Linköping University Medical Dissertations
No 1442

Structured management of patients with suspected acute appendicitis

Manne Andersson

Linköping University
FACULTY OF HEALTH SCIENCES

Division of Surgery and Clinical Oncology
Department of Clinical and Experimental Medicine
Faculty of Health Sciences
Linköping University, Sweden

Linköping 2015

Cover illustration: A farmer from Devon, England, 1942, with the divine power of finding water using only his hands and a hazel twig.
With permission, © Imperial War Museum (D 9817)

Printed by LiU-Tryck, Linköping University, Sweden, 2011

ISBN 978-91-7519-137-9

ISSN 0345-0082

To my family

Malin, Einar, Line and Elmer

This study has been carried out with the support of
Futurum - Academy for Health and Care, Jönköping County Council, Sweden and
FORSS - Medical Research Council of Southeast Sweden

CONTENTS

LIST OF PUBLICATIONS

This thesis is based on the following papers, which will be referred to by their Roman numerals as indicated below:

I. Andersson M, Andersson RE.
The Appendicitis inflammatory response score: A tool for the diagnosis of acute appendicitis that outperforms the Alvarado score
World Journal of Surgery. 2008;32(8):1843-49

II. Andersson M, Rubér M, Ekerfelt C, Hallgren HB, Olaison G, Andersson RE.
Can new inflammatory markers improve the diagnosis of acute appendicitis?
World Journal of Surgery. 2014;38(11):2777-83

III. Andersson M, Kolodziej B, Andersson RE.
Structured management of patients with suspected acute appendicitis using a clinical score and selective imaging (STRAPPSCORE)
Manuscript

IV. Andersson M, Andersson RE.
Routine versus selective diagnostic imaging in patients with intermediate probability of acute appendicitis. A randomised controlled multicentre study
Submitted manuscript

ABSTRACT

Background. Acute appendicitis ("appendicitis") is one of the most common abdominal surgical emergencies worldwide. In spite of this, the diagnostic pathways are highly variable across countries, between centres and physicians. This has implications for the use of resources, exposure of patients to ionising radiation and patient outcome. The aim of this thesis is to construct and validate a diagnostic appendicitis score, to evaluate new inflammatory markers for inclusion in the score, and explore the effect of implementing a structured management algorithm for patients with suspected appendicitis. Also, we compare the outcome of management with routine diagnostic imaging versus observation and selective imaging in equivocal cases.

Methods. In study I, the Appendicitis Inflammatory Response (AIR) score was constructed from eight variables with independent diagnostic value (right lower quadrant pain, rebound tenderness or muscular defence, WBC count, proportion of polymorphonuclear granulocytes, CRP, body temperature and vomiting). Its diagnostic properties were evaluated and compared with the Alvarado score. In study II, we performed an external validation and evaluation of novel inflammatory markers for inclusion in the score on patients with suspected appendicitis at two Swedish hospitals. In study III we externally validated and evaluated the impact of an AIR-score-based algorithm assigning patients to a low or high risk of having appendicitis in an interventional multicentre study involving 25 Swedish hospitals and 3791 patients. In study IV, we compared the efficiency of routine diagnostic imaging with repeated clinical assessment followed by selective imaging in a randomised trial of 1028 patients with equivocal signs of appendicitis, as indicated by an intermediate AIR score, from study III.

Main results. In study I we found that the AIR score could assign 63% of the patients to either a high- or low-risk group of appendicitis with an accuracy of 97%, which compared favourably with the Alvarado score. In study II, the diagnostic properties of the AIR score proved to be reproducible, but the inclusion of novel inflammatory markers did not improve the diagnostic accuracy. In study III, the AIR-score-based algorithm led to a reduction in negative explorations, operations for non-perforated appendicitis and hospital admissions in the low-risk group and reduced use of imaging in both low- and high-risk groups. In study IV, routine imaging led to more operations for non-perforated appendicitis but had no effect on negative explorations or perforated appendicitis.

Conclusions. The AIR score was found to have promising diagnostic properties that were not improved further with the inclusion of novel inflammatory variables. Structured management of patients with suspected appendicitis according to an AIR-score-based algorithm may improve outcome while reducing hospital admissions and use of imaging. Patients with equivocal signs of appendicitis do not benefit from routine imaging which may lead to an increased detection of, and treatment for, uncomplicated cases of appendicitis that are otherwise allowed to resolve spontaneously.

ABBREVIATIONS

AAS	Adult appendicitis score
AIR score	Appendicitis inflammatory response score
AUC	Area under the ROC curve
CCL	Chemokine (C-C-motif) ligand 2
CFR	Case fatality rate
CI	Confidence interval
CONSORT	Consolidated standards of reporting trials
CT	Computerised tomography
CXCL8	Chemokine (C-X-C motif) ligand 8
ED	Emergency department
IL-6	Interleukin 6
IQR	Inter quartile range
ITT	Intention-to-treat
LR–	Negative likelihood ratio
LR+	Positive likelihood ratio
MAR	Missing at random
MCAR	Missing completely at random
MI	Multiple imputation
MMP-9	Matrix metalloproteinase 9

MPO	Myeloperoxidase
MRI	Magnetic resonance imaging
NMAR	Not missing at random
PMN	Polymorphonuclear granulocytes
PV–	Negative predictive value
PV+	Positive predictive value
RIF	Right inferior fossa
ROC	Receiver operating characteristic
SAA	Serum amyloid A
SD	Standard deviation
SMR	Standardized mortality ratio
Th1, Th2	T-helper 1, T-helper 2
US	Ultrasonography

BACKGROUND

ANATOMY

The appendix and caecum develop from the midgut of the human embryo, starting in the sixth week of development[1]. The appendix elongates from the posteromedial tip of the caecum and assumes an average length of about 9 cm in adults[2]. Its position is highly variable between individuals, but it is usually located in the right lower abdominal fossa. In a study of the appendix position in 10 000 subjects, the most common positions were retro-caecal (65%), pelvic or psoas-near (31%) and sub-caecal (2%) (Fig. 1)[3]. Its position with regard to the visceral peritoneum ranges from completely intra- to completely retroperitoneal, which has implications for the clinical findings and surgery for the inflamed appendix. The arterial and venous supply originates from the superior mesentery vessels via the appendiceal branch of the ileocolic artery and vein. The lymphatic vessels drain into lymph nodes surrounding the ileocolic vessels. The efferent sympathetic innervation of the appendix is brought in from the superior mesenteric plexus (T10–L1), and the afferent parasympathetic innervation is derived from elements of the vagus nerve[4].

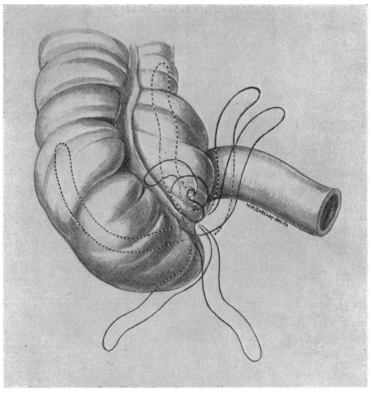


Fig. 1. Drawing from 1933 showing various positions of the appendix in relation to the distal ileum and caecum[3].

HISTOLOGY

The wall of the appendix can be divided into five principal layers from outer to inner surface: the serosa, which is an extension of the peritoneum, the muscularis propria, submucosa, muscularis mucosae and finally the mucosa. The mucosal layer resembles that of the large intestine, but the crypts are irregular, and they each contain a small number or argentaffine cells. Between the crypts and the muscularis mucosae are neuroendocrine complexes[4 5]. The mucosal layer of the appendix contains prominent lymphoid nodules consisting of a follicle centre, surrounded by a mantle of lymphocytes. The muscularis mucosae is impinged by the lymphocytes surrounding the follicles[6]. The lymphatic tissue in the appendix develops during the first year of life and continues to increase until adulthood, after which it gradually atrophies[7].

PHYSIOLOGY

The question "What good does the appendix do?" can be answered in short: Nobody knows for sure. It has been suggested that the appendix is simply a vestigial organ of evolutionary development. However, there are other theories:

The primitive sensory organ theory. Some suggest that the appendix was originally the immune system's sensory-perception organ, at least before the more sophisticated sensory-perceptive functions of our species were developed[8].

Sampling theory. Very few lymphatic follicles are present in the appendix of the newborn, although the intestine become colonised almost immediately after birth. At a few weeks of age, both follicles and germinal centres increase in size and numbers, reaching a peak in adulthood. Interestingly, this is paralleled by an absence of bacterial translocation in the appendix wall during the first two weeks, followed by an increase during the next few months. Appendix is a part of the gut-associated lymphatic tissue, and is involved in the production of IgA-, IgM- and IgG-type immunoglobulins[9]. The location of the appendix, near the ileocaecal valve, and the presence of lymphatic tissue, support the hypothesis that the appendix can help the immune system with antigenic data acquisition – the theory of the appendix being a sampling organ[10].

Safe house theory. The epithelium of the appendix is covered by a mucinous biofilm containing secretory IgA that may enhance the survival of commensal microorganisms. The regular shedding and regeneration of the biofilm may help regenerate the normal bacterial flora in the event that the large bowel becomes infected by pathogens[11].

Others have proposed that the appendix may work as a pacemaker for gastrointestinal synchronised motor function[12]. Although there are neuroendocrine cells in the appendix, and it secretes up to 2 ml of mucin-containing fluid, there is no strong evidence regarding specific endocrine or exocrine functions.

APPENDICEAL CARCINOMAS

Many types of primary and secondary tumours have been found in the appendix. Primary neoplasms are uncommon, being found in less than 1% of appendectomies, and they account for about 0.4–1% of all gastrointestinal neoplasms[13]. The majority are benign, or require no other treatment than appendectomy, and some are incidental[14-16]. Nevertheless, malignant and semi-malignant tumours exist. In the following, three entities will be briefly addressed: malignant carcinoid and mucinous neoplasms and primary intestinal-type adenocarcinoma.

Carcinoids

Carcinoids are neuroendocrine tumours with a peak incidence during the fourth decade of life[17]. The appendix is the most common site for gastrointestinal carcinoids, which are found in 0.3–0.9% of all appendectomies[18]. Appendiceal carcinoids are usually located at the tip of the appendix and are often incidental findings at appendectomy for appendicitis. They rarely cause metastatic disease[19]. Consequently, the carcinoid syndrome (systemic hormonal symptoms of flushing and bronchoconstriction secondary to liver metastases) is seldom seen. Goblet cell carcinoid is a rare form of tumour, with a peak incidence a little later in life and has a more aggressive behaviour than the conventional carcinoids[20].

Prognosis and management. Serosal involvement is common, but is not considered to predict aggressive behaviour[17]. On the other hand, tumour size is an important prognostic factor. For carcinoids smaller than 15–20 mm without signs of lymph-node or appendix-base involvement, appendectomy is considered a sufficient treatment. For larger tumours, right hemicolectomy is recommended[21 22]. While the prognostic value of the mitotic activity is well established for neuroendocrine tumours of other origins, the evidence is not as strong for primary appendix carcinoids. Nevertheless, some regard proliferation markers as complementary tools in the decision making regarding these patients[18]. Follow-up for patients operated for carcinoids larger than 10–20mm involve plasma chromogranin A-screening, computerised tomography (CT) imaging in cases of elevated chromogranin A-levels, and octreotide scintigraphy for diagnosing metastatic disease. Overall, the prognosis of appendiceal carcinoids is favourable, with a five-year survival of 90–100%.

Mucinous adenocarcinoma

Mucinous adenomas and adenocarcinomas of the appendix arise from dysplastic mucinous epithelium. A "mucocele" refers to an appendix that is dilated and filled with mucus, regardless of underlying cause (simple obstruction, benign mucinous adenomas, mucinous adenocarcinomas etc.).The histological distinction between non-invasive and invasive disease is difficult. A wide spectrum of histological features have been described, and consequently there is some confusion in the histological and clinical classification[13 23]. If mucin and neoplastic cells are found not only in the lumen of the appendix, but also on the outer surface of the appendix, the neoplasm is regarded as invasive. This underlines the importance of careful handling of the specimen by the surgeon to avoid the risk of seeding. Mucinous adenocarcinoma is characterised by the presence of neoplastic cells or lakes of mucin in the appendiceal wall, or on the outer surface[24]. A finding of acellular mucin on the exterior of the appendix represents a special diagnostic difficulty. This suggests either a local contamination of intraluminal mucin during the removal of an appendix with a non-invasive mucinous neoplasm, or a well differentiated hypocellular mucinous adenocarcinoma[25].

Pseudomyxoma peritonei syndrome. A mucinous neoplasm originating from the appendiceal epithelium, with dissemination beyond the primary site and production of mucinous peritoneal ascites is a clinical condition classified as the pseudomyxoma peritonei syndrome[23].

Incidental findings at operation. Mucinous neoplasms may present with symptoms suggestive of appendicitis. Hence, during the surgical exploration, the surgeon may find anything from an unusually pronounced swelling of the appendix to an abdominal cavity containing copious amounts of mucinous ascites. In the latter scenario, an extensive procedure aiming at complete cytoreduction, combined with intraperitoneal chemotherapy is required[26 27]. If, on the other hand, there is a small (<2cm) tumour in the appendix, not involving the base and no evidence of extra appendiceal disease, appendectomy may be a sufficient treatment. However, if the base of the appendix is involved, or the non-perforated neoplasm is larger than 2 cm, a right hemicolectomy is recommended[13 28].

Intestinal-type adenocarcinoma

Non-mucinous adenocarcinomas of the appendix are even less common than mucinous neoplasms, and develop from tubulous or tubulovillous adenomas[29]. Like other neoplasms of the appendix, the majority are incidentally found at operation for appendicitis symptoms. These tumours resemble colorectal adenocarcinomas, hence the names colonic-type or intestinal-type adenocarcinoma. The intestinal type adenocarcinoma seems to have a different biology from the mucinous adenocarcinoma, with a higher proportion of low-differentiated tumours, and a higher proportion of lymphatic node involvement[30]. Some suggest right hemicolectomy for localised disease, regardless of size, and others propose local resection (appendectomy) for tumours less than 20 mm, not involving the base of appendix and with no signs of lymphatic spread[13 30].

APPENDICITIS

Historical aspects

The appendix was first described at the beginning of the 16th century, and was given the name "appendix vermiformis"(Lat. wormlike) by Vidius Vidius, an anatomy teacher in Pisa, Italy[9]. The first appendectomy was performed in 1736 by Claudius Amyand, as he encountered a perforated appendix in a hernia with a faecal fistula. The term appendicitis was coined by Reginald H. Fitz, professor of pathology at Harvard University, USA[4]. This moved the precedent focus from the caecum, and "typhlitis", towards the appendix, and appendicitis. The London surgeon Robert Lawson Tait performed the first intentional appendectomy for appendicitis in 1880, and in 1889 Charles McBurney published his report on appendicitis[31]. He also described what was ever since referred to as "McBurneys point", which he defined as a point "1½–2 inches inside of the right anterior superior spinous process of the ileum on a line drawn to the umbilicus"[32]. A few years later he published on the "gridiron incision", still the standard incision used for open appendectomies[33].

Although initially considered a self-limiting disease, McBurney, among others, advocated early exploratory laparotomy[32 34]. In 1889, appendectomy was introduced in Sweden by Karl Gustav Lennander in Uppsala. The number of appendectomies in Sweden increased steadily during the first part of the 20th century. The mortality from appendicitis did not decrease until the middle of the century, which was coincidental with the introduction of antibiotic and intravenous fluid therapy[35].

Epidemiology

More than 10 500 appendectomies are performed annually in Sweden, and about 9500 of these patients are found to have appendicitis[36]. The lifetime risk of appendicitis is estimated at about 7% for females and 9% for males. The risk of having an appendectomy is higher, making it the most commonly performed emergency procedure in general surgery[37 38]. The appendicitis incidence is approximately 100 per 100 000 persons per year. Since the 1960s there was a decrease until the 1990s[39 40]. This seems mainly attributed to a decrease in non-perforated appendicitis, whereas the incidence of about 20/100 000 for perforated

appendicitis is more stable over time[37 41-44], and across all age groups[44 45]. In contrast, the incidence of non-perforated appendicitis shows a secular decreasing trend and is strongly age-dependent, with a peak in the second decade of life[37 44-46]. A slight increase of non-perforated appendicitis after the mid 1990s is seen, which may be the result of an increased detection rate following the introduction of CT and diagnostic laparoscopy[41 47 48].

Aetiology

Many theories have been presented over the years regarding the causes of appendicitis. A hypothesis involving the combination of immunological characteristics of the individual, and local conditions in the appendix, such as obstruction seems to have some support in the literature.

Obstruction. The appendix secretes small amounts of mucin and contains bacteria that grow continuously. An outlet obstruction could therefore increase the intraluminal pressure which may cause swelling, decreased blood supply, subsequent bacterial translocation and necrosis of the wall[49]. Animal models with ligation of the proximal appendix have shown that obstruction does indeed induce inflammatory changes much like those seen in acute appendicitis in man[50]. The obstruction may be caused by fecaliths or calculi, foreign bodies, faecal obstruction, fibrous bands or lymphoid hyperplasia[49 51-53].

The role of obstruction was more or less directly addressed in one study by inserting a pressure-measuring needle into inflamed appendices[54]. As most inflamed appendices in that study had normal intraluminal pressure, the conclusion was that obstruction was not likely the primary step in pathogenesis, but may occur secondarily as a *result* of the inflammatory process. Furthermore, fecaliths, which are most commonly believed to be the main cause of obstruction, only exist in a minority of inflamed appendices, and are also seen in uninflamed appendices[55].

Infection. Seasonal variations with an increase during summer months have been described, and clusters of appendicitis among children, which may indicate an infectious aetiology[56 57]. Others have failed to show any correlation with viral infections[58]. The role of infections as a cause of appendicitis is not fully understood.

Diet and hygiene. The low-fibre diet of the western world correlates to some extent with geographical differences in acute appendicitis, which has led to theories of dietary causes of appendicitis[59]. That theory has also been challenged, however, as the sharp decline in appendicitis incidence since the 1950s is not mirrored by changes in dietary intake. Instead, the improved sanitation and living standard, causing a change in immune response have been suggested as another hypothesis[60].

Immunological. The T helper (Th) and cytotoxic T cells are the principal types of T cells involved in the adaptive immune response of humans[61]. The Th cells (also called "CD4+ T cells") are further subdivided into, among others, Th1 and Th 2 cells depending on their cytokine-secreting pattern[62]. Different individuals' immune reaction to a certain stimulus is complex, and may have a constitutional preference for either a Th 1 or Th 2 cell-mediated inflammatory response. The immune response in patients with Crohn's disease and ulcerative colitis is characterised by an exuberant Th1 and Th2-like pro-inflammatory activity, respectively[63]. Interestingly, perforated appendicitis is correlated with Crohns disease, whereas an inverse correlation between appendicitis and ulcerative colitis exists[64-66]. The hypothesis of a protective effect of a Th 2-like immune response predominance on the development of appendicitis is further strengthened by the coincidental drop in appendicitis incidence and shift towards a Th 2-like immune response in the third trimester of pregnancy[67-70]. On the other side is the association between the Th 1-dominated immune response seen in Crohn's disease and perforated appendicitis, which is also supported by the effects of an excessive Th 1 response, namely tissue damage and necrosis[71].

Definition of appendicitis

There is no gold standard definition of appendicitis. Neither is there consensus with regard to the nomenclature or histopathological criteria for different grades of appendicitis[72]. The issue is further complicated by the non-operative management of subgroups of appendicitis patients. For obvious reasons, no histopathological diagnosis will be obtained in those cases.

Histological diagnosis

Mucosal inflammation. Histological changes of inflammation not involving the muscularis propria are variably classified as "mild", "limited", "early", or "superficial" appendicitis in the literature. However, mucosal inflammation is found as often in appendices from incidental appendectomies as in primary appendectomies[73]. Furthermore, patients operated for suspected appendicitis with histopathological findings of inflammation depth limited to the mucosa do not differ in any clinical characteristics from patients without any microscopic evidence of inflammation at all[74]. In this study, mucosal inflammation alone is *not* considered an appendicitis criterion.

Definition of appendicitis in this study

- *Phlegmonous appendicitis*. Inflammation with transmural infiltration of neutrophil granulocytes[75]. Micro-abscesses, oedema or vascular thrombi may, or may not, be present. Transmural necrosis or perforation is not present.
- *Gangrenous appendicitis*. Inflammation with transmural infiltration of neutrophil granulocytes and transmural necrosis of the appendiceal wall[76]. No signs of perforation are detected at surgical removal.
- *Perforated appendicitis*. Macroscopic findings of perforation at operation and transmural inflammation at histopathological examination.
- *Appendiceal abscess*. A collection of pus surrounding a perforated appendix found during operation, or as indicated by diagnostic imaging in non-operated patients.
- *Non-perforated appendicitis*. Phlegmonous or gangrenous appendicitis.
- *Advanced appendicitis*. Gangrenous or perforated appendicitis, or appendiceal abscess.
- *Antibiotic treated appendicitis*. Non-operated appendicitis that was verified by unequivocal signs of appendicitis at diagnostic imaging and treated with antibiotics.
- *Non-treated appendicitis*. Unequivocal findings of appendicitis at diagnostic imaging in a patient who received no surgical or antibiotic treatment.

Natural history

The view of the prognosis and optimal treatment strategy of appendicitis has changed over time. During the 20th century, however, the predominant understanding has been that more or less invariably, the disease progresses from onset to perforation with associated serious adverse effects or death. Consequently,

an aggressive surgical attitude promoting early surgical exploration on wide indications ("when in doubt, cut it out"), has been adopted. This dogma has been challenged over the last few decades.

Spontaneous resolution

The natural history of appendicitis in general, and spontaneous resolution in particular, is an ongoing controversy that is unlikely to settle in the near future; spontaneous resolution is questioned by proponents of histopathologic evidence of appendicitis as a prerequisite for proving a subsequent disease resolution. This is, by definition, impossible to obtain (unless the appendix is re-implanted after removal), and is consequently a circular chain of evidence.

On the other hand, there is a growing body of indirect evidence supporting the view that spontaneous resolution occurs. In retrospective studies comparing "aggressive" and "expectant" management of patients with suspected appendicitis, fewer patients require appendectomy in the expectant management group[44 75]. Observational studies have also compared groups of patients managed with liberal or restrictive use of CT. The restrictive use of imaging is correlated with a lower number of patients operated for appendicitis in adults, as well as in children[77-79]. This correlation is further supported on an epidemiological level in a longitudinal study showing that the 25-year nationwide decrease in non-perforated appendicitis incidence rate in the USA has been replaced by an increase coincidental with the introduction and increased use of CT[41 80]. Case-series of resolving ultrasonography (US)- or CT-verified appendicitis provides further evidence, but one should underline that this is reported in a limited number of cases[81-84]. Prospective randomised controlled trials comparing the management of routine diagnostic laparoscopy with expectant management have shown an increase in the number of patients operated for appendicitis in the laparoscopy group[47 48]. Again, this supports spontaneous resolution in uncomplicated cases in the expectant group.

Perforated and non-perforated appendicitis

The different patterns with regard to incidence and response to diagnostic and interventional efforts for perforated and non-perforated appendicitis suggest that they are two separate entities with different natural history (Fig.2).

Perforated appendicitis. Clearly, the inflammation of the appendix will progress and cause perforation in a number of appendicitis patients. Although the proportion of perforations varies between different studies, the *incidence* of perforation is relatively constant, regardless of age, sex, surgical management, or use of diagnostic imaging[41 44]. The incidence of perforation is also strikingly constant over time, according to epidemiological studies[41 44 45 80]. Furthermore, it has been reported that patients with perforated appendicitis often have a long pre-hospital delay, but are usually identified early as candidates for immediate surgery, which is mirrored by a short in-hospital delay in this group[85 86]. Although this is true for the majority of cases, one has to underline that the presentation of patients with perforated appendicitis can be elusive. Failure to diagnose and provide appropriate care may lead to severe deterioration of the patient's condition.

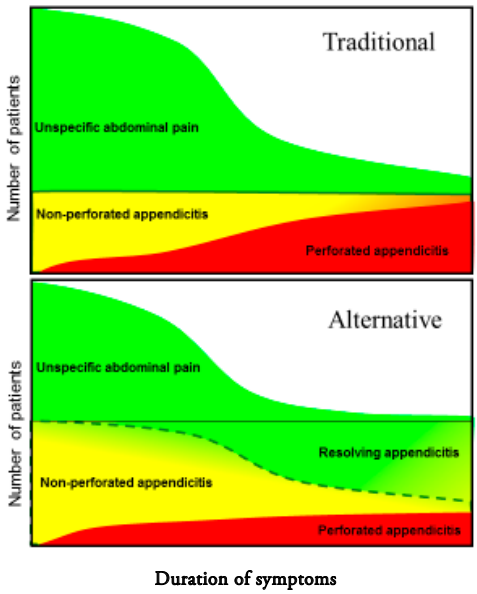


Fig. 2. Traditional and alternative understanding of the natural history of appendicitis.
Revised version, based on an illustration made by Andersson, R.E

Non-perforated appendicitis. The stable incidence of perforated appendicitis contrasts with the highly variable incidence of non-perforated appendicitis according to age, surgical management and use of diagnostic imaging[41 44 45].

