Prospective Validation of the Pediatric Appendicitis Score in a Canadian Pediatric Emergency Department

Maala Bhatt MD, FRCPC

The Department of Epidemiology, Biostatistics and Occupational Health
McGill University, Montreal, May 2008

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Masters in Epidemiology © Maala Bhatt 2008
Table of Contents

Preface:  
Acknowledgements i  
Statement of Originality ii  
Abstract iii

Chapter 1:  Introduction  
1.1 Background 1  
1.2 Thesis Objectives 3

Chapter 2:  Literature Review  
2.1 The Appendix 5  
2.2 The Development of Appendicitis 5  
2.3 Symptoms of Appendicitis 6  
2.4 The Diagnostic Challenge in Childhood Appendicitis 7  
2.5 The Consequences of a Delayed Diagnosis 8  
2.6 The Therapeutic Dilemma 9  
2.7 Laboratory Investigations in the Diagnosis of Appendicitis  
   2.7.1 White Blood Cell Count 10  
   2.7.2 C-Reactive Protein 12  
2.8 Imaging investigations in the Diagnosis of Appendicitis  
   2.8.1 Ultrasound 14  
   2.8.2 Computed Tomography 19  
2.9 Clinical Scoring Systems in the Diagnosis of Appendicitis 31  
2.10 Summary of Literature Review 44

Chapter 3:  Materials and Methods 46

Chapter 4:  Results 50  
4.1 Tables 53  
4.2 Figures 60

Chapter 5:  Discussion 62

Chapter 6:  Bibliography 67

Chapter 7:  Appendices  
A. Human Ethics Approval  
B. Information & Consent and Assent Forms  
C. Data Collection Form
Acknowledgements

I am grateful for the tremendous guidance and support that I have received from my thesis supervisors and colleagues at the Montreal Children’s Hospital from the time of this project’s conception to its culmination. I am indebted to my thesis supervisor, Dr. Lawrence Joseph, for his trust, invaluable advice and careful reviews, and without whom this thesis would still be a work in progress. To my committee member, Dr. Francine Ducharme, for her critical revisions to the intellectual content of the thesis; to my friend, colleague and mentor, Dr. David McGillivray for his direction in conducting this investigation and to Dr. Harley Eisman, for his indulgence in providing me with dedicated time to complete my degree.

I would also like to acknowledge the emergency department research assistants who helped identify patients for recruitment, the many attending physicians, residents and fellows who took the time to recruit patients during busy emergency department shifts and the patients and parents who selflessly participated in this study. Without the dedication and altruism of these groups, the completion of this study would not have been possible.

Finally, I wish to thank my family for all of their support during this seemingly never-ending journey of higher education: to my sister Reena for showing me that it is possible to achieve your dreams and for her tenacity in ensuring that I didn’t become a lawyer; to my partner Dan for his love of the ocean and his spirited pursuit of ideas; and most importantly, to my parents, Saroj and Jay, for inspiring academic achievement in a loving home environment rich with experiences and filled with laughter.
Statement of Originality

The research presented in this thesis represents original work on behalf of the author and other contributors. Dr. Bhatt was responsible for the conception of this project, the development of the research protocol, data entry and analysis, the literature review and drafting of the thesis. Dr. McGillivray assisted with the development of the research protocol and the conduct of the investigation. Dr. Joseph contributed greatly in the drafting of the thesis. Dr. Ducharme made critical revisions to the intellectual content of the thesis.
Abstract

Objective: Clinical scoring systems attempt to improve the diagnostic accuracy of pediatric appendicitis. The Pediatric Appendicitis Score (PAS) was the first score created specifically for children and showed excellent performance in the derivation study when administered by pediatric surgeons. Our objective was to validate the score in a non-referred population by medical physicians.

Methods: A convenience sample of children, 4 - 18 years old with suspected appendicitis, presenting to the Montreal Children’s Hospital emergency department was prospectively evaluated. Score components were collected on standardized forms by emergency physicians who were blind to final diagnosis. Outcomes were ascertained by review of pathologic specimens for children undergoing surgery and by telephone follow-up for children who were discharged home. Patients with histology confirmed appendicitis and those without appendicitis (includes patients who underwent appendectomy but who had negative histology) were compared. We present information about the diagnostic accuracy of the Pediatric Appendicitis Score.

Results: Of the enrolled children who met inclusion criteria (n=246), 83 (34%) had pathology proven appendicitis. Using the single cut-point suggested in the derivation study (PAS 5), yielded an unacceptably high number of false positives (37.6%) and negatives (7.2%). The score’s performance improved when two cut-points were used. When children with PAS ≤ 4 were discharged home without further investigations, the sensitivity was 97.6% with NPV 97.7%. When a PAS of ≥ 8 determined the need for appendectomy, the score’s specificity was 95.1% with PPV 85.2%. Using this strategy, the negative appendectomy rate was 8.8%, missed appendicitis rate was 2.4% and 41% of imaging investigations would have been avoided.

Conclusion: The PAS is a useful tool in the evaluation of children with possible appendicitis. Scores ≤ 4 help rule out appendicitis while scores ≥8 help predict appendicitis. Patients with PAS 5-7 may need further radiological evaluation.
Résumé

Objectif: Les systèmes de scores cliniques tentent d'améliorer l'exactitude du diagnostic de l'appendicite aiguë chez l'enfant. Le Score d'Appendicite Pédiatrique (PAS) était le premier score créé spécifiquement pour les enfants et avait démontré une excellente performance dans l'étude de dérivation lorsqu'il était administrée par des chirurgiens pédiatriques. Notre objectif était de valider le score dans une population non-référée par les médecins médicaux.

Méthodes: Un groupe de convenance d'enfants, âgés de 4 à 18 ans, se présentant à l'urgence de l'Hôpital de Montréal pour Enfants avec une suspicion d'appendicite a été évalué prospectivement. Les composantes du score étaient collectées sur des formulaires standardisés par les médecins d'urgence, à aveugle. Les résultats ont été établis par l'examen des spécimens pathologiques pour les enfants subissant la chirurgie et par suivi de téléphonique pour les enfants qui ont eu leur congé de l'hôpital. Les patients présentant l'histologie ont confirmé l'appendicite et ceux sans appendicite (inclus des patients qui a subi l'appendectomy mais qui a eu l'histologie négative) ont été comparés.

Nous présentons des informations sur l'exactitude diagnostique du Score d'Appendicite Pédiatrique.

Résultats: Des enfants inscrits qui ont répondu aux critères d'inclusion (n=246), 83 (34%) ont eu la pathologie appendicite prouvée. En utilisant le seuil suggéré dans l'étude de dérivation (PAS 5), rapporté un nombre exagérément élevé de faux positifs (37.6%) et de faux négatifs (7.2%) a été obtenu. La performance du score était améliorée lorsque deux seuils étaient employés. Quand des enfants avec un score PAS de ≤ 4 ont eu leur congé sans investigations supplémentaires, la sensibilité était 97.6% avec une valeur négatif prédictive de 97.7%. Quand un PAS ≥ 8 a déterminé la nécessité d'une appendicectomy, la spécificité du score était 95.1% avec une valeur positif prédictive de 85.2%. Utilisant cette stratégie, le taux négatif d'appendicetomie était de 8.8%, le taux manqué d'appendicites était 2.4% et 41% des investigations par imagerie auraient été évitées.

