What is the effectiveness of crowdsourcing for healthcare information verification?

Crowdsourcing achieves accuracy comparable to expert verification for healthcare information tasks—with crowd-expert agreement matching expert-expert agreement in several domains—while completing tasks 5 to 150 times faster and at a fraction of the cost, though effectiveness is optimized when tasks are well-structured, self-contained, and employ appropriate quality control measures such as limiting assessments to three per item.

Abstract

Crowdsourcing demonstrates comparable accuracy to expert verification for healthcare information tasks, with crowd-expert agreement (κ=0.58) matching expert-expert agreement (κ=0.59) for ontology verification and correlation coefficients reaching 0.95 for surgical skills assessment . Citation screening achieves near-perfect sensitivity (98.9-100%) , while pharmacovigilance annotation exceeds 90% accuracy compared to expert-curated datasets . Knowledge base development shows moderate recall (62-66%) with high precision (87-98%) . Crowdsourcing offers substantial efficiency advantages, completing tasks 4.8 to 150.9 times faster than expert evaluation at approximately one-quarter the cost .

Effectiveness depends on task characteristics and implementation parameters. Simple, self-contained verification tasks with clear criteria achieve highest accuracy , while tasks requiring synthesis or extensive clinical judgment show more variable performance . Three assessments per citation optimizes sensitivity without sacrificing specificity for screening tasks , and qualification sets of 10-15 items adequately evaluate crowd performance . Worker qualifications matter less than task design, with no differences observed based on training background for structured tasks . Crowdsourcing is most effective when expert-level tasks are reformulated for crowd suitability and when appropriate quality control mechanisms are implemented .

Paper search

We performed a semantic search using the query ”What is the effectiveness of crowdsourcing for healthcare information verification?” across over 138 million academic papers from the Elicit search engine, which includes all of Semantic Scholar and OpenAlex.

We retrieved the 50 papers most relevant to the query.

Screening

We screened in sources based on their abstracts that met these criteria:

- **Healthcare Information Verification Focus**: Does this study evaluate crowdsourcing methods specifically applied to healthcare information verification?
- **Effectiveness Measures**: Does this study report quantitative or qualitative measures of effectiveness (such as accuracy rates, time to verification, cost analysis, or user experience measures)?
- **Empirical Study Design**: Is this an empirical study (including randomized controlled trials, quasi-experimental studies, observational studies, case studies, mixed-methods research, or systematic reviews/meta-analyses)?
- **Crowdsourcing Approach Definition**: Does this study clearly define their crowdsourcing approach?
- **Verification vs Data Collection**: Does this study focus on information verification rather than solely on healthcare data collection without verification components?

- **Healthcare Domain**: Does this study examine crowdsourcing within healthcare domains rather than non-healthcare domains?
- **Publication Quality and Empirical Data**: Is this study a full research paper with empirical data and sufficient methodological detail (rather than a theoretical paper, opinion piece, editorial, conference abstract, poster, or brief communication)?

We considered all screening questions together and made a holistic judgement about whether to screen in each paper.

Data extraction

We asked a large language model to extract each data column below from each paper. We gave the model the extraction instructions shown below for each column.

- **Verification Target**:

  Extract what specific healthcare information was being verified through crowdsourcing, including:

- Type of information (clinical knowledge, skills assessment, ontologies, facility data, etc.)
- Complexity level (simple yes/no verification vs complex judgment tasks)
- Domain expertise required
- Volume/scale of information processed

- **Crowdsourcing Method**:

  Extract details about how crowdsourcing was implemented, including:

- Platform used (Amazon Mechanical Turk, custom platform, etc.)
- Worker qualifications/selection criteria
- Compensation structure and amounts
- Task design and instructions given to workers
- Quality control mechanisms (training, validation questions, etc.)
- Number of workers per task

- **Comparison Standard**:

  Extract what the crowdsourced results were compared against, including:

- Type of gold standard (expert judgment, existing databases, manual verification, etc.)
- Number and qualifications of expert evaluators
- Inter-expert agreement measures when reported
- How discrepancies were resolved

- **Accuracy Results**:

  Extract all quantitative measures of crowdsourcing accuracy and reliability, including:

- Correlation coefficients with expert/gold standard
- Sensitivity and specificity values
- Agreement measures (Cronbach's alpha, kappa, etc.)
- Error rates or percentage of correct classifications
- Confidence intervals where provided

