

Machine Learning Methods in Systematic Reviews: Identifying Quality Improvement Intervention Evaluations



Agency for Healthcare Research and Quality
Advancing Excellence in Health Care • www.ahrq.gov

Research White Paper

Machine Learning Methods in Systematic Reviews: Identifying Quality Improvement Intervention Evaluations

Prepared for:

Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

Contract No. 290-2007-10062-I

Prepared by:

Southern California Evidence-based Practice Center, RAND Corporation
Santa Monica, CA

Investigators:

Susanne Hempel, Ph.D.
Kanaka D. Shetty, M.D.
Paul G. Shekelle, M.D., Ph.D.
Lisa V. Rubenstein, M.D.
Marjorie S. Danz, M.D.
Breanne Johnsen, B.A.
Siddhartha R. Dalal, Ph.D.

AHRQ Publication No. 12-EHC125-EF
September 2012

This report is based on research conducted by the Southern California Evidence-based Practice Center under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. 290-2007-10062-I). The findings and conclusions in this document are those of the author(s), who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

The information in this report is intended to help health care decisionmakers—patients and clinicians, health system leaders, and policymakers, among others—make well-informed decisions and thereby improve the quality of health care services. This report is not intended to be a substitute for the application of clinical judgment. Anyone who makes decisions concerning the provision of clinical care should consider this report in the same way as any medical reference and in conjunction with all other pertinent information, i.e., in the context of available resources and circumstances presented by individual patients.

This report may be used, in whole or in part, as the basis for development of clinical practice guidelines and other quality enhancement tools, or as a basis for reimbursement and coverage policies. AHRQ or U.S. Department of Health and Human Services endorsement of such derivative products may not be stated or implied.

This document is in the public domain and may be used and reprinted without permission except those copyrighted materials that are clearly noted in the document. Further reproduction of those copyrighted materials is prohibited without the specific permission of copyright holders.

Persons using assistive technology may not be able to fully access information in this report. For assistance, contact EffectiveHealthCare@ahrq.hhs.gov.

None of the investigators have any affiliation or financial involvement that conflicts with the material presented in this report.

Suggested citation: Hempel S, Shetty KD, Shekelle PG, Rubenstein LV, Danz MS, Johnsen B, Dalal SR. Machine Learning Methods in Systematic Reviews: Identifying Quality Improvement Intervention Evaluations. Research White Paper (Prepared by the Southern California Evidence-based Practice Center under Contract No. 290-2007-10062-I). AHRQ Publication No. 12-EHC125-EF. Rockville, MD: Agency for Healthcare Research and Quality. September 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

Preface

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-based Practice Centers (EPCs), sponsors the development of evidence reports and technology assessments to assist public- and private-sector organizations in their efforts to improve the quality of health care in the United States. The reports and assessments provide organizations with comprehensive, science-based information on common, costly medical conditions and new health care technologies. The EPCs systematically review the relevant scientific literature on topics assigned to them by AHRQ and conduct additional analyses when appropriate prior to developing their reports and assessments.

To improve the scientific rigor of these evidence reports, AHRQ supports empiric research by the EPCs to help understand or improve complex methodologic issues in systematic reviews. These methods research projects are intended to contribute to the research base and be used to improve the science of systematic reviews. They are not intended to be guidance to the EPC program, although may be considered by EPCs along with other scientific research when determining EPC program methods guidance.

AHRQ expects that the EPC evidence reports and technology assessments will inform individual health plans, providers, and purchasers; as well as the health care system as a whole by providing important information to help improve health care quality. The reports undergo peer review prior to their release as a final report.

We welcome comments on this Methods Research Project. They may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, or by email to epc@ahrq.hhs.gov.

Carolyn M. Clancy, M.D.
Director
Agency for Healthcare Research and Quality

Jean Slutsky, P.A., M.S.P.H.
Director, Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Stephanie Chang, M.D., M.P.H.
Director, Evidence-based Practice Program
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Suchitra Iyer, Ph.D.
Task Order Officer
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Acknowledgements

We would like to thank Suchitra Iyer, AHRQ Project Officer, for her patience and helpful comments throughout the project; Aneesa Motala and Tanja Perry for assistance with data management and manuscript; Jeremy Miles for statistical assistance; Roberta Shanman for the literature searches; and Miriam de Kleijn for contributions to team discussions on quality improvement.

Machine Learning Methods in Systematic Reviews: Identifying Quality Improvement Intervention Evaluations

Structured Abstract

Background: Electronic searches typically yield far more citations than are relevant, and reviewers spend a substantial amount of time screening titles and abstracts to identify potential studies eligible for inclusion in a review. This is of particular relevance in complex research fields such as quality improvement. We tested a semiautomated literature screening process applied to the title and abstract screening stage of systematic reviews. A machine learning approach may allow literature reviewers to screen only a fraction of a search output and to use a predictive model to learn and then emulate the reviewers' decisions. Once learned, the model can apply the selection process to an essentially unlimited number of citations.

Method: Two independent literature reviewers screened 1,591 quasi-randomly selected citations in a training dataset used to predict decisions on the remaining citations in a MEDLINE search output of 9,395 citations. We explored different prediction algorithms and tested results against reference samples screened by experts in quality improvement. Qualitative (relevance cutoff determined in ROC curve) and quantitative predictions (probability rank order of citations) were determined.

Results: The agreement between independent literature reviewers ranged from $\kappa = 0.55$ to 0.57. Across two reference samples, the predictive performance of the machine learning approach demonstrated 90.1 percent sensitivity, 43.9 percent specificity, and 32.1 percent PPV. This translates to a reduction of 36.1 percent in citation screening if applied. The predictive performance was affected by reviewer disagreements: a subgroup analysis restricted to citations both reviewers agreed on showed a sensitivity of 98.8 percent (specificity 43.9 percent).

Conclusion: Machine learning approaches may assist in the title and abstract inclusion screening process in systematic reviews of complex, steadily expanding research fields such as quality improvement. Increased reviewer agreement appeared to be associated with improved predictive performance.

Contents

Background	1
Objectives and Key Questions	2
Methods	4
Dataset	4
Initial Procedure	4
Machine Learning Approach	4
Predictions	5
Predictive Performance Test	6
Reference Samples	6
Analysis	6
Results	8
Reviewer Characteristics	8
Predictions	9
Discussion	14
Conclusion	17
References	18

Tables

Table 1. Inter-Rater Agreement Training Sample, 400 Abstracts, Original Reviewer Set	8
Table 2. Inter-Rater Agreement Validation Sample, 400 Abstracts, Original Reviewer Set	8
Table 3. Inter-Rater Agreement Validation Sample, 400 Abstracts, New Reviewer Set	8
Table 4. Confusion Matrix Original Reviewer Pair (LR, PS)	11
Table 5. Confusion Matrix New Reviewer Pair (MD, SH)	11
Table 6. Confusion Matrix Subsample Full Agreement Between Reviewers	13

Figures

Figure 1. Prediction Performance Training Data (in-sample)	10
--	----

Appendixes

Appendix A. Analytic Framework Machine Learning in Systematic Reviews	
Appendix B. Search Strategy	
Appendix C. Inclusion Screening Form for Publications Obtained as Full Text	
Appendix D. Machine Learning Confusion Matrix, Text Terms Distinguishing Relevant and Irrelevant Citations, and Reviewer Disagreements	

Background

Outputs from electronic database searches for systematic reviews may contain tens of thousands of publications, despite progress in electronic search filter development. The reasons for this include the general increase in publication volumes, the complexity of the review questions requiring a diverse set of studies, and review areas that do not allow the restriction to high level of evidence studies such as randomized controlled trials (RCTs). These electronic searches typically yield far more citations than are relevant and selected for full text scrutiny.

Literature reviewers spend a substantial amount of time screening titles and abstracts to identify potentially relevant studies. Screening large search outputs is a lengthy process requiring considerable reviewer resources for a repetitive task that nonetheless requires conscientious and diligent work from expert reviewers. In this first step of inclusion screening, the title and abstract stage, reviewers do not make decisions about inclusion in a review. The cognitive decision reviewers have to make at this stage is whether a publication is worth screening as a full text publication because there is a chance that the publication will meet inclusion screening criteria. The decision at this stage of the review is whether to obtain full text or whether to discard. For the most part, this decision is an educated guess because the reviewers do not have the full information accessible that is available in the full text of the publication. At this review stage, reviewers are over-inclusive, and disagreements between different independent reviewers are typically not reconciled. The primary aim of this screening process is to ensure that pertinent studies are not missed. An incomplete set of eligible studies threatens the validity of the systematic review and the integrity of the review process. However, in systematic reviews additional sources are available to identify pertinent studies (e.g., reference mining of included studies, apart from the electronic database search).

In this project we assess a literature screening process that uses machine learning. Information technology approaches are being integrated in a number of research fields and the general purpose and approach vary. Altman et al. (2008) summarize opinions from leading scientists in biology; their responses highlighted the possibility of fusing literature and biological databases through text mining. Wei et al. (2008) published an automatic tool for screening human genetic association publication records exported from PubMed. Aphinyanaphongs and Aliferis (2007) used text categorization models to identify Web pages that make unproven cancer treatment claims. Information retrieval technologies are also being developed and tested for systematic review applications, primarily aiming to reduce the workload for literature reviewers by reducing the number of citations that have to be screened manually (e.g., Cohen et al., 2010; Cohen et al., 2009; Wallace, Small, et al., 2010; Wallace, Trikalinos, et al., 2010). Cohen et al. (2006) described an automated method to classify citations potentially relevant to 15 systematic drug class reviews. The model attempted to predict which articles contained high-quality, drug class-specific evidence. Results varied by topic; in some cases a 50 percent reduction of articles needing manual review was achieved. Balk et al. (2011) used a computerized screening approach to reduce the number of abstracts needed to manually screen prior to starting subsequent steps of the systematic review by 50 percent; all abstracts were eventually screened during the course of the review but no further relevant abstracts were identified.

Different approaches to differentiate citations have been explored. Wallace, Trikalinos, et al. (2010) used a Support Vector Machine (SVM) approach and reported that the number of manually screened publications can be reduced by between 40 and 50 percent for three systematic review topics. Multiple regression models and discriminant function approaches are

popular in related fields trying to refine search filters for electronic databases (e.g., Glanville et al., 2006). Other text categorization approaches use Bayesian logistic regression (Genkin et al., 2007) and latent semantic indexing (LSI; Deerwester et al., 1990).

Machine learning methods may be particularly important in facilitating reviews of complex research areas such as quality improvement. Quality improvement interventions aim to improve the way health care is delivered. Quality improvement is a research field that is characterized by heterogeneous interventions, a wide range of study designs, and study outcomes (e.g., Danz et al., 2010; O’Neil et al., 2011; Rubenstein et al., 2008), and it is more challenging to search than (for example) classic drug studies. Electronic search strategies for quality improvement are particularly challenging, producing large search outputs with limited success in retrieving pertinent studies when restricted to a manageable search yield, such as 50,000 citations per database (Hempel et al., 2011). Electronic search filter development has been successful for methodological study designs such as RCTs (see e.g., Jenkins, 2004), but quality improvement evaluations employ a variety of study designs and searches that cannot be restricted to RCTs alone (see e.g., Rubenstein et al., 2008). Quality improvement interventions can encompass a large number of approaches, and literature searches are often designed to identify well-known as well as novel approaches and not a set of prespecified interventions. In addition, outcomes are often institution and intervention-specific and can also encompass a large number of possible measures. Finally, there is no MEDLINE Subject Heading (MeSH) term that allows the parsimonious identification of quality improvement publications in the most relevant database (MEDLINE).

We applied a machine learning approach that has been developed in our institution as part of a research program to explore the facilitation of research with novel information technology. Variations of the approach have been utilized for diverse research questions and applications (see e.g., Shetty and Dalal, 2011). The particular approach is designed to assist the process of ongoing systematic reviews. A machine learning approach has also recently been tested to assist in updates of systematic reviews (Dalal et al., in press). Systematic reviews require frequent updates given that the evidence base can change quickly in a short amount of time with medical developments and the publication of new research studies. The approach indicated reduced workload associated with screening updated search results for all relevant efficacy/effectiveness and adverse event articles by more than 50 percent with minimal or no loss of relevant articles. The validation process for the review update application concentrates on the model’s predictions of the full text inclusion screening decision in comparison to literature reviewers’ decisions. However, a machine learning approach may also be able to assist the regular inclusion screening process in systematic reviews.

Objectives and Key Questions

The aim of this project was to test a machine learning approach for literature reviewers’ decisions at the title and abstract inclusion screening stage of a systematic review to identify published articles evaluating quality improvement—which citations from an electronic search output should be obtained as full text and screened further. To test the validity of the approach we compared the computer-generated prediction algorithm against reference samples screened by experts in quality improvement. The project aimed to answer the following question:

- What are the performance outcomes of a machine learning process when assisting in the title and abstract inclusion screening process of systematic reviews?

In particular, we were interested in the following questions:

- To what extent is the algorithm able to model reviewers' decisions in the training citation dataset?
- To what extent is the algorithm able to extrapolate from the training sample to a large citation dataset?
- How well will a machine learning approach work in a complex and heterogeneous research field?

Appendix A Figure 1 outlines the proposed approach and its validity testing in detail. In the first step, literature reviewers screen a sample of the search output and decide which titles and abstracts should be obtained and screened as full text and which should be discarded. In a second step, an algorithm is developed that aims to learn and construct a model based on the reviewers' decisions. Different methods to differentiate relevant and irrelevant publications and reviewer effects were explored. Once learned, the selection model can be applied to an essentially unlimited number of publications. In a third step, the algorithm is applied to the entire search output to predict which titles and abstracts will be selected. The performance in the final citation dataset may be markedly different from the training set performance because the algorithm has to extrapolate from the training set decision to a different, larger sample. Novel interventions not included in the training sample may be missed in the final citation dataset as the model will not have had the opportunity to learn about these. In a fourth step to validate the machine learning approach, the literature reviewers sample the remaining titles and abstracts (publications from the entire search output that were not part of the training set) and screen these for inclusion (that is, decide whether they should be obtained and screened as full text or discarded). In a final step, the algorithm's predictions are compared with the reviewers' decisions; for the training dataset as well as the remaining, out of sample, titles and abstracts.

Methods

Dataset

We used a PubMed and MEDLINE (Ovid) based search output that aimed to identify quality improvement and continuous quality improvement publications. The search strategy is documented in Appendix B. It combined several strategies to identify quality improvement citations by using a general search strategy to identify novel approaches as well as searching for specific, known quality improvement interventions and components. It aimed to identify approaches covered by the Cochrane Effective Practice and Organisation of Care Group registry (EPOC, 2012) and used continuous quality improvement terms to identify relevant studies. It covered a 5 year publication period (2004 to 2009) and was limited to English-language publications but was not limited to specific clinical fields, interventions, or outcomes.

The search strategy produced 9,395 hits of which only a small number will be relevant to quality improvement in general, and to evaluations of quality improvement interventions in particular. This filter problem is typical for the research field and a low base rate of potential includes has to be assumed (Hempel et al., 2011). The search yield is stored in an EndNote library containing all downloadable information available from PubMed and MEDLINE accessed through the Ovid interface.

Initial Procedure

The initial procedure for this project involved having two reviewers (Drs. Lisa Rubenstein [LR] and Paul Shekelle [PS]), both experts in quality improvement research, screen a fraction of the search output to identify potentially relevant publications, in this case evaluations of quality improvement interventions. Each reviewer's decision was recorded for the sample; the decision as this point is only whether to obtain the full text of the publication or to discard it. Discarding the citation means that the reviewers deemed it unlikely that the full text publication will meet the full text inclusion criteria for quality improvement evaluation publications shown in Appendix C. The full text inclusion criteria require a publication to report on a health care delivery organization; to include data on the effectiveness, impacts, or success of an intervention; to report on patient, provider behavior, or process of care health outcomes; and to evaluate an intervention that aimed to change how delivery of care is routinely structured.

The selected training sample represents a small percentage (17 percent) of the total search output (1,600 titles and abstracts, where abstracts were available), quasi-randomly selected (ordered by first page number, i.e. 1, 11, 111 etc.) from all titles and abstracts identified in the electronic searches. The number of titles and abstracts was chosen based on workload considerations and were presented to reviewers in batches of 400 titles and records. The sample citations were screened by one or both of the expert reviewers independently. LR screened 800 titles and abstracts, PS 1,200, both reviewers screened 400 titles in duplicate; 9 duplicate citations were subsequently found and removed, leaving 1,591 unique citations in the training set.

Machine Learning Approach

The overall machine learning approach applied to this project approach uses the information from the literature reviewers to learn and to predict the reviewer's decisions. A training set (1,591 coded abstracts) was used to empirically model screening decisions as a function of

article characteristics (i.e., article feature vectors); the resulting model is then used to predict citation screening decisions in the test set. The text analysis program had access to all information that can be downloaded from the database MEDLINE accessed through Ovid, and a PubMed search including database categorizations (MeSH terms). We considered two approaches for constructing these feature vectors: one based solely on MeSH and another based on text characteristics.

The first approach (“Augmented MeSH Approach”) had achieved good performance in testing on drug comparative effectiveness reviews (Dalal et al., in press). It primarily relies on extracting variables from a limited number of relevant MeSH headings and subheadings, as well as a few metadata of the citation (e.g., whether “quality improvement” is in the title) as binary feature vectors. We located relevant MeSH terms and subheadings by matching terms within the search strategy to terms within the MeSH database.

