Original Article

A pilot validation study of crowdsourcing systematic reviews: update of a searchable database of pediatric clinical trials of high-dose vitamin D

Nassr Nama[1], Klevis Iliriani[2*], Meng Yang Xia[1*], Brian P. Chen[1*], Linghong Linda Zhou[1*], Supichaya Pojsupap[3*], Coralea Kappel[1*], Katie O'Hearn[3], Margaret Sampson[4], Kusum Menon[1,3], James Dayre McNally[1,3]

[1]Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada; [2]School of Medicine, Trinity College, Dublin, Ireland; [3]Department of Pediatrics, [4]Department of Volunteers Communication and Information Resources, Children's Hospital of Eastern Ontario, Ottawa, ON, Canada

Contributions: (I) Conception and design: N Nama, M Sampson, K Menon, JD McNally; (II) Administrative support: N Nama, M Sampson, K Menon, JD McNally; (III) Provision of study materials or patients: N Nama, JD McNally; (IV) Collection and assembly of data: N Nama, K Iliriani, MY Xia, BP Chen, LL Zhou, S Pojsupap, C Kappel, K O'Hearn, M Sampson, JD McNally; (V) Data analysis and interpretation: N Nama, MY Xia, M Sampson, K Menon, JD McNally; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

*These authors contributed equally to this work.

Correspondence to: James D. McNally, MD, PhD. Division of Critical Care, Department of Pediatrics, Children's Hospital of Eastern Ontario, 401 Smyth Road, Ottawa, Ontario, K1H 8L1, Canada. Email: dmcnally@cheo.on.ca.

Background: Completing large systematic reviews and maintaining them up to date poses significant challenges. This is mainly due to the toll required of a small group of experts to screen and extract potentially eligible citations. Automated approaches have failed so far in providing an accessible and adaptable tool to the research community. Over the past decade, crowdsourcing has become attractive in the scientific field, and implementing it in citation screening could save the investigative team significant work and decrease the time to publication.

Methods: Citations from the 2015 update of a pediatrics vitamin D systematic review were uploaded to an online platform designed for crowdsourcing the screening process (http://www.CHEORI.org/en/CrowdScreenOverview). Three sets of exclusion criteria were used for screening, with a review of abstracts at level one, and full-text eligibility determined through two screening stages. Two trained reviewers, who participated in the initial systematic review, established citation eligibility. In parallel, each citation received four independent assessments from an untrained crowd with a medical background. Citations were retained or excluded if they received three congruent assessments. Otherwise, they were reviewed by the principal investigator. Measured outcomes included sensitivity of the crowd to retain eligible studies, and potential work saved defined as citations sorted by the crowd (excluded or retained) without involvement of the principal investigator.

Results: A total of 148 citations for screening were identified, of which 20 met eligibility criteria (true positives). The four reviewers from the crowd agreed completely on 63% (95% CI: 57–69%) of assessments, and achieved a sensitivity of 100% (95% CI: 88–100%) and a specificity of 99% (95% CI: 96–100%). Potential work saved to the research team was 84% (95% CI: 77–89%) at the abstract screening stage, and 73% (95% CI: 67–79%) through all three levels. In addition, different thresholds for citation retention and exclusion were assessed. With an algorithm favoring sensitivity (citation excluded only if all four reviewers agree), sensitivity was maintained at 100%, with a decrease of potential work saved to 66% (95% CI: 59–71%). In contrast, increasing the threshold required for retention (exclude all citations not obtaining 3/4 retain assessments) decreased sensitivity to 85% (95% CI: 65–96%), while improving potential workload saved to 92% (95% CI: 88–95%).

Conclusions: This study demonstrates the accuracy of crowdsourcing for systematic review citations screening, with retention of all eligible articles and a significant reduction in the work required from the

investigative team. Together, these two findings suggest that crowdsourcing could represent a significant advancement in the area of systematic review. Future directions include further study to assess validity across medical fields and determination of the capacity of a non-medical crowd.

Keywords: Crowdsourcing; systematic reviews; citation screening; vitamin D

Submitted Aug 07, 2016. Accepted for publication Oct 28, 2016.
doi: 10.21037/tp.2016.12.01
View this article at: http://dx.doi.org/10.21037/tp.2016.12.01

Introduction

Systematic reviews are considered one of the cornerstones of evidence-based medicine, and can often support or refute the importance of a treatment or research idea with a higher level of confidence than an individual study (1). There are, however, recognized challenges to performing a well done systematic review with many never being finished, requiring years to complete, missing eligible trials, or becoming rapidly out of date (2).

To ensure complete identification of the relevant evidence base, investigators must search a variety of sources including multiple electronic databases (e.g., MEDLINE, Embase). This results in the retrieval of thousands or tens of thousands of citations (3-5), with only a small percentage (3–5%) ultimately meeting eligibility criteria (6). Accepted practice has each of the potential citations being evaluated in duplicate (independently). The time required to identify the eligible studies is considerable (7), and will only continue to increase given the rapid growth in scientific literature (8). Recent work suggests that researchers may already be utilizing search and screen approaches that negatively impact the systematic review process (9,10). Alternative methodological avenues that maintain, or possibly enhance, the validity of the systematic review processes have been recognized as desirable (11-13). For example, automated computer screening has been considered (14), where abstracts are ranked based on specific keywords. This method has failed to gain momentum due to inadequate validation, need for computer science expertise, and the fact that many investigators view abstract screening as a human intelligence task (14,15).

