

INPLASY

Has the Appropriateness of Laboratory Testing Improved? A Systematic Review and Meta-Analysis

INPLASY202670037

doi: 10.37766/inplasy2026.7.0037

Received: 13 July 2026

Published: 13 July 2026

Hernandez, MM; Kirmaier, A; Morgan, A; Ahmed, AM; Mishra, M; Farren, M; Lee, GR; Kirby, JE; Arnaout, R.

Corresponding author:

Matthew Hernandez

matt.hernandez@stanford.edu

Author Affiliation:

Beth Israel Deaconess Medical Center; Stanford University School of Medicine.

ADMINISTRATIVE INFORMATION**Support** - Reagan-Udall Foundation for the FDA and by FDABAA-23-00123/FY23C1JWP13 (to Ramy Arnaout) and Stanford University 1292781-100-DHPIR (to Matthew M. Hernandez).**Review Stage at time of this submission** - Completed but not published.**Conflicts of interest** - Matthew M. Hernandez is supported by Gilead Science Inc. and receives reagent support from DiaSorin S. p. A. for projects unrelated to this study.**INPLASY registration number:** INPLASY202670037**Amendments** - This protocol was registered with the International Platform of Registered Systematic Review and Meta-Analysis Protocols (INPLASY) on 13 July 2026 and was last updated on 19 October 2026.**INTRODUCTION**

Review question / Objective To conduct a systematic review and meta-analysis of laboratory testing appropriateness across chemistry, haematology, microbiology, transfusion, and molecular testing published between 2012 and 2022. To apply our novel OU8 framework to standardize utilisation measurements in an effort to address existing methodological limitations.

Condition being studied A systematic review of existing literature identified that the last major meta-analysis of laboratory test utilisation (1997–2012) reported overall rates of 20·6% for overuse and 44·8% for underuse. Since then, no comprehensive meta-analysis has simultaneously reassessed both overuse and underuse, leaving the current status of the highest volume medical activity in doubt. Subsequent reviews have been limited in scope or methodology, frequently

focusing solely on overuse or omitting pooled estimates due to heterogeneity. Inconsistent definitions of appropriateness, variable measurement strategy, and heterogeneous reporting across the literature have precluded meaningful comparisons and trend assessments. Consequently, it remains unclear if inappropriate testing has improved over the last decade, despite the increased adoption of electronic health records, usage audits, and stewardship programmes—leaving the value of these many activities uncertain, from the perspective of improving testing. Furthermore, underuse remains substantially understudied despite its potentially severe clinical consequences.

METHODS

Search strategy This study was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)

guidelines. We searched Medline, EMBASE, BIOSIS, and CINAHL for laboratory test utilisation studies published between 1 Jan 2012 and 1 Jan 2023. The Medline database was searched on 28 Apr 2023; detailed search terms are provided below. As searches of BIOSIS, CINAHL, and EMBASE yielded no unique results beyond those identified in Medline, the latter served as the sole source for subsequent analysis.

Medline: ("laboratory test"[tiab] OR ("serologic test"[tiab]) OR ("serum test"[tiab]) OR ("diagnostic"[tiab] AND "test"[tiab]) OR "genetic test"[tiab] OR "Medical test"[tiab] OR "Drug level"[tiab] OR "Blood test"[tiab] OR "Antigen test"[tiab] OR ("Biomarker"[tiab] AND "test"[tiab]) OR ("biochemical"[tiab] AND "test"[tiab]) OR ("glycemic"[tiab] AND "test"[tiab]) OR "Western blot"[tiab] OR "Protein assay"[tiab] OR "Tumor marker"[tiab] OR "Tumour marker"[tiab] OR "Trace element test"[tiab] OR ("serology test"[tiab]) OR ("serological test"[tiab]) OR "PCR test"[tiab] OR "MALDI"[tiab] OR "Antimicrobial susceptib"[tiab] OR "Antibiotic susceptib"[tiab] OR "Antimicrobial sensitiv"[tiab] OR "Antibiotic sensitiv"[tiab] OR "urine test"[tiab] OR "urinary test"[tiab] OR "lab test"[tiab] OR "Screening test"[tiab] OR "Confirmatory test"[tiab]) AND ("underutiliz*" [tiab] OR "underutilis*" [tiab] OR "underus*" [tiab] OR "overus*" [tiab] OR "misus*" [tiab] OR "overutiliz*" [tiab] OR "overutilis*" [tiab] OR "optimi*" [tiab] OR "optima*" [tiab] OR "Inappropriat*" [tiab] OR "Not appropriate" [tiab] OR "repeat*" [tiab] OR "contraindicat*" [tiab] OR "duplicat*" [tiab] OR ("clinical" [tiab] AND "audit" [tiab]))).