The second approach, which we call “Text Mining Approach,” solely uses words from the articles' titles and abstracts as feature vectors. In this approach, each title-abstract is represented by a vector of frequencies of each of the words in the title-abstract. This document corpus yielded a vocabulary of 20,000+ words; to reduce the dimensionality of these feature vectors, we removed various common verbs and nouns (e.g., “is,” “are,” “the”). Further, since there can be many different forms of words, we reduced them to their common grammatical roots (e.g., follow, follows, following, etc., becomes follow) using word stemming (Porter 1980). Finally, we discarded all the terms which occurred less than four times in the training data (the 1,591 manually reviewed titles and abstracts). Based on this we created a term document matrix consisting of 3,072 terms (the feature vectors) and 9,395 documents (all of the abstracts and titles in the test dataset).

Due to the size of the feature vectors we applied two prominent algorithms for our analysis and tested their performance: the lasso shrinkage and selection method (Friedman et al., 2010; Tibshirani, 1996), and a method for finding an optimal separating hyper plane (SVM, see Vapnik, 1995). The lasso algorithm minimizes the usual sum of squared errors, with a bound on the sum of the absolute values of the coefficients. This method shrinks coefficients of less important variables to zero, resulting in fewer variables with better predictive power. The shrinkage factor is determined by a tenfold cross-validation of the entire training data. We also tested an SVM-based approach, which distinguishes between relevant and irrelevant articles by finding the optimal separating hyperplane for distinguishing between the two clusters of feature vectors (relevant and irrelevant). As with lasso, this was developed on training data and then used to predict out-of-sample citations in the reference samples (see also Hastie et al., 2009; Vapnik, 1995).

Predictions

The predictive model was first developed to differentiate relevant and irrelevant titles and abstracts in the training set (1,591 citations). In a next step, the model was then applied to the remaining citations (7,804 citations after removal of duplicates). Two parameters were generated:

- Quantitative prediction
 - A likelihood score expressing the level of relevance for each citation in the search output based on the selected prediction model
- Qualitative prediction

- A cutoff score determining for each included citation whether it should be discarded or obtained for full text review based on the selected prediction model

The quantitative score takes the word frequency and similarities of relevant studies into account when predicting from the training set. Some citations will be more clearly irrelevant than others or are more likely to be relevant than others. The determined cutoff divides the sample qualitatively into a class of relevant, and a class of irrelevant citations.

A machine learning approach may put emphasis onto different kinds of predictive performance measures depending on its purpose and the cutoff can be chosen accordingly. Individual performance parameters are not independent of each other and there will be a trade-off between sensitivity and specificity. The selected approach may primarily aim to weed out irrelevant citations (specificity) to reduce workload, accepting that a noticeable number of relevant studies will be missed. Another application may put more emphasis on ensuring that no true positives will be missed, allowing a greater threshold of irrelevant citations (false positives) so that truly relevant citations will not be missed. The chosen cutoff aimed to reduce workload while maximizing sensitivity to the extent possible. In this instance, we deemed a sensitivity of 85 percent to be acceptable for the study purpose.

Predictive Performance Test

Reference Samples

The critical test for the machine learning approach is whether the model can extrapolate from the training sample to the remaining large number of citations. Therefore we sampled the remaining titles and abstracts and literature reviewers screened these for relevance, and then compared the reviewer decisions with the prediction model results.

The first set of literature reviewers were the same quality improvement experts that provided the training set decisions. The reviewers screened the citations in a random order, blind to the model predictions, for the quantitative as well as the qualitative score. Four hundred citations were screened in duplicate. Both literature reviewers reviewed the citations in the reference samples independently. An additional set of reviewers also screened a literature sample. Drs. Danz (MD) and Hempel (SH), both familiar with full text screening of quality improvement abstracts but not involved in the training data, also independently screened a sample of 400 citations from the remaining dataset.

Analysis

We compared the predictions with the reviewer's decision creating a confusion matrix, a 2×2 table showing correctly and incorrectly classified titles and abstracts, i.e., true and false negatives and positives. The literature reviewers form the gold standard which is compared with the algorithm's prediction according to a given threshold. The basic table based on the accumulated reviewers' decisions is shown in Appendix D Table 1. True negatives are citations correctly predicted to be discarded, false negatives are citations predicted to be discarded but one or both reviewers indicated they should be obtained, true positives are those predicted and eventually chosen by one or both reviewers to be obtained as full text, and false positives indicate a prediction to obtain the citation when in fact literature reviewers discarded the citation. The confusion matrix cells are not balanced due to the large number of irrelevant publications in each

search output. We estimated from previous searches that only about 10 to 20 percent of titles and abstracts tend to be relevant and are pursued further; all other citations are discarded.

In the main validity analysis a citation is considered relevant (decision: obtain as full text) if at least one of the two reviewers has deemed it relevant. This includes a higher number of relevant titles and abstracts than based on the individual reviewer's decisions given that individual reviewers do not perfectly overlap in their relevance decisions at the title and abstract stage. In addition, we explored the effect of individual reviewers and the reliability of the reference standard (the accumulated reviewer decisions). We calculated the rater agreement by computing the total agreement as well as the kappa statistic for this agreement between the two quality improvement experts. In addition, performance based on individual reviewer decisions was explored. Sources of disagreements between literature reviewers were identified, further explored, and reconciled. The impact of reviewer disagreements was explored through a subgroup analysis of citations with full agreement between independent reviewers.

We computed information retrieval performance measures for the sample used to train the model as well as for the remaining citations in the search output. The performance from the training data shows the potential of the algorithm in modeling reviewers' decisions. The performance data for the remaining large dataset compared with reviewer screening of selected samples, provides information whether the algorithm can extrapolate from the training data sample to a large number of citations.

To document the predictive performance, we computed common composite scores, such as the sensitivity (recall, proportion of true positives correctly detected), the specificity (proportion of true negatives correctly detected), and the positive predictive value (precision, ratio of true positives to combined true and false positives) to characterize the machine learning method's performance. Sensitivity is a central variable in this context because the primary aim of the title and abstract screening phase in systematic reviews is obtaining all pertinent studies. However, perfect sensitivity can be reached by declaring all citations relevant, which would entail obtaining almost 10,000 publications for full text review in this case. Therefore, we (and other systematic review researchers) consider the balance between true and false positives in both manual title and abstract screening and machine learning applications. Finally, we determined the proportion of workload (literature search screening of both relevant and irrelevant citations) that might have been avoided had the predictive model been used exclusively.

Results

Reviewer Characteristics

In the training sample, LR's overall acceptance rate was 9.6 percent (of 795 citations screened in total) and PS's was 12 percent (1,195 citations screened). Table 1 shows the rater agreement for a subsample of 400 citations that were screened in duplicate by the two reviewers.

Table 1. Inter-rater agreement training sample, 400 abstracts, original reviewer set

	LR Discard	LR Obtain Full Text	Total
PS Discard	357	15	372
PS Obtain Full Text	8	19	27
Total	365	34	399

The independent reviewers agreed on 376/399 citations (94 percent). The inter-rater agreement (kappa) was 0.57. The overall acceptance rate in the training sample consisting of 1,591 unique citations was 12.6 percent. This is based on the cumulative reviewer decision; for 200 citations at least one reviewer indicated that the full text publication should be obtained.

In the primary validation sample of 400 citations, LR and PS served again as literature reviewers. In this reference sample, LR's acceptance rate was 19.3 percent and PS's was 11.8 percent. The independent reviewers agreed on 376/399 citations (94 percent). The inter-rater agreement (kappa) was 0.57. Table 2 shows the agreement between reviewers.

Table 2. Inter-rater agreement validation sample, 400 abstracts, original reviewer set

	LR Discard	LR Obtain Full Text	Total
PS Discard	315	38	353
PS Obtain Full Text	8	39	47
Total	323	77	400

The overall acceptance rate (one or both reviewers indicated that the citation should be obtained as full text rather than discarded) in this sample was 21.3 percent. This was a statistically significantly higher acceptance rate than in the training sample ($p < 0.001$).

To explore whether the reviewer behavior and prediction performance was unique to the training sample reviewer pair we added a second pair of reviewers (MD, SH) screening 400 different citations from the remaining dataset. In this second validation sample, MD's acceptance rate was 17.5 percent and SH's was 18 percent. Table 3 shows the agreement between reviewers for the 400 citations.

Table 3. Inter-rater agreement validation sample, 400 abstracts, new reviewer set

	MD Discard	MD Obtain Full Text	Total
SH Discard	303	25	328
SH Obtain Full Text	27	45	72
Total	330	70	400

Both independent reviewers agreed on 348/400 citations (87 percent). The inter-rater agreement (kappa) was 0.55. The overall acceptance rate (one or both reviewers want to see the full text publication) in this reference sample was 24.3 percent.

Predictions

The machine learning approach used the revFiewer decisions in the training sample to learn and to predict decisions. In a first step, the algorithm identifies terms that differentiate relevant from irrelevant citations. The process determines which terms are over- and which terms are under-represented in relevant citations (relevant citations are those that were selected by reviewers for further full text inclusion screening). The program has access to all fields and content of citations downloaded from electronic databases including MeSH tags.

Included citations used the terms “improving,” “patients,” “interventions,” “implementing,” and “staff” more often than citations that were considered irrelevant by the reviewers. The selection of terms may reflect that reviewers aimed to identify evaluations of interventions, i.e., studies reporting empirical data on (typically) patient outcomes. The terms “adequate,” “public,” “search,” “evolving,” and “consensus” were underrepresented in relevant citations. This selection appears to correlate with the exclusion of discussion papers and opinion pieces as reflected in the inclusion criteria. A list of the most statistically significant over- and underrepresented terms are shown in Appendix D Table 2.

To what extent is the algorithm able to model reviewers’ decisions in the training citation dataset?

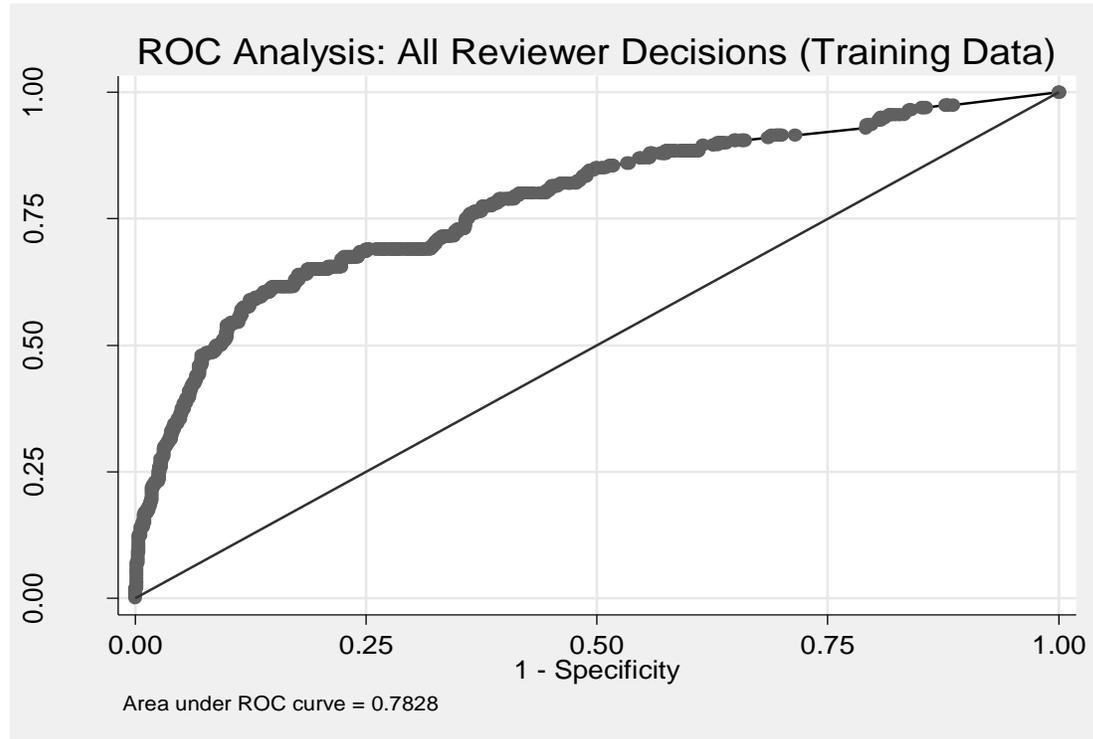
We tested a number of statistical prediction algorithms and a number of strategies to identify the best model, i.e., maximizing specificity while maintaining sensitivity at an acceptable level. For the training data modeling purposes, a citation was declared relevant if at least one of the reviewers indicated that it should be obtained as full text.

The Augmented MeSH Approach did not work well in our application even on training data, since the MeSH feature vectors did not support the differentiation of relevant and irrelevant citations in the chosen quality improvement dataset. The approach's poor performance may have stemmed from the lack of designated MeSH terms for quality improvement publications. Furthermore, the remaining MeSH terms in the search output were not specific to quality improvement and were not useful to differentiate the citation pool.

However, both the lasso approach and SVM algorithms worked relatively well on the training data using the Text Mining Approach. SVM performed much better on the training data, while did rather poorly on test data. This may be due to the fact that SVM on the training data had 1,276 non-zero coefficients, while lasso had only 78 non-zero coefficients. Thus, SVM is over-fitting on the training data, which may have negatively affected its predictions. Reported results are therefore based on the lasso approach.

The final prediction model based on lasso with the threshold chosen to reduce work load (correctly detecting truly irrelevant citations) while maintaining a sensitivity of at least 85 percent (maintaining a high level of correctly identifying relevant citations – citations that reviewers chose to be obtained as full text). Figure 1 shows the receiver operating characteristic (ROC) curve for the training sample (1,591 citations screened by at least one reviewer).

Figure 1. Prediction performance training data (in-sample)



The final model using the full dataset was able to predict the reviewers' decisions with sufficient accuracy. However, when dividing the in-sample (training) data into subsets and cross-validating predictions across subsets, the mean achieved sensitivity was 83.5 percent with a corresponding specificity of 51.5 percent (PPV 19.9 percent).

To what extent is the algorithm able to extrapolate from the training sample to a large citation dataset?

The prediction model was then applied to extrapolate predictions for the remaining 7,804 titles and abstracts in the dataset. The model determined for each title and abstract whether the citation was likely to be discarded by a literature reviewer or selected to be obtained as full text. To validate the use of machine learning in systematic reviews this “out-of-sample” set was tested against reviewer decisions.

The following Table 4 shows the results for the original set of reviewers, LR and PS who also provided literature reviewer decisions for the training set. The reference set data are based on the combined decision across reviewers—at least one reviewer selected the citation as relevant (to be obtained as full text) so a true discard would have been rejected by both independent reviewers whereas a true obtain could have been selected by one or both reviewers. The prediction model was developed in the training, not this validation sample and reviewers were blind to the model predictions.

Table 4. Confusion matrix original reviewer pair (LR, PS)

	Predicted Discard	Predicted Obtain Full Text	Total
True Discard	126	189	315
True Obtain Full Text	8	77	85
Total	134	266	400

The sensitivity and corresponding specificity were 90.6 percent and 40 percent with a PPV of 28.9 percent. The model missed 8/85 relevant citations. However, the model predicted that many more citations would be classified as “obtain” by the reviewers (266 predicted vs. 85 actually selected citations). The value in terms of work saved was 33.5 percent—a third of the search output could be discarded while maintaining a sensitivity of 90.6 percent.

A further application of the approach is to use the quantitative score to rank order the citations by relevance, for example to enable citations with a higher probability to be processed first. The algorithm determines a quantitative probability score for each of the identified citations. Ordering the citations in the primary reference sample (LR, PS) based on the probability score, 26/85 relevant citations (31 percent) would be found in the first 50 citations while 48/50 of the last citations are irrelevant (2 relevant citation would be missed). Screening the first 100 rank ordered citations will identify over half the relevant citations (45/85, 53 percent) while 97/100 of the last citations are irrelevant (3 citations missed). The quantitative score does not provide a perfect map of the citation sample but allows us to divide the dataset into parts with better odds of finding relevant citations.

We also tested how well the model predicted selections of independent reviewers not involved in the selection of the training data set to investigate how reviewer-specific the predictions are. Reviewers MD and SH screened a different random sample. The performance in this reference sample was very similar to the original reviewer set. Table 5 shows the confusion matrix.

Table 5. Confusion matrix new reviewer pair (MD, SH)

	Predicted Discard	Predicted Obtain Full Text	Total
True Discard	145	158	303
True Obtain full text	10	87	97
Total	155	245	400

The resulting sensitivity was 89.7 percent with corresponding specificity of 47.9 percent and a PPV of 35.5 percent. The performance was only marginally less sensitive than for the reviewer pair that formed the basis of the training set. The results correspond to 38.7 percent of work saved.

The sensitivity and specificity for individual reviewers did not indicate that performance was notably better or worse for individual reviewers in the reference samples. Sensitivity ranged from 90.0 percent (corresponding specificity 44.8 percent, MD) to 97.9 percent (corresponding specificity 37.7 percent, PS).

The quantitative score results were replicated in this second reference sample (MD, SH): 31/97 relevant citations (32 percent) were included in 50 citations with the highest probability score while 48/50 of the lowest ranking citations are irrelevant (2 relevant citation would be missed) when ordering the citations by the quantitative prediction score. Screening the first 100 rank ordered citations will identify over half the relevant citations (54/97, 56 percent) while 95/100 of the last citations are irrelevant (5 citations missed).

Across both reference samples, the performance was 90.1 percent sensitivity, 43.9 percent specificity, 32.1 percent PPV and 36.1 percent work saved.

How well will a machine learning approach work in a complex and heterogeneous research field?

