Developing Criteria and Associated Instructions for Consistent and Useful Quality Improvement Study Data Extraction for Health Systems

Adrian V. Hernandez, MD, PhD[1,2], Yuani M. Roman, MD, MPH[1,2], and C. Michael White, Pharm. D., FCP, FCCP[1,2,3]

[1]School of Pharmacy, University of Connecticut Evidence-based Practice Center, Storrs, CT, USA; [2]Department of Pharmacy, Hartford Hospital, Hartford, CT, USA; [3]University of Connecticut School of Pharmacy, Storrs, CT, USA.

BACKGROUND: The Agency for Healthcare Research and Quality (AHRQ) could devote resources to collate and assess quality improvement studies to support learning health systems (LHS) but there is no reliable data on the consistency of data extraction for important criteria.

METHODS: We identified quality improvement studies and evaluated the consistency of data extraction from two experienced independent reviewers at three time points: baseline, first revision (where explicit instructions for each criterion were created), and final revision (where the instructions were revised). Six investigators looked at the data extracted by the two systematic reviewers and determined the extent of similarity on a scale of 0 to 10 (where 0 represented no similarity and 10 perfect similarity). There were 42 assessments for baseline, 42 assessments for the first revision, and 42 assessments for the final revision. We asked two LHS participants to assess the relative value of our criteria.

RESULTS: The consistency of extraction improved from 1.17 ± 1.85 at baseline to 6.07 ± 2.76 after revision 1 ($P < 0.001$) and to 6.81 ± 1.94 out of 10 for the final revision ($P < 0.001$). However, the final revision was not significantly improved over the first revision ($P = 0.14$). One key informant rated the difficulty in finding and using quality improvement studies a 6 (moderately difficult) while the other a 4 (moderately difficult). When asked how valuable it would be if AHRQ found and collated the demographic information about the health systems and the interventions used in published quality improvement studies, they rated it a 9 (highly valuable) and a 6 (moderately valuable).

CONCLUSION: Creating explicit instructions for extracting data for quality improvement studies helps enhance the consistency of data extraction. This is important because it is difficult for LHS to vet these quality improvement studies on their own and they would value AHRQ's support in that regard.

KEY WORDS: quality improvement; data extraction; learning health system.

Electronic supplementary material *The online version of this article (https://doi.org/10.1007/s11606-020-06098-1) contains supplementary material, which is available to authorized users.*

Received August 26, 2019
Revised March 13, 2020
Accepted July 30, 2020
Published online August 17, 2020

J Gen Intern Med 35(Suppl 2):S802–S7
DOI: 10.1007/s11606-020-06098-1

INTRODUCTION

Quality patient care is an important part of health system accreditation and is increasingly tied to revenue through the hospital value-based purchasing program.[1–3] In 2020, 55% of health systems will see added net revenue from hitting performance metrics while others will receive a financial penalty.[2] In its report, Best Care at Lower Cost: The Path to Continuously Learning Health Care in America, the Health and Medicine Division of the National Academy of Sciences proposed the concept of the learning health system to improve healthcare quality and outcomes.[4] Regular health systems become learning health systems when they internally identify areas of quality weakness and strive to utilize the best knowledge to improve them.[5, 6]

The Agency for Healthcare Research and Quality (AHRQ) created a working group in 2017 that interviewed nine leaders in quality and safety improvement in health systems to explore how they used evidence.[7] Health systems looked to other institutions within their buying group, institutions in a formal consortium, or institutions reporting their experiences in the biomedical literature, for quality improvement approaches. However, finding the full spectrum of quality improvement projects completed in an area from the medical literature is difficult. There are many ways to improve the quality of care including purchasing new equipment or software, dedicating staff to champion changes, educational programs, email reminders, electronic health record reminders, internal protocols and guidelines, prior authorization, and limiting utilization to a specialist or specialty service. Faced with multiple approaches to improve, health systems need to identify approaches that overlap with their strengths and are realistic with their fiscal and staffing reality. The best approach in one health system may not translate well to another and, in some cases, may not even be feasible.