Conclusion: Le PAS est un outil utile dans l'évaluation des enfants avec une appendicite possible. Des scores de ≤ 4 éliminent l'appendicite tandis que l'aide du ≥ 8 de points prédissent une appendicite. Les patients présentant un PAS entre 5 et 7 ont besoin d'évaluation radiologique.
Appendicitis is the most common atraumatic surgical abdominal disorder in children over 2 years of age[1-3], with approximately 70,000 pediatric appendectomies performed each year in the United States. The diagnosis of appendicitis continues to be problematic in young children because of overlapping signs and symptoms with other common childhood illnesses. Misdiagnosis rates are high; it is estimated that one-third of children with appendicitis have been previously evaluated by a physician for their symptoms[4]. Delayed or missed diagnoses have the potential to result in significant morbidity from appendiceal perforation, abscess formation, wound infection and wound dehiscence[5-7]. The resulting hospital stays tend to be longer, more complicated and have an important impact on the child and family. Conversely, performing appendectomies on all children with a suspicion of appendicitis, to avoid this delay in diagnosis, carries risks as well. Children who have an appendectomy without evidence of appendicitis on pathologic review (termed a negative appendectomy) are unnecessarily subject to the risks associated with a general anesthetic. The risks of adverse outcomes secondary to general anesthesia or the surgery itself are small, especially in healthy children, but they are not insignificant. The anesthetic related mortality rate is approximately 1 in 100,000 and the morbidity rate is estimated between 1% and 4% of all sedations[8].

Unfortunately, no error-free diagnostic test exists for appendicitis. Laboratory investigations (eg. complete blood count, C-reactive protein and urinalysis) are often ordered during the initial evaluation, but are known to be imperfect indicators of appendicitis when used on their own. A prospective study by Cardall et al. evaluating the discriminatory value of the total WBC in patients with suspected appendicitis, found that patients with a WBC greater than 10,000 had a sensitivity
of 76% (95% CI: 65%, 84%) and a specificity of 52% (95% CI: 45%, 60%) for the diagnosis of appendicitis[9]. A combination of clinical judgment and laboratory investigation is also non-specific, with an average negative appendectomy rate of 10% – 20%[9, 10].

Imaging with ultrasound has been used to aid diagnosis since the 1980s. This technique is inexpensive, rapid, non-invasive and requires no patient preparation. The reported sensitivity and specificity each exceed 90% when the operator is experienced[11]. However, other reports of the sensitivity and specificity of ultrasound in children with suspected appendicitis have ranged from 44% to 94% and 47% to 95%, respectively[12-18]. As evidenced by the variability in reports of ultrasound sensitivities and specificities, either not all operators are experienced, or these reports are optimistic in some settings. The accuracy of the test is further limited by several patient factors such as obesity, pain which does not allow compression of the abdomen, appendicitis complicated by abscess formation and perforation of the appendix[19].

In the mid-1990s computed tomography (CT) overtook ultrasound as the principal imaging technique for appendicitis. It proved to be an extremely accurate diagnostic tool that was less influenced by operator experience or by patient factors. A recent meta-analysis reported an overall sensitivity of 94% (95% CI: 91%, 95%) and specificity of 95% (95% CI: 93%, 96%)[20]. However, serious concerns have recently been raised about the association between the exposure to ionizing radiation (such as is found in CT scanning) in childhood and lifetime mortality risk from cancer[21]. This has led practitioners in search of an alternate diagnostic tool to CT – a tool that will be accurate, reliable and valid. This search has led to renewed interest in clinical scoring systems.

In general, clinical scoring systems have been shown to increase the diagnostic accuracy in a time-efficient and cost-effective manner[22]. Several scoring systems exist specifically for appendicitis, however none have been routinely applied in pediatric clinical practice, largely because of a failure to achieve high diagnostic accuracy in repeated validation studies[23-29].
Until recently, most of the scores were retrospectively developed, were not specific to pediatrics and were not clinically optimal because of large zones where the diagnosis of appendicitis was uncertain. In 2002, Samuel developed the first appendicitis score that was specific to children[30]. It was prospectively developed on 1170 children between 4 and 15 years old with abdominal pain who were referred to surgeons to rule out a diagnosis of appendicitis. Samuel created the Pediatric Appendicitis Score consisting of 3 signs, 3 symptoms and 2 laboratory investigations for a total score of 10. When a cutoff of 5 was used in the derivation study, the score had a sensitivity of 100% and a specificity of 92%. This was exceptional performance for a diagnostic score and, as a result, has the potential to be a superior alternative to imaging with CT and ultrasound. Since abdominal pain is one of the most common presenting complaints to pediatric emergency departments[1-3], having a tool that could reliably differentiate appendicitis from the numerous other less acute causes would be extremely valuable from a clinical standpoint. The Pediatric Appendicitis Score, which was developed and internally validated by surgeons had never been externally validated and never validated by medical physicians at the time this project was initiated.

1.2 Thesis Objectives

The primary objective of this thesis is to validate the Pediatric Appendicitis Score in the hands of emergency physicians. Since emergency physicians, not surgeons, are the first to encounter patients with possible appendicitis in the hospital setting, it was felt that the score would be most useful in this context.

We conducted a prospective observational cohort study in the Montreal Children’s Hospital emergency department on all children between the ages of 4 and 18 years who presented with acute abdominal pain suspicious of appendicitis. The Montreal Children’s Hospital is an urban tertiary care hospital with an annual emergency department census of 65,000 patients per year. The Pediatric Appendicitis Score was calculated for each patient with and without appendicitis. This score was tabulated based on blinded data collected by medical physicians at the time of the
initial evaluation, before laboratory or imaging investigations were ordered. The gold standard for final diagnosis was histological examination of the pathological specimen for those undergoing surgery and telephone follow-up for those who did not undergo surgery. The specific objectives of our study were as follows:

1. To assess the overall performance of the Pediatric Appendicitis Score for predicting appendicitis using a receiver operator characteristic (ROC) curve
2. To examine the accuracy of the optimal cut-point defined by Samuel for predicting appendicitis in our population and clinical setting.
3. To determine the cut-point that maximizes the performance of the score for predicting appendicitis in our population. Sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) was calculated for each score cut-point to fulfill this objective.
4. To determine the potential impact of the Pediatric Appendicitis Score on patient outcomes (negative appendectomy rate and missed appendectomy rate) and reduction of imaging investigations by retrospectively applying the cut-points determined in objective three to our population.

This thesis consists of five sections. In Chapter 2 we will present a detailed literature review on appendicitis, concentrating on the challenges in diagnosis. As a part of this review a discussion of the accuracy of commonly used and novel diagnostic techniques will be provided. In Chapter 3, we will present the study methods and plan for statistical analysis, while Chapter 4 contains the results. In Chapter 5, we will present a discussion of our results.
Chapter 2

Literature Review

Arriving at an early, accurate diagnosis in childhood appendicitis has been a longstanding challenge for clinicians with important consequences for the patient when this does not occur. In this literature review we will present a brief description of the relevant anatomy, the pathophysiology of appendicitis and its common presenting signs and symptoms. We will then describe the importance of and the challenges in finding a highly sensitive and specific diagnostic test in pediatric appendicitis. Finally, we will present a review of the diagnostic testing options that have been evaluated in children. Laboratory investigations, imaging with ultrasound and computed tomography and clinical scoring systems will be discussed.

