



University of Zurich  
Zurich Open Repository and Archive

Winterthurerstr. 190  
CH-8057 Zurich  
<http://www.zora.unizh.ch>

---

*Year: 2002*

---

## Systematic quantitative overviews of the literature to determine the value of diagnostic tests for predicting acute appendicitis: study protocol

Bachmann, L M; Bischof, D B; Bischofberger, S A; Bonani, M G; Osann, F M;  
Steurer, J

Bachmann, L M; Bischof, D B; Bischofberger, S A; Bonani, M G; Osann, F M; Steurer, J. Systematic quantitative overviews of the literature to determine the value of diagnostic tests for predicting acute appendicitis: study protocol. *BMC Surg* 2002, 2:2.

Postprint available at:  
<http://www.zora.unizh.ch>

Posted at the Zurich Open Repository and Archive, University of Zurich.  
<http://www.zora.unizh.ch>

Originally published at:  
*BMC Surg* 2002, 2:2

# Systematic quantitative overviews of the literature to determine the value of diagnostic tests for predicting acute appendicitis: study protocol

## Abstract

**BACKGROUND:** Suspected acute appendicitis is the most frequent cause for emergency operations in visceral surgery worldwide. In approximately twenty percent of all cases however, the diagnosis is incorrect and patients undergo surgery without having acute appendicitis. Operations of bland appendices put patients at risk and entail a serious waste of resources. Several highly accurate tests have been introduced to diagnose acute appendicitis. The false positive rate however, has not changed over the last twenty years. Given the variation that exists in both practice and research, the uncertainty regarding the quality of the underlying evidence, there is a clear need for comprehensive, systematic and quantitative overviews of the diagnostic value of the various tests purported to be predictive of acute appendicitis. **METHODS:** Literature will be identified searching general bibliographic databases (MEDLINE and EMBASE), specialist computer databases (DARE, Cochrane Database of Systematic Reviews, conference proceedings, MEDION, SCISEARCH, BIOSIS) without language restrictions. We will contact experts and the manufacturers of tests. Hand-searching will complete our searches. Identified articles will be selected according to populations, tests, outcomes and study design. Papers meeting the selection criteria will be appraised to rate their methodological quality. Analysis will include exploration of heterogeneity in results. We will conduct meta-analyses to generate summary estimates of test accuracy measures and summary ROC curves where appropriate. If meta-analysis is considered to be inappropriate, we will describe the identified evidence in the context of appraised quality. **DISCUSSION:** These reviews should lead to formulation of recommendations for current practice and future research.

Study protocol

## Systematic quantitative overviews of the literature to determine the value of diagnostic tests for predicting acute appendicitis: study protocol

Lucas M Bachmann\*<sup>1,2</sup>, Dominique B Bischof<sup>1</sup>, Stephan A Bischofberger<sup>1</sup>, Marco G Bonani<sup>1</sup>, Franziska M Osann<sup>1</sup> and Johann Steurer<sup>1</sup>

Address: <sup>1</sup>Horten Centre, University of Zurich, Switzerland and <sup>2</sup>Academic Department of Obstetrics and Gynaecology, University of Birmingham, UK

E-mail: Lucas M Bachmann\* - lucas.bachmann@evimed.ch; Dominique B Bischof - dbischof@gmx.ch; Stephan A Bischofberger - knut@access.unizh.ch; Marco G Bonani - marcobonani@access.unizh.ch; Franziska M Osann - fransl@gmx.ch; Johann Steurer - polstf@usz.unizh.ch

\*Corresponding author

Published: 10 April 2002

Received: 7 March 2002

*BMC Surgery* 2002, 2:2

Accepted: 10 April 2002

This article is available from: <http://www.biomedcentral.com/1471-2482/2/2>

© 2002 Bachmann et al; licensee BioMed Central Ltd. Verbatim copying and redistribution of this article are permitted in any medium for any purpose, provided this notice is preserved along with the article's original URL.