Abstract screening and full text evaluation are usually performed by a small group of highly trained experts. With considerable other demands on their time, this approach frequently leads to significant delays. An alternative would be to have a significant portion of the screening process performed by a large group of individuals with less specialized training and subject expertise. If feasible, this approach could significantly speed up the systematic review process. In essence, this idea amounts to crowdsourcing or "the process of obtaining needed service, ideas, or content by soliciting contributions from an online community rather than from traditional employees or suppliers." Over the past decade, this concept has been gaining in importance, with the introduction of Wikipedia as evidence of feasibility. In the biomedical areas, crowdsourcing has been used with success to gain wide input on clinical trials designs (16) and to assist in the prediction of complex biological structures (17). Crowdsourcing of abstract screening has been utilized in a previous project, although the accuracy of this process was not validated (18).

The objective of this project was to determine whether a crowd with no project specific training or expertise could accurately determine study eligibility for a systematic review.

Methods

We performed a validation study comparing the results of a systematic review performed through crowdsourcing to the findings generated using the gold-standard, trained experts approach. The systematic review was an update to a previously published study (9), and the protocol for this study was established a priori (PROSPERO protocol registration number: CRD42016038178). Results are reported according to the PRISMA guidelines for systematic reviews (*Table S1*) (19).

Identification of studies

The previously reported MEDLINE search strategy (9) was used. Our previously published systematic review included all citations up to January 2015. In this update, all

148 citations published between January 2015 and January 2016 and indexed on MEDLINE were included. This update was restricted to MEDLINE as 98% (166/169) of all eligible publications from the prior systematic review and 100% (79/79) of trials from the past 5 years were identified through MEDLINE. The search strategy (Appendix S1) was developed by a librarian (Margaret Sampson) and peer reviewed by a second (Lorie Kloda, MLIS, PhD), using the PRESS (Peer Review of Electronic Search Strategies) standard (20).

Screening was conducted using an online platform "CrowdScreen SR" designed for this study (http://www.CHEORI.org/en/CrowdScreenOverview). This program allows both the abstract and full text of citations to be uploaded so reviewers can assess eligibility at multiple screening levels. Study inclusion criteria were identical to those previously reported (*Table S2*) (9). At each level, reviewers were instructed to place citations into one of three groups: (I) retain; (II) exclude; or (III) unclear—I cannot assess this citation. When a citation was categorized as exclude the reviewer was prompted to indicate one or more eligibility criteria that were not met.

Review of citations by two experts (accepted gold standard approach) was performed as previously described (9). Data was extracted from eligible articles independently and in duplicates by two authors and entered into REDCap (21). The methods for stratification of study populations and vitamin D dosing regimens were consistent with the original systematic review (9). Each study was assessed using Cochrane risk of bias tool (22).

Crowd screening

Review of citations by the crowdsourcing arm proceeded in parallel. For this initial study we sought individuals with post-secondary education and a medical background (e.g., medical school, nursing) who had not provided input into the design of the systematic review protocol and had not received training sessions by the investigators on how to screen citations. These individuals were recruited at the Children's Hospital of Eastern Ontario and the Medical School at the University of Ottawa, by notifying members of a pediatric interest group. Reviewers had unique usernames and passwords, allowing separate tracking and evaluation of their progress. Initially, each reviewer was given access to a demo module for practice assessments on 16 abstracts and 9 full-text citations from the original systematic review. During the demo immediate feedback was provided on whether the reviewer's assessment of the abstract was accurate. Afterwards, reviewers started the formal screening process. Reviewers were not assigned to a fixed number of citations but were offered the flexibility to screen as many citations as they could at each of the screening levels. Citations were randomly distributed among reviewers.

Data collection and analysis

For this study, the decision was made to evaluate each citation a minimum of four times. At both abstract and full text screening levels, the assessments for each citation were categorized as shown in *Table S3*. In brief: (I) group 1: three or more retain assessments; (II) group 2: three or more exclude assessments; (III) group 3: any other combination of four assessments. The investigative team was only required to review citations that belonged to the third group, as well as the finally retained citations after the three levels of screening.

The outcome of primary interest was sensitivity, calculated by determining the number of trials retained by the investigators (true positives) that were also retained by the crowd after both abstract and full text screening. The second outcome of interest was the number of abstracts and full text assessments that the investigative team did not review, or work saved. This was calculated as number of citations retained or excluded solely by the crowd at the first two levels, and those excluded at the third (under the assumption that the investigative team would confirm full eligibility of all retained studies) (*Table S3*). This was presented as the percentage of all abstracts and full texts being reviewed, and was labeled as work saved, consistent with other reports in the field. Jeffreys interval was used to calculate 95% confidence for sensitivity, specificity and work saved (23).

Results

Systematic review update

Figure 1 demonstrates the flow of studies identified by the search strategy, as per the gold standard screening method (i.e., two trained reviewers). A total of 148 unique records were retrieved from the electronic search, of which 99 were excluded at level one, with an additional 14 excluded at level two screening. In total, we identified 35 publications that reported on the results of a clinical trial

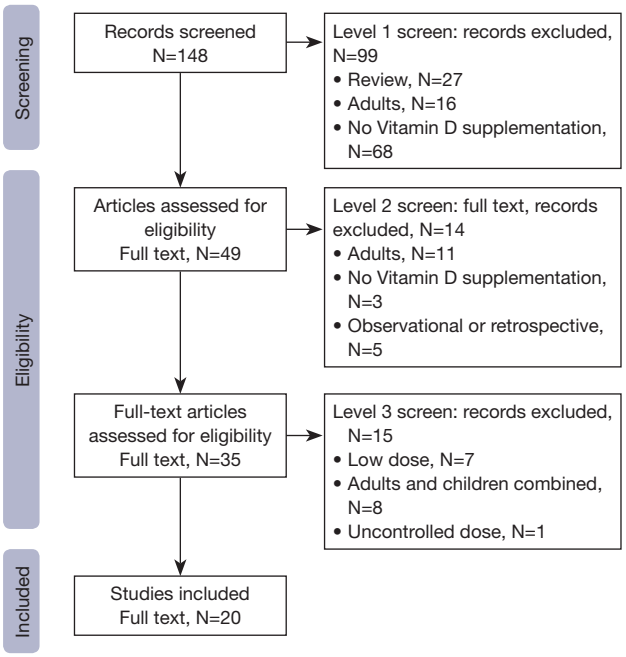

Figure 1 Flow chart of study selection based on inclusion and exclusion criteria. Numbers will exceed the total because studies could be excluded for multiple reasons.

administering ergocalciferol or cholecalciferol to children. From these, 20 articles met eligibility criteria for high dose supplementation (Appendix S2), representing 17 distinct trials (*Table 1*). Reviewing cited references failed to identify any additional trials. Evaluation of the 17 included studies showed that 19 different populations and 39 distinct arms were described (*Figure S1*). Details on population, dosing and methodology are provided in *Table S4*.