- **Efficiency Results**:

  Extract comparative efficiency metrics, including:

- Time to completion (crowdsourced vs traditional methods)
- Cost comparisons (total cost, cost per task, cost-effectiveness ratios)
- Throughput/scalability metrics
- Speed multipliers (e.g., 'X times faster than experts')
- Resource requirements

- **Effectiveness Moderators**:

  Extract factors that influenced crowdsourcing effectiveness, including:

- Task characteristics that affected performance
- Worker characteristics or qualifications that mattered
- Contextual factors (geographic, demographic, institutional)
- Implementation factors that improved/hindered results
- Threshold effects or optimal parameters identified

- **Practical Outcomes**:

  Extract real-world applications and utility of the crowdsourced verification, including:

- How results were integrated into workflows or systems
- Practical benefits achieved (error detection, knowledge base improvement, etc.)
- Scalability demonstrated or potential
- Limitations or failure modes identified
- Recommendations for implementation

Characteristics of Included Studies

The 10 sources examined crowdsourcing across diverse healthcare information verification tasks, ranging from ontology verification to surgical skills assessment to systematic review screening.

Study	Full text retrieved?	Study Type	Verification Target	Platform	Crowd Characteristics
J. Mortensen et al., 2015	Yes	Primary study	SNOMED CT ontological relationships	CrowdFlower	Diverse workforce, 25 workers per task
A. McCoy et al., 2015	Yes	Primary study	Problem-medication pairs	Clinicians using EHR system	Clinical expertise required
Nassr Nama et al., 2017	Yes	Primary study	Systematic review citation eligibility	Custom platform (CHEORI)	Medical background, 4 assessments per citation

Study	Full text retrieved?	Study Type	Verification Target	Platform	Crowd Characteristics
J. Dai et al., 2017	No	Systematic review	Surgical technical skills assessment	Not specified	Moderate scale, 13 studies reviewed
Alex Gartland et al., 2020	Yes	Primary study	Pharmacovigilance data (adverse events, indications)	Amazon Mechanical Turk	USA-based workers, 1-3 per task
J. Mortensen et al., 2013	No	Primary study	SNOMED CT CORE subset relationships	Amazon Mechanical Turk	Simple, short tasks
D. Holst et al., 2015	Yes	Primary study	Robotic surgical skills (GEARS tool)	Amazon Mechanical Turk	50 workers per video, >100 HITs, >95% approval
Nassr Nama et al., 2020	Yes	Secondary analysis	Citation screening for systematic reviews	Not specified	Volunteers without content expertise, up to 3 per citation
A. McCoy et al., 2012	Yes	Primary study	Problem-medication pairs	Allscripts Enterprise EHR	867 clinicians
Minki Kim et al., 2014	Yes	Primary study	Pediatric clinic quality	Naver and Daum online communities	User-generated content analysis

The studies employed heterogeneous crowdsourcing approaches. Three studies used Amazon Mechanical Turk as the primary platform , while others utilized custom platforms , electronic health record systems , or existing online communities . Task complexity varied substantially, from simple yes/no verification tasks to complex judgment tasks requiring clinical knowledge . The number of crowd workers per task ranged from 1 to 50 , reflecting different approaches to achieving consensus and quality control.

Accuracy Outcomes

Agreement with Expert Standards

Study	Comparison Standard	Correlation/Agreement	Sensitivity	Specificity	Other Accuracy Metrics
J. Mortensen et al., 2015	Expert panel (5 clinical experts)	κ=0.58 (crowd-expert)	Not reported	Not reported	AUC=0.83 ; Expert inter-rater κ=0.59

