK[51] and the Netherlands.[52] The rates of overuse varied between the two settings: between 77% (Netherlands) and 92% (UK). The overuse of urinary cultures for uncomplicated urinary tract infections was studied in the USA,[53 54] Spain[55] and Sweden.[56] The rate of overuse varied from 57% to 77% in the USA, compared with approximately 50% in Sweden and 36% in Spain. Overuse of upper endoscopy was studied widely (n=11) in Australia,[57 58] Saudi Arabia,[59 60] UK,[61] Italy,[62–64] USA[50 65] and Malaysia.[66] The overuse varied markedly, from 7.5% to 54% (n=11), respectively (figure 3, online supplementary file 3 table 1). Similarly, the inappropriate overuse of colonoscopy varied substantially from 8% in Australia[58] to 52% in Malaysia.[67] None of the above studies were considered high risk of bias and thus results did not change on further analysis excluding high-risk studies.

Our results also suggest that the inappropriate overuse of CT and MRI scans for non-red flag headache (a headache without symptoms suggesting a malignant underlying pathology) has more than doubled in the last 10 years in the USA (*2000*: 6.7% (95% CI 5.4% to 8.2%), *2010*: 14% (95% CI 12.% to 16%)) (see online supplementary file 4 table 1).[48] Conversely, the rate of inappropriate overuse of radiology tests for non-red flag LBP was consistently low, with all (n=18 measures) but two measure showing inappropriate overuse less than 25% of the time (see online supplementary file 4 table 1). One of these studies[68] estimated overuse to be about 50% but was conducted in 2001 and thus may reflect improvements over time. The other study is current but used a small sample size.[69] None of these studies were considered high risk of bias and thus results did not change on further analysis excluding high-risk studies.

Variation of inappropriateness against the same guideline recommendation

Eleven different guideline recommendations were studied more than once. There was significant heterogeneity

Figure 3 Rates of overuse. GORD, gastro-oesophageal reflux disease; GP, general practitioner; LBP, low back pain; NHMRC, National Health and Medical Research Council; NICE, National Institute for Health and Care Excellence; U/S, ultrasound; UTI, urinary tract infection.

(I^2 >50%) in nine of these pooled measures. Significant heterogeneity may have occurred for several reasons: (1) vastly different populations (for instance, one study measured the inappropriateness of upper endoscopy in Saudi Arabia[60] using the American Gastroenterological Association recommendations, whereas another study used the same recommendations in the USA[70]); (2) contrasting healthcare systems[71 72]; (3) relevance and applicability of one country's national guideline to another country[73]; (4) a low number of measures for meta-analysis[46]; and/or (5) significant heterogeneity, reflecting significant variation in inappropriate ordering.

DISCUSSION

There is marked variation in the rate of underuse and overuse of diagnostic tests from many primary care settings across the world. This variation suggests improvement can be made in the rate of appropriate diagnostic test ordering.

Primary care use of echocardiography is consistently poor. Echocardiography is inappropriately underused for some clinical situations, for example, confirming a diagnosis of heart failure, and inappropriately overused in others, for example, perioperative assessment. This was consistent across the countries where appropriateness of echocardiogram has been studied. This is of concern given the expertise and resource requirements to perform the test and the increasing availability of direct access ordering for primary care physicians.

For four tests, we found marked variation in the rate of inappropriate use. Underuse of PFTs varied by >50%, whereas overuse of urinary cultures, upper endoscopy and colonoscopy all varied by around 40%.

Radiology tests for both non-red flag LBP and non-red flag headache were frequently *not* overused, but the rate of overuse of imaging for non-red flag headache showed concerning trends, more than doubling from 2000 to 2010 (see online supplementary file 4 table 1).

Implications and future research

Two principle conclusions can be drawn from our results: (1) ordering of echocardiograms from primary care appears to require improvement and (2) markedly varying rates of inappropriate use for PFTs (underuse), colonoscopy (overuse), upper endoscopy (overuse) and urinary cultures (overuse) suggest that ordering can be improved.