Crowdsourcing

The crowdsourcing arm was composed of eight reviewers: six medical students, one pediatrics subspecialty fellow and one nurse. One student withdrew from the study prior to starting the screening process due to time constraints. All the medical students were early in their training (two were first year and three were in their second year). On average, the remaining reviewers assessed 12.3 demo citations (range, 2–37), and correctly classified 91% of demos (range, 70–100%). During the formal screening process, the seven reviewers evaluated an average of 170 citations (range, 74–233) when abstract and full text screening were combined. Four reviewers contributed to all three stages of screening, two reviewers assisted with two levels, and the remaining reviewer contributed to only abstract screening. For all three levels of screening, each citation was classified according to the distribution of reviewer assessments (*Figure 2*). For comparison, the number of eligible citations as determined by the experts is also shown.

Crowd sensitivity

Sensitivity of the crowd for retaining eligible studies was 100% (20/20, 95% CI: 88–100%) when only those citations receiving 3 or 4 exclude assessments were discarded (*Table 2* and *Table S5*). Crowdsourcing reviewers agreed completely on the assessment (i.e., 4 retain or 4 exclude) in 65% (96/148) of the citations at the abstract stage and 60% (50/84) at full text review. Only three eligible articles required review by the PI for level 3 due to disagreement among the crowd. Otherwise, the remaining 17 articles passed the screening without any review from the PI. When all three screening stages were considered, specificity of the crowd was estimated at 99% (127/128, 95% CI: 96–100%). The lone ineligible article that was categorized as eligible with three retain assessments was a published protocol for a randomized controlled trial (RCT) that would have met systematic review eligibility once completed (24). Due to the small number of participants it was not possible to analyze for differences in performance by subgroup (25).

Crowd efficiency

Citations that were sorted by the crowd without involvement of the principal investigator were considered as potential work saved to the investigative team. With this approach, only the citations not receiving three or more congruent assessments and the final set of retained citations remained as work requiring assessment by the investigative group (*Figure 2*). As such, the crowd automatically classified (retained or excluded) 84% (124/148) of the abstracts (level 1) and 71% (35/49) of full texts in level 2 (*Table S5*). In the final screening stage 31% (11/35) of remaining citations were excluded by the crowd. Combined, the work saved throughout the whole screening process was 73% (170/232, 95% CI: 67–79%). Additionally, we assessed the change in sensitivity and work saved that occurred with modification of the threshold for retaining or excluding citations at all three levels of screening (

Table 1 Descriptive characteristics of eligible studies. Study and population descriptions of the 17 original trials identified in the screening process, representing studies of high-dose vitamin D supplementation in pediatrics, published between Jan 2015–Jan 2016

Author, year	Location	Study design	Population, age	n, N[†]	High-dose regimens	Primary outcome	Risk of bias
Vehapoglu, 2015	Middle East	Single arm	Healthy with growing pain, 4–12 y	148, 148	Single dose PO: 150,000 IU (<6 years) or 300,000 IU (>6 years)	25 OHD, pain	High
Aytac, 2016	Middle East	Paralleled	Renal disease, 10–17 y	65, 65	300,000 IU single D3 PO	Cardiovascular	High
Le Roy, 2015	C/S. America	RCT	Cerebral palsy, 6–14 y	15, 30	100,000 IU single D3 PO	25 OHD	Low
Cayir, 2014	Middle East	Paralleled	Migraines, 8–16 y	27, 53	800 or 5,000 IU daily PO	Migraine attacks	Medium
Rajakumar, 2015	N. America	RCT	Healthy/VDD, 8–14 y	78, 157	1,000 IU daily D3 PO	25 OHD	Low
Mayes, 2015	N. America	RCT	Burns, 6 m–19 y	26, 39	100 IU/Kg daily PO, D2 *vs*. D3	Fracture risk	Medium
Shah, 2015	N. America	RCT	Obesity, 11–18 y	20, 40	150,000 IU Q3 months D2 PO	25 OHD	Low
Simek, 2016	N. America	RCT	IBD, 8–21 y	34, 40	5,000 IU/10 kg or 10,000 IU/10 kg weekly D3 PO	25 OHD	Low
Tan, 2015	Australia	RCT	Healthy/VDD, 5–9 y	37, 73	Daily D3 PO: 5,000 IU (25 OHD <27.5), 2,500 IU (25 OHD >27.5) *vs*. single D3 PO: 200,000 IU (25 OHD <27.5), 100,000 (25 OHD >27.5)	25 OHD	Medium
Hanson, 2015	N. America	RCT	Premature newborns	32, 32	400 or 800 IU daily D3 PO	25 OHD	Low
Dougherty, 2015	N. America	RCT	Sickle cell, 5–20 y	44, 44	4,000 or 7,000 IU daily D3 PO	25 OHD	Low
Galli, 2015	Europe	RCT	Eczema, 6 m–16 y	41, 89	2,000 IU daily D3 PO	25 OHD	Low
Morandi, 2015	Europe	Single arm	Healthy/VDD, 3–15 y	33, 33	100,000 IU monthly D3: IM (25 OHD <10) or PO (25 OHD 10–20); 25,000 IU monthly D3 PO (25 OHD 20–30)	Pain	High
Moodley, 2015	C/S. America	RCT	Healthy newborns	18, 51	50,000 IU single D3 PO	25 OHD	Low
Eltayeb, 2015	Middle East	RCT	Hepatitis C, 7–14 y	31, 60	2,000 IU daily D3 PO	HCV RNA level	Medium
Dubnov-Raz, 2015	Middle East	RCT	Healthy/VDD, 12–21 y	28, 55	2,000 IU daily D3 PO	Respiratory marker	Low
Steenhoff, 2015	Africa	RCT	HIV, 5–51 y	60, 60	4,000 or 7,000 IU daily D3 PO	25 OHD	Low