The AHRQ Evidence-based Practice Center (EPC) program created a report on closing the quality gap in asthma care.[8, 9]

The EPC program consists of institutions in the USA and Canada that review relevant scientific literature on a wide spectrum of clinical and health services topics to produce various types of evidence reports. These reports may be used for informing and developing coverage decisions, quality measures, educational materials and tools, clinical practice guidelines, and research agendas.[8] In this asthma report, they had to screen 3843 titles and abstracts to find the 171 relevant quality improvement projects that had been conducted. The interventions included self-monitoring, patient or caregiver education, provider education, organizational change, auditing of records and feedback, provider reminders, patient reminders, and financial or nonfinancial incentives.[9] The interventions spanned from 1 to 60 months and some were in outpatient primary care clinics, pulmonary clinics, home, school, community centers, and simultaneously applied multiple settings. The interventions were devised or provided by different people including physicians, nurses, pharmacists, health educators, and healthcare teams. In this example of a single quality improvement systematic review, the effort needed to find relevant data was immense as was the heterogeneity of the intervention types, where the interventions took place, and who provided or championed the interventions. These factors make it difficult for individual health systems to identify the types of quality improvement studies that would be especially relevant to them.[9]

EPCs also conduct research on the methodology of evidence synthesis and tagged the quality improvement literature as needed additional assessment and understanding. In 2018, a multi-EPC and AHRQ working group sought to develop a process to assess and present useful information from quality improvement projects to learning health systems.[8] There was progress in defining quality improvement studies and generating a list of 33 candidate criteria (Table 1) describing features of the interventions and the environment in which they were conducted. Unfortunately, there was little similarity to what was being extracted by experienced systematic reviewers, which was an impediment to progress, and workgroup activities were postponed. For example, for the criteria entitled "Who is receiving the intervention?" had one data extractor saying "Nurses" as compared with another that said "Advanced Practice Nurses in the Surgical Intensive Care Unit" or for the criteria "Duration of the intervention" had one data extractor stating the duration was the extra time needed to clean the patient's room (e.g., 15 min) while another stated it was the entire time the new protocol had been in place (e.g., several months). Furthermore, these criteria were not assessed by learning health system participants to identify the ones of greatest value. Without this input, it is unclear whether having 33 criteria overwhelms the reader without providing value or the relative value of each of the criteria.

The University of Connecticut (UConn) EPC sought to develop data extraction instructions for each criterion that can enhance the consistency of data extraction among different systematic reviewers and assess the difficulty health systems have with identifying and using quality improvement studies.

METHODS

Assessing the Consistency of Data Extraction

We utilized studies contained in the AHRQ report entitled "Closing The Quality Gap: A Critical Analysis of Quality Improvement Strategies: Volume 5—Asthma Care" as our data sources.[9] We chose this report because it was an extensive review of quality improvement projects that had been conducted by an EPC and the use of our criteria is envisioned to result in enhancements to subsequent EPC reports like this one should it turn out to be valuable. In the first phase of the study, we used the criteria (Table 1) that the 2018 AHRQ working group identified as potentially valuable for health systems to understand how to use and implement QI studies.[8] These criteria were selected from previous work in the quality improvement and implementation literature.[10–13]

Our methodology is displayed visually in Figure 1. At baseline, two experienced EPC systematic reviewers (AVH and YMR, named authors in this article) extracted data for all 33 criteria from two different studies (Kamps et al. 2003 and 2004, and Brown 2004).[9] There was only a rudimentary description of each criterion at this stage. When examining

Table 1 Criteria to Assess Quality Improvement Studies

1. Who is delivering the intervention (e.g., provider types)?	10. Duration of the effect of the intervention	19. Leadership commitment and involvement	28. Population needs/ burden of illness
2. Who is receiving the intervention (e.g., patient types)	11. Team composition (people delivering the intervention)	20. Clinical champion involvement	29. Geographic location
3. Provider demographics	12. External policies and incentives required	21. Physical environmental changes required	30. External factors
4. Recipient demographics	13. Required skills/training	22. Incentives	31. Organizational history of change
5. Active vs. passive components	14. Number and description of components	23. Implementation strategies	32. Fidelity
6. Discretionary vs. mandatory components	15. A priori components vs. added later/final	24. Types of intervention effects	33. Intervention adaptation
7. Duration of the intervention	16. Theoretical foundation	25. Organizational setting	
8. Frequency of intervention	17. Which interventions are independent	26. Financial setting	
9. Intensity of the intervention	18. Cost of implementation	27. Organizational receptivity/ readiness	