Participant or population Studies that measure the (in)appropriate utilisation of diagnostic laboratory tests.

Intervention Not applicable.

Comparator Not applicable.

Study designs to be included Retrospective study, prospective study.

Eligibility criteria We included original research containing discrete numerical data on the appropriateness of laboratory testing. For studies evaluating interventions to improve

appropriateness, only pre-intervention data were used to better reflect established clinical practice. Studies were excluded if they: (i) were meta-analyses, reviews, case reports, or editorials; (ii) examined only radiographic imaging or anatomic/surgical pathology; (iii) addressed laboratory quality control rather than appropriateness; (iv) overlapped with previously identified publications; or (v) were not published in English.

Information sources MEDLINE, EMBASE, BIOSIS, and CINAHL were searched for eligible publications. Searches were conducted on April 28, 2023, for studies published between January 1, 2012, and January 1, 2023. Searches of BIOSIS, CINAHL, and EMBASE yielded no unique results beyond those identified in MEDLINE, which therefore supplied the records used for subsequent analysis. Full search details are provided under Search strategy. Data from the previous laboratory-utilization meta-analysis by Zhi et al. (PMID: 24260139) were additionally used for between-review comparisons.

Main outcome(s) The main outcomes were the prevalence of laboratory test overuse and underuse, quantified from measurement-specific numerators and denominators and classified using the OU8 framework. Overuse represented testing performed when not indicated; underuse represented indicated testing that was not performed. The OU8 framework distinguished four overuse metrics and four underuse metrics according to the populations represented by their numerators and denominators. Overall and subgroup-specific pooled proportions were estimated separately for overuse and underuse. Additional analyses evaluated associations with measurement type, test timing, appropriateness-criteria characteristics, test volume, laboratory department, and geographic location. Between-review analyses compared current estimates with reanalyzed data from the previous (PMID: 24260139) meta-analysis.

Quality assessment / Risk of bias analysis Data extraction and quality assessment were performed in three stages. First, two authors (AMA and MM) independently extracted author names, publication year, PMID, study period, location, test type, appropriateness criteria, and relevant numerators and denominators. Second, two additional authors (AK and MMH) independently reviewed the data to confirm accuracy, further characterize study metrics, and identify gaps. Finally, a panel of four authors (AK, AM, MMH, and RA) convened to resolve discrepancies and finalize the utilization metrics used for analysis. During this final review,

metrics were excluded if they: (i) involved radiology or surgical pathology; (ii) occurred after a diagnostic stewardship intervention; (iii) lacked concrete numerators or denominators; or (iv) used denominators that precluded classification into an appropriateness metric. Qualifying metrics were then categorized as either overuse or underuse. Methodological quality was appraised using the Joanna Briggs Institute (JBI) Critical Appraisal Checklist for Studies Reporting Prevalence Data, adapted to nine domains emphasizing the threats most relevant to estimating appropriateness. Each measure was appraised independently by two reviewers (AK and MMH), with adjudication by a third and fourth reviewer (AM and RA). Publication bias was assessed using Peters' regression test for funnel-plot asymmetry and Doi plots with Luis Furuya-Kanamori (LFK) indices (R metasens package, version 1.5.3).

Strategy of data synthesis Overuse and underuse were analyzed separately. Principal pooled estimates were obtained using three-level random-effects models of logit-transformed proportions and their sampling variances, with measurements nested within studies. Three-level models distinguished sampling error, within-study heterogeneity across measurements, and between-study heterogeneity. Models were fitted using restricted maximum likelihood (REML) with the `rma.mv()` function in the `metafor` package (version 4.8.0). Overall proportions and 95% confidence intervals were obtained by back-transforming fitted mean logits and their normal-theory confidence limits. Individual measurement confidence intervals were calculated using the Clopper-Pearson method. Three-level models were compared with reduced two-level models retaining measurement-level heterogeneity but constraining between-study variance to zero. Model fit was assessed using likelihood-ratio tests and the Akaike and Bayesian information criteria (AIC and BIC), with lower information-criterion values indicating better fit. These comparisons supported retaining three-level models for the current-review overuse and underuse analyses. Heterogeneity was summarized using within-study and between-study variance components and their respective contributions to total I^2 , calculated relative to the sum of heterogeneity and typical sampling variance.