### Abstract

**Background:** Suspected acute appendicitis is the most frequent cause for emergency operations in visceral surgery worldwide. In approximately twenty percent of all cases however, the diagnosis is incorrect and patients undergo surgery without having acute appendicitis. Operations of bland appendices put patients at risk and entail a serious waste of resources. Several highly accurate tests have been introduced to diagnose acute appendicitis. The false positive rate however, has not changed over the last twenty years. Given the variation that exists in both practice and research, the uncertainty regarding the quality of the underlying evidence, there is a clear need for comprehensive, systematic and quantitative overviews of the diagnostic value of the various tests purported to be predictive of acute appendicitis.

**Methods:** Literature will be identified searching general bibliographic databases (MEDLINE and EMBASE), specialist computer databases (DARE, Cochrane Database of Systematic Reviews, conference proceedings, MEDION, SCISEARCH, BIOSIS) without language restrictions. We will contact experts and the manufacturers of tests. Hand-searching will complete our searches. Identified articles will be selected according to populations, tests, outcomes and study design. Papers meeting the selection criteria will be appraised to rate their methodological quality. Analysis will include exploration of heterogeneity in results. We will conduct meta-analyses to generate summary estimates of test accuracy measures and summary ROC curves where appropriate. If meta-analysis is considered to be inappropriate, we will describe the identified evidence in the context of appraised quality.

**Discussion:** These reviews should lead to formulation of recommendations for current practice and future research.

## Introduction

Suspected acute appendicitis is the most frequent cause for emergency operations in visceral surgery worldwide. In the UK 37,289 patients had an emergency excision of the appendix in the year 2000 [1]. In approximately twenty percent of all cases however, the diagnosis is incorrect and patients undergo surgery without having acute appendicitis at all [2–5]. Operations of bland appendices may lead to morbidity in 4.6 percent [6] and to mortality in 0.14 percent [6] of cases. Despite the introduction of reports of highly accurate diagnostic procedures for the diagnosis of acute appendicitis a big retrospective cohort study [7] concluded that the rate of misdiagnosis (the false positive rate) has not changed over the last twenty years. One potential explanation of that finding might be, that studies reporting on test accuracy overestimate the true potential of correct classification due to inappropriate methodology and bias of reported results since primary research on evaluation of tests is generally poor in quality [8–10].

Online searches of the electronic databases revealed a number of broad reviews, commentaries and recommendations on tests for predicting acute appendicitis but there was a dearth of focused, rigorous diagnostic overviews of the available evidence. These publications showed that there are several prediction rules and tests or markers purported to be predictive of acute appendicitis. However, they offer only limited guidance for practice because traditional literature reviews evaluating tests for acute appendicitis have not applied the scientific strategies to assemble, appraise, and synthesize relevant evidence, which have been embodied in the criteria for high quality reviews.

Given the variation that exists in both practice and research, the uncertainty regarding the quality of the underlying evidence, and the importance of early prediction of acute appendicitis in view of the available effective treatments, there is a clear need for a comprehensive, systematic and quantitative overview of the diagnostic value of the various tests purported to be predictive of acute appendicitis.

At present there is a dearth of such reviews and in this commentary, we will describe how we are using such a systematic approach to collate and critically appraise the available literature in the diagnosis of acute appendicitis.

## Methods

### Study identification

Non-comprehensive search strategies can lead to significant bias in the retrieval of relevant literature. This weakens the strength of inferences from systematic reviews and poses a particular problem in reviews of diagnostic tests [11,12]. Therefore we will identify literature via general

bibliographic databases including MEDLINE and EMBASE, specialist computer databases such as DARE and MEDION (a database of diagnostic test reviews set up by Dutch and Belgian researchers), the Cochrane Database of Systematic Reviews, relevant specialist registers of the Cochrane Collaboration, conference proceedings and BIOSIS without language restrictions. In addition we will contact individual experts and those with an interest in this field to uncover grey literature and we will contact the manufacturers of tests. Hand-searching of selected specialist journals, checking of reference lists and SCISEARCH to identify frequently cited articles will complete our searches. In cases of duplicate publication, the most recent and complete versions will be selected. A comprehensive database of relevant articles will be constructed – a preliminary search has been carried out in order to estimate the size of the relevant literature. MEDLINE Searches located 800 potentially relevant citations. Expanding search to other databases, hand searching, reference list searching and or contact with authors might add another 100% citations, so the total is likely to be 1600. Letters will be sent to major centres and the first author of each shortlisted selected paper published in the last five years, asking them whether they know of any published or unpublished relevant studies not included on our list. The search strategy used to identify articles in MEDLINE is shown in: appendix.doc.