Study	Comparison Standard	Correlation/Agreement	Sensitivity	Specificity	Other Accuracy Metrics
A. McCoy et al., 2015	Lexi-Comp drug database	Not reported	Not reported	Not reported	Recall=62.3%, Precision=87.5%
Nassr Nama et al., 2017	Two trained reviewers	63% complete agreement	100% (95% CI: 88-100%)	99% (95% CI: 96-100%)	Not reported
J. Dai et al., 2017	Expert evaluations	Pearson's r=0.59-0.95	Not reported	Not reported	Cronbach's α=0.32-0.92
Alex Gartland et al., 2020	Pharmacovigilance experts (13 curators)	Not reported	Not reported	Not reported	Overall accuracy >90% ; Phase I: 92.8% ; Phase II: 91.8%
J. Mortensen et al., 2013	Manual expert inspection	Not reported	Not reported	Not reported	86% correct verification
D. Holst et al., 2015	Seven expert robotic surgeons	r=0.95	Not reported	Not reported	Cronbach's α=0.93 ; Expert inter-rater reliability=0.89 ; R^2=0.91
Nassr Nama et al., 2020	Investigative team judgments	Correlation at 15 citations: 0.86 (sensitivity), 0.75 (specificity)	98.9-100.0%	49-87.0%	Sensitivity at 3 assessments: >99%
A. McCoy et al., 2012	Two expert investigators	κ=0.68 (expert inter-rater)	65.8% (combined KB)	97.9% (combined KB)	Manual links: sensitivity 42.8%, specificity 99.6%
Minki Kim et al., 2014	HIRA government database	Kendall tau and Spearman rho used	Not reported	Not reported	Regional variation in congruence

Crowdsourcing demonstrated strong agreement with expert standards across most verification tasks. The inter-rater agreement between crowd and experts (κ=0.58) was nearly indistinguishable from agreement among experts themselves (κ=0.59) for ontology verification . In surgical skills assessment, correlation coefficients reached as high as 0.95 with Cronbach's alpha of 0.93, indicating excellent agreement . The systematic review of surgical crowdsourcing found correlations ranging from good to excellent (Pearson's r=0.59-0.95) across multiple studies .

For citation screening tasks, crowdsourcing achieved exceptionally high sensitivity (98.9-100%) while maintaining variable specificity (49-87%) . This pattern suggests crowdsourcing reliably identifies relevant items but may be more conservative in exclusions. Pharmacovigilance annotation achieved overall accuracy exceeding 90% compared to expert-curated reference datasets .

The knowledge base development studies showed more moderate performance, with recall of 62.3% and precision of 87.5% for problem-medication pairs , though this represented an improvement over pilot study results (recall 46.9%, precision 83.3%) . The combined knowledge base achieved sensitivity of 65.8% and specificity of 97.9% .

Efficiency Outcomes

Study	Time Comparison	Cost Comparison	Throughput Metrics
J. Mortensen et al., 2015	Not explicitly reported	$0.50/relationship (crowd) vs $2.00/relationship (expert)	Large-scale processing implied
A. McCoy et al., 2015	Not mentioned	Described as ”inexpensive”	Not reported
Nassr Nama et al., 2017	Not mentioned	Not mentioned	84% work saved at abstract screening; 73% overall
J. Dai et al., 2017	4.8-150.9x faster than expert evaluation	Described as ”cost effective”	Not reported
Alex Gartland et al., 2020	~5% of expert time (20x faster)	Not explicitly reported	5000 posts in 33 hours
J. Mortensen et al., 2013	Described as ”quickly”	Described as ”relatively cheaply”	Not reported
D. Holst et al., 2015	<5 hours (crowd) vs 14 days (experts)	$0.75/assessment	”Boundless” assessor availability
Nassr Nama et al., 2020	Not mentioned	Not mentioned	Not reported
A. McCoy et al., 2012	Not mentioned	Described as ”inexpensive”	Not reported
Minki Kim et al., 2014	Not mentioned	Not mentioned	173,748 messages analyzed

Crowdsourcing demonstrated substantial efficiency advantages over traditional expert-based verification. The surgical skills assessment study found crowd workers completed evaluations in less than 5 hours compared to 14 days for expert surgeons . Across multiple surgical education studies, non-expert evaluation was consistently 4.8 to 150.9 times faster than expert evaluation . Pharmacovigilance annotation was completed in approximately 5% of the time required for expert curation .

Cost comparisons were favorable, with crowdsourcing costing approximately one-quarter of expert verification for ontology tasks ($0.50 vs $2.00 per relationship) . Individual assessments on Amazon Mechanical Turk cost as little as $0.75 per evaluation . For citation screening, crowdsourcing reduced investigative team workload by 73-84% .

Moderating Factors

Several factors influenced crowdsourcing effectiveness across studies.