This suggests that, in effect, health care providers have limited possibilities to influence the *number* of perforations. A more intense diagnostic workup will rather lead to the detection and operations for potentially resolving appendicitis, which will reduce the proportion of perforations due to an increase of the denominator. Therefore, the *proportion* of perforated appendicitis is an inappropriate quality measure.

TREATMENT OF APPENDICITIS

After McBurney introduced the gridiron incision at the end of the 19th century, open appendectomy became the standard procedure for patients with suspected appendicitis for 100 years. Most open appendectomies are performed through a muscle-splitting incision in the right lower quadrant of the abdomen. In the 1980s diagnostic laparoscopy and laparoscopic appendectomy were introduced and became increasingly popular during the 1990s. A recent development is the single-incision laparoscopy, which theoretically could achieve better cosmetic results[87 88]. The latest surgical technique, which is not in routine use, is the natural orifice transluminal endoscopic surgery (NOTES), sometimes referred to as "scarless surgery"[89]. During the last decade antibiotic treatment has been proposed as an alternative to surgical treatment.

Open or laparoscopic operation

The proportion of appendectomies that are performed with laparoscopic technique is increasing in both Europe and the USA[36 90 91]. A number of randomised controlled trials have been performed on adults and children to illuminate the outcomes of open *vs* laparoscopic appendectomy. The latter have various benefits, including fewer wound infections, reduced postoperative pain and shorter hospital stay at the cost of a higher risk of intra-abdominal abscesses and longer operating time[92]. Moreover, if the preliminary diagnosis is wrong, laparoscopy (i.e. *diagnostic* laparoscopy) enables the surgeon to perform a full inspection of the entire peritoneal cavity and make an alternative diagnosis in many cases, which is particularly true for women[93]. Primary laparoscopic approach in fertile women, with only diagnostic laparoscopy performed in negative cases, is associated with a reduction in negative appendectomies and a higher proportion of patients receiving a definitive diagnosis[92].

In conclusion, the differences in outcome between open and laparoscopic appendectomy are small, and in clinical practice it seems reasonable to choose the surgical method depending on the experience of the surgeon as well as patient preference and clinical context[36 92].

Appendectomy or antibiotic treatment

Antibiotics as a first line treatment of appendicitis have been tested in several randomised trials[94-99]. The results are conflicting, which also applies to the many meta-analyses of randomised trials on this topic[100-103]. Overall, antibiotic treatment cures approximately 75% of the patients within two weeks without recurrence or other major complications within one year, as compared with over 97% for primary appendectomy. However, this difference was found to be inconclusive in a systematic review and meta-analysis[103]. In retrospective studies, the presence of an appendicolith demonstrated on diagnostic imaging is reported to predict increased risk of recurrent appendicitis[104 105]. In a recent prospective study, diagnostic appendicitis scores, which are discussed more in detail in the following sections of the thesis, were independent predictors of failure of antibiotic treatment of patients with suspected appendicitis[106].

For patients with a palpable mass, appendiceal phlegmone or abscess demonstrated at imaging, surgical treatment is associated with increased morbidity and need for more extensive bowel resection due to distorted anatomy by advanced inflammatory changes[107]. Antibiotic treatment with percutaneous drainage in cases of abscess is successful in more than 90% of the cases, with a risk of recurrence of less than 10%, and a reduced morbidity and hospital stay compared with surgical treatment[108]. However, non-operatively managed patients should be followed up in order to exclude Crohn's disease and malignancies.

Morbidity and mortality

Postoperative morbidity and mortality are influenced by several factors such as disease severity, operative technique, use of antibiotics, the patient's age and comorbidity. Infectious complications are the most common postoperative complications, but small bowel obstruction is also a matter of concern in the long run.

Wound infection and intra-abdominal abscess. Infectious complications occur more often after operation for advanced appendicitis[109]. In a systematic review of randomised controlled trials comparing open appendectomy with laparoscopic appendectomy, the risk of wound infection for adults and adolescents was 7.4% and 3.1% for open and laparoscopic operation, respectively. In contrast, the risk of intra-abdominal abscess was doubled for laparoscopic compared with open appendectomy (1.8% vs 0.95%).

The same trend was seen for children, but the risk of intra-abdominal abscess was not significantly higher in the laparoscopic group[92]. The use of prophylactic antibiotics is reported to reduce postoperative infectious complications[110].

Small bowel obstruction and intestinal damage. The reported risk of small bowel obstruction after appendectomy varies, but is approximately 1.5% during the first 15 postoperative years according to a population-based cohort study[36]. In this study the risk was lower following primary laparoscopic appendectomy than open appendectomy during the first two years, but after that no difference remained. Others have reported a lower risk of small bowel obstruction for laparoscopic appendectomy than for open appendectomy, especially in non-randomised studies, which may be attributed to differences in case-mix or comorbidity as these findings have not been supported in a meta-analysis of randomised controlled trials[111 112].

Mortality. As previously mentioned in the historical section, the annual death rate from appendicitis did not decline during the first decades of the 20th century in spite of an enormous increase in appendectomies, and was about 15 per 100 000 in the United States during the 1930s[113]. Only in the middle of the century, after the introduction of intravenous fluid therapy and antibiotics, did the mortality decrease[35 60]. Mortality following appendicitis, and appendectomy, in the modern era is low, but is influenced by several factors. The case fatality rate (CFR) within 30 postoperative days in a population-based study of all appendectomies during a 10-year period in Sweden was 2.44 per 1000 operations, with a sharp increase among the oldest patients[114], which is in keeping with the overall CFR in England during a 10-year period between 1996 and 2996 (CFR 2.4 per 1000 emergency appendectomies)[91]. Interestingly, the standardised mortality ratio was higher for negative appendectomies (9.1) than for perforated appendicitis (6.5) and non-perforated appendicitis (3.5)[114]. The most important causes of mortality after appendectomy today are probably an interaction between comorbidity, anaesthesiosurgical trauma and diagnostic failure[115].

DIAGNOSING APPENDICITIS

Signs and symptoms

The clinical diagnosis of appendicitis is sometimes straightforward, with a "schoolbook" presentation of initial vague abdominal pain, followed during the next 24 to 48 hours by elevated body temperature, nausea, pain migration towards the right lower quadrant of the abdomen, and signs of localised peritonitis at clinical examination. But often the symptoms and disease history are more ambiguous. The diagnosis is especially challenging in small children, in the elderly and during pregnancy. A rich flora of signs and symptoms are seen in appendicitis patients, some of which are more closely associated with the disease and are presented in this section.

Disease history and symptoms

Abdominal pain. There is no universal definition of *acute* abdominal pain. For the purpose of this thesis, we have considered pain duration of less than five days as acute. The initial pain in appendicitis is often described as dull or diffuse, which is probably due to the stimulation of visceral afferent nerve fibres. As the disease progresses during the next hours or day(s), inflammatory changes extend to the serosa of the appendix and parietal peritoneum in the region. This is thought to cause the characteristic relocation or "migration" of pain towards the location of the appendix, usually in the right lower quadrant (RLQ)[1]. Patients typically complain about aggravation of pain with sudden movements[116 117]. Occasionally, the pain history is more dramatic, with a more sudden onset of intense pain, which may be attributed to the presence of an obstructive fecalith[118 119]. However, intense abdominal pain should also make the clinician consider other diagnoses, in order to avoid false positive decisions[120].

Tenderness. Tenderness over the location of the appendix, most often in the RLQ, or even over McBurney's point, is a common finding at clinical examination in both children and adults[121 122].

Nausea and vomiting. Appendicitis often causes gastric upset with anorexia, nausea and vomiting[123 124]. These symptoms may cause patients, or health care personnel, to misinterpret the condition as gastroenteritis.

Elevated body temperature. Fever usually develops at some stage as a part of the systemic inflammatory response, but rarely precedes the development of abdominal pain[124].

Rebound tenderness and muscular defence. Rebound tenderness refers to pain elicited by removal of pressure during palpation of the abdomen. Muscular defence is characterised by involuntary muscular contraction, or *guarding*, upon applying pressure. Percussion tenderness and indirect tenderness (pain in the RLQ upon palpating left lower quadrant; Rovsing's sign) may also be present. These findings are all considered signs of local peritonitis[122 124 125].

Biochemical inflammatory markers

The immune system has evolved to defend us against microbes. Simply speaking, it consists of three defence lines with increasing specificity[61]:

- Barriers, such as mucous membranes or skin. An unspecific physical or chemical barrier.
- The innate immune system. Provides the host with a rapid response to pathogens or signals from injured cells at the expense of specificity.
- The adaptive system, which develops a targeted response to specific antigens, and has the ability to generate an "immunological memory."

These defence lines induce complex cascades of actions and counter actions, involving potent regulating mechanism in order to orchestrate the inflammatory process. Failure to regulate the immunological response appropriately can cause negative effects on the host, such as tissue damage and/or persistent inflammation.

Appendicitis, by definition, involves inflammation, either as a primary cause or as response to a stimulus preceding the condition. Inflammatory cells, acute phase reactants and other components of the immune system are linked with the condition. Some are used as predictors in clinical routine healthcare while the value of others is unclear.

Inflammatory markers used in routine health care

White blood cell count (WBC). White blood cells, or leucocytes, are involved in the innate and adaptive immune response. Consequently, in appendicitis, an increase in WBC is seen in both children and adults[72 124 126 127]. In experimental models, as well as in observational studies, a local and systemic increase in the number of white blood cells is seen at an early stage of disease[128 129]. Pregnancy is associated with a physiological increase in WBC, which should be considered when interpreting the result of a case of suspected appendicitis in pregnancy[130].

Polymorphonuclear granulocytes (PMN). PMN and mononuclear cells are the two main subgroups of white blood cells. The vast majority of PMN are neutrophil granulocytes. Eosinophil and basophil granulocytes constitute a few percent of the total PMN count[131]. Normally, PMN constitute 40–60% of the circulating white blood cell population, which approximately applies to the proportion of neutrophil granulocytes as well[132]. Appendicitis is accompanied by an increase in neutrophil count, and in the proportion of neutrophils [72 124 133].

C-reactive protein (CRP). Inflammation promotes the release of Cytokines, Chemokines and stress hormones. As a response, the hepatocytes of the liver produce a variety of acute-phase proteins. CRP, the dominating acute phase-protein, is an early marker of inflammation and can increase by up to a 1000-fold within 24–72 hours in response to acute stimuli[134-136]. It was first discovered in the 1930s in the sera of patients with pneumonia, and was reported to bind to the "Fraction C" polysaccharide component of the pneumococcal wall[137]. The physiologic role of CRP is still not fully known, but it takes part in activation of complement, opsonising of microbes, regulation of coagulation in sepsis but also in modulation of inflammatory response[61 138 139]. Elevated serum levels are associated with appendicitis in patients of all ages, and in accordance with the dynamics of CRP production, the correlation grows stronger with the time after onset of symptoms[72 124 129 140].

Other inflammatory markers

A large range of cytokines, chemokines and acute-phase reactants which are not in routine use for diagnosing appendicitis have been evaluated with regard to their discriminating and predictive properties[141-150]. Although some of them have promising diagnostic properties when used alone, so far none of them has proven to

provide additional predictive value when used in combination with the established predictors described above.

Diagnostic imaging

Few surgical procedures with a 10–20% risk of finding an unaffected target organ are considered acceptable today. A substantial number of negative explorations were previously regarded as a positive quality measure as surgical exploration on liberal grounds was thought to reduce the risk of appendiceal perforation[151]. Today, efforts are made to diagnose patients more accurately with the intent to minimise negative appendectomies, fast-track surgery and, if possible, to reduce the number of perforations using diagnostic imaging techniques.

Computed tomography (CT)

The use of CT has increased tremendously during the last few decades in Europe, the United States and Japan[152]. During the 1980s, the diagnostic properties of CT with regard to appendicitis were published[153 154]. Appendiceal CT was first adopted in the United States and became more widely used in the diagnosis of appendicitis during the 1990s[41 155 156]. Today, multi detector helical ("spiral") 5mm section standard dose CT with or without enteral or intravenous contrast enhancement is often used, but "low-dose" CT has also been proposed as a feasible alternative[157]. In high income countries, round-the-clock availability of CT is high, which makes it a useful diagnostic tool in emergency care. Furthermore it can provide an alternative diagnosis when the clinical diagnosis of appendicitis is incorrect.

Diagnostic criteria of appendicitis. There is no general agreement with regard to diagnostic CT criteria of appendicitis. The outer diameter threshold is set at 6–10mm in different studies[158 159]. The presence of contrast-enhanced and thickened appendiceal wall, periappendiceal fat stranding, extra-luminal air, "arrowhead sign" and absence of intraluminal air are all regarded as signs of appendicitis, although the latter sign is unspecific[157 159 160].

Diagnostic properties. In general, the diagnostic properties of appendiceal CT are favourable. In a meta-analysis of prospective studies evaluating CT in suspected appendicitis in adolescents and adults, the pooled estimates for sensitivity and

specificity were 0.94 and 0.94, respectively[161]. In a more recent meta-analysis restricted to prospective comparative studies of CT and US, the sensitivity and specificity for CT were 0.91 and 0.90, respectively[162]. Also, the level of inter-observer agreement in diagnosing appendicitis with CT is reported to be good, although it may influence the diagnostic accuracy at least as much as the type of CT protocol used[163 164].

Areas of controversy. While many have reported excellent outcomes when using CT for diagnosing appendicitis, which is reflected by the results of a recent meta-analysis, others have failed to correlate the increased use of CT with improved diagnostic accuracy on a population level[165-167]. Furthermore, only a few randomised controlled studies have compared the use of CT with clinical assessment, and the results are conflicting[168-170].

Ionising radiation. Children are inherently more sensitive to radiation exposure and have more years left at risk of developing cancer. Since the majority of patients with appendicitis are young, the increased use of CT in patients with suspected appendicitis have raised concerns regarding the potential harmful effects of ionising radiation (Fig. 3)[171]. These may be elicited on either deterministic or stochastic grounds. Extrapolation of data from atomic bomb survivors and nuclear industry workers show a linear correlation between radiation dose and cancer risk [152 172 173]. Epidemiological studies have found a small excess cancer risk attributed to a single CT scan, which is highest for abdominal CT, and for exposure early in life[174]. However, the technical development of CT hardware and optimised low-dose protocols is continuously reducing the radiation dose, which will be beneficial unless counteracted by a corresponding increase in CT scan incidence.

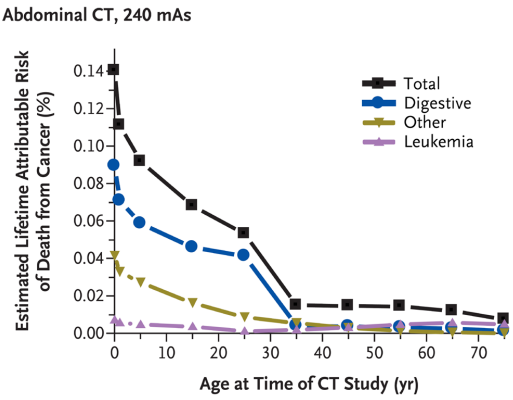


Fig. 3. Estimated lifetime risk of death attributed to the radiation from a single CT scan.
Reproduced with permission from Brenner DJ, Hall EJ. Computed tomography– an increasing source of radiation exposure. The New England journal of medicine 2007;**357**(22):2277-84, Copyright © Massachusetts Medical Society

Ultrasonography (US)

Appendiceal US was introduced in the 1980s[175]. The term "graded compression" was coined by Puylaert in 1986 to refer to the pressure applied to the transducer by the US operator in order to displace overlying bowel and to reduce gas artefacts[176]. The reduced distance from the transducer to the appendix also allows the use of high-frequency transducers that yield higher resolution. This technique was combined with colour doppler and curved transducers in the 1990s, which further improved the diagnostic properties[177].

Diagnostic criteria of appendicitis. Although there is an overlap with regard to the outer diameter of a normal and that of an inflamed appendix, 6 mm is in general used as a threshold for a positive test[178]. A normal appendix can be compressed easily and has no visible colour-doppler flow. Thus, a non-compressible appendix with visible colour-doppler flow is suggestive of appendicitis[179 180].

Diagnostic properties. US has the advantage over CT in that it does not expose patients to ionising radiation, but has a longer learning curve and higher operator dependency[181]. While the interpretation of appendiceal CT is facilitated by abdominal fat, the opposite is true for US; thin patients or patients with normal habitus are easier to examine. Overall, US performs inferior in terms of diagnostic accuracy than CT. In two meta-analyses the sensitivity and specificity were 0.78–0.86 and 0.81–0.83, respectively[161 162]. However, in experienced hands, and for both paediatric and adult patients, diagnostic accuracy well over 90% is reported[182 183].

Magnetic resonance imaging (MRI)

The use of MRI in appendicitis cases was reported in the early 1980s[184]. The initial problems with MRI included time-consuming data acquisition and the low resolution of the images. Today, the resolution is high, contrast enhancement is available and the data acquisition is quicker, albeit not as quick as helical CT. Therefore, in order to avoid motion artefacts, the patient must be co-operable. MRI is not as readily available as CT and US in most centres, which limits its use in emergency cases. However, it is regarded by many as the modality of choice for pregnant women, in particular if US is inconclusive[185].

Diagnostic criteria of appendicitis. An outer appendix diameter of more than 6 or 7 mm and increased wall thickness is suggestive of appendicitis, together with oedema of the appendix wall and surrounding fat[186 187].

Diagnostic accuracy. In general, the diagnostic accuracy of MRI is reported to be higher than US, close to that of CT. In a meta-analysis including both prospective and retrospective studies, the pooled sensitivity and specificity were 97% and 95%, respectively[187].

Clinical scores

The suspicion of appendicitis is usually raised by the clinician as a result of a synthesis of the patient's disease history, clinical signs, symptoms and basic biochemical markers. This is a complex process that is dependent on the individual physician's previous clinical experience. However, many physicians involved in the primary management of patients with acute abdominal pain are in the beginning of their career, and thus clinical scores have been proposed to provide a condensation of conclusions and experiences drawn from a large number of similar cases. A clinical score can be used as a triage test at the emergency department, but can also be repeated for cases that are observed over a period of time. Repeated scoring may detect signs of resolution or progression of the disease. Ultimately, this enables the clinician to determine the prognosis of the present patient with suspected appendicitis, and to manage the patient accordingly.

Construction of diagnostic scores

In order to avoid spectrum or verification bias, diagnostic scores should be developed and evaluated on the group of patients they are designed to serve, namely patients with abdominal pain and suspected appendicitis[188]. Relevant variables with independent predictive values should be included and the weight of the variables should be determined using an appropriate mathematical model. Finally, the score should be user-friendly and have high discriminating capacity and predictive value.

Appendicitis scores

A large number of scores have been proposed. Some are exclusively designed for children, others for adults and some for patients of all ages. The Alvarado score, the Lintula score, and the Pediatric Appendicitis Score (PAS) are among the most well-known and widely used[189-191]. The Appendicitis Inflammatory Response (AIR) score, which is constructed and validated as a part of this thesis, was published in 2008[192].The construction of new, refined diagnostic scores has continued; the most recent score to date was published in 2014[193].

Impact on patient outcome. Most scores have been reported to yield high diagnostic accuracy in the original reports, but this is not necessarily confirmed in external validation studies [194 195]. Preferably, the diagnostic accuracy of the score should be externally validated in cross-sectional studies with prospectively collected data. If the score is intended to support clinical decision-making or replace diagnostic imaging, it should also undergo interventional and/or randomised studies in order to define the diagnostic score's impact on patient outcome[196 197].

Ohmann et al. conducted a large pre-post interventional study during which a clinical appendicitis score developed earlier, and Mán et al. recently published a study in which patients were assigned in a weekly alternation to an intervention group (management according to Alvarado score) and a control group (clinical assessment and US)[198 199 200]. The results of these studies were discouraging, with no positive effect attributed to the use of the scores in question. In 2008, Lintula et al. published on a randomised study in which children were assigned to either a score-based algorithm or standard clinical management[191 201]. They found an improved diagnostic accuracy in the score-group, but when same score-based algorithm was applied to adults, no effect on diagnostic accuracy was found[202]. Unlike the AIR score, the Lintula score does not include biochemical markers, which may limit its diagnostic properties.

Others have demonstrated either a decrease in the use of CT or an increase in diagnostic accuracy when implementing structured clinical pathways in pre-post interventional studies, which is in keeping with results presented in this thesis (study III)[203 204 205].

PRESENTATION OF DIAGNOSTIC PROPERTIES

Diagnostic tests are rarely definitive, which is particularly true for diagnostic scores and other quantitative tests. They generally do not provide a binary "yes" or "no" result. Therefore it makes sense to present the results according to a three-zone partition generated by a low- and high cut-off[206]. The low-risk zone should exclude the disease (or need for treatment of disease), the intermediate-risk zone indicates an equivocal diagnosis and the high-risk zone should confirm the disease. Consequently, the diagnostic properties of each cut-off, as well as the proportion of patients in the grey (i.e. intermediate) zone comprise the overall diagnostic properties of the diagnostic test. Whether discriminating capacity in terms of area under the receiving operator characteristic (ROC) curve, sensitivity and specificity, or likelihood ratios and predictive values are the most appropriate measures of the score's diagnostic performance is not a clear-cut case. In the following section, these metrics will be defined.

Measures of diagnostic characteristics

The performance of a test can be measured and presented in many different ways, but what we really like to know is that the test can:

- Correctly identify individuals with the disease of interest
- Correctly identify individuals without the disease of interest

These two dimensions are interconnected, so for one-dimensional metrics paired statistics should be presented (e.g. sensitivity *and* specificity) in order to allow meaningful interpretation of the results. Some metrics are not very intuitive, and are prone to induce confusion for the individual clinician, or researcher (e.g. sensitivity), while others seem easy to understand, but are actually inherently difficult to generalise from (e.g. positive and negative predictive value, which are true only for a specific disease prevalence).

In the following, the measures used in this thesis are derived, with the understanding that the terms "diseased" and "non-diseased" are used instead of diseased or non-diseased as determined by the gold standard (Fig. 4).

Fig. 4. Cross-tabulation of disease status and test result

		Gold standard	
		Diseased	Non-diseased
Test result	Positive	**A** True positive	**B** False positive
	Negative	**C** False negative	**D** True negative

Sensitivity

Sensitivity is defined as the proportion of diseased subjects with a positive test result: A/(A+C). Another way to put it is that sensitivity is the probability that a diseased subject will get a positive test result. A negative test result of a test with high sensitivity will therefore rule out the disease. It is important to recognise that sensitivity expresses the test performance in those who have the disease, regardless of its performance in those who do not have the disease.

Specificity

Specificity is defined as the proportion of non-diseased subjects with a negative test result: D/(B+D).

In other words, specificity is the probability that a non-diseased subject will get a negative test result. A positive test result of a test with high specificity will rule in the disease. In contrast to sensitivity, the specificity expresses the test performance in those who do not have disease, regardless of its performance in those with the disease.

Positive predictive value

Positive predictive value (PV+) is defined as the proportion of subjects with a positive test result that are diseased: A/(A+B)

PV+ represents the probability that a patient with a positive test result actually has the disease, regardless of the properties of a negative test. PV+ is dependent on the disease prevalence in the population subjected to the test.

Negative predictive value

Negative predictive value (PV–) is the proportion of subjects with a negative test result that are non-diseased: D/(C+D)

Thus, the PV– reflects the probability that a patient with a negative test actually does not have the disease. Again, this does not imply anything with regard to the diagnostic properties of a positive result of the same diagnostic test. In accordance with PV+, PV– is specific for the disease prevalence of the population in question, and is not transferable to another population with different disease spectrum.

Likelihood ratio

Likelihood ratios (LR) report the direction as well as the magnitude of a test result's impact on the probability of the condition tested for. LR is calculated by dividing the likelihood of a test result for diseased subjects by the likelihood of the same test result for non-diseased subjects. LR can be used with Fagan's nomogram in order to convert pre-test probability of the condition to post-test probability according to Bayes' theorem[207].

Positive likelihood ratio. The positive likelihood ratio (LR+) is the ratio of the proportion of diseased subjects with a positive test and the proportion of non-diseased subjects with a positive test:

- [A/(A+C)]/[B/(B+D)]
- or sensitivity/(1-specificity)
- or proportion true positives/proportion false positives

LR+ thus reflects how many times more likely a positive test result is for diseased than for non-diseased subjects. This is true regardless of the disease prevalence (pre-test probability) of the population taking the test.

Negative likelihood ratio. The negative likelihood ratio (LR–) is the ratio of the proportion of diseased subjects with a negative test and the proportion of non-diseased subjects with a negative test:

- [C/(A+C)]/[D/(B+D)]
- or (1-sensitivity)/specificity
- or proportion false negatives/proportion true negatives

LR– thus reflects how many times more likely a negative test result is for diseased than for non-diseased subjects. As for LR+, this is transferable across populations regardless of the disease prevalence.

A rule of thumb is that a LR+ over 10 and a LR– less than 0.1 represents a highly useful test because it alters the pre- to post-test probability by a multiple of 10.