The tested machine learning approach aimed to predict decisions based on titles and abstracts, i.e., the first step in a systematic review after citations have been identified in electronic searches. Title and abstract screening is ultimately less reliable than full text decisions which is based on explicit inclusion and exclusion criteria typically reviewed in duplicate with conflict resolution. The applied prediction model for the title and abstract screening process uses the accumulated information across both independent reviewers—the final decision was “obtain” if at least one reviewer identified the citation as relevant. There were several disagreements amongst reviewers (training set: 23/399 citations; reference set LR, PS: 46/400; reference set MD, SH: 52/400). At the title and abstract screening stage, reviewers have to process a large amount of information in a short amount of time without defined inclusion criteria and based on very limited information (citation only). Hence, the reviewer’s decisions may be an imperfect gold standard.

We explored reviewer disagreements in a set of citations screened in duplicate (LR, PS; reference sample). Appendix D Table 3 shows the full citation plus the abstract of the 46 citations two reviewers disagreed on, meaning one reviewer indicated to obtain the citation as full text while the other decided to discard it. The disagreements were discussed in our quality improvement research team. The discussions showed that the source of disagreement was typically not a simple reviewer error (e.g., one of the two reviewers overlooked a piece of relevant information). An example is PMID 17530519, a publication on asthma treatment protocols in the emergency department which reported an electronic search indicating that the article is a literature review, not an evaluation of an individual study. The disagreement was sometimes based on the lack of information and both reviewers interpreting the limited information differently. An example is the decision whether PMID 17851382, a publication on caring for patients dying at home from heart failure, is likely to report on patient (or caregiver), provider behavior, or process of care health outcomes that can be linked to a quality improvement intervention in a health care delivery organization. However, in several cases reviewers conceptually disagreed whether the publication would meet inclusion criteria if obtained as full text based on the presented information. PMID 18410318, a study evaluating a program to enhance after-hour telephone communication between nurses and physicians, was eventually classified to be discarded but intense discussions evolved around the question whether the reported outcome “the percentage of calls that resulted in immediate evaluation by a physician” should be classified as a process of care health outcome. All reasons for discarding citations rather than obtaining these as full text are documented in Appendix D Table 3. The most frequent reason to discard publications after team discussions was the reported outcome, the lack of patient, provider behavior, or process of care health outcomes.

In order to explore the reliability of the reference standard and whether its reliability influences the validity in terms of test performance, we performed a subgroup analysis for those citations where both independent reviewers agreed and independently came to the same decision. Table 6 shows the confusion matrix for a subsample of citations with full agreement between reviewers across both reference samples (LR, PS and MD, SH).

Table 6. Confusion matrix subsample full agreement between reviewers

	Predicted Discard	Predicted Obtain Full Text	Total
True Discard	271	347	618
True Obtain Full Text	1	83	83
Total	272	430	702

Note: Based on both reference samples (LR, PS; MD, SH)

The sensitivity increased to 98.8 percent when restricting to citations with full agreement amongst reviewers. The specificity remained at 43.9 percent and the PPV was 19.3 percent. The workload saving estimate was similar to the main analysis (38.7 percent).

Discussion

The project aimed to answer the following question: What are the performance outcomes of a machine learning process when assisting in the title and abstract inclusion screening process of a systematic review on quality improvement? We found that the predictive performance was 90.1 percent sensitivity, 43.9 percent specificity, and 32.1 percent PPV. The corresponding work saved was 36.1 percent if applied. The performance might be associated with reviewer agreement: a subgroup analysis restricted to titles and abstracts that both reviewers agreed on showed a sensitivity of 98.8 percent (specificity 43.9 percent).

In particular, we set out to answer the following questions:

- To what extent is the algorithm able to model reviewers' decisions in the training citation dataset?
- To what extent is the algorithm able to extrapolate from the training sample to a large citation dataset?
- How well will a machine learning approach work in a complex and heterogeneous research field?

The questions arise from the nature of the data (titles and abstract inclusion screening decisions), the process (reviewers screening an arbitrary fraction of the search output), and the nature of the research field (quality improvement).

First, in this project we attempted to predict literature reviewers' decisions at the title and abstract stage. Reviewers did not yet have the chance to see the full text publication but need to decide whether the citation has potential to meet full text inclusion criteria. Modeling reviewers' decisions at the title and abstract stage is distinctly different from modeling full text screening decisions. The achieved results show less promise than achieved when predicting full text decisions (see Dalal et al., in press). The gold standard and data used to train the prediction model are less reliable than full-text decisions which are based on decisions reconciled between two reviewers or a review team and screened against documented, explicit inclusion and exclusion criteria. Citations are selected based on their potential, which involves educated guesses about the nature of the full text publication. Individual reviewers may not necessarily agree and come to different decisions about the citation. In the presented work, disagreements between reviewers were not primarily based on errors by individual reviewers. Disagreement evolved from different interpretations of the limited information (title and abstract only) and in addition, conceptual disagreements between reviewers about the reported information were apparent and remained to some extent even after extensive team discussions. This suggests lack of reliability of the gold standard which may well influence the predictive validity of the approach; according to classical test theory, reliability is a necessary condition for validity. As shown, we explored the impact of the reliability of the gold standard in a subgroup analysis by restricting to citations with reviewer agreement and found that the sensitivity of the prediction model increases considerably. While some amount of disagreement cannot be avoided and is likely to be rooted in the nature of the data (decisions have to be made based on limited information), for the purpose of training a prediction model, reviewers may have to spend time on reconciling conceptual disagreements to increase the reliability and validity of the model. We have shown elsewhere that implicit agreement on quality improvement cannot be assumed (Danz et al., 2010). Performance results may improve in a less heterogeneous research field, for example where quality improvement interventions evaluations are limited to a set of specific interventions, study designs, or outcomes.

In addition to modeling decisions in the training set, the main task of the machine learning approach is to extrapolate to the remaining search output. Reviewers screen only a fraction of the search output to train a prediction model. The results indicate reasonable sensitivity but sensitivity coefficients and estimated workload savings were moderate; machine learning approaches have been shown to achieve a 50 percent workload reduction including our own work (e.g., Cohen et al., 2006; Dalal et al., in press; Wallace, Trikalinos, et al., 2010). However, predictive performance was not worse than estimated in the training dataset when applying cross-validation tests by artificially dividing the initial sample into training and test tests. Future machine learning applications may be able to incorporate a flexible rather than a static approach to determine a prediction model. Wallace, Small, et al. (2010) report the use of an interactive training approach where reviewers train the model interactively by providing decisions on titles and abstracts that have been selected as being most informative rather than using a random sample of titles and abstracts. Furthermore, we used a fixed number of titles and abstracts for the training dataset based on reviewer burden. A flexible approach may allow frequent feedback regarding the success of the prediction model, information whether more training data are needed, or predictions regarding the maximum achievable model success. Thus, reviewers would know early on whether to proceed with the training process: some topics tend to be more suitable for machine learning approaches than others (see also Cohen et al., 2006).

We also explored the extent of specific reviewer calibration. Two specific reviewers served as trainers in the training set. The prediction model aimed to learn and predict the decisions of two specific reviewers and the model may be distinctly calibrated towards these two experts. However, we found that the predictive performance was not noticeably better for reviewers that served as trainers; results were very similar for a second set of expert reviewers. There was however evidence that reviewer behavior had changed between the training and the test stage. Most noticeably we saw that the acceptance rate of citations (the proportion of citations deemed relevant) was higher in the reference samples than in the original training sample. By the time reference samples were drawn, reviewers may have had a refined understanding of the kind of studies that should be included in the review.

We chose the example of quality improvement and aimed to identify intervention evaluations that aimed to improve the way health care is delivered. We have shown in prior publications that quality improvement is difficult to define; the field is characterized by heterogeneous interventions, a wide range of study designs, and study outcomes (Danz et al., 2010; O'Neil et al., 2011; Rubenstein et al., 2008). Even when features of quality improvement are agreed, the operationalization of features can remain a source of disagreements (see e.g., Danz et al., 2010). We explored a set of reviewer disagreements to understand whether the training process when developing the prediction model can be improved. Extensive team discussion showed that disagreements are often based on conceptual issues, for example about the nature of eligible outcomes, rather than simple reviewer errors, for example overlooking a detail in the title or abstract of the citation. The subgroup analysis for citations with full agreements showed that predictive validity can increase, suggesting that in these kind of machine learning approaches, false negatives (missing pertinent studies) as well as false positives (selecting citations that are not eligible for inclusion in the review) should be explored and reconciled where possible to improve the training process.

Unlike in previous work (Dalal et al., in press), in the current machine learning application MeSH terms were not used to differentiate relevant from irrelevant citations; the selected prediction model was based on text words. The dataset was the result of a carefully developed

electronic search strategy already designed to maximize sensitivity and specificity where possible. Were MeSH terms easily available, these would have been considered in the original search strategy and integrated as a further filter. However, there is no designated MeSH term for quality improvement publications, which generally hinders accessibility of quality improvement evidence (Hempel et al., 2011). The reliance on MeSH terms can be a problem for newer citations and non-PubMed citations because such citations will need to be evaluated by systematic review researchers regardless of when such controlled language is added to the citation; therefore, a text-only strategy can be advantageous in machine learning assisted reviews.

The evaluation of the predictive model focused primarily on the determined qualitative cutoff (the citation is either predicted to be relevant or to be irrelevant) and resulting confusion matrix results and sensitivity and specificity. However, the developed approach can also determine a score for each citation quantifying the likelihood of relevance. This feature enables a better management of citations and the likelihood score can be used to prioritize citations or to determine the order in which citations are screened. Similarly, Balk et al. (2011) used a computerized screening program to prioritize citations identified in a search on the diagnosis and treatment of sleep apnea so that most relevant citations were processed first in the course of the review.

Our study tested a general approach of integrating machine learning in a systematic review process. The developed prediction model may be used as an additional, refined search algorithm for quality improvement intervention studies complementing electronic searches. However, the general tested machine learning approach was designed to assist an ongoing systematic review process, acknowledging that future reviews will synthesize other literature and will have other review-specific requirements which an equally tailored machine learning approach will need to address. How machine learning approaches are applied in practice has to be decided for each review. Applications may follow a static or interactive training model and different algorithms may need to be tested to find the most promising approach for each review topic. The machine learning decisions may be used to prioritize the screening order, may replace one of the two literature reviewers in the inclusion screening process, or may replace both reviewers after training if sufficient confidence in the automated process can be established.

Conclusion

Machine learning approaches applied in systematic reviews of complex research fields such as quality improvement may assist in the title and abstract inclusion screening process. Machine learning approaches are of particular interest considering steadily increasing search outputs and accessibility of the existing evidence is a particular challenge of the research field quality improvement. Increased reviewer agreement appeared to be associated with improved predictive performance.

References

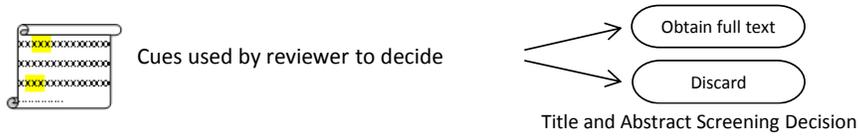
- Altman RB, Bergman CM, Blake J, et al. Text mining for biology—the way forward: opinions from leading scientists. *Genome Biology*. 2008;9(Suppl. 2):S7.
- Aphinyanaphongs Y, Aliferis C. Text categorization models for identifying unproven cancer treatments on the web. *Stud Health Technol Inform*. 2007;129:968-72
- Balk EM, Moorthy D, Obadan NO, et al. Diagnosis and Treatment of Obstructive Sleep Apnea in Adults. Comparative Effectiveness Review No. 32. (Prepared by Tufts Evidence-based Practice Center under Contract No. 290-2007-100551). AHRQ Publication No. 11-EHC052-EF. Rockville, MD: Agency for Healthcare Research and Quality. 2011. www.effectivehealthcare.ahrq.gov/reports/final.cfm
- Cohen AM, Ambert K, McDonagh M. Cross-topic learning for work prioritization in systematic review creation and update. *J Am Med Inform Assoc* 2009;16(5):690-704.
- Cohen AM, Ambert K, McDonagh M. A Prospective Evaluation of an Automated Classification System to Support Evidence-based Medicine and Systematic Review. AMIA Annual Symposium proceedings. 2010;121-5.
- 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(2):206-19.
- Dalal SR, Shekelle PG, Hempel S, et al. A Pilot Study Using Machine Learning and Domain Knowledge To Facilitate Comparative Effectiveness Review Updating. Methods Research Report (Prepared by the Southern California Evidence-based Practice Center under Contract No. 290-2007-10062-I). AHRQ Publication No. 12-EHC069-EF. Rockville, MD: Agency for Healthcare Research and Quality. September 2012.
- Danz M, Rubenstein L, Hempel S, et al. Identifying quality improvement intervention evaluations –is consensus achievable? *Qual Saf Health Care*. 2010;19(4):279-83.
- Deerwester S, Dumais S, Furnas G, et al. Indexing by Latent Semantic Analysis. *J Am Soc Inf Sci*. 1990;41(6):391-407.
- EPOC: Cochrane Effective Practice and Organisation of Care Group. Welcome to the EPOC Ottawa Website. <http://epoc.cochrane.org>. Accessed April 2012.
- Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33(1):1-22.
- Genkin A, Lewis DD, Madigan D. Large-scale Bayesian logistic regression for text categorization. *Techmetric* 2007;49(3):201-304.
- Glanville JM, Lefebvre C, Miles JN, et al. How to identify randomized controlled trials in MEDLINE: ten years on. *J Med Libr Assoc*. 2006;94:130-6.
- Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. New York: Springer Verlag; 2009.
- Hempel S, Rubenstein L, Shanman R, et al. Identifying quality improvement intervention publications. *Implement Sci*. 2011;6:85.
- Jenkins M. Evaluation of methodological search filters—a review. *Health Info Libr J*. 2004;21:148-63.
- O’Neil S, Rubenstein L, Danz M, et al. Identifying continuous quality improvement publications: what makes an improvement intervention ‘CQI’? *BMJ Qual Saf*. 2011 Dec;20(12):1011-9. Epub 2011 Jul 4.
- Porter M. An algorithm for suffix stripping. *Program* 1980;14(3):130-7.
- Rubenstein LV, Hempel S, Farmer M, et al. Finding order in heterogeneity: Types of quality improvement intervention publications. *Qual Saf Health Care*. 2008;17(6):403-8.
- Shetty KD, Dalal SR. Using information mining of the medical literature to improve drug safety. *J Am Med Inform Assoc*. 2011;18:668-74.
- Tibshirani R. Regression shrinkage and selection by Lasso. *J R Stat Soc Ser C Appl Stat*. 1996;Series B 58:267-88.

- Vapnik VN. The Nature of statistical learning theory. New York: Springer-Verlag; 1995.
- Wallace BC, Small K, Brodley C. et al. Active learning for biomedical citation screening. KDD '2010. Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining. Washington, DC: ACM; 2010.
- Wallace BC, Trikalinos TA, Lau J, et al. Semi-automated screening of biomedical citations for systematic reviews. BMC Bioinformatics. 2010;11:55.
- Wei Y, Clyne M, Dolan SM, et al. GAPscreener: an automatic tool for screening human genetic association literature in PubMed using the support vector machine technique. BMC Bioinformatics. 2008; 9:205.

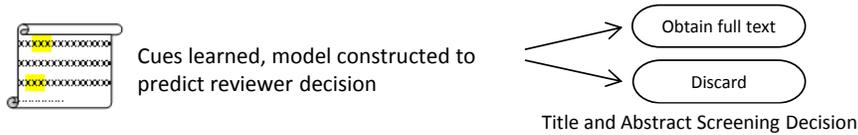
Appendix A. Analytic Framework Machine Learning in Systematic Reviews

Appendix Figure 1: Analytic Framework Machine Learning in Systematic Reviews

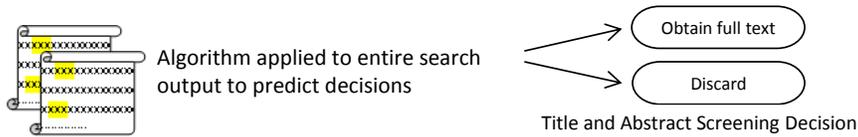
Stage 1: Sample of search output selected for reviewer screening



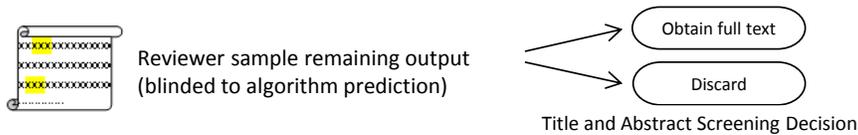
Stage 2: Algorithm learns and predicts reviewer decisions



Stage 3: Total search output extrapolation



Stage 4: Reviewer screening of reference sample



Stage 5: Comparison of decisions

Algorithm prediction			Reviewer decision	
ID	Prediction		ID	Decision
Training set			Training set	
1	Obtain	↔	1	Obtain
2	Discard		2	Discard
3	Discard		3	Discard
...
Remaining search output			Remaining search output	
1	Obtain	↔	1	Obtain
2	Discard		2	Discard
3	Discard		3	Discard
...