Current-review analyses were conducted separately from the historical dataset. For between-review comparisons, measurements from the previous meta-analysis by Zhi et al. (PMID: 24260139) were combined with the current dataset, with review origin included as a

moderator. Between-review differences were evaluated using two-sided Wald contrasts of fitted mean logits. For overuse, three heterogeneity specifications were compared: common within-study and between-study variances (Model 0); review-specific between-study variance with common within-study variance (Model 1); and review-specific variances at both levels (Model 2). Allowing review-specific between-study heterogeneity improved fit by likelihood-ratio testing, although information-criterion support was modest and mixed. Differences in periods, settings, and underlying data represented by the two reviews provided additional justification for this specification. Subsequent overuse comparisons therefore used review-specific between-study variance and common within-study variance. The same three-level modeling approach was used to estimate and compare overall underuse, including re-estimation of the previous overall underuse rate. Although neither the historical underuse model comparison nor the between-review heterogeneity comparisons clearly supported additional complexity, a three-level model with review-specific between-study variance and common within-study variance was retained for consistency with the overuse analysis. Unless otherwise indicated, estimates attributed to the previous review reflect this reanalysis.

Benjamini-Hochberg adjustment was applied separately to omnibus moderator tests and distinct pairwise comparisons within each outcome analysis. For between-review overuse comparisons, adjustment covered all estimable category-specific review contrasts across the included moderators, excluding the overall review comparison. Raw and adjusted P values were reported; confidence intervals were not adjusted for multiplicity. Statistical significance was defined as adjusted $P < 0.05$ for multiplicity-adjusted comparisons and unadjusted $P < 0.05$ otherwise. Analyses were performed using R version 4.4.1.

Subgroup analysis Subgroup associations were evaluated using three-level mixed-effects meta-regression, fitting one categorical moderator at a time. Moderators included OU8 measurement type, test timing, criteria objectivity, criteria permissiveness or restrictiveness, criteria origin, test volume, laboratory department, and geographic location, where estimable. Category-specific estimates were derived from the corresponding moderator model. Omnibus Wald tests assessed overall moderator associations, and two-sided Wald z-tests evaluated pairwise comparisons.

For between-review overuse analyses, category-specific contrasts were estimated from models containing the observed review-by-category combinations, with the selected heterogeneity structure shared across categories within each moderator model. Limited historical underuse data precluded moderator-specific between-review comparisons.

The proportion of estimated total heterogeneity explained by each moderator was quantified using pseudo- R^2 , calculated as one minus the ratio of summed within-study and between-study variance components in the moderator model to those in the intercept-only model. Exploratory joint moderator models evaluated the combined explanatory contribution of selected moderators.

Percentile 95% confidence intervals were obtained from 1,000 study-cluster bootstrap replicates, resampling studies with all their constituent measurements (R boot package, version 1.3.30). Confidence intervals were not reported when fewer than 95% of attempted replicates produced valid estimates.

Sensitivity analysis Measurements with extreme observed rates of 0% or 100% were considered potentially artifactual because some could reflect selection of inappropriate testing instances or denominator definitions that constrained the observed rate. Following visual inspection and LFK threshold interpretation, 29 overuse and 9 underuse measurements were identified as outliers and excluded from the primary analyses, resulting in the exclusion of two studies.

Their influence on pooled estimates was evaluated by comparing full datasets, containing all eligible measurements, with refined datasets excluding measurements with rates of 0% or 100%. Pooled estimates were obtained using binomial-logistic random-effects models fitted by maximum likelihood without study-level clustering, using the same specification for both datasets. This preliminary sensitivity analysis was distinct from the three-level analyses used for the principal estimates.

Pooled proportions, 95% confidence intervals, heterogeneity, and variance estimates were compared before and after exclusion. Peters' regression and Doi-plot assessments were also examined before and after exclusion.

Country(ies) involved United States of America.

Keywords Clinical laboratory testing; diagnostic stewardship; laboratory utilization; overuse; underuse; testing appropriateness.

Contributions of each author

Author 1 - Matthew Hernandez - Data analysis, manuscript drafting, critical revision.

Email: matt.hernandez@stanford.edu

Author 2 - Andrea Kirmaier - Data analysis, manuscript drafting.

Email: andrea.x.kirmaier@kp.org

Author 3 - Alexandra Morgan - Manuscript drafting.

Email: amorgan2@bidmc.harvard.edu

Author 4 - Ahmed Ahmed - Data collection.

Email: aahmed18@bidmc.harvard.edu

Author 5 - Meghana Mishra - Data collection.

Email: mmishra2@bwh.harvard.edu

Author 6 - Martina Farren - Data analysis.

Email: martina.farren@stanford.edu

Author 7 - Ghee Rye Lee - Manuscript drafting.

Email: gheerye.lee@duke.edu

Author 8 - James Kirby - Critical revision.

Email: jekirby@bidmc.harvard.edu

Author 9 - Ramy Arnaut - Study conception and design, manuscript drafting, critical revision, supervision and funding acquisition.

Email: rarnaout@bidmc.harvard.edu