ROC curves

ROC methodology was first developed in the 1950s in the context of radar signal detection[208]. In short, a ROC curve illustrates the trade-off between the sensitivity and specificity at every possible threshold (or cut-off) of a continuous or discrete variable. Statisticians find it hard to understand why clinicians invariably like to ruin beautiful continuous variables by dichotomisation (i.e. introducing a cut-off that separates test positives from test negatives). On the other hand, this is often tempting from the clinician's point of view, in order to facilitate decision making. A ROC-curve describes the *discriminating* capacity of a test across all possible cut-offs.

ROC curve. The ROC curve is plotted in a coordinate system with sensitivity on the y-axis and 1-specificity on the x-axis (Fig. 5). For each possible threshold, the tied sensitivity and specificity are plotted. Provided that a test value above the threshold is considered a positive test, a very low threshold will yield a very high sensitivity at the expense of specificity (as illustrated by point "E", Fig. 5). For each step that we increase the threshold, the sensitivity will drop and the specificity will increase (point "D" would be the extreme scenario). Curve "B" is created by plotting 13 ties of sensitivity and specificity, indicating that it reflects the discriminating capacity of a discrete ordinal variable with 13 scale steps (e.g. the AIR score) The position and shape of the ROC curve is determined by the

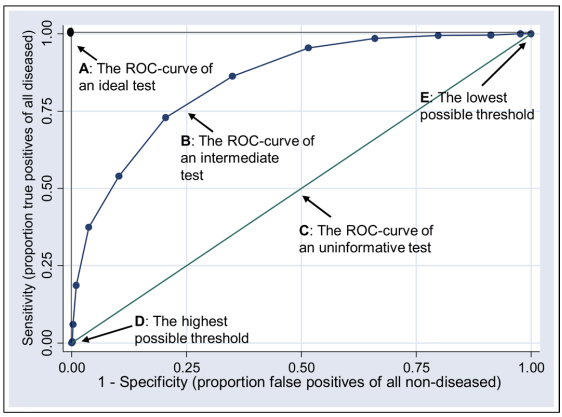


Fig. 5. Illustration of ROC curves

degree of separation, and by the degree of variability, of the test measurements for diseased and non-diseased subjects[209].

Area under the curve. A common way to assess a test's global accuracy is to calculate the area under the ROC curve (AUC). An AUC of 0.5 is obtained by a diagonal line from origo to the upper right corner of the coordinate system, and is the result of a completely useless test (as in curve "C", Fig. 5). In contrast, an AUC of 1.0 is obtained by the ideal test for which the sensitivity and specificity is 1.0 for each possible threshold (as in curve "A"). The direct interpretation of the AUC is that if we take a random diseased subject and a random non-diseased subject, the AUC corresponds to the probability that the diseased subject will have a higher test value than the non-diseased subject (assuming that large test values are indicative of the disease).

MISSING VALUES

In clinical research, data collection is universally imperfect in the sense that some variables will usually be missing for a number of study subjects[210]. Some study designs are more prone to "generate" missing values (e.g. retrospective studies) than others, but the clinical context and complexity and nature of the data will also have implications. The easiest, most common way of dealing with missing values is to perform complete-case analysis (i.e. analyse whatever is available). Another approach is to "impute" or "replace" the missing data, which should only be done if the dataset meets certain prerequisites with regard to missing pattern.

The pattern of missing data

There are three principal types of missing data patterns[211 212].

- Missing completely at random (MCAR): This uncommon pattern means that the individuals and variables with missing data are a pure random sample of the full sample. For instance, if the study subjects all roll a die and if the result is "1" they refrain from answering a survey question. This is an uncommon behaviour.

- Missing at random (MAR): This is by far the most common pattern of missingness assumed in medical data. It means that, in contrast to data MCAR, the mechanism of missingness is not pure random, but we can observe the variables that are associated with missingness. It is important to recognise that MAR can be assumed even though the process leading to missing values is NOT random like tossing a coin. MAR can be assumed as long as the variables leading to missingness are recorded. Let us suppose that patients arriving at the emergency department late at night are too tired to fill in the formula that contains a question about the duration of symptoms. This is clearly not completely random. However, if we have recorded the time of arrival of the patients at the emergency department, and the age, pain intensity and body temperature as well as other variables that can be associated with exhaustion, the probability that a variable is missing depends on information available in our database.

- Not missing at random (NMAR): This is when missingness depends on unobserved predictors or if it depends on the missing value itself. An example would be if blood samples are not drawn for children under 10 years, and age is not registered in our database, or abdominal tenderness is systematically not evaluated, or recorded, for patients complaining of abdominal pain.

Imputing of missing values

If we intend to use our collected data to illuminate the predictive properties of the variables, we are likely to employ regression analysis, which by default will disregard subjects with any missing predictor value. Unless the variable is MCAR this can induce bias and underestimation of standard errors, and regardless of missingness pattern, it will lead to loss of efficiency[212 213]. There are several ways of imputing missing values, but some (e.g. imputation of the mean of a continuous variable) will underestimate the variability between subjects, and this is suboptimal in that it is over-optimistic. A more sophisticated method of imputing missing values is multiple imputation (MI), which will be briefly explained in the following paragraph. It requires the missing values to be MAR or MCAR.

Multiple imputation. Briefly, MI takes into account all available data, and takes advantage of the correlation of the missing variable with all variables included in the imputation model. Even variables that seem extraneous, known as auxiliary variables, should be included in the model, because they may be correlates of missingness[214]. A failure to include such variables will hinder a MAR assumption as described above. The first step in MI is to replace missing values and create the first imputed dataset using a series (iterations) of multiple regression equations in forward (imputing) and posterior steps (finding new estimates for the parameters in the model, deliberately inducing deviation from the previous estimate). This is repeated *m* times, so that *m* number of imputed datasets are created. Secondly, the analysis chosen for making inference on our material is conducted at each imputed dataset, treating each of them as a complete datasets. Finally, the estimates from the *m* analyses are combined while taking into account the combined variance of the estimates within each imputed dataset *and* across all *m* datasets[212 214].

BOOTSTRAP

Baron Münchhausen used his own bootstraps to pull himself out of a swamp according to an old tale. This would, if it was true, clearly violate some of the basic laws of nature. Bootstrap resampling, named after the tale, is a non-parametric technique that draws new samples of subjects (or values) from the original sample, while replacing the subject last selected. So, if we have a sample of five subjects (a, b,

c, d, e), a new sample of five subjects is drawn from it. But as we replace each selected subject immediately, each subject can be selected from 0–5 times; the first resample could for example comprise study subjects a, a, b, d, d, the second could comprise b, b, d, e, e, and so on. This is typically repeated thousands of times, and the statistic of interest is calculated from each resample. The idea is that the estimates will vary in each resample, and the distribution is used to construct confidence intervals. Bootstrapping can be used for a variety of purposes, e.g. to calculate standard errors, bias-corrected standard deviations and confidence intervals and estimation of the optimism in the performance estimation of prediction models[212]. In this thesis, we have used bootstrapping with the intent to create bias-corrected confidence intervals only.

AIMS OF THE THESIS

- To construct and validate a clinical appendicitis score.

- To investigate whether the inclusion of novel inflammatory markers improves the diagnostic properties of the AIR score.

- To test the hypothesis that the implementation of an AIR-score-based clinical algorithm for patients with suspected appendicitis can improve diagnostic accuracy and reduce the use of diagnostic imaging and hospital admissions.

- To test the hypothesis that the routine use of imaging will decrease the number of negative appendectomies at the cost of an increase in operations for appendicitis that can resolve spontaneously, in comparison with in-hospital observation followed by selective imaging.

PATIENTS AND METHODS

OVERVIEW

An overview of the study design, participants, setting and outcomes is presented in Table 1.

Table 1. Study overview

	Study	Design	Participants	Setting	Comparison	Main outcome measures
	I	Cross-sectional observation study	Patients admitted for suspected appendicitis 1992–93, and 1997 (n=751)	Hospitals in Jönköping, Eksjö, Motala and Kalmar, Sweden	N/A	Sensitivity/specificity and predictive values ROC area for the AIR score compared with the Alvarado score
	II	Cross-sectional observation study	Patients presenting at ED with suspected appendicitis 2003–2005 (n=432)	Hospitals in Jönköping and Linköping, Sweden	AIR score vs extended AIR score	Sensitivity/specificity and likelihood ratio ROC area for the AIR score alone and combined with novel inflammatory markers
The STRAPPSCORE study	III	Single arm, pre-post interventional multicentre study	Patients presenting at ED with suspected appendicitis 2009–2012 (n=4320)	25 hospitals in Sweden	AIR-score-based clinical algorithm vs routine clinical management	Effect on diagnostic accuracy Use of imaging Number of hospital admissions
	IV	Nested unrestricted 1:1 randomised controlled multicentre study	Patients presenting at ED with suspected appendicitis 2010–2012 (n=1383)	21 hospitals in Sweden	Routine early, imaging vs in-hospital observation and selective imaging	Number of negative explorations Operations for perforated and non-perforated appendicitis. Number of patients treated for appendicitis

ED Emergency department

STUDY DESIGN

Studies I and II are cross-sectional observation studies. The STRAPPSCORE study consists of a single arm, pre-post interventional multicentre study (study III) and a nested unrestricted 1:1 randomised controlled multicentre study (study IV)

PATIENTS AND SETTING

In study I, consecutive patients, 10 years and older, admitted for suspected appendicitis to the county hospitals in Jönköping and Kalmar, and to the general hospitals in Eksjö and Motala, Sweden, were included. In study II and in the STRAPPSCORE study, patients presenting at the emergency department with suspected appendicitis, were considered for inclusion. Study II was conducted at the University Hospital in Linköping and the County Hospital Ryhov in Jönköping, Sweden, from 1 December 2003 to 17 August 2005. Studies III and IV were conducted at 21 and 25 Swedish hospitals, respectively, during the period of 1 September 2009 to 1 January 2012. Study III started with an initial baseline period, followed by an intervention period, in which the randomised study IV was nested.

Exclusion criteria

Pregnancy and prior appendectomy were exclusion criteria in all studies. In studies II–IV, patients with pain duration of more than five days were not considered for inclusion. In Studies III and IV, patients younger than five years were not eligible for inclusion.

METHODS

Data collection

Studies I–IV

Details regarding disease history (anorexia, vomiting, right lower quadrant pain and pain migration), clinical findings (muscular defence, rebound tenderness and body temperature), gender, and laboratory variables (leucocyte count, PMN or neutrophil count and CRP) were prospectively recorded in a study protocol on admission to hospital at the emergency department by the attending physician. A repeat examination was carried out for patients who were observed at the ward after a period of four to eight hours. The AIR score variables at each examination were recorded in the study protocol, and in study II blood samples for later analysis were drawn. At discharge, information on clinical diagnosis, type and time of surgical procedures and use of imaging and antibiotics (studies III–IV) was recorded.

Study III–IV

During studies III–IV, the completed protocols were transferred to a web-based database. The primary investigator checked the quality of registrations regularly and gave feedback to the responsible investigator at each hospital when any defects were found.

Diagnosis

Non-operated patients, and patients who had a negative exploration, were classified as no appendicitis. Participating centres were instructed to send all excised appendices for histopathologic diagnosis. "Phlegmonous appendicitis" was defined by transmural granulocyte infiltration. The criteria for "advanced appendicitis" were microscopic evidence of transmural gangrene, an operative finding of perforation or a localised abscess. "Perforated appendicitis" was defined by the operative (macroscopic) finding of perforation. "All appendicitis" refers to phlegmonous *or* advanced appendicitis. The histopathological diagnoses of the specimens in studies I and II were re-evaluated by one pathologist blinded to the score results, using a structured protocol. This also applies for the specimens of patients in the low- and high-risk groups of the STRAPPSCORE study.

Studies I–II

The histopathologic diagnosis of all specimens was re-evaluated by one consultant pathologist, blinded to the previous clinical and histopathologic diagnosis.

Studies III–IV

The same method of re-evaluation applied to the specimens of patients in the low- and high-risk groups of study III, whereas the histopathologic report was made according to a standardised protocol for patients in study IV, but no re-evaluation was done. Patients who needed a therapeutic operation in spite of an uninflamed appendix found at surgical exploration were classified as "other" diagnosis.

A non-operated patient that needed readmission and an operation for appendicitis at any Swedish hospital within seven days after the index admission was considered a *missed appendicitis* and the outcome of the patient was changed according to appendectomy diagnosis.

Study IV.

Antibiotic-treated appendicitis was defined as a non-operated appendicitis that was verified by imaging and treated with antibiotics. *Non-treated appendicitis* was defined by unequivocal findings of appendicitis at diagnostic imaging in a patient who received no surgical or antibiotic treatment.

Biochemical analyses

Biochemical inflammatory markers used in routine healthcare (WBC, PMN, neutrophils and CRP) were analysed by routine methods applied at the hospital in question at the time of the study.

In study II, the blood samples were centrifuged within one hour after sampling. Serum and plasma were stored at –70°C. Six inflammatory markers with potentially high discriminating capacity (interleukin[IL]-6, chemokine C-C motif ligand[CCL]2,

matrix metalloproteinase[MMP]-9, chemokine C-X-C motif ligand[CXCL] 8, serum amyloid A[SAA] and myeloperoxidase[MPO]) were analysed using a multiplex bead array (LUMINEX). Details of this method have been published elsewhere[147].

Construction of the AIR score and the extended score (studies I–II)

Multivariable logistic regression analyses

The variables with an independent diagnostic value in a previous study[124], and the Alvarado score variables[189] of the randomly selected construction sample (n=316) in study I, were entered into a weighted ordered multivariable logistic regression model. The outcome was coded as 0 for no appendicitis, 1 for phlegmonous appendicitis and 2 for advanced appendicitis, with the weight of 1 for no appendicitis, 2 for phlegmonous and 5 for advanced appendicitis. Continuous variables were stratified into 10 intervals. Dummy variables representing adjacent intervals with similar regression coefficients were successively manually combined, until a set of intervals with distinct and significant regression coefficients were obtained. Variables and intervals with a p-value <0.10 were kept in the final model. The regression coefficients of the final, most parsimonious model were used as scoring points. The same methodology was applied in study II, with the exception that the logarithm was used for non-normally distributed variables and that the AIR score variables were combined with one, two and finally all new inflammatory markers.

AIR score. The AIR score was constructed by rounding up the regression coefficients to the nearest integer. The exceptions were right iliac fossa pain and light muscular defence/rebound tenderness which were both rounded down to obtain a trend of increasing risk of appendicitis with increasing intensity of this variable (Table 3).

Extended score. The same method was applied in study II, where an extended version of the AIR score with the integration of the inflammatory marker CCL-2 was constructed.

Cut-off levels

For the AIR score, as well as for the Alvarado (study I), and the extended score (study II), we defined a "low" and "high cut-off" to obtain three diagnostic test zones; one "low-risk zone" with high sensitivity for detecting advanced appendicitis, used to rule out advanced appendicitis. One "high-risk zone" for detecting appendicitis of any severity with high specificity, used to rule in appendicitis, and finally an "intermediate-risk zone".

Validation of the AIR score (studies I–III)

The internal validation was conducted on the randomly selected validation sample of 229 patients in studies I–II. The external validation was performed on the patients included during the baseline and intervention periods of study III. The discriminating capacity and predictive values were determined for detecting all appendicitis and advanced appendicitis separately. The cut-off levels for the Alvarado and the extended scores were chosen to match the sensitivity and specificity of the corresponding cut-off levels for the AIR score.

Outcome measures

Discriminating capacity. The overall discriminating capacity of the AIR score, Alvarado score (study I) and the extended score (study II) were compared by the corresponding AUC for detecting all appendicitis and advanced appendicitis. Given the larger sample in study III, comparisons of the AUC between predefined subgroups of patients were also performed.

Diagnostic accuracy and predictive value. The sensitivity and specificity were determined for the high and low cut-off for the AIR score (studies I–III), as compared with the Alvarado score (study I) and the extended score (study II). In addition, the positive predictive value was calculated in study I and the likelihood ratios were calculated in studies II–III. Also, for all scores, the proportion of patients with inconclusive score results was determined.

Interventions of the STRAPPSCORE study

Study III

Baseline period. The AIR score parameters were registered prospectively, but the AIR score criteria were not provided to the clinician, thus the AIR score was not determined. The patients were managed according to the local conventional standard of care.

Intervention period. During the intervention period, the AIR score sum was calculated at the emergency department, and an AIR-score-based clinical algorithm was provided to the clinician (Fig. 6). The management of the patient in accordance with the algorithm was not imperative, but the physician was asked to note the reason for non-compliance in the study protocol.

For patients with an AIR score <5 ("low risk group"), the algorithm proposed outpatient management, with a follow-up within 24 hours. For patients with an AIR score of >8 points ("high risk group") the algorithm proposed surgical exploration.

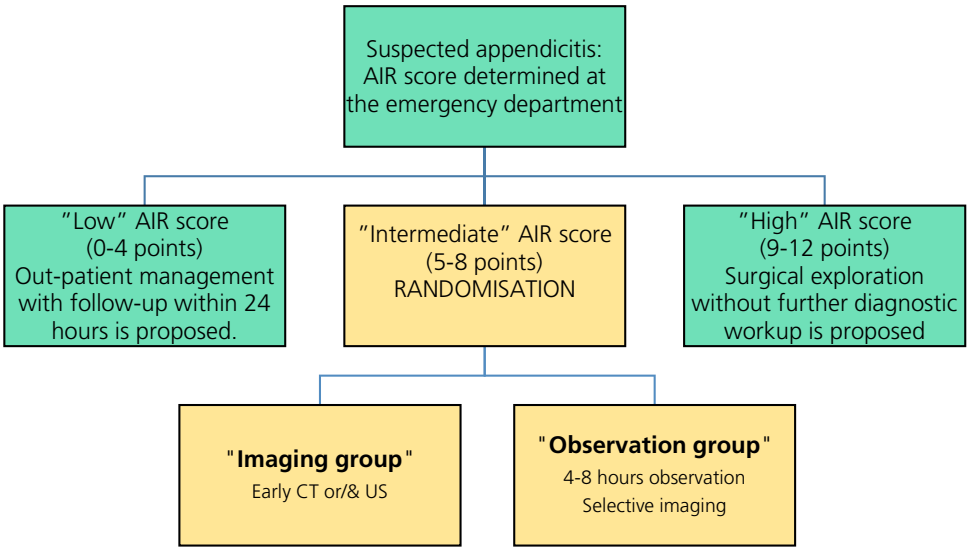


Fig. 6 The Intervention period of the STRAPPSCORE study, i.e. study III (■), and study IV (■). Management algorithm based on AIR score at the emergency department

Study IV

Patients with intermediate AIR score results (5–8 points) were randomised using thoroughly shuffled sealed opaque envelopes, between early diagnostic imaging (US or CT); "Imaging group" or observation with clinical re-assessment with selective imaging for patients with persistent equivocal findings; "Observation group" (Fig. 6). For both groups, the decision to admit, operate or discharge the patient, and the choice of surgical method, was left to the discretion of the surgeon.

Imaging group. The patient was immediately referred for an abdominal CT or US, which was performed in accordance with the local standard protocols, as soon as possible, by the radiologist on call.

Observation group. For patients in the Observation group, an AIR-score-based re-evaluation was scheduled in four to eight hours. If the score decreased to less than five points, the protocol proposed discharge of the patient. In contrast, for patients with an increase of the score to more than eight points, surgical exploration was proposed. For patients with a persisting score of 5–8 points, further observation, imaging or laparoscopy was suggested.

Outcome measures of the STRAPPSCORE study

Study III

The outcome measures in this study were the proportions of negative appendectomies and patients operated for phlegmonous and advanced appendicitis, the proportion of patients admitted for 24 hours or more and the proportion of patients that returned to the emergency department or were readmitted within 30 days after the index admission. All outcomes were analysed for the high- and low-risk groups only.

Study IV

The primary outcomes were the number of patients treated for appendicitis, and the number of negative explorations and operations for perforated and non-perforated appendicitis. Secondary outcomes were the number of admissions, time from arrival to operation, duration of stay, number of missed appendicitis and readmissions for any reason within 30 days after discharge.

Follow-up

Studies I-II. A follow-up after one month (study I) and six months (study II) of all non-operated patients was conducted by reviewing the patients' files. None had developed appendicitis during the follow-up period.

Studies III-IV. All patients included in the STRAPPSCORE study were followed-up for a minimum of 30 days through linkage of their Swedish national identification number with the Swedish national patient register.

Statistical methods

Statistical analysis was performed using Stata 7–13, StataCorp. 2000–2014. *Stata Statistical Software: Release 7–13* College Station, TX: StataCorp LP.

Study I

Differences in proportions were analysed using the chi-square test and Fisher's test, as appropriate. Differences in normally distributed continuous variables were analysed using Student's *t* test. The sensitivity and specificity were compared between the scores with McNemar's test.

Study II

Some 135 missing values, distributed in 95 patients, were judged missing at random, and were imputed by multiple iterated chained equations, using the "ICE" command in STATA 11[215]. The results of all analyses for the imputed datasets (n=20) were combined according to Rubin's rule[211]. Bootstrapping was used in order to get bias-corrected confidence intervals. In addition, complete case analyses were performed for comparison. The two proportion Z-test was used to compare differences between the scores in diagnostic accuracy, AUC and proportions of patients with indeterminate score points.

Studies III-IV

Comparisons between groups. Categorical variables were compared by Fisher's exact test or chi-square test, as appropriate. Differences in means of normally and non-normally distributed continuous variables were compared using t-test or Mann-Whitney-U test. The comparison of sensitivity and specificity for subgroups of patients in study III was made with chi-square test. For comparisons across more than two ordinal groups, we used the chi-square statistic for the trend. ROC areas were compared using the "roccomp" command in STATA. All outcomes in study IV were analysed according to intention-to-treat.

Sample size calculation. Assuming a decrease of negative explorations from 15% to 10% in the Imaging group, and accepting a two-sided alpha error of 0.05 and a beta error of 0.20, a target enrolment of 686 patients in each arm was required in study IV. We also hypothesised that routine imaging would increase the number of operations for non-perforated appendicitis. An increase of 20% would require 580 patients in each arm, with the same alpha and beta error levels applied. In study I, 37% of the patients had an intermediate AIR score. Thus, in order to randomise 1372 patients from the STRAPPSCORE study, we would have to enrol a minimum of 3708 patients overall.

Meta-analysis. In study IV, the combined results of previously published randomised controlled studies are presented as forest plots. The effect estimates were analysed using the Mantel-Haenszel method in *Review Manager (RevMan) [Computer program]. Version 5.3*. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration, 2014.

ETHICS

Study I

For this study, which was based on the anonymised database of a previously published observational study published in 1999[124], we did not apply for ethic committee approval as the data were analysed, the results were compiled, and the manuscript was drafted before the Swedish act concerning the Ethical Review of Research Involving Humans (2003:460) was implemented on 1 January 2004.

Studies II–IV

The studies were approved by the regional ethics committee at the University of Linköping, Sweden. Written informed consent was a prerequisite for inclusion in the study.

RESULTS

DEMOGRAPHIC OVERVIEW

The basic demographic characteristics of studies I–IV are compiled in Table 2.

Table 2. Overview: Study demographics and appendicitis prevalence

	Study		Age		Sex			Appendicitis prevalence
			Mean		Women	Men	*(% Women)*	(%)
	I	Construction sample	25.9		171	145	*(54.1)*	36.4
		Validation sample	23.4		124	105	*(54.1)*	33.2
			Median	*(range)*				
	II		21.0	*4–84*	203	225	*(47.4)*	41.4
STRAPPSCORE	III	Baseline	26.7	*5–95*	666	486	*(57.8)*	32.1
		Intervention	25.3	*5–96*	1386	1253	*(52.5)*	36.1
		p-value	0.72				0.003	0.019
	IV	Observation	26.1	*5–90*	251	274	*(47.8)*	46.3
		Imaging	28.3	*5-92*	264	279	*(48.6)*	55.4

EXCLUDED PATIENTS

Study I. Patients with incomplete datasets (n=206) were not included in the analysis. Hence, the complete datasets of 545 patients were analysed in this study.

Study II. Four patients with more than three missing AIR score parameters were excluded. The remaining 428 patients were included in the analysis. Missing values for these patients were imputed.

Study III. In total, 529 patients with incomplete datasets, for whom no AIR score sum was recorded, were excluded from analysis.

Study IV. The Consolidated Standards of Reporting Trials (CONSORT) diagram is presented in Fig. 7. In this study, 244 eligible patients were excluded due to non-randomisation or lack of informed consent. Furthermore, one centre (n=71) was excluded from analysis due to consistently low registration quality and non-compliance with randomisation status.