[†], "N" refers to the total number of patients enrolled in the study, while "n" includes only those who have received high dose regimens. 25 OHD, 25-Hydroxycholecalciferol; C/S. America, Central and South America; D2, ergocalciferol; D3, cholecalciferol; HCV, hepatitis C virus; HIV, human immunodeficiency virus; IBD, inflammatory bowel disease; IM, intramuscular; IU, international units; N. America, North America; PO, oral; Q3 months, every 3 months; RCT, randomized controlled trial; VDD, vitamin D deficiency.

Citation assessment profile			Abstract screening		Full text review			
			Level 1		Level 2		Level 3	
Retain	Exclude	Unclear	Citations	Eligible	Citations	Eligible	Citations	Eligible
4	0	0	25	15	26	17	11	11
3	0 or 1	0 or 1	11	5	6	3	7	6
Other combination			24		14		6	3
0 or 1	3	0 or 1	17		0		1	
0	4	0	71		3		10	
Total			148		49		35	

Figure 2 Assessment of 148 citations by the crowd at each screening level. Four reviewers assessed each citation and selected one of three options (retain, exclude, unclear). Citations were stratified depending on the combination of the four assessments. Number of eligible citations refers to the true positives identified by the gold-standard approach (two trained experts).

Table 2 Contingency table comparing assessment by the crowd to the gold-standard two trained experts

Crowd	Experts		Value (95% CI)
	Retain	Exclude	
Retain	20	1	PPV =95% (80–99%)
Exclude	0	127	NPV =100% (98–100%)
Value (95% CI)	Sensitivity =100% (88–100%)	Specificity =99% (67–79%)	–

Sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) are provided as a percentage with a 95% confidence interval.

). Using a more conservative approach and only discarding citations with four exclude assessments maintained the sensitivity at 100% (20/20, 95% CI: 88–100%) and decreased work saved to 66% (152/232, 95% CI: 59–71%). When the approach was changed to remove all citations unless they received 3 or 4 retain assessments, the workload saved improved to 92% (214/232, 95% CI: 88–95%), although sensitivity declined to 85% (17/20, 95% CI: 65–96%).

Discussion

The challenges associated with completing and updating large reviews are well described (11-13). In this pilot study, we sought to demonstrate whether it was feasible to crowdsource important time consuming steps of the systematic review process. Our main study finding were that a crowd of individuals with no subject-specific expertise and no input into the systematic review protocol was able to retain 100% of the eligible citations through screening, while reducing the work required of the investigative team to just 27%.

The ability to identify and retain studies that meet systematic review eligibility criteria is the most important outcome when evaluating an alternative or adjunctive methodology for abstract and full text screening (14). The ability of the crowd to retain 100% of eligible studies throughout the entire screening process in our study exceeds the 95% cut-off utilized in the computerized automated text screening literature to identify promising algorithms (12,26). As a role for crowdsourcing in systematic reviews is a novel idea, there are no published validation studies for comparison. The only published study in this area, by Brown and colleagues, had an online crowd complete a systematic review in the area of nutrition, but did not compare crowd responses to gold standard (18). Although unpublished, the most relevant findings for comparison have been described as part of an ongoing Embase project where volunteers screen abstracts to identify those representing RCTs on humans. As part of this work the investigators performed a nested validation study and reported 99% sensitivity, further supporting the idea (27,28). Although suggestive, findings from the Embase project have limited applicability as the abstracts may have been preselected, were not evaluated against a full set of systematic review eligibility criteria, and full text screening was not evaluated.

Although high sensitivity is essential, crowdsourcing is only valuable if it also reduces investigator workload. A recent systematic review of studies evaluating computerized text recognition identified work saved as one of the

most common measures evaluated (14). In our study, we calculated that crowdsourcing would have reduced investigator workload by 84% for abstract screening, exceeding estimates in all but a few of the text mining studies (14). It is important to recognize that based on comfort level and the size of literature, investigators may choose alternative approaches that prioritize either sensitivity or work saved. When considering an algorithm that prioritized sensitivity, the work saved at the abstract stage was 73%. Although lower, it still outperformed many of the computerized text recognition studies. In contrast, an algorithm that prioritized work saved was able to further decrease the work required across three levels by the investigative team to just 8%, but reduced sensitivity to 85%. Although sensitivity of 85% may be a cause for concern, it is important to consider that there is increasing evidence that current search and screen approaches may achieve sensitivities well below 85%. For example, a recent analysis by Créquit *et al*. of 29 systematic reviews on lung cancer showed that these reviews missed 46% (n=34) of trials and 30% (n=8,486) of patients that were eligible and published prior to publication date (10). Furthermore, our recent systematic review identifying all high dose vitamin D trials in children demonstrated that individual systematic reviews missed 28% of eligible trials (9). Even if perfect sensitivity is not achieved in future studies, crowdsourcing may ultimately improve on the proportion of eligible trials identified if investigator groups are able to incorporate less specific search terms. In addition to initial systematic reviews, crowdsourcing may facilitate updates of previously published reviews, or contribute to real-time up-to-date online "living systematic reviews" (29,30).