Appendix B. Search Strategy

DATABASE SEARCHED & TIME PERIOD COVERED: MEDLINE ON OVID – 2004-2009

LIMITERS: ENGLISH ONLY

SEARCH #1 + MESH TERMS FOR “QUALITY”:

(quality and (improv* or enhance*) and (interven* or initiative* or strateg* or program* or collaborative* or process*)).mp. [mp=title, original title, abstract, name of substance word, subject heading word]

AND

(quality of health care or quality assurance, health care or quality indicators, health care or health plan implementation).sh. or organizational change.mp. or organizational structure.mp. or organizational innovation/ [mp=title, original title, abstract, name of substance word, subject heading word]

NUMBER OF RESULTS: 4628

SEARCH #2:

149. (quality adj5 (improv* or enhance*) adj5 (interven* or initiative* or strategy or program or programme or collaborative* or process)).mp. [mp=title, original title, abstract, name of substance word, subject heading word]

150. limit 149 to (english language and yr="2004 -Current")

NUMBER OF RESULTS: 1693

EPOC SEARCH FILTER:

2. exp *education, continuing/
3. (education\$ adj2 (program\$ or intervention? or meeting? or session? or strateg\$ or workshop? or visit?)).tw.
4. (behavio?r\$ adj2 intervention?).tw. or pamphlets/ or (leaflet? or booklet? or poster? or pamphlet?).tw.
5. (((written or printed or oral) adj information) or (information\$ adj2 campaign)).tw.
6. (education\$ adj1 (method? or material?)).tw.
7. advance directives/ or outreach.tw. or ((opinion or education\$ or influential) adj1 leader?).tw.
8. (facilitator? or academic detailing or consensus conference?).tw. or *guideline adherence/ or practice guideline?.tw.
9. (guideline? adj2 (introduc\$ or issu\$ or impact or effect? or disseminat\$ or distribut\$)).tw.
10. ((effect? or impact or evaluat\$ or introduc\$ or compar\$) adj2 training program\$).tw.
11. *reminder systems/ or reminder?.tw. or (recall adj2 system\$).tw. or (prompter? or prompting).tw. or algorithm?.tw.
12. *feedback/ or feedback.tw. or chart review\$.tw. or ((effect? or impact or records or chart?) adj2 audit).tw.
13. (compliance or marketing).tw.
14. 6 or 11 or 3 or 7 or 9 or 12 or 2 or 8 or 4 or 13 or 10 or 5
16. exp *reimbursement mechanisms/
17. fee for service.tw. or *capitation fee/ or *"deductibles and coinsurance"/ or cost shar\$.tw.
18. (copayment? or co payment? or (prepay\$ or prepaid or prospective payment?)).tw.
19. *hospital charges/ or formular\$.tw. or fundhold\$.tw. or *medicaid/ or *medicare/ or blue cross.tw.

20. 18 or 19 or 16 or 17
21. *nurse clinicians/ or *nurse midwives/ or *nurse practitioners/ or (nurse adj (rehabilitator? or clinician? or practitioner? or midwi\$)).tw.
22. *pharmacists/ or clinical pharmacist?.tw. or paramedic?.tw. or *patient care team/ or exp *patient care planning/ or (team? adj2 (care or treatment or assessment or consultation)).tw.
23. (integrat\$ adj2 (care or service?)).tw.
24. ((care adj2 (coordinat\$ or program\$ or continuity)) or (case adj1 management)).tw.
25. exp *ambulatory care facilities/ or *ambulatory care/
26. exp *ambulatory care facilities/
27. 25 or 22 or 21 or 24 or 26 or 23 or 20
28. *home care services/ or *hospices/ or *nursing homes/ or *office visits/ or *house calls/
29. *day care/ or *aftercare/ or *community health nursing/ or (chang\$ adj1 location?).tw.
30. (domiciliary or (home adj1 treat\$) or day surgery).tw. or *medical records/ or *medical records systems, computerized/ or (information adj2 (management or system?)).tw.
31. *peer review/ or *utilization review/ or exp *health services misuse/
32. 28 or 30 or 31 or 29
33. *physician's practice patterns/ or quality assurance.tw. or *process assessment/ [health care]
34. *program evaluation/ or *length of stay/ or (early adj1 discharg\$).tw. or discharge planning.tw.
35. (((offset or triage).tw. or exp *"Referral and Consultation"/) and "consultation"/) or *drug therapy,computer assisted/
36. near patient testing.tw. or *medical history taking/ or *telephone/ or (physician patient adj (interaction? or relationship?)).tw.
37. *health maintenance organizations/ or managed care.tw. or (hospital? adj1 merg\$).tw.
38. 35 or 33 or 34 or 36 or 37
39. ((standard or usual or routine or regular or traditional or conventional or pattern) adj2 care).tw.
40. (program\$ adj2 (reduc\$ or increas\$ or decreas\$ or chang\$ or improv\$ or modify\$ or monitor\$ or care)).tw.
41. (program\$ adj1 (health or care or intervention?)).tw.
42. ((effect? or impact or evaluat\$ or introduc\$ or compar\$) adj2 treatment program\$).tw.
43. ((effect? or impact or evaluat\$ or introduc\$ or compar\$) adj2 care program\$).tw.
44. ((effect? or impact or evaluat\$ or introduc\$ or compar\$) adj2 screening program\$).tw.
45. ((effect? or impact or evaluat\$ or introduc\$ or compara\$) adj2 prevent\$ program\$).tw.
46. (computer\$ adj2 (dosage or dosing or diagnosis or therapy or decision?)).tw.
47. ((introduc\$ or impact or effect? or implement\$ or computer\$) adj2 protocol?).tw.
48. ((effect? or impact or introduc\$) adj2 (legislation or regulations or policy)).tw.
49. 39 or 40 or 41 or 48 or 47 or 42 or 46 or 45 or 43 or 44
50. (randomized controlled trial or controlled clinical trial).pt. or intervention studies/ or experiment\$.tw.
51. ((time adj series) or (pre test or pretest or posttest or post test)).tw.
52. random allocation/ or impact.tw. or intervention?.tw. or chang\$.tw. or evaluation studies/ or evaluat\$.tw. or effect?.tw.
53. comparative study.pt.

54. 52 or 50 or 53 or 51
55. animal/
56. human/
57. 55 NOT (55 and 56)
58. 38 or 27 or 32 or 49 or 20 or 14
59. 54 not 57
60. 59 and 58
61. limit 60 to review
62. 60 not 61
63. meta-analysis.pt.
64. 62 not 63
65. (quality of health care or quality assurance, health care or quality indicators, health care or health plan implementation).sh.
66. health plan implementation.sh.
67. 66 or 65
68. 67 and 64
69. 68
70. limit 69 to yr="2004 -Current"
83. (quality and (improv* or enhance*) and (interven* or initiative* or strateg* or program* or collaborative* or process*)).mp.
84. limit 83 to (english language and yr="2004 - 2009")
85. limit 68 to (english language and yr="2004 -Current")
104. from 85 keep 3801-3943
105. 38 or 27 or 32 or 49 or 20 or 14
106. 105 and 54
107. 106 not 57
108. limit 107 to review
109. 107 not 108
110. meta-analysis.pt.
111. 109 not 110
AND
112. (quality of health care or quality assurance, health care or quality indicators, health care or health plan implementation).sh. or organizational change.mp. or organizational structure.mp. or organizational innovation/ [mp=title, original title, abstract, name of substance word, subject heading word]
113. 112 and 111
114. limit 113 to (english language and yr="2004 -Current")
NUMBER OF RESULTS: 4587
CQI SEARCH
161. pdsa.ti,ab. or plan-do-study-act.mp. or plan do study act.mp. [mp=title, original title, abstract, name of substance word, subject heading word]
162. limit 161 to (english language and yr="2004 -Current")
163. pdca.ti,ab. or plan-do-check-act.mp. or plan do check act.mp. [mp=title, original title, abstract, name of substance word, subject heading word]

164. define-measure-analyze-improve-control.mp. or dmaic.ti,ab. or dmadv.ti,ab. or define-measure-analyze-design-verify.mp. [mp=title, original title, abstract, name of substance word, subject heading word]

165. ((iterative adj cycle) or (rapid adj cycle) or (small adj test adj2 change)).mp. or deming.ti,ab. or taguchi.ti,ab. or kansei.ti,ab. [mp=title, original title, abstract, name of substance word, subject heading word]

166. (six-sigma or (six adj sigma)).mp. [mp=title, original title, abstract, name of substance word, subject heading word]

167. total quality management.ti,ab.

168. ((quality adj function adj deployment) or (house adj2 quality) or (quality adj circle) or (breakthrough adj series)).mp. [mp=title, original title, abstract, name of substance word, subject heading word]

169. ((institute adj2 healthcare adj improvement) or (iso adj "9004") or (iso adj 15594*)).mp. [mp=title, original title, abstract, name of substance word, subject heading word]

170. kaizen.ti,ab. or (toyota adj production adj system).mp. or (toyota adj a3).mp. [mp=title, original title, abstract, name of substance word, subject heading word]

171. ((lean adj manufacturing) or (lean adj production) or (lean adj healthcare) or (lean adj health adj care) or (lean adj health adj service) or (lean adj healthcare adj service) or (lean adj health adj care adj service)).mp. [mp=title, original title, abstract, name of substance word, subject heading word]

172. ((inventive adj problem adj solving) or (inventive adj problem-solving) or (inventive adj problemsolving) or (business adj process adj reengineering) or (business adj process adj re-engineering)).mp. [mp=title, original title, abstract, name of substance word, subject heading word]

173. (business adj process adj engineering).ti,ab.

174. (IHI or (Institute adj Healthcare adj Improvement)).mp. [mp=title, original title, abstract, name of substance word, subject heading word]

175. (system* adj redesign).mp. [mp=title, original title, abstract, name of substance word, subject heading word]

179. 170 or 161 or 165 or 168 or 167 or 163 or 172 or 166 or 174 or 173 or 164 or 169 or 175 or 171

180. limit 179 to (english language and yr="2004 -Current")

NUMBER OF RESULTS: 696

CQI SEARCH #2:

176. ((process adj improvement) or (improvement adj collaborative*)).mp.

177. (advanced adj clinical adj access).mp. [mp=title, original title, abstract, name of substance word, subject heading word]

178. cqi.ti,ab.

186. 178 or 177 or 176

187. limit 186 to (english language and yr="2004 -Current")

NUMBER OF RESULTS: 203

ALL SEARCHES WERE COMBINED INTO ONE ENDNOTE FILE AND DUPLICATES WERE REMOVED. TOTAL NUMBER OF RESULTS IN COMBINED FILE: 9426 (9,395 after further duplicate removal based on the PubMed identifier)

Appendix C. Inclusion Screening Form for Publications Obtained as Full Text

Full Text 4Q Screener

Article #: _____ Reviewer: _____ Date of Review: _____

First Author, Year: _____

1st 2 Words in Title: _____

REVIEW FULL ARTICLE:	Yes	No/Don't Know
<p>1. <i>Setting</i>: Did the article report on an intervention that was implemented in or by a healthcare delivery organization or organizational unit? <i>(This item requires a) an identifiable, purposefully introduced intervention, and b) an organizational setting. Studies of naturalistic evolution of care are excluded. For organization or organizational unit, include settings such as hospitals, clinics, provider offices, nursing homes, and community settings (e.g., home health services, school-based clinics, public health clinics. Exclude research studies that recruit patients outside of a healthcare delivery organization or organizational unit.)</i></p> <p>2. <i>Design</i>: Were qualitative or quantitative data on the <u>effectiveness, impacts, or success</u> of the intervention reported in the article? <i>(This item requires the presence of data on whether or not the intervention worked. Exclude meta-analyses, systematic reviews, stories, tool development articles. Also exclude articles reporting data only on relationships a) between one element of context and another, or b) between context and intervention process or outcomes.</i></p> <p>3. <i>Outcome</i>: Did the article report on patient (or care-giver), provider behavior, or process of care health outcomes? <i>(This item requires the presence of data on a health-related outcome(s). Exclude articles that report only cost data, or data on provider knowledge or satisfaction in the absence of impact on patient (or care-giver) health outcomes.)</i></p> <p>4. <i>Intervention</i>: Did the article suggest that the intervention aimed to change how delivery of care is routinely structured within a specific organization or organizational unit? <i>(This item requires that the article suggests a) that members of the subject organization or organizational unit were actively involved in carrying out the intervention, and b) there is an intention to continue the intervention as an ongoing activity in the subject organization, if successful and within cost constraints. Include any change in delivery of care (e.g., introduction of checklist) if used in routine or daily activities. Exclude one time educational interventions.)</i></p>		

Appendix D. Machine Learning Confusion Matrix, Text Terms Distinguishing Relevant and Irrelevant Citations, and Reviewer Disagreements

Appendix Table 1: Machine Learning Confusion Matrix

Prediction	Reviewer's decision: Irrelevant (discard citation)	Reviewer's decision: Relevant (obtain citation as full text)
Irrelevant (discard citation)	True negative	False negative
Relevant (obtain citation as full text)	False positive	True positive

Appendix Table 2: Text Terms Distinguishing Relevant and Irrelevant Citations

Terms Overrepresented in Relevant Citations	Terms Underrepresented in Relevant Citations
"improving"	"adequate"
"patients"	"public"
"interventions"	"search"
"implementing"	"evolving"
"staff"	"consensus"
"month"	"panel"
"programs"	"response"
"significantly"	"next"
"after"	"right"
"hospitalizations"	"still"
"process"	"transparency"
"controlled"	"balanced"
"initial"	"ratio"
"team"	"personnel"
"before"	"multivariable"
"impact"	"private"
"prospective"	"purchasers"
"with"	"linear"
"rate"	"literature"
"described"	"characterized"
"increased"	"compromise"
"changing"	"know"
"practice"	"agreement"
"projects"	"definitions"
"outcomes"	"campaign"
"department"	"boards"
"reducing"	"systematic"
"had"	"principally"
"care"	"typically"
"through"	"derived"
"baseline"	"sources"
"led"	"maintenance"
"feedback"	"turn"
"keeping"	"decade"
"detailing"	"century"
"diabetes"	"says"
"up"	"fact"
"success"	"suitable"

Appendix Table 3: Reviewer Disagreements

Citation with disagreement (included by one out of two independent reviewers)	Final decision and reason for discarding
<p>Integrating palliative and critical care: evaluation of a quality-improvement intervention. Am J Respir Crit Care Med. 2008 Aug 1;178(3):269-75. Epub 2008 May 14. Curtis JR, Treece PD, Nielsen EL, Downey L, Shannon SE, Braungardt T, Owens D, Steinberg KP, Engelberg RA. Division of Pulmonary and Critical Care, Box 359762, Harborview Medical Center, University of Washington, Seattle, WA 98104, USA. jrc@u.washington.edu RATIONALE: Palliative care in the intensive care unit (ICU) is an important focus for quality improvement. OBJECTIVES: To evaluate the effectiveness of a multi-faceted quality improvement intervention to improve palliative care in the ICU. METHODS: We performed a single-hospital, before-after study of a quality-improvement intervention to improve palliative care in the ICU. The intervention consisted of clinician education, local champions, academic detailing, feedback to clinicians, and system support. Consecutive patients who died in the ICU were identified pre- (n = 253) and postintervention (n = 337). Families completed Family Satisfaction in the Intensive Care Unit (FS-ICU) and Quality of Dying and Death (QODD) surveys. Nurses completed the QODD. The QODD and FS-ICU were scored from 0 to 100. We used Mann-Whitney tests to assess family results and hierarchical linear modeling for nurse results. MEASUREMENTS AND MAIN RESULTS: There were 590 patients who died in the ICU or within 24 hours of transfer; 496 had an identified family member. The response rate for family members was 55% (275 of 496) and for nurses, 89% (523/590). The primary outcome, the family QODD, showed a trend toward improvement (pre, 62.3; post, 67.1), but was not statistically significant (P = 0.09). Family satisfaction increased but not significantly. The nurse QODD showed significant improvement (pre, 63.1; post, 67.1; P < 0.01) and there was a significant reduction in ICU days before death (pre, 7.2; post, 5.8; P < 0.01). CONCLUSIONS: We found no significant improvement in family-assessed quality of dying or in family satisfaction with care, we found but significant improvement in nurse-assessed quality of dying and reduction in ICU length of stay with an intervention to integrate palliative care in the ICU. Improving family ratings may require interventions that have more direct contact with family members. PMID: 18480429 [PubMed—indexed for MEDLINE]</p>	<p>Obtain</p>