### Study selection

Studies will be selected for inclusion in the review in a two-stage process using the selection criteria based on those shown in Table 1. First, a comprehensive database of the literature search will be constructed. The citations will be scrutinised by two reviewers to obtain copies of full manuscripts of all citations that are likely to meet the selection criteria. Two reviewers will then independently select the studies, which meet predefined, and explicit criteria regarding populations, tests, outcomes and study design. These criteria will be pilot tested using a sample of papers and agreement between reviewers will be measured. When disagreements occur the two reviewers will meet. Experience suggests that often the cause of the disagreement is a simple oversight on the part of one of the reviewers. When this is not the case the issue will be resolved by consensus involving a third reviewer.

### Study validation

Papers meeting the selection criteria will be appraised to rate their methodological quality. In addition to using ratings of study quality as possible explanations for differences in results, the extent to which primary research met methodological standards is important per se for assessing the strength of any conclusions that are reached. There is an ongoing debate over what constitutes the best quality assessment tool for diagnostic test studies. We will evalu-

**Table 1: Study Selection Criteria.**


---

**Population:** Patients suspicious to have acute appendicitis

**Diagnostic tests:**

- Prediction rules
- Inflammatory markers (C-reactive protein, leucocytes count)
- Transabdominal ultrasound
- Computer Tomography (CT)
- Magnetic Resonance Imaging (MRI)
- Scintigraphy
- Diagnostic laparoscopy clinical history and physical examination

**Outcome measures:** Test accuracy, morbidity and mortality from misdiagnosis

**Study design:** Diagnostic test studies will be selected. They consist of observational studies (prospective or cross-sectional) of defined non-randomised populations in which the results of the diagnostic test of interest are compared with the results of a reference standard allowing generation of a 2 × 2 table to compute indices of diagnostic accuracy.

---

**Table 2: Criteria for study validation.**


---

**Population:** Consecutive recruitment of an appropriate spectrum of eligible patients will be considered ideal. Convenience sampling, i.e. arbitrary recruitment or non-consecutive recruitment will be deemed inadequate. In the absence of any explicit information on the method of recruitment, the article will be categorised as unclear reported population enrolment.

**Diagnostic test:** The description of the diagnostic test will be considered ideal if the methodology is reported together with the measurement parameter and the cut-off level for an abnormal result. In the absence of any of the above information in the manuscript, then the diagnostic intervention will be considered as unclear reported.

**Outcome measures:** Blinding will be considered ideal if it is clearly reported that the results of the various tests were not divulged. Information on the number of patients recruited into the study and those whose outcome data were known will also be sought from the manuscripts. Withdrawal of patients from the study, missing data and lack of outcome data outwith the study hospital will be categorised as lost to follow-up. In particular we will look for evidence of verification bias where the rates follow-up and confirmation of outcome are different in patients with positive test results compared to those with negative test results.

Any available randomised trials will be assessed for validity separately to the diagnostic accuracy studies considering factors associated with bias in such trials, e.g. concealment of randomisation, sequence generation, blinding and follow-up.

---

ate elements of study design, which are likely to have a direct relationship to bias in a diagnostic test study [10][13][14][15]. The items shown in Table 2 will be used for methodological quality assessment. Agreement for the quality assessments will be calculated, and disagreement resolved, in the same fashion as for the assessment of study selection. We will evaluate the agreement between the two reviewers using percentage agreement and weighted kappa statistics [16].

### Data collection

The extraction of study's findings will be conducted in duplicate using a pre-designed and piloted data extraction form to avoid any errors. Given the extent of insufficient reporting in the medical literature, we propose to obtain missing information from investigators whenever possible. It is otherwise impossible to distinguish between what was done but not reported and what was not done. A template of data extraction form is shown in: appendix.doc.