In addition to appropriate sensitivity and work saved, it is important to acknowledge that crowdsourcing will not become an established mainstream methodology without attention to feasibility. Feasibility has been acknowledged as one of the factors preventing adoption of automated text recognition for abstract screening, as most investigators are uncomfortable with the technology and individuals with appropriate expertise are scarce (14). For crowdsourcing to be widely adopted, it will be necessary to create a software platform that allows investigators to upload citations, define eligibility criteria, individualize parameters for retaining citations, and provide access to a large crowd of online individuals. Although both the Amazon Mechanical Turk and Embase projects do support feasibility, neither presents a resource that could be easily adapted by others to their specific study (18). The second aspect of feasibility that needs exploration is crowd motivation. Experience with crowdsourcing in other fields suggests that this may not be a problem. For example, the 57,000 crowd members from the general public helped in the FoldIt project, by participating in an online game aimed at determining the most stable structure of specific proteins (17). With respect to systematic reviews, individuals and groups may be motivated to assist for a number of reasons including personal interest in the topic, research experience, educational credit (course work), altruism (volunteers), financial benefit, or academic advancement (authorship). While the feasibility study by Brown might be taken to suggest that individuals will only do the work if paid, it is important to note that the crowd received only a relatively small payment ($0.07 per citation) (18). In our study, although authorship was eventually offered to those participants who met criteria, the initial advertisement requested volunteers, and many individuals willing to participate were turned away. Finally, and although unpublished, the greatest evidence to support feasibility comes from the Embase project where an online community of volunteers has assessed approximately 100,000 abstracts for free (27,28).

Despite findings that support the ability to crowdsource parts of the systematic review, it is important to highlight study limitations. First, despite calculating 100% sensitivity, given the relatively small number of eligible citations in the study the true sensitivity may be lower. Second, our results are based on citations and criteria from a single systematic review, making it unclear how to generalize findings to other fields and research questions. Third, this study focused on evaluating sensitivity and work saved among a crowd of individuals with medical training that lacked content expertise and training on the screening process. Although the size of the online community without medical training is much larger, and therefore would be much more powerful, it was felt that this smaller crowd should be evaluated first. Consequently, it is important that our findings must not be extrapolated to individuals with little medical or scientific training until those studies have been completed. The validity of this approach in a wider variety of methodologies and fields remains to be assessed. Even with further validation in other setting, some investigators may dismiss the innovation over concern that characteristics and complexity related to their review will make the approach invalid. Consequently, crowdsourcing software should be designed to evaluate individual reviewer performance before and throughout study participation.

Due to the large and increasing body of published

literature, investigators are struggling to publish comprehensive up-to-date systematic reviews (2). Crowdsourcing has the potential to lead to faster and more complete knowledge synthesis efforts by simultaneously allowing for the use of broader search terms, increasing the speed of citation screening, and freeing up investigative team time to focus on other aspects of the project. In comparison, our study provides initial proof of concept and additional larger scale studies should be performed to confirm or refute these promising results. Future directions include assessing the validity of crowdsourcing in a variety of medical and scientific fields, the capacity of different crowds (healthcare professionals, undergraduate students, hospital volunteers), an evaluation of how individual education and experience influences accuracy, and exploration of the educational benefits of crowdsourcing.

Acknowledgements

We thank Ms. Tharshika Thangarasa and Mrs. Colleen Fitzgibbons for assistance in citations screening.

Footnote

Conflicts of Interest: The authors have no conflicts of interest to declare.

References

1. Pearson A, Wiechula R, Court A, et al. The JBI model of evidence-based healthcare. Int J Evid Based Healthc 2005;3:207-15.
2. Shojania KG, Sampson M, Ansari MT, et al. How quickly do systematic reviews go out of date? A survival analysis. Ann Intern Med 2007;147:224-33.
3. Martin A, Saunders DH, Shenkin SD, et al. Lifestyle intervention for improving school achievement in overweight or obese children and adolescents. Cochrane Database Syst Rev 2014;(3):CD009728.
4. Fletcher-Watson S, McConnell F, Manola E, et al. Interventions based on the Theory of Mind cognitive model for autism spectrum disorder (ASD). Cochrane Database Syst Rev 2014;(3):CD008785.
5. Lavoie MC, Verbeek JH, Pahwa M. Devices for preventing percutaneous exposure injuries caused by needles in healthcare personnel. Cochrane Database Syst Rev 2014;(3):CD009740.
6. Sampson M, Tetzlaff J, Urquhart C. Precision of healthcare systematic review searches in a cross-sectional sample. Res Synth Methods 2011;2:119-25.
7. Allen IE, Olkin I. Estimating time to conduct a meta-analysis from number of citations retrieved. JAMA 1999;282:634-5.
8. Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? PLoS Med 2010;7:e1000326.
9. Nama N, Menon K, Iliriani K, et al. A systematic review of pediatric clinical trials of high dose vitamin D. PeerJ 2016;4:e1701.
10. Créquit P, Trinquart L, Yavchitz A, et al. Wasted research when systematic reviews fail to provide a complete and up-to-date evidence synthesis: the example of lung cancer. BMC Med 2016;14:8.
11. Matwin S, Kouznetsov A, Inkpen D, et al. A new algorithm for reducing the workload of experts in performing systematic reviews. J Am Med Inform Assoc 2010;17:446-53.
12. Cohen AM, Hersh WR, Peterson K, et al. Reducing workload in systematic review preparation using automated citation classification. J Am Med Inform Assoc 2006;13:206-19.
13. Jonnalagadda S, Petitti D. A new iterative method to reduce workload in systematic review process. Int J Comput Biol Drug Des 2013;6:5-17.
14. O'Mara-Eves A, Thomas J, McNaught J, et al. Using text mining for study identification in systematic reviews: a systematic review of current approaches. Syst Rev 2015;4:5.
15. Thomas J, McNaught J, Ananiadou S. Applications of text mining within systematic reviews. Res Synth Methods 2011;2:1-14.
16. Leiter A, Sablinski T, Diefenbach M, et al. Use of crowdsourcing for cancer clinical trial development. J Natl Cancer Inst 2014;106. pii: dju258.
17. Good BM, Su AI. Games with a scientific purpose. Genome Biol 2011;12:135.
18. Brown AW, Allison DB. Using Crowdsourcing to Evaluate Published Scientific Literature: Methods and Example. PLoS One 9:e100647.
19. 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.
20. Sampson M, McGowan J, Cogo E, et al. An evidence-based practice guideline for the peer review of electronic search strategies. J Clin Epidemiol 2009;62:944-52.
21. Harris PA, Taylor R, Thielke R, et al. Research electronic