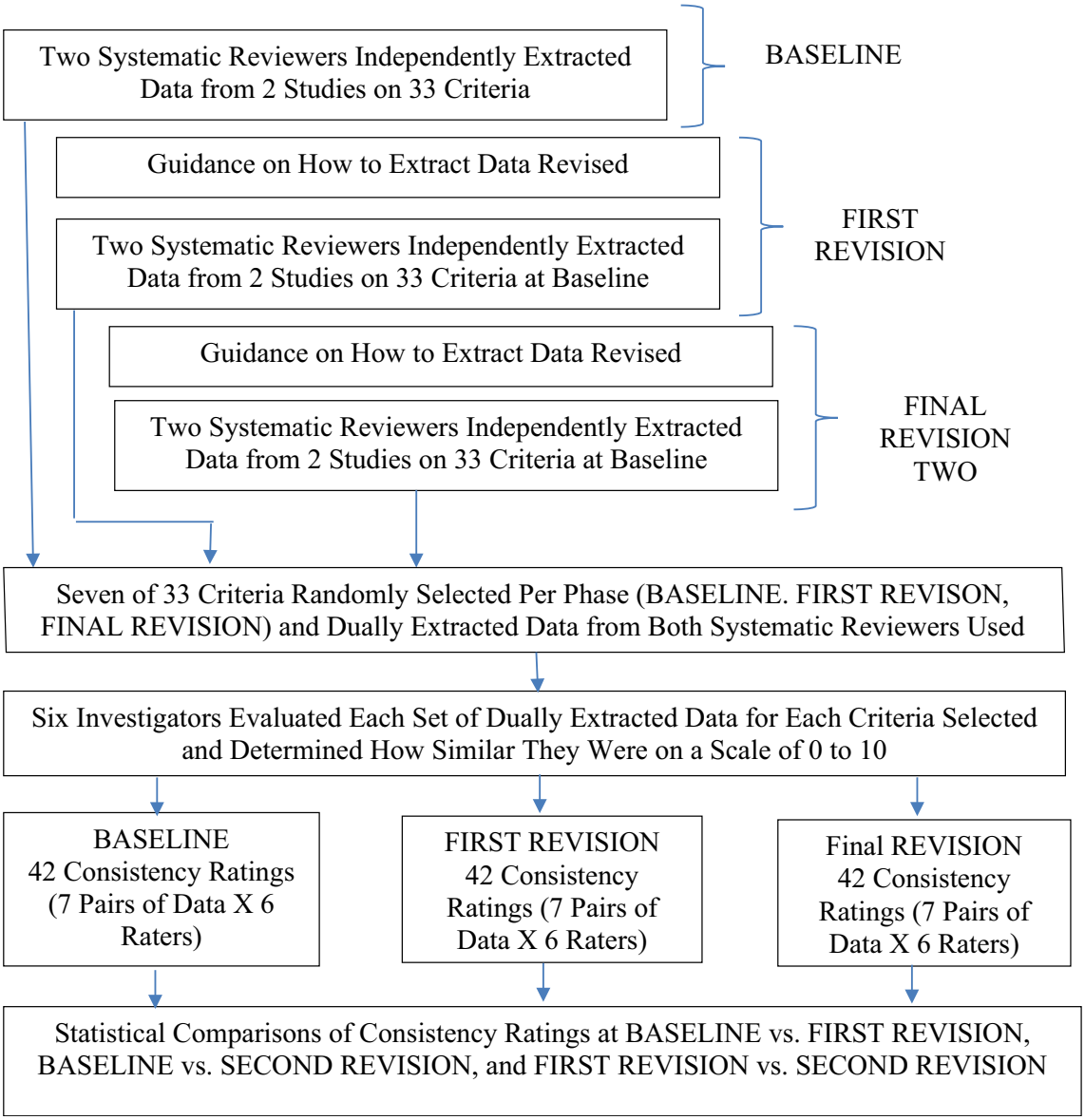

Figure 1 Schematic of methods to assess extent of data extraction consistency

the consistency of the results, three investigators mutually developed detailed instructions on how to extract study data for each criterion. Two systematic reviewers then independently extracted data for all 33 criteria from two other source studies utilizing these instructions (first revision). Reviewing inconsistencies among extractors led to research team to refine the instructions for each criterion and reduce ambiguity. Using the refined instructions (final revision), data was independently extracted by two systematic reviewers for each of the 33 criteria from two other studies. For example, criterion #7 Duration of Intervention was confusing at baseline as it was not clearly described the different between duration of intervention and length of intervention. After the first revision, we added the following extra text: "Duration of intervention should specify total hours/days of intervention. It comes from the multiplication of frequency and intensity. Length of intervention is the calendar time the intervention has been in place." During the final revision, criterion #7 was not a source of inconsistency anymore.