Future research should focus on: determining the reasons for deviation from guidelines, assessing the quality of guidelines supporting diagnostic test use and systematic reviews quantifying inappropriate screening and monitoring tests. Furthermore, investigators wishing to undertake primary studies measuring inappropriate use should focus on developing objective data extraction methods for assessing patient notes and define clearly the interval they (investigators) will consider a test ordered for a particular symptom or disease.

Strengths in relation to other studies

Compared with other studies of inappropriate use of healthcare resources, we used data from real clinical encounters. This allowed a more robust assessment of diagnostic test inappropriateness, where other studies used surveys and hypothetical clinical vignettes.[19 74 75] Furthermore, we quantified the appropriateness of all types of diagnostic tests, rather than focusing on a specific test or specific disease (such as only laboratory tests[29]). Our paper is the first systematic review of studies that measured inappropriateness of all diagnostic tests ordered from primary care. Zhi *et al*[29] quantified the mean rates of overuse and underuse of laboratory tests in secondary care and focused on quantifying an overall rate of overuse and underuse. They estimated that overuse and underuse of laboratory tests was around 21% and 45%, respectively.[29] We choose not to quantify an overall rate of overuse and underuse because we feel the results would not be representative; we would be combining data from multiple different healthcare settings and data captured only the studied selection of diagnostic tests available in primary care.

Our use of guideline recommendations as the metric of appropriateness allowed a direct measure of diagnostic test appropriateness. Other studies that have assessed temporal and geographical variation in the use of diagnostic tests[76 77] have noted substantial differences in diagnostic practices across different regions, irrespective of disease prevalence and patient characteristics.[77] These studies, however, could not quantify what percentage of the temporal increase in the use of a diagnostic test is inappropriate and what percentage of variation between regions is inappropriate. We have quantified the percentage of inappropriate testing.

Although beyond the scope of our review, ultimately, interventions should be implemented to improve test use. A 2015 systematic review[78] concluded that 'Interventions such as educational strategies, feedback and changing test order forms may improve the efficient use of laboratory tests in primary care'. Thus, doctors, academics and policy makers can use our results to identify diagnostic tests in their particular healthcare settings that may benefit from intervention.

Limitations

The use of guidelines to quantify appropriateness of diagnostic tests could be considered a limitation of this study. Guidelines are often criticised for varying quality[25–27 79] and panel members' conflicts of interests.[80] However, clinical practice guidelines have been shown to improve both care outcomes and processes of care,[24] allow assessment of care on a population level, inform health policy,[81 82] set the standard of care across many healthcare settings[21 22] and provide a medicolegal framework.[23] One major medical insurance company advises that 'doctors must be prepared to explain and justify their decisions and actions, especially if they depart from guidelines produced by a nationally recognised body'.[23] Furthermore, guidelines have been used to measure appropriateness of the use of tests in other published peer-reviewed studies.[29] There will always be times when it is appropriate to depart from guidelines, but dramatic, consistent variation from guidelines requires investigation and is unlikely to be caused entirely by the quality of guidelines.

Furthermore, our study includes only a selection of diagnostic tests and is thus not an all-encompassing reflection of clinical practice. The data reflect the use of a specific test, sometimes for a particular clinical situation, in a particular country's healthcare system. Thus, policy makers and those interested in improving the quality of primary care diagnostic test use can use our results as a resource to identify tests in their healthcare setting that require improvement and/or investigation to decipher why such deviation from guidelines exists. Our conclusions from this paper, however, are not generalisable to all primary care settings nor all primary care diagnostic tests.

Lastly, caution must be taken when comparing results that measured inappropriateness using different denominators. The results from studies that measured inappropriateness using patients who had undergone a diagnostic test as a denominator should be interpreted differently to studies that used patients with a diagnosis or symptoms as a denominator (and vice versa).

CONCLUSION

There is marked variation in underuse and overuse of appropriate diagnostic test use in primary care across the world. From the available data, echocardiograms are ordered particularly poorly, while the substantial variation in appropriate ordering of PFTs, colonoscopy,

upper endoscopy and urinary cultures suggests a need for improvement.

Acknowledgements We would like to thank Kate Roche and Jason Hendry for comments on the draft and figures. We also thank the peer reviewers for their constructive feedback.

