OPEN ACCESS

Overuse of diagnostic testing in healthcare: a systematic review

Joris L J M Müskens, Rudolf Bertijn Kool, Simone A van Dulmen, Gert P Westert

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjqs-2020-012576).

IQ healthcare, Radboud University Medical Center, Radboud Institute for Health Sciences, Nijmegen, The Netherlands

Correspondence to
Mr Joris L J M Müskens, IQ Healthcare, Radboudumc, Nijmegen 6500 HB, The Netherlands;
Joris.muskens@radboudumc.nl

Received 19 October 2020
Revised 8 April 2021
Accepted 19 April 2021

To cite: Müskens JLJM, Kool RB, van Dulmen SA, *et al*. *BMJ Qual Saf* Epub ahead of print: [please include Day Month Year]. doi:10.1136/bmjqs-2020-012576

ABSTRACT

Background Overuse of diagnostic testing substantially contributes to healthcare expenses and potentially exposes patients to unnecessary harm. Our objective was to systematically identify and examine studies that assessed the prevalence of diagnostic testing overuse across healthcare settings to estimate the overall prevalence of low-value diagnostic overtesting.

Methods PubMed, Web of Science and Embase were searched from inception until 18 February 2020 to identify articles published in the English language that examined the prevalence of diagnostic testing overuse using database data. Each of the assessments was categorised as using a patient-indication lens, a patient-population lens or a service lens.

Results 118 assessments of diagnostic testing overuse, extracted from 35 studies, were included in this study. Most included assessments used a patient-indication lens (n=67, 57%), followed by the service lens (n=27, 23%) and patient-population lens (n=24, 20%). Prevalence estimates of diagnostic testing overuse ranged from 0.09% to 97.5% (median prevalence of assessments using a patient-indication lens: 11.0%, patient-population lens: 2.0% and service lens: 30.7%). The majority of assessments (n=85) reported overuse of diagnostic testing to be below 25%. Overuse of diagnostic imaging tests was most often assessed (n=96). Among the 33 assessments reporting high levels of overuse (≥25%), preoperative testing (n=7) and imaging for uncomplicated low back pain (n=6) were most frequently examined. For assessments of similar diagnostic tests, major variation in the prevalence of overuse was observed. Differences in the definitions of low-value tests used, their operationalisation and assessment methods likely contributed to this observed variation.

Conclusion Our findings suggest that substantial overuse of diagnostic testing is present with wide variation in overuse. Preoperative testing and imaging for non-specific low back pain are the most frequently identified low-value diagnostic tests. Uniform definitions and assessments are required in order to obtain a more comprehensive understanding of the magnitude of diagnostic testing overuse.

INTRODUCTION

In modern medicine, diagnostic tests, including laboratory tests, imaging and more invasive procedures, figure prominently in clinical decision making surrounding a new diagnosis.[1 2] However, the use of a diagnostic test is not always appropriate, as it may generate false positives, produce downstream cascades of more testing, expose patients to radiation or other harms, and create unnecessary patient anxiety, and could therefore be considered of low value.[3–7] Recent studies show that low-value diagnostic tests are still widely used and account for a substantial portion of the total amount of low-value healthcare expenses.[8–12] However, despite the potential avoidance of both costs and patient harms, the full quantification of low-value diagnostic testing has been difficult to achieve.

Understanding the prevalence of low-value diagnostic testing is essential to spur doctors, health systems and policymakers to take action to reduce its use. Most assessments of low-value diagnostic testing to date have been performed in the USA, Canada and Australia.[5 13–17] Only a few assessments have been completed in Europe.[18–21] Although multiple assessments of diagnostic testing overuse exist, only a small fragment of the problem has been uncovered.

One systematic review assessing the prevalence of diagnostic testing overuse and underuse in the primary care setting has previously been published by O'Sullivan *et al*.[22] Previous assessments demonstrate that overtesting is not limited to the primary care setting.[16 18 19] We therefore chose not to limit our study to one healthcare setting but rather to include all assessments of overtesting irrespective of the healthcare setting in which they were conducted. Furthermore, it is often hard to distinguish primary from secondary care practices due to differences in definitions of primary and secondary care procedures between countries and healthcare systems. In the present study, we therefore chose a more specific approach to the examination of the problem of low-value diagnostic testing compared with O'Sullivan. We narrowed down our

scope to studies of similar study design that only quantified the overuse, and not underuse, of diagnostic testing (eg, overtesting) using database data and used guidelines to distinguish appropriate from inappropriate testing to obtain a more uniform overview of the problem. This review might help policymakers and healthcare providers in their efforts to reduce overuse of diagnostic testing and can also help identify new knowledge gaps.

METHODS

This systematic review was performed and is reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines[23] (online supplemental file 1) and Meta-analyses of Observational Studies in Epidemiology statements,[24] no protocol has been registered. PubMed, Web of Science, and Embase were searched on 18 February 2020, for studies, of any design, assessing overuse of diagnostic tests. We did not restrict our search with respect to publication start date. The search can be summarised as (Medical Overuse OR Low-value care OR wasteful care OR wasteful healthcare) AND (Diagnosis) AND (Variation OR Volume OR Prevalence OR Frequency) (see online supplemental file 2 for the full strategy). We limited our search to human studies and studies published in English. The reference list of each included study was also searched for potentially relevant studies.

Study selection

Full texts were independently screened for eligibility by two reviewers. We included studies that quantified the overuse of diagnostic tests using database data, described a prevalence assessment and mentioned the relevant guideline(s) used to distinguish appropriate from inappropriate diagnostic testing. For the purpose of this study, we defined low-value diagnostic testing (or overtesting) as the overuse of diagnostic practices which are unlikely to benefit the patient given the harms, cost, available alternatives or preferences of the patient.[25] We excluded studies that did not quantify or assess provision of low-value diagnostic services; measured against a local guideline only (eg, did not use a guideline published by a government or professional society, but rather used a guideline that is only applicable locally); used survey data as the principal data source; were not published in English; used data derived from countries not in the Organisation for Economic Co-operation and Development (OECD); were intervention studies; or assessed (non-diagnostic) routine (population) screening tests as defined by Wald and Law[26] We only included studies using data from countries that are part of the OECD because of the comparability of the social–economic characteristics of the populations. Disagreements regarding the eligibility of studies were discussed by three of the study authors until consensus was reached.

Data extraction

The following information was extracted from each article: author and year, country, study population demographics (age/sex), the guideline used to determine the (in)appropriateness, the low-value care definition used, data sources, collected parameters, healthcare setting in which the assessment was conducted (primary/secondary care/both/unclear), type of low-value diagnostic test examined and the study outcome (prevalence estimate(s)). Assessments of diagnostic imaging procedure(s) were assigned to one of six categories based on the imaging modality they examined: cardiac test, combination, endoscopy, scan, ultrasound and X-ray (see online supplemental file 3) for an overview of the different imaging modalities in each category). The combination category contains assessments of multiple imaging modalities, but which did not report the individual outcomes for each included modality. For example, some studies examined the use of X-ray, MRI and CT in the examination of low back pain but did not report the individual outcome measures for the different modalities, but solely reported the combined outcome. When studies contained assessments for more than one unique diagnostic test, data for each test were collected and presented as an individual assessment. In case assessments were carried out over multiple time periods, only the data from the most recent time period were extracted. Each of the extracted assessments was assigned an assessment lens based on the classification proposed by Chalmers *et al* in 2017.[27] Chalmers *et al* concluded that different lenses are used to assess low-value care, each of which produces distinct outcomes. In general, Chalmers *et al* distinguishes two types of lenses that are used: service-centric and patient-centric lenses. Assessments using the service lens focus on the proportion of diagnostic tests that are of low value, while assessments using the patient lens focus on the proportion of the patient population that received the low-value diagnostic tests. Assessments using a patient centric lens can be further subdivided into assessments using either a patient-indication or patient-population lens. Which of the two patient centric lenses is applicable depends on the type of denominator that is used.[27] Assessment using a *patient-indication lens* only include patients with a specific indication in their denominator, while assessments using a *patient-population lens* include the entire cohort in their denominator.

The process of assigning lenses to the different assessments was performed in the following manner. One of the authors drafted an initial proposal regarding the applicable lens for each of the included assessments. This proposal was then critically appraised by two other authors, which was followed by multiple rounds of discussion until all authors agreed on the lens used.

Quality assessment

Risk of bias was assessed by three researchers using a modified version of the Hoy risk of bias tool. The Hoy risk of bias tool is a validated tool for the assessment of both internal and external validity of prevalence studies.[28] The tool was modified in the following manner:

1. Three domains (points 4, 7 and 9) from the original tool were found to be not applicable with respect to the identified studies. These domains either required information which is not applicable to retrospective research involving (electronic) database data or examined study designs which were not included in our study. Domain 7 was considered to be not applicable, since we did not grade the underlying evidence of the guidelines used in each of the included assessments. These domains were therefore removed after internal discussion among three authors. Online supplemental file 4 contains the original and modified tool, including more detailed reasoning for removal of each of the three domains.
2. The wording was adjusted to reflect the prevalence of low-value diagnostic testing instead of the prevalence of disease. Studies were considered at high risk of bias when they scored at least two 'high' and one 'unclear' among the seven risk of bias criteria. The original Hoy risk of bias tool does not provide a definition of high risk of bias. We therefore decided to use the aforementioned cut-off value for high risk after internal discussion among the authors. The process of grading risk of bias was similar to the one we used to assign lenses to the different assessments. One author drafted an initial proposal regarding the risk of bias scores of the included studies, which was followed by critical appraisal by two other authors and followed by multiple rounds of discussion until consensus was reached regarding the risk of bias score for each of the studies.

Statistical analysis

The primary outcome of this study is the prevalence of overuse of diagnostic tests across all healthcare settings. Descriptive statistics and median prevalences were calculated across all assessments for diagnostic imaging, laboratory testing and electroencephalogram categories and for the different assessment lenses within those categories. Analysis was performed using R V.3.6.3,[29] and data visualisation was done using the R package ggplot2.[30] Random-effect meta-analysis with 95% CIs (Clopper-Pearson), according to the DerSimonian and Laird method, was performed on similar assessments applying the same lens using the Meta[31] and Metafor package.[32] Variance was stabilised using the double arcsine transformation. The among-study heterogeneity was assessed using the I^2 statistic. The I^2 statistic represents the percentage of total variation across studies that is attributable to heterogeneity rather than change.[33 34] When applicable, data from similar assessments were pooled based on the lens that was used to assess the prevalence.

RESULTS

Article characteristics

Our search strategy identified 2542 articles. Of these, 2459 were excluded based on the title or abstract. Thirty-four studies met the eligibility criteria and were included. One additional eligible study was identified through screening of the reference lists of the included studies. A PRISMA flow diagram of the selection procedure is shown in figure 1. From the included studies, seven conducted their assessments in the primary care setting (7/35), five in the secondary setting (5/35) and nine in both settings (9/35). The remaining 14 studies (14/35) did not provide a clear indication as to the setting in which their assessments were conducted and therefore labelled as unclear (also see online supplemental file 6). The included studies were conducted in eight different countries and contained 118 assessments of low-value diagnostic tests. Most studies were conducted in the USA (n=23). The 118 identified assessments are divided into imaging procedures (n=96) and other diagnostic tests (n=22), which included laboratory tests (n=19), and electroencephalography procedures (n=3) (as shown in table 1). The majority of the assessments used a patient-indication lens (n=67, 57%), followed by the service lens (n=27, 23%) and patient-population lens (n=24, 20%). Among the studies included, three studies assessed overuse among different insurance populations,[35–37] and one study assessed overuse across two different time periods.[38] Of note, since we were interested in the most recent measurements of low-value diagnostic overtesting, we decided to only include the most recent measurements from the study by Flaherty *et al*.[38]

Risk of bias

Using the Hoy risk of bias tool, we assessed the risk of bias of the included studies based on eight criteria (online supplemental file 4 contains the used modified Hoy risk of bias tool). Assessment of risk of bias revealed 25 studies as low risk of bias and 10 studies as high risk of bias (eg, scoring at least two categories high and one unclear). Almost all studies graded as high risk of bias, scored as being of high risk on the following two criteria: 'the examined population being a close representation of the national population' and 'the use of a clear case definition of the low-value diagnostic test examined' (online supplemental file 5 contains a detailed description of the risk of bias assessment outcome.