To assess the impact of the instructions on the consistency of data extraction, an investigator randomly selected dually extracted data from 7 of the 33 specific criteria (termed a data field) listed in the Appendix at each phase using a random permutations generator (Randomization.com – Third Generator). In all, seven dually extracted data fields were selected at baseline, seven after the first revision of the instructions were completed (first revision), and seven after the final instructions were completed (final revision). Six investigators looked at the data extracted by one of the systematic reviewers and then the other for the same field and determined the extent of similarity on a scale of 0 to 10 (where 0 represented no similarity and 10 perfect similarity of extraction). This meant there were 126 assessments of the similarity of extracted data (42 assessments for baseline, 42 assessments for first revision, and 42

assessments for the final revision) as displayed in Figure 1. This non-parametric data on extraction consistency was compared between the groups (baseline vs. first revision, baseline vs. final revision, first revision vs. final revision) using the related-samples Wilcoxon signed-rank test (SPSS). We provide mean and median extraction consistency values for baseline, first revision, and final revision and a P value of < 0.05 considered statistically significant. The final instruction set for the criteria are included in the Appendix.

Assessing the Relative Value of the Criteria

Two quality improvement clinicians from Hartford Healthcare and the University of Connecticut Health Center were first asked to rate how difficult it is to find and use quality improvement studies in health system quality improvement endeavors. They were then asked how valuable it would be if AHRQ found and collated the background information about the interventions and the health systems from published quality improvement studies. For both questions, people were asked to use a 0 to 10 rating scale where 0 denotes no difficulty or no value and 10 denotes great difficulty or great value, respectively. While this is a subjective, and therefore imperfect, rating system, it does provide some insight into something that is very difficult to quantify otherwise. Our quality improvement clinicians (identified in the Acknowledgments section) have years of experience in quality improvement and come from both a small rural academic medical center and a large, urban and suburban, community health system to maximize applicability.

RESULTS

Impact of Iterative Instruction Revision on Consistency

For our primary aim, Table 2 delineates the degree to which raters found similarities in data extracted from studies for each of the randomly selected criteria between our two independent extractors. At baseline, we found very little extraction similarity for the criteria in the source studies by our two independent systematic reviewers. After we refined the instructions about how to extract data for the criteria the first time, we dramatically improved the consistency of extraction scores between the two reviewers by 5.2-fold ($P < 0.001$). The second and final revision of the instructions again increased the rating of similarity between the two reviewers versus baseline by 5.8-fold ($P < 0.001$) but only nominally different from that achieved after the first revision ($P = 0.14$) (Table 2). The mean and median agreement scores of 6.81 and 7 represent a moderate level of agreement with the final set of instructions for extracting data for these criteria.

Table 2 Data Extraction Consistency Ratings

	Mean (standard deviation)	Median (25th–75th percentile)
Baseline ratings	1.17 (1.85)	0 (0–3)
First revision ratings	6.07 (2.76)*	6 (3–9)
Final revision ratings	6.81 (1.94)*,	7 (6–8)

**P < 0.0001 vs. baseline, P = 0.14 vs. first revision. Scores could range from 0 to 10 where 0 = no consistency of extraction between systematic reviewers and 10 = perfect consistency*

For example, at baseline for the criterion #7 "Duration of Intervention," reviewer A extracted "Duration of intervention: two face-to-face group meetings lasting ~2.5 h each. Length of time of intervention: 2 years." and reviewer B extracted "2-3 weeks, 4 months." After the first revision, both reviewers extracted "Duration of intervention: Three 1-day learning sessions. Length of intervention: 12 months" and "Duration of the intervention: 3 1-day sessions; length: 12 months," respectively. Finally, after the final revision, both reviewers extracted: "Duration of intervention: First component (Five 3-h sessions over 5 months, plus 2 additional 3-h sessions at end of 1st year); second component (3-h in first year); third component (no time specified for monthly visits up to 2y fup). Total for first and second component was 24 h; unknown for third component. Length of intervention: 2 years." and "First and fourth component: 3 hours x 7=21 hours; second component: 3 hours; third component: duration of intervention not specified. Length of intervention: 2 years." respectively.