data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009;42:377-81.
22. Higgins JP, Green S. Front Matter, in Cochrane Handbook for Systematic Reviews of Interventions: Cochrane Book Series. John Wiley & Sons, Ltd, Chichester, UK 2008.
23. Brown LD, Cai TT, DasGupta A. Interval Estimation for a Binomial Proportion. Statistical Science 2001;16:101-17.
24. McNally JD, O'Hearn K, Lawson ML, et al. Prevention of vitamin D deficiency in children following cardiac surgery: study protocol for a randomized controlled trial. Trials 2015;16:402.
25. Hrynaszkiewicz I, Norton ML, Vickers AJ, et al. Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. BMJ 2010;340:c181.
26. Cohen AM. Optimizing feature representation for automated systematic review work prioritization. AMIA Annu Symp Proc 2008:121-5.
27. Tsertsvadze A, Chen YF, Moher D, et al. How to conduct systematic reviews more expeditiously? Syst Rev 2015;4:160.
28. Elliott J, Sim I, Thomas J, et al. #CochraneTech: technology and the future of systematic reviews. Cochrane Database Syst Rev 2014;(9):ED000091.
29. Elliott JH, Turner T, Clavisi O, et al. Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. PLoS Med 2014;11:e1001603.
30. Badgett RG, Vindhyal M, Stirnaman JT, et al. A Living Systematic Review of Nebulized Hypertonic Saline for Acute Bronchiolitis in Infants. JAMA Pediatr 2015;169:788-9.

Cite this article as: Nama N, Iliriani K, Xia MY, Chen BP, Zhou LL, Pojsupap S, Kappel C, O'Hearn K, Sampson M, Menon K, McNally JD. A pilot validation study of crowdsourcing systematic reviews: update of a searchable database of pediatric clinical trials of high-dose vitamin D. Transl Pediatr 2017;6(1):18-26. doi: 10.21037/tp.2016.12.01

Supplementary

Appendix S1 MEDLINE search strategy

1. exp Vitamin d/
2. (vitamin adj (d or d2 or d3)).tw.
3. Calcifediol/
4. calcidiol.tw.
5. Ergocalciferols/
6. Ergocalciferol$.tw.
7. Cholecalciferol/
8. Cholecalciferol$.tw.
9. calciferol.tw.
10. Vitamin D Deficiency/dh, dt
11. or/1-10
12. (25-hydroxyvitamin D or 25-hydroxy vitamin d or Plasma vitamin D).tw.
13. 64719-49-9.rn.
14. 25OHD3.tw.
15. "25(OH)D3".tw.
16. 25-OHD3.tw.
17. "25-(OH)D3".tw.
18. 25OHD.tw.
19. "25(OH)D".tw.
20. 25-OHD.tw.
21. "25-(OH)D".tw.
22. (25-hydroxycholecalciferol or 25-hydroxyergocalciferol).tw.
23. Calcium/bl, ur
24. plasma calcidiol.tw.
25. (Urine calcium or (calcium adj3 ratio)).tw.
26. or/12–25
27. exp Vitamin D Deficiency/ not Vitamin D Deficiency/dh, dt
28. (avitaminosis and (d or d2 or d3)).tw.
29. Vitamin D/to
30. No-Observed-Adverse-Effect Level/
31. upper limit$.tw.
32. UL.tw.
33. (excess$ or toxic$).tw.
34. (noael or noel).tw.
35. (no observed adj2 effect$).tw.
36. Calcification, Physiologic/de
37. Hypercalcemia/
38. Kidney Calculi/
39. Nephrocalcinosis/
40. Urinary Calculi/
41. Bladder Calculi/
42. Ureteral Calculi/
43. Calcinosis/
44. Hypercalcemi$.tw.
45. (Burnett$ adj2 syndrome$).tw.
46. Hypercalciuri$.tw.
47. exp Vitamin d/ae or Calcifediol/ae or Ergocalciferols/ae or Cholecalciferol/ae
48. (Side effect* or adverse effect$).tw.
49. or/27–48
50. 11 and (26 or 49)
51. ((randomized controlled trial or controlled clinical trial).pt. or randomized.ab. or placebo.ab. or clinical trials as topic.sh. or randomly.ab. or trial.ti.) not (exp animals/ not humans.sh.)
52. (Single arm or pilot or cross-over or n-of-1).tw.
53. Double-blind Method/ or Single-blind Method/
54. (clin$ adj25 trial$).ti,ab.
55. ((singl$ or doubl$ or trebl$ or tripl$) adj25 (blind$ or mask$)).ti,ab.
56. Placebos/
57. 50 and (or/51–56)
58. limit 50 to clinical trial, all
59. or/57–58
60. ((single adj2 dose) or bolus or stoss* or single day or mega*).tw.
61. Dose-Response Relationship, Drug/
62. 60 or 61
63. 11 and 62
64. 59 or 63
65. 64 and (child* or adolescent or infan*).mp.
66. 64 and ((Infan* or newborn* or new-born* or perinat* or neonat* or baby or baby* or babies or toddler* or minors or minors* or boy or boys or boyfriend or boyhood or girl* or kid or kids or child or child* or children* or schoolchild* or schoolchild).mp. or school child.ti,ab. or school child*.ti,ab. or (adolescen* or juvenil* or youth* or teen* or under*age* or pubescen*).mp. or exp pediatrics/ or (pediatric* or paediatric* or peadiatric*).mp. or school.ti,ab. or school*.ti,ab. or (prematur* or preterm*).mp.)
67. limit 66 to ("in data review" or in process or "pubmed not medline")
68. 65 or 67