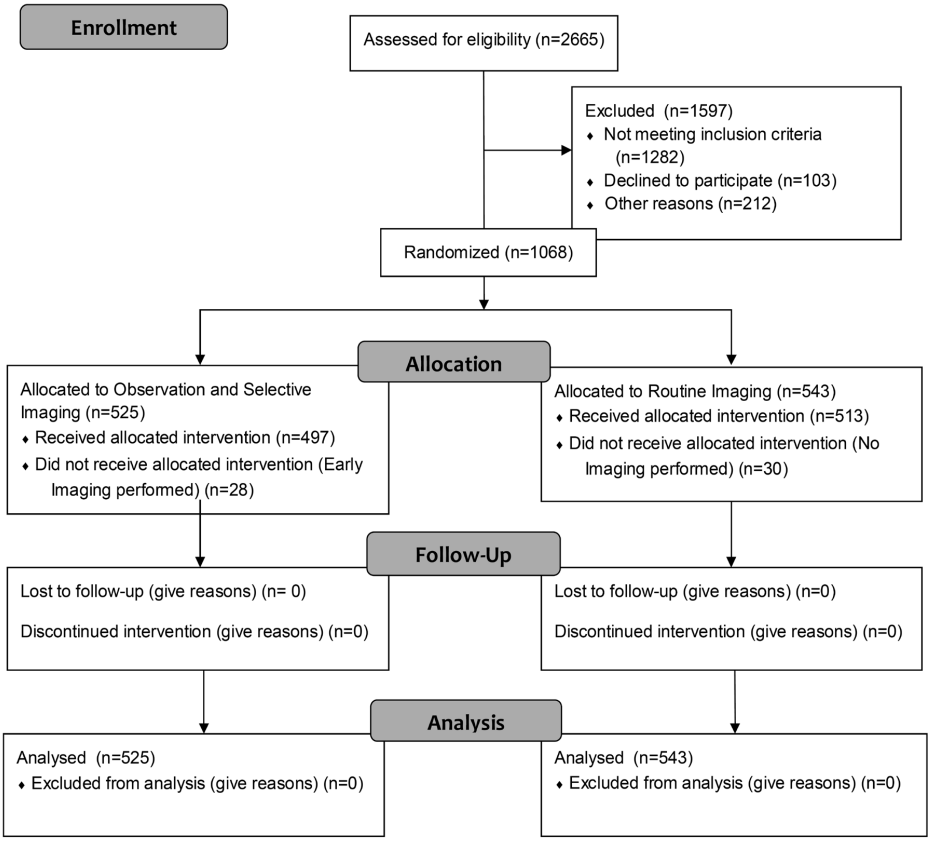


Fig. 7. CONSORT diagram

STUDY I

Demographic characteristics are presented in Table 2. In total, 250 of the 545 patients (45.9%) underwent surgery. The prevalence of appendicitis was 36.4% and 33.2% in the construction and validation sample, and the prevalence of advanced appendicitis was 13.9% and 13.1%, respectively. The proportion of negative appendectomies was 23.8% in the construction sample and 23.2% in the validation sample.

Construction of the score

The regression score, simplified (i.e. the AIR-) score and the Alvarado score are presented in Table 3. Eight variables reflecting peritoneal irritation and inflammatory response were kept in the final model (vomiting, right-lower-quadrant pain, rebound tenderness, muscular defence, body temperature, WBC, proportion PMN and CRP concentration). Migration of pain, nausea and male gender were not included as they had a p-value of >0.1 in the weighted ordered logistic regression analysis. The AIR score has a range from 0–12 points.

Table 3. The regression score, simplified (AIR) score and Alvarado score.

		Proposed score		Alvarado
		Regression	Simplified (AIR)	score
Relocation of pain		ns	–	1
Vomiting		0.45	1	1
Pain in RIF		1.12	1	2
Anorexia		ns	–	1
Male sex		ns	–	–
Rebound tenderness or muscular defence	None	0	0	0
	Light	1.54	1	1
	Medium	1.90	2	1
	Strong	2.32	3	1
Body temperature	37.5–37.9	0	0	1
	38.0–38.4	0	0	1
	≥38.5	0.85	1	1
Proportion PMN	70–74%	0.92	1	0
	75–84%	0.92	1	1
	≥85%	1.41	2	1
WBC count	10.0–14.9*10^9/L	0.96	1	2
	≥15.0*10^9/L	1.46	2	2
CRP concentration	10–49 mg/L	1.04	1	–
	≥50 mg/L	2.35	2	–
Range		–	0–12	0–10

RIF Right inferior fossa

Validation of the score

The regression-based score and the AIR score were evaluated and compared with the Alvarado score on the validation sample of 229 patients. In all analyses, the AIR score performed at least as well as the regression-based score. Therefore, the results are presented for the AIR score, as compared with the Alvarado score only.

Discriminating capacity

The AIR score had better discriminating capacity for all appendicitis and advanced appendicitis than the Alvarado score (AUC 0.93 versus 0.88 and 0.97 versus 0.92, respectively). We found no difference in the AIR score's discriminating capacity with regard to age or sex.

Diagnostic accuracy and predictive value

The diagnostic properties of the scores with regard to the low and high cut-off, for all and advanced appendicitis are presented in Table 4.

As previously mentioned, the low cut-off for the Alvarado score was chosen to match the corresponding cut-off for the AIR score with regard to the sensitivity for appendicitis, and the high cut-off to match the AIR score's specificity for appendicitis. Thus, the sensitivity at the low cut-off and the specificity at the high cut-off were similar for the two scores.

The specificity for appendicitis at the low cut-off was higher for the AIR score than for the Alvarado score (0.73 vs 0.61), which also applied to the sensitivity for advanced appendicitis at the high cut-off (0.67 vs 0.40).

The predictive values for the AIR score and the Alvarado score were similar. The negative predictive value of an AIR score <5 points was 0.97 for all- and 1.0 for advanced appendicitis. The positive predictive value of an AIR score >8 points was 0.97 for all appendicitis and 0.95 for advanced appendicitis.

A smaller proportion of patients was classified to the intermediate risk group by the AIR score compared with the Alvarado score (37.1% and 48.0%, respectively). Consequently, compared with the Alvarado score, the AIR score classified a larger proportion of patients to either the high-risk group or the low risk group (62.9% vs 52.0%), with similar diagnostic accuracy (0.97 and 0.98, respectively).

Table 4. Diagnostic characteristics of the AIR score according to the cut-off points compared with the Alvarado score

	Cut-off point			
	AIR score		Alvarado score	
Diagnostic value	>4points	>8 points	>4 points	>8 points
All appendicitis				
Sensitivity	0.96	0.37	0.97	0.28
Specificity	0.73†	0.99	0.61†	0.99
PV+	0.64	0.97	0.56	0.91
NV–	0.97	0.76	0.98	0.73
Advanced appendicitis				
Sensitivity	1.00	0.67‡	1.00	0.40‡
Specificity	0.73†	0.99	0.61†	0.99
PV+	0.42	0.95	0.34	0.86
PV–	1.00	0.94	1.00	0.89

† $p<0.001$, ‡ $p<0.005$

PV+ positive predictive, *PV*– negative predictive value

STUDY II

Basic demographic characteristics are presented in Table 2. The prevalence of appendicitis was 41.4% in this study. The prevalence of phlegmonous and advanced appendicitis was 25.5% and 15.9%, respectively. Of all 206 operated patients, 29 (14.1%) had a negative appendectomy. Four patients were excluded due to multiple missing values. Of the remaining 428 patients, a total of 135 missing values distributed in 95 patients were imputed.

Discriminating capacity of new inflammatory markers

Univariable analysis. The discriminating capacity of the new inflammatory markers, as well as the AIR score variables, were tested for both advanced appendicitis and all appendicitis in weighted ordered logistic regression models, and presented as AUC. The leukocyte count was the strongest discriminator for advanced appendicitis (AUC 0.89, CI 0.84–0.93) whereas defence or rebound tenderness was the strongest discriminator for all appendicitis (AUC 0.84, CI 0.80–0.88). Among the new inflammatory markers, SAA was the strongest discriminator for advanced appendicitis (AUC 0.80, CI 0.75–0.85) and SAA, MPO and MMP-9 were the strongest discriminators for all appendicitis (AUC 0.71)

Multivariable analysis. The discriminating capacity for the new inflammatory variables when combined with the AIR score variables is presented in Table 5. Although the discriminating capacity was similar, the combination of CCL-2 and the AIR score variables had the highest point estimate, taking both advanced and all appendicitis into account (AUC 0.98 [CI 0.96–0.99] and 0.93 [CI 0.90–0.95], respectively). The combination of the AIR score variables with more than one new inflammatory marker did not increase the discriminating capacity further.

Table 5. Discriminating capacity of the new inflammatory markers combined one by one with the AIR score variables, expressed as the mean ROC areas for all appendicitis and advanced appendicitis.

	All appendicitis		Advanced appendicitis	
	AUC	CI(95%)	AUC	CI(95%)
AIR score	0.92	*0.90–0.95*	0.97	*0.95–0.99*
AIR+SAA	0.92	*0.90–0.95*	0.97	*0.96–0.99*
AIR+MPO	0.93	*0.90–0.95*	0.97	*0.96–0.99*
AIR+MMP9	0.92	*0.90–0.95*	0.97	*0.95–0.99*
AIR+CCL2	**0.93**	***0.90–0.95***	**0.98**	***0.96–0.99***
AIR+CXCL8	0.92	*0.90–0.95*	0.97	*0.96–0.99*
AIR+IL6	0.92	*0.90–0.95*	0.97	*0.96–0.99*

Construction of the extended score

In a previous study, we found that CCL2, SAA and IL-6 had the highest independent discriminating capacity for appendicitis[147]. We now tested extended versions of the AIR score by the inclusion of these biochemical markers one by one. The versions including SAA and IL-6 did not perform as well as the one including CCL2. Therefore, only the results of the extended version including CCL2 are presented in the following.

Validation of the extended score

Discriminating capacity.

The discriminating capacity of the AIR score and the extended score (range 0-14 points) for all appendicitis and advanced appendicitis was similar (AUC of 0.91 [95% CI 0.89–0.94] vs 0.92 [95% CI 0.89–0.95], p=0.60 and 0.96 [95% CI 0.94–0.98] vs 0.97 [95% CI 0.95–0.99], p=0.43).

For clarity, the ROC curves from one imputed dataset, representing the discriminating capacity for all appendicitis and advanced appendicitis of the AIR and Extended scores, are presented in Fig. 8.

Fig. 8 a) Comparison of the ROC curve for diagnosing all appendicitis for the AIR score and the extended score including CCL2.

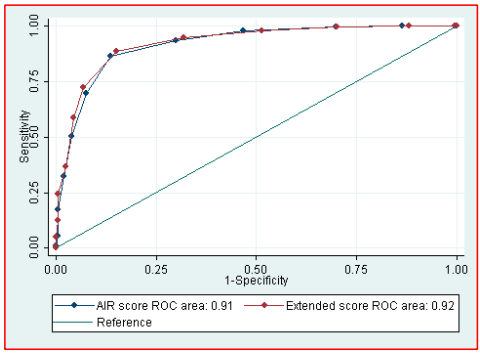
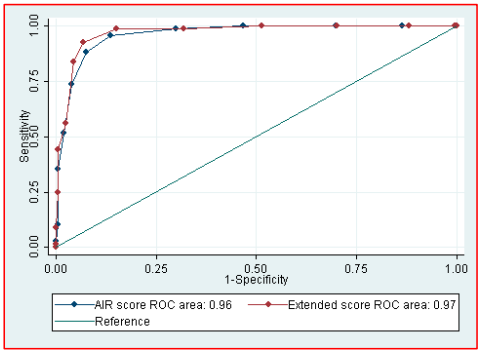


Fig. 8 b) Comparison of the ROC curve for diagnosing advanced appendicitis for the AIR score and the extended score including CCL2.



Diagnostic accuracy

The diagnostic properties of the AIR score and extended score are presented in Table 6. In accordance with the validation of the AIR score in study I, a low and high cut-off point was selected for the extended score (>5 and >9 points). As described previously, these were chosen to match the sensitivity and specificity of the AIR score at the low and high cut-off to allow comparison of the scores' diagnostic properties. We found no differences between the AIR score and the extended score with regard to sensitivity, specificity, likelihood ratios, or the proportion of patients assigned to the high, low or intermediate probability group of having appendicitis.

Table 6. Comparison of the diagnostic properties of the AIR score and the extended score including CCL2.

	All appendicitis						Advanced appendicitis					
	Sensitivity	*CI(95%)*	Specificity	*CI(95%)*	LR+	LR–	Sensitivity	*CI(95%)*	Specificity	*CI(95%)*	LR+	LR–
Low cut-off												
AIR score>4	0.93	*0.89-0.97*	0.71	*0.65-0.77*	3.2	0.1	0.99	*0.96-1.0*	0.71	*0.65-0.77*	3.4	0.01
Extended score >5	0.95	*0.92-0.98*	0.69	*0.63-0.75*	3.1	0.07	0.99	*0.96-1.0*	0.69	*0.63-0.75*	3.2	0.01
High cut-off												
AIR score>8	0.31	*0.23-0.38*	0.98	*0.96-1.00*	15.5	0.7	0.49	*0.36-0.62*	0.98	*0.96-1.00*	24.5	0.5
Extended score >9	0.37	*0.30-0.45*	0.97	*0.95-1.00*	12.3	0.7	0.56	*0.44-0.69*	0.97	*0.95-1.00*	18.7	0.5

STUDY III

In this study, a total of 3791 patients from 25 Swedish hospitals had a full set of AIR score parameters recorded, and were included in the analysis. Demographic characteristics are presented in Table 2. The median duration of symptoms was 24 (interquartile range, IQR, 24 to 48) hours and the mean AIR score of the patients was 5.0 (standard deviation, SD, 2.46) points.

The AIR score assigned 41.9% (1590) of the patients to the low-risk group and 8.8% (333) patients to the high-risk group. The remaining 49.3% (1868) of the patients were assigned to the intermediate risk group. Overall, the prevalence of appendicitis was 34.9%, and the proportion of operated patients that had negative findings was 11.5% (Table 7).

Table 7. The distribution of diagnoses according to AIR score.

AIR score	No appendicitis				Appendicitis				Other		Total	
	Non-operated		Operated negative		Phlegmonous		Advanced					
	n	%	n	%	n	%	n	%	n	%	n	%
Low (0–4p)	1416	*89.1*	36	*2.3*	110	*6.9*	25	*1.6*	*3*	*0.2*	1590	*100*
Intermediate (5–8p)	823	*44.1*	123	*6.6*	579	*31.0*	331	*17.7*	*12*	*0.6*	1868	*100*
High (9–12p)	34	*10.2*	15	*4.5*	64	*19.2*	213	*64.0*	*7*	*2.1*	333	*100*
Total	2273	*60.0*	174	*4.6*	753	*19.9*	569	*15.0*	*22*	*0.6*	3791	*100*

Discriminating capacity

The discriminating capacity of the AIR score is presented in Table 8. The AIR score has higher discriminating capacity for detecting advanced appendicitis than for all appendicitis (AUC 0.89 vs 0.83, respectively). The discriminating capacity for all appendicitis is higher for women (AUC 0.84), for all and advanced appendicitis for patients under the age of 15 years (AUC 0.87 and 0.94, respectively), and for patients with a symptom duration of 36 hours or more (AUC 0.87 and 0.92, respectively). We found no difference in the discriminating capacity of the AIR score according to the level of experience of the examining physician. The AUC was similar for all participating hospitals (data not shown).

Table 8. Discriminating capacity of the AIR score for all appendicitis and advanced appendicitis, according to subgroups

		Total	All appendicitis			Advanced appendicitis		
		n	n	AUC	(*95% CI*)	n	AUC	(*95% CI*)
Sex	Women	2052	547	0.84	*0.83–0.86*	255	0.90	*0.88–0.91*
	Men	1739	775	0.80	*0.78–0.82*	314	0.88	*0.86–0.90*
	p-value			0.004			0.17	
Age (years)	<15	527	181	0.87	*0.84–0.90*	63	0.94	*0.91–0.96*
	15–49	2686	892	0.83	*0.81–0.84*	345	0.89	*0.88–0.91*
	≥50	565	245	0.76	*0.73–0.80*	160	0.81	*0.77–0.85*
	p-value			<0.001			<0.001	
Duration of symptoms (h)	<12	804	236	0.79	*0.75–0.82*	62	0.83	*0.78–0.88*
	12–35	1375	577	0.80	*0.78–0.82*	221	0.86	*0.84–0.89*
	≥36	1109	281	0.87	*0.85–0.89*	168	0.92	*0.90–0.94*
	p-value			<0.001			<0.001	
Competence of examining physician	PRHO*	1522	556	0.82	*0.80–0.84*	231	0.88	*0.86–0.91*
	Surgical registrar	1094	370	0.83	*0.80–0.85*	165	0.89	*0.87–0.92*
	General surgeon	502	191	0.82	*0.78–0.86*	78	0.88	*0.84–0.92*
	Other	439	155	0.81	*0.77–0.85*	71	0.88	*0.83–0.92*
	p-value			0.96			0.83	
Total		3791	1322	0.83	*0.81–0.84*	569	0.89	*0.87–0.90*

*Pre-registration house officer/ Junior doctor

Predictive capacity

The diagnostic properties of the AIR score for subgroups, with regard to the low and high cut-off, are compiled in Table 9.

Low-risk group. The sensitivity for advanced appendicitis in the low-risk group is 0.96 and the negative likelihood ratio (LR–) is 0.068. We found no difference in this regard between women and men or different age groups. The sensitivity for advanced appendicitis does, however, increase with duration of symptoms (0.89 for duration <12hours, 0.96 for duration between 12 and 36 hours, and 0.98 for duration>36 hours).

High-risk group. The specificity for all appendicitis is 0.98, and the positive likelihood ratio (LR+) is 10.5 with no difference between men and women. The specificity is higher for younger patients (0.98 for <50 years and 0.93 for ≥50 years of age) and for patients with short duration of symptoms (0.99 for <12 hours and 0.97 for ≥12 hours duration of symptoms). The LR+ for advanced appendicitis at the high cut-off is 18.5. The highest LR+ was found for children under the age of 15 (LR+ 28.0), for patients with duration of symptoms less than 12 hours (LR+ 23.0) and for women (LR+ 19.0).

Table 9. The sensitivity, specificity and likelihood ratios of the high- and low-risk groups for subgroups

		Total	All Appendicitis					Advanced Appendicitis				
Low-risk group (<5p)		n	n	Sens	Spec	LR+	LR–	n	Sens	Spec	LR+	LR–
	Total	1590	135	0.90	0.59	2.20	0.17	**25**	**0.96**	**0.59**	**2.34**	**0.068**
Study period	Baseline	586	62	0.83	0.67	2.52	0.25	**7**	**0.95**	**0.67**	**2.88**	**0.075**
	Intervention	1004	73	0.92	0.55	2.04	0.15	**18**	**0.96**	**0.55**	**2.13**	**0.073**
	*p-value**			*<0.001*	*<0.001*				***0.80***	***<0.001***		
Sex	Women	977	53	0.90	0.61	2.31	0.16	**8**	**0.97**	**0.61**	**2.49**	**0.049**
	men	613	82	0.89	0.55	1.98	0.20	**17**	**0.95**	**0.55**	**2.11**	**0.091**
	*p-value**			*0.60*	*0.002*				***0.19***	***0.002***		
Age (yr)	5–15	252	18	0.90	0.68	2.81	0.15	**1**	**0.98**	**0.68**	**3.06**	**0.029**
	15–49	1167	94	0.89	0.60	2.23	0.18	**15**	**0.96**	**0.60**	**2.4**	**0.067**
	≥50	167	23	0.91	0.45	1.65	0.20	**9**	**0.94**	**0.45**	**1.71**	**0.13**
	*p-value***			*0.80*	*<0.001*				***0.21***	***<0.001***		
Duration of symptoms (h)	<12	378	38	0.84	0.60	2.10	0.27	**7**	**0.89**	**0.60**	**2.25**	**0.18**
	12–35	500	58	0.90	0.55	2.00	0.18	**9**	**0.96**	**0.55**	**2.13**	**0.073**
	≥36	504	27	0.92	0.63	2.49	0.13	**5**	**0.98**	**0.63**	**2.65**	**0.032**
	*p-value***			*0.002*	*0.20*				***0.006***	***0.20***		
High-risk group (>8p)		n	n	Sens	Spec	LR+	LR–	n	Sens	Spec	LR+	LR–
	Total	**333**	**277**	**0.21**	**0.98**	**10.50**	**0.81**	213	0.37	0.98	18.50	0.64
Study period	Baseline	**81**	**67**	**0.18**	**0.98**	**9.00**	**0.84**	56	0.38	0.98	19.00	0.63
	Intervention	**252**	**210**	**0.22**	**0.98**	**11.00**	**0.80**	157	0.37	0.98	18.50	0.64
	*p-value**			**0.11**	**0.28**				0.85	0.28		
Sex	Women	**159**	**128**	**0.23**	**0.98**	**11.5**	**0.79**	96	0.38	0.98	19.00	0.63
	men	**174**	**149**	**0.19**	**0.97**	**6.33**	**0.84**	117	0.37	0.97	12.33	0.65
	*p-value**			***0.066***	***0.39***				*0.93*	*0.39*		
Age (yr)	<15	**54**	**48**	**0.27**	**0.98**	**13.5**	**0.74**	35	0.56	0.98	28.00	0.45
	15–49	**190**	**162**	**0.18**	**0.98**	**9.00**	**0.84**	119	0.34	0.98	17.00	0.67
	≥50	**87**	**66**	**0.27**	**0.93**	**3.86**	**0.78**	58	0.36	0.93	5.14	0.69
	*p-value***			***0.58***	***<0.001***				*0.057*	*<0.001*		
Duration of symptoms (h)	<12	**26**	**23**	**0.10**	**0.99**	**10.00**	**0.91**	14	0.23	0.99	23.00	0.78
	12–35	**128**	**101**	**0.18**	**0.97**	**6.0**	**0.85**	71	0.32	0.97	10.67	0.70
	≥36	**142**	**121**	**0.35**	**0.97**	**11.67**	**0.67**	102	0.48	0.97	16.00	0.54
	*p-value***			***<0.001***	***0.016***				*<0.001*	*0.016*		

*p-value: Chi-square or Fisher's exact test as appropriate, **p-value: Test for trend regarding sensitivity and specificity

Effect of the AIR-score-based algorithm

Patients were managed according to the AIR-score-based algorithm during the intervention period of the STRAPPSCORE study. Routine clinical management was applied during the baseline period. The outcomes for the high- and low-risk groups were compared between the baseline and intervention periods (Table 10).

Low-risk group. The use of the AIR-score-based algorithm reduced the proportion of hospital admissions compared with routine management (29.5% vs 42.8%). Furthermore, the use of US and CT was lower (19.2% vs 34.5%) and the proportion of negative explorations and operations for phlegmonous appendicitis was smaller (1.6% vs 3.4%, p=0.019 and 5.5% vs 9.4%).

High-risk group. During the intervention period, a smaller proportion of patients had an US or CT done compared with the baseline period (38.5% vs 53.1%). We found no other differences in outcomes between the intervention period and the baseline period.

Follow up. We found no difference between the baseline and the intervention period regarding return to emergency department or readmissions within the 30-day follow-up period.

Table 10. Outcome measures for the low-risk and high-risk groups presented for the baseline and intervention periods

		Baseline period n=1152		Intervention period n=2639		p-value
Low-risk group (AIR score <5)	n (%)	n=586	*(50.9)*	n=1004	*(38.0)*	<0.001
AIR score	mean (*SD*)	2.59	*(1.21)*	2.72	*(1.15)*	0.049
Negative explorations	n (*%*)	20	*(3.4)*	16	*(1.6)*	0.019
AIR score among negative explorations	mean (*SD*)	3.50	*(0.69)*	3.56	*(0.82)*	0.54
Phlegmonous appendicitis	n (%)	55	*(9.4)*	55	*(5.5)*	0.003
Advanced appendicitis	n (%)	7	*(1.2)*	18	*(1.8)*	0.41
Admissions to hospital	n (*%*)	251	*(42.8)*	296	*(29.5)*	<0.001
Antibiotic treatment only	n (%)	3	*(0.5)*	12	*(1.2)*	0.28
Imaging procedures	n (%)	202	*(34.5)*	193	*(19.2)*	<0.001
Return to ED within 30d	n (%)	50	*(8.5)*	92	*(9.2)*	0.67
Readmissions within 30d	n (%)	9	*(1.5)*	22	*(2.2)*	0.45
Missed appendicitis	n (%)	1	*(0.17)*	4	*(0.40)*	0.66
High-risk group (AIR score >8)	n(%)	81	*(7.0)*	252	*(9.6)*	0.012
AIR score	mean (*SD*)	9.62	*(0.86)*	9.54	*(0.72)*	0.76
Negative explorations	n (*%*)	4	*(4.9)*	11	*(4.4)*	0.77
AIR score among negative explorations	mean (*SD*)	9.75	*(1.50)*	9.36	*(0.67)*	0.87
Phlegmonous appendicitis	n (%)	11	*(13.6)*	53	*(21.0)*	0.14
Advanced appendicitis	n (%)	56	*(69.1)*	157	*(62.3)*	0.27
Admissions to hospital	n (*%*)	80	*(98.8)*	247	*(98.0)*	0.66
Antibiotic treatment only	n (%)	4	*(4.9)*	17	*(6.8)*	0.79
Imaging procedures	n (%)	43	*(53.1)*	97	*(38.5)*	0.021
Return to ED within 30d	n (%)	8	*(9.9)*	15	*(6.0)*	0.23
Readmissions within 30d	n (%)	1	*(1.2)*	4	*(1.6)*	>0.99
Missed appendicitis	n (%)	0	*(0)*	2	*(0.79)*	>0.99

STUDY IV

During the period January 2010 to January 2012, 2665 patients were assessed for eligibility at 21 hospitals participating in the STRAPPSCORE study. Some 543 patients were randomised to the Imaging group and 525 patients were randomised to the Observation group (Fig. 7). The age and sex distribution, mean AIR score, level of inflammatory markers- and duration of symptoms were similar in the two groups. The non-randomised patients did not differ in any of these baseline characteristics or outcomes compared with the 1068 randomised patients.