2.1 The Appendix

The appendix is a blind-ending worm-shaped tube, measuring about 5 - 10 centimeters that arises from the first part of the large bowel called the cecum. The cecum and the base of the appendix are located in the right lower quadrant of the abdomen. The base of the appendix is consistently located at McBurney’s point, one third the distance between the anterior superior iliac spine and the umbilicus. Although the tip of the appendix is classically located in close proximity to the base, it is frequently “hidden” behind the small bowel, large bowel or projected downwards into the pelvis. Histologically, the appendix contains lymphoid follicles (where white blood cells are formed), located in the deep connective tissue layer. The appendix does not serve any known physiologic function for the body[10, 31].

2.2 The Development of Appendicitis

Appendicitis, an inflammation of the appendix, begins with an obstruction of the appendiceal lumen. This obstruction can result from an enlarged lymphoid follicle (most often from
gastroenteritis or another viral illness), inspissated fecal matter or another foreign body. Obstruction of the lumen leads to overgrowth of the normal bacterial flora and continued mucus secretion. As a result of the obstruction, the intraluminal pressure increases, causing distention of the appendix. At this stage, simple acute appendicitis is present. If the distention of the appendix continues, obstruction to venous and lymphatic flow ensues, resulting in an increasingly swollen and necrotic appendix. When full-thickness necrosis of the appendiceal wall occurs, the bacteria inside the appendix move through the wall, into the peritoneal cavity. At this stage, the appendix is gangrenous and there is a localized infection in the abdominal cavity surrounding the appendix. Without intervention, the gangrenous appendix will perforate, spilling the appendiceal contents into the peritoneal cavity. The bacterial contents can cause a widespread infection of the abdominal cavity called peritonitis and a walled-off peri-appendiceal abscess may form. Peritonitis and peri-appendiceal abscesses can lead to death if the inflamed appendix is not removed and the patient is not treated with antibiotics[10, 32].

2.3 Symptoms of Appendicitis

Appendicitis typically begins with a generalized, vague abdominal pain that coincides with the obstruction of the appendiceal lumen. In the classic description, with bacterial proliferation and appendiceal distention, the pain slowly intensifies and becomes localized to the right lower quadrant over a period of 12 to 24 hours. The pain is exaggerated by any movement that causes irritation to the abdomen, such as, walking, coughing or hopping. Fever, nausea, vomiting and decreased appetite are typically present at this stage. If the appendicitis progresses to perforation, the patient may experience a brief reprieve from the pain as the pressure is relieved from inside the appendix. If there is still no intervention, when peritonitis develops, the fever heightens and the abdominal pain intensifies, becoming generalized to the left and right lower quadrants.
In the classic examination, the abdomen is soft and there is localized tenderness to palpation of the right lower quadrant of the abdomen, at McBurney’s point. Additional signs are pain felt in the right lower quadrant with palpation of the left lower quadrant (Rovsing's sign), an increase in pain from passive extension of the right hip joint that stretches the iliopsoas muscle (psoas sign), or pain produced by passive internal rotation of the flexed thigh (obturator sign). Rebound tenderness, the intensification of pain on release of palpation, may be present when there is a localized infection in the abdominal cavity surrounding the appendix. With perforation and peritonitis development, the patient has involuntary guarding to palpation, known as a rigid abdomen[31, 32].

2.4 The Diagnostic Challenge in Childhood Appendicitis

It has long been recognized that diagnosing appendicitis can present a significant challenge in the pediatric population[33]. Less than 50% of children with appendicitis have a classic presentation[34, 35] and many children present with signs and symptoms mimicking other common but self-limited causes of abdominal pain. Atypical locations of the appendix contribute, at least in part, to the atypical findings in early childhood appendicitis. Symptoms of gastroenteritis, constipation, urinary tract infection and ovarian problems are all typical of atypical presentations of appendicitis[36]. This tremendously varied presentation compounded by a child’s inability to clearly express their symptoms and the fact that the majority of children with a suspected diagnosis of appendicitis on initial clinical presentation do not have the disease, is at the root of the diagnostic challenge[37].

The problem of establishing an early, accurate diagnosis is evidenced by high misdiagnosis rates in the pediatric population. Although it has traditionally been taught that 70% - 80% of cases of appendicitis should be diagnosed after the initial clinical examination, based on the presence of classic signs and symptoms (abdominal pain, nausea, history of pain migration and signs of local peritonitis), this is not true in the pediatric population[38-40]. It is estimated between 28% and
57% of children with appendicitis do not have this classic constellation of attributes, and as such almost one-third of children with an ultimate diagnosis of appendicitis have been previously evaluated by a physician for their symptoms[41, 42].

2.5 The Consequences of a Delayed Diagnosis

All children with appendicitis will eventually receive a correct diagnosis because as their disease evolves, the child appears more unwell and the symptoms of appendicitis become more obvious and specific. At this late stage, the majority of children will have a perforation of their appendix.

The consequences of prolonged observation as a diagnostic approach are significant. Pledger and Stringer[43] reported twelve deaths in children from acute appendicitis between 1993-1997 in England and Wales and cited the main factor contributing to death to be a delay in diagnosis. In 2000, Graff and his colleagues[7] showed that the occurrence of perforated appendicitis was directly correlated with a delay in diagnosis. They conducted a retrospective review of 2163 children presenting to 12 Connecticut area hospitals in 1996. One thousand forty-five patients had appendicitis and 1118 had acute abdominal pain of varying etiologies. They showed that the proportion of patients with perforation and abscess formation was higher if a diagnosis of appendicitis was missed at the initial evaluation [31.6% (95%CI: 22.2%, 40.4%)] and [13.7% (95%CI: 7.3%, 21.0%)] respectively compared to the group where a correct diagnosis was made at the time of the initial evaluation [17.0% (95%CI: 14.5%, 19.9%) and 4.8% (95%CI: 3.5%, 6.6%) respectively]. These findings directly correlate with the duration of a patient’s symptoms – the longer a patient is symptomatic, the more likely they are to develop an abscess or perforation.

The morbidity from acute appendicitis is almost entirely related to perforation and its complications[44-46]. Fifty percent of children with perforated appendicitis experience a complication such as peri-appendiceal abscess formation, peritonitis, sepsis, wound infection and wound dehiscence[47]. These problems lead to increased pain and suffering for the patient and
result in longer more complicated and costly hospital stays. Additionally, these patients have a predisposition for long-term complications such as bowel obstruction from adhesions and impaired fertility in female patients[35, 48]. In combination, these factors have an important impact on the child and family.

For the physician, a delayed diagnosis of Appendicitis also has an important impact. In the US, it has been shown that a missed diagnosis of appendicitis is the most frequently successful malpractice claim against Emergency Room Physicians[49].

2.6 The Therapeutic Dilemma

The persistent challenge is to diagnose appendicitis early enough to prevent the progression to perforation while minimizing the number of negative appendectomies (appendectomies that are performed on children who are found not to have the disease on gross and histologic review). In order to detect appendicitis early enough, the physician must have a low “threshold” for establishing the diagnosis, at the expense of performing unnecessary appendectomies on children who are ultimately found to have another etiology for their abdominal pain, but who, on presentation, had symptoms mimicking appendicitis. When global physician assessment was the only means to diagnose appendicitis, it was accepted that negative appendectomy rates between 15% and 25% were an indication of good practice[50]. When the diagnosis was in doubt, performing a negative appendectomy was viewed as preferable to delaying the diagnosis and risk experiencing the complications of perforation[51, 52].