### Analysis

By analysis we mean synthesis of results from individual studies (meta-analysis), and exploration of variation in results from study to study (heterogeneity) and generation of the most useful combination of tests. We will conduct meta-analyses to generate summary estimates of sensitivities, specificities, predictive values, likelihood ratios (LRs) and receiver operating characteristic (ROC) curves where appropriate [13,14,17]. If meta-analysis is considered to be inappropriate, we will describe the identified evidence in the context of appraised quality. If a meta-analysis is considered appropriate, we will examine the correlation between true positive rates and false positive rates in individual studies. If the correlation is poor, we will use LR as the main accuracy measure. If we find a correlation then we will generate a summary ROC curve [18] in addition to pooling of LRs. Many authorities considered this the preferred method of pooling test results from primary studies [13,14,17]. The summary ROC plot provides a way of summarising the performance of a test from the results of several studies over a range of test thresholds. However, our preference for LRs is based on the published recommendations that LRs are more clinically meaningful as measures of diagnostic accuracy [15]. Our experience has been that the true positive rates and false positive rates in individual studies are poorly correlated in which case it is not feasible to generate a summary ROC curve. Moreover, when the outcome of a test is of binary nature (positive or negative) LRs are more clinically meaningful than ROC curves. One disadvantage of analysis using LR is that it generates two measures for each test, one for a positive result and another for a negative result. A ratio of LRs will be used to generate a single measure called diagnostic odds ratio, which is more suitable for statistical analysis. For the purpose of meta-analysis, we will weight the logLR from each study in inverse proportion to its variance in order to combine the LRs from each study. To demonstrate the practical application of the summary LRs generated, we will calculate posttest probabilities for acute appendicitis using Bayes' theorem. An estimate of the pretest probability will be obtained by calculating the prevalence of the outcome event in the population studied. The following algorithm of equations will be used for calculating posttest probability:

pretest probability = prevalence of acute appendicitis

pretest odds = pretest probability / (1 - pretest probability)

posttest odds = likelihood ratio × pretest odds

posttest probability = posttest odds / (1 + posttest odds)

In order to deal with the uncertainty of the estimate, we will generate 95% confidence intervals around the point estimate. Approximate variance for the posttest odds will be obtained by adding the variances of the combined LRs and pretest odds, enabling the calculation of its 95% confidence intervals. The 95% confidence intervals for the posttest probabilities will then be generated by converting the limits of the posttest odds to their respective probabilities.

Heterogeneity of results between different studies will be formally assessed using the Breslow-Day test which compares for each study the ratio of the odds of having the outcome of interest when the test result is positive to the odds of having the same outcome when the test result is negative [19]. To explore causes of heterogeneity in the estimates of diagnostic accuracy of the tests for acute appendicitis, we will conduct a sensitivity analysis. This will be carried out by subgroup analyses to see whether variations in population, intervention, outcomes and study quality will affect the estimate of diagnostic accuracy. Results of pooled analyses will be provided within cogent patient groups.

## Discussion

In summary, systematic reviews of diagnostic literature to predict acute appendicitis allow us to assess the quality of the available evidence and to identify specific tests (including history, physical examination and tests) that have diagnostic value. These reviews should lead to formulation of recommendations for current practice and future research. Just as an evidence-based culture in delivery of health care has been supported by systematic reviews of literature on therapeutic interventions, we can expect to see an extension of this approach in the area of care involving use of diagnostic and screening tests.

## Competing interests

none declared

## Authors' Contributions

LMB and JS initiated the project and wrote the protocol. DBB, SAB, MGB and FMO screened the pilot searches, all authors commented on earlier drafts and approved the final manuscript.

## Additional material

### Appendix.doc

Click here for file  
[http://www.biomedcentral.com/content/supplementary/1471-2482-2-2-S1.doc]

## Acknowledgements

The authors would like to thank Gill Richie and Julie Glanville of the Centre for Reviews and Dissemination in York (UK) for searching the databases.