Interventions. A CT and/or US was performed in 512 (94.3%) patients in the Imaging group and in 163 (31.0%) patients in the Observation group. The median time to imaging was shorter in the Imaging group (5.2 hours) than in the Observation group (12.3 hours).

Primary outcomes

The primary outcomes are presented in Table 11. More patients were diagnosed with appendicitis (301 (55.4%) vs 243(46.3%)) and more patients were treated for appendicitis with surgery, drainage or antibiotics, in the Imaging group than in the Observation group (290 (*53.4*%) vs 243 (*46.3*%)). More patients were operated for non-perforated appendicitis in the Imaging group (237, 43.6%) than in the Observation group (197, 37.5%). We found no difference between the Imaging and Observation group in the number of patients with an operation for perforated appendicitis, negative appendectomies, or operations for other conditions.

Table 11. Primary outcomes and diagnoses for patients assigned to the Observation- and Imaging groups

Outcome	Observation group n=525		Imaging group n=543		p-value
	n	%	n	%	
Appendicitis diagnosis	243	(*46.3*)	301	(*55.4*)	0.003
Treated for appendicitis	243	(*46.3*)	290	(*53.4*)	0.020
Operated	233	(*44.4*)	275	(*50.6*)	0.040
–Non-perforated	197	(*37.5*)	237	(*43.6*)	0.042
–Perforated	36	(*6.9*)	38	(*7.0*)	0.93
Percutaneous drainage	2	(*0.4*)	1	(*0.2*)	0.62
Antibiotics treated	8	(*1.5*)	14	(*2.6*)	0.28
Non-treated appendicitis*	0	(*0*)	11	(*2.0*)	0.001
Non appendicitis diagnoses	282	(*53.7*)	242	(*44.6*)	0.003
Operated	40	(*7.6*)	36	(*6.6*)	0.53
Negative appendectomy	34	(*6.5*)	35	(*6.4*)	0.98
–Other	6	(*1.1*)	1	(*0.2*)	0.07
Non-operated	242	(*46.1*)	206	(*37.9*)	0.007
–Diverticulitis	14	(*2.7*)	16	(*2.9*)	0.78
–NSAP**	181	(*34.5*)	127	(*23.4*)	<0.001
–Other	47	(*9.0*)	63	(*11.6*)	0.15
Any operation	273	(*52.0*)	311	(*52.3*)	0.083

*Patient with unequivocal signs of appendicitis at diagnostic imaging receiving no operative or antibiotic treatment. **Non-specific abdominal pain

Secondary outcomes

Median time from arrival to operation was shorter in the Imaging group (13.7 hours) than in the Observation group (15.5 hours). We found no difference between the groups with regard to the duration of hospital stay, number of patients admitted to hospital, or number of readmissions or return to the emergency department within the 30-day follow-up (Table 12).

Table 12. Secondary outcomes and diagnostic imaging for patients assigned to the Observation- and Imaging groups

	Observation group n=525		Imaging group n=543		p-value
Diagnostic imaging, n (%)					
–CT only	97	(*18.5*)	308	(*56.7*)	
–Ultrasonography only	58	(*11.0*)	166	(*30.6*)	
–CT and Ultrasonography	8	(*1.5*)	38	(*7.0*)	
–None	362	(*69.0*)	31	(*5.7*)	<0.001
Admissions					
Number of admissions (%)	475	(*90.5*)	480	(*88.4*)	0.27
Length of stay(d), Mean, (SD)	1.89	(*1.57*)	1.94	(*1.70*)	0.77
Durations h (IQR)					
Median time to first imaging	12.3	(*7.5–18.3*)	5.2	(*3.6–7.5*)	<0.001
Median time to surgery	15.5	(*10.9–23.3*)	13.7	(*9.0–21.7*)	0.009
30 day follow-up, n (%)					
Return to ED	81	(*15.4*)	87	(*16.0*)	0.79
Readmission	28	(*5.3*)	27	(*5.0*)	0.79
Return to ED after operation	18	(*3.4*)	14	(*2.6*)	0.42
Missed appendicitis *	1	(*0.2*)	3	(*0.6*)	0.63

*Appendectomy within 7 d after discharge from hospital

DISCUSSION

The main finding of this study is that the AIR score may be a valuable tool in selecting patients with suspected appendicitis for outpatient management or surgical exploration and identifying those that benefit from further diagnostic work-up. Adding new inflammatory variables did not improve the diagnostic efficiency of the AIR score. The use of a score-based management algorithm improved outcome for low-risk patients in terms of less negative explorations in spite of a reduced use of diagnostic imaging and in-hospital care without altering the number of patients with perforated appendicitis. Selective imaging in the intermediate risk group was associated with a lower number of patients that were treated for appendicitis compared with routine imaging, suggesting that spontaneous resolution occurred in uncomplicated cases.

The framework of test development and evaluation

Diagnostic tests used in medicine are typically developed with the intent to achieve better outcome for the patient or population, preferably in a cost-efficient way. The evaluation of a new test involves a number of steps starting with the assessment of technical accuracy; does the test provide useful information under highly controlled conditions[197 216 217]? After the "technical accuracy" of the test is established, its place in a diagnostic pathway is defined; is this a triage test, will it replace an existing test or will it be added on an existing pathway[197 218]? The evaluation continues with a focus on diagnostic accuracy, the test's ability to detect and exclude the target disorder in the population of interest[197 219]. The last two principal steps are assessment of the test's impact on patient outcome, and finally a cost-effectiveness analysis which represents a broader perspective of costs and benefits for society induced by the testing[188 197 217].

In the perspective of this thesis, the technical accuracy of the individual parameters in the AIR score was already evaluated in a previous study, and their correlation with appendicitis when combined was evaluated in study I[124]. The diagnostic accuracy was explored internally in study I and externally validated in studies II and III. The impact on patient outcome was evaluated in study III (low- and high-risk groups) and study IV (intermediate group). The last step, cost-effectiveness, is yet to be established based on the STRAPPSCORE study, but requires a longer follow-up than the 30 days reported in the study.

METHODOLOGICAL CONSIDERATIONS

Study design

Studies I and II

Studies I and II are cross-sectional studies, which is a reasonable option for evaluation of the association between test results and disease status[196]. Some aspects of cross-sectional study design merits attention, however.

A common design in cross-sectional diagnostic accuracy studies is the "case-referent" design, which requires the disease status to be known in order to select a number of cases and a convenient number of referents (non-diseased subjects). The disease prevalence of the sample can thus be chosen at the discretion of the study designer. While this design is appropriate for early phase studies of novel or potentially harmful tests, it does not provide answers as to the diagnostic properties of the test in real life[197]. The results are likely to be overoptimistic due to the fact that the test performance is studied in two extreme groups of non-diseased and diseased. In fact, in real life, such ideal circumstances would by definition mean that further testing could not possibly establish the correct diagnosis any better[219]. Nevertheless, many diagnostic accuracy studies of diagnostic scores, CT, US and laboratory parameters for diagnosing appendicitis are of case-referent design, which is inappropriate if claims regarding the test's properties in clinical practice are made.

We have studied cohorts of patients from relevant populations, namely patients with suspected appendicitis, presenting at the emergency department. However, study I is based only on patients with suspected appendicitis who were *admitted* to hospital, whereas in study II (and studies III–IV), patients were eligible for inclusion regardless of whether they were admitted or not. The former is not optimal as one can expect that the disease spectrum and other characteristics can differ between patients that are admitted to hospital and those that are not[188]. The external validation was performed on all patients with suspected appendicitis presenting at the emergency department, which we regard as the target population of the AIR score.

Study III

Study III is a single-arm pre-post interventional study. The term "single arm" refers to the fact that only one group of patients is studied, and "pre-post" refers to the fact that the outcome is studied before and after an intervention is introduced.[220] Considering that the aim of study III was to explore the effect of an intervention on outcome for patients with suspected appendicitis one can argue that a randomised controlled study would provide stronger evidence.

While this is true in theory, the nature of the intervention (i.e. the AIR-score-based management algorithm) raises some obstacles in this regard. First, a randomised parallel group design would most likely result in carry-over effects between groups. It is probably overoptimistic to believe that physicians would be able to completely disregard the AIR-score-based algorithm in patients randomised to routine clinical management. Second, in order to overcome carry-over effects, a cluster randomisation on hospital level could be performed. However, a fundamental problem in this design would be the pre-existing differences between participating centres with regard to the control group; routine management varies between hospitals (which actually is one of the main motives of performing this study). In effect, the nature of the intervention, and the underlying health care problem it is hypothesised to reduce, have forced us to adopt the pre-post interventional design.

The main drawback of the study design we have chosen is that it does not control for different types of bias. In other words, differences between the "pre-group" and the "post-group" may be attributed to differences in the populations studied in the two different periods rather than the effect of the intervention. Furthermore, as the pre- and post-groups are included over a period of time, secular trends in management are not controlled for, and could affect the observed differences between groups. However, as the AIR score parameters were recorded throughout the study, it is possible to describe the populations with regard to the severity of symptoms and the probability of appendicitis. Also, to the best of our knowledge, no major changes in diagnostic or therapeutic aids were introduced during the study period that could have inflicted a change of routine management.

Study IV

Study IV is an unrestricted parallel 1:1 randomised controlled study, nested in study III. An alternative to the nested design would be to randomise all patients with suspected appendicitis, regardless of disease probability (i.e. AIR score sum). While this was considered, such a design would rather compare a supra-normal use of imaging with a more judicious use of imaging in the large group of patients with a low suspicion of appendicitis. Furthermore, in the group of patients with high probability of appendicitis, the mandatory use of diagnostic imaging would potentially delay surgery for patients that display signs of severe, generalised peritonitis. In conclusion, we did not regard this as a feasible or attractive, study design in the context of Swedish health care.

We were not able to blind the physicians, nor the study subjects as to the assigned intervention. In those cases where the histopathological examination determined the diagnosis (i.e. negative exploration and non-perforated appendicitis), the pathologist was not aware of the allocation of the patient. However, for patients with perforated appendicitis, and for non-operated patients, the surgeon or the physician discharging the patient decided on the diagnosis and was likely to know the allocation of the patient. This has a potential to reduce the internal validity of the study. However, even with an external expert panel, two problems would still occur; perforated appendicitis was defined as operative findings of a perforation, and this would be difficult to evaluate in retrospect. Furthermore, the allocation would be revealed in those cases where the randomisation status was commented on in the record, and indirectly if an imaging procedure was done soon after admission, or even done at all.

One concern in randomised controlled studies is lack of external validity (or generalisability). This may be less of a problem in this study due to its pragmatic design. In short, the large number of participating hospitals at various levels, as well as the broad selection of clinicians involved in the inclusion of patients and interventions, should increase the external validity of the results[221]. The external validity is further supported as an otherwise unselected sample of patients with equivocal signs of appendicitis presenting at emergency departments day or night with a wide age span, of both sexes, were eligible for inclusion.

Data collection

Studies I–IV

In cross-sectional studies (studies I and II), the collection of data can be either prospective or retrospective. In all studies, we have used pre-specified data collection forms for each patient included in the studies, thus the data collection has been forward in time from the start of the study (i.e. prospective). Although prospective data collection normally reduces the amount of missing data, as compared with retrospective data retrieval, there are alternatives to the methods used in these studies.

A computerised or online patient protocol would allow mandatory information boxes, so that the registration cannot be completed without all required information. This would not have been possible for study I as the data were already collected during the 1990:s. For studies II–IV it was considered, but because of the clinical context (i.e. emergency care) we were concerned that any extra step in the management of the patient, such as logging on to a computer and getting access to the correct web page, would reduce the proportion of eligible patients that would be included, especially during busy hours. The assumption was that a prepared data collection form handed to the physician on call would add a minimum of extra workload, and therefore increase the possibility of including patients round the clock. Today most, if not all, hospitals in Sweden have computerised patient records. This, in combination with smartphones and "pads", would probably help overcome the previous obstacles with digital data collection in future studies.

During the data collection for study I, the proportion of PMN was reported by the laboratories, without separating neutrophils from basophils and eosinophils. In the subsequent studies (II–IV), the proportion PMN refers to the proportion of neutrophils only. Although this is by definition two different measures, the combined attribution of basophils and eosinophils to the total PMN count is only a few percent, and in cases of an inflammatory response elicited by acute appendicitis the proportion of netrophils is most likely even more dominant.

Missing values

The use of paper forms as discussed above, and the setting of the studies in routine emergency health care has contributed to missing data in all studies. A more stringent supervision of the documentation of data in real time would have the potential to limit the amount of missing data. This would require the presence of administrative staff or nurses connected to the study group round the clock during the study period.

In study I, 206 subjects (27%) were excluded from analysis due to missing values. These patients had only one missing variable, which was PMN in 75% of the cases. In study II, 99 subjects (23%) had study protocols with at least one missing value, again the majority of missing data was attributed to neutrophils. In contrast to study I, we chose to perform multiple imputation in this study. In retrospect, it is reasonable to say that this could have been done in study I as well. Complete case analysis in study I led to the loss of all information regarding 27% of the cases, which may have resulted in reduction of statistical power, and increased risk of biased results in the internal validation[212 213]. However, as the missing values equally affected both the AIR and Alvarado score, the risk of severe bias is probably small.

There is no such thing as a general rule for when to impute missing values, but the combination of a smaller proportion of missing data as well as the clinical context and study design made us refrain from imputing missing values in studies III and IV.

Data analysis

While weighted ordered logistic regression analysis was used in the construction of the AIR score (study I) and the extended score (study II), the fundamentals of logistic regression analysis is technically advanced. The author does not claim to possess more than elementary knowledge, but some basic aspects are discussed in the following.

Regression analysis. For construction of the scores in studies I and II we have used logistic regression analysis. Considering the consequences of making false negative decisions with regard to advanced appendicitis, we weighted the regression analysis towards increasing weight for increased disease severity. Although stepwise regression analysis can induce bias, the retrograde manual stepwise regression analysis we have adopted

is to the best of our knowledge reasonable for exploratory purposes, as in the construction of a new score.

Reliability of predictors. Well-defined predictors, with low inter-observer variability, are in general preferable. While this is fulfilled for WBC, neutrophil and CRP measurements, it is not necessarily true for the interpretation of rebound tenderness or muscular defence. We have therefore categorised this more subjective variable into four categories ("none", "light", "medium" and "strong"), with the assumption that it should serve to increase inter-examiner agreement and reduce the loss of information of a binary "yes" or "no" variable[222].

Furthermore, the seemingly objective variables body temperature, WBC, neutrophils and CRP may vary due to differences in analytical methods and apparatus. Another concern that accompanies multivariable regression models, is the power for reliable predictions. Testing too many predictors on a dataset with few events (in this case appendicitis) can overfit the prediction model. A rough rule of thumb is that the sample should contain a minimum of 10 events per variable included in the model, which is adopted in study I[212 223].

Interventions of the STRAPPSCORE study

Study III

The AIR score is externally validated in study II and in three international cross-sectional studies, and it has been evaluated for the selection of patients for non-operative antibiotic treatment in one prospective study[106 193 224 225]. In study III, the aim was to evaluate the outcome after implementing an AIR-score-based intervention (i.e. the AIR-score-based algorithm, Fig. 6).

As discussed in the study design section, we did not consider a randomised design appropriate. The results of a non-randomised study with the intention to explore the change in outcome after introducing a clinical decision aid can be difficult to interpret, however. The structured recording of data regarding the patients can induce positive effects in the sense that it draws attention towards important signs, symptoms and laboratory parameters for diagnosing appendicitis. Thus, there are most likely two mechanisms of intervention; the structured data collection per se as well as the AIR-

score-based algorithm. A baseline period, with only structured data collection, followed by an intervention period should make it possible to analyse the impact on outcomes attributed to the AIR-score-based algorithm per se.

It was not imperative to follow the AIR-score-based algorithm. Both the clinician and the patient were allowed to overrule the assignment. Consequently, most patients were managed according to the algorithm during the intervention period, but some were not. This may reduce the differences between the intervention and baseline period, leading to type II statistical errors. However, when considering the alternative, it was not deemed reasonable, or ethical, to have the AIR-score-based algorithm override the clinical judgement of participating clinicians in the first interventional study of the algorithm ever. Furthermore, even after successful and thorough external validation, any clinical decision aid should be used to *aid* the clinician's decision making, not replace it.

Study IV

For patients randomised to the Imaging group, the instruction to the clinician was to immediately refer the patient to an appropriate "early imaging" procedure. The type of modality was restricted to CT or US, but the CT or US protocol, the use of contrast agents, the type of scanners or software were not standardised. This can be criticised as CT and US have different diagnostic properties[161 162].

"Early imaging", as described in the study protocol can be accused of being, at best, ill-defined. However, even if only CT *or* US were allowed, a proper standardisation would require the use of equivalent hardware, CT-protocols and contrast administration across participating centres. Furthermore, the level of experience of the US operator, as well as the CT interpreter, would need standardisation. This would require a nationwide adoption of round the clock imaging standards with rather substantial implications for resource allocation, revised imaging protocols and local guidelines. This was considered unrealistic. Thus, we have used a pragmatic approach with regard to the definition of "early imaging"; immediate referral for US or CT as soon as possible.

For the Observation group, the clinician was instructed to perform a clinical re-assessment within 4–8 hours, and record the AIR score parameters again. In this group, selective imaging was advocated for patients with persisting intermediate AIR score. Again, the type of imaging was not standardised, but decided at the discretion of the clinician depending on patient characteristics and local availability.

In accordance with study III, and for much the same reasons, compliance with randomisation status was not imperative. The clinician was asked to note the reason for non-compliance in the data collection form. Also, whether to operate or institute non-operative treatment was decided at the discretion of the surgeon for both groups. Again, compulsory compliance with the imaging result or AIR-score-based re-assessment would be more scientifically efficient in theory, but it was not considered compatible with the clinical context of this multicentre study. Nevertheless, 95% of the patients in the Observation group had an initial observation period of 4–8 hours and 94% of the patients in the Imaging group had an imaging procedure done within a median waiting time of five hours.

Randomisation (study IV)

Randomisation involves sequence generation, concealment and implementation of the allocation. Unrestricted sequence generation is considered superior to all other methods in bias prevention and unpredictability of allocation, but has the disadvantage of not controlling for the risk of obtaining imbalanced groups[226 227]. On the other hand, this is a risk that is diminished as sample size grows, and in a study with over 500 participants in each group it should be small for the total sample, but on hospital level imbalance may be a problem[227].

Since the interventions (i.e. imaging procedures) were not standardised across participating centres, imbalanced groups on hospital level could potentially induce bias. Therefore, we employed 1:1 randomisation restricted only on a hospital level. In other words, each hospital received an equal number of thoroughly shuffled and opaque envelopes for the two assignments. The surgical emergency room nurse was instructed to ask the physician about eligibility in the study for each patient admitted for abdominal pain, and to hand out the next envelope to the physician if the patient was deemed eligible and included in the study. The process was enhanced by a runner kept with the envelopes, in which every included patient was registered together with

the randomisation status, thus making it possible to match the number of envelopes with the number of randomised patients.

Admittedly, this is not the optimal sequence generation or concealment strategy. However, opaque sealed envelopes may give adequate concealment according to a recent meta-epidemiological study[228]. The Imaging and Observation groups were similar with regard to the baseline characteristics which also suggests that randomisation was achieved (Table 2).

A central online sequence generator, preferably linked to an online data collection form for each patient, would be a more robust strategy. As previously mentioned, we were concerned about the trade-off with regard to the risk of bias due to lower inclusion rates during busy hours inflicted by additional workload for the clinician. Furthermore, the risk of bias is considered small with the type of concealment we have used[229]. Also, the nature of the endpoints of the study has implications in this regard; objective endpoints are less at risk of bias than subjective endpoints[230].

PRINCIPAL FINDINGS AND INTERPRETATION

Internal validation and comparison with the Alvarado score

Internal validation is an evaluation of the accuracy of the test in the sample used to develop it[231]. The AIR score was validated on a random subsample of 229 patients that were included in study I. The prevalence of appendicitis in this sample was 33%, which corresponds to the prevalence of appendicitis of 35% in the larger multicentre study III. Thus, it seems that the internal validation sample comprises a relevant cohort with regard to the group of patients that the AIR score is intended to serve, namely patients presenting at the emergency department with acute abdominal pain, eliciting reasonable suspicion of appendicitis. Also, this is important as the analytical specificity (the tendency to yield false positive results due to other conditions or components) would otherwise not be assessed properly, and diagnostic accuracy would most likely be overestimated[188 197].

Discriminating capacity and predictive properties

In general, when evaluating diagnostic tests, we are interested in two dimensions of diagnostic accuracy: discriminating capacity and predictive value. Discriminating capacity refers to the test's ability to separate diseased from non-diseased persons, and predictive value implies the probability that a person with a certain test result is diseased (or non-diseased). Detection and expeditious treatment of perforated appendicitis has higher priority than non-perforated appendicitis, as the latter may represent a different entity with the potential of spontaneous resolution in uncomplicated cases[41 44 45 232]. Therefore, AUC and predictive metrics are reported for all appendicitis and advanced appendicitis, separately.

Discriminating capacity. The AUC was larger for the AIR score in detecting advanced compared with all appendicitis (0.97 vs 0.93, respectively) which was better than the Alvarado score (0.92 vs 0.88). Although the favourable outcome of the AIR score in the head-to-head comparison with the Alvarado score is encouraging at first sight, one may also infer this comparison to be unfair: In this study the AIR score is internally validated and, in effect, the Alvarado score is externally validated, which inherently may result in a seemingly poor performance[233 231].

Regarding the AIR score, there was a trend towards better discriminating capacity among women and young patients, but this was not statistically significant. However, it elicited the planning of stratified analyses with regard to age and sex in subsequent external validation of the AIR score.

Predictive value. A low and a high cut-off was defined for both the AIR score and the Alvarado score in order to allow comparison of the scores' predictive values and proportion of patients with intermediate, "grey zone", test results as well as the proportion of patients allocated to the high- and low-risk groups, respectively (Table 4).

Low-risk group (<5p): The score-based algorithm suggests out-patient management for patients in the low-risk group. Consequently, the sensitivity for advanced appendicitis is considered the most important diagnostic aspect for this group. The sensitivity and PV– for advanced appendicitis was 1.0 for the AIR score, which also applies to the Alvarado score. These results suggest that both scores can rule out advanced appendicitis for patients with a low score.

High-risk group (>8p): The score-based algorithm suggests operation without any further investigations in the high-risk group, and therefore the specificity for all appendicitis is regarded as the highest priority here. The AIR score and Alvarado score had similar specificity (0.99), and the PV+ for all appendicitis at this cut-off was 0.97 and 0.91, respectively. This can also be expressed in another way: 99% of the non-appendicitis patients had a score of eight points or less (specificity), and 97% of the patients with an AIR score higher than eight points had appendicitis (PV+).

Intermediate-risk group (5-8p): In order to illuminate the potential value of the AIR score in a clinical context, the diagnostic accuracy metrics have to be interpreted in combination with the proportion of patients that are defined as being at intermediate or "grey zone" risk of having appendicitis. This proportion was 37% for the AIR score. Consequently, 63% of the patients were allocated to a high- or low-risk group with high diagnostic accuracy. The corresponding proportions for the Alvarado score were less favourable, but again, this must be interpreted with caution as the AIR score is evaluated on patients representing the same population as it was constructed on in this study.

External validation

Regardless of how well a diagnostic test performs in internal validation, it is useless if it cannot predict outcome outside the material it was developed in. External validation refers to the generalisability of the results obtained during internal validation tests. Generalisability comprises the *reproducibility* of the test's diagnostic performance on patients who were not included in the development of the test, and *transportability*, i.e. the test's ability to yield accurate predictions in different settings, in plausibly related populations (populations that would be reasonable to apply the test to)[212 231].

External validation of the AIR score

The AIR score has to date been externally validated in two studies from our group (studies II and III), and three studies from international research groups. The main results are compiled in Table 13. The global measure of discrimination is the AUC, and the most important predictive properties are the ability to rule out advanced appendicitis at the low cut-off (sensitivity) and to rule in appendicitis, i.e. rule out negative explorations, at the high cut-off (specificity).

Table 13. Diagnostic properties of the AIR score in validation studies for all and advanced appendicitis (AA).