In an effort to minimize the occurrence of these two adverse events, many diagnostic testing options have been investigated. Early studies examined the utility of laboratory investigations in diagnosing appendicitis, while more recent efforts have focused on imaging investigations. Finally, clinical scoring systems have received increased attention over the past two decades.
**2.7 Laboratory Investigations in the Diagnosis of Appendicitis**

Laboratory tests have been extensively investigated as a potential diagnostic aid in acute appendicitis because they are fast, economical and universally available. The white blood cell count (WBC) and C-reactive protein (CRP), a marker of inflammation, have received the most attention in the literature, but other markers of inflammation such as the erythrocyte sedimentation rate (ESR), tumor necrosis factor alpha (TNF-α), α₁-glycoprotein, endotoxin, interleukin-6 (IL-6) and interleukin-8 (IL-8) have also been investigated. Single tests as well as repeated measures over time have been investigated. Unfortunately, no single test has been shown to be a reliable early predictor of early acute appendicitis.

**2.7.1 White Blood Cell Count (WBC)**

Early studies by Doraiswamy (1979)[53] and Peltola et al. (1986)[54] investigated the accuracy of the WBC as a diagnostic test for appendicitis. They showed an association between an elevated WBC and appendicitis, though the estimates of the strength of this association varied greatly. An elevated WBC (defined as WBC > 10,000 to 15,000 cells/mL) was shown to be 42% to 60.5% sensitive and 84% to 100% specific for appendicitis. These studies were retrospective and included only patients undergoing an appendectomy, therefore, the specificity must be interpreted with caution. This methodologic flaw is a form of selection bias, called verification bias. In general terms, this occurs when a study only investigates patients who have a higher or lower suspicion of disease to undergo testing with the gold standard. To avoid this, a study should include all consecutive patients at risk for a disease and not only a subset who underwent definitive testing.

More recent studies by Wang et al.[55] and Cardall et al.[9] used improved designs which reduced the selection bias found in the previous studies by evaluating patients presenting to Emergency Departments with abdominal pain in whom the physician suspected appendicitis on the basis of the clinical examination. Wang and her colleagues evaluated 410 children, aged 1 to
19 years, who presented to a pediatric emergency department with acute non-traumatic abdominal pain and who had symptoms of appendicitis on clinical examination. They found that of the 74 cases of confirmed appendicitis, 32% of these children had a normal WBC. An elevated WBC (defined as WBC > 17,000 between ages 1-3.9 years; WBC > 14,000 ages 4 – 11.9 years; and WBC > 13,000 ages 12-19 years) was 67.5% sensitive and 79% specific in detecting acute appendicitis. Cardall and his colleagues found similar results. They evaluated a mixed pediatric and adult population in whom the treating physician suspected appendicitis. A standardized data collection form was completed on all 293 enrolled patients and appendicitis was confirmed in 92 patients by histologic examination. They found that 24% of all patients with appendicitis had a normal WBC and reported that an elevated WBC (defined as WBC > 10,000 cells/mL) was 76% sensitive and 52% specific.

In 2005 Kharbanda and his colleagues developed a clinical decision rule to identify children who are at low risk for appendicitis[56]. They collected 24 demographic, historical, physical examination and laboratory variables on 601 children who underwent a surgical consultation for possible appendicitis in a pediatric emergency department. By univariate recursive partitioning, they determined that a WBC of 8,850 cells/mL was the ideal cut-point for the prediction of appendicitis in their sample. Intuitively, this WBC value, which is in the normal range, should not have a high discriminative value between those patients with and without appendicitis. This was confirmed by the study findings, as the WBC was not retained as one of the six predictors in the final decision rule.

In summary, these studies show that the WBC on its own does not accurately discriminate between patients with and without appendicitis and it performs no better, and possibly worse than clinical examination alone.
2.7.2 C-Reactive Protein (CRP)

CRP is an acute phase protein that is synthesized in the liver in response to tissue damage or inflammation. It can rise within 2 to 6 hours after the onset of inflammation and, because of its relatively short half-life of 19 hours, it falls quickly after the inflammatory stimulus has been removed[57]. A normal CRP for a child is < 6 mg/L. These desirable properties have led investigators to study the usefulness of CRP as a diagnostic test for appendicitis. The results have not been consistent.

A study by Juan Carlos Rodriguez-Sanjuan and his colleagues in 1999 investigated the diagnostic accuracy of CRP in childhood appendicitis[58]. They performed a retrospective review of 124 children who were operated on with a clinical diagnosis of appendicitis (history, physical examination, WBC and CRP). The removed appendix underwent histological review to confirm or refute the pre-operative diagnosis and the patients were analyzed on the basis of these findings – those with appendicitis and those without. They determined that a CRP of 17mg/L had the highest positive predictive value, however, the sensitivity and specificity at this cut-point were low, with values of 58% and 80%, respectively.

Using identical methodology, in 2001, JM Gronroos studied 100 patients with histology confirmed appendicitis and 100 patients who underwent a negative appendectomy[59]. He found that the CRP could not adequately separate those patients with and without appendicitis with mean CRP values of 30 ± 4 mg/L in those with appendicitis and 31 ± 4 mg/L in those without appendicitis. In this study the CRP had a sensitivity of 48% and a specificity of 57%.

Thus, the CRP performs poorly as a diagnostic test for appendicitis in these studies where only a select group of patients were included and its discriminatory value is likely to be even worse in a clinical setting where all patients with acute abdominal pain, suspicious of appendicitis, are tested. Other non-appendicidal causes of acute abdominal pain may be inflammatory in etiology, making it difficult to separate those with and without appendicitis on the basis of the CRP. Further, these studies are subject to the same selection bias as the early WBC studies by Doraiswamy and Peltola and may overestimate the discriminative value of the CRP in appendicitis as they only included patients who underwent appendectomy. This limited population makes the study results difficult to generalize to the clinical setting.

Although not useful in diagnosing early appendicitis, in 1996 Chung and his colleagues showed that the CRP is valuable in the diagnosis of advanced appendicitis[61]. They measured CRP levels in 56 patients with perforated appendicitis and 22 patients with simple appendicitis. They found the CRP was significantly higher in patients with perforated appendicitis and when a cut-point of 50mg/L was used, the CRP had a sensitivity of 76% with a specificity of 82% in diagnosing perforated appendicitis.

These findings were supported by Okomoto et al.[63] and Sack et al.[64] in 2006. Okomoto and his colleagues conducted a retrospective review of 289 patients with confirmed appendicitis and separated the patients into 2 groups – those with simple appendicitis (222 patients) and those with advanced disease (67 patients). Advanced disease was defined as appendiceal perforation, abscess formation or the presence of purulent peritoneal fluid on operative report. Mean CRP values were significantly different in those with simple appendicitis (19 ± 23 mg/L) and those with advanced disease (103 ± 68 mg/L). A Receiver Operator Characteristic (ROC) curve created to assess the discriminatory value of the CRP in separating these two groups had an area under the curve of 0.896 with an optimal CRP cut-point of 39 mg/L. At this point, the CRP had a sensitivity of 82% (95% CI: 77.5%, 86.5%) and a specificity of 83.3% (95% CI: 79%, 87.6%). Sack and his colleagues measured several blood inflammatory markers, including CRP, on 211 children
suspected of having appendicitis. The patients were classified into 4 groups a) non-surgical abdominal pain (22 patients) b) early/absent appendicitis (81 patients) c) phlegmonous appendicitis (78 patients) and d) perforated appendicitis (30 patients). The authors found that CRP was unable to distinguish between children with non-surgical abdominal pain and those with early appendicitis, but was significantly different in children with early appendicitis and those with phlegmonous or perforated appendicitis.