Task Design and Complexity

Task reformulation significantly affected performance. Complex expert-level tasks needed to be adapted for crowd workers . Simple, short tasks were more amenable to crowdsourcing , while tasks requiring synthesis or background

knowledge not directly provided were less suitable . Binary outcome questions performed well for pharmacovigilance tasks , and the use of standardized assessment tools like GEARS facilitated effective crowdsourcing for surgical skills .

Number of Assessments and Quality Control

The optimal number of crowd assessments per item emerged as a critical parameter. Studies found that three assessments per citation achieved excellent sensitivity (>99%) for citation screening , with additional assessments providing diminishing returns and actually decreasing specificity . For surgical skills assessment, 50 unique workers per video achieved satisfactory agreement with expert grades .

Qualification sets with as few as five eligible and five ineligible citations effectively evaluated crowd member performance . Correlation coefficients plateaued at approximately 15 citations (r=0.86 for sensitivity, r=0.75 for specificity) .

Worker Characteristics

Worker qualifications showed varying importance across tasks. For surgical skills assessment, requiring completion of 100+ HITs with >95% approval rating improved quality . No differences in performance were noted based on training background or prior research experience for citation screening . Using "Master Turkers" with positive feedback could potentially improve accuracy . Medical background without specific training was sufficient for systematic review screening .

Compensation and Incentives

Higher pay increased time efficiency but did not significantly affect accuracy in pharmacovigilance annotation . Studies emphasized the importance of appropriate compensation to avoid creating an "Internet sweatshop" perception while maintaining quality .

Contextual Factors

Geographic and demographic factors influenced effectiveness for community-based crowdsourcing. Higher birthrates, larger variance in educational attainment, higher population density, and more doctors per clinic were associated with greater congruence between crowdsourced information and official data . Institutional context and clinical terminology differences affected generalizability of knowledge base approaches .

Practical Outcomes and Applications

Study	Primary Application	Practical Benefits	Limitations Identified
J. Mortensen et al., 2015	Ontology error detection	Identified 39 previously undiscovered critical errors	Need to reformulate expert tasks for crowd suitability
A. McCoy et al., 2015	Clinical decision support integration	EHR integration for patient summarization and e-prescribing	Terminology and granularity differences across settings

Study	Primary Application	Practical Benefits	Limitations Identified
Nassr Nama et al., 2017	Systematic review acceleration	Faster knowledge synthesis with broader search capability	Limited generalizability validation needed
J. Dai et al., 2017	Surgical education feedback	Cost-effective alternative to expert assessment	Inconsistency in expert evaluations as baseline
Alex Gartland et al., 2020	Machine learning training data	Accurate annotation for pharmacovigilance automation	Generalizability beyond social media unclear
J. Mortensen et al., 2013	Large-scale ontology auditing	Quick, inexpensive ontology verification	Not specified
D. Holst et al., 2015	Trainee advancement and QI	Scalable skills assessment for residency programs	Needs validation in human surgery contexts
Nassr Nama et al., 2020	SR workflow optimization	Reduced workload, maintained accuracy	Additional assessments beyond three detrimental
A. McCoy et al., 2012	Problem list maintenance	Automated linking and patient safety improvement	Potential for incorrect links requiring continuous evaluation
Minki Kim et al., 2014	Health policy guidance	Identification of regions vulnerable to misinformation	Text mining performance limitations

Crowdsourcing enabled practical outcomes including identification of 39 previously undiscovered critical errors in SNOMED CT , integration into EHR workflows for clinical decision support , and development of scalable surgical skills assessment systems . The technology showed potential for integration into ontology development environments , systematic review platforms , and pharmacovigilance automation pipelines .

Synthesis

The evidence consistently demonstrates that crowdsourcing achieves accuracy comparable to expert verification across diverse healthcare information tasks, with substantial efficiency gains. However, apparent heterogeneity in performance metrics across studies can be explained by examining task characteristics, implementation parameters, and comparison standards.

Reconciling Variation in Agreement Measures

Studies reporting highest agreement ($r>0.90$) employed well-structured assessment tools (GEARS) and tasks with clear evaluation criteria. Studies with moderate agreement ($\kappa=0.58$-0.68) involved more subjective judgment tasks such as ontology verification and problem-medication linking . Notably, in these moderate-agreement scenarios, crowd-expert agreement matched or approached expert-expert agreement (0.58 vs 0.59 for ontology verification) , suggesting the limitation lies in inherent task ambiguity rather than crowd capability.