Appendix S2 List of 20 trials of high-dose pediatrics vitamin D supplementation (31–50)

31. Cayir A, Turan MI, Tan H. Effect of vitamin D therapy in addition to amitriptyline on migraine attacks in pediatric patients. Braz J Med Biol Res 2014;47:349-54.
32. Gallo S, Comeau K, Sharma A, et al. Redefining normal

bone and mineral clinical biochemistry reference intervals for healthy infants in Canada. Clin Biochem 2014;47:27-32.
33. Steenhoff AP, Schall JI, Samuel J, et al. Vitamin D3supplementation in Batswana children and adults with HIV: a pilot double blind randomized controlled trial. PLoS One 2015;10:e0117123.
34. Vehapoglu A, Turel O, Turkmen S, et al. Are Growing Pains Related to Vitamin D Deficiency? Efficacy of Vitamin D Therapy for Resolution of Symptoms. Med Princ Pract 2015;24:332-8.
35. Eltayeb AA, Abdou MA, Abdel-aal AM, et al. Vitamin D status and viral response to therapy in hepatitis C infected children. World J Gastroenterol 2015;21:1284-91.
36. Dubnov-Raz G, Rinat B, Hemilä H, et al. Vitamin D supplementation and upper respiratory tract infections in adolescent swimmers: a randomized controlled trial. Pediatr Exerc Sci 2015;27:113-9.
37. Galli E, Rocchi L, Carello R, et al. Serum Vitamin D levels and Vitamin D supplementation do not correlate with the severity of chronic eczema in children. Eur Ann Allergy Clin Immunol 2015;47:41-7.
38. Morandi G, Maines E, Piona C, et al. Significant association among growing pains, vitamin D supplementation, and bone mineral status: results from a pilot cohort study. J Bone Miner Metab 2015;33:201-6.
39. Moodley A, Spector SA. Single high-dose vitamin D at birth corrects vitamin D deficiency in infants in Mexico. Int J Food Sci Nutr 2015;66:336-41.
40. Tan JK, Kearns P, Martin AC, et al. Randomised controlled trial of daily versus stoss vitamin D therapy in Aboriginal children. J Paediatr Child Health 2015;51:626-31.
41. Dougherty KA, Bertolaso C, Schall JI, et al. Safety and Efficacy of High-dose Daily Vitamin D3 Supplementation in Children and Young Adults With Sickle Cell Disease. J Pediatr Hematol Oncol 2015;37:e308-15.
42. Shah S, Wilson DM, Bachrach LK. Large Doses of Vitamin D Fail to Increase 25-Hydroxyvitamin D Levels or to Alter Cardiovascular Risk Factors in Obese Adolescents: A Pilot Study. J Adolesc Health 2015;57:19-23.
43. Rajakumar K, Moore CG, Yabes J, et al. Effect of Vitamin D3 Supplementation in Black and in White Children: A Randomized, Placebo-Controlled Trial. J Clin Endocrinol Metab 2015;100:3183-92.
44. Dubnov-Raz G, Livne N, Raz R, et al. Vitamin D Supplementation and Physical Performance in Adolescent Swimmers. Int J Sport Nutr Exerc Metab 2015;25:317-25.
45. Hanson C, Lyden E, Nelson A, et al. Response of vitamin D binding protein and free vitamin D concentrations to vitamin D supplementation in hospitalized premature infants. J Pediatr Endocrinol Metab 2015;28:1107-14.
46. Mayan I, Somech R, Lev A, et al. Thymus Activity, Vitamin D, and Respiratory Infections in Adolescent Swimmers. Isr Med Assoc J 2015;17:571-5.
47. Le Roy C, Meier M, Witting S, et al. Effect of supplementation with a single dose of vitamin D in children with cerebral palsy. Preliminary randomised controlled study. Rev Chil Pediatr 2015;86:393-8.
48. Mayes T, Gottschlich MM, Khoury J, et al. Investigation of Bone Health Subsequent to Vitamin D Supplementation in Children Following Burn Injury. Nutr Clin Pract 2015;30:830-7.
49. Simek RZ, Prince J, Syed S, et al. Pilot Study Evaluating Efficacy of 2 Regimens for Hypovitaminosis D Repletion in Pediatric Inflammatory Bowel Disease. J Pediatr Gastroenterol Nutr 2016;62:252-8.
50. Aytaç MB, Deveci M, Bek K, et al. Effect of cholecalciferol on local arterial stiffness and endothelial dysfunction in children with chronic kidney disease. Pediatr Nephrol 2016;31:267-77.

Table S1 PRISMA 2009 checklist

Section/topic	Item#	Checklist item	Reported on page #
Title			
Title	1	Identify the report as a systematic review, meta-analysis, or both	18
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	18-19
Introduction			
Rationale	3	Describe the rationale for the review in the context of what is already known	19
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS)	19
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	19
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	*Table S2*
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	19-20
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated	Appendix S1
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)	20
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	20
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made	*Table 1*
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	20
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means)	*Table S4*
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	20-23
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)	*Table 1*
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified	21, 23
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	*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	*Table 1*
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)	*Table 1*
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (I) simple summary data for each intervention group; (II) effect estimates and confidence intervals, ideally with a forest plot	*Table 1*
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency	N/A
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see item 15)	*Table S4*
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)	23-25
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)	24
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research	24-25
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	25

From: 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. #, number; N/A, not available.