Overuse of diagnostic tests

Online supplemental file 6 provides an overview of the studies, characteristics and outcomes. Prevalence estimates of diagnostic testing overuse ranged from 0.09% to 97.5% (median prevalence of assessments using a patient-indication lens: 11.0%, a patient-population lens: 2.0% and a service lens: 30.7%). The majority of included assessments of low-value

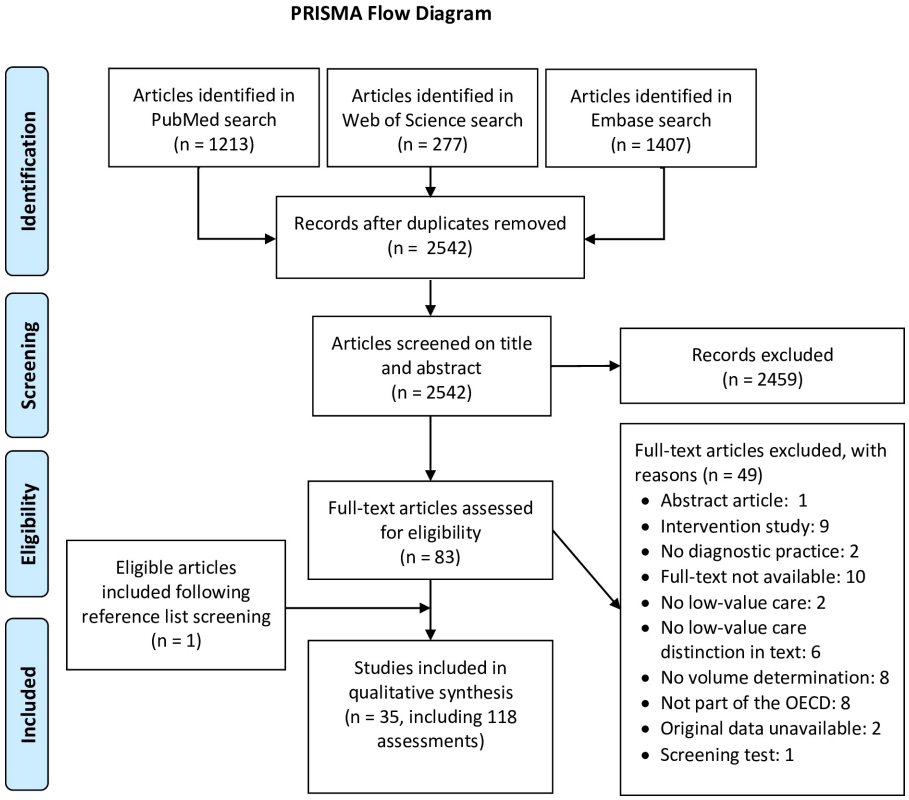


Figure 1 PRISMA Flow-diagram. OECD, Organisation for Economic Co-operation and Development ; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

diagnostic testing (n=85) report overuse to be below 25%. Among the 33 assessments reporting high levels of overuse (≥25%), imaging for uncomplicated low back pain (n=6) and preoperative testing (n=7), such as preoperative baseline lab tests, echocardiography or (cardiac) stress tests,were most commonly assessed. Overuse of diagnostic imaging procedures was most often assessed (n=96), with prevalence of overuse varying between 0.09% and 97.5% (median prevalence of assessments using a patient-indication lens: 11.2%, a patient-population lens: 1.2% and a service lens: 22.0%), as shown in figure 2. Prevalence assessments in the 'other diagnostic tests' category (n=22) varied between 0.10% and 78.6%, as shown in figure 3A,B. This category contained two distinct categories: laboratory tests (n=19, median prevalence of assessments using a patient-indication lens: 16.3%, a patient-population lens: 3.5 % and a service lens: 47.5%: 14.0%) and electroencephalography (n=3, median prevalence of assessments using a patient-indication lens: 0.2% and a patient-population lens: 0.1%).

The highest prevalence of overuse was reported in the following five diagnostic practices: use of electrocardiograms, chest X-rays or pulmonary function tests in low-risk patients having low-risk surgery (97.5%); imaging for low back pain within the first 6 weeks of symptom onset in the absence of red flags (86.2%); knee arthroscopy for meniscal derangements (81.7%); baseline lab tests for low-risk patients receiving low risk surgery (78.6%); and knee arthroscopy for osteoarthritis (71.7%). Overall, imaging in case of non-specific low back pain (15/118) and preoperative tests (14/118), such as preoperative baseline lab tests, echocardiography or exercise stress tests, were most often assessed diagnostic practices identified in this study. Figures 2 and 3 show that a large variation in assessment outcomes of similar diagnostic tests, irrespective of assessment lens used, exists. For example, Bouck *et al*,[39] Schwartz *et al*[13] and Mafi *et al*[17] yielded vastly different results in their respective studies. Bouck *et al*[39] used a patient-indication lens and reported 30.70% of the identified imaging procedures to be considered as overuse, while Schwartz *et al*[13] used a patient-population lens and found 4.1% to be considered as overuse. On the other hand, Mafi *et al* used a service lens in their assessment and reported the level of overuse to be 86.2%.

Variation among assessments of similar procedures

For the two types of diagnostic tests, multiple assessments using similar lenses were identified among the included studies. These included short-interval repeat bone densitometry testing (dual-energy X-ray absorptiometry) and the use of imaging procedures for non-specific low back pain. Considerable heterogeneity was observed between the extracted assessments for both groups ($I^2 \geq 100\%$) (see online supplemental file

Table 1 Overview of study characteristics

Countries where the studies were conducted	Studies, n (%)
Australia	3 (9)
Austria	1 (3)
Canada	4 (11)
Italy	1 (3)
Netherlands	1 (3)
Spain	1 (3)
Switzerland	1 (3)
USA	23 (66)
Total	35 (100)
Type of diagnostic test	**Assessments, n (%)**
Imaging	96 (81)
Cardiac test	14 (12)
Combination	14 (12)
Endoscopy	11 (9)
Scan	34 (29)
Ultrasound	6 (5)
X-ray	17 (14)
Other diagnostic tests	22 (19)
Laboratory tests	19 (16)
Electroencephalography	3 (3)
Total	118 (100)
Type of assessment lens used	**Assessments, n (%)**
Patient indication	67 (57)
Patient population	24 (20)
Service	27 (23)
Total	118 (100)

7 for the generated forest plot). We therefore chose to forgo generating pooled estimates since pooling heterogenous studies could lead to invalid results. In particular, assessments of overuse of imaging for non-specific low back pain showed substantial variation, irrespective of the assessment lens used.

DISCUSSION

In this systematic review, we identified and summarised the outcomes of studies assessing the prevalence of overuse of diagnostic tests. The majority of the 118 identified assessments examined the overuse of diagnostic imaging procedures (n=96), followed by the category other diagnostic tests, which included laboratory tests (n=19) and electroencephalography tests (n=3). Assessments of low-value diagnostic testing using a patient-indication lens were most common (n=67, 57%), followed by assessments that used a service lens (n=27, 23%) and the patient-population lens (n=24, 20%). Major variation between prevalence estimates was observed, irrespective of the assessment lens used. Prevalence estimates of diagnostic testing overuse ranged from 0.09% to 97.5% (median prevalence of assessments using a patient-indication lens: 11.0%, a patient-population lens: 2.0% and a service lens: 30.7%), although 85 of the included assessments reported the prevalence of overuse to be below 25%. Among the 33 assessments reporting high levels of overuse (ie, ≥25%), multiple assessments exploring the overuse of imaging for uncomplicated low back pain (five assessments) or preoperative tests (seven assessments) were present. Additionally, 11 of the 33 measurements reporting high levels of overuse were extracted from eight studies considered at high risk of bias. Similar to the review of O'Sullivan *et al*,[22] we found substantial variation in overuse among diagnostic services. However, our study adds to this finding by illustrating that variation is not limited to the primary care setting. Substantial overuse of diagnostic testing was also observed among diagnostic services often used in the secondary care setting, such as short interval of bone mineral density testing or non-indicated cardiac testing before low-risk surgery. Through implementation of the concept of the assessment lenses to the included assessments, as proposed by Chalmers *et al*,[27] we were able to better compare the different assessment outcomes for similar diagnostic tests. Comparison of the different assessment outcomes regarding similar tests revealed that the observed variation could in part be explained by the use of different assessment lenses, an aspect which O'Sullivan *et al* did not account for in their study.[22] Furthermore, we found that distinguishing primary from secondary care practices is often difficult and not always straightforward. Reasons are that many diagnostic practices are often provided in both the primary and secondary settings, and the setting in which these practices are provided often differs between countries and their respective healthcare systems.

For the two types of low-value diagnostic testing, that is, short-interval repeat bone densitometry testing and imaging for non-specific low back pain without the presence of red flags, several similar assessments were extracted from the included literature. We tried to pool those similar assessments but refrained from doing so after observing significant among-study heterogeneity ($I^2 \geq 100\%$). We therefore chose to report the results of the individual studies instead. The high levels of heterogeneity observed warrant further examination through means of subgroup analysis. However, the examination of potential sources of heterogeneity was hampered by the limited number of assessments present in each group. The limited number of assessments in each group also prevented us from reliably testing for publication bias.[40] Although we could not examine the heterogeneity through means of statistical subgroup analysis, we have tried to find possible explanations for the observed heterogeneity in the available literature and comparison of the studies. As mentioned before, substantial variation among the extracted assessments of overuse was observed among the assessments included in our study. This variation could be caused by differences in study