Contributors Conception and design: JWO, RP and CH. Search strategy: NR and JWO. Screening, extraction and risk of bias: JWO, AA and BDN. Analysis and interpretation of the data: JWO, RP, JA and CH. Drafting of the article: JWO (all authors critically reviewed and approved manuscript). Statistical expertise: RP. Clinical expertise: JWO, BDN, JA and CH. JWO is the guarantor.

Funding This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Data extracted from the included studies in this review are available on request from the corresponding author.

REFERENCES

1. Foot C, Naylor C, Imison C. The quality of GP diagnosis and referral. 2010. http://amapro.isabelhealthcare.com/pdf/Kings_Fund_Diagnosis_and_Referral_2010.pdf
2. Koch H, van Bokhoven MA, ter Riet G, *et al*. Ordering blood tests for patients with unexplained fatigue in general practice: what does it yield? Results of the VAMPIRE trial. *Br J Gen Pract* 2009;59:93–100.
3. Heneghan C, Glasziou P, Thompson M, *et al*. Diagnostic strategies used in primary care. *BMJ* 2009;338:b946.
4. Hobbs FDR, Bankhead C, Mukhtar T, *et al*. Clinical workload in UK primary care: a retrospective analysis of 100 million consultations in England, 2007-14. *Lancet* 2016;387:2323–30.
5. Centers for Disease Control and Prevention, National Center for Health Statistics. National ambulatory medical care survey: 2012 summary tables. 2012;5 http://www.cdc.gov/nchs/data/ahcd/namcs_summary/2010_namcs_web_tables.pdf
6. Alderwick H, Robertson R, Appleby J, *et al*. *Better value in the NHS The role of changes in clinical practice*, 2015.
7. Fisher ES, Bynum JP, Skinner JS. Slowing the growth of health care costs--lessons from regional variation. *N Engl J Med* 2009;360:849–52.
8. Appleby J, Thompson J, Jabbal J. Quarterly Monitoring Report: How is the NHS performing? *King's Fund* 2016:1–42.
9. Epner PL, Gans JE, Graber ML. When diagnostic testing leads to harm: a new outcomes-based approach for laboratory medicine. *BMJ Qual Saf* 2013;22 (Suppl):ii6–ii10.
10. Gandhi TK, Kachalia A, Thomas EJ, *et al*. Annals of Internal Medicine Article Missed and Delayed Diagnoses in the Ambulatory Setting. *Ann Intern Med* 2006;145:488–96.
11. Katzberg RW, Lamba R. Contrast-induced nephropathy after intravenous administration: fact or fiction? *Radiol Clin North Am* 2009;47:789–800.
12. Lumbreras B, Donat L, Hernández-Aguado I. Incidental findings in imaging diagnostic tests: a systematic review. *Br J Radiol* 2010;83:276–89.
13. Welch HG, Schwartz L, Woloshin S. Overdiagnosed: Making people sick in the pursuit of health. *Beacon Press* 2011;2011.
14. Moynihan R, Doust J, Henry D. Preventing overdiagnosis: how to stop harming the healthy. *BMJ* 2012;344:e3502.
15. Berwick DM, Hackbarth AD. Eliminating waste in US health care. *JAMA* 2012;307:1513.
16. Cecchini M, Lee S. Tackling wasteful spending on healthcare. 2017. http://networks.sustainablehealthcare.org.uk/sites/default/files/resources/Tackling Wasteful Spending on Health.pdf#page=117
17. Department of Health. *NHS 2010–2015: from good to great. Preventative, people-centred, productive*. London, 2010–2015.
18. Esmail A, Neale G, Elstein M, *et al*. *Case studies in litigation: claims reviews in four specialties*. Manchester, 2004.
19. Sirovich BE, Woloshin S, Schwartz LM. Too Little? Too Much? Primary care physicians' views on US health care: a brief report. *Arch Intern Med* 2011;171:1582–5.
20. Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS Med* 2010;7:e1000326.
21. Garber AM. Evidence-based guidelines as a foundation for performance incentives. *Health Aff* 2005;24:174–9.
22. Ransohoff DF, Pignone M, Sox HC, *et al*. How to decide whether a clinical practice guideline is trustworthy. *JAMA* 2013;309:139.
23. Fryar C. Doctors can depart from guidelines in patients' best interests. *BMJ* 2015;350:h841.
24. Grimshaw JM, Russell IT. Effect of clinical guidelines on medical practice: a systematic review of rigorous evaluations. *Lancet* 1993;342:1317–22.
25. 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.
26. Grilli R, Magrini N, Penna A, *et al*. Practice guidelines developed by specialty societies: the need for a critical appraisal. *Lancet* 2000;355:103–6.
27. Lenzer J. Why we can't trust clinical guidelines. *BMJ* 2013;346:f3830.
28. Spyridonidis D, Calnan M. Opening the black box: a study of the process of NICE guidelines implementation. *Health Policy* 2011;102:117–25.
29. Zhi M, Ding EL, Theisen-Toupal J, *et al*. The landscape of inappropriate laboratory testing: a 15-year meta-analysis. *PLoS One* 2013;8:e78962.
30. McGlynn E, Asch S, Adams J, *et al*. Quality of health care delivered to adults in the United States. *N Engl J Med* 2003;349:1866–8.
31. Sheldon TA, Cullum N, Dawson D, *et al*. What's the evidence that NICE guidance has been implemented? Results from a national evaluation using time series analysis, audit of patients' notes, and interviews. *BMJ* 2004;329:999.
32. National Health Service. NHS Imaging and Radiodiagnostic activity in England. 2013;7:1 http://www.england.nhs.uk/statistics/wp-content/uploads/sites/2/2013/04/KH12-release-2012-13.pdf
33. Moher D, Liberati A, Tetzlaff J, *et al*. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 2009;339:b2535.
34. Stroup DF, Berlin JA, Morton SC, *et al*. Meta-analysis of Observational Studies in Epidemiology. *JAMA* 2008;2000:283.
35. Wald NJ. Guidance on terminology. *J Med Screen* 2008;15:50.
36. Raffle A, Gray J. *Screening: Evidence and Practice*: Oxford University Press, 2007.
37. Glasziou P, Irwig L, Aronson J. *Evidence-based medical monitoring: from principles to practice*. Oxford UK: Blackwell Publishing, BMJ books, 2008.
38. Hoy D, Brooks P, Woolf A, *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.
39. Belletti D, Liu J, Zacker C, *et al*. Results of the CAPPS: COPD--assessment of practice in primary care study. *Curr Med Res Opin* 2013;29:957–66.
40. Bertella E, Zadra A, Vitacca M, *et al*. COPD management in primary care: is an educational plan for GPs useful? *Multidiscip Respir Med* 2013;8:24.
41. Chavez PC, Shokar NK. Diagnosis and management of chronic obstructive pulmonary disease (COPD) in a primary care clinic. *COPD* 2009;6:446–51.
42. Lange P, Rasmussen FV, Borgeskov H, *et al*. The quality of COPD care in general practice in Denmark: the KVASIMODO study. *Prim Care Respir J* 2007;16:174–81.
43. Ulrik CS, Sørensen TB, Højmark TB, *et al*. Adherence to COPD guidelines in general practice: impact of an educational programme delivered on location in Danish general practices. *Prim Care Respir J* 2013;22:23–8.
44. Barendregt JJ, Doi SA, Lee YY, *et al*. Meta-analysis of prevalence. *J Epidemiol Community Health* 2013;67:974–8.
45. Doi SA, Barendregt JJ, Khan S, *et al*. Advances in the meta-analysis of heterogeneous clinical trials I: The inverse variance heterogeneity model. *Contemp Clin Trials* 2015;45:130–8.