<p>A quality improvement program to enhance after-hours telephone communication between nurses and physicians in a long-term care facility. J Am Geriatr Soc. 2008 Jun;56(6):1080-6. Epub 2008 Apr 11. Whitson HE, Hastings SN, Lekan DA, Sloane R, White HK, McConnell ES. Department of Medicine, Division of Geriatric Medicine, Duke University Medical Center, Durham, North Carolina 27710, USA. heather.whitson@duke.edu OBJECTIVES: To determine whether satisfaction of on-site nurses with after-hours telephone communication with off-site physicians improved in one long-term care (LTC) facility after a nurse-oriented intervention. DESIGN: Longitudinal quality improvement study. SETTING: Extended Care and Rehabilitation Center (ECRC), Durham Veterans Affairs Medical Center. PARTICIPANTS: Eighteen registered nurses. INTERVENTION: Communicating Health Assessments by Telephone (Project CHAT), a program of individualized training sessions and decision support tools to aid LTC nurses with symptom assessment and communication of health information over the telephone. MEASUREMENTS: Nurses completed six satisfaction surveys (three surveys in the 3 months before Project CHAT and three surveys in the 3 months after Project CHAT). RESULTS: The nurses' average satisfaction scores increased on several items, including those that assessed whether the nurse was pretty sure what pieces of information the physician was going to ask for (P=.04), felt that the amount of patient information the physician asked for seemed reasonable (P=.03), felt prepared to answer the questions the physician asked (P=.01), and felt that the process of gathering patient information for the physician was easy (P=.01). The percentage of calls that resulted in immediate evaluation by a physician (on-site or in the emergency department) increased from 2.0% in the period before Project CHAT to 8.6% in the period after Project CHAT (P=.01). CONCLUSION: Nurses' satisfaction with several aspects of after-hours telephone medicine improved after an inexpensive, education-based intervention in one LTC facility. Further research is needed to determine how similar interventions might affect other quality measures, including patient outcomes. PMID: 18410318 [PubMed—indexed for MEDLINE]</p>	<p>Discard. Reason: Outcome (provider satisfaction)</p>
<p>Improving program documentation quality through the application of continuous improvement processes. J Contin Educ Nurs. 2007 Nov-Dec;38(6):271-6. Lovlien CA, Johansen M, Timm S, Eversman S, Gusa D, Twedell D. Department of Nursing, Mayo Clinic-Rochester, Rochester, Minnesota 55905, USA. Maintaining the integrity of record keeping and retrievable information related to the provision of continuing education credit creates challenges for a large organization. Accurate educational program documentation is vital to support the knowledge and professional development of nursing staff. Quality review and accurate documentation of programs for nursing staff development occurred at one institution through the use of continuous improvement principles. Integration of the new process into the current system maintains the process of providing quality record keeping. PMID: 18050984 [PubMed—indexed for MEDLINE]</p>	<p>Discard. Reason: Outcome (documentation of education)</p>
<p>The evaluation of Healthy Families Arizona: a multisite home visitation program. J Prev Interv Community. 2007;34(1-2):109-27. Krysik J, LeCroy CW. Arizona State University, AZ, USA. jkrysik@msn.com</p> <p>Healthy Families Arizona is a broadly implemented home visitation program aimed at preventing child abuse and neglect, improving child health and development, and promoting positive parent/child interaction. The program began as a pilot in two sites in 1991 and by 2004 had grown to 48 sites located in urban, rural, and tribal regions of the state. The unique administrative structure of the program and collaboration between evaluation and quality assurance have helped overcome many of the problems familiar to home visitation programs. This paper describes how a systematic focus to improve processes and outcomes has positioned the program for a randomized longitudinal study. Key components of the program are described and evaluation results are presented. PMID: 17890196 [PubMed—indexed for MEDLINE]</p>	<p>Obtain</p>
<p>Results of a 12-month quality-circle prescribing improvement programme for GPs. Br J Gen Pract. 2007 Jul;57(540):574-6. Riou F, Piette C, Durand G, Chaperon J. Department de Santé Publique, Faculté de Médecine, Université de Rennes, Rennes, France. francoise.riou@univ-rennes1.fr</p> <p>BACKGROUND: The impact of a quality-circle GP prescribing improvement programme, implemented in France in 2001-2002, was assessed by a controlled study. The study involved all 27 GPs of three semi-rural areas of Brittany, France. Practice data (overall prescribing cost and markers of prescribing efficiency) were collated in an intention-to-treat analysis, using the Mann-Whitney U test. Twenty-four GPs attended the meetings regularly. The reduction in drug expenditure exceeded the cost of the programme, although variations in size effects were observed among the settings. PMCID: PMC2099641 PMID: 17727751 [PubMed—indexed for MEDLINE]</p>	<p>Discard. Reason: Outcome (not health related)</p>

<p>Recent trends in the care of patients with non-ST-segment elevation acute coronary syndromes: insights from the CRUSADE initiative. Arch Intern Med. 2006 Oct 9;166(18):2027-34. Mehta RH, Roe MT, Chen AY, Lytle BL, Pollack CV Jr, Brindis RG, Smith SC Jr, Harrington RA, Fintel D, Fraulo ES, Califf RM, Gibler WB, Ohman EM, Peterson ED. Duke Clinical Research Institute, Duke University Medical Center, Durham, NC 27715, USA. mehta007@dcric.duke.edu</p> <p>BACKGROUND: The extent to which national health quality improvement initiatives have altered reported treatment gaps among patients with non-ST-segment elevation acute coronary syndromes (NSTEMI ACS) is unknown. We sought to determine recent trends in adherence to guideline-based therapies for NSTEMI ACS. METHODS: We evaluated the treatment of patients with high-risk (positive cardiac markers and/or ischemic ST-segment changes) NSTEMI ACS enrolled in the Can Rapid Risk Stratification of Unstable Angina Patients Suppress Adverse Outcomes With Early Implementation of the ACC/AHA (American College of Cardiology/American Heart Association) Guidelines (CRUSADE) Quality Improvement Initiative from 2002 through 2004 (a total of 113 595 patients over 11 calendar quarters). We analyzed adherence to guideline-recommended therapies, including medications used in the acute care period (<24 hours after presentation), invasive procedures, in-hospital outcomes, and discharge therapies and interventions. RESULTS: The use of each class I guideline recommendation, as well as overall adherence to the guidelines, improved significantly (P<.001) during the study period. In the acute care setting, the use of antiplatelet agents increased by 5% and beta-blockers by 12%; at hospital discharge, the use of antiplatelet agents increased by 3% and beta-blockers by 8%. Heparin use in the acute care period increased by 6%, largely owing to a 9% increase in the use of low-molecular-weight heparin. Use of glycoprotein IIb/IIIa inhibitors in the acute care period also increased by more than 13%. At discharge, clopidogrel use increased by 22%, lipid-lowering agents by 11%, and angiotensin-converting enzyme inhibitors by 5%. While adherence improved, many patients still failed to receive 100% indicated treatments at the end of the study period. CONCLUSIONS: During the 4 years since the initial release of the ACC/AHA guidelines for NSTEMI ACS, adherence to class I recommendations has significantly improved among hospitals participating in CRUSADE. Still, further improvements are needed for optimal implementation of these guidelines. PMID: 17030838 [PubMed—indexed for MEDLINE]</p>	<p>Obtain</p>
<p>The impact of quality-reporting programs on hospital operations. Health Aff (Millwood). 2006 Sep-Oct;25(5):1412-22. Pham HH, Coughlan J, O'Malley AS. Center for Studying Health System Change, Washington, DC, USA. mpham@hschange.org Comment in Health Aff (Millwood). 2007 Jan-Feb;26(1):294-5. Health Aff (Millwood). 2007 Jan-Feb;26(1):294.</p> <p>We used data from the 2005-06 Community Tracking Study site visits to examine the impact of quality reporting on hospitals' data collection and review processes, feedback and accountability mechanisms, quality improvement activities, and resource allocation. Individual hospitals participate in multiple, varied reporting programs with distinct effects on hospital operations. Reporting programs play complementary roles in encouraging quality improvement but are poorly coordinated and command sizable resources, in large part because of inadequate information technology. Policy should be directed at encouraging formal assessments of how individual and combinations of programs affect quality outcomes, and the development of adaptable information systems. PMID: 16966741 [PubMed—indexed for MEDLINE]</p>	<p>Discard. Reason: Design (no intervention but correlational effect investigated)</p>

<p>Detection and prevention of medication errors using real-time bedside nurse charting. <i>J Am Med Inform Assoc.</i> 2005 Jul-Aug;12(4):390-7. Epub 2005 Mar 31. Nelson NC, Evans RS, Samore MH, Gardner RM. Department of Medical Informatics, LDS Hospital, Salt Lake City, UT 84143, USA. ldnnelso@ihc.com</p> <p>OBJECTIVE: Charting systems with decision support have been developed to assist with medication charting, but many of the features of these programs are not properly used in their clinical application. An analysis of medication error reports at LDS Hospital revealed the occurrence of errors that should have been detected and prevented by decision support features if real-time entry at the bedside had taken place. The aim of this study was to increase the real-time bedside charting behavior of nurses. DESIGN: A quasiexperimental before and after design was used. The study took place in two 40-bed surgical units, one of which served as the study unit, the other as control unit. The study unit received educational intervention about error avoidance through real-time bedside charting, and 12 weeks of monitoring and performance feedback. The real-time and bedside charting rates for the study and control units were measured before and after the intervention. RESULTS: Before the intervention on the study unit, the real-time charting rate was 59% and the bedside rate was 40%. At the conclusion of a 12-week intervention period, the real-time rate increased to 73% and the bedside rate increased to 63%. Postintervention real-time rates were 75% after eight weeks and remained at 75% after one year. Equivalent control unit real-time rates varied from 53% to 57%, and bedside rates varied from 34% to 44% during the same intervals. CONCLUSION: Targeted educational intervention and monitored feedback yielded measurable improvements in the effective use of the computerized medication charting system and must be an ongoing process. PMCID: PMC1174883 PMID: 15802486 [PubMed—indexed for MEDLINE]</p>	Obtain
<p>A multimethod quality improvement intervention to improve preventive cardiovascular care: a cluster randomized trial. <i>Ann Intern Med.</i> 2004 Oct 5;141(7):523-32. Ornstein S, Jenkins RG, Nietert PJ, Feifer C, Roylance LF, Nemeth L, Corley S, Dickerson L, Bradford WD, Litvin C. Department of Family Medicine, Medical University of South Carolina, Charleston, South Carolina 29425, USA. ornstesm@musc.edu Summary for patients in <i>Ann Intern Med.</i> 2004 Oct 5;141(7):I53.</p> <p>BACKGROUND: Research is needed to validate effective and practical strategies for improving the provision of evidence-based medicine in primary care. OBJECTIVE: To determine whether a multimethod quality improvement intervention was more effective than a less intensive intervention for improving adherence to 21 quality indicators for primary and secondary prevention of cardiovascular disease and stroke. DESIGN: 2-year randomized, controlled clinical trial with the practice as the unit of randomization. SETTING: 20 community-based family or general internal medicine practices in 14 states. All used the same electronic medical record. PARTICIPANTS: 44 physicians, 17 midlevel providers, and approximately 200 staff members; data from the electronic medical records of 87,291 patients. INTERVENTIONS: All practices received copies of practice guidelines and quarterly performance reports. Intervention practices also hosted quarterly site visits to help them adopt quality improvement approaches and participated in 2 network meetings to share "best practice" approaches. MEASUREMENTS: The percentage of indicators at or above predefined targets and the percentage of patients who had achieved each clinical indicator. RESULTS: Intervention practices improved 22.4 percentage points (from 11.3% to 33.7%) in the percentage of indicators at or above the target; control practices improved 16.4 percentage points (from 6.3% to 22.7%). The 6.0-percentage point absolute difference between the intervention and control group was not statistically significant ($P > 0.2$). Patients in intervention practices had greater improvements than those in control practices for diagnoses of hypertension (improvement difference, 15.7 percentage points [95% CI, 5.2 to 26.3 percentage points]) and blood pressure control in patients with hypertension (improvement difference, 8.0 percentage points [CI, 0.0 to 16.0 percentage points]). LIMITATIONS: The study involved a small number of practices and lacked a pure control group. CONCLUSIONS: Primary care practices that use electronic medical records and receive regular performance reports can improve their adherence to clinical practice guidelines for cardiovascular disease and stroke prevention. PMID: 15466769 [PubMed—indexed for MEDLINE]</p>	Obtain

<p>Practice based education to improve delivery systems for prevention in primary care: randomised trial. BMJ. 2004 Feb 14;328(7436):388. Epub 2004 Feb 6. Margolis PA, Lannon CM, Stuart JM, Fried BJ, Keyes-Elstein L, Moore DE Jr. University of North Carolina at Chapel Hill, North Carolina Center for Children's Healthcare Improvement, 730 Airport Rd, Ste 104, CB#7226, Chapel Hill, NC 27599, USA. Peter_Margolis@med.unc.edu</p> <p>OBJECTIVE: To examine the effectiveness of an intervention that combined continuing medical education with process improvement methods to implement "office systems" to improve the delivery of preventive care to children. DESIGN: Randomised trial in primary care practices. SETTING: Private paediatric and family practices in two areas of North Carolina. PARTICIPANTS: Random sample of 44 practices allocated to intervention and control groups. INTERVENTION: Practice based continuing medical education in which project staff coached practice staff in reviewing performance and identifying, testing, and implementing new care processes (such as chart screening) to improve delivery of preventive care. MAIN OUTCOME MEASURE: Change over time in the proportion of children aged 24-30 months who received age appropriate care for four preventive services (immunisations, and screening for tuberculosis, anaemia, and lead). RESULTS: The proportion of children per practice with age appropriate delivery of all four preventive services changed, after a one year period of implementation, from 7% to 34% in intervention practices and from 9% to 10% in control practices. After adjustment for baseline differences in the groups, the change in the prevalence of all four services between the beginning and the end of the study was 4.6-fold greater (95% confidence interval 1.6 to 13.2) in intervention practices. Thirty months after baseline, the proportion of children who were up to date with preventive services was higher in intervention than in control practices; results for screening for tuberculosis (54% v 32%), lead (68% v 30%), and anaemia (79% v 71%) were statistically significant (P < 0.05). CONCLUSION: Continuing education combined with process improvement methods is effective in increasing rates of delivery of preventive care to children. PMCID: PMC341391 PMID: 14766718 [PubMed—indexed for MEDLINE]</p>	Obtain
<p>Nursing home care: changes after supervision. J Adv Nurs. 2004 Feb;45(3):269-79. Hansebo G, Kihlgren M. Senior Lecturer, Department of Neurotec, Karolinska Institutet, Ersta Sköndal University College, Stockholm, Sweden. gorel.hansebo@euc.ersta.se</p> <p>BACKGROUND: An intervention project was conducted in three nursing home wards in Sweden. Most patients had severe dementia. The intervention consisted of supervision for individualized and documented nursing care, based on multidimensional assessment. AIM: To illuminate changes in carers' approach after the intervention. METHODS: Several data collections were conducted across the intervention and consisted of nursing documentation, patient life stories as told by carers, video recorded interactions, stimulated recall interviews and a questionnaire. Both quantitative and qualitative methods were used in the analyses. FINDINGS: The findings from the different methods mirrored each other and added to the credibility of the intervention. Communicated knowledge about patients improved in nursing documentation and also as told by carers. Carers were differently skilled in managing the complexity of nursing care situations before as well as after the intervention, but the intervention contributed to developing carers in 'confirming nursing care'. They also improved in their ability to verbalize reflections about their everyday life with patients with dementia. CONCLUSION: Supervision made it possible for carers to share their lived experiences about their day-to-day life with patients, which could promote personal and professional development and thus improve care quality. It also appeared that a detailed assessment tool used as part of the nursing process contributed to seeing a patient as a real person behind a dementia surface. PMID: 14720244 [PubMed—indexed for MEDLINE]</p>	Obtain

<p>Continuous monitoring of adverse events: influence on the quality of care and the incidence of errors in general surgery. Rebaso P, Mora L, Luna A, Montmany S, Vallverdú H, Navarro S. Corporacio Sanitària Parc Tauli, Parc Tauli s/n, 08208, Sabadell, Spain. prebaso@cspt.es</p> <p>BACKGROUND: This study was designed to determine the incidence of adverse events and errors in the care of surgical patients and to demonstrate that continuous prospective collection of data on adverse events can improve quality of care and reduce the number of errors. Retrospective studies find adverse events in approximately 5% of patients admitted. Prospective studies publish figures of approximately 30%. No studies to date have tried to use continuous collection of data on adverse events to reduce the incidence of errors. METHODS: Longitudinal prospective surveillance of adverse events in patients admitted to the Surgery Service during a 22-month period. Sequelae after discharge and errors during hospital stay were evaluated by peer review. RESULTS: A total of 3,807 patients were controlled: 1,177 patients presented 2,193 adverse events (30.9% of admissions); 330 adverse events due to errors were detected in 258 patients (6.9% of admissions). Thirty-four deaths were considered due to adverse events (0.89% of admissions), and in 11 cases mortality was deemed avoidable (0.29% of admissions). The incidence of adverse events remained constant during the study period, but errors decreased from 11.1% to 4.5% (P = 0.005). CONCLUSIONS: This is the first attempt to determine the prevalence of errors in surgery. Introducing systematic programs for recording adverse events can reduce error rates and promote a culture of patient safety in a General Surgery Department. PMID: 19082657 [PubMed—indexed for MEDLINE]</p>	<p>Obtain</p>
<p>Relationship between the presence of practice systems and the quality of care for depression. Am J Med Qual. 2008 Nov-Dec;23(6):420-6. Solberg LI, Asche SE, Margolis KL, Whitebird RR, Trangle MA, Wineman AP. HealthPartners Research Foundation and Medical Group, Minneapolis, Minnesota 55440-1524, USA. leif.i.solberg@healthpartners.com</p> <p>A valid measure of practice systems for improving chronic disease care is needed as a guide for both improvement and public accountability. We tested whether a new survey measure of the presence of practice systems (the PPC-R) is associated with performance measure rates for depression among 40 medical groups in Minnesota. These PPC-R scores were compared with standardized medical group measures of antidepressant persistence. Only 54% of potentially important systems were present, and there was high variability. However, there was a positive correlation between systems and quality on the 90-day measure of antidepressant persistence, both overall (r = .33, P = .04) and for the Chronic Care Model domains of decision support (r = .38, P = .02) and delivery system redesign (r = .31, P = .05). Thus, practice systems overall and several domains of the Chronic Care Model appear to be associated with higher quality care for depression. This questionnaire may help practices identify particular systems to improve. PMID: 19001099 [PubMed—indexed for MEDLINE]</p>	<p>Discard. Reason: Design (not effect of intervention tested)</p>