## References

1. **Hospital Episode Statistics Department of Health Available at:** 2001 [http://www.doh.gov.uk/hes/standard\_data/available\_tables/total\_operations/tb01099a.pdf]
2. Reynolds SL: **Missed appendicitis in a pediatric emergency department.** *Pediatr Emerg Care* 1993, **9**:1-3
3. Rothrock SG, Skeoch G, Rush JJ, Johnson NE: **Clinical features of misdiagnosed appendicitis in children.** *Ann Emerg Med* 1991, **20**:45-50
4. Rothrock SG, Green SM, Dobson M, Colucciello SA, Simmons CM: **Misdiagnosis of appendicitis in nonpregnant women of child-bearing age.** *J Emerg Med* 1995, **13**:1-8
5. McCallion J, Canning GP, Knight PV, McCallion JS: **Acute appendicitis in the elderly: a 5-year retrospective study.** *Age Ageing* 1987, **16**:256-260
6. Velanovich V, Satava R: **Balancing the normal appendectomy rate with the perforated appendicitis rate: implications for quality assurance.** *Am Surg* 1992, **58**:264-269
7. Flum DR, Morris A, Koepsell T, Dellinger EP: **Has misdiagnosis of appendicitis decreased over time? A population-based analysis.** *JAMA* 2001, **286**:1748-1753
8. Sheps SB, Schechter MT: **The assessment of diagnostic tests. A survey of current medical research.** *JAMA* 1984, **252**:2418-2422
9. Reid MC, Lachs MS, Feinstein AR: **of methodological standards in diagnostic test research. Getting better but still not good.** *JAMA* 1995, **274**:645-651
10. Lijmer JG, Mol BW, Heisterkamp S, Bossel GJ, Prins MH, van de Meulen JH, et al: **Empirical evidence of design-related bias in studies of diagnostic tests.** *JAMA* 1999, **282**:1061-1066
11. Irwig L, Macaskill P, Glasziou P, Fahey M: **Meta-analytic methods for diagnostic test accuracy.** *J Clin Epidemiol* 1995, **48**:119-130
12. Vamvakas EC: **Meta-analyses of studies of the diagnostic accuracy of laboratory tests: a review of the concepts and methods.** *Arch Pathol Lab Med* 1998, **122**:675-686
13. **Cochrane Methods Group on Systematic Review of Screening and Diagnostic Tests: Recommended Methods, last updated on 9 February 1998.** 1996 [http://www.cochrane.org/cochrane/sadtdoc1.htm]
14. Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC, et al: **Guidelines for meta-analyses evaluating diagnostic tests.** *Ann Intern Med* 1994, **120**:667-676
15. Jaeschke R, Guyatt G, Sackett DL: **Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group.** *JAMA* 1994, **271**:389-391
16. Cohen J: **A coefficient of agreement for nominal scales.** *Educ Psychol Meas* 1960, **20**:27-46
17. Midge AS, Stukel TA, Littenberg B: **A meta-analytic method for summarizing diagnostic test performances: receiver-operating-characteristic-summary point estimates.** *Med Decis Making* 1993, **13**:253-257
18. Moses LE, Shapiro D, Littenberg B: **Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations.** *Stat Med* 1993, **12**:1293-1316

19. Breslow NE, Day NE: **Statistical methods in cancer research. Volume I – The analysis of case-control studies.** IARC Sci Publ 1980, 5-338

### Pre-publication history

The pre-publication history for this paper can be accessed here:

<http://www.biomedcentral.com/1471-2482/2/2/prepub>

Publish with **BioMed Central** and every scientist can read your work free of charge

*"BioMedcentral will be the most significant development for disseminating the results of biomedical research in our lifetime."*

Paul Nurse, Director-General, Imperial Cancer Research Fund

Publish with **BMC** and your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours - you keep the copyright

Submit your manuscript here:

<http://www.biomedcentral.com/manuscript/>



**BioMedcentral.com**

[editorial@biomedcentral.com](mailto:editorial@biomedcentral.com)