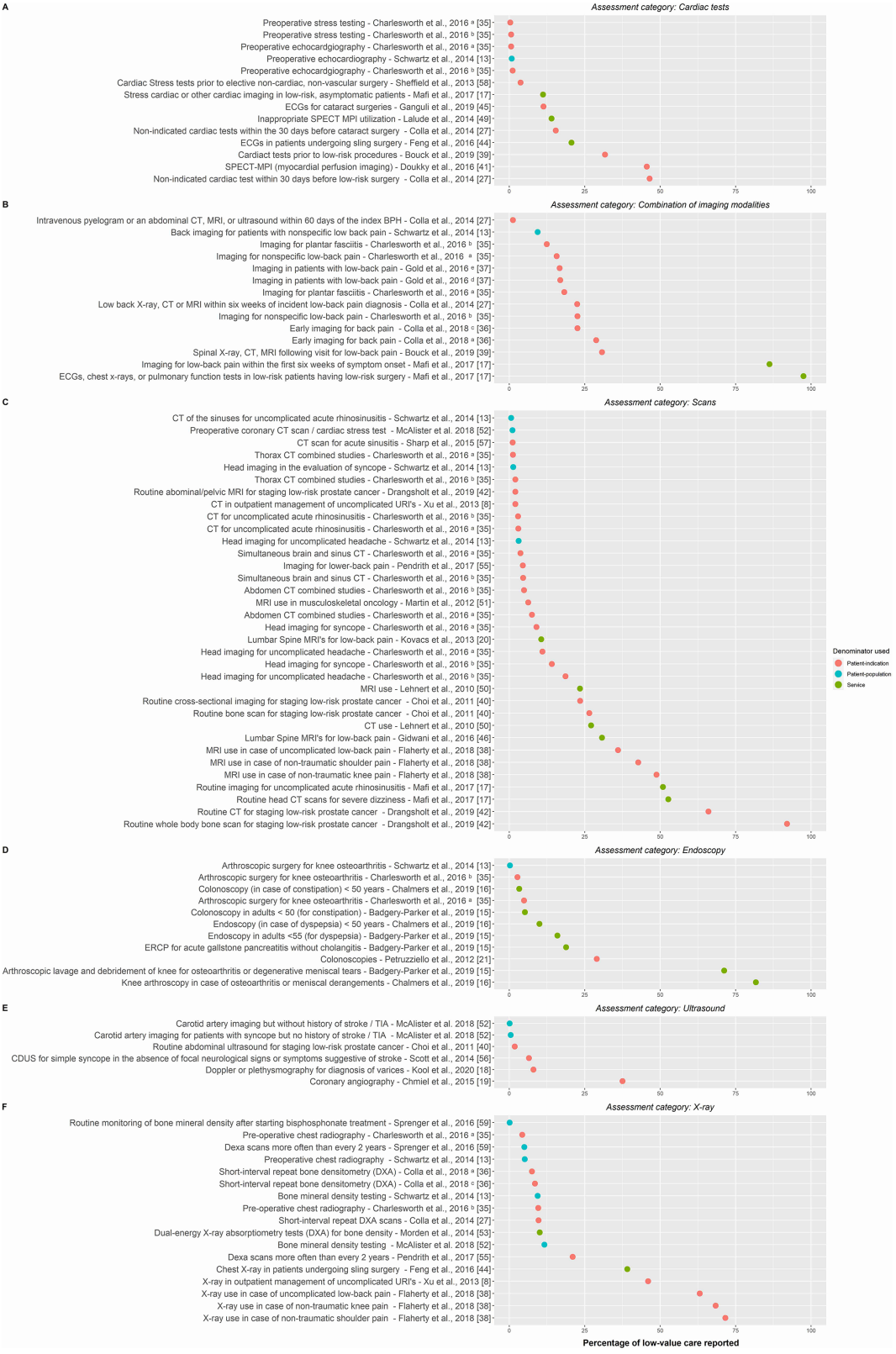


Figure 2 Assessment outcomes regarding the prevalence of low-value diagnostic tests for all assessments included in the diagnostic imaging category: (A) cardiac tests, (B) combination, (C) scans, (D) endoscopy, (E) ultrasound and (F) X-ray. Among the included studies, some studies contained multiple assessments undertaken in different cohorts. These assessments are distinguished by the following: (A) assessment performed among a commercially insured population, (B) assessment performed among Medicaid beneficiaries, (C) assessment performed among Medicare beneficiaries, (D) assessment performed using Kaiser Permanente EPIC Electronic Healthcare Records data, (E) assessment performed using data derived from the Oregon Community Health Information Network. BPH, benign prostatic hyperplasia; CDUS, colour duplex ultrasound scan; CKD, chronic kidney disease; ERCP, endoscopic retrograde cholangiography; PSA, prostate-specific antigen; SPECT MPI, single-photon emission CT–myocardial perfusion imaging; TIA, transient Ischaemic attack; URI, upper respiratory infection.

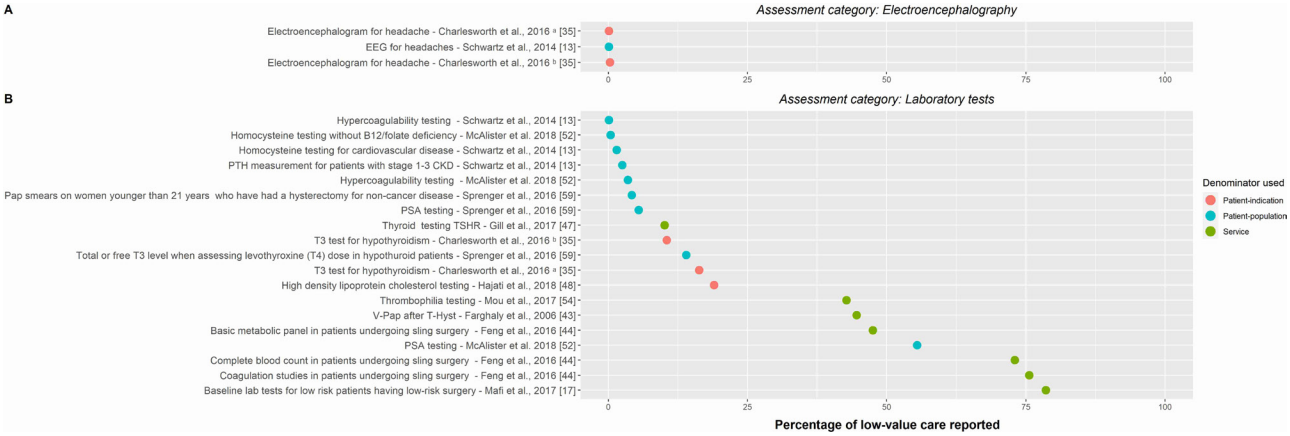


Figure 3 Assessment outcomes regarding the prevalence of low-value diagnostic tests for all assessments included in the other diagnostic tests category: A) laboratory tests and (B) electroencephalography tests. Among the included studies, some studies contained multiple assessments undertaken in different cohorts. These assessments are distinguished by the following: (A) assessment performed among a commercially insured population and (B) assessment performed among Medicaid beneficiaries. CKD, chronic kidney disease; EEG, electroencephalography; PSA, prostate-specific antigen; PTH, parathyroid hormone; T-Hyst, total hysterectomy; TSHR, thyroid-stimulating hormone reflexive testing; V-Pap, vaginal Pap smear.

design, cohort size or operationalisation of guidelines. Additionally, previous research has shown that factors such as population characteristics, healthcare systems and insurance systems can greatly affect the amount of overuse.[7 13 15 16 18 25 35 36 39] For example, the studies by both Bouck *et al*[39] and Pendrith *et al*[41] examined the overuse of imaging for low back pain in Canada. However, each study used different data sources (Patient-Level Physician Billing Repository, Discharge Abstract Database, National Ambulatory Care Reporting System versus Ontario Health Insurance Plan claims database, respectively) and therefore used different codes to identify the included cohort. Furthermore, Pendrith *et al*[41] included all visits to the primary care physician of adult patients (age>18 years) in their examination, while Bouck *et al*[39] included only the first family physician visit. Although such differences appear small, they can drastically alter the patients included in the cohort and therefore influence the final prevalence estimate. The observed differences in estimates could also be caused by differences in definitions of low-value diagnostic testing. Most assessments are based on recommendations derived from initiatives such as the National Institute for Health and Care Excellence (England) or Choosing Wisely. However, no standardised definitions of low-value procedures or assessments, for the specific countries, exist. The absence of standardised definitions for specific countries could result in different cohorts and thus different prevalence estimates.

Finally, the use of different methods to assess overuse can explain the observed differences in outcomes. Some articles used the method as proposed by Schwartz *et al*,[13] which proposes the use of narrow (high specificity and low sensitivity) and broad indicators (low specificity and high sensitivity) to assess low-value care.[13] Narrow definitions are more tightly formulated, resulting in a more distinct cohort of patients/services that is included as compared with the cohorts created using broad definitions. Through a combination of both assessments, a more complete understanding of the problem is obtained. However, while using both narrow and broad indicators appears to be a good way to provide an estimate of the amount of overuse of low-value practices, it was only employed in three of the included articles.[13 15 16] In our analysis, we only used the broad assessments from those studies since the underlying definitions of those more closely resembled the original recommendations. Therefore, broad assessment outcomes are more suitable for comparison to the outcomes of studies that directly used the relevant original recommendations in their assessments.

Strengths and limitations

A strength of this study is that we did not limit our review to a single type of diagnostic testing or disease. Additionally, we did not limit the search to a particular setting; as a result, we present prevalence estimates for a wide range of diagnostic tests across all healthcare settings. Furthermore, we included only direct measures of diagnostic testing overuse acquired from data collected from databases.

Our study also faces some limitations. First, we recognise that the measurement of low-value care is often biased. Most existing measurements of low-value care target practices that are easily measured using existing data. These measurements clearly distinguish high-value from low-value services. However, most guidelines do not provide such a clear distinction. Detailed clinical information is often required to accurately distinguish high-value from low-value care but is often not present in the available data.[13 36 42–45] Because of these reasons, only a relatively small part of the total amount of low-value services has been examined so far. Unfortunately, we

were unable to reliably test for publication bias due to the limited number of similar assessments which used the same scope present in our study.[40] Publication bias might be present among assessments of low-value practices because reports of the presence of substantial overuse are undesirable for most parties involved in such assessments. However, while our overview contains such a wide range of assessment outcomes, we have attempted to reduce the publication bias where possible. Second, although we attempted to include all relevant keywords in our search strategy, our strategy may have missed some relevant terms and thus overlooked some studies assessing overuse of diagnostic services. Additionally, we incorporated several terms, such as overuse and low-value care in our search, which have been added to the lexicon relatively recently. Also, our search strategy identifies only studies that explicitly acknowledge the examined tests as representing overuse or low-value care. It is therefore possible our search might have missed studies which did not use these terms yet or that included some appropriate services alongside inappropriate ones in their assessment. Third, we included only studies that assessed overuse in relation to a specific guideline. Although this is a commonly used criterion and seen as an objective method to assess overuse, it is prone to underestimation of the actual prevalence of the problem. Yet, there is a risk of missing patients who do not exactly fit the specific guideline(s) used or falsely classifying a test as (in)appropriate due to the clinical complexity of the patient involved. Furthermore, by requiring an assessment to be performed against a guideline, we did not capture all assessments of low-value diagnostic practices. Different methods are also used to distinguish appropriate from inappropriate care, such as expert opinion, Delphi or RAND appropriateness methods.[46] Because we included only the assessment to require a guideline, our study therefore does not capture the full scope of assessments of low-value diagnostic overtesting. Fourth, we used a modified version of the Hoy risk of bias tool.[28] This is a validated tool for the assessment of risk of bias in prevalence studies, although we had to slightly adjust it to make it suitable to our research. However, while we tried to keep the tool as original as possible, we do need to consider that the modifications made to the original tool might have affected the outcome of our risk of bias assessment. Lastly, each of the included studies used their own definition of overuse in their assessments. Due to these differences in definitions of overuse, it is often difficult to directly compare assessments of similar procedures since these differences are in part responsible for the differences in outcome. However, by assigning assessment lenses to the included assessment of similar practices, we were able to group assessments using similar definitions of overuse and compare those to one another.

Implications for practice and future research

Most studies included in our review were conducted in the USA, and only a few studies examining diagnostic testing overuse have been conducted in Europe. Findings from one country (such as the USA) are often not generalisable to other countries due to differences in (patient) population characteristics, healthcare and insurance systems. Additional assessments of overuse from different countries are needed to gain further insight into the magnitude of the overuse problem. Insight into the prevalence of diagnostic testing overuse is required to create a sense of urgency among (local) physicians and policymakers and to help develop effective strategies to tackle low-value diagnostic overtesting.[47 48] Assessments should be repeated to monitor the problem of overuse of diagnostic testing over time and the effects of implemented strategies and interventions. In our review, only one study assessed overuse across multiple time periods.[38] The overview of assessment outcomes generated in this review could be used by both policymakers and care providers as a source of inspiration for (future) assessments in their own organisation(s) and (subsequently) as comparison material for their assessment outcomes.