Table S2 Screening criteria

Level	Screening criteria
Level 1	The citation is not a review article (or case report)
	The study is on humans and children are included
	The study administers at least one dose of vitamin D (cholecalciferol, ergocalciferol) to the patient
	The citation does not represent a conference abstract
Level 2	At least one study arm (group) includes children
	At least one dose of vitamin D (ergocalciferol and/or cholecalciferol) was administered
	Vitamin D was administered at one or more doses determined by the investigators
	The citation is in English, French, German or Spanish
Level 3	One or more study arms provide vitamin D supplementation at a dose that r exceeds the IOM age specific Recommended Dietary Allowance (RDA) or Adequate Intake (AI)
	If the study included adults, the information for pediatrics population is presented separately
	Vitamin D was not administered mixed with food and in uncontrolled volume

Table S3 Citations disposition based on crowd's assessment

Retain	Exclude	Unclear	Citation disposition
4	0	0	Retain
3	1	0	Retain
3	0	1	Retain
2	2	0	Citation reviewed by investigative team
2	1	1	Citation reviewed by investigative team
2	0	2	Citation reviewed by investigative team
1	2	1	Citation reviewed by investigative team
1	1	2	Citation reviewed by investigative team
1	0	3	Citation reviewed by investigative team
0	2	2	Citation reviewed by investigative team
0	1	3	Citation reviewed by investigative team
0	0	4	Citation reviewed by investigative team
1	3	0	Exclude
0	3	1	Exclude
0	4	0	Exclude

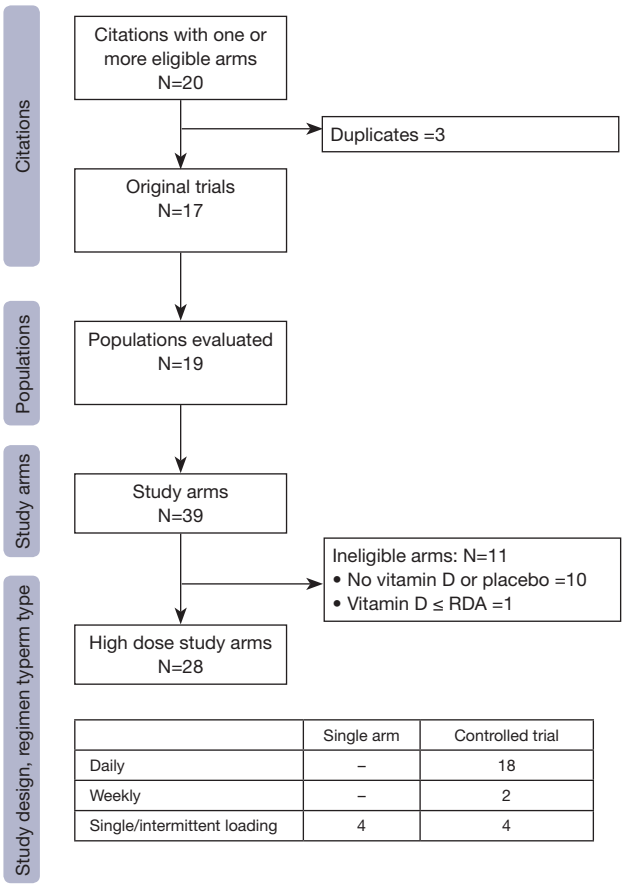

Figure S1 Flow of study arms.

Table S4 Assessment of study design, populations, supplementation and methodological quality

Study characteristics	n	%
Study design[†]		
RCT/qRCT	13	76
Single arm	3	18
Controlled, other	1	6
Randomized trial quality[†,‡]		
Low risk	11	65
Medium risk/unclear	3	18
High risk	3	18
Age groups[§]		
Neonates	4	22
Infants	4	22
Toddlers	6	33
Schoolers	16	89
Adolescents	15	83
Population[§]		
Healthy/subclinical VDD	8	42
Classical diseases	3	16
Non-classical diseases	8	42
Dosing regimen[¶]		
Constant	18	64
Variable	10	36
Dosing groups[¶]		
RDA/AI-999	3	11
1,000–3,999	5	18
4,000–39,999	11	39
≥40,000	9	32
Frequency[¶]		
Daily	18	64
Intermittent/single dose	8	29
Weekly/biweekly	2	7

(q)RCT, (Quasi) randomized controlled trial. [†], values represent the number of trials, and the percentage out of the 17 identified trials; [‡], studies were assessed using Cochrane risk of bias tool (Higgins & Green, 2008); [§], numbers of populations out of 19. For age, numbers will add up to more than 19 populations as some included children from two or more groups; [¶], number of high-dose vitamin D arms out of 28 identified in eligible studies.

Table S5 Sensitivity and work saved by the crowd at each screening level

Measure	Estimated value (n, 95% CI)
Sensitivity	100% (20/20, 95% CI: 88–100%)
Specificity	99% (127/128, 95% CI: 96–100%)
Work saved	
Level 1	84% (124/148, 95% CI: 77–89%)
Level 2	71% (35/49, 95% CI: 58–83%)
Level 3	31% (11/35, 95% CI: 18–48%)
Total	73% (170/232, 95% CI: 67–79%)

Table S6 Comparison of three different approaches for study disposition based on crowd's assessments

Approach	Crowd's assessments	Sensitivity	Work saved[†]
Low-risk	Exclude =4, retain ≥3, PI for other combinations	100% (20/20)	66% (152/232)
Medium-risk	Exclude ≥3, retain ≥3, PI for other combinations	100% (20/20)	73% (170/232)
High-risk	Retain ≥3, exclude the rest	85% (17/20)	92% (214/232)

[†], potential work saved to the investigative team, calculated as the percentage of trials that were excluded at any level as well as those that were retained through the first two levels without input from PI.