		Prevalence of appendicitis (%)		AUC		Sensitivity low cut-off		Specificity high cut-off		Intermediate score results
	n	All	AA	All	AA	All	AA	All	AA	(%)
Study I[192]	229	33.2	13,1	0.93	0.97	0.96	1.0	0.99	0.99	37
Study II[234]	428	41.4	15.9	0.91	0.96	0.93	0.99	0.98	0.98	42
Study III	3791	34.9	15.0	0.83	0.89	0.90	0.96	0.98	0.98	49
de Castro et al[224]	941	36.8	10.8	0.96	0.96	0.93	0.93	1.0	1.0	40
Kollár et al[225]	182	36.8	8.8	0.85	*	0.94	1.0	0.97	0.92	45
Sammalkorpi et al[193]	829	47.3	11.3	0.81	*	0.83	*	0.97	*	*

*Estimate not specified for the AIR score.

Study II. The data collection of this study was conducted several years after study I. It involved one external centre, and in contrast to study I, patients presenting at the emergency department regardless of admission to hospital were included. Consequently, study II should have the potential to stress the secular, geographical and spectrum transportability of the AIR score[231 235].

The prevalence of appendicitis in this study was 41%, as compared with the prevalence of 33% in study I. The AUC for detecting all appendicitis and advanced appendicitis was 0.91 and 0.96 respectively, which is only marginally lower than in study I (0.93 and 0.97). This surprisingly small difference may be attributed to the fact that some patients in study II were rescored after an observation period in accordance with the study protocol, and the last available AIR score was used in the analysis. Repeated evaluation may have helped to yield a higher discriminating capacity than expected in the external validation[236].

The proportion of neutrophils was used in study II rather than the proportion of PMN, as in study I. In this regard, one can debate whether study II is an external validation of the AIR score, or a new internal validation of a slightly altered AIR score.

Study III. All 3791 patients from 25 centres participating in the STRAPPSCORE study were included in the external validation of the AIR score's discriminating capacity and predictive value. The overall prevalence of appendicitis was 34.9%, which is nearly identical with the prevalence in study I. In this study, the patients' AIR score upon arrival at the emergency department were used to analyse its discriminating and predictive properties. The ROC area was 0.83 for all appendicitis and 0.89 for advanced appendicitis (Table 8).

The larger sample in this study allowed an analysis of discriminating capacity in subgroups. The AUC was higher in younger patients regardless of disease severity. This was also true for patients with a longer duration of symptoms on arrival at the emergency department. In contrast, a higher discriminating capacity in women was only confirmed for all appendicitis, but it did not reach significance for advanced appendicitis. Interestingly, no difference was found with regard to the level of experience of the examining physician. This implies that the AIR score may help to bridge the gap between less experienced and senior physicians. However, with four subgroups, the failure to find any differences in this regard could be the result of a

type II error; no sample size calculation was done for the sub-group analysis. Furthermore, between 0.3% (age) and 13.2% (duration of symptoms) of the patients did not have the subgroup variable recorded, and these were not multiply imputed, which may give biased results.

The sensitivity of advanced appendicitis for the low-risk group and the specificity for all appendicitis for the high-risk group was 0.96 and 0.98, respectively. This implies that the AIR score can rule out advanced appendicitis in the low-risk group, and rule in appendicitis in the high-risk group with high, but not absolute, certainty. According to the subgroup analysis, caution should be applied in ruling out advanced appendicitis in the low-risk group for patients with short duration of symptoms, older patients and men, and in ruling in appendicitis in the high-risk group for older patients (Table 9).

Also, if all patients with an AIR score of more than eight points were scheduled for surgery, the proportion of negative appendectomies could reach 14% (assuming that all non-operated patients in the high-risk group did not have appendicitis), which is no longer regarded as acceptable[237]. This implies that patients with a high AIR score should invariably have a formal surgical consultation. If the patients has a relatively unaltered general condition and moderate signs of local peritonitis on clinical examination, initial management with a period of close in-hospital observation or diagnostic imaging, rather than immediate surgery, may be the way to proceed. It is important to emphasise that a clinical score, or any other diagnostic method, does not replace clinical judgement; the patient's characteristics and the clinical context have to be taken into consideration.

External reports. Two international groups have reported on the diagnostic properties of the AIR score applied to both children and adults in diagnostic accuracy studies, and one has reported on the diagnostic properties when applied to adults only (Table 13)[193 224 225].

In the Netherlands, De Castro et al. conducted a cross-sectional study in 2012 that evaluated the AIR score and the Alvarado score on prospectively collected data of 941 subjects[224 238]. The definition of phlegmonous appendicitis relied on histopathological criteria, and advanced appendicitis (i.e. gangrenous and perforated appendicitis) was based on macroscopic appearance. The prevalence of appendicitis in this study was 37%, close to the prevalence in study I. The AIR score's discriminating capacity expressed as AUC was 0.96 for all appendicitis and advanced appendicitis, which was

higher than for the Alvarado score (0.82). The AIR score's sensitivity for advanced appendicitis at the low cut-off was 0.93 and the specificity for all appendicitis at the high cut-off was 1.0.

In 2014, Kollár et al. published their results from a cross-sectional study from Ireland of 182 consecutive patients with suspected appendicitis[225]. The diagnostic properties of the AIR score, Alvarado score and the clinical impression of a senior surgeon were assessed. The appendicitis diagnosis was established by histological evidence of transmural inflammation (phlegmonous appendicitis) or transmural gangrene or perforation (advanced appendicitis). As in the study by de Castro et al., the prevalence of appendicitis was 37%. The discriminatory capacity of the AIR score, as determined by the corresponding AUC, was 0.85, which was almost identical with that of the Alvarado score and the senior surgeon's assessment (0.84 and 0.86, respectively). The sensitivity for advanced appendicitis at the low cut-off was the same for both scores (1.0) and not significantly better than the assessment of a senior surgeon (0.88). The specificity of the AIR score with regard to all appendicitis at the high cut-off was 0.97, which was higher than the Alvarado score (0.76) but not higher than the assessment of a senior surgeon (0.91).

Recently, Sammalkorpi et al. reported on the construction of the new adult appendicitis score (AAS) and comparison with the AIR and Alvarado score in 829 patients, 16 years and older, in Finland[193]. In contrast to the AIR score and the Alvarado score, the AAS includes information on the patient's sex and duration of symptoms. The histopathological criterion for appendicitis was transmural infiltration of neutrophils. Complicated appendicitis was defined as perforation or abscess. In this study, the distinction between phlegmonous and gangrenous or perforated appendicitis was omitted for the presentation of diagnostic accuracy. The prevalence of appendicitis was 47.4%, which is higher than in the previous internal and external validation studies of the AIR score. The performance of the AAS was superior to that of the AIR score and the Alvarado score. The AUC of the AIR score was 0.81, which was not significantly better than the Alvarado score (0.79), but was inferior to that of the AAS (0.88). Sensitivity for all appendicitis at the low cut-off was 0.83 and specificity at the high cut-off was 0.97 for the AIR score. The corresponding results for the Alvarado score were 0.98 and 0.94, respectively.

Assessment of new inflammatory markers (study II)

We evaluated the potential of several inflammatory markers not in routine use for inclusion in the AIR score in this study. Others have studied their individual diagnostic properties, but their incremental contribution in diagnosing appendicitis when combined with a clinical score is unclear. We constructed and tested an extended version of the AIR score, incorporating CCL-2, which was the model with the highest point estimate in the multivariable analysis. The extended score and the AIR score were found to have similar predictive properties (Table 6). This implies that improvement of appendicitis scores probably has to rely on the inclusion of discriminators that are stronger than those known today.

Effect on outcome (studies III and IV)

The final step in evaluating the diagnostic properties of a diagnostic test after internal and external validation studies is to assess the effect on outcome in prospective interventional studies[188 197]. In study III, outcome measures were compared between the baseline and the intervention period. As all patients with intermediate risk of appendicitis were eligible for study IV, the evaluation could only be applied to the low- and high-risk groups.

In the low-risk group, the intervention resulted in a reduction of negative explorations and operations for phlegmonous appendicitis, fewer admissions and 45% reduction in the use of diagnostic imaging, in accordance with the management algorithm. This was in spite of a higher mean AIR score among patients included during the intervention period (Table 10).

In the high-risk group there was a 28% reduction in the use of diagnostic imaging indicating a high compliance with the AIR-score-based algorithm in this regard. This was not accompanied by any changes in the proportion of negative explorations, which was 5% in both periods, or operations for phlegmonous or advanced appendicitis.

The total number of readmissions and the frequency of missed appendicitis were the same in both periods. These results suggest that the AIR score is an efficient and safe decision support for clinicians in selecting low-risk patients for outpatient management and high-risk patients for surgical consultation. The lower proportion of operations for non-perforated appendicitis in the low-risk group during the intervention period is in accordance with the hypothesis of resolving appendicitis. The reduced use of imaging in both the low- and the high-risk groups may have contributed to the reduction in operations for non-perforated appendicitis in the low-risk group and did not lead to an increase in the negative appendectomy rate in either group. This is in accordance with recommendations to avoid diagnostic imaging in populations with low or high prevalence of appendicitis due to the high risk of false positive and false negative results, respectively.[161 162]

Although the study design has inherent limitations, these results suggest that the AIR-score-based algorithm may reduce the risk of negative explorations and allow non-operative management of uncomplicated cases of phlegmonous appendicitis. In the low-risk group, few patients were treated with antibiotics during the baseline (0.5%) and intervention periods (1.2%). Thus, the reduction in operations for non-perforated appendicitis seen during the intervention period is probably not related to antibiotic treatment.

Effect of routine imaging versus observation and selective imaging in the intermediate–risk-group (AIR score 5–8)

In contrast to our hypothesis, routine imaging did not reduce the proportion of negative explorations. Although we did not reach the target enrolment due to falling recruitment rate this only reduced the statistical power to detect the projected decrease of negative appendectomies from 15% to 10% from the desired 80% to about 70%.

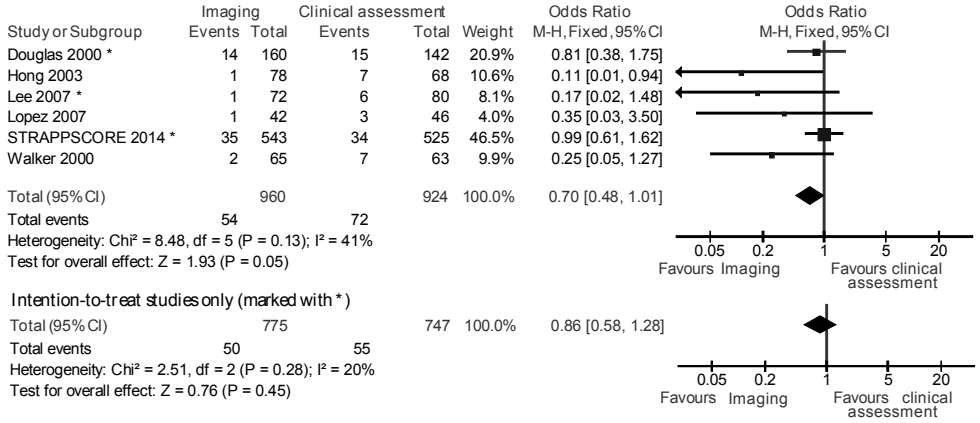
Routine imaging was associated with an increase in treatment of non-perforated appendicitis supporting an increased detection, and treatment, of uncomplicated cases of appendicitis that would otherwise resolve spontaneously. While this assumption would have been regarded as speculative previously, circumstantial evidence of spontaneous resolution has been accumulating in later years[41 44 47 48 77-79 81].

No other difference of clinical importance was found. In-hospital observation and selective imaging was thus associated with fewer operations for non-perforated appendicitis, a reduction ionising radiation exposure and no adverse effects compared with routine imaging.

These results diverge from the current general belief. Possible explanations are that the clinical diagnosis has been underestimated, that the efficiency of imaging has been overestimated or a combination of both. The diagnostic efficiency of imaging is undisputed, but most reports come from specialised, high-quality single centres, and the real-world situation may be different. There is very little knowledge of the diagnostic efficiency of modern clinical management.

Only five randomised trials have been published comparing the impact of routine imaging with clinical assessment or selective imaging in patients with suspected appendicitis, and the results are conflicting[168-170 239 240]. A meta-analysis of these studies, including the STRAPPSCORE study, shows that routine diagnostic imaging is associated with a slight reduction of negative explorations (Fig. 9 a). However, in a sensitivity analysis of the randomised studies, including only those that analyse their results according to intention-to-treat (ITT), there was no significant reduction of negative explorations, but an increase in the number of non-perforated appendicitis (Fig. 9 a-b). There was no difference with regard to the number of operations for perforated appendicitis, but it was only specified in three studies[169 170 239].

(a)



(b)

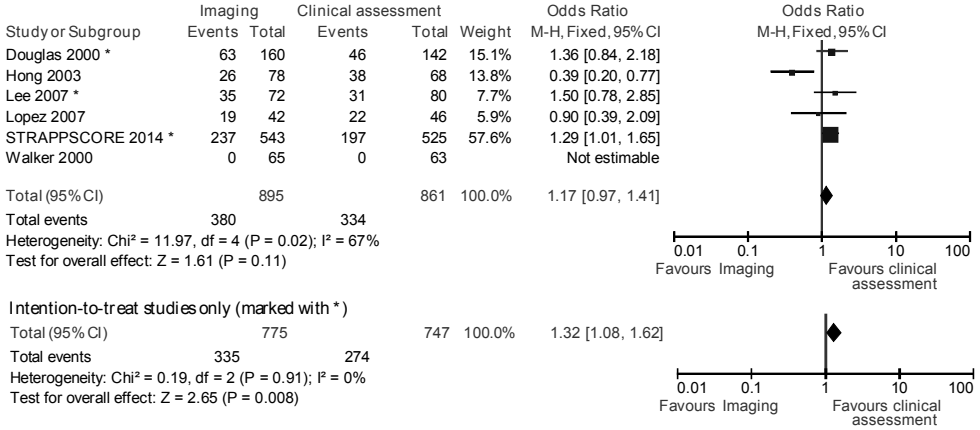
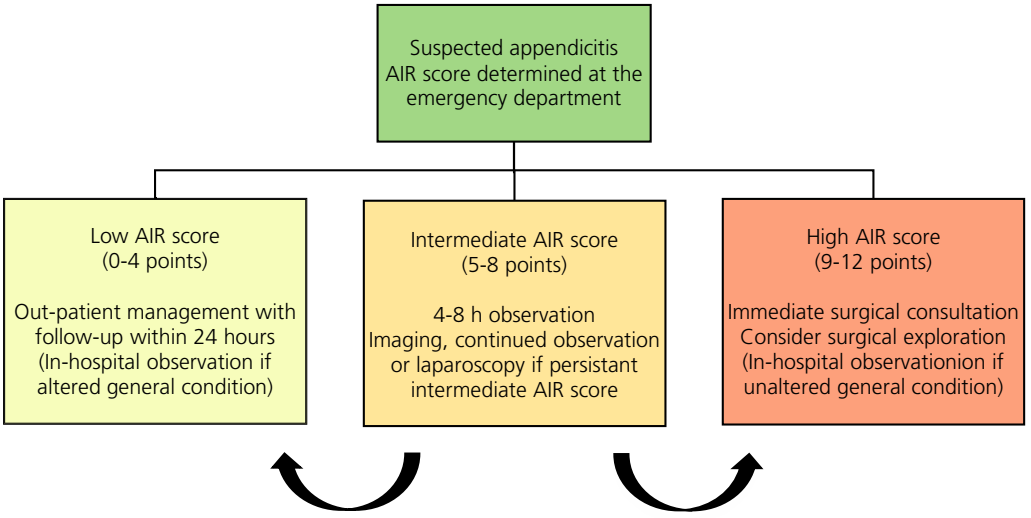


Fig. 9. Forest plots of randomised controlled studies comparing routine diagnostic imaging ("Imaging") with clinical assessment or selective diagnostic imaging ("Clinical assessment") for patients with suspected appendicitis, including a sensitivity analysis of only ITT studies. (a) Negative explorations. (b) Non-perforated appendicitis.

PROPOSED ALGORITHM



CONCLUSIONS

- The AIR score proved to have favourable diagnostic properties in the internal and external validation studies
- Combining the AIR score with novel inflammatory markers did not improve its diagnostic properties further; in order to do so, stronger discriminators than those known today need to be identified.
- The implementation of an AIR-score-based clinical algorithm can reduce the need for diagnostic imaging in patients with suspected appendicitis. Implementation of the algorithm was associated with a reduction of hospital admissions and increasing diagnostic accuracy in patients with limited signs and symptoms suggestive of appendicitis.
- Patients with suspected appendicitis and equivocal clinical findings do not benefit from routine diagnostic imaging compared with repeat clinical examination and selective imaging. The latter was associated with fewer operations for non-perforated appendicitis, which supports the hypothesis of resolving appendicitis.

FUTURE PERSPECTIVES

We have not made a health-economic assessment of the AIR-score-based clinical algorithm. Hypothetically, a reduction in negative explorations and operations for phlegmonous appendicitis as well as a reduction in hospital admissions and use of diagnostic imaging could reduce health care spending in this large group of emergency patients. However, long-term follow-up is appropriate in order to include costs attributed to recurrent symptoms or postoperative conditions.

The rapid adoption of computerised systems in routine health care will most likely make it possible to apply more advanced yet user-friendly decision aids. These will not lose information due to truncation or dichotomisation of continuous variables and may take into consideration non-linear and multidimensional correlations.

The evolution of diagnostic imaging will certainly continue, which ideally will establish the diagnosis of patients with higher accuracy and without today's drawbacks of ionising radiation and operator dependency. Within a foreseeable period of time, however, it will not replace the need for initial clinical evaluation, selecting the appropriate modality and time frame of diagnostic imaging for patients who are triaged to further diagnostic work-up.

Simple decision aids, (e.g. the AIR score) should be validated for the health-care environment of low-income countries, short of advanced technique in terms of computers and diagnostic imaging.

Most importantly, a better understanding of the disease spectrum is fundamental for adequate management of the largest group of patients in abdominal emergency surgery of today: Which appendicitis patient benefit from surgery, who will benefit most from antibiotic treatment, and who does not need any treatment at all?

SAMMANFATTNING PÅ SVENSKA

(Summary in Swedish)

Blindtarmsinflammation är en vanlig åkomma som sammanlagt drabbar ca 7% av alla kvinnor och 9% av alla män. Patienter med buksmärtor som framkallar misstanke om blindtarmsinflammation utgör en betydande del av patienterna på en kirurgisk akutmottagning, och att ställa korrekt diagnos är ibland svårt. Detta leder till att patienter vissa gånger opereras i onödan (så kallad negativ exploration) eller att patienter med blindtarmsinflammation felaktigt skickas hem från akutmottagningen för att inom kort återkomma med ibland allvarlig sjukdomsbild.

Traditionellt sett har handläggningen av patientgruppen präglats av en stor benägenhet att operera bort blindtarmen även vid begränsade symtom, då uppfattningen varit att alla inflammerade blindtarmar förr eller senare brister. Detta har lett till en stor andel (25-40%) negativa explorationer. Ett alternativt synsätt på sjukdomens naturalförlopp har uppkommit då det visat sig att en brusten blindtarmsinflammation oftast föreligger redan när patienten kommer till sjukhus och att observation av övriga fall inte verkar öka antalet blindtarmar som brister. Sannolikt representerar brusten- och icke-brusten blindtarmsinflammation i viss utsträckning två olika sjukdomar: en snabbt progredierande typ som leder till brusten blindtarm och svår sjukdomsbild, och en annan mer stillsam typ som sällan leder till brusten blindtarm, utan snarare tenderar att läka spontant eller som resultat av antibiotikabehandling.

Under senare decennier har den kliniska diagnostiken kompletterats med analys av inflammatoriska parametrar, framför allt vita blodkroppar, C-reaktivt protein (CRP) och polymorfkärniga neutrofila granulocyter. Dessutom har bilddiagnostiska metoder såsom skiktröntgen (CT) och ultraljud (UL) börjat användas alltmer. CT har generellt sett god diagnostisk skärpa, men utsätter patienten för joniserande strålning, vilket framför allt är ett problem för yngre individer och barn. UL är helt ofarligt men har något sämre diagnostiska egenskaper och är mer undersökarberoende. För både CT och UL gäller att de presterar allra bäst i patientgrupper med medelhög förekomst av blindtarmsinflammation, och sämre i grupper med mycket hög eller mycket låg sjukdomsförekomst.

Upprepad klinisk bedömning och värdering av inflammatoriska parametrar kan bidra till att skärpa diagnostiken när symtombilden initialt inte är övertygande. Klinisk diagnostik utgör dock en komplex sammanvägning av delvis subjektiva variabler som kräver erfarenhet. Kliniska scorer har inte fått något nämnvärt genomslag i rutinsjukvård, vilket kan bero på bristande användarvänlighet eller att de utgjort ett otillräckligt beslutsstöd. En klinisk score har dock de teoretiska fördelarna att den sammanväger kliniska parametrar avseende den aktuella patienten och ger en prognostisk information som kan utgöra en bas för den fortsatta handläggningen.

Det vore fel att påstå att det råder koncensus avseende vilket naturalförlopp som korrekt återspeglar verkligheten, i vilken utsträckning man kan lita på klinisk diagnostik eller vilka patienter som har nytta av bilddiagnostiska metoder. Handläggningen av denna stora patientgrupp präglas därför av stor variation.

Målet med den här studien är att konstruera och validera en ny klinisk score (Appendicitis Inflammatory Response, AIR, score) samt utvärdera om nya biokemiska inflammationsmarkörer kan förbättra dess diagnostiska förmåga. Vi vill också undersöka om en AIR-scorebaserad strukturerad handläggning av patienter med misstänkt blindtarmsinflammation kan förbättra diagnostiken och minska behovet av inläggningar på sjukhus och bilddiagnostik. Slutligen vill vi också pröva hypotesen att handläggning med rutinmässig bilddiagnostik (CT eller UL) minskar antalet negativa explorationer till kostnaden av ett ökat antal operationer av lindrig blindtarmsinflammation som kan spontanläka, jämfört med observation på sjukhus med selektivt använd bilddiagnostik.

Delarbete I

I denna observationsstudie nyttjade vi prospektivt insamlade data beträffande kliniska symtom och tecken av samt inflammatoriska parametrar för 545 patienter som lagts in på fyra sjukhus i södra Sverige under misstanke om blindtarmsinflammation. Vi delade i efterhand upp patienterna i två grupper, där data avseende den ena gruppen användes för konstruktion av scoren med hjälp av multivariabel regressionsanalys. Data från den andra gruppen användes för validering AIR-scoren och jämförelse av dess diagnostiska kapacitet med en annan klinisk score (Alvarado-scoren).

AIR-scoren konstruerades av åtta variabler med oberoende prediktiv förmåga (smärta i höger fossa, släppömhet eller muskelförsvar, feber, kräkning, förhöjt antal vita

blodkroppar, ökad andel polymorfkärniga granulocyter och ökade CRP-nivåer). Vi fann att en låg AIR-scoresumma (0-4 poäng poäng) kunde användas för att utesluta avancerad blindtarmsinflammation och att en hög AIR-scoresumma (9-12 poäng) kunde användas för att säkerställa att patienten hade blindtarmsinflammation med hög säkerhet. I jämförelse med Alvarado-scoren föreföll AIR-scoren ha starkare diagnostisk förmåga.

Delarbete II

Denna observationsstudie inkluderade 432 patienter som sökt på akutmottagningarna vid Länssjukhuset Ryhov i Jönköping och Universitetssjukhuset i Linköping med misstänkt blindtarmsinflammation. AIR-scoreparametrarna registrerades och nya biokemiska inflammationsmarkörer analyserades på blodprov från patienterna. På likartat sätt som i studie I konstruerades nu en ny score som även innehöll den mest lovande av de nya markörerna (CCL-2).

Vi fann att de lovande diagnostiska egenskaperna för AIR-scoren från studie I kunde reproduceras i denna studie, men att addera ytterligare nya markörer förbättrade inte scoren ytterligare.

Delarbete III och IV (STRAPPSCORE-studien)

Delarbete III är en prospektiv interventionsstudie i vilken 3791 patienter med misstänkt blindtarmsinflammation inkluderades vid 25 olika svenska sjukhus. Studien inleddes med en fas under vilken AIR-scoreparametrar registrerades passivt, men själva scoren tillhandahölls inte och räknades inte ut. Under nästa fas infördes en strukturerad AIR-scorebaserad handläggning av patienterna i enlighet med en studiealgoritm som förespråkade hemgång med poliklinisk uppföljning för patienter med låg AIR-score (0-4 poäng) eller operation utan ytterligare diagnostik för patienter med hög AIR-score(9-12p). Patienter i mellangruppen (5-8 poäng) blev erbjudna att delta i en randomiserad studie med lottning mellan rutinmässig bilddiagnostik alternativt inläggning för observation och selektiv bilddiagnostik vid fortsatt oklar klinisk bild (delarbete IV).

Vi fann i delarbete III att AIR-scoren hade något svagare diagnostisk styrka än i delarbete I och II när den testades på alla inkluderade patienter i bägge faser. Däremot visade sig en strukturerad AIR-scorebaserad handläggning minska antalet negativa explorationer och operationer för okomplicerad blindtarmsinflammation, samt minska antalet inläggningar på sjukhus för patienter med låg AIR-score. För patienter med låg- eller hög AIR-score minskades dessutom behovet av bilddiagnostik.