International agreement on low-value service definitions and standardisation assessment methods (eg, identical denominators, similar lenses and scopes) could contribute to prevalence estimates that are comparable across countries. An example would be the recently completed study which compared the overuse of laboratory testing in USA to that in Canada.[49] However, while it would certainly help to have unified definitions and methods for the assessment of low-value care, it would certainly be an ambitious goal to set. Hence, each of the different assessments included in this study were conducted in different contexts and with slightly different purposes in mind. However, what they all do have in common is that they were performed to gain insight into the (local) problem of low-value diagnostic practices. These assessments therefore are crucial first steps in the process of reducing low-value diagnostic practices (locally).

Lastly, it might be of interest to include cost estimates in future assessments because it is known that cost differences exist across countries and healthcare systems. Another reason why costs estimates would be of interest would be that previous research has indicated that low-cost services are predominantly overused.[17] We therefore suggest that future studies should include the associated costs of low-value diagnostic tests (possibly including additional downstream costs due to performance of low-value diagnostic tests) in their assessments. However, we would like to emphasise that while cost is an important argument in the discussion of addressing low-value testing, it is not the only and certainly not the most important potential harm of unnecessary testing.

CONCLUSION

This study shows that there is substantial overuse of diagnostic testing present across all healthcare settings, with much variation among similar diagnostic services. Preoperative testing and imaging for non-specific low back pain are the most frequently assessed and overused low-value diagnostic tests. Effective strategies to tackle the overuse of diagnostic testing must be developed and implemented by health systems, providers, policymakers and others. Additionally, more uniform definitions and assessments of low-value diagnostic tests are required in order to obtain a better understanding of the magnitude of diagnostic testing overuse.

Twitter Rudolf Bertijn Kool @tijnkool

Acknowledgements The authors would like to thank Eve A. Kerr, M.D., M.P.H. from the University of Michigan and Mandi L. Klamerus, M.P.H. from CCMR Center for Clinical Management Research, department of Veteran Affairs, for their critical review and contributions to the paper.

Contributors JM, RBK, SAvD and GPW all contributed to the design of the study and its development. JM, RBK and SAvD screened the articles and aided in the data extraction. JM took the lead in the principal analysis of the data and writing of the manuscript. RBK, SAvD and GPW contributed to the interpretation and presentation of the results. All authors provided critical feedback and helped shape the research, analysis and manuscript.

Funding This work was funded by ZonMW, the Dutch Organization for Health Research and Development (grant number 80-83920-98-803). ZonMW did not contribute to the design of the study, collection, analysis and interpretation of data and in writing the manuscript.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as supplementary information.

ORCID iDs
Joris L J M Müskens http://orcid.org/0000-0002-9440-7703
Rudolf Bertijn Kool http://orcid.org/0000-0003-3134-487X
Simone A van Dulmen http://orcid.org/0000-0003-4003-8540
Gert P Westert http://orcid.org/0000-0003-3744-8207

REFERENCES

1 Heneghan C, Glasziou P, Thompson M, *et al*. Diagnostic strategies used in primary care. *BMJ* 2009;338:b946.
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:e93–100.
3 McAlister FA, Lin M, Bakal J. Frequency of low-value care in Alberta, Canada: a retrospective cohort study. *BMJ Quality & Safety* 2018;27:340.
4 Berlin L. Overdiagnosed: making people sick in pursuit of health. *JAMA* 2011;305:1356–9.
5 Brownlee S, Chalkidou K, Doust J, *et al*. Evidence for overuse of medical services around the world. *Lancet* 2017;390:156–68.
6 Brodersen J, Schwartz LM, Heneghan C, *et al*. Overdiagnosis: what it is and what it isn't. *BMJ Evidence-Based Medicine* 2018;23:1–3.
7 Colla CH, Mainor AJ, Hargreaves C, *et al*. Interventions aimed at reducing use of low-value health services: a systematic review. *Med Care Res Rev* 2017;74:507–50.
8 Xu KT, Roberts D, Sulapas I, *et al*. Over-prescribing of antibiotics and imaging in the management of uncomplicated URIs in emergency departments. *BMC Emerg Med* 2013;13:7.
9 Xu S, Hom J, Balasubramanian S, *et al*. Prevalence and predictability of low-yield inpatient laboratory diagnostic tests. *JAMA Netw Open* 2019;2:e1910967.
10 Bruce Alexander C. Message from the President: reducing healthcare costs through appropriate test utilization. *Critical Values* 2015;5:6–9.
11 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.
12 Yeh DD. A clinician's perspective on laboratory utilization management. *Clinica Chimica Acta* 2014;427:145–50.
13 Schwartz AL, Landon BE, Elshaug AG, *et al*. Measuring low-value care in Medicare. *JAMA Intern Med* 2014;174:1067–76.
14 Elshaug AG. Over 150 potentially low-value health care practices: an Australian study. *Med J Aust* 2013;198:85.
15 Badgery-Parker T, Pearson S-A, Chalmers K, *et al*. Low-Value care in Australian public hospitals: prevalence and trends over time. *BMJ Qual Saf* 2019;28:205–14.
16 Chalmers K, Pearson S-A, Badgery-Parker T, *et al*. Measuring 21 low-value Hospital procedures: claims analysis of Australian private health insurance data (2010–2014). *BMJ Open* 2019;9:e024142.
17 Mafi JN, Russell K, Bortz BA, *et al*. Low-Cost, high-volume health services contribute the most to unnecessary health spending. *Health Aff* 2017;36:1701–4.
18 Kool RB, Verkerk EW, Meijs J, *et al*. Assessing volume and variation of low-value care practices in the Netherlands. *Eur J Public Health* 2020;30:236–40.
19 Chmiel C, Reich O, Signorell A, *et al*. Appropriateness of diagnostic coronary angiography as a measure of cardiac ischemia testing in non-emergency patients - a retrospective cross-sectional analysis. *PLoS One* 2015;10:e0117172.
20 Kovacs FM, Arana E, Royuela A, *et al*. Appropriateness of lumbar spine magnetic resonance imaging in Spain. *Eur J Radiol* 2013;82:1008–14.
21 Petruzziello L, Hassan C, Alvaro D, *et al*. Appropriateness of the indication for colonoscopy: is the endoscopist the 'gold standard'? *J Clin Gastroenterol* 2012;46:590–4.

22 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.
23 Moher D, Liberati A, Tetzlaff J, *et al*. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med* 2009;6:e1000097.
24 Stroup DF, Berlin JA, Morton SC, *et al*. Meta-Analysis of observational studies in epidemiology: a proposal for reporting. meta-analysis of observational studies in epidemiology (moose) group. *JAMA* 2000;283:2008–12.
25 Colla CH, Morden NE, Sequist TD, *et al*. Choosing wisely: prevalence and correlates of low-value health care services in the United States. *J Gen Intern Med* 2015;30:221–8.
26 Nicholas W, Malcolm L. *Medical screening Oxford textbook of medicine*. Oxford, UK: Oxford University Press.
27 Chalmers K, Pearson S-A, Elshaug AG. Quantifying low-value care: a patient-centric versus service-centric lens. *BMJ Qual Saf* 2017;26:855–8.
28 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.
29 R Core Team. *R: a language and environment for statistical computing. v3.6.3 ED*. Vienna, Austria: R foundation for Statistical Computing, 2019.
30 Wickham H. *ggplot2: elegant graphics for data analysis*. New York: Springer, 2009.
31 Balduzzi S, Rücker G, Schwarzer G. How to perform a meta-analysis with R: a practical tutorial. *Evid Based Ment Health* 2019;22:153–60.
32 Viechtbauer W. Conducting Meta-Analyses in *R* with the metafor Package. *J Stat Softw* 2010;36:1–48.
33 Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002;21:1539–58.
34 Higgins JPT*et al*. Measuring inconsistency in meta-analyses. *BMJ* 2003;327:557–60.
35 Charlesworth CJ, Meath THA, Schwartz AL, *et al*. Comparison of low-value care in Medicaid vs commercially insured populations. *JAMA Intern Med* 2016;176:998–1004.
36 Colla CH, Morden NE, Sequist TD, *et al*. Payer type and low-value care: comparing choosing wisely services across commercial and Medicare populations. *Health Serv Res* 2018;53:730–46.
37 Gold R, Esterberg E, Hollombe C. Low back imaging when not indicated: a descriptive Cross-System analysis. *Perm J* 2016;20:25–33.
38 Flaherty S, Zepeda ED, Mortele K. Magnitude and financial implications of inappropriate diagnostic imaging for three common clinical conditions. *Int J Qual Health Care* 2018.
39 Bouck Z, Pendrith C, Chen X-K, *et al*. Measuring the frequency and variation of unnecessary care across Canada. *BMC Health Serv Res* 2019;19:446.
40 Page MJ HJ, Sterne JAC. Chapter 13: Assessing risk of bias due to missing results in a synthesis. In: Higgins JPT TJ, Chandler J, Cumpston M, *et al*, eds. *Cochrane; 2020*, 2020.
41 Pendrith C, Bhatia M, Ivers NM, *et al*. Frequency of and variation in low-value care in primary care: a retrospective cohort study. *CMAJ Open* 2017;5:E45–51.
42 Elshaug AG, McWilliams JM, Landon BE. The value of low-value Lists. *JAMA* 2013;309:775–6.
43 Bhatia RS, Levinson W, Shortt S, *et al*. Measuring the effect of choosing wisely: an integrated framework to assess campaign impact on low-value care. *BMJ Qual Saf* 2015;24:523–31.
44 Morgan DJ, Leppin AL, Smith CD, *et al*. A practical framework for understanding and reducing medical overuse: Conceptualizing overuse through the Patient-Clinician interaction. *J Hosp Med* 2017;12:346–51.
45 Chalmers K, Badgery-Parker T, Pearson S-A, *et al*. Developing indicators for measuring low-value care: mapping choosing wisely recommendations to hospital data. *BMC Res Notes* 2018;11:163.
46 Kathryn F. The Rand/UCLA appropriateness method user's manual: Santa Monica : Rand, 2001 2001.
47 Parchman ML, Henrikson NB, Blasi PR, *et al*. Taking action on overuse: creating the culture for change. *Health Care* 2017;5:199–203.
48 Grimshaw JM, Patey AM, Kirkham KR, *et al*. De-implementing wisely: developing the evidence base to reduce low-value care. *BMJ Qual Saf* 2020;29:409–17.
49 Henderson J, Bouck Z, Holleman R, *et al*. Comparison of payment changes and choosing wisely recommendations for use of low-value laboratory tests in the United States and Canada. *JAMA Intern Med* 2020;180:524.

Supplementary file 1: PRISMA checklist.

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2-3
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	4
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	N.A.
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	No review protocol has been registered
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	5-6
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	5
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	5 & Supplementary file 2
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	5-6
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	6-7

Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	5-7
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	7-8 & Supplementary file 4
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	8
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	8
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	7-8
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	N.A.
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	9 & Figure 1
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	9-10 & Supplementary file 6
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	10 & Supplementary file 5
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	9-12, Supplementary file 6 & Figure 2, Figure 3
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	Supplementary file 7
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	10 & Supplementary file 5
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	N.A.

DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	13-15
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	15-17
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	17-19
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	21

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

Supplementary file 2: search strategy

1. Pubmed:

("Medical Overuse"[Mesh] OR Low-value care [tiab] OR low-value hospital care [tiab] OR low-value healthcare [tiab] OR wasteful care [tiab] OR wasteful healthcare [tiab] OR wasteful hospital care [tiab] OR overuse of healthcare [tiab] OR overuse procedure* [tiab] OR medical overuse [tiab] OR inappropriate healthcare [tiab] OR inappropriate care [tiab] OR unwanted healthcare [tiab] OR unwanted care [tiab] OR unnecessary healthcare [tiab] OR unnecessary care [tiab] OR Overdiagnos* [tiab] OR Over diagnos* [tiab] OR ineffective care [tiab] OR ineffective healthcare [tiab]) AND ("Diagnosis"[Mesh] OR diagnos* [tiab]) AND "statistics and numerical data" [Subheading] OR Variation* [tiab] OR Volume* [tiab] OR Prevalence* [tiab] OR Cost [tiab] OR costs [tiab] OR Frequenc* [tiab] AND "Guidelines as Topic"[Mesh] OR Guideline* [tiab] OR Choosing Wisely [tiab] OR policy [tiab] OR policies [tiab]

2. Web of science:

TOPIC: (“Low-value care” OR “low-value hospital care” OR “low-value healthcare” OR “wasteful care” OR “wasteful healthcare” OR “wasteful hospital care” OR “overuse of healthcare” OR “overuse procedure*” OR “medical overuse” OR “inappropriate healthcare" OR “inappropriate care” OR “unwanted healthcare” OR “unwanted care” OR “unnecessary healthcare” OR “unnecessary care” OR “Overdiagnos*” OR “Over diagnos*” OR “ineffective care” OR “ineffective healthcare”) AND (diagnos*) AND (Variation* OR Volume* OR Prevalence* OR Cost OR costs OR frequenc*) AND (Guideline* OR “Choosing Wisely” OR policy OR policies)

3. Embase:

(exp clinical effectiveness/ OR (Low-value care OR low-value hospital care OR low-value healthcare OR wasteful care OR wasteful healthcare OR wasteful hospital care OR overuse of healthcare OR overuse procedure* OR medical overuse OR inappropriate healthcare OR inappropriate care OR unwanted healthcare OR unwanted care OR unnecessary healthcare OR unnecessary care OR Overdiagnos* OR Over diagnos* OR ineffective care OR ineffective healthcare).ti,ab,kw.) AND (exp diagnosis/ OR (diagnos*).ti,ab,kw.) AND frequency/ OR exp health statistics/ OR (Variation* OR Volume* OR Prevalence* OR Cost OR costs OR frequenc*).ti,ab,kw. AND exp Health statistics/ AND exp Practice guideline/ OR (Guideline* OR Choosing Wisely OR policy OR policies).ti,ab,kw.

Supplementary file 3. Overview of imaging modalities per category.

Cardiac test	Endoscopy	Scan	Ultrasound	X-ray	Combination
Stress electrocardiogram	Arthroscopy	CT	Carotid artery imaging	X-ray	Combination of multiple modality categories
Echocardiogram	Endoscopy	MRI	Doppler	Dual Energy X-ray absorptiometry	
Cardiac nuclear medicine imaging	Colonoscopy	Whole body scan	Ultrasound		
Cardiac MRI/CT angiography		Bone scan	Plethysmography		
Single photon emission computed tomography myocardial perfusion imaging (SPECT-MPI)					

Supplementary file 4: Modified Hoy risk of bias tool.

Name of author(s):_____ Year of publication:____

Name of paper/study:_____

This tool is designed to assess the risk of bias in population-based prevalence studies. Please read the additional notes for each item when initially using the tool. Note: If there is insufficient information in the article to permit a judgement for a particular item, please answer **No (HIGH RISK)** for that particular item.

Risk of bias item	Adjusted criteria for answers (or reason for a criteria being not applicable (N.A.)):	Criteria for answers (please circle one option)	Additional notes and examples
External Validity			
1. Was the study's target population **a close representation** of the national population in relation to relevant variables, e.g. age, sex, occupation?	• Original tool used.	• **Yes (LOW RISK):** The study's target population was a close representation of the national population. • **No (HIGH RISK)**: The study's target population was clearly NOT representative of the national population.	The **target population** refers to the group of people or entities to which the results of the study will be generalised. Examples: • The study was a national health survey of people 15 years and over and the sample was drawn from a list that included all individuals in the population aged 15 years and over. The answer is: **Yes (LOW RISK).** • The study was conducted in one province only, and it is not clear if this was representative of the national population. The answer is: **No (HIGH RISK).** • The study was undertaken in one village only and it is clear this was not representative of the national population. The answer is: **No (HIGH RISK).**
2. Was the sampling frame a **true or close representation** of the target population?	• **Yes (LOW RISK):** The sampling frame was a **true or close** representation of the target population described in the guideline. • **No (HIGH RISK)**: The sampling frame was NOT a **true or close** representation of the target population described in the guideline. Limited to no information regarding the demographics of the study population was provided.	• **Yes (LOW RISK):** The sampling frame was a true or close representation of the target population. • **No (HIGH RISK):** The sampling frame was NOT a true or close representation of the target population.	The **sampling frame** is a list of the sampling units in the target population and the study sample is drawn from this list. Examples: • The sampling frame was a list of almost every individual within the target population. The answer is: **Yes (LOW RISK)**. • The cluster sampling method was used and the sample of clusters/villages was drawn from a list of all villages in the target population. The answer is: **Yes (LOW RISK)**. • The sampling frame was a list of just one particular ethnic group within the overall target population, which comprised many groups. The answer is: **No (HIGH RISK).**
3. Was some form of **random selection** used to select the sample, OR, was a census undertaken?	• Original tool used.	• **Yes (LOW RISK):** A census was undertaken, OR, some form of random selection was used to select the sample (e.g. simple random sampling, stratified random sampling, cluster sampling, systematic sampling). • **No (HIGH RISK)**: A census was NOT undertaken, AND some form of random selection was NOT used to select the sample.	A census collects information from every unit in the sampling frame. In a survey, only part of the sampling frame is sampled. In these instances, random selection of the sample helps minimise study bias. Examples: • The sample was selected using simple random sampling. The answer is: **Yes (LOW RISK)**. • The target population was the village and every person in the village was sampled. The answer is: **Yes (LOW RISK)**. • The nearest villages to the capital city were selected in order to save on the cost of fuel. The answer is: **No (HIGH RISK).**

4. Was the likelihood of **non-response bias minimal?**	• N.A. No survey studies have been included. All data is obtained from electronic databases or file studies; in which we have made the assumption that all described data was collected from the aforementioned databases.	• **Yes (LOW RISK):** The response rate for the study was >/=75%, OR, an analysis was performed that showed no significant difference in relevant demographic characteristics between responders and non-responders • **No (HIGH RISK)**: The response rate was <75%, and if any analysis comparing responders and non-responders was done, it showed a significant difference in relevant demographic characteristics between responders and non-responders.	Examples: • The response rate was 68%; however, the researchers did an analysis and found no significant difference between responders and non-responders in terms of age, sex, occupation and socio- economic status. The answer is: **Yes (LOW RISK)**. • The response rate was 65% and the researchers did NOT carry out an analysis to compare relevant demographic characteristics between responders and non-responders. The answer is: **No (HIGH RISK)**. • The response rate was 69% and the researchers did an analysis and found a significant difference in age, sex and socio-economic status between responders and non-responders. The answer is: **No (HIGH RISK)**.
Internal Validity			
5. Were data collected **directly from the subjects** (as opposed to a proxy)?	• Original tool used.	• **Yes (LOW RISK):** All data were collected directly from the subjects. • **No (HIGH RISK)**: In some instances, data were collected from a proxy.	A proxy is a representative of the subject. Examples: • All eligible subjects in the household were interviewed separately. The answer is: **Yes (LOW RISK)**. • A representative of the household was interviewed and questioned about the presence of low back pain in each household member. The answer is: **No (HIGH RISK)**.
6. Was an acceptable case definition used in the study?	• **Yes (LOW RISK):** An acceptable case definition was used. A clear definition or reference to an appropriate guideline is presented. • **No (HIGH RISK)**: An acceptable case definition was NOT used. It is unclear which definition or guideline has been used.	• **Yes (LOW RISK):** An acceptable case definition was used. • **No (HIGH RISK)**: An acceptable case definition was NOT used.	• For a study on low back pain, the following case definition was used: "Low back pain is defined as activity-limiting pain lasting more than one day in the area on the posterior aspect of the body from the bottom of the 12th rib to the lower gluteal folds." The answer is: **Yes (LOW RISK)**. • For a study on back pain, there was no description of the specific anatomical location „back" referred to. The answer is: **No (HIGH RISK)**. • For a study on osteoarthritis, the following case definition was used: "Symptomatic osteoarthritis of the hip or knee, radiologically confirmed as Kellgren-Lawrence grade 2-4". The answer is: LOW RISK.
7. Was the study instrument that measured the parameter of interest (e.g. prevalence of low back pain) shown to have **reliability and validity (if necessary)**?	• N.A. none of the studies included used an instrument to measure the parameter of interest, all studies used data directly derived from databases.	• **Yes (LOW RISK):** The study instrument had been shown to have reliability and validity (if this was necessary), e.g. test-re-test, piloting, validation in a previous study, etc. • **No (HIGH RISK)**: The study instrument had NOT been shown to have reliability or validity (if this was necessary).	• The authors used the COPCORD questionnaire, which had previously been validated. They also tested the inter-rater reliability of the questionnaire. The answer is: **Yes (LOW RISK)**. • The authors developed their own questionnaire and did not test this for validity or reliability. The answer is: **No (HIGH RISK)**.

8. Was the **same mode of data collection** used for all subjects?	• **Yes (LOW RISK):** The same mode of data collection was used for all subjects. All data was derived from the same database. • **No (HIGH RISK)**: The same mode of data collection was NOT used for all subjects. Data collected for the different participants was obtained from different databases.	• **Yes (LOW RISK):** The same mode of data collection was used for all subjects. • **No (HIGH RISK)**: The same mode of data collection was NOT used for all subjects.	The mode of data collection is the method used for collecting information from the subjects. The most common modes are face-to- face interviews, telephone interviews and self-administered questionnaires. Examples: • All eligible subjects had a face-to-face interview. The answer is: **Yes (LOW RISK)**. • Some subjects were interviewed over the telephone and some filled in postal questionnaires. The answer is: **No (HIGH RISK)**.
9. Was the **length of the shortest prevalence period** for the parameter of interest appropriate?	• N.A. no prevalence period for the parameter of interest was used in our situation. We selected studies which use database data and thus are recorded at the time of treatment/consultation.	• **Yes (LOW RISK):** The shortest prevalence period for the parameter of interest was appropriate (e.g. point prevalence, one-week prevalence, one-year prevalence). • **No (HIGH RISK)**: The shortest prevalence period for the parameter of interest was not appropriate (e.g. lifetime prevalence)	The prevalence period is the period that the subject is asked about e.g. "Have you experienced low back pain over the previous year?" In this example, the prevalence period is one year. The longer the prevalence period, the greater the likelihood of the subject forgetting if they experienced the symptom of interest (e.g. low back pain). Examples: • Subjects were asked about pain over the past week. The answer is: **Yes (LOW RISK)**. • Subjects were only asked about pain over the past three years. The answer is: **No (HIGH RISK)**.
10. Were the **numerator(s) and denominator(s)** for the parameter of interest appropriate?	• Original tool used.	• **Yes (LOW RISK):** The paper presented appropriate numerator(s) AND denominator(s) for the parameter of interest (e.g. the prevalence of low back pain). • **No (HIGH RISK)**: The paper did present numerator(s) AND denominator(s) for the parameter of interest but one or more of these were inappropriate.	There may be errors in the calculation and/or reporting of the numerator and/or denominator. Examples: • There were no errors in the reporting of the numerator(s) AND denominator(s) for the prevalence of low back pain. The answer is: **Yes (LOW RISK)**. • In reporting the overall prevalence of low back pain (in both men and women), the authors accidentally used the population of women as the denominator rather than the combined population. The answer is: **No (HIGH RISK).**
11. Summary item on the overall risk of study bias			
• **LOW RISK OF BIAS:** Further research is very unlikely to change our confidence in the estimate. • **HIGH RISK OF BIAS:** Further research is very likely to have an important impact on our confidence in the estimate and is likely to change the estimate.			