<p>Implementing after-hours pharmacy coverage for critical access hospitals in northeast Minnesota. Am J Health Syst Pharm. 2008 Sep 15;65(18):1727-34. Stratton TP, Worley MM, Schmidt M, Dudzik M. Department of Pharmacy Practice and Pharmaceutical Sciences, University of Minnesota College of Pharmacy, Duluth, MN 55812, USA. tstratto@d.umn.edu <tstratto@d.umn.edu> Comment in Am J Health Syst Pharm. 2009 Apr 1;66(7):612; author reply 614.</p> <p>PURPOSE: A project that used health information technology (IT) to provide after-hours pharmacy coverage to critical access hospitals in northeast Minnesota is described. SUMMARY: SISU Medical Systems was established to address the health care IT needs of the Wilderness Health Care Coalition hospitals. Administrators and nursing and pharmacy leaders at several Wilderness Coalition hospitals were interested in obtaining after-hours pharmacy services to optimize patient safety. Eight of the Wilderness Coalition critical access hospitals obtained the technology necessary to allow pharmacy staff at St. Luke's Hospital (the hub hospital) in Duluth, Minnesota, to electronically enter orders into the rural hospitals' patient electronic medical records. The system placed the orders into the patients' medication profiles on automated dispensing machines located at seven of the eight rural hospitals. The pharmacy computer system allowed for medication order processing, drug interaction checking, medication dispensing via automated dispensing cabinets at the rural hospital sites, and formulary and inventory management. Medications that were not available in a rural hospital's automated dispensing cabinet were obtained from the locked pharmacy by the nurse supervisor. Round-the-clock pharmacy coverage was almost achieved. Participating rural hospitals received 24-hour coverage from the hub hospital during weekends and holidays, but no after-hours (4 a.m.-7 a.m.) coverage was provided on weekdays. The staff at the rural hospitals determined from their experiences that new orders were less likely to be written during these hours. CONCLUSION: Using Internet-based health IT, pharmacists from a metropolitan (hub) hospital with round-the-clock pharmacist coverage participated in the care of patients at a number of small, rural hospitals and helped ensure that those patients received safe and effective medication therapy. The coverage provided by pharmacists at the hub hospital improved nursing satisfaction with the overall quality of pharmacy services provided by both the hub hospital and the local onsite pharmacists. PMID: 18768999 [PubMed—indexed for MEDLINE]</p>	<p>Discard. Reason: Outcome (provider satisfaction)</p>
<p>From ethics to palliative care: a community hospital experience. Nurs Clin North Am. 2008 Sep;43(3):469-76, x. Robley LR. Siegel Institute for Leadership, Ethics, and Character, Kennesaw State University, Mailbox 5500, 1000 Chastain Road, Kennesaw, GA 30144, USA. lrobley@kennesaw.edu</p> <p>One of the most vulnerable and voiceless groups of patients within American hospitals and institutions today are those who are dying. Health care institutions struggle with the challenge of providing excellent palliative and end-of-life care to patients while providing curative therapies at the same time. This article describes the efforts and accomplishments of the ethics committee of a community hospital system to provide for the palliative and end-of-life needs of its patients. PMID: 18674676 [PubMed—indexed for MEDLINE]</p>	<p>Obtain</p>

<p>Anthroposophic health care--different and home-like. <i>Scand J Caring Sci.</i> 2008 Sep;22(3):357-66. Arman M, Ranheim A, Rehnsfeldt A, Wode K. Department of Neurobiology, Care Sciences and Society, Division of Nursing, Karolinska Institutet, Huddinge, Sweden. maria.arman@ki.se</p> <p>Anthroposophic health care is rooted in the work of Steiner and Wegman in Switzerland during the 1920s. The Swedish hospital in this study offers integrated conventional and anthroposophic health care therapies which are conceptualized as an extended and integrative variant of health care and not as CAM. In anthroposophic care, health is viewed as a matter of body, soul and spirit in balance. Therapeutic resources include nursing care, therapeutic conduct (art and body therapies) and medicines based on natural remedies. This study aims to deepen the understanding of what constitutes good care from a patient's perspective to alleviate patients' suffering and to identify clinical markers for good care. As anthroposophic care is associated with theory and holistic ideas, this study aims at exploring whether or not anthroposophic care has a beneficial effect. A qualitative method was used, and the analysis was conducted with a phenomenological hermeneutic approach. Sixteen former patients, of whom nine were diagnosed with various kinds of cancer and seven with burnout syndrome, were interviewed regarding their experience of anthroposophic care. Patients especially noted the benefits of the holistic caring environment; the empathetic approach and true caring offered, as well as the peaceful atmosphere and rest. A turning point or shift in perspectives, implying a home coming in relation to inner aspects was discussed as an outcome. Although patients in general were overwhelmingly impressed and positive they were also ambivalent. One interpretation is that there is a gap between the anthroposophic and conventional paradigm that affects patients negatively. As mutual scepticism still prevents any real integration between integrative and conventional care, the onus appears to be on the patient to take the risk and act as bridge-builder. From a caring science perspective, the study shows that appropriation of specific values and theory makes it possible to create a true caring culture. PMID: 18840219 [PubMed—indexed for MEDLINE]</p>	<p>Discard. Reason: Setting (not specific health care delivery organization)Obtain</p>
<p>Mother-child relationship quality and effective discipline as mediators of the 6-year effects of the New Beginnings Program for children from divorced families. <i>J Consult Clin Psychol.</i> 2008 Aug;76(4):579-94. Zhou Q, Sandler IN, Millsap RE, Wolchik SA, Dawson-McClure SR. Department of Psychology, University of California, Berkeley, CA 94720-1650, USA. qingzhou@berkeley.edu</p> <p>This study examines whether program effects on mother-child relationship quality and effective discipline mediated the 6-year longitudinal effects of the New Beginnings Program (NBP) to improve mental health and competence outcomes in 218 adolescents from divorced families in a randomized experimental trial. The NBP is a theory-based and parenting-focused preventive intervention to help children adjust to divorce, and it has previously shown significant main and/or Program x Baseline Risk interaction effects to reduce adolescents' mental health and social adaptation problems and to promote competence. Mediation analyses were conducted using single- and two-group (high and low baseline risk) structural equation modeling. A multiple-methods and multiple-informants approach was used to assess the putative mediators and adolescents' outcomes. Results indicated that program-induced improvement in maternal effective discipline at posttest mediated the intervention effect on adolescents' GPA at the 6-year follow-up. Moreover, program-induced improvement in mother-child relationship quality mediated the intervention effect on adolescents' mental health problems for those with high baseline risk for maladjustment. The discussion focuses on the implications of the mediation findings for advancing the developmental theories that informed the design of the NBP and the implications for implementation of the NBP in community settings. Copyright 2008 APA, all rights reserved. PMID: 18665687 [PubMed—indexed for MEDLINE]</p>	<p>Discard. Reason: Intervention (therapy, not different delivery method)</p>

<p>Impact of hospital policies on breastfeeding outcomes. <i>Breastfeed Med.</i> 2008 Jun;3(2):110-6. Rosenberg KD, Stull JD, Adler MR, Kasehagen LJ, Crivelli-Kovach A. Office of Family Health, Oregon Public Health Division, Portland, Oregon 97232, USA. ken.d.rosenberg@state.or.us</p> <p>OBJECTIVE: The objective of this study was to explore the association between the Ten Steps of the Baby Friendly Hospital Initiative (BFHI) of the World Health Organization (Geneva, Switzerland) and breastfeeding at 2 days and 2 weeks. METHODS: A 65-question institutional survey assessing compliance with the Ten Steps was used to determine an overall breastfeeding Support Score for each of Oregon's 57 birthing hospitals. Hospital breastfeeding outcomes were obtained from the newborn metabolic screening forms. RESULTS: Hospitals' overall breastfeeding Support Scores ranged from 49.4 to 98.2 out of a possible total score of 100. Hospital compliance with individual Steps ranged from 5.3% for Step 2 (staff training) to 93% for Step 4 (helping with breastfeeding initiation) and Step 8 (encouraging feeding on demand). After controlling for institutional differences (by multivariate linear regression) we found that increases in overall hospital breastfeeding Support Scores were associated with increases in breastfeeding percentage at 2 days ($p = 0.021$) and at 2 weeks postpartum ($p = 0.011$). In analyzing each Step individually, however, only the presence of a written hospital policy was independently associated with breastfeeding percent ($p = 0.028$). CONCLUSIONS: This institutional-level evaluation corroborates previous findings demonstrating that increased implementation of the Ten Steps is associated with increased breastfeeding. Further, it suggests that hospitals with comprehensive breastfeeding policies are likely to have better breastfeeding support services and better breastfeeding outcomes. Hospitals may consider using these results to prioritize breastfeeding support services through development of hospital breastfeeding policies and to utilize institutional surveys as a component of breastfeeding quality improvement initiatives. PMID: 18563999 [PubMed—indexed for MEDLINE]</p>	<p>Obtain</p>
<p>Clarifying values and transforming attitudes to improve access to second trimester abortion. <i>Reprod Health Matters.</i> 2008 May;16(31 Suppl):108-16. Turner KL, Hyman AG, Gabriel MC. Ipas, Chapel Hill, NC, USA. turnerk@ipas.org</p> <p>Access to safe second trimester abortion services is poor in many countries, sometimes despite liberal laws and policies. Addressing the myriad factors hindering access to safe abortion care requires a multi-pronged strategy. Workshops aimed at clarifying values are useful for addressing barriers to access stemming from misinformation, stigmatization of women and providers, and negative attitudes and obstructionist behaviours. They engage health care providers and administrators, policymakers, community members and others in a process of self-examination with the goal of transforming abortion-related attitudes and behaviours in a direction supportive of women seeking abortion. This is especially important for women seeking second trimester abortion, which tends to be even more stigmatized than first trimester abortion. This paper reports on some promising experiences and results from workshops in Viet Nam, Nepal and South Africa. Some recommendations that emerge are that values clarification should be included in abortion training, service delivery and advocacy programmes. Evaluations of such interventions are also needed. PMID: 18772091 [PubMed—indexed for MEDLINE]</p>	<p>Discard: Reason: Intervention (one-time education)</p>

<p>Interprofessional relationships and communication in primary palliative care: impact of the Gold Standards Framework. Br J Gen Pract. 2008 Apr;58(549):256-63. Mahmood-Yousuf K, Munday D, King N, Dale J. Health Services Research Institute, Warwick Medical School, University of Warwick, Coventry, UK. Comment in Br J Gen Pract. 2008 Apr;58(549):231-4.</p> <p>BACKGROUND: High quality end-of-life care in the community is achieved with effective multidisciplinary teamwork, interprofessional communication between GPs and district nurses, and early referral of patients to district nurses. These aspects of palliative care are highlighted in the Gold Standards Framework, a programme recently established in UK primary care. AIM: To investigate the extent to which the framework influences interprofessional relationships and communication, and to compare GPs' and nurses' experiences. DESIGN OF STUDY: Qualitative interview case study. SETTING: Fifteen participating practices from three primary care trusts in England. METHOD: Thirty-eight semi-structured interviews were undertaken with GPs, district nurses, Macmillan nurses, and framework facilitators. RESULTS: Adoption of the framework often resulted in earlier referral of palliative care patients to district nurses. Multidisciplinary team meetings enabled communication for sharing knowledge, discussing management problems, and keeping colleagues informed; however, arranging and maintaining such meetings was often problematic. Nurses particularly valued formal meetings while GPs generally preferred informal ad hoc dialogue. GPs largely maintained control of the mode of multidisciplinary working. The best functioning teams used a mixture of formal and informal meetings with a relatively non-hierarchical working style. CONCLUSION: Implementing the framework enabled processes of communication associated with high quality palliative care in general practice, but there was marked variation in how this worked in individual teams. In general, hierarchical doctor-nurse relationships persisted. PMID: 18387229 [PubMed—indexed for MEDLINE]</p>	<p>Discard. Reason: Outcome (provider interview data only)</p>
<p>Do hospitalists or physicians with greater inpatient HIV experience improve HIV care in the era of highly active antiretroviral therapy? Results from a multicenter trial of academic hospitalists. Clin Infect Dis. 2008 Apr 1;46(7):1085-92. Schneider JA, Zhang Q, Auerbach A, Gonzales D, Kaboli P, Schnipper J, Wetterneck TB, Pitrak DL, Meltzer DO. Department of Medicine, Tufts-New England Medical Center, Boston, Massachusetts, USA. jschneil@medicine.bsd.uchicago.edu</p> <p>BACKGROUND: Little is known about the effect of provider type and experience on outcomes, resource use, and processes of care of hospitalized patients with human immunodeficiency virus (HIV) infection. Hospitalists are caring for this population with increasing frequency. METHODS: Data from a natural experiment in which patients were assigned to physicians on the basis of call cycle was used to study the effects of provider type—that is, hospitalist vs. nonhospitalist—and HIV-specific inpatient experience on resource use, outcomes, and selected measures of processes of care at 6 academic institutions. Administrative data, inpatient interviews, 30-day follow-up interviews, and the National Death Index were used to measure outcomes. RESULTS: A total of 1207 patients were included in the analysis. There were few differences in resource use, outcomes, and processes of care by provider type and experience with HIV-infected inpatients. Patients who received hospitalist care demonstrated a trend toward increased length of hospital stay compared with patients who did not receive hospitalist care (6.0 days vs. 5.2 days; P = .13). Inpatient providers with moderate experience with HIV-infected patients were more likely to coordinate care with outpatient providers (odds ratio, 2.40; P = .05) than were those with the least experience with HIV-infected patients, but this pattern did not extend to providers with the highest level of experience. CONCLUSION: Provider type and attending physician experience with HIV-infected inpatients had minimal effect on the quality of care of HIV-infected inpatients. Approaches other than provider experience, such as the use of multidisciplinary inpatient teams, may be better targets for future studies of the outcomes, processes of care, and resource use of HIV-infected inpatients. PMID: 18444829 [PubMed—indexed for MEDLINE]</p>	<p>Discard. Reason: Design (no intervention evaluated)</p>

<p>Beyond fighting fires and chasing tails? Chronic illness care plans in Ontario, Canada. <i>Ann Fam Med.</i> 2008 Mar-Apr;6(2):146-53. Russell G, Thille P, Hogg W, Lemelin J. CT Lamont Primary Health Care Research Centre, Elisabeth Bruyère Research Institute, University of Ottawa, Ontario, Canada. grussell@scohs.on.ca</p> <p>PURPOSE: Recent work has conceptualized new models for the primary care management of patients with chronic illness. This study investigated the experience of family physicians and patients with a chronic illness management initiative that involved the joint formulation of comprehensive individual patient care plans. METHODS: A qualitative evaluation, framed by phenomenology, immediately followed a randomized controlled trial examining the effect of external facilitators in enhancing the delivery of chronic condition care planning in primary care. The study, set in Ontario family practices, used semistructured in-depth interviews with a purposive sample of 13 family physicians, 20 patients, and all 3 study facilitators. Analysis used independent transcript review and constant comparative methods. RESULTS: Despite the intervention being grounded in patient-centered principles, family physicians generally viewed chronic illness management from a predominantly biomedical perspective. Only a few enthusiasts viewed systematic care planning as a new approach to managing patients with chronic illness. Most family physicians found the strategy to be difficult to implement within existing organizational and financial constraints. For these participants, care planning conflicted with preexisting concepts of their role and of their patient's abilities to become partners in care. The few patients who noticed the process spoke favorably about their experience. CONCLUSIONS: Although the experiences of the enthusiastic family physicians were encouraging, we found important individual-level barriers to chronic illness management in primary care. These issues seemed to transcend existing organizational and resource constraints. PMCID: PMC2267432 PMID: 18332407 [PubMed—indexed for MEDLINE]</p>	<p>Discard. Reason: Outcome (provider satisfaction / opinion)</p>
<p>Quality assessment tools: ACOG Voluntary Review of Quality of Care Program, Peer Review Reporting System. <i>Obstet Gynecol Clin North Am.</i> 2008 Mar;35(1):147-62, x. Lichtmacher A. Women's Health, Lovelace Health System, Albuquerque, NM 87109, USA. abraham.lichtmacher@lovelace.com</p> <p>The Voluntary Review of Quality of Care Program is the American College of Obstetricians and Gynecologists' program of peer review. It is dedicated to quality improvement, patient safety, and peer review, which is conducted by a national professional specialty organization. Since the program's inception, 236 hospitals providing obstetric and gynecologic services in this country have been reviewed. This article presents the results of these reviews. Common problems are identified and possible corrective action is recommended. This program represents a useful model of national peer review activity that can impact patient care. PMID: 18319133 [PubMed—indexed for MEDLINE]</p>	<p>Obtain</p>
<p>Group intervention to improve quality of life in haemodialysis patients. <i>J Clin Nurs.</i> 2007 Nov;16(11C):268-75. Lii YC, Tsay SL, Wang TJ. Mackey College of Nursing, Taipei, Taiwan.</p> <p>AIM AND OBJECTIVE: The purpose of this study was to investigate the effects of group intervention on depression, self-efficacy and quality of life in haemodialysis patients. BACKGROUND: Chronic renal failure and haemodialysis treatment is a long-term process; patients need to have an appropriate adaptive strategy to confront the stress stemming from the disease and subsequent haemodialysis treatment. The application of group intervention for haemodialysis patients has been limited. DESIGN: This study applied an experimental design methodology. Patients were selected at two haemodialysis units of major medical centres in northern Taiwan. METHODS: Out of 60 original patients randomly assigned into experimental or control groups, 48 completed the study. Twenty patients in the experimental group received group psychosocial intervention. The therapy ran for two hours per week for two months. Twenty-eight patients in the comparison group received routine unit care and a self-care booklet. Instruments included the Strategies Used by People to Promote Health, the Beck Depression Inventory and the Short Form-36. Data were collected at pretest and one month following the therapy. RESULTS: The findings demonstrated that self-care self-efficacy, depression and quality of life significantly improved statistically for patients in the therapy group, compared with patients in the comparison group. CONCLUSIONS: The study indicated that group psychosocial intervention significantly reduced depression, improved self-care self-efficacy and quality of life in haemodialysis patients. RELEVANCE TO CLINICAL PRACTICE: The present study was conducted with a group of outpatients, did not require expensive resources and was not time intensive, making it a viable therapy, clinically suitable for haemodialysis patients. PMID: 17931320 [PubMed—indexed for MEDLINE]</p>	<p>Obtain</p>