Assessing the Value of the Criteria

The two health system representatives rated how difficult it is to find and use quality improvement studies in health system quality improvement endeavors in general. On a scale of 0 to 10, the first key informant rated it a 6 (moderately difficult) while the second reviewer rated it a 4 (moderately difficult), justified by the paucity of published quality improvement literature that they can locate and the time involved in evaluating them. They were then asked how valuable it would be if AHRQ found and collated the demographic information about the health systems and the interventions used in published quality improvement studies. On a scale or 0 to 10, the first key informant rated it a 9 (highly valuable) and the second rated it a 6 (moderately valuable).

DISCUSSION

This is the first study, which we are aware of, that explicitly looked at the consistency of data extraction from quality improvement studies. Our systematic reviewers are trained members of EPCs with ample experience but in the absence of detailed instructions, the consistency of data extraction from quality improvement studies in our study was very poor. Fortunately, we found that heterogeneous data extraction is surmountable with explicit instructions developed in an iterative fashion. We went through two refinements of the data extraction instructions for each criterion and were able to improve the consistency of extraction from baseline to the

final revision from 1.17 to 6.81 out of 10. Since the second refinement only increased the consistency of extraction slightly over the first, further refinements are unlikely to provide appreciable enhancements. We believe that the uniqueness of quality improvement studies requires this standardized approach to data extraction versus more traditional observational studies and randomized trials. Since we only looked at one disease state and a relatively few number of studies, there are limitations to our approach and perhaps some issues of applicability as well. As such, further research looking at other disease states would be beneficial.

Our criteria overlaps with that of the Quality Improvement - Minimum Quality Criteria Set (QI-MQCS), a pared down version of the criteria created by the Standards for Quality Improvement Reporting Excellence (SQUIRE) group.[10, 14] The QI-MQCS tool, developed with input from nine expert panelists, selected 14 criteria that the panelists gave a mean rating of 2.0 or greater in terms of importance (scale from 1 to 3 where 3 denoted it should be included, 2 denoted it may be included, and 1 denoted it should not be included) and two additional criteria not vetted through the expert panel. Of the 16 QI-MQCS criteria, our criterion set includes 12 of them (organizational motivation, organizational readiness, intervention, intervention rationale, organizational characteristics, implementation, timing, adherence/fidelity, penetration or reach, sustainability, comparator, and data source). Our criteria did not include the study design, health outcomes, ability for the intervention to be replicated, or inclusion of study limitations criteria. However, our quality improvement applicability table would accompany the standard information presented in EPC evidence reviews where the study design, health outcomes, and qualitative or quantitative synthesis of the results appear. In total, our criterion encompasses and expands on their criteria. This is not surprising since we both relied on the SQUIRE 2.0 criteria while we also used other sources to identify criteria that we felt were valuable.[10–13]

Testing the criterion instructions using quality improvement instructions in other diseases is a valuable next step as is more fully vetting the 33 criteria for their usefulness and completeness.

CONCLUSIONS

Our study suggests that learning health systems need support in identifying quality improvement studies and the key features of the interventions and the institutions that carried them out. In the absence of explicit and detailed instructions, there is very high heterogeneity in data extraction among independent reviewers that improves considerably with the refinement of the criteria using an explicit process. Now that consistency of extraction has been enhanced for each of our candidate criteria, a future study should determine the relative value of each criterion to learning health systems.

Acknowledgments:

We acknowledge Christina M. Polomoff, Pharm.D., BCACP, BCGP, from Integrated Care Partners, Hartford Healthcare's physician-led clinically integrated network, and Kevin Chamberlin, Pharm.D., FASCP, from UConn Health's John Dempsey Hospital for their contributions as health system key informants.