Supplementary file 6. Details of low-value care assessments extracted from the included studies.

Study	Country	Type of low-value care examined	Care setting assessed	Lens used	Guideline(s)/recommendations used	Cohort size	Amount LVC	Prevalence estimate (95%CI)	Type
Badgery-Parker et al., 2019 [15]	Australia	- Endoscopy in adults <55 (for dyspepsia)	Secondary care	Service	CW Australia, RACP EVOLVE and NICE 'do not do' guidelines/recommendations	14 813	2 360	15.93 [15.35 ; 16.53]	Imaging
		- Arthroscopic lavage and debridement of knee for osteoarthritis or degenerative meniscal tears				4 218	3 002	71.17 [69.78 ; 72.53]	Imaging
		- Colonoscopy in adults < 50 (for constipation)				11 790	608	5.16 [4.76 ; 5.57]	Imaging
		- ERCP (endoscopic retrograde cholangiopancreatography) for acute gallstone pancreatitis without cholangitis				420	79	18.80 [15.18 ; 22.88]	Imaging
Bouck et al., 2019 [39]	Canada	- Spinal X-ray, CT, MRI following visit for low-back pain.	Primary care	Patient-indication	CW Canada guidelines/recommendations from the Canadian Anesthesiologists' Society (CAS), Canadian Cardiovascular Society (CCS) & Canadian Society of Internal Medicine (CSIM)	97 740	30 006	30.70 [30.41 ; 30.99]	Imaging
		- Cardiac tests (electrocardiogram, chest x-ray, stress test, or transthoracic echocardiogram) prior to low-risk procedures				527 691	167 278	31.70 [31.60 ; 31.80]	Imaging
Chalmers et al., 2019 [16]	Australia	- Knee arthroscopy in case of osteoarthritis or meniscal derangements	Secondary care	Service	RACP EVOLVE, CW Australia, CW USA, CW Canada, CW UK guidelines/recommendations	3 620	2 958	81.70 [80.41 ; 82.96]	Imaging
		- Endoscopy in case of dyspepsia below 50 years				5 021	501	9.97 [9.16 ; 10.84]	Imaging
		- Colonoscopy in case of constipation below 50 years				4 017	133	3.31 [2.78 ; 3.91]	Imaging
Charlesworth et al., 2016 [35] [a]	United States	- Imaging for nonspecific low-back pain	Both	Patient-indication	CW USA , NICE guidelines/recommendations and QualityNet measures from the Center for Medicaid & Medicare Services (CMS)	82 659	12 977	15.70 [15.45 ; 15.95]	Imaging
		- Head imaging for uncomplicated headache				65 931	7 252	11.0 [10.76 ; 11.24]	Imaging
		- Head imaging for syncope				7 466	672	9.0 [8.36 ; 9.67]	Imaging
		- Imaging for plantar fasciitis				10 986	1 999	18.20 [17.48 ; 18.93]	Imaging
		- T3 test for hypothyroidism				31 228	5 090	16.30 [15.89 ; 16.71]	Laboratory test
		- Pre-operative chest radiography				39 169	1 684	4.30 [4.10 ; 4.5]	Imaging
		- Abdomen CT combined studies				28 075	2 106	7.50 [7.20 ; 7.82]	Imaging
		- Simultaneous brain and sinus CT				9 742	360	3.70 [3.33 ; 4.09]	Imaging
		- CT for uncomplicated acute rhinosinusitis				59 961	1 799	3.0 [2.87 ; 3.14]	Imaging

		- Arthroscopic surgery for knee osteoarthritis				23 190	1 136	4.90 [4.62 ; 5.18]	Imaging
		- Thorax CT combined studies				9 287	111	1.20 [0.98 ; 1.44]	Imaging
		- Preoperative echocardiography				39 169	235	0.60 [0.53 ; 0.68]	Imaging
		- Preoperative stress testing				39 169	118	0.30 [0.25 ; 0.36]	Imaging
		- Electroencephalogram for headache				59 936	60	0.10 [0.08 ; 0.13]	Other test
Charlesworth et al., 2016 [35] [b]	United States	- Imaging for nonspecific low-back pain	Both	Patient-indication	CW USA , NICE guidelines/recommendations and QualityNet measures from the Center for Medicaid & Medicare Services (CMS)	18 871	4 625	22.60 [22.01 ; 23.2]	Imaging
		- Head imaging for uncomplicated headache				23 211	4 317	18.60 [18.10 ; 19.11]	Imaging
		- Head imaging for syncope				3 174	448	14.10 [12.92 ; 15.37]	Imaging
		- Imaging for plantar fasciitis				1 450	180	12.40 [10.76 ; 14.22]	Imaging
		- T3 test for hypothyroidism				5 891	619	10.50 [9.74 ; 11.32]	Laboratory test
		- Pre-operative chest radiography				7 848	753	9.60 [8.95 ; 10.27]	Imaging
		- Abdomen CT combined studies				13 416	657	4.90 [4.54 ; 5.28]	Imaging
		- Simultaneous brain and sinus CT				7 761	357	4.60 [4.14 ; 5.09]	Imaging
		- CT for uncomplicated acute rhinosinusitis				11 992	348	2.90 [2.61 ; 3.22]	Imaging
		- Arthroscopic surgery for knee osteoarthritis				6 143	166	2.70 [2.31 ; 3.14]	Imaging
		- Thorax CT combined studies				3 822	76	2.0 [1.57 ; 2.48]	Imaging
		- Preoperative echocardiography				7 848	86	1.10 [0.88 ; 1.35]	Imaging
		- Preoperative stress testing				7 848	47	0.60 [0.44 ; 0.8]	Imaging
		- Electroencephalogram for headache				20 091	60	0.30 [0.23 ; 0.38]	Other test
Chmiel et al., 2015 [19]	Switzerland	- Performance of a diagnostic coronary angiography without previous non-invasive ischemia testing	Unclear	Patient-indication	The American heart association (AHA), NICE, the European Society of Cardiology (ESC) & Swiss Society of Cardiology (SGK) guidelines/recommendations	2 714	1 018	37.50 [35.68 ; 39.36]	Imaging
Choi et al., 2011 [40]	United States	- Routine cross-sectional imaging (CT, MRI, endorectal coil MRI) for staging low-risk prostate cancer	Unclear	Patient-indication	American College of Radiology (ACR) & National Comprehensive Cancer Network (NCCN) guidelines/recommendations	2 330	548	23.50 [21.81 ; 25.3]	Imaging
		- Routine bone scan for staging low-risk prostate cancer				2 330	617	26.50 [24.70 ; 28.32]	Imaging
		- Routine abdominal ultrasound for staging low-risk prostate cancer				2 330	42	1.80 [1.30 ; 2.43]	Imaging
Colla et al., 2014 [27]	United States	- Low back X-ray, CT or MRI within six weeks of incident low-back pain diagnosis	Both	Patient-indication	American Academy of Family Physicians (AAFP), American College of Physicians (ACP) & North American Spine Society	8 440 000	1 899 000	22.50 [22.48 ; 22.52]	Imaging
		- Intravenous pyelogram or an abdominal CT, MRI, or ultrasound within 60 days of the				75 000 000	900 000	1.20 [1.20 ; 1.20]	Imaging

		index diagnosis benign prostatic hyperplasia (BPH)			(NASS) guidelines/recommendations				
		- DXA scans performed on female beneficiaries at low risk for fracture within 23 months of a previous scan				147 420 000	14 299 740	9.70 [9.70 ; 9.70]	Imaging
		- Non-indicated cardiac test, including stress tests, echocardiograms, electrocardiograms and advanced cardiac imaging in the 30 days before cataract surgery				6 490 000	999 460	15.40 [15.37 ; 15.43]	Imaging
		- Non-indicated cardiac test, including stress tests, echocardiograms, electrocardiograms, CTs, MRIs or PETs within 30 days before low-risk surgery				3 550 000	1 650 750	46.50 [46.46 ; 46.54]	Imaging
Colla et al., 2018 [36] [a]	United States	- Early imaging inf patients with uncomplicated, incident low-back pain who received nonindicated low-back-pain imaging in the 6 weeks following diagnosis	Both	Patient-indication	American Academy of Family Physicians (AAFP)& American College of Physicians (ACP) guidelines/recommendations	3 110 000	895 680	28.80 [28.76 ; 28.84]	Imaging
		- Short-interval repeat bone densitometry (DXA)				8 000 000	600 000	7.50 [7.48 ; 7.52]	Imaging
Colla et al., 2018 [36] [c]	United States	- Early imaging inf patients with uncomplicated, incident low-back pain who received nonindicated low-back-pain imaging in the 6 weeks following diagnosis	Both	Patient-indication	American Academy of Family Physicians (AAFP)& American College of Physicians (ACP) guidelines/recommendations	8 400 000	1 898 400	22.60 [22.58 ; 22.62]	Imaging
		- Short-interval repeat bone densitometry (DXA)				143 000 000	12 155 000	8.50 [8.50 ; 8.5]	Imaging
Doukky et al., 2016 [41]	United States	- SPECT-MPI (myocardial perfusion imaging)	Unclear	Patient-indication	The American College of Cardiology (ACC)& American Society of Nuclear Cardiology (ASNC) revised appropriate use criteria (AUC) for SPECT MPI of 2009	1 511	688	45.53 [22.58 ; 22.62]	Imaging
Drangsholt et al., 2019 [42]	United States	- Routine imaging for staging low-risk prostate cancer (Whole body scan)	Secondary care	Patient-indication	National Comprehensive Cancer Network (NCCN) guidelines/recommendations	414	381	92.0 [88.99 ; 94.45]	Imaging
		- Routine imaging for staging low-risk prostate cancer (CT)				414	273	66.0 [61.15 ; 70.5]	Imaging