I delarbete IV fann vi att strategin med rutinmässig bilddiagnostik inte minskade risken för negativa explorationer eller påverkade antalet patienter som opereras för brusten blindtarmsinflammation. Däremot noterades i enlighet med vår hypotes att rutinmässig bilddiagnostik ökade antalet operationer för icke brusten blindtarmsinflammation, jämfört med gruppen av patienter som lottades till observation och selektiv bilddiagnostik. Vi tolkar det som att i den senare gruppen tillåts okomplicerade fall spontanläka. Rutinmässig bilddiagnostik ledde till kortare väntan på operation utan att påverka den totala vårdtiden för patienten.

Konklusion

AIR-score har goda diagnostiska egenskaper som inte förbättras när den kombineras med nya inflammationsmarkörer. Strukturerad AIR-scorebaserad handläggning av patienter med misstänkt blindtarmsinflammation kan leda till förbättrad diagnostik och ett minskat behov av bilddiagnostiska undersökningar. Rutinmässig bilddiagnostik för patienter med oklar klinisk bild tycks inte innebära någon uppenbar fördel jämfört med observation och selektiv bilddiagnostik.

ACKNOWLEDGEMENTS

Roland Andersson, my main supervisor. For patience and for generously sharing your scientific and clinical knowledge with me. I would also like to thank you for the moments we have spent talking about anything from flash drives to religions. Also, this would probably be the proper place to comment on your splendid sense of humour and seemingly supernatural intelligence, but I have decided not to, as it would jeopardise your humility.

Conny Wallon, my co-supervisor, for support and encouraging comments delivered in native "Östgötska", and for catalysing this thesis by making the Department of Surgery at Linköping University Hospital participate in the studies.

Co-authors **Christina Ekerfelt**, **Gunnar Olaison**, **Blanka Kolodziej** and **Hanna Björnsson Hallgren** for important co-operation in the design and implementation of the studies and for wise comments regarding the interpretation of the results. In particular **Marie Rubér**, who in addition to the aforementioned contributions also performed the analyses of new inflammatory markers, merits special attention.

Previous and present head of the Department of Surgery, Ryhov County Hospital; **Johannes Järhult**, **Axel Ros** and **Erik Wellander**, for continuously encouraging scientific work in our department and for enduring my sometimes unnecessarily straight forward way of verbalising my divergent opinion in various matters.

My **colleagues**, **staff** and **friends** at the Department of Surgery, in particular the members of the vascular team **Håkan Åstrand**, **Erik Wellander**, **Francis Rezk**, **Magnus Rydh**, **Burkhard Lotz**, **Veronica Skoog** and the generous vascular interventionists **Berne Åsberg** and **Werner Puskar**, for the countless challenging, exciting, sad and happy moments we have shared throughout the years. I could not have been more fortunate.

My (somewhat) retired colleagues and highly distinguished surgeons **Rudolf Schiöler**, **Anders Hugander**, **Rune Gustavsson**, **Reine Gustavsson** and **Johannes Järhult** for your intense efforts to teach me some hard core "reality based surgery", commonly referred to by others as the art of medicine.

All **colleagues** and **staff** at the Unit of Vascular Surgery and Interventional Radiology, Sahlgrenska University Hospital in Göteborg, for letting me experience your contagious pursuit of perfection in evidence based vascular surgery. In particular I would like to thank **Håkan Roos** and **Klas Österberg**. The former accidently happened to recruit me to spend a year with their vascular team, and both surprisingly found themselves recruited to the arctic mountain marathon team.

The exceptional **staff** at the Medical Library, Ryhov County Hospital, for reference retrieval.

All members of the **STRAPPSCORE studygroup** for making the STRAPPSCORE study possible.

Pär Lindblom, my best friend, with family. For great memories, long talks, long distance runs and skiing adventures in thin air. **Karl Hagman,** my very best friend, with family. For constantly providing new perspectives on life on thin ice, for laughs and medium distance runs. **Anders Jonsson**, undoubtedly my best friend, with family. Not exactly for running, but for friendship, unprecedented social skills and great vacations together. There are more to come. **Johan** "STATA" **Mårtensson**, my best friend ever, with family. For short distance runs, long-distance breakfasts at your summer house and for your exuberant generosity. The **Malin and Johannes Bengnér** family. For, needless to say, global excellence and electrifying friendship.

My older brothers **Mikkel** and **Mattias,** with families, who supposedly both took an active part in my upbringing in Arjeplog back in the 1970s. In short, for being brothers.

My parents, **Jane and Christer**, for love and support, and for being such great playmates for our children.

Kristina, my mother in law. What would we have done without you? **Bengt-Erik**, **Maria** and **Mats**, with extended families; my father-, sister-, and brother in law, for bringing back memories, cousins and unforgettable moments over and over again.

Einar, **Line** och **Elmer**. Ni är det största som hänt, det bästa jag vet och det finaste vi har!

Malin, my love. Words are not enough. 66° 33′ 39″N, 17° 7′ 41″E

REFERENCES

1. Schumpelick V, Dreuw B, Ophoff K, et al. Appendix and cecum. Embryology, anatomy, and surgical applications. The Surgical clinics of North America 2000;**80**(1):295-318.

2. Ajmani ML, Ajmani K. The position, length and arterial supply of vermiform appendix. Anatomischer Anzeiger 1983;**153**(4):369-74.

3. Wakeley CP. The Position of the Vermiform Appendix as Ascertained by an Analysis of 10,000 Cases. Journal of anatomy 1933;**67**(Pt 2):277-83.

4. Soybel DI. Appendix. In: norton HA, ed. Surgery Basic Sciense and Clinical Evidence. Second ed. New York: Springer Science+Business Media, LCC, 2008:991-1010.

5. Papadaki L, Rode J, Dhillon AP, et al. Fine structure of a neuroendocrine complex in the mucosa of the appendix. Gastroenterology 1983;**84**(3):490-7.

6. Spencer J, Finn T, Isaacson PG. Gut associated lymphoid tissue: a morphological and immunocytochemical study of the human appendix. Gut 1985;**26**(7):672-9.

7. Dasso JF, Obiakor H, Bach H, et al. A morphological and immunohistological study of the human and rabbit appendix for comparison with the avian bursa. Developmental and comparative immunology 2000;**24**(8):797-814.

8. Bazar KA, Lee PY, Joon Yun A. An "eye" in the gut: the appendix as a sentinel sensory organ of the immune intelligence network. Medical hypotheses 2004;**63**(4):752-8.

9. Glover W. The Human Vermiform Appendix. The Human Vermiform Appendix A General Surgeon's Reflections 1988. https://answersingenesis.org/human-body/vestigial-organs/the-human-vermiform-appendix/.

10. Gebbers JO, Laissue JA. Bacterial translocation in the normal human appendix parallels the development of the local immune system. Annals of the New York Academy of Sciences 2004;**1029**:337-43.

11. Randal Bollinger R, Barbas AS, Bush EL, et al. Biofilms in the large bowel suggest an apparent function of the human vermiform appendix. Journal of theoretical biology 2007;**249**(4):826-31.

12. Wangensteen OH, Buirge RE, Dennis C, et al. Studies in the Etiology of Acute Appendicitis: The Significance of the Structure and Function of the Vermiform Appendix in the Genesis of Appendicitis a Preliminary Report. Annals of surgery 1937;**106**(5):910-42.

13. Murphy EM, Farquharson SM, Moran BJ. Management of an unexpected appendiceal neoplasm. The British journal of surgery 2006;**93**(7):783-92.

14. Esmer-Sanchez DD, Martinez-Ordaz JL, Roman-Zepeda P, et al. [Appendiceal tumors. Clinicopathologic review of 5,307 appendectomies]. Cirugia y cirujanos 2004;**72**(5):375-8.

15. Tchana-Sato V, Detry O, Polus M, et al. Carcinoid tumor of the appendix: a consecutive series from 1237 appendectomies. World journal of gastroenterology : WJG 2006;**12**(41):6699-701.

16. Connor SJ, Hanna GB, Frizelle FA. Appendiceal tumors: retrospective clinicopathologic analysis of appendiceal tumors from 7,970 appendectomies. Diseases of the colon and rectum 1998;**41**(1):75-80.

17. McCusker ME, Cote TR, Clegg LX, et al. Primary malignant neoplasms of the appendix: a population-based study from the surveillance, epidemiology and end-results program, 1973-1998. Cancer 2002;**94**(12):3307-12.

18. Goede AC, Caplin ME, Winslet MC. Carcinoid tumour of the appendix. The British journal of surgery 2003;**90**(11):1317-22.

19. Roggo A, Wood WC, Ottinger LW. Carcinoid tumors of the appendix. Annals of surgery 1993;**217**(4):385-90.

20. Kanthan R, Saxena A, Kanthan SC. Goblet cell carcinoids of the appendix: immunophenotype and ultrastructural study. Archives of pathology & laboratory medicine 2001;**125**(3):386-90.

21. Anderson JR, Wilson BG. Carcinoid tumours of the appendix. The British journal of surgery 1985;**72**(7):545-6.

22. Goede AC, Winslet MC. Surgery for carcinoid tumours of the lower gastrointestinal tract. Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland 2003;**5**(2):123-8.

23. Sugarbaker PH. The natural history, gross pathology, and histopathology of appendiceal epithelial neoplasms. European journal of surgical oncology : the journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology 2006;**32**(6):644-7.

24. Tang LH. Epithelial neoplasms of the appendix. Archives of pathology & laboratory medicine 2010;**134**(11):1612-20.

25. Yantiss RK, Shia J, Klimstra DS, et al. Prognostic significance of localized extra-appendiceal mucin deposition in appendiceal mucinous neoplasms. The American journal of surgical pathology 2009;**33**(2):248-55.

26. Gonzalez-Moreno S, Brun E, Sugarbaker PH. Lymph node metastasis in epithelial malignancies of the appendix with peritoneal dissemination does not reduce survival in patients treated by cytoreductive surgery and perioperative intraperitoneal chemotherapy. Annals of surgical oncology 2005;**12**(1):72-80.

27. Gonzalez-Moreno S, Sugarbaker PH. Right hemicolectomy does not confer a survival advantage in patients with mucinous carcinoma of the appendix and peritoneal seeding. The British journal of surgery 2004;**91**(3):304-11.

28. Stocchi L, Wolff BG, Larson DR, et al. Surgical treatment of appendiceal mucocele. Archives of surgery 2003;**138**(6):585-9; discussion 89-90.

29. O'Donnell ME, Badger SA, Beattie GC, et al. Malignant neoplasms of the appendix. International journal of colorectal disease 2007;**22**(10):1239-48.

30. Benedix F, Reimer A, Gastinger I, et al. Primary appendiceal carcinoma--epidemiology, surgery and survival: results of a German multi-center study. European journal of surgical oncology : the journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology 2010;**36**(8):763-71.

31. McBurney C. Experience with early operative interference in cases of disease of the vermiform appendix. N Y Med J 1889;**50**:676-84.

32. McBurney C. II. The Indications for Early Laparotomy in Appendicitis. Annals of surgery 1891;**13**(4):233-54.

33. McBurney C. IV. The Incision Made in the Abdominal Wall in Cases of Appendicitis, with a Description of a New Method of Operating. Annals of surgery 1894;**20**(1):38-43.

34. Spencer WG. On Indications for Immediately Opening the Abdomen in Acute Cases. British medical journal 1909;**2**(2556):1789-92.

35. Wenckert A, Robertson B. Oxytetracycline (terramycin) in the treatment of appendicitis with peritonitis. A clinical study of 5,564 consecutive appendectomies. Acta chirurgica Scandinavica 1960;**120**:79-87.

36. Andersson RE. Short-term complications and long-term morbidity of laparoscopic and open appendicectomy in a national cohort. The British journal of surgery 2014;**101**(9):1135-42.

37. Addiss DG, Shaffer N, Fowler BS, et al. The epidemiology of appendicitis and appendectomy in the United States. American journal of epidemiology 1990;**132**(5):910-25.

38. Flum DR, Koepsell T. The clinical and economic correlates of misdiagnosed appendicitis: nationwide analysis. Archives of surgery 2002;**137**(7):799-804; discussion 04.

39. Williams NM, Jackson D, Everson NW, et al. Is the incidence of acute appendicitis really falling? Annals of the Royal College of Surgeons of England 1998;**80**(2):122-4.

40. Blomqvist P, Ljung H, Nyren O, et al. Appendectomy in Sweden 1989-1993 assessed by the Inpatient Registry. Journal of clinical epidemiology 1998;**51**(10):859-65.

41. Livingston EH, Woodward WA, Sarosi GA, et al. Disconnect between incidence of nonperforated and perforated appendicitis: implications for pathophysiology and management. Annals of surgery 2007;**245**(6):886-92.

42. McCahy P. Continuing fall in the incidence of acute appendicitis. Annals of the Royal College of Surgeons of England 1994;**76**(4):282-3.

43. Korner H, Soreide JA, Pedersen EJ, et al. Stability in incidence of acute appendicitis. A population-based longitudinal study. Dig Surg 2001;**18**(1):61-6.

44. Andersson R, Hugander A, Thulin A, et al. Indications for operation in suspected appendicitis and incidence of perforation. Bmj 1994;**308**(6921):107-10.

45. Luckmann R. Incidence and case fatality rates for acute appendicitis in California. A population-based study of the effects of age. American journal of epidemiology 1989;**129**(5):905-18.

46. Soreide O. Appendicitis--a study of incidence, death rates and consumption of hospital resources. Postgraduate medical journal 1984;**60**(703):341-5.

47. Decadt B, Sussman L, Lewis MP, et al. Randomized clinical trial of early laparoscopy in the management of acute non-specific abdominal pain. The British journal of surgery 1999;**86**(11):1383-6.

48. Morino M, Pellegrino L, Castagna E, et al. Acute nonspecific abdominal pain: A randomized, controlled trial comparing early laparoscopy versus clinical observation. Annals of surgery 2006;**244**(6):881-6; discussion 86-8.

49. Wagner JM, McKinney WP, Carpenter JL. Does this patient have appendicitis? JAMA : the journal of the American Medical Association 1996;**276**(19):1589-94.

50. Pieper R, Kager L, Tidefeldt U. Obstruction of appendix vermiformis causing acute appendicitis. An experimental study in the rabbit. Acta chirurgica Scandinavica 1982;**148**(1):63-72.

51. Jones BA, Demetriades D, Segal I, et al. The prevalence of appendiceal fecaliths in patients with and without appendicitis. A comparative study from Canada and South Africa. Annals of surgery 1985;**202**(1):80-2.

52. Nitecki S, Karmeli R, Sarr MG. Appendiceal calculi and fecaliths as indications for appendectomy. Surgery, gynecology & obstetrics 1990;**171**(3):185-8.

53. Klingler PJ, Seelig MH, DeVault KR, et al. Ingested foreign bodies within the appendix: A 100-year review of the literature. Digestive diseases (Basel, Switzerland) 1998;**16**(5):308-14.

54. Arnbjornsson E, Bengmark S. Role of obstruction in the pathogenesis of acute appendicitis. American journal of surgery 1984;**147**(3):390-2.

55. Chang AR. An analysis of the pathology of 3003 appendices. The Australian and New Zealand journal of surgery 1981;**51**(2):169-78.

56. Andersson R, Hugander A, Thulin A, et al. Clusters of acute appendicitis: further evidence for an infectious aetiology. Int J Epidemiol 1995;**24**(4):829-33.

57. Stein GY, Rath-Wolfson L, Zeidman A, et al. Sex differences in the epidemiology, seasonal variation, and trends in the management of patients with acute appendicitis. Langenbeck's archives of surgery / Deutsche Gesellschaft fur Chirurgie 2012;**397**(7):1087-92.

58. Alder AC, Fomby TB, Woodward WA, et al. Association of viral infection and appendicitis. Archives of surgery 2010;**145**(1):63-71.

59. Morris J, Barker DJ, Nelson M. Diet, infection, and acute appendicitis in Britain and Ireland. Journal of epidemiology and community health 1987;**41**(1):44-9.

60. Barker DJ. Acute appendicitis and dietary fibre: an alternative hypothesis. British medical journal (Clinical research ed) 1985;**290**(6475):1125-7.

61. Janeway CAT, Paul. Walport, Mark. Shlomchik, M.J. *Immuno Biology, The immune system in health and disease*. New York, USA: Garland Science Publisher, 2005.

62. Gor DO, Rose NR, Greenspan NS. TH1-TH2: a procrustean paradigm. Nature immunology 2003;**4**(6):503-5.

63. Strober W, Fuss IJ. Proinflammatory cytokines in the pathogenesis of inflammatory bowel diseases. Gastroenterology 2011;**140**(6):1756-67.

64. Andersson RE, Olaison G, Tysk C, et al. Appendectomy is followed by increased risk of Crohn's disease. Gastroenterology 2003;**124**(1):40-6.

65. Frisch M, Pedersen BV, Andersson RE. Appendicitis, mesenteric lymphadenitis, and subsequent risk of ulcerative colitis: cohort studies in Sweden and Denmark. Bmj 2009;**338**:b716.

66. Andersson RE, Olaison G, Tysk C, et al. Appendectomy and protection against ulcerative colitis. The New England journal of medicine 2001;**344**(11):808-14.

67. Marzi M, Vigano A, Trabattoni D, et al. Characterization of type 1 and type 2 cytokine production profile in physiologic and pathologic human pregnancy. Clinical and experimental immunology 1996;**106**(1):127-33.

68. Saito S, Sakai M, Sasaki Y, et al. Quantitative analysis of peripheral blood Th0, Th1, Th2 and the Th1:Th2 cell ratio during normal human pregnancy and preeclampsia. Clinical and experimental immunology 1999;**117**(3):550-5.

69. Andersson RE, Lambe M. Incidence of appendicitis during pregnancy. Int J Epidemiol 2001;**30**(6):1281-5.

70. Zingone F, Sultan AA, Humes DJ, et al. Risk of Acute Appendicitis in and Around Pregnancy: A Population-based Cohort Study From England. Annals of surgery 2015;**261**(2):332-7.

71. Wan YY. Multi-tasking of helper T cells. Immunology 2010;**130**(2):166-71.

72. Andersson RE. Meta-analysis of the clinical and laboratory diagnosis of appendicitis. The British journal of surgery 2004;**91**(1):28-37.

73. Campbell JS, Fournier P, Dasilva T. When is the appendix normal? A study of acute inflammations of the appendix apparent only upon histologic examination. Canadian Medical Association journal 1961;**85**:1155-7.

74. Pieper R, Kager L, Nasman P. Clinical significance of mucosal inflammation of the vermiform appendix. Annals of surgery 1983;**197**(3):368-74.

75. Howie JG. Too Few Appendicectomies? Lancet 1964;**1**(7345):1240-2.

76. Carr NJ. The pathology of acute appendicitis. Annals of diagnostic pathology 2000;**4**(1):46-58.

77. Petrosyan M, Estrada J, Chan S, et al. CT scan in patients with suspected appendicitis: clinical implications for the acute care surgeon. European surgical research Europaische chirurgische Forschung Recherches chirurgicales europeennes 2008;**40**(2):211-9.

78. Rao PM, Rhea JT, Rattner DW, et al. Introduction of appendiceal CT: impact on negative appendectomy and appendiceal perforation rates. Annals of surgery 1999;**229**(3):344-9.

79. Andersson RE. Resolving appendicitis is common: further evidence. Annals of surgery 2008;**247**(3):553; author reply 53.

80. Anderson JE, Bickler SW, Chang DC, et al. Examining a common disease with unknown etiology: trends in epidemiology and surgical management of appendicitis in California, 1995-2009. World journal of surgery 2012;**36**(12):2787-94.

81. Kirshenbaum M, Mishra V, Kuo D, et al. Resolving appendicitis: role of CT. Abdominal imaging 2003;**28**(2):276-9.

82. Cobben LP, de Van Otterloo AM, Puylaert JB. Spontaneously resolving appendicitis: frequency and natural history in 60 patients. Radiology 2000;**215**(2):349-52.

83. Hahn HB, Hoepner FU, Kalle T, et al. Sonography of acute appendicitis in children: 7 years experience. Pediatric radiology 1998;**28**(3):147-51.

84. Ooms HW, Koumans RK, Ho Kang You PJ, et al. Ultrasonography in the diagnosis of acute appendicitis. The British journal of surgery 1991;**78**(3):315-8.

85. Temple CL, Huchcroft SA, Temple WJ. The natural history of appendicitis in adults. A prospective study. Annals of surgery 1995;**221**(3):278-81.

86. White JJ, Santillana M, Haller JA, Jr. Intensive in-hospital observation: a safe way to decrease unnecessary appendectomy. The American surgeon 1975;**41**(12):793-8.

87. Esposito C. One-trocar appendectomy in pediatric surgery. Surgical endoscopy 1998;**12**(2):177-8.

88. Antoniou SA, Koch OO, Antoniou GA, et al. Meta-analysis of randomized trials on single-incision laparoscopic versus conventional laparoscopic appendectomy. American journal of surgery 2014;**207**(4):613-22.

89. Palanivelu C, Rajan PS, Rangarajan M, et al. Transvaginal endoscopic appendectomy in humans: a unique approach to NOTES--world's first report. Surgical endoscopy 2008;**22**(5):1343-7.

90. Masoomi H, Nguyen NT, Dolich MO, et al. Laparoscopic appendectomy trends and outcomes in the United States: data from the Nationwide Inpatient Sample (NIS), 2004-2011. The American surgeon 2014;**80**(10):1074-7.

91. Faiz O, Clark J, Brown T, et al. Traditional and laparoscopic appendectomy in adults: outcomes in English NHS hospitals between 1996 and 2006. Annals of surgery 2008;**248**(5):800-6.

92. Sauerland S, Jaschinski T, Neugebauer EA. Laparoscopic versus open surgery for suspected appendicitis. The Cochrane database of systematic reviews 2010(10):CD001546.

93. Moberg AC, Ahlberg G, Leijonmarck CE, et al. Diagnostic laparoscopy in 1043 patients with suspected acute appendicitis. The European journal of surgery = Acta chirurgica 1998;**164**(11):833-40; discussion 41.

94. Eriksson S, Granstrom L. Randomized controlled trial of appendicectomy versus antibiotic therapy for acute appendicitis. The British journal of surgery 1995;**82**(2):166-9.

95. Styrud J, Eriksson S, Nilsson I, et al. Appendectomy versus antibiotic treatment in acute appendicitis. a prospective multicenter randomized controlled trial. World journal of surgery 2006;**30**(6):1033-7.

96. Hansson J, Korner U, Khorram-Manesh A, et al. Randomized clinical trial of antibiotic therapy versus appendicectomy as primary treatment of acute appendicitis in unselected patients. The British journal of surgery 2009;**96**(5):473-81.

97. Malik AA, Bari SU. Conservative management of acute appendicitis. Journal of gastrointestinal surgery : official journal of the Society for Surgery of the Alimentary Tract 2009;**13**(5):966-70.

98. Vons C, Barry C, Maitre S, et al. Amoxicillin plus clavulanic acid versus appendicectomy for treatment of acute uncomplicated appendicitis: an open-label, non-inferiority, randomised controlled trial. The Lancet 2011;**377**(9777):1573-79.

99. Turhan AN, Kapan S, Kutukcu E, et al. Comparison of operative and non operative management of acute appendicitis. Ulusal travma ve acil cerrahi dergisi = Turkish journal of trauma & emergency surgery : TJTES 2009;**15**(5):459-62.

100. Kirby A, Hobson RP, Burke D, et al. Appendicectomy for suspected uncomplicated appendicitis is associated with fewer complications than conservative antibiotic management: A meta-analysis of post-intervention complications. The Journal of infection 2015;**70**(2):105-10.

101. Varadhan KK, Neal KR, Lobo DN. Safety and efficacy of antibiotics compared with appendicectomy for treatment of uncomplicated acute appendicitis: meta-analysis of randomised controlled trials. Bmj 2012;**344**:e2156.

102. Mason RJ, Moazzez A, Sohn H, et al. Meta-analysis of randomized trials comparing antibiotic therapy with appendectomy for acute uncomplicated (no abscess or phlegmon) appendicitis. Surgical infections 2012;**13**(2):74-84.

103. Wilms IM, de Hoog DE, de Visser DC, et al. Appendectomy versus antibiotic treatment for acute appendicitis. The Cochrane database of systematic reviews 2011(11):CD008359.

104. Shindoh J, Niwa H, Kawai K, et al. Predictive factors for negative outcomes in initial non-operative management of suspected appendicitis. Journal of gastrointestinal surgery : official journal of the Society for Surgery of the Alimentary Tract 2010;**14**(2):309-14.

105. Tsai HM, Shan YS, Lin PW, et al. Clinical analysis of the predictive factors for recurrent appendicitis after initial nonoperative treatment of perforated appendicitis. American journal of surgery 2006;**192**(3):311-6.