46. Higgins JP, Thompson SG, Deeks JJ, *et al*. Measuring inconsistency in meta-analyses. *BMJ* 2003;327:557–60.
47. Mafi JN, McCarthy EP, Davis RB, *et al*. Worsening trends in the management and treatment of back pain. *JAMA Intern Med* 2013;173:1573–81.
48. Mafi JN, Edwards ST, Pedersen NP, *et al*. Trends in the ambulatory management of headache: analysis of NAMCS and NHAMCS data 1999-2010. *J Gen Intern Med* 2015;30:548–55.
49. Williams CM, Maher CG, Hancock MJ, *et al*. Low back pain and best practice care: a survey of general practice physicians. *Arch Intern Med* 2010;170:271–7.
50. Cai JX, Campbell EJ, Richter JM. Concordance of outpatient esophagogastroduodenoscopy of the upper gastrointestinal tract with evidence-based guidelines. *JAMA Intern Med* 2015;175:1563–4.
51. Gurzun MM, Ionescu A. Appropriateness of use criteria for transthoracic echocardiography: are they relevant outside the USA? *Eur Heart J Cardiovasc Imaging* 2014;15:450–5.
52. van Gurp N, Boonman-De Winter LJ, Meijer Timmerman Thijssen DW, *et al*. Benefits of an open access echocardiography service: a Dutch prospective cohort study. *Neth Heart J* 2013;21:399–405.
53. Johnson JD, O'Mara HM, Durtschi HF, *et al*. Do urine cultures for urinary tract infections decrease follow-up visits? *J Am Board Fam Med* 2011;24:647–55.
54. Grover ML, Bracamonte JD, Kanodia AK, *et al*. Assessing adherence to evidence-based guidelines for the diagnosis and management of uncomplicated urinary tract infection. *Mayo Clin Proc* 2007;82:181–5.
55. Llor C, Rabanaque G, López A, *et al*. The adherence of GPs to guidelines for the diagnosis and treatment of lower urinary tract infections in women is poor. *Fam Pract* 2011;28:294–9.
56. Lindbäck H, Lindbäck J, Melhus Å. Inadequate adherence to Swedish guidelines for uncomplicated lower urinary tract infections among adults in general practice. *APMIS* 2017;125:816–21.
57. Leon P, Catherine K, Mark N, *et al*. Gastro-oseophageal reflux disease. The impact of guidelines on GP management. 2008.
58. Hughes-Anderson W, Rankin SL, House J, *et al*. Open access endoscopy in rural and remote Western Australia: does it work? *ANZ J Surg* 2002;72:699–703.
59. Aljebreen AM, Alswat K, Almadi MA. Appropriateness and diagnostic yield of upper gastrointestinal endoscopy in an open-access endoscopy system. *Saudi J Gastroenterol* 2013;19:219–22.
60. Azzam NA, Almadi MA, Alamar HH, *et al*. Performance of American Society for Gastrointestinal Endoscopy guidelines for dyspepsia in Saudi population: prospective observational study. *World J Gastroenterol* 2015;21:637–43.
61. Elwyn G, Owen D, Roberts L, *et al*. Influencing referral practice using feedback of adherence to NICE guidelines: a quality improvement report for dyspepsia. *Qual Saf Health Care* 2007;16:67–70.
62. Cardin F, Zorzi M, Bovo E, *et al*. Effect of implementation of a dyspepsia and Helicobacter pylori eradication guideline in primary care. *Digestion* 2005;72:1–7.
63. Cardin F, Zorzi M, Terranova O. Implementation of a guideline versus use of individual prognostic factors to prioritize waiting lists for upper gastrointestinal endoscopy. *Eur J Gastroenterol Hepatol* 2007;19:549–53.
64. Hassan C, Bersani G, Buri L, *et al*. Appropriateness of upper-GI endoscopy: an Italian survey on behalf of the Italian Society of Digestive Endoscopy. *Gastrointest Endosc* 2007;65:767–74.
65. Fiorenza JP, Tinianow AM, Chan WW. The initial management and endoscopic outcomes of dyspepsia in a low-risk patient population. *Dig Dis Sci* 2016;61:2942–8.
66. Chan YM, Goh KL. Appropriateness and diagnostic yield of EGD: a prospective study in a large Asian hospital. *Gastrointest Endosc* 2004;59:517–24.
67. Chan TH, Goh KL. Appropriateness of colonoscopy using the ASGE guidelines: experience in a large Asian hospital. *Chin J Dig Dis* 2006;7:24–32.
68. Eccles M, Steen N, Grimshaw J, *et al*. Effect of audit and feedback, and reminder messages on primary-care radiology referrals: a randomised trial. *Lancet* 2001;357:1406–9.
69. Tahvonen P, Oikarinen H, Niinimäki J, *et al*. Justification and active guideline implementation for spine radiography referrals in primary care. *Acta Radiol* 2017;58:586–92.
70. Majumdar SR, Soumerai SB, Farraye FA, *et al*. Chronic acid-related disorders are common and underinvestigated. *Am J Gastroenterol* 2003;98:2409–14.
71. Basu S, Andrews J, Kishore S, *et al*. Comparative performance of private and public healthcare systems in low- and middle-income countries: a systematic review. *PLoS Med* 2012;9:e1001244.
72. Ridic G, Gleason S, Ridic O. Comparisons of health care systems in the United States, Germany and Canada. *Mater Sociomed* 2012;24:112–20.
73. Gagliardi AR, Brouwers MC. Do guidelines offer implementation advice to target users? A systematic review of guideline applicability. *BMJ Open* 2015;5:e007047.
74. Kachalia A, Berg A, Fagerlin A, *et al*. Overuse of testing in preoperative evaluation and syncope: a survey of hospitalists. *Ann Intern Med* 2015;162:100–8.
75. Swennen MH, Rutten FH, Kalkman CJ, *et al*. Do general practitioners follow treatment recommendations from guidelines in their decisions on heart failure management? A cross-sectional study. *BMJ Open* 2013;3:e002982.
76. Parker L, Levin DC, Frangos A, *et al*. Geographic variation in the utilization of noninvasive diagnostic imaging: national medicare data, 1998-2007. *AJR Am J Roentgenol* 2010;194:1034–9.
77. Song Y, Skinner J, Bynum J, *et al*. Regional variations in diagnostic practices. *N Engl J Med* 2010;363:45–53.
78. Cadogan SL, Browne JP, Bradley CP, *et al*. The effectiveness of interventions to improve laboratory requesting patterns among primary care physicians: a systematic review. *Implement Sci* 2015;10:167.
79. Burgers JS, Fervers B, Haugh M, *et al*. International assessment of the quality of clinical practice guidelines in oncology using the Appraisal of Guidelines and Research and Evaluation Instrument. *J Clin Oncol* 2004;22:2000–7.
80. Gale EA. Conflicts of interest in guideline panel members. *BMJ* 2011;343:d5728.
81. IoM C to A the PHS on CPG. *Clinical practice guidelines: directions for a new program*. Washington, 1990.
82. Browman GP, Snider A, Ellis P. Negotiating for change. The healthcare manager as catalyst for evidence-based practice: changing the healthcare environment and sharing experience. *Healthc Pap* 2003;3:10–22.