		- Routine imaging for staging low-risk prostate cancer (abdominal/pelvic MRI)				414	8	2.0 [0.84 ; 3.77]	Imaging
Farghaly et al., 2006 [43]	United States	- V-Pap after T-Hyst	Unclear	Service	US Preventive Services Task Force (USPTF) & American Society for Colposcopy and Cervical Pathology (ASCCP) guidelines/recommendations	1 303	581	44.59 [41.87 ; 47.34]	Laboratory test
Feng et al., 2016 [44]	United States	- Preoperative ECGs in patients undergoing sling surgery	Unclear	Service	Summary guidelines from the American Academy of Family Physicians (AAFP)	63	13	20.60 [11.47 ; 32.7]	Imaging
		- Preoperative chest X-ray (CXR) in patients undergoing sling surgery				23	9	39.10 [19.71 ; 61.46]	Imaging
		- Preoperative basic metabolic panel in patients undergoing sling surgery				59	28	47.50 [34.30 ; 60.88]	Laboratory test
		- Preoperative complete blood count determination in patients undergoing sling surgery				63	46	73.0 [60.35 ; 83.43]	Laboratory test
		- Preoperative coagulation studies in patients undergoing sling surgery				41	31	75.60 [59.70 ; 87.64]	Laboratory test
Flaherty et al., 2018 [38]	United States	- MRI use in case of uncomplicated LBP	Both	Patient-indication	American College of Radiology (ACR) guidelines	5 103	1 838	36.01 [59.70 ; 87.64]	Imaging
		- MRI use in case of non-traumatic Knee pain				6 935	3 384	48.79 [47.61 ; 49.98]	Imaging
		- MRI use in case of non-traumatic shoulder pain				9 388	4 011	42.72 [41.72 ; 43.73]	Imaging
		- X-ray use in case of uncomplicated LBP				10 540	6 650	63.09 [62.16 ; 64.02]	Imaging
		- X-ray use in case non-traumatic Knee pain				37 543	25 668	68.37 [67.90 ; 68.84]	Imaging
		- X-ray use in case of non-traumatic shoulder pain				36 453	26 100	71.60 [71.13 ; 72.06]	Imaging
Ganguli et al., 2019 [45]	United States	- EKGs for cataract surgeries	Unclear	Patient-indication	Choosing Wisely recommendations from the American Academy of Ophthalmology & American Society for Clinical Pathology	110 183	12 451	11.30 [11.11 ; 11.49]	Imaging
Gidwani et al., 2016 [46]	United States	- Lumbar Spine MRI's for non-specific or nonpersistent low-back pain	Unclear	Service	The American College of Physicians (ACP) and the American Association of Neurological Surgeons (AANS) guidelines/recommendations and	110 661	33 973	30.70 [30.43 ; 30.97]	Imaging

					the National Quality forum (NQF)-endorsed CMS criteria				
Gill et al., 2017 [47]	Canada	- Thyroid testing	Unclear	Service	Clinical practice guidelines and Choosing Wisely guidelines/recommendations from the American Society for Clinical Pathology (ASCP), American Association for Clinical Chemistry (AACC), American Association of Clinical Endocrinologists (AACE), Canadian Society of Endocrinology and Metabolism (CSEM), Endocrine society.	752 217	75 974	10.10 [10.04 ; 10.16]	Laboratory test
Gold et al., 2016 [37] [d]	United States	- Imaging in patients with non-specific low-back pain (X-ray, CT, MRI)	Primary care	Patient-indication	American Academy of Family Physicians National Quality Measures Clearinghouse (NQMC) guidelines/recommendations	19 503	3 288	16.86 [16.34 ; 17.39]	Imaging
Gold et al., 2016 [37] [e]	United States	- Imaging in patients with non-specific low-back pain (X-ray, CT, MRI)	Primary care	Patient-indication	American Academy of Family Physicians National Quality Measures Clearinghouse (NQMC) guidelines/recommendations	2 694	449	16.67 [15.28 ; 18.13]	Imaging
Hajati et al., 2018 [48]	Australia	- High density lipoprotein cholesterol testing more often than once every 12 months for high-risk groups	Unclear	Patient-indication	Royal Australian College of General Practitioners (RACGP) guidelines	1 628 477	309 411	19.0 [18.95 ; 19.05]	Laboratory test
Kool et al., 2020 [18]	The Netherlands	- Doppler or plethysmography for diagnosis of varices	Primary care	Patient-indication	Dutch general practitioner, CW USA & CW Canada guidelines/recommendations	15 990	1 279	8.0 [7.58 ; 8.43]	Imaging
Kovacs et al., 2013 [20]	Spain	- Lumbar Spine MRI's for low-back pain without red flags.	Unclear	Service	National Institute for Clinical Excellence (NICE), the American College of Physicians and the American College of Radiologists guidelines/recommendations	602	64	10.60 [8.28 ; 13.37]	Imaging

Lalude et al., 2014 [49]	United States	- Inappropriate Single photon emission computed tomography myocardial perfusion imaging (SPECT MPI) utilization	Unclear	Service	The American College of Cardiology (ACC)& American Society of Nuclear Cardiology (ASNC) revised appropriate use criteria (AUC) for SPECT MPI of 2009	54	8	14.0 [6.62 ; 27.12]	Imaging
Lehnert et al., 2010 [50]	United States	- CT use - MRI use	Primary care	Service	American College of Radiology (ACR) Appropriateness Criteria and other evidence-based guidelines/recommendations	284 175	77 41	27.11 [17.37 ; 30.41] 23.43 [22.03 ; 32.68]	Imaging Imaging
Mafi et al., 2017 [17]	United States	- Baseline lab tests for low risk patients having low-risk surgery - Stress cardiac or other cardiac imaging in low-risk, asymptomatic patients - Routine head CT scans for Emergency Department visits for severe dizziness - EKGs, chest x-rays, or pulmonary function tests in low-risk patients having low-risk surgery - Routine imaging for uncomplicated acute rhinosinusitis - Imaging for low-back pain within the first six weeks of symptom onset, in absence of red flags	Unclear	Service	CW USA, US Preventive Services Task Force (USPTF), Medicare's Healthcare Effectiveness Data and Information Set (HEDIS) criteria and other clinical guidelines/recommendations	595 552 244 487 29 816 33 754 14 196 48 857	468 104 27 383 15 713 32 910 7 226 48 857	78.60 [78.51 ; 78.69] 11.20 [11.08 ; 11.32] 52.70 [52.13 ; 53.27] 97.50 [97.33 ; 97.66] 50.90 [50.08 ; 51.73] 86.20 [85.89 ; 86.51]	Laboratory test Imaging Imaging Imaging Imaging Imaging
Martin et al., 2012 [51]	United States	- MRI use in musculoskeletal oncology	Both	Patient-indication	Referral guidelines for patients with cancer	320	20	6.25 [3.86 ; 9.49]	Imaging
McAlister et al. 2018 [52]	Canada	- PSA testing for men 75 or older with no history of prostate cancer - Bone mineral density testing within 2 years of prior scan - Hypercoagulability testing in patients with first Deep Vein Trombosis/ Pulmonary Embolism - Preoperative coronary CT scan or cardiac stress test before non-cardiac surgery - Homocysteine testing without B12/folate	Both	Patient-population	CW and NICE guidelines/recommendations	100 227 271 854 21 311 698 683 2 585 832	55 596 31 617 744 7 266 10 602	55.47 [55.16 ; 55.78] 11.63 [11.52 ; 11.74] 3.49 [3.25 ; 3.75] 1.04 [1.02 ; 1.06] 0.41 [0.40 ; 0.42]	Laboratory test Imaging Laboratory test Imaging Laboratory

		deficiency							test
		- Carotid artery imaging but without history of stroke / TIA				3 162 394	2 846	0.09 [0.09 ; 0.09]	Imaging
		- Carotid artery imaging for patients with syncope but no history of stroke / TIA				74 060	355	0.48 [0.43 ; 0.53]	Imaging
Morden et al., 2014 [53]	United States	- Short-interval (repeated in under 2 years) dual-energy X-ray absorptiometry tests (DXAs) for bone density	Unclear	Service	The American College of Rheumatology guidelines/recommendations	13 800 000	1 393 800	10.10 [10.08 ; 10.12]	Imaging
Mou et al., 2017 [54]	United States	- Thrombophilia in adult patients with venous thromboembolism occurring in the setting of major transient risk factors (surgery, trauma, or prolonged immobility)	Secondary care	Service	The American Society of Hematology (ASH) and Clinical guidelines/recommendations	1 817	777	42.76 [40.47 ; 45.08]	Laboratory test
Pendrith et al., 2017 [55]	Canada	- Imaging for low back pain in absence of red flags - DEXA scans repeated under 2 years	Primary care	Patient-indication	CW Canada, CW USA	271 588 2 229 113	12 222 468 114	4.5 [4.42 ; 4.58] 21.0 [20.95 ; 21.05]	
Petruzziello et al., 2012 [21]	Italy	- Colonoscopy use	Both	Patient-indication	The American Society for Gastrointestinal Endoscopy (ASGE) guidelines/recommendations	432	125	29.0 [24.70 ; 33.46]	Imaging
Schwartz et al., 2014 [13]	United States	- Bone mineral density test <2 y after prior bone mineral density test	Both	Patient-population	CW USA, the US Preventive Services Task Force “D”, NICE “do not do”, Canadian Agency for Drugs and Technologies in Health (CADTH) guidelines/ recommendations and guidelines from peer-reviewed medical literature	1 360 908	127 925	9.40 [9.35 ; 9.45]	Imaging
		- Homocysteine testing for cardiovascular disease				1 360 908	20 414	1.50 [1.48 ; 1.52]	Laboratory test
		- Hypercoagulability testing for patients with deep vein thrombosis				1 360 908	1 361	0.10 [0.09 ; 0.11]	Laboratory test
		- PTH (parathyroid hormone) measurement for patients with stage 1-3 CKD (chronic kidney disease)				1 360 908	34 023	2.50 [2.47 ; 2.53]	Laboratory test
		- Preoperative chest radiograph specified as a preoperative assessment or occurring within 30 days before a low- or intermediate-risk noncardiothoracic surgical procedure				1 360 908	69 406	5.10 [5.06 ; 5.14]	Imaging
		- Preoperative echocardiography before a low- or intermediate-risk noncardiothorical surgery				1 360 908	10 887	0.80 [0.78 ; 0.82]	Imaging

		- CT of the sinuses for uncomplicated acute rhinosinusitis				1 360 908	8 165	0.60 [0.59 ; 0.61]	Imaging
		- Head imaging in the evaluation of syncope				1 360 908	17 692	1.30 [1.28 ; 1.32]	Imaging
		- Head imaging for uncomplicated headache				1 360 908	42 188	3.10 [3.07 ; 3.13]	Imaging
		- EEG for headaches				1 360 908	1 361	0.10 [0.09 ; 0.11]	Other test
		- Back imaging for patients with nonspecific low back pain				1 360 908	127 925	9.40 [9.35 ; 9.45]	Imaging
		- Arthroscopic surgery for knee osteoarthritis				1 360 908	2 722	0.2 [0.19 ; 0.21]	Imaging
Scott et al., 2014 [56]	United States	- Carotid duplex ultrasound (CDUS)for simple syncope in the absence of focal neurological signs or symptoms suggestive of stroke	Unclear	Patient-indication	European Society of Cardiology's Task Force on Syncope & American College of Physicians' Clinical (ACP) Efficacy Assessment Project guidelines/recommendations	137 424	8 933	6.50 [6.37 ; 6.63]	Imaging
Sharp et al., 2015 [57]	United States	- CT scan for acute sinusitis	Both	Patient-indication	American Academy of Family Physicians (AAFP), American Academy of Allergy, Asthma and Immunology (AAAAI), American Academy of Otolaryngology-Head and Neck surgery (AAO-HNS) CW guidelines/recommendations	152 774	1681	1.1 [1.05 ; 1.15]	Imaging
Sheffield et al., 2013 [58]	United States	- Cardiac Stress tests prior to elective non-cardiac, non-vascular surgery: Echocardiograms, Myocardial nuclear imaging, Exercise treadmill or pharmacological stress tests	Secondary care	Patient-indication	American College of Cardiology (ACC)& American Heart Association guidelines (AHA) guidelines/recommendations	74 785	2 804	3.75 [3.61 ; 3.89]	Imaging
Sprenger et al., 2016 [59]	Austria	- DEXA scans more often than every 2 years	Primary care	Patient-population	CW USA, CW Canada, NICE recommendations/guidelines and contradicted medical practices and potentially low-value health care practices identified by the studies Prasad et al., 2013 and Elshaug et al., 2012	246 131	12 282	4.99 [4.91 ; 5.07]	Imaging
		- Total or free T3 level when assessing levothyroxine (T4) dose in hypothyroid patients				246 131	34 434	13.99 [13.86 ; 14.12]	Laboratory test
		- Men 75 to 80 years old with a PSA less than 3ng/ml are unlikely to die or experience aggressive prostate cancer during their remaining life, suggesting that PSA testing may be safely discontinued in these men.				246 131	13 390	5.44 [5.35 ; 5.53]	Laboratory test
		- Pap smears on women younger than 21				246 131	10 387	4.22 [4.14 ; 4.3]	Laboratory

		years who have had a hysterectomy for non-cancer disease.							test
		- Routine monitoring of bone mineral density after starting bisphosphonate treatment				246 131	271	0.11 [0.10 ; 0.12]	Imaging
Xu et al., 2013 [8]	United States	- X-ray in outpatient management of uncomplicated Upper Respiratory Infections (URI's)	Primary care	Patient-indication	Centres for Disease Control and Prevention (CDC) guidelines	2 200 000	1 012 000	46.0 [45.95 ; 46.05]	Imaging
		- CT in outpatient management of uncomplicated Upper Respiratory Infections (URI's)				2 200 000	44 000	2.0 [1.98 ; 2.02]	Imaging