Although CRP appears to discriminate between early and advanced appendicitis, it does not provide any assistance with distinguishing between patients with appendicitis and those with abdominal pain of other etiologies.

2.8 Imaging Investigations in the Diagnosis of Appendicitis

2.8.1 Ultrasound

Ultrasonography, first reported as a diagnostic aid for appendicitis in 1986[65], is an ideal non-invasive means to image the abdominal cavity. It is relatively inexpensive, poses no ionizing radiation risk to the patient and can be performed rapidly with little or no patient preparation. Graded compression ultrasonography, a technique first described by Puyalaert in 1986, involves examination of the intra-abdominal contents after slow compression of the anterior abdominal wall[65]. This technique, when performed by an experienced ultrasonographer, minimizes discomfort, maximizes relaxation of the abdominal muscles and should improve visualization of the appendix compared to standard ultrasonography by displacing gas-filled bowel loops. The appendix is identified as a tubular structure that shows no peristaltic activity. The most widely used criteria to diagnose appendicitis on ultrasound are a) a non-compressible appendix with a maximal cross-sectional diameter of greater than 6mm b) presence of an appendicolith, defined as a calcified deposit within the appendix which can result in obstruction of the lumen or c) evidence of a complex fluid collection in the periappendiceal area suggestive of appendiceal
perforation[35]. The criteria used to indicate the absence of appendicitis, which will be referred to as a “negative study” varies from author to author.

Several authors have reported a high diagnostic accuracy for ultrasound in childhood appendicitis. Prospective studies by Vignault (1990)[15], Rubin (1990)[66], Hahn (1998)[67] and Schulte (1998)[68] evaluated the accuracy of ultrasound in all children with abdominal pain suspicious of appendicitis. These authors reported that the sensitivity of ultrasound in diagnosing appendicitis in their populations ranged from 89% to 94% and the specificity ranged from 89% to 98%. All studies, with the exception of Schulte’s, used the definition described above for a positive ultrasound. Schulte used a more liberal definition of appendicitis, describing the characteristics of an inflamed appendix without requiring the diameter to exceed a certain threshold. This does not appear to have had a significant impact on their results as their false positive rate of 9.8% is comparable to that of other studies that showed false positive rates ranging between 11.1% and 11.4%. Of note, these rates, which would result in a negative appendectomy, are lower than rates reported based purely on clinical examination.

The principal flaw in these studies is that they considered a scan where the appendix was not visualized to be negative. Vignault and his colleagues reported that the appendix was not seen in 71.5% of ultrasounds that were considered to be negative[15]. Hahn and his colleagues have documented the risk of employing this rationale[67]. In their series, a non-visualized appendix accounted for 98% of the false-negative ultrasounds and accounted for a missed appendicitis rate of 10% of their population. Using “appendix not visualized” as sufficient criteria for a negative scan likely overstates the usefulness of ultrasound as a diagnostic test. If these “appendix not visualized” ultrasounds had been classified as “equivocal”, these patients would require further testing to confirm or refute the diagnosis.

In spite of this flaw, the false negative rates (6% - 11%) in these studies were low. This is most likely due to the fact that all of the ultrasounds were performed or supervised by staff radiologists.
who were experienced in ultrasonography. It is known that the likelihood of visualizing the appendix on ultrasound increases with operator experience as it takes practice to master the technique of graded compression ultrasonography and to be able to distinguish the appendix from the surrounding structures[18]. Because of this, the sensitivities and specificities reported in these studies are likely overstated and could not necessarily be generalized to a teaching hospital, for example, where residents perform unsupervised ultrasounds during the evenings and on weekends or to a hospital where pediatric ultrasound is performed infrequently.

Several other studies evaluating the use of ultrasound in children who had an equivocal diagnosis of appendicitis based on clinical evaluation have been performed. These studies excluded those children who either were discharged home without further testing because they had a very low suspicion of appendicitis, or who were taken directly to surgery without further testing because they had a very high suspicion of appendicitis. Three prospective studies[69-71] and three retrospective studies[16, 18, 72] have been reported. Again, all of the ultrasounds were performed by experienced pediatric radiologists and if the appendix was not visualized, the scan was considered as negative. This latter point may have contributed to the lower sensitivities reported in these studies (range 85% to 91%) when compared to the studies that evaluated “all comers” with suspected appendicitis. The higher number of false negative studies in this population of children with atypical or unusual findings highlights the importance of visualizing the entire appendix in order to make an accurate diagnosis on ultrasound. The specificities reported are similar to those described in the previous set of studies with a range of 88% to 98%. Two prospective studies by Rice et al.[69] and Lessin et al.[70] evaluated the accuracy of clinical impression in a group of children with atypical findings and found that the sensitivity of a clinical diagnosis was poor (38% and 50%) while the specificity was better but still inferior to ultrasound examination (85% and 95%). Sivit and his colleagues found that ultrasound was more sensitive in children with low or intermediate risk of appendicitis based on history and physical examination compared to those with a high clinical suspicion[14, 71].
Despite the high diagnostic accuracy of ultrasound in childhood appendicitis established by these studies, authors have been unable to demonstrate improved patient outcomes when ultrasound is used to diagnose appendicitis compared to traditional evaluation using history, physical examination and laboratory investigations. A randomized controlled trial by Douglas and his colleagues compared the outcomes of 100 children with possible appendicitis whose treatment decisions were based on clinical assessment (control group, 34 patients) or ultrasonography and clinical assessment (intervention group, 66 patients)[73]. Only patients with equivocal diagnoses underwent mandatory ultrasound in the intervention group. The study contained a mix of adult and pediatric patients, however, the results discussed here pertain only to the 100 patients in the pediatric subgroup. The authors observed a lower negative appendectomy rate in the intervention group (4.6%) compared to the control group (8.8%). Although this represents a an absolute difference of 4.4% in negative appendectomy rates between groups the 95% confidence interval (-5.5%, 18.7%) is very wide and contains zero, indicating no effect. They also observed a lower absolute difference in mean time to therapeutic operation in the intervention group (7.0 hours) compared with the control group (13.1 hours) but, again, the 95% confidence interval for the difference in means (-1.0, 13.2) is wide and includes zero, indicating no effect. The proportion of patients with perforated appendicitis and the mean duration of stay in hospital was similar between the intervention and control groups. The study was powered to detect a 9% difference in negative appendectomy rates, 15.2 hour difference in mean duration of stay and a 3.3 hour difference in time to operation, which may be the reason for the negative results observed for the first two outcomes. The authors concluded that the diagnosis of appendicitis by ultrasound did not produce better patient outcomes than clinical diagnosis alone. We are unable to comment on the similarity between the 2 groups at baseline and whether this may have influenced the results as the authors do not provide this information for the pediatric subgroup.