106. Di Saverio S, Sibilio A, Giorgini E, et al. The NOTA Study (Non Operative Treatment for Acute Appendicitis): prospective study on the efficacy and safety of antibiotics (amoxicillin and clavulanic acid) for treating patients with right lower quadrant abdominal pain and long-term follow-up of conservatively treated suspected appendicitis. Annals of surgery 2014;**260**(1):109-17.

107. Lane JS, Schmit PJ, Chandler CF, et al. Ileocecectomy is definitive treatment for advanced appendicitis. The American surgeon 2001;**67**(12):1117-22.

108. Andersson RE, Petzold MG. Nonsurgical treatment of appendiceal abscess or phlegmon: a systematic review and meta-analysis. Annals of surgery 2007;**246**(5):741-8.

109. Margenthaler JA, Longo WE, Virgo KS, et al. Risk factors for adverse outcomes after the surgical treatment of appendicitis in adults. Annals of surgery 2003;**238**(1):59-66.

110. Andersen BR, Kallehave FL, Andersen HK. Antibiotics versus placebo for prevention of postoperative infection after appendicectomy. The Cochrane database of systematic reviews 2005(3):CD001439.

111. Markar SR, Penna M, Harris A. Laparoscopic approach to appendectomy reduces the incidence of short- and long-term post-operative bowel obstruction: systematic review and pooled analysis. Journal of gastrointestinal surgery : official journal of the Society for Surgery of the Alimentary Tract 2014;**18**(9):1683-92.

112. Wei B, Qi CL, Chen TF, et al. Laparoscopic versus open appendectomy for acute appendicitis: a metaanalysis. Surgical endoscopy 2011;**25**(4):1199-208.

113. Loveland JE. Reginald Heber Fitz, The Exponent of Appendicitis. The Yale journal of biology and medicine 1937;**9**(6):509 b1-20.

114. Blomqvist PG, Andersson RE, Granath F, et al. Mortality after appendectomy in Sweden, 1987-1996. Annals of surgery 2001;**233**(4):455-60.

115. Andersson MN, Andersson RE. Causes of short-term mortality after appendectomy: a population-based case-controlled study. Annals of surgery 2011;**254**(1):103-7.

116. Ashdown HF, D'Souza N, Karim D, et al. Pain over speed bumps in diagnosis of acute appendicitis: diagnostic accuracy study. Bmj 2012;**345**:e8012.

117. Golledge J, Toms AP, Franklin IJ, et al. Assessment of peritonism in appendicitis. Annals of the Royal College of Surgeons of England 1996;**78**(1):11-4.

118. Alaedeen DI, Cook M, Chwals WJ. Appendiceal fecalith is associated with early perforation in pediatric patients. Journal of pediatric surgery 2008;**43**(5):889-92.

119. Grimes C, Chin D, Bailey C, et al. Appendiceal faecaliths are associated with right iliac fossa pain. Annals of the Royal College of Surgeons of England 2010;**92**(1):61-4.

120. Andersson RE, Hugander AP, Ghazi SH, et al. Why does the clinical diagnosis fail in suspected appendicitis? The European journal of surgery = Acta chirurgica 2000;**166**(10):796-802.

121. Kraemer M, Franke C, Ohmann C, et al. Acute appendicitis in late adulthood: incidence, presentation, and outcome. Results of a prospective multicenter acute abdominal pain study and a review of the literature. Langenbeck's Archives of Surgery 2000;**385**(7):470-81.

122. Klein MD. Clinical approach to a child with abdominal pain who might have appendicitis. Pediatric radiology 2007;**37**(1):11-4.

123. Jahn H, Mathiesen FK, Neckelmann K, et al. Comparison of clinical judgment and diagnostic ultrasonography in the diagnosis of acute appendicitis: experience with a score-aided diagnosis. The European journal of surgery = Acta chirurgica 1997;**163**(6):433-43.

124. Andersson RE, Hugander AP, Ghazi SH, et al. Diagnostic value of disease history, clinical presentation, and inflammatory parameters of appendicitis. World journal of surgery 1999;**23**(2):133-40.

125. Wray CJ, Kao LS, Millas SG, et al. Acute appendicitis: controversies in diagnosis and management. Current problems in surgery 2013;**50**(2):54-86.

126. Doraiswamy NV. Leucocyte counts in the diagnosis and prognosis of acute appendicitis in children. The British journal of surgery 1979;**66**(11):782-4.

127. Raftery AT. The value of the leucocyte count in the diagnosis of acute appendicitis. The British journal of surgery 1976;**63**(2):143-4.

128. Tsuji M, McMahon G, Reen D, et al. New insights into the pathogenesis of appendicitis based on immunocytochemical analysis of early immune response. Journal of pediatric surgery 1990;**25**(4):449-52.

129. Eriksson S, Granstrom L, Carlstrom A. The diagnostic value of repetitive preoperative analyses of C-reactive protein and total leucocyte count in patients with suspected acute appendicitis. Scand J Gastroenterol 1994;**29**(12):1145-9.

130. Branch DW. Physiologic adaptations of pregnancy. American journal of reproductive immunology (New York, NY : 1989) 1992;**28**(3-4):120-2.

131. Mölne JW, A. *Inflammation*. 1st ed. Stockholm: Liber AB, 2007.

132. Wright HL, Moots RJ, Bucknall RC, et al. Neutrophil function in inflammation and inflammatory diseases. Rheumatology 2010;**49**(9):1618-31.

133. Hallan S, Asberg A, Edna TH. Additional value of biochemical tests in suspected acute appendicitis. The European journal of surgery = Acta chirurgica 1997;**163**(7):533-8.

134. Hurlimann J, Thorbecke GJ, Hochwald GM. The liver as the site of C-reactive protein formation. The Journal of experimental medicine 1966;**123**(2):365-78.

135. Zimmerman MA, Selzman CH, Cothren C, et al. Diagnostic implications of C-reactive protein. Archives of surgery 2003;**138**(2):220-4.

136. Steel DM, Whitehead AS. The major acute phase reactants: C-reactive protein, serum amyloid P component and serum amyloid A protein. Immunology today 1994;**15**(2):81-8.

137. Tillett WS, Francis T. Serological Reactions in Pneumonia with a Non-Protein Somatic Fraction of Pneumococcus. The Journal of experimental medicine 1930;**52**(4):561-71.

138. Lelubre C, Anselin S, Zouaoui Boudjeltia K, et al. Interpretation of C-reactive protein concentrations in critically ill patients. BioMed research international 2013;**2013**:124021.

139. Xia D, Samols D. Transgenic mice expressing rabbit C-reactive protein are resistant to endotoxemia. Proceedings of the National Academy of Sciences of the United States of America 1997;**94**(6):2575-80.

140. Paajanen H, Mansikka A, Laato M, et al. Are serum inflammatory markers age dependent in acute appendicitis? Journal of the American College of Surgeons 1997;**184**(3):303-8.

141. Yu CW, Juan LI, Wu MH, et al. Systematic review and meta-analysis of the diagnostic accuracy of procalcitonin, C-reactive protein and white blood cell count for suspected acute appendicitis. The British journal of surgery 2013;**100**(3):322-9.

142. Rivera-Chavez FA, Wheeler H, Lindberg G, et al. Regional and systemic cytokine responses to acute inflammation of the vermiform appendix. Annals of surgery 2003;**237**(3):408-16.

143. Kharbanda AB, Cosme Y, Liu K, et al. Discriminative accuracy of novel and traditional biomarkers in children with suspected appendicitis adjusted for duration of abdominal pain. Academic emergency medicine : official journal of the Society for Academic Emergency Medicine 2011;**18**(6):567-74.

144. Yoon DY, Chu J, Chandler C, et al. Human cytokine levels in nonperforated versus perforated appendicitis: molecular serum markers for extent of disease? The American surgeon 2002;**68**(12):1033-7.

145. Yildirim O, Solak C, Kocer B, et al. The role of serum inflammatory markers in acute appendicitis and their success in preventing negative laparotomy. Journal of investigative surgery : the official journal of the Academy of Surgical Research 2006;**19**(6):345-52.

146. Lycopoulou L, Mamoulakis C, Hantzi E, et al. Serum amyloid A protein levels as a possible aid in the diagnosis of acute appendicitis in children. Clinical chemistry and laboratory medicine : CCLM / FESCC 2005;**43**(1):49-53.

147. Ruber M, Andersson M, Petersson BF, et al. Systemic Th17-like cytokine pattern in gangrenous appendicitis but not in phlegmonous appendicitis. Surgery 2010;**147**(3):366-72.

148. Reed JL, Strait RT, Kachelmeyer AM, et al. Biomarkers to distinguish surgical etiologies in females with lower quadrant abdominal pain. Academic emergency medicine : official journal of the Society for Academic Emergency Medicine 2011;**18**(7):686-91.

149. Allister L, Bachur R, Glickman J, et al. Serum markers in acute appendicitis. The Journal of surgical research 2011;**168**(1):70-5.

150. Sack U, Biereder B, Elouahidi T, et al. Diagnostic value of blood inflammatory markers for detection of acute appendicitis in children. BMC surgery 2006;**6**:15.

151. Velanovich V, Satava R. Balancing the normal appendectomy rate with the perforated appendicitis rate: implications for quality assurance. The American surgeon 1992;**58**(4):264-9.

152. Hall EJ, Brenner DJ. Cancer risks from diagnostic radiology. The British journal of radiology 2008;**81**(965):362-78.

153. Fish B, Smulewicz JJ, Barek L. Role of computed tomography in diagnosis of appendiceal disorders. New York state journal of medicine 1981;**81**(6):900-4.

154. Balthazar EJ, Megibow AJ, Hulnick D, et al. CT of appendicitis. AJR American journal of roentgenology 1986;**147**(4):705-10.

155. McDonald GP, Pendarvis DP, Wilmoth R, et al. Influence of preoperative computed tomography on patients undergoing appendectomy. The American surgeon 2001;**67**(11):1017-21.

156. Weyant MJ, Eachempati SR, Maluccio MA, et al. Interpretation of computed tomography does not correlate with laboratory or pathologic findings in surgically confirmed acute appendicitis. Surgery 2000;**128**(2):145-52.

157. Keyzer C, Tack D, de Maertelaer V, et al. Acute appendicitis: comparison of low-dose and standard-dose unenhanced multi-detector row CT. Radiology 2004;**232**(1):164-72.

158. Benjaminov O, Atri M, Hamilton P, et al. Frequency of visualization and thickness of normal appendix at nonenhanced helical CT. Radiology 2002;**225**(2):400-6.

159. Rao PM, Rhea JT, Novelline RA. Sensitivity and specificity of the individual CT signs of appendicitis: experience with 200 helical appendiceal CT examinations. Journal of computer assisted tomography 1997;**21**(5):686-92.

160. Choi D, Park H, Lee YR, et al. The most useful findings for diagnosing acute appendicitis on contrast-enhanced helical CT. Acta Radiologica 2003;**44**(6):574-82.

161. Terasawa T, Blackmore CC, Bent S, et al. Systematic review: computed tomography and ultrasonography to detect acute appendicitis in adults and adolescents. Annals of internal medicine 2004;**141**(7):537-46.

162. van Randen A, Bipat S, Zwinderman AH, et al. Acute appendicitis: meta-analysis of diagnostic performance of CT and graded compression US related to prevalence of disease. Radiology 2008;**249**(1):97-106.

163. van Randen A, Lameris W, Nio CY, et al. Inter-observer agreement for abdominal CT in unselected patients with acute abdominal pain. European radiology 2009;**19**(6):1394-407.

164. Keyzer C, Cullus P, Tack D, et al. MDCT for suspected acute appendicitis in adults: impact of oral and IV contrast media at standard-dose and simulated low-dose techniques. AJR American journal of roentgenology 2009;**193**(5):1272-81.

165. Krajewski S. Impact of computed tomography of the abdomen on clinical outcomes in patients with acute right lower quadrant pain: a meta-analysis. Canadian Journal of Surgery 2011;**54**(1):43-53.

166. Flum DR, McClure TD, Morris A, et al. Misdiagnosis of appendicitis and the use of diagnostic imaging. Journal of the American College of Surgeons 2005;**201**(6):933-9.

167. Flum DR, Morris A, Koepsell T, et al. Has misdiagnosis of appendicitis decreased over time? A population-based analysis. JAMA : the journal of the American Medical Association 2001;**286**(14):1748-53.

168. Walker S, Haun W, Clark J, et al. The value of limited computed tomography with rectal contrast in the diagnosis of acute appendicitis. American journal of surgery 2000;**180**(6):450-4; discussion 54-5.

169. Hong JJ, Cohn SM, Ekeh AP, et al. A prospective randomized study of clinical assessment versus computed tomography for the diagnosis of acute appendicitis. Surgical infections 2003;**4**(3):231-9.

170. Lee CC, Golub R, Singer AJ, et al. Routine versus selective abdominal computed tomography scan in the evaluation of right lower quadrant pain: a randomized controlled trial. Academic emergency medicine : official journal of the Society for Academic Emergency Medicine 2007;**14**(2):117-22.

171. Brenner DJ, Hall EJ. Computed tomography--an increasing source of radiation exposure. The New England journal of medicine 2007;**357**(22):2277-84.

172. Preston DL, Pierce DA, Shimizu Y, et al. Effect of recent changes in atomic bomb survivor dosimetry on cancer mortality risk estimates. Radiation research 2004;**162**(4):377-89.

173. Cardis E, Vrijheid M, Blettner M, et al. Risk of cancer after low doses of ionising radiation: retrospective cohort study in 15 countries. Bmj 2005;**331**(7508):77.

174. Mathews JD, Forsythe AV, Brady Z, et al. Cancer risk in 680,000 people exposed to computed tomography scans in childhood or adolescence: data linkage study of 11 million Australians. Bmj 2013;**346**:f2360.

175. Deutsch A, Leopold GR. Ultrasonic demonstration of the inflamed appendix: case report. Radiology 1981;**140**(1):163-4.

176. Puylaert JB. Acute appendicitis: US evaluation using graded compression. Radiology 1986;**158**(2):355-60.

177. Quillin SP, Siegel MJ. Appendicitis in children: color Doppler sonography. Radiology 1992;**184**(3):745-7.

178. Rettenbacher T, Hollerweger A, Macheiner P, et al. Outer diameter of the vermiform appendix as a sign of acute appendicitis: evaluation at US. Radiology 2001;**218**(3):757-62.

179. Rettenbacher T, Hollerweger A, Macheiner P, et al. Ovoid shape of the vermiform appendix: a criterion to exclude acute appendicitis--evaluation with US. Radiology 2003;**226**(1):95-100.

180. Jeffrey RB, Jain KA, Nghiem HV. Sonographic diagnosis of acute appendicitis: interpretive pitfalls. AJR American journal of roentgenology 1994;**162**(1):55-9.

181. Strouse PJ. Pediatric appendicitis: an argument for US. Radiology 2010;**255**(1):8-13.

182. Kaiser S, Frenckner B, Jorulf HK. Suspected appendicitis in children: US and CT--a prospective randomized study. Radiology 2002;**223**(3):633-8.

183. Rettenbacher T, Hollerweger A, Gritzmann N, et al. Appendicitis: should diagnostic imaging be performed if the clinical presentation is highly suggestive of the disease? Gastroenterology 2002;**123**(4):992-8.

184. Butler H, Bryan PJ, LiPuma JP, et al. Magnetic resonance imaging of the abnormal female pelvis. AJR American journal of roentgenology 1984;**143**(6):1259-66.

185. Lam M, Singh A, Kaewlai R, et al. Magnetic resonance of acute appendicitis: pearls and pitfalls. Current problems in diagnostic radiology 2008;**37**(2):57-66.

186. Cobben L, Groot I, Kingma L, et al. A simple MRI protocol in patients with clinically suspected appendicitis: results in 138 patients and effect on outcome of appendectomy. European radiology 2009;**19**(5):1175-83.

187. Barger RL, Jr., Nandalur KR. Diagnostic performance of magnetic resonance imaging in the detection of appendicitis in adults: a meta-analysis. Academic radiology 2010;**17**(10):1211-6.

188. Knottnerus JA, van Weel C, Muris JW. Evaluation of diagnostic procedures. Bmj 2002;**324**(7335):477-80.

189. Alvarado A. A practical score for the early diagnosis of acute appendicitis. Ann Emerg Med 1986;**15**(5):557-64.

190. Samuel M. Pediatric appendicitis score. Journal of pediatric surgery 2002;**37**(6):877-81.

191. Lintula H, Pesonen E, Kokki H, et al. A diagnostic score for children with suspected appendicitis. Langenbeck's archives of surgery / Deutsche Gesellschaft fur Chirurgie 2005;**390**(2):164-70.

192. Andersson M, Andersson RE. The appendicitis inflammatory response score: a tool for the diagnosis of acute appendicitis that outperforms the Alvarado score. World journal of surgery 2008;**32**(8):1843-9.

193. Sammalkorpi HE, Mentula P, Leppaniemi A. A new adult appendicitis score improves diagnostic accuracy of acute appendicitis--a prospective study. BMC gastroenterology 2014;**14**:114.

194. Ohmann C, Yang Q, Franke C. Diagnostic scores for acute appendicitis. Abdominal Pain Study Group. The European journal of surgery = Acta chirurgica 1995;**161**(4):273-81.

195. Kulik DM, Uleryk EM, Maguire JL. Does this child have appendicitis? A systematic review of clinical prediction rules for children with acute abdominal pain. Journal of clinical epidemiology 2013;**66**(1):95-104.

196. Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. Journal of clinical epidemiology 2003;**56**(11):1118-28.

197. Van den Bruel A, Cleemput I, Aertgeerts B, et al. The evaluation of diagnostic tests: evidence on technical and diagnostic accuracy, impact on patient outcome and cost-effectiveness is needed. Journal of clinical epidemiology 2007;**60**(11):1116-22.

198. Man E, Simonka Z, Varga A, et al. Impact of the Alvarado score on the diagnosis of acute appendicitis: comparing clinical judgment, Alvarado score, and a new modified score in suspected appendicitis: a prospective, randomized clinical trial. Surgical endoscopy 2014;**28**(8):2398-405.

199. Ohmann C, Franke C, Yang Q, et al. [Diagnostic score for acute appendicitis]. Der Chirurg; Zeitschrift fur alle Gebiete der operativen Medizen 1995;**66**(2):135-41.

200. Ohmann C, Franke C, Yang Q. Clinical benefit of a diagnostic score for appendicitis: results of a prospective interventional study. German Study Group of Acute Abdominal Pain. Archives of surgery 1999;**134**(9):993-6.

201. Lintula H, Kokki H, Kettunen R, et al. Appendicitis score for children with suspected appendicitis. A randomized clinical trial. Langenbeck's archives of surgery / Deutsche Gesellschaft fur Chirurgie 2009;**394**(6):999-1004.

202. Lintula H, Kokki H, Pulkkinen J, et al. Diagnostic score in acute appendicitis. Validation of a diagnostic score (Lintula score) for adults with suspected appendicitis. Langenbeck's archives of surgery / Deutsche Gesellschaft fur Chirurgie 2010;**395**(5):495-500.

203. Fleischman RJ, Devine MK, Yagapen MA, et al. Evaluation of a novel pediatric appendicitis pathway using high- and low-risk scoring systems. Pediatric emergency care 2013;**29**(10):1060-5.

204. Russell WS, Schuh AM, Hill JG, et al. Clinical practice guidelines for pediatric appendicitis evaluation can decrease computed tomography utilization while maintaining diagnostic accuracy. Pediatric emergency care 2013;**29**(5):568-73.

205. Antevil JL, Rivera L, Langenberg BJ, et al. Computed tomography-based clinical diagnostic pathway for acute appendicitis: prospective validation. Journal of the American College of Surgeons 2006;**203**(6):849-56.

206. Coste J, Pouchot J. A grey zone for quantitative diagnostic and screening tests. Int J Epidemiol 2003;**32**(2):304-13.

207. Fagan TJ. Letter: Nomogram for Bayes theorem. The New England journal of medicine 1975;**293**(5):257.

208. Metz CE. ROC methodology in radiologic imaging. Investigative radiology 1986;**21**(9):720-33.

209. Macascill PG, C. Deeks, J.J. Harbord, R.M. Takwoingi, Y. *Analysing and Presenting Results*: The Cochrane Collaboration, 2010.

210. Barnard J, Meng XL. Applications of multiple imputation in medical studies: from AIDS to NHANES. Statistical methods in medical research 1999;**8**(1):17-36.

211. DB R. Inference and missing data. Biometrika 1976(63):581-92.

212. Steyerberg EW. *Clinical Prediction Models. A Practical Approach to Development, Validation, and Updating*. New York: Springer Science+Business Media, LCC, 2009.

213. Stuart EA, Azur M, Frangakis C, et al. Multiple imputation with large data sets: a case study of the Children's Mental Health Initiative. American journal of epidemiology 2009;**169**(9):1133-9.

214. Enders CK. A primer on the use of modern missing-data methods in psychosomatic medicine research. Psychosomatic medicine 2006;**68**(3):427-36.

215. P R. Multiple imputation of missing data. The Stata Jounal 2004(4):227-41.

216. Guyatt GH, Tugwell PX, Feeny DH, et al. A framework for clinical evaluation of diagnostic technologies. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne 1986;**134**(6):587-94.

217. Ledley RS, Lusted LB. Reasoning foundations of medical diagnosis; symbolic logic, probability, and value theory aid our understanding of how physicians reason. Science 1959;**130**(3366):9-21.

218. Bossuyt PM, Irwig L, Craig J, et al. Comparative accuracy: assessing new tests against existing diagnostic pathways. Bmj 2006;**332**(7549):1089-92.

219. Sackett DL, Haynes RB. The architecture of diagnostic research. Bmj 2002;**324**(7336):539-41.

220. Thiese MS. Observational and interventional study design types; an overview. Biochemia medica 2014;**24**(2):199-210.

221. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet 2005;**365**(9453):82-93.

222. Altman DG, Royston P. The cost of dichotomising continuous variables. Bmj 2006;**332**(7549):1080.

223. Moons KG, Royston P, Vergouwe Y, et al. Prognosis and prognostic research: what, why, and how? Bmj 2009;**338**:b375.

224. de Castro SM, Unlu C, Steller EP, et al. Evaluation of the appendicitis inflammatory response score for patients with acute appendicitis. World journal of surgery 2012;**36**(7):1540-5.

225. Kollar D, McCartan DP, Bourke M, et al. Predicting Acute Appendicitis? A comparison of the Alvarado Score, the Appendicitis Inflammatory Response Score and Clinical Assessment. World journal of surgery 2015;**39**(1):104-9.

226. Schulz KF, Grimes DA. Generation of allocation sequences in randomised trials: chance, not choice. Lancet 2002;**359**(9305):515-9.

227. Lachin JM, Matts JP, Wei LJ. Randomization in clinical trials: conclusions and recommendations. Controlled clinical trials 1988;**9**(4):365-74.

228. Herbison P, Hay-Smith J, Gillespie WJ. Different methods of allocation to groups in randomized trials are associated with different levels of bias. A meta-epidemiological study. Journal of clinical epidemiology 2011;**64**(10):1070-5.

229. Schulz KF, Grimes DA. Allocation concealment in randomised trials: defending against deciphering. Lancet 2002;**359**(9306):614-8.

230. Wood L, Egger M, Gluud LL, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. Bmj 2008;**336**(7644):601-5.

231. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Annals of internal medicine 1999;**130**(6):515-24.

232. Kraemer M, Kremer K, Leppert R, et al. Perforating appendicitis: is it a separate disease? Acute Abdominal Pain Study Group. The European journal of surgery = Acta chirurgica 1999;**165**(5):473-80.

233. Altman DG, Vergouwe Y, Royston P, et al. Prognosis and prognostic research: validating a prognostic model. Bmj 2009;**338**:b605.

234. Andersson M, Ruber M, Ekerfelt C, et al. Can new inflammatory markers improve the diagnosis of acute appendicitis? World journal of surgery 2014;**38**(11):2777-83.

235. Steyerberg EW, Eijkemans MJ, Boersma E, et al. Applicability of clinical prediction models in acute myocardial infarction: a comparison of traditional and empirical Bayes adjustment methods. American heart journal 2005;**150**(5):920.

236. Graff L, Radford MJ, Werne C. Probability of appendicitis before and after observation. Ann Emerg Med 1991;**20**(5):503-7.

237. Drake FT, Florence MG, Johnson MG, et al. Progress in the diagnosis of appendicitis: a report from Washington State's Surgical Care and Outcomes Assessment Program. Annals of surgery 2012;**256**(4):586-94.

238. Unlu C, de Castro SM, Tuynman JB, et al. Evaluating routine diagnostic imaging in acute appendicitis. Int J Surg 2009;**7**(5):451-5.

239. Douglas CD, Macpherson NE, Davidson PM, et al. Randomised controlled trial of ultrasonography in diagnosis of acute appendicitis, incorporating the Alvarado score. Bmj 2000;**321**(7266):919-22.

240. Lopez PP, Cohn SM, Popkin CA, et al. The use of a computed tomography scan to rule out appendicitis in women of childbearing age is as accurate as clinical examination: a prospective randomized trial. The American surgeon 2007;**73**(12):1232-6.

APPENDIXES

Papers

The articles associated with this thesis have been removed for copyright reasons. For more details about these see:
http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-113766