<p>Nutritional status among patients with hip fracture in relation to pressure ulcers. Clin Nutr. 2007 Oct;26(5):589-96. Epub 2007 Jul 26. Hommel A, Bjorkelund KB, Thorngren KG, Ulander K. Department of Health Sciences, Lund University, P.O. Box 157, SE-221 00 Lund, Sweden. ami.hommel@med.lu.se</p> <p>BACKGROUND &#38; AIMS: Patients with a hip fracture often have a poor nutritional status that is associated with increased risk of complications, morbidity and mortality. The aim of this study was to investigate the effects of an improved care intervention in relation to nutritional status and pressure ulcers. An intervention of best practices for patients with hip fracture was introduced, using the available resources effectively and efficiently with a not too complicated or expensive intervention. METHODS: A quasi-experimental study of 478 patients consecutively included between April 1, 2003 and March 31, 2004. A new evidence-based clinical pathway was introduced on October 1, 2003. The results from the first 210 patients in the control group and the last 210 patients in the intervention group are presented in this article. RESULTS: The total number of patients with a hospital-acquired pressure ulcer was in the intervention group, 19 patients, and in the control group, 39 patients (p = 0.007). No patient younger than 65 years developed a pressure ulcer. There were no statistical significant differences between the groups with respect to blood biochemical variables at inclusion. Patients in the control group had higher arm muscle circumference (AMC) (p = 0.05), calf circumference (CC) (p = 0.038) and body mass index (BMI) (p = 0.043) values. Abnormal anthropometrical tests of BMI, triceps skin fold (TSF) <10th percentile and AMC <10th percentile were found in 12 patients in the control group and in 4 patients in the intervention group. None of the 4 patients in the intervention group developed pressure ulcers. However, 2 of the 12 patients in the control group were affected. CONCLUSIONS: It is possible to reduce the development of hospital-acquired pressure ulcers among elderly patients with a hip fracture even though they have poor prefracture nutritional status. Results in this study indicate the value of the new clinical pathway, as number of patients who have developed pressure ulcers during their stay in hospital has been reduced by 50%. PMID: 17662510 [PubMed—indexed for MEDLINE]</p>	<p>Obtain</p>
<p>The New Jersey radiographic quality assurance program at 5 years. J Am Coll Radiol. 2007 Oct;4(10):691-8. Timins J, Orlando P, Lipoti J. Christ Hospital, Jersey City, New Jersey, USA.</p> <p>PURPOSE: Five years ago, the New Jersey Bureau of Radiological Health decided to develop and implement a proactive, outcome-based quality assurance and quality control program. The specific goals were reducing unnecessary patient radiation exposure and improving image quality (IQ). METHODS: From inception, input was obtained from the regulated community, which included all medical facilities providing diagnostic radiographic services. The program monitored entrance skin exposure (ESE) and quantitative IQ for 3 radiographic examinations: posteroanterior chest, anteroposterior lumbar spine, and anteroposterior foot. RESULTS: Within 1 year, the mean ESE for postanterior chest x-rays had dropped by 39%, the ESE for anteroposterior lumbar spine films had decreased by 18%, and the ESE for anteroposterior foot x-rays had declined by 48%. After the second year, the decreases in ESE were 44%, 29%, and 61%, respectively. There was a corresponding improvement in quantitative IQ. CONCLUSION: Reductions in mean patient radiation exposure can be accomplished for specific radiographic examinations while achieving significant improvement in quantitative IQ. This requires the education of and cooperation with the regulated community. Details of the program and the response of the regulated community are presented. PMID: 17903753 [PubMed—indexed for MEDLINE]</p>	<p>Discard. Reason: Setting (not health care delivery organization)</p>
<p>Caring for patients dying at home from heart failure: a new way of working. Int J Palliat Nurs. 2007 Jun;13(6):266-71. Pooler J, Yates A, Ellison S. Hospice Good Shepherd, Backford, Chester, UK. jackie@hospiceotgs.com</p> <p>This article explores the difficult journey that heart failure patients frequently experience when trying to access palliative care. It describes how a team of Macmillan and heart failure nurse specialists attempted to address the problem using the specialist role to effect change. Individual and group learning needs were identified and addressed while the use of reflective practice and group working helped the nurses to manage and implement change. This project, with management support, empowered the specialist teams to think creatively about nursing practice and improve patient care. It has encouraged working with clinical nurse specialists from other disciplines, thus avoiding a narrowness of outlook. Although this project initially focused on a small number of patients, it has enabled the teams to become established in partnership working; the collaborative approach to providing palliative care for end-stage cardiac failure patients has since continued to grow and flourish. It is hoped that, in the future, further studies can take place to gain more detailed information from patients and their families about how partnership working can continue to meet the needs of this group. PMID: 17851382 [PubMed—indexed for MEDLINE]</p>	<p>Obtain</p>

<p>Asthma treatment protocols in the emergency department: are they effective? J Asthma. 2007 May;44(4):243-8. Self TH, Usery JB, Howard-Thompson AM, Sands C. University of Tennessee Health Science Center, Memphis, TN, USA. tself@utmem.edu</p> <p>Management of asthma in emergency departments (ED) has been well documented to be deficient over many years, despite national and international guidelines. This review summarizes the effect of ED protocols aimed at improving the assessment and treatment of asthma in the ED. We performed a PubMed search of the English literature for ED asthma protocols published from 1986 to 2006 and identified 11 studies. Protocols were effective in improving at least some areas of management, including use of appropriate patient assessment, drug therapy per national guidelines, and patient education. A small number of protocols with the specific aims of reducing the length of stay in the ED as well as rates of hospital admission and return visits were effective. Persistent education of ED staff regarding protocols based on current management guidelines and adoption of easy-to-use forms can facilitate improved care of patients with asthma in the ED. PMID: 17530519 [PubMed—indexed for MEDLINE]</p>	<p>Discard. Reason: design (review)</p>
<p>Peer mentoring in long-term care: rationale, design, and retention. Gerontol Geriatr Educ. 2007;28(2):77-90. Hegeman C, Hoskinson D, Munro H, Maiden P, Pillemer K. Foundation for Longterm Care, 150 State Street, Suite 391, Albany, NY 12207, USA. chegeman@nyahsa.org</p> <p>This article describes two successful peer-mentoring programs designed to improve the quality of care in nursing homes and retention rates among direct care staff. The first program, "Growing Strong Roots," examined CNA retention rates and the second program, "Peer Mentoring for Long Term Charge Nurses," examined RN and LPN retention rates. These programs include a training component, an on-the-job mentoring component, and reinforcing booster sessions. This article includes content and operational issues, recommendations for further program development and research, and replication information. PMID: 18032192 [PubMed—indexed for MEDLINE]</p>	<p>Discard. Reason: outcome (no health outcome)</p>
<p>Quality of consultation and the project 'Support and Consultation on Euthanasia in the Netherlands' (SCEN). Health Policy. 2007 Jan;80(1):97-106. Epub 2006 Apr 18. Jansen-van der Weide MC, Onwuteaka-Philipsen BD, van der Wal G. Department of Public and Occupational Health and EMGO Institute, Vrije Universiteit Medical Centre, Van der Boechorststraat 7, 1081 BT, Amsterdam, The Netherlands.</p> <p>OBJECTIVE: Consultation of another physician is one of the requirements for prudent practice. The project 'Support and Consultation on Euthanasia in the Netherlands' (SCEN) is aimed at professionalizing consultation. The objective of this study is to assess whether the quality of consultation was improved through SCEN. METHOD: In four districts all general practitioners (GPs) received a pre-test questionnaire approximately six weeks before the start of the project in the period (n=1224, response 71%). In the period from April 2000 to December 2002, all GPs in districts in which SCEN had been implemented received a written post-test questionnaire one and a half years after the start of the project. This post-test questionnaire was returned by 60% of the GPs (n=3614). RESULTS: In SCEN consultations the attending physicians has no specific relation to the attending physician in 85% of consultations, while this is the case for 31% of other consultations. While before the start of SCEN in 71% of consultations six or seven of the seven criteria for good consultation were met, in SCEN consultations 83% of cases six or seven of these requirements were met. GPS who had consulted a SCEN physician generally were more positive about different aspects than those who consulted another consultant, such as considering the consultant to be able to make an independent judgment (totally agree 74% vs. 59%). CONCLUSION: Although the quality of consultation appears to be high for both SCEN physicians and other consultants, the SCEN project further contributed to the quality of consultation. Since GPs attach importance to judgment of SCEN physician and have the intention to use it in future, and the quality of consultation stays high over time, this project is expected to maintain its value. PMID: 16621122 [PubMed—indexed for MEDLINE]</p>	<p>Obtain</p>

<p>Civilian doctors in military clinics--outsourcing for better medicine. <i>Mil Med.</i> 2007 Jan;172(1):75-8. Dankner R, Rieck J, Bentacur AG, Bar Dayan Y, Shahar A. The Gertner Institute for Epidemiology and Health Policy Research, Tel-Hashomer, 52621, Israel.</p> <p>OBJECTIVES: To determine whether outsourcing of medical consulting services could improve the quality of medical treatment in military primary care clinics. METHODS: Data were collected prospectively over 2 months in two regular army clinics manned by ordinary army doctors and in two intervention clinics also staffed with senior civilian doctors. The causes for doctor visits, diagnoses, and other patient data were collected. RESULTS: Information was recorded from 4970 soldier visits in the four clinics. Although a prescription of rest days was similar in both types of clinics, the level of tertiary referrals was lower by one-third in the intervention clinics compared with the regular clinics. Surrogate markers for quality of care, such as increased use of planned follow-up and reduced antibiotic use, were significantly better in the intervention clinics, and so was overall patient satisfaction. CONCLUSIONS: Integration of specialist civilian physicians in the military primary care system is highly beneficial and provides better care and saves costs. PMID: 17274271 [PubMed—indexed for MEDLINE]</p>	Obtain
<p>Developing and implementing quality initiatives in the ICU: strategies and outcomes. <i>Crit Care Nurs Clin North Am.</i> 2006 Dec;18(4):469-79, ix. Harrigan S, Hurst D, Lee C, Christie V, Wolfe RB, Morrical D, Kleinpell R. Ball Memorial Hospital, Cardinal Health System, Muncie, IN, USA. SHarrigan@chsmail.org</p> <p>The Institute of Medicine identifies quality health care as care that is safe, time, effective, efficient, equitable, and patient centered. In the ICU, where patients who have complex, high-acuity are at increased risk of complications, morbidity, and mortality, promoting quality-focused care is especially important. This article describes several performance-improvement initiatives that were developed and implemented a Midwestern community hospital during a 4-year period. The initiatives to reduce catheter-associated blood stream infections, to identify early sepsis, and to promote evidence-based care. PMID: 17118301 [PubMed—indexed for MEDLINE]</p>	Obtain
<p>Preventing low birth weight in Illinois: outcomes of the family case management program. <i>Matern Child Health J.</i> 2006 Nov;10(6):481-8. Silva R, Thomas M, Caetano R, Aragaki C. The University of Texas Health Science Center at Houston School of Public Health, Houston, Texas, USA.</p> <p>OBJECTIVES: In the mid 1980's the federal government passed legislation allowing states to expand their Medicaid programs for pregnant women. States were also offered matching funds for "enhanced" prenatal care services. The Illinois Family Case Management (FCM) Program targets low-income women and aims to reduce barriers to prenatal care and infant health care utilization and also provides health education. We evaluated the outcome of the Illinois Family Case Management Program (FCM) in preventing low birth weight in Winnebago County. METHODS: A total of 6,440 participants were included in this study. Logistic regression was used to test whether number of visits or total hours of visitation were significant protective factors against low birth weight. RESULTS: While participating in the FCM Program resulted in a lower rate of low birth weight delivery, neither increasing time with a family case manager nor increasing number of visits showed statistically significant additional protection against low birth weight delivery after adjustment for potential confounding factors. CONCLUSION: In order to further improve program outcomes, efforts need to include improving quality of interventions or developing new interventions rather than simply increasing the amount of current intervention for each participant. The cost effectiveness of shifting FCM Program efforts away from infants (aged 0-1 year) towards improved prenatal interventions should be evaluated. PMID: 16865536 [PubMed—indexed for MEDLINE]</p>	Obtain

<p>Day care laparoscopic cholecystectomy: a feasibility study in a public health service hospital in a developing country. World J Surg. 2006 Sep;30(9):1690-5; discussion 1696-7. Chauhan A, Mehrotra M, Bhatia PK, Baj B, Gupta AK. Department of Surgery, Base Hospital, Delhi Cantt, Delhi 110010, India. bolubonkey@rediffmail.com</p> <p>BACKGROUND: Day care laparoscopic cholecystectomy (DCLC) has been shown to be safe in centers with adequate infrastructure for day care surgery in economically advanced countries. However, the feasibility of applying this concept in public health service hospitals in less developed and developing nation needs to be studied. Unique protocols need to be developed and tested, taking into account local conditions and infrastructural constraints. PATIENTS AND METHODS: Patients less than 60 years old, graded I and II on the American Society of Anesthesiologists (ASA) physical status score, living within one hour traveling time and willing to make their own arrangements for a return to hospital in case of problems, were selected for DCLC. RESULTS: 291 cases (78%) out of 373 laparoscopic cholecystectomies done in one calendar year were found suitable for DCLC. The most common cause for omitting from DCLC was that the patient lived out of the defined area (57%). Four of 291 (1.3%) cases were cancelled due to medical condition; 270/287 (96.1%) were discharged the same evening as surgery; 6 patients were converted to open surgery; and 11 did not meet the necessary discharge criteria. Eight of 270 (2.9%) required readmission out of which 3 (1.1%) required intervention. Overall, incidence of complication rate was 3.4%. Analysis of data showed that results were comparable to previously published studies, hence extrapolating that inclusion and discharge criteria used in the study are valid. However, there are certain social constraints which hinder truly universal application of DCLC. CONCLUSIONS: DCLC is a safe and technically feasible concept, even in public health service centers without dedicated ambulatory surgery units. It has potential for much economical and social benefit in these countries. PMID: 16902738 [PubMed—indexed for MEDLINE]</p>	<p>Obtain</p>
<p>The New Zealand mobile surgical bus service: what is it achieving? N Z Med J. 2006 Jun 23;119(1236):U2025. Bax K, Shedd S, Frizelle F. Department of Surgery, Christchurch Hospital, Christchurch. Comment in N Z Med J. 2006;119(1237):U2074; discussion U2074.</p> <p>AIM: Equitable access and provision of health care is a cornerstone of New Zealand Government health planning. Recent closures of rural hospitals have led to difficulties with access to surgical services. The mobile surgical service has been developed to help; partly to address this issue as well as to address several other stated goals in the provision of rural health. This study aims to audit the goals set out for the mobile surgical service and determine if they have been achieved. METHOD: The following outcome measures were assessed: number and type of procedure, length of stay, complications, services for Maori, upskilling for rural staff, social benefits, impact on child health, improved training with telepresence surgery, and the cost. RESULTS: Over the first 2 years (1 March 2002 to 28 February 2004) of service provision, 1901 procedures were undertaken; 57 patients had complications. The most common complication was wound infection, which occurred in 5% of operations. One in 3 treated patients were Maori and 40% of those treated were 15 years of age or younger. The mobile surgical bus service also appears to be meeting its social benefit, upskilling goals, and educational goals. CONCLUSIONS: The provision of specialist services to the rural communities is a difficult problem faced not only in New Zealand. Though still on a trial basis, the mobile surgical bus service appears to be meeting its stated goals to be addressing one of the important goals of the Government health policy: equitable access and provision to surgical care. PMID: 16807569 [PubMed—indexed for MEDLINE]</p>	<p>Obtain</p>