Roosevelt and Reynolds (1998)[74] and Emil et al. (2001)[75] performed retrospective studies that supported these conclusions. Roosevelt and Reynolds studied 231 children with pathologically confirmed appendicitis and analyzed the patients in 2 groups: those who
underwent ultrasound examination prior to surgery (100 patients) and those who did not (131 patients). The groups were similar at presentation for mean temperature, WBC count, percentage with vomiting, diarrhea, abdominal tenderness or guarding, however those who did not undergo ultrasound examination were more often male (71%) and more frequently had RLQ pain (65%). They found the rate of perforation was similar between the two groups, and that ultrasound delayed the time to surgery. They did not comment on the negative appendectomy rate. Using similar methodology, Emil and his colleagues retrospectively reviewed 454 children who underwent ultrasonography (191 patients) or clinical evaluation alone (263 patients). The authors conclude that patients in the ultrasonography group had worse outcomes with higher negative appendectomy rates [13.1% vs. 5.1% (95%CI: 2.9%, 14.0%)] and more post-operative abscesses [4.4% vs. 1.2% (95%CI: 0.07%, 7.0%)]. Strong conclusions cannot be drawn from the 95% confidence interval for the difference in post-operative abscess formation between groups since the interval is wide and includes values very close to zero that are not highly clinically relevant.

The results of these 3 studies are limited by the fact that they compared children who had an equivocal diagnosis of appendicitis to those who were taken directly to the operating room. This latter group likely represents children with a classic or straightforward presentation of appendicitis. In these patients it is much easier to make a definitive diagnosis of appendicitis, making it an unfair comparison group for those with an uncertain or atypical presentation. Even in the randomized study by Douglas and his colleagues, only children with an equivocal diagnosis of appendicitis in the intervention group underwent ultrasound while children with varying degrees of diagnostic certainty were evaluated by clinical examination alone. An improved methodology for these studies would have been to compare patients with similar baseline characteristics. For example, if patients with an equivocal diagnosis of appendicitis were randomized to evaluation with clinical examination or evaluation with ultrasound and clinical examination, more accurate conclusions about the utility of ultrasound in a clinical setting could be reached. Therefore,
although these studies conclude that ultrasound does not improve patient outcomes, their results must be interpreted with care.

Ultrasound as a diagnostic tool for appendicitis has three chief limitations: the need for an experienced operator, the difficulty in visualizing a non-inflamed appendix and the requirement for the absence of certain patient characteristics. Hahn and his colleagues identified inappropriate technique by the ultrasonographer and incorrect interpretation of the images as the most common reasons for obtaining 50 (10%) false negative results, in their population of children[67]. In addition, patient characteristics such as superimposed air or feces, patient obesity, excessive pain not allowing adequate abdominal compression and a non-cooperative child also contribute to erroneous diagnoses. Hormann and his colleagues found that if a child was overweight (classified by body mass index (BMI) for age percentiles), then the appendix was much less likely to be visualized than if a child was under weight or normal weight[19]. They suggested that ultrasound may delay definitive treatment in this patient population and should not be the imaging investigation of choice as the likelihood of visualizing the appendix, even with an experienced operator, is low. The most common reason for a false positive diagnosis on ultrasound identified by Hahn and his colleagues was the presence of lymphoid tissue hyperplasia. This could just represent early appendicitis because it is certainly in the early pathophysiology of acute appendicitis.

2.8.2 Computed Tomography

Although ultrasound has been shown to have a high diagnostic accuracy with experienced operators, the need for this expertise, the difficulty in visualizing a non-inflamed appendix and the requirement for the absence of certain patient characteristics in order to produce consistent, high quality results has led authors to investigate computed tomography (CT) as an alternative imaging modality to ultrasound in the diagnosis of childhood appendicitis. Contrast-enhanced scans (administration of intravenous, oral or rectal contrast material to the patient prior to CT
scan) are most commonly used in the diagnosis of appendicitis, as they improve the visualization of the appendix[76, 77]. Definitive diagnosis of appendicitis can be made when an inflamed, non-opacified (with contrast) appendix is seen with a maximal cross-sectional diameter greater than 6mm. Other secondary signs such as the identification of a) an appendicolith b) focal thickening of the apex of the cecum c) an arrowhead-shaped collection of contrast material in the cecum that points to an occluded appendiceal lumen d) stranding of the periappendiceal fat and e) a periappendiceal mass suggestive of phlegmon, abscess or thickening of adjacent atonic bowel loops aid in a diagnosis when definitive signs are not present[78-80]. All of the studies discussed in this section use these diagnostic criteria, unless otherwise specified.

CT scans are operator independent diagnostic tools that are performed using standardized procedures by trained technicians. The interpretation of their results is not limited to experienced staff radiologists. In a study by Albano and her colleagues, 2nd and 3rd year radiology residents had 96% agreement with staff radiologists in the interpretation of 103 CT scans performed for suspected appendicitis on a mixed adult and pediatric population[81]. In an exclusively pediatric population, Lowe and her colleagues showed that although residents were less confident about their interpretation of 75 unenhanced scans compared to the staff radiologists (89% vs 93% 95%CI:-4.2%, 12.2%), they agreed on 91% of the scan results with a kappa coefficient of 0.73 ± 0.095[82]. The residents incorrectly interpreted 9 scans – 7 false negatives and 2 false positives. The good to substantial agreement shown in this study may have been improved by the administration of contrast material to the patients prior to CT scanning. By enhancing the visualization of the appendix, the number of false negative interpretations by the residents should theoretically decrease as the appendix is more easily identified on contrast-enhanced scans.

CT is of additional value as a diagnostic tool as radiologists have been shown to have more confidence of in the interpretation of the test when compared to ultrasound. Pena and Taylor conducted a prospective study on 139 children and young adults (< 21 years old) who had ultrasound and/or CT scans for suspected appendicitis[83]. CT scans were only performed on
children with persistent symptoms who had negative or equivocal findings on ultrasound (108 patients). Radiologists were asked to record their level of confidence (very low, low, medium, high, very high) in their interpretation of the ultrasounds and CT scans immediately after each examination. The authors showed that radiologists had “high” or “very high” confidence in the interpretations of the scans in 91.6% of CT scans, but in only 57.5% of ultrasounds. When a CT scan was interpreted as negative, 95% of radiologists had “high” or “very high” confidence in their results, whereas only 54% of radiologists had this level of confidence when they interpreted an ultrasound as negative. In this study, CT had a superior sensitivity to ultrasound (96% vs. 44%) and a comparable specificity (96% vs. 93%). These findings are important. Confidence in the interpretation of CT results allows physicians to make definitive management decisions for patients. This is not true for ultrasound where a negative scan has a considerable false negative rate. As such, the confidence in negative scans is low and often requires prolonged observation or further diagnostic testing before management decisions can be made.