Sensitivity-Specificity Trade-offs

Citation screening studies consistently achieved near-perfect sensitivity (98.9-100%) but variable specificity (49-87%) . This asymmetry reflects appropriate calibration for systematic review screening, where missing relevant studies (false negatives) is more consequential than including irrelevant ones (false positives). The conservative approach of requiring all reviewers to agree before exclusion maintains sensitivity at 100% while accepting reduced specificity.

Optimal Implementation Parameters

Convergent evidence identifies three assessments per citation as optimal for screening tasks , with additional assessments decreasing specificity without improving sensitivity . Qualification sets of 10-15 true positives and negatives provide adequate performance evaluation . Higher worker numbers (50 per video) prove necessary for complex judgment tasks like surgical assessment , while single reviewers achieve 91.8% accuracy for simpler binary classifications .

Task Suitability

Crowdsourcing performs optimally for self-contained verification tasks where necessary information can be provided within the task interface . Tasks requiring synthesis, extensive background knowledge, or clinical judgment show more variable performance . Simple yes/no verification tasks achieve highest accuracy , while complex problem-medication linking shows moderate recall (62-66%) but high precision (87-98%) .

Generalizability Considerations

Results generalized across healthcare settings for knowledge base development, with validation in both academic and community-based health systems . For community-sourced information (online health communities), demographic factors including birthrate, educational variance, population density, and healthcare provider density significantly predicted congruence with official data , suggesting that effectiveness of passive crowdsourcing varies by population characteristics.

References

A. McCoy, A. Wright, A. Laxmisan, M. Ottosen, Jacob A. McCoy, David Butten, and Dean F. Sittig. "Development and Evaluation of a Crowdsourcing Methodology for Knowledge Base Construction: Identifying Relationships Between Clinical Problems and Medications." *J. Am. Medical Informatics Assoc.*, 2012.

A. McCoy, Adam Wright, M. Krousel-Wood, Eric J. Thomas, Jacob A. McCoy, and Dean F. Sittig. "Validation of a Crowdsourcing Methodology for Developing a Knowledge Base of Related Problem-Medication Pairs." *Applied Clinical Informatics*, 2015.

Alex Gartland, A. Bate, Jeffery L. Painter, tim. a. casperson, and G. Powell. "Developing Crowdsourced Training Data Sets for Pharmacovigilance Intelligent Automation." *Drug Safety*, 2020.

D. Holst, Timothy M. Kowalewski, Lee W. White, T. Brand, J. Harper, Mathew D. Sorensen, M. Truong, et al. "Crowd-Sourced Assessment of Technical Skills: Differentiating Animate Surgical Skill Through the Wisdom of Crowds." *Journal of Endourology*, 2015.

J. Dai, T. Lendvay, and Mathew D. Sorensen. "Crowdsourcing in Surgical Skills Acquisition: A Developing Technology in Surgical Education." *Journal of Graduate Medical Education*, 2017.

J. Mortensen, E. Minty, Michael Januszyk, T. Sweeney, A. Rector, Natasha Noy, and M. Musen. "Using the Wisdom of the Crowds to Find Critical Errors in Biomedical Ontologies: A Study of SNOMED CT." *J. Am. Medical Informatics Assoc.*, 2015.

J. Mortensen, M. Musen, and Natasha Noy. "Crowdsourcing the Verification of Relationships in Biomedical Ontologies." *American Medical Informatics Association Annual Symposium*, 2013.

Minki Kim, Yuchul Jung, Dain Jung, and Cinyoung Hur. "Investigating the Congruence of Crowdsourced Information With Official Government Data: The Case of Pediatric Clinics." *Journal of Medical Internet Research*, 2014.

Nassr Nama, Klevis Iliriani, Meng Yang Xia, B. P. Chen, L. Zhou, Supichaya Pojsupap, C. Kappel, et al. "A Pilot Validation Study of Crowdsourcing Systematic Reviews: Update of a Searchable Database of Pediatric Clinical Trials of High-Dose Vitamin D." *Translational Pediatrics*, 2017.

Nassr Nama, N. Barrowman, Katie O'Hearn, M. Sampson, R. Zemek, and J. McNally. "Quality Control for Crowdsourcing Citation Screening: The Importance of Assessment Number and Qualification Set Size." *Journal of Clinical Epidemiology*, 2020.