<p>The value of monitoring frozen section-permanent section correlation data over time. Arch Pathol Lab Med. 2006 Mar;130(3):337-42. Raab SS, Tworek JA, Souers R, Zarbo RJ. Department of Pathology, University of Pittsburgh, Pittsburgh, PA 15232, USA. raabss@upmc.edu</p> <p>CONTEXT: The effectiveness of the long-term monitoring of errors detected by frozen section-permanent section correlation is unknown. OBJECTIVE: To determine factors important in laboratory improvement in frozen section-permanent section discordant and deferral rates by participation in a multi-institutional continuous quality improvement program. DESIGN: Participants in the College of American Pathologists Q-Tracks program self-reported the number of anatomic pathology frozen-permanent section discordant and deferred cases in their laboratories by prospectively performing secondary review of intraoperative consultations. Laboratories participated in the program for 1 to 5 years and reported their data every quarter. We calculated mean and median discordant and deferred case frequencies and used mixed linear modeling to determine if length of participation in the program was associated with improved performance. PARTICIPANTS: One hundred seventy-four laboratories self-reported data. MAIN OUTCOME MEASURES: Mean frozen-permanent section discordant and deferred diagnostic frequencies and changes in these frequencies over time were measured. RESULTS: The mean and median frozen-permanent section discordant frequencies were 1.36% and 0.70%, respectively. The mean and median deferred diagnostic frequencies were 2.35% and 1.20%, respectively. Longer participation in the Q-Tracks program was significantly associated (P = .04) with lower discordant frequencies; 4- or 5-year participation showed a decrease in discordant frequency of 0.99%, whereas 1-year participation showed a decrease in discordant frequency of 0.84%. Longer participation in the Q-Tracks monitor was associated with lower microscopic sampling frequencies for discordant diagnoses (P = .04). Increased length of participation in the Q-Tracks program was significantly associated (P = .04) with lower deferred diagnostic frequencies. CONCLUSIONS: Long-term monitoring of frozen-permanent section correlation is associated with sustained improvement in performance. PMID: 16519561 [PubMed—indexed for MEDLINE]</p>	<p>Discard. Reason: Setting (not health care delivery organization)</p>
<p>Field testing of the ACCESS materials: a portfolio of materials to assist health workers to identify children with disabilities and offer simple advice to mothers. Int J Rehabil Res. 2005 Dec;28(4):293-302. Wirz S, Edwards K, Flower J, Yousafzai A. Centre for International Child Health, Institute Child Health, University College London, UK. s.wirz@ich.ucl.ac.uk</p> <p>Children in less-developed countries with mild to moderate disabilities often remain unidentified until school age. Delayed identification leads to less successful interventions for most children and risks secondary disabilities. The disability group at the Centre for International Child Health was funded to address this issue by developing a screening portfolio. The field testing of this portfolio is reported here. The results collected through quantitative analysis of the children brought for screening, and the fact that the field workers identified disabilities in children over the age of 2 years with over 82% accuracy when compared with professionals, suggests that health workers can be taught to use the portfolio effectively. Among younger children the accuracy is poorer. Field workers can additionally be trained to give advice to mothers and/or refer where appropriate. Equally important, results from focus group discussions with both health workers and parents, and questionnaires to health workers, demonstrate that both groups found the process clear and useful. Parents liked the process and found the advice materials helpful. Several health workers made comments about how the portfolio's use helped to develop positive attitudes towards disability and improved their own self-confidence. PMID: 16319554 [PubMed—indexed for MEDLINE]</p>	<p>Obtain</p>
<p>Comparisons of activity measurements with radionuclide calibrators-A tool for quality assessment and improvement in nuclear medicine. Appl Radiat Isot. 2005 Oct;63(4):493-503. Oropesa P, Hernández AT, Serra R, Varela C. Centro de Isótopos (CENTIS), P.O. Box 3415, San José de las Lajas, Habana, Cuba. poropesa@centis.edu.cu</p> <p>A national program of ongoing comparisons for assaying gamma-emitting radiopharmaceuticals for amount of radioactivity using radionuclide calibrators was begun in 2000. Nuclides of the most wide-spread use in Cuban nuclear medicine, ¹³¹I, ²⁰¹Tl and (^{99m}Tc), as well as two measurement geometries, glass vials and plastic syringes, were employed. In this paper, the participants' performance is assessed by mean of a statistical analysis of the reported data. Performance tables have been obtained and a chi² statistic is calculated from observed and expected frequencies, with the aim of testing the hypothesis about the independence of some characteristics of the comparison results, at a significance level alpha=0.05. The proportion of satisfactory results in the years 2002-2004 were found to be at the same level, but higher than in 2000. It reveals an improvement of the measurement quality since 2002. The causes of improvement were investigated using the statistical treatment of several data available as supplementary information. PMID: 16039865 [PubMed—indexed for MEDLINE]</p>	<p>Discard. Reason: Setting (not health care delivery organization)</p>

<p>Improving oral care in patients receiving mechanical ventilation. Am J Crit Care. 2005 Sep;14(5):389-94. Cutler CJ, Davis N. Clinical Excellence Department, Advocate Health Care, Oak Brook, IL, USA.</p> <p>BACKGROUND: Comprehensive oral care is an evidence-based prevention strategy to reduce the risk of ventilator-associated pneumonia in patients receiving mechanical ventilation. Until recently, no comprehensive guidelines or standards existed to define necessary tasks, methods, and frequency of oral care to provide patients with optimal results. OBJECTIVES: To observe current practice of, define best practice for, and measure compliance with standardized comprehensive oral care. METHODS: This observational study was part of a larger research study performed at 5 acute care hospitals. Time blocks of 4 hours were randomized over 8 intensive care units and the 7 days of the week. Baseline data were collected before implementation of multifaceted education on an oral-cleansing protocol; interventional data were collected afterward. RESULTS: Oral care practices were observed for 253 patients. During the baseline period, oral cleansing was primarily via suction swabs. Toothbrushing and moisturizing of the oral tissues were not observed. Only 32% of the patients had suctioning to manage oral secretions. During the interventional period, 33% of patients had their teeth brushed, 65% had swab cleansing, and 63% had a moisturizer applied to the oral mucosal tissues. A total of 61% had management of oral secretions; 38% had oropharyngeal suctioning via a special catheter. CONCLUSIONS: Implementation of an evidence-based oral cleansing protocol improved the care of patients receiving mechanical ventilation. Multifaceted education and implementation strategies motivated staff to increase oral care practices. PMID: 16120890 [PubMed—indexed for MEDLINE]</p>	<p>Obtain</p>
--	---------------

<p>The development of quality assurance programs for radiotherapy within the German Hodgkin Study Group (GHSg). Introduction, continuing work, and results of the radiotherapy reference panel. <i>Strahlenther Onkol.</i> 2005 Sep;181(9):557-66. Müller RP, Eich HT. Department of Radiation Oncology, University of Cologne, Joseph-Stelzmann-Strasse 9, 50924 Köln, Germany. rolf-peter.mueller@medizin.uni-koeln.de</p> <p>BACKGROUND AND PURPOSE: The German Hodgkin Study Group (GHSg), including more than 500 participating centers, established a central radiotherapy (RT) reference center to improve quality of treatment, starting with the first study generation in 1978. More than 11,000 patients with Hodgkin's lymphoma (HL) have been enrolled into these trials. Extensive continuing quality assurance programs (QAPs) during the study generations have been performed. The purpose of the present article is to summarize the experiences and results of the performed and ongoing QAPs. MATERIAL AND METHODS: A panel of expert radiation oncologists (second study generation HD4-6, 1988-1994, and third study generation HD7-9, 1993-1998) retrospectively evaluated the adequacy of treatment fields, applied radiation doses, treatment time, and technical parameters. Furthermore, a detailed analysis of relapses in correlation with the performed RT was conducted. For the fourth study generation (HD10-12, 1998-2002), the RT reference center changed from Munich to Cologne. New RT QAPs were initiated according to the demands of the new trials and former programs were enhanced: (1) central prospective radiation oncologic review of cross-sectional imaging (HD10, HD11) to create the individual radiation treatment plan; (2) retrospective analysis of the adequacy of the performed involved-field (IF) RT (HD10, HD11); (3) the multidisciplinary HD12 panel (radiation oncologists, medical oncologists, diagnostic radiologists); (4) initiation and integration of a teleradiotherapy network into the GHSg trials. RESULTS: A strong achievement of these activities in the era of extended-field RT was to show that major deviations of radiation treatment portals and radiation dose from prospective treatment prescriptions revealed to be unfavorable prognostic factors for patients with early-stage HL (HD4). The central prospective radiation oncological review of all diagnostic imaging (HD10, HD11) showed that corrections of disease involvement in 49% of patients (593/1,214) with early stages (HD10) and in 67% of patients (936/1,397) with intermediate stages (HD11) were necessary. These procedures had a significant impact on the correctness of stage definition, allocation to treatment groups and on the extension of the IF treatment volume. Until now, 1,080 patients in HD10 and HD11 have been evaluated retrospectively with regard to the adequacy of the performed IF-RT. Although the participating institutions got a precise RT prescription, interim results reveal deviations in a significant number of cases. In the HD12 trial (advanced stages), a multidisciplinary panel of radiation oncologists, radiologists and medical oncologists reviewed all the diagnostic imaging from diagnosis throughout the treatment in comparison to the documentation forms. For patients with poor response to chemotherapy, the panel recommended RT independent of the randomization. This procedure ensured that patients with a poor response to chemotherapy received additional RT. 1,080 of 1,594 randomized patients (68%) could be analyzed. After chemotherapy, 599 patients (56%) showed residual disease (> 1.5 cm), and in 145/1,080 patients (13.5%) the panel recommended additional RT independent of the randomization arm. The introduction of electronic image transfer optimized and simplified the workflow of the QAPs. Rapid online consultation and real-time teleconferences regarding disease involvement, patient management and communication of the RT prescription with connected hospitals proved to be extremely helpful. CONCLUSION: Today, radiation oncologists in the GHSg perform a continuous and efficient QAP to improve treatment quality of study patients. For early favorable and unfavorable HL a central prospective review of all diagnostic imaging is performed by expert radiation oncologists to control the disease extension and to define the IF treatment volume. Retrospective analysis of RT portals by an expert panel detects faults in the applied irradiation. Participants are trained on the definition of IF-RT by workshops on the occasion of annual GHSg meetings and on the annual meetings of the German Society of Therapeutic Radiation Oncology (DEGRO). For the advanced stages a multidisciplinary panel evaluates the treatment response to chemotherapy. Patients with a poor response receive additional RT due to the panel's recommendation. The introduction of teleradiotherapy into the GHSg trials improves the dialogue between the central RT reference center and study participants and thus contributes to high RT quality for study patients. PMID: 16170482 [PubMed—indexed for MEDLINE]</p>	<p>Obtain</p>
---	---------------

<p>A quality assurance program in dental radiographic units in western Greece. Oral Surg Oral Med Oral Pathol Oral Radiol Endod. 2005 May;99(5):622-7. Stavrianou K, Pappous G, Pallikarakis N. University of Patras School of Medicine, Patras, Greece.</p> <p>OBJECTIVE: This survey reports on successful efforts to establish a quality assurance (QA) program for 50 intraoral x-ray units in the public and private sectors in the region of Achaia, Greece. It was conducted in 2 phases, in 1996 and in 2003, including on-site inspections, QA tests, and standard questionnaires. The aim of the study was to assess equipment conditions, knowledge, and adoption of radiographic QA guidelines by general dentists and, more importantly, the impact of the recommendations and training provided. STUDY DESIGN: The tested parameters focused on radiation protection, equipment maintenance, film speed used, film processing conditions, and radiologic characteristics such as voltage, radiation leakage, type of collimation, source-to-skin distance, timer accuracy, and entrance dose. RESULTS: The data gathered in 1996 demonstrate minimal compliance with equipment function requirements and radiation safety measures. 1 The comparative evaluation of all the parameters gathered from the 2 surveys, indicated that in 2003 the vast majority of the dentists followed the recommendations given in 1996. Only 2 dentists persisted in neglecting the guidelines. PMID: 15829888 [PubMed—indexed for MEDLINE]</p>	<p>Discard. Reason: Design (no intervention, survey / correlational study)</p>
<p>Improving surgical wound classification--why it matters. AORN J. 2004 Aug;80(2):208-9, 212-23. Devaney L, Rowell KS. National Surgical Quality Improvement Program, Center for Clinical Effectiveness in Surgery, Massachusetts General Hospital, Boston, USA.</p> <p>Surgical wound classification is an important predictor of the risk of postoperative surgical site infections. Wound classification also is used to analyze clinical, economic, and educational outcomes in national reports on quality. As integral members of the health care team, nurses and physicians need to ensure that their data are correct, consistent, and reliable. This article delineates how one institution developed a multifaceted education program that resulted in a 26% improvement in the rate of correctly classified wounds. The education program provided regular feedback of results that helped identify opportunities for improvement on a widespread level. PMID: 15382594 [PubMed—indexed for MEDLINE]</p>	<p>Obtain</p>
<p>Public-private mix for DOTS implementation: what makes it work? Bull World Health Organ. 2004 Aug;82(8):580-6. Epub 2004 Sep 13. Lönnroth K, Uplekar M, Arora VK, Juvekar S, Lan NT, Mwaniki D, Pathania V. Stop TB Department, World Health Organization, Geneva, Switzerland. lonnrothk@who.int</p> <p>OBJECTIVE: To compare processes and outcomes of four public-private mix (PPM) projects on DOTS implementation for tuberculosis (TB) control in New Delhi, India; Ho Chi Minh City, Viet Nam; Nairobi, Kenya; and Pune, India. METHODS: Cross-project analysis of secondary data from separate project evaluations was used. Differences among PPM project sites in impact on TB control (change in case detection, treatment outcomes and equity in access) were correlated with differences in chosen intervention strategies and structural conditions. FINDINGS: The analysis suggests that an effective intervention package should include the following provider-side components: (1) orienting private providers (PPs) and the staff of the national TB programme (NTP); (2) improving the referral and information system through simple practical tools; (3) the NTP adequately supervising and monitoring PPs; and (4) the NTP providing free anti-TB drugs to patients treated in the private sector. CONCLUSION: Getting such an intervention package to work requires that the NTP be strongly committed to supporting, supervising and evaluating PPM projects. Further, using a local nongovernmental organization or a medical association as an intermediary may facilitate collaboration. Investing time and effort to ensure that sufficient dialogue takes place among all stakeholders is important to help build trust and achieve a high level of agreement. PMCID: PMC2622937 PMID: 15375447 [PubMed—indexed for MEDLINE]</p>	<p>Discard. Reason: Design (correlational study, not intervention evaluation)</p>
<p>Robotically assisted left ventricular epicardial lead implantation for biventricular pacing: the posterior approach. Ann Thorac Surg. 2004 Apr;77(4):1472-4. Derosé JJ Jr, Belsley S, Swistel DG, Shaw R, Ashton RC Jr. Division of Cardiothoracic Surgery, St. Luke's-Roosevelt Hospital Center and Columbia University College of Physicians and Surgeons, New York, New York, USA. jjd11@columbia.edu</p> <p>Patients with congestive heart failure and altered interventricular conduction enjoy improvements in quality of life and ventricular function after successful resynchronization therapy with biventricular pacing. Technical limitations owing to individual coronary sinus and coronary venous anatomy result in a 10% to 15% failure rate of left ventricular (LV) lead placement through percutaneous approaches. To provide a minimally invasive option for these patients with LV lead failures, we developed a technique of endoscopic, epicardial LV lead implantation with the use of the da Vinci robotic system. The surgical approach targets the posterolateral wall through a novel posterior approach. PMID: 15063301 [PubMed—indexed for MEDLINE]</p>	<p>Discard. Reason: Intervention (different surgical technique)</p>

<p>Benefits of multidisciplinary case conferencing using audiovisual compared with telephone communication: a randomized controlled trial. J Telemed Telecare. 2004;10(6):351-4. Wilson SF, Marks R, Collins N, Warner B, Frick L. Ambulatory Care Continuum, Macarthur Health Service, Campbelltown, New South Wales, Australia. stephen.wilson@swhs.nsw.gov.au</p> <p>Multidisciplinary case conferencing using a video-link was compared with multidisciplinary case conferencing by telephone. One hundred patients were randomized to either videoconferencing (intervention group, 50 patients) or audioconferencing (control group, 50 patients). The effectiveness of the intervention compared with the control was evaluated in terms of: the number of conferences per patient, average length of conference, length of treatment, number of occasions of service, degree of multidisciplinary team involvement, recorded level of communication, quality of the management plan (in terms of the number of points contained in it) and staff satisfaction. The intervention and control groups showed significant differences on only two of the outcome measures: the mean number of case conferences per patient was less for the intervention group, and the intervention group had a shorter length of treatment (6 days) than the control group (10 days). The study did not demonstrate any significant differences in occasions of service or time commitment, which might have resulted in lower costs. However, the introduction of case conferencing by video-link was accompanied by a high level of satisfaction on the part of the 14 team members who were interviewed. PMID: 15603634 [PubMed—indexed for MEDLINE]</p>	Obtain
<p>Promoting evidence-based practice: an internship for staff nurses. Worldviews Evid Based Nurs. 2004;1(4):215-23. Cullen L, Titler MG. Department of Nursing Services and Patient Care, University of Iowa Hospitals and Clinics, Iowa City, IA 52242-1009, USA. laura-cullen@uiowa.edu</p> <p>Implementing evidence-based practice is a complex but valued process that requires support for nurses to make it a reality in care delivery. To address this, an Evidence-Based Practice Staff Nurse Internship was developed at the University of Iowa Hospitals and Clinics in the United States. PROGRAM OVERVIEW: The objective of this internship is to promote use of evidence by staff nurses to improve patient outcomes. Through a competitive application process, six nurses are accepted in each cohort. The program provides didactic content and dedicated work time on topics interns select (e.g., family pet visitation, sedation management, bowel sounds assessment, guided imagery, family transition to pediatric floor). Interns receive paid clinical release time for participating. ROLES AND RESPONSIBILITIES: Teams include the staff nurse, the nurse manager, and an advanced practice nurse. Responsibilities for each step are assigned to the person with the appropriate expertise. EVALUATION: Interns report understanding the process, appreciate the opportunity for professional growth, and report their objectives are being met. Participants evaluated the program very positively and also provided recommendations for revision (e.g., revising class content). The program resulted in improved quality of care such as increased patient and family satisfaction, decreased length of stay, and cost savings. IMPLICATIONS: Programs that support practitioners through the evidence-based practice process are needed for use in a variety of settings internationally. This unique program supports staff nurses in making evidence-based practice a reality for their work and patients and might be transferable across settings. PMID: 17166151 [PubMed—indexed for MEDLINE]</p>	Obtain

Notes: N=46 citations. For each citation one original reviewer included the study initially, the second reviewer excluded it. Reviewer disagreements are shown for validation set 1, LR-PS. The final decision is based on a team discussion.