CT has been shown to be a highly accurate diagnostic tool in pediatric appendicitis. Doria and her colleagues performed a meta analysis of all pediatric and adult studies published between January 1986 and December 2004 assessing CT and/or ultrasound as a diagnostic test for appendicitis[20]. The pediatric and adult studies were analyzed separately. This discussion will only contain information about the pediatric studies using CT (alone or in addition to ultrasound). Eight pediatric studies met the inclusion criteria (prospective or retrospective study design, information available about patient follow-up or surgical pathology results, data available to compute accuracy and information about the criteria used to define a positive study). Three studies (Pena and Taylor 2000, Kaiser et al. 2002 and Lowe et al. 2001)[83-85] were prospective and five (Kaiser et al. 2004, Sivit et al. 2000, Fefferman et al. 2001, Karakas et al. 2000 and Mullins et al. 2001)[86-90] were retrospective. Kasier et al. (2002) and Pena and Taylor (2000) investigated the use of CT only when ultrasound was negative or inconclusive. Lowe et al. (2001), Fefferman et al. (2001), Mullins et al. (2001) and Kaiser et al. (2004) investigated the accuracy of CT as the sole imaging modality while Sivit et al. (2000) and Karakas et al. (2000)
investigated the use of CT in comparison to US. The quality of the studies was assessed by 3 reviewers (2 blinded and 1 un-blinded), using a standardized checklist. The median quality score of all pediatric studies (including those investigating the use of US alone) was modest at 34.4% (out of a maximum possible score of 100%) with an intraclass correlation coefficient of 0.7 (95% CI: 0.46, 0.84), indicating good agreement between reviewers. The principal flaw of the 8 pediatric CT studies was that the majority were retrospective and as such, did not provide information about patient follow-up when the scans were interpreted as negative. Although this could lead to an overstatement of the sensitivity, as discussed in the previous sections, confidence in a negative CT scan should be high, and therefore, should not have an important impact on the sensitivity. In the meta-analysis, the pooled sensitivity [94% (95% CI: 92%, 97%)] and specificity [95% (95% CI: 94%, 97%)] for CT scan are very high, indicating excellent performance as a diagnostic tool. These numbers suggest that the principal benefit offered by CT over ultrasound is the important decrease in the false negative rate and consequent improvement in the sensitivity. However, given that the majority of the studies were retrospective, and some were subject to indication bias[83], the sensitivity of CT must be interpreted with care. Using similar search criteria as the meta-analysis, no further pediatric studies investigating the accuracy of CT as a diagnostic tool in appendicitis were found since the publication of the meta-analysis.

It is widely accepted that CT is an accurate diagnostic tool, however, specific CT techniques such as the use of contrast and its optimal delivery method remain controversial. Contrast-enhanced scans are theoretically more sensitive than unenhanced scans because of the improved visualization of the appendix, however there are several disadvantages to administering contrast to patients, such as a time delay between the patient receiving the contrast and performing the CT, the potential for adverse reactions and poor tolerance of oral or rectal contrast in ill patients. Two investigators have studied the usefulness of unenhanced scans as a diagnostic tool for pediatric appendicitis. Hoecker and her colleagues conducted a retrospective review of all patients undergoing CT scans for suspected appendicitis during a 6 month period in 2000[91].
According to institutional practice, all scans performed on pediatric patients during this study period were unenhanced. One hundred and eleven patients were identified and their CT, pathology results and follow-up information was abstracted by a blinded reviewer. They found that unenhanced CT scans had a sensitivity of 87.5% (95% CI: 75.8%, 94.8%) and specificity of 93.7% (95% CI: 85.4%, 98.0%). The sensitivity of these unenhanced scans is good but inferior to that reported for enhanced scans in Doria’s meta analysis. The authors do not have complete follow-up information for the patients who were discharged with negative CT scan results. If patients did not re-present to the same hospital, they were assumed not to have appendicitis. This methodologic flaw may underestimate the false negative rate and overestimate the sensitivity reported in this study. Kaiser and her colleagues also found that the sensitivity of unenhanced scans (66%) was inferior to enhanced scans (90%) in a retrospective review of all patients undergoing CT scanning for a clinical suspicion of appendicitis[86]. Each patient first received a limited unenhanced scan (images of the lower abdomen only) followed by an IV contrast enhanced scan of the entire abdomen. Again, in this study, patients who were discharged with negative CT scans and did not re-present to the same hospital with a recurrence of symptoms were assumed not to have appendicitis. However, since there are 2 groups, we can assume that the risk of underestimating the false negative rate in both groups would be similar, thus allowing a comparison of the reported sensitivities and conclusion that enhanced scans are more sensitive than unenhanced scans.

Contrast for CT scans may be administered orally, rectally or intravenously. Oral and rectal contrast create bowel opacification which leads to easier identification of the appendix. Rectal contrast is favoured over oral contrast by many authors[90, 92], because it offers more consistent cecal opacification, shorter delay between the administration of the contrast and presumed opacification and, although it’s delivery via rectal tube is uncomfortable, it is better tolerated in patients with nausea and vomiting[77]. There are no pediatric studies comparing oral and rectal contrast in children undergoing CT scans for suspected appendicitis and no clear consensus prevails in the literature. Intravenous contrast enhances inflammatory lesions, which helps
differentiate between fluid collections and bowel wall. Intravenous contrast is generally given as an adjunct to oral or rectal contrast, but there is a single study by Kharbanda and his colleagues in 2007 which retrospectively compared the diagnostic performance of CT with rectal + IV contrast to CT with IV contrast alone for children with suspected appendicitis[93]. The study population contained 2 sequential cohorts – those presenting between April 2003 and February 2004 received CT with rectal + IV contrast and those presenting between February and December 2004 received CT with IV contrast only, according to a change in institutional protocol. CT reports and surgical pathology results were abstracted from the medical record. In patients who did not undergo surgery, information from a follow-up telephone call made 2-4 weeks after the emergency department visit was available from a previous prospective study by the same authors. They found that the sensitivity [92% (95% CI: 85%, 97%)] and specificity [87% (95% CI: 79%, 92%)] of rectal+IV enhanced scans were comparable to those of IV contrast only scans [sensitivity 93% (95%CI: 84%, 97%) and specificity 92% (95%CI: 85%, 96%)]. There is a paucity of studies comparing the efficacy of different contrast delivery methods and no conclusions can be drawn from the existing literature.