EBM Critical Appraisal Worksheet—SYSTEMATIC REVIEW

Questions	Article	Pts
1. Did the review ask a clearly focused clinical question? What is the question?		3
2. Is the intervention feasible?		2
ARE THE RESULTS VALID?		
3. Did the reviewers try to identify all relevant studies? *Describe their approach.*		4
4. Were the criteria used to select articles for inclusion done *a priori*, clearly stated, and appropriate? *Describe*.		4
5. Did the reviewers perform an official validity assessment of the included studies? What was done? Is it reproducible?		4
6. Were the results similar from study to study? Did they use an appropriate analysis model? *Explain*.		4
7. Were the populations, interventions, outcomes, and outcome measurements combined in a way that makes sense? *Why or why not?*		4
8. Could publication bias have occurred? *How do you know?*		2
WHAT ARE THE RESULTS?		
9. What are the main results? How are they presented? How precise are they? *(What is the confidence interval? p-value?)*		4
10. Were all clinically relevant outcomes considered? *Discuss*.		5
11. Are the results generalizable? *Why or why not?*		3
WILL THE RESULTS HELP ME IN CARING FOR MY PATIENTS?		
12. Are the results statistically *and* clinically significant? *Explain*.		4
13. Are the benefits worth the harms and costs? *Why or why not?*		4
What level of evidence would you assign to this article and why?		3

Adapted from: Andrew D. Oxman, Deborah J. Cook, Gordon H. Guyatt, for the Evidence Based Medicine Working Group. Users' Guide to the Medical Literature. How to Use an Article About Overview. JAMA. (1994 Nov 2;272(17):1367-71).

04/24/16