Among the included studies, some studies contained multiple assessments undertaken in different cohorts. These assessments are distinguished by the following: [a] assessment performed among a commercially insured population, [b] assessment performed among Medicaid beneficiaries, [c] assessment performed among Medicare beneficiaries, [d] assessment performed using Kaiser Permanente Epic EHR data, [e] assessment performed using data derived from the Oregon Community Health Information Network. Abbreviations: CT, Computed Tomography; CW, Choosing Wisely; DEXA or DXA, Dual Energy X-ray absorptiometry; EKG, Electrocardiogram; LBP, low-back pain; MRI, Magnetic Resonance Imaging; NICE, National Institute for Health and Care Excellence; PSA, Prostate-Specific Antigen; RACP, Royal Australasian College of Physicians; SPECT MPI, Single-Photon Emission computed tomography; TIA, Transient Ischemic Attack; T-hyst, Total Hysterectomy.

Supplementary file 5: Risk of bias assessment outcome.

Author	1. Was the study"s target population **a close representation** of the national population in relation to relevant variables, e.g. age, sex, occupation?	2. Was the sampling frame a **true or close representation** of the target population?	3. Was some form of **random selection** used to select the sample, OR, was a census undertaken?	4. Was the likelihood of **non-response bias minimal?**	5. Were data collected **directly from the subjects** (as opposed to a proxy)?	6. Was an acceptable case definition used in the study?	7. Was the study instrument that measured the parameter of interest (e.g. prevalence of low back pain) shown to have **reliability and validity (if necessary)?**	8. Was the **same mode of data collection** used for all subjects?	9. Was the **length of the shortest prevalence period** for the parameter of interest appropriate?	10. Were the numerator(s) and denominato r(s) for the parameter of interest appropriate?	***11. Summary item on the overall risk of study bias: - 2x high: High - 1x high + 1x unclear: High - Rest of combinations: low***
Badgery-Parker et al., 2019 [15]	Low	Low	Low	N.A.	High	Low	N.A.	Low	N.A.	Low	Low
Bouck et al., 2019 [39]	Low	Low	Low	N.A.	High	Low	N.A.	Low	N.A.	Low	Low
Chalmers et al., 2019 [16]	Low	Low	Low	N.A.	High	Low	N.A.	Low	N.A.	Low	Low
Charlesworth et al., 2016 [35]	Low	Low	Low	N.A.	High	Low	N.A.	Low	N.A.	Low	Low
Chmiel et al., 2015 [19]	Low	Low	Low	N.A.	High	Unclear	N.A.	Low	N.A.	Low	Low
Choi et al., 2011 [40]	Low	Low	Low	N.A.	High	Low	N.A.	Low	N.A.	Low	Low
Colla et al., 2014 [27]	Low	Low	Low	N.A.	High	Low	N.A.	Low	N.A.	Low	Low
Colla et al., 2018 [36]	Low	Low	Low	N.A.	High	Low	N.A.	Low	N.A.	Low	Low
Doukky et al., 2016 [41]	High	High	Low	N.A.	High	Unclear	N.A.	Low	N.A.	Low	High
Drangsholt et al., 2019 [42]	High	High	Low	N.A.	High	Low	N.A.	Low	N.A.	Low	High
Farghaly et al., 2006 [43]	Low	Unclear	Low	N.A.	High	Low	N.A.	Low	N.A.	Low	Low
Feng, et al., 2016 [44]	High	Low	Low	N.A.	High	High	N.A.	Low	N.A.	Low	High
Flaherty et al., 2018 [38]	Low	Low	Low	N.A.	High	Low	N.A.	Low	N.A.	Low	Low
Ganguli et al., 2019 [45]	Low	Low	Low	N.A.	High	Low	N.A.	Low	N.A.	Low	Low
Gidwani et al., 2016 [46]	High	High	Low	N.A.	High	Low	N.A.	Low	N.A.	Low	High
Gill et al., 2017 [47]	Low	Low	Low	N.A.	High	Low	N.A.	Low	N.A.	Low	Low
Gold et al., 2016 [37]	Low	Low	Low	N.A.	High	Low	N.A.	Low	N.A.	Low	Low

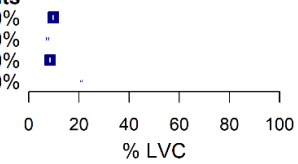
Hajati et al., 2018 [48]	Low	Low	Low	N.A.	High	Low	N.A.	Low	N.A.	Low	Low
Kool et al., 2020 [18]	Low	Low	Low	N.A.	High	Low	N.A.	Low	N.A.	Low	Low
Kovacs et al., 2013 [20]	Low	Unclear	Low	N.A.	High	Low	N.A.	Low	N.A.	Low	Low
Lalude et al., 2014 [49]	Unclear	Low	Low	N.A.	High	High	N.A.	Low	N.A.	Low	High
Lehnert et al., 2010 [50]	High	Unclear	Low	N.A.	High	High	N.A.	Low	N.A.	Low	High
Mafi et al., 2017 [17]	Low	Unclear	Low	N.A.	High	Low	N.A.	Low	N.A.	Low	Low
Martin et al., 2012 [51]	High	Low	Low	N.A.	High	Unclear	N.A.	Low	N.A.	Low	High
McAlister et al. 2018 [52]	Low	Low	Low	N.A.	High	Low	N.A.	Low	N.A.	Low	Low
Morden et al., 2014 [53]	Low	Low	Low	N.A.	High	Low	N.A.	Low	N.A.	Low	Low
Mou et al., 2017 [54]	High	Low	Low	N.A.	High	High	N.A.	Low	N.A.	Low	High
Pendrith et al., 2017 [55]	Low	Low	Low	N.A.	High	Low	N.A.	Low	N.A.	Low	Low
Petruzziello et al., 2012 [21]	Unclear/high	Low	Low	N.A.	High	High	N.A.	Low	N.A.	Low	High
Schwartz et al., 2014 [13]	Low	Low	Low	N.A.	High	Low	N.A.	Low	N.A.	Low	Low
Scott et al., 2014 [56]	Low	Low	Low	N.A.	High	Low	N.A.	Low	N.A.	Low	Low
Sharp et al., 2015 [57]	High	Low	Low	N.A.	High	High	N.A.	Low	N.A.	Low	High
Sheffield et al., 2013 [58]	Low	Low	Low	N.A.	High	Low	N.A.	Low	N.A.	Low	Low
Sprenger et al., 2016 [59]	Low	Low	Low	N.A.	High	Low	N.A.	Low	N.A.	Low	Low
Xu et al., 2013 [8]	Low	Low	Low	N.A.	High	Low	N.A.	Low	N.A.	Low	Low

Supplementary file 7: Forest plots of assessment outcomes from studies using the same lens to assess overuse of similar low-value care (LVC) diagnostic tests. Among the included studies, some studies contained multiple assessments undertaken in different cohorts. These assessments are distinguished by the following: [a] assessment performed among a commercially insured population, [b] assessment performed among Medicaid beneficiaries, [c] assessment performed among Medicare beneficiaries, [d] assessment performed using Kaiser Permanente (KP) Epic EHR data, [e] assessment performed using data derived from the Oregon Community Health Information Network (OCHIN).

A. Short-interval repeat bone densitometry: patient-indication lens

Study	Cases	Total	% LVC	95% C.I.	Weights
Colla et al., 2014	14299740	147420000	9.70	[9.70; 9.70]	25.0%
Colla et al., 2018 [a]	600000	8000000	7.50	[7.48; 7.52]	25.0%
Colla et al., 2018 [c]	12155000	143000000	8.50	[8.50; 8.50]	25.0%
Pendrith et al., 2017	468114	2229113	21.00	[20.95; 21.05]	25.0%

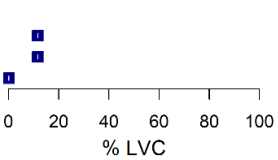
Heterogeneity:Tau^2 = 0.0006; Chi^2 = 409644.19, df = 3 (P = 0); I^2 = 100%



B. Short-interval repeat bone densitometry: patient-population lens

Study	Cases	Total	% LVC	95% C.I.	Weights
McAlister et al. 2018	31617.0	271854	11.63	[11.51; 11.75]	33.3%
Schwartz et al., 2014	127925.4	1094374	11.69	[11.63; 11.75]	33.3%
Sprenger et al., 2016	271.0	246131	0.11	[0.10; 0.12]	33.3%

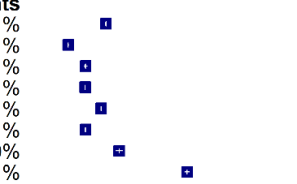
Heterogeneity:Tau^2 = 0.0264; Chi^2 = 83058.45, df = 2 (P = 0); I^2 = 100%



C. Imaging for low-back pain: patient-indication lens

Study	Cases	Total	% LVC	95% C.I.	Weights
Bouck et al., 2019	30006	97740	30.70	[30.41; 30.99]	9.1%
Charlesworth et al., 2016 [a]	12977	82659	15.70	[15.45; 15.95]	9.1%
Charlesworth et al., 2016 [b]	4265	18871	22.60	[22.01; 23.20]	9.1%
Colla et al., 2014	1899000	8440000	22.50	[22.47; 22.53]	9.1%
Colla et al., 2018 [a]	895680	3110000	28.80	[28.75; 28.85]	9.1%
Colla et al., 2018 [c]	1898400	8400000	22.60	[22.57; 22.63]	9.1%
Flaherty et al., 2018	1838	5103	36.02	[34.71; 37.34]	9.0%
Flaherty et al., 2018	6650	10540	63.09	[62.17; 64.01]	9.1%
Gold et al., 2016 [d]	3288	19503	16.86	[16.34; 17.39]	9.1%
Gold et al., 2016 [e]	449	2694	16.67	[15.28; 18.10]	8.9%
Pendrith et al., 2017	12222	271588	4.50	[4.42; 4.58]	9.1%

Heterogeneity:Tau^2 = 0.0031; Chi^2 = 159518.30, df = 10 (P = 0); I^2 = 100%



D. Imaging for low-back pain: service lens

Study	Cases	Total	% LVC	95% C.I.	Weights
Gidwani et al., 2016	33973	110661	30.70	[30.43; 30.97]	33.4%
Kovacs et al., 2013	64	602	10.63	[8.29; 13.23]	33.3%
Mafi et al., 2017	42115	48857	86.20	[85.89; 86.50]	33.4%

Heterogeneity:Tau^2 = 0.1809; Chi^2 = 49746.87, df = 2 (P = 0); I^2 = 100%