Role of the Funder: *A representative from AHRQ served as a Contracting Officer's Technical Representative and provide technical assistance during the conduct of the full evidence report and provided comments on draft versions of the full evidence report. AHRQ did not directly participate in the literature search, determination of study eligibility criteria, data analysis or interpretation, or preparation, review, or approval of the manuscript for publication.*

Corresponding Author: *C. Michael White, Pharm. D., FCP, FCCP; University of Connecticut School of Pharmacy, Storrs, CT, USA (e-mail: Charles.white@uconn.edu).*

Funding Information *This project was funded under Contract No. HHSA290-2015-00012I Task Order I from the Agency for Healthcare Research and Quality (AHRQ), U.S. Department of Health and Human Services (HHS).*

Compliance with Ethical Standards:

Conflict of Interest: *The authors declare that they do not have a conflict of interest.*

Disclaimer: *The authors of this manuscript are responsible for its content. Statements in the manuscript do not necessarily represent the official views of or imply endorsement by AHRQ or HHS.*

REFERENCES

1. The Joint Commission. Quality Accreditation Standards Information. Available at: https://www.jointcommission.org/standards_information/jcfaq.aspx. Accessed, 8/6/18.
2. Sullivan T. According to CMS 1,500 hospitals will receive bonus payments in 2020. December 4, 2019. Available at: https://www.policymed.com/2019/12/according-to-cms-1500-hospitals-will-receive-bonus-payments-in-2020.html Accessed 2/21/20.
3. Health Services Advisory Group. National Impact Assessment of the Centers for Medicare & Medicaid Services (CMS) Quality Measures Report. Feb 28, 2018. Available at: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityMeasures/Downloads/2018-Impact-Assessment-Report.pdf. Accessed 8/6/18.
4. Institute of Medicine. 2013. Best Care at Lower Cost: The Path to Continuously Learning Health Care in America. Washington, DC: The National Academies Press. Available at: https://doi.org/10.17226/13444. Accessed. 8/6/18
5. The Learning Healthcare Project. Available at: http://www.learninghealthcareproject.org/section/background/learning-healthcare-system. Accessed 8/6/18.
6. **McGinnis JM** Evidence-based medicine - engineering the Learning Healthcare System. Stud Health Technol Inform. 2010;153:145-57.
7. **White CM, Sanders-Schmidler GD, Butler M, et al.** Understanding health systems' use of and need for evidence to inform decisionmaking. Agency for Healthcare Research and Quality (US); 2017. Report 17(18)-EHC035-EF. Rockville, MD
8. AHRQ Quality Improvement Working Group. Synthesizing Evidence for Quality Improvement. Agency for Healthcare Research and Quality. August 8, 2018. Available at: https://effectivehealthcare.ahrq.gov/topics/health-systems/quality-improvement. Accessed 6/10/2019.
9. **Bravata DM, Sundaram V, Lewis R, Gienger A, Gould MK, McDonald KM, Wise PH, Holty J-EC, Hertz K, Paguntalan H, Sharp C, Kim J, Wang E, Chamberlain L, Shieh L, Owens DK**. Asthma Care. Vol 5 of: Shojania KG, McDonald KM, Wachter RM, Owens DK, editors. Closing the Quality Gap: A Critical Analysis of Quality Improvement Strategies.

Technical Review 9 (Prepared by the Stanford University-UCSF Evidence-based Practice Center under Contract No. 290-02-0017). AHRQ Publication No. 04(07)-0051-5. Rockville, MD: Agency for Healthcare Research and Quality. January 2007.

10. **Ogrinc G, Davies L, Goodman D, Batalden P, Davidoff F, Stevens D.** Standards for Quality Improvement Reporting Excellence 2.0: revised publication guidelines from a detailed consensus process. J Surg Res 2016;200(2):676–682.
11. **Pinnock H, Barwick M, Carpenter CR for the StaRI Group, et al.** Standards for Reporting Implementation Studies (StaRI): explanation and elaboration document. BMJ Open. 2017;7:e013318. https://doi.org/10.1136/bmjopen-2016-013318.
12. **Hoffmann T, Glasziou P, Boutron I, et al.** Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ 348: g1687. 2014.
13. **Guise J-M, Butler ME, Chang C, Viswanathan M, Pigott T, Tugwell P.** AHRQ series on complex intervention systematic reviews—paper 6: PRISMA-CI extension statement and checklist. J Clin Epidemiol. 2017;90:43–50.
14. **Hempel S, Shekelle PG, Liu JL, et al.** Development of the Quality Improvement Minimum Quality Criteria Set (QI-MQCS): a tool for critical appraisal of quality improvement intervention publications. BMJ Qual Saf 2015;24:796-804.