INPLASY

INPLASY202670037

doi: 10.37766/inplasy2026.7.0037

Received: 13 July 2026

Published: 13 July 2026

Corresponding author:
Matthew Hernandez

matt.hernandez@stanford.edu

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

Has Laboratory Testing Improved? A Systematic Review and Meta-analysis Using the OU8 Framework

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

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 13 July 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 utlisation 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 Corresponding authors can be contacted at matt.hernandez@stanford.edu (Matthew M. Hernandez) or rarnaout@bidmc.harvard.edu (Ramy Arnaout).

Main outcome(s) We identified over 30 million diagnostic tests across 278 study measures. Overall, 26·3% of tests were unnecessary, reflecting overuse (95% CI: 22·3–30·6%), while 42·5% of indicated tests were not performed, indicating underuse (95% CI: 32·0–53·7%). Mixed-effects meta-regression showed that overuse remained unchanged from the previous decade (p=0·1765), whereas overuse of tests based on permissive (broad) criteria increased from 5·3% (95% CI: 3·6–7·6%) to 18·6% (95% CI: 14·4–23·8%) (p=0·04). Underuse remains understudied relative to its prevalence. Initial testing was more frequently overused than repeat testing (45·1% vs 13·6%, p<0·0001), whereas repeat testing was more frequently underused (73·6% vs 34·0%, p=0·0059). The OU8 framework revealed wide variations in overuse (14·5–43·4%) and underuse (26·9–45·4%) depending on the type of measurement, underscoring the need for standardised approaches.We identified over 30 million diagnostic tests across 278 study measures. Overall, 26·3% of tests were unnecessary, reflecting overuse (95% CI: 22·3–30·6%), while 42·5% of indicated tests were not performed, indicating underuse (95% CI: 32·0–53·7%).

Quality assessment / Risk of bias analysis Data extraction and quality assessment were performed in three stages. First, two authors (AA 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 characterise study metrics, and identify gaps. Finally, a panel of four authors (AK, AM, MMH, and RA) convened to resolve discrepancies and finalise the utilisation 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/denominators; or (iv) used denominators that precluded classification into an appropriateness metric. Qualifying metrics were then categorised as either overuse or underuse. Publication bias was assessed using Peters' regression test for funnel plot asymmetry13 and Doi plots with Luis Furuya-Kanamori (LFK) indices (R metasens package, v1.5.3).

Strategy of data synthesis Initial over- and underuse rates were estimated via random-intercept logistic regression with logit transformation (R meta package, v8.2.1). Study metric 95% confidence intervals (CI) were calculated using the Clopper-Pearson method. Heterogeneity (I2) was quantified via Cochran's Q statistic. Between-study variance (τ2) was estimated using maximum-likelihood estimation, with 95% CI computed using the t-distribution with Hartung-Knapp adjustment.

Note, for longitudinal comparison with our previous meta-analysis, data from PMID: 24260139 were pooled with the current dataset (totalling 409 measures: 342 overuse, 67 underuse), using the review origin ("Then" (PMID: 24260139) vs "Now" (current meta-analysis)) as a moderator variable; differences in proportions were assessed via two-sample z-tests. Three-level models were supported for both over- and underuse analyses in this pooled dataset. Because these datasets were constructed independently, we evaluated whether between-study heterogeneity (τ2) should be constrained as equal across cohorts or estimated separately. For overuse, a significant likelihood ratio test ($\chi^2(1)=5{\cdot}73$, $p=0{\cdot}017$) and AIC ($\Delta AIC=3{\cdot}73$) favoured the unconstrained model with separate τ2 estimates per cohort. For underuse, all three criteria converged on the constrained model ($\chi^2(1)=0{\cdot}046$, $p=0{\cdot}831$; $\Delta AIC=1{\cdot}95$; $\Delta BIC=4{\cdot}16$), and a common τ2 was retained. Unless otherwise noted, all rates attributed to PMID: 24260139 are the values resulting from this re-analysis.

Subgroup analysis Random-effects meta-analyses generated pooled estimates for specified subgroups. To test for rate differences across subgroups, we compared nested two-level and three-level random-effects models: two-level models partitioned variance at the metric level(within studies), while three-level models additionally accounted for between-study variance. Models were fitted using the rma.mv() function (R package metafor, v4.8.0) with restricted maximum likelihood (REML) estimation. Model selection was based on likelihood ratio tests (ANOVA with χ2 statistic; df=1), the Akaike Information Criterion (AIC), and the Bayesian Information Criterion (BIC), with absolute differences >10 indicating strong evidence for the preferred model. Convergence across all three criteria consistently favoured three-level models for both over- and underuse estimates.

Subsequently, three-level mixed-effects meta-regression was used to evaluate subgroup differences; omnibus tests of moderators (QM statistic) assessed overall significance, while pairwise comparisons were evaluated using two-tailed Wald z-tests against a designated reference category. Meta-regression heterogeneity was quantified by calculating the variance (σ2) at both within-study and between-study levels. The proportion of total variance explained by each moderator was quantified using a pseudo R2 statistic, comparing intercept-only models to moderator-inclusive models; 95% CI were obtained via percentile bootstrapping (1,000 iterations; R package boot, v1.3.30).

Sensitivity analysis Following visual inspection and LFK threshold interpretation, 29 overuse and four underuse measurements were identified as outliers (extreme rates of 0% or 100%) and excluded as likely artefacts, resulting in the exclusion of two studies. Pooled estimates, heterogeneity (I2), and variance (τ2) were evaluated prior to and after exclusion of outliers and compared.

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 Arnaout - Study conception and design, manuscript drafting, critical revision, supervision and funding acquisition.
Email: rarnaout@bidmc.harvard.edu