Despite the high diagnostic accuracy of CT in childhood appendicitis, the results from studies of patient outcomes when CT is used to diagnose appendicitis have not been strongly positive. In 2004 Kaiser and her colleagues conducted a retrospective review of the first 150 appendectomies performed on children in 1991, 1994, 1997 and 2000[94]. They found that the negative appendectomy rate decreased over the years (23.0% in 1991, 8.7% in 1994, 8.0% in 1997 and 4.0% in 2000) with no change in the perforation rates (32% in 1991, 34% in 1994, 34% in 1997 and 29% in 2000). Although this decrease in the negative appendectomy rate (18% reduction; 95% CI: 11.2%, 26.0%) is meaningful, the largest decrease occurred from 1991 to 1994 (14.3% difference; 95%CI: 5.8%, 22.1%), which corresponds to the increased use of ultrasound (1.3% in 1991 vs. 41% in 1994). The negative appendectomy rate did not change from 1994 to 2000 with the introduction of CT (4.7% reduction; 95%CI: -1.0%, 10.7%). During these years, CT utilization increased from 0% in 1994 to 60% in 2000. Interestingly the authors showed that the perforation
rate did not change over these years, even with the increased use of imaging (overall reduction 3.0% 95%CI: -7.6%, 14.6%). Rao et al. (1999) and Partrick et al. (2003) showed similar results[95, 96]. Rao and his colleagues performed a review of patients undergoing appendectomy prior to the introduction of CT (July 1992 to September 1995, 129 patients) and after the implementation of CT (January to December 1997, 59 patients). CT was introduced, refined and investigated at the study hospital during 1996. The study population was mixed adult and pediatric, but the results discussed here pertain only to the pediatric patients (188 patients). Before the introduction of CT, the negative appendectomy rate was 13% (17 of 129 children, 95% CI: 8.4%, 20.0%) compared to 7% (4 of 59 patients, 95% CI: 2.7%, 16.2%) after the introduction of CT. Similarly, the perforation rates decreased from 23% (95% CI: 16.4%, 31.8%) before the introduction of CT to 15% (95% CI: 7.5%, 26.2%) after. Although there has been an 8% absolute decrease in perforation rate, the 95% confidence intervals for the group difference is (-3.7%, 19.7%) which includes zero and precludes a definitive conclusion. In 2003, Partrick and his colleagues conducted a retrospective review of 616 children who underwent an appendectomy between 1997 and 2001. In their population, there was a natural decrease in US use (40% in 1997 to 7% in 2001) and an increase in CT use (1% in 1997 to 58% in 2001) with an increase in the total number of imaging investigations performed over the study period (40% in 1997 to over 60% in 2001). The authors showed that the negative appendectomy rate fell from 12% in 1997 to 7% in 2001, however, the greatest decline occurred from 1997 (12%) to 1998 (8%). In 1998, the rate of CT utilization was < 10%, so the decrease in negative appendectomies cannot solely be attributed to increased CT use. The perforation rate also decreased during the study period from 39% in 1997 to 33% in 2001. We are unable to calculate confidence intervals around these proportions as the actual number of patients per year of study is not provided. Martin et al. (2004), Bendeck et al. (2002) and McGory et al. (2005) found similar results[97-99]. With such high sensitivities reported for CT scans, it is interesting to note that the negative appendectomy rates have not changed substantially since the more widespread use of CT. This may be due to the longitudinal design of these studies. The differences in outcomes reported over the years may be due to other factors that may or may not be related to the introduction of imaging but are
separate from the results of the test. For example, systematic increases in surgical consultations, increased imaging of all children with a suspicion of appendicitis, instead of only those with equivocal presentations and later presentation of patients with appendicitis to hospital due to changes in insurance systems could influence the results. In any case, the concern expressed by some authors that imaging patients prior to appendectomy delays the operation and results in increased perforation rates is not supported by these results.

The chief limitation to using CT as a diagnostic tool for childhood appendicitis is the exposure to ionizing radiation. It is now known that a single abdominal CT increases a child’s lifetime risk of a malignancy and this risk is further amplified the younger a child is at the time of the examination. Brenner and his colleagues produced a landmark study on the assessment of the lifetime cancer mortality risks attributable to radiation from pediatric CT[21]. They showed that the estimated lifetime risk of a malignancy for children undergoing CT scan when \( \leq 5 \) years old was 0.15 – 0.23% and for children undergoing CT scan between 5 and 15 years old was 0.11 – 0.15%. Compared to the cancer risk for adults, the risk for children is equivalent to an additional case of cancer per 1,100 abdominal CT scans performed. Although this represents a relative risk of 1.35% above the natural background incidence of cancer, the absolute numbers are more meaningful with the increased number of pediatric CTs being ordered[100]. Pediatric CT scans account for only 4% of the total number of scans done in the US each year, however, they are estimated to contribute to approximately 20% of the total potential cancer mortality. These findings with grave implications have been widely disseminated within pediatric emergency medicine and pediatric radiology resulting in more judicious use of CT scanning in pediatric patients. A long standing radiation-safety principle called ALARA (“as low as reasonably achievable”) has now reached the forefront and pediatric radiologists are using as low a dose of radiation that will yield acceptable images. Fefferman and her colleagues performed a computerized simulation of 100 abdominal scans using a standard dose of radiation to produce an image quality equivalent of what would result from low dose (20mA) radiation[100]. They found that the radiologist’s interpretation of these low dose scans was inferior compared to the
interpretation of standard images (sensitivity 77% vs. 91.5%). The dose of radiation was the only risk factor for the increase in false negative results. Although using the simulated images using a lower dose of radiation did not produce clinically acceptable results, the authors conclude that these types of computer simulations are important for future work as there may be a radiation dose between 20mA and the standard dose that will produce acceptable images and result in comparable accuracy data.

2.9 Clinical Scoring Systems

Although imaging as a diagnostic aid for appendicitis has received the most attention over the past decade, there has been renewed interest in clinical scoring systems because of the variable performance of ultrasound and the important concern of malignancy risk with CT scans in children. Clinical scoring systems, a type of clinical decision rule, quantify the individual contributions that various components of the history, physical examination and simple tests make towards diagnosis, prognosis or likely response to treatment. They aim to help clinicians cope with the uncertainty of medical decision-making and help with efficiency by standardizing the collection and interpretation of clinical data. They are inexpensive, time-efficient diagnostic tools that have the potential to improve patient outcomes.

2.9.1 Methodologic Standards for the Development and Validation of Clinical Decision Rules

Methodologic standards for the development of clinical decision rules were originally described by Wasson et al.[101]. More recently, Steill and his colleagues assessed clinical decision rule articles in four major medical journals and proposed modifications to Wasson et al’s original methodologic standards[102]. They have identified eight criteria upon which to judge the adequacy of a clinical prediction rule.
First, the outcome assessed by the decision rule should be clinically important and clearly defined. Clinically important outcomes should be objective and reproducible in other settings. For example, confirmation of appendicitis by the presence of certain criteria on histologic examination of the pathological specimen is an objective measure that can be assessed in multiple settings, whereas a behavioural outcome such as admission to hospital may be dependent of local factors and difficult to replicate. To avoid observation bias, the outcome measure should be assessed without knowledge of the predictor variables.

Second, all potential predictor variables of the outcome should be clearly defined and collected in a prospective fashion. Investigators should ensure that the physicians in their study have been adequately trained to evaluate the patients and that the collection of data is standardized. Clinical data is most reliable when collected prospectively and recorded on a data collection form designed specifically for the decision rule study. Data collected from a review of the medical record lacks precision and there are often large amounts of missing information. This method of data collection is generally unacceptable other than for the assessment of feasibility.

Third, the assessment of the predictor variables must be reliable. This refers to the consistency or reproducibility of the findings by the same clinician or by different clinicians. Decision rules are highly dependent on the findings from the clinical examination, therefore, the findings must be reliable in order for the rule to be dependable. Only predictor variables with good agreement (beyond that expected by chance alone) should be considered for inclusion in a decision rule.

Fourth, the study subjects must be well described in terms of inclusion criteria, method of selection, clinical and demographic characteristics and the study setting. Explicit inclusion criteria allows the reader to understand what types of patients were studied and therefore, to which patients the rule may be applicable. The method of patient selection should be free of bias, so that the study subjects encompass a wide clinical and demographic spectrum and are representative of all patients seen at the site with the designated condition.
Fifth, the authors should justify the number of subjects enrolled in the study. Of particular importance is that the sample size must be appropriate for the statistical technique chosen. There may be problems with overfitting the data if there are too few outcome events per predictor variable. A commonly used “rule of thumb” is that there should be at least 10 patients with positive outcomes per independent variable in the prediction rule. For example if a clinical decision rule for appendicitis was composed of 5 variables, there should be at least 50 cases of appendicitis in the study population. Another important consideration in choosing the sample size is the degree of precision desired around the measure of accuracy.

Sixth, the mathematical techniques used to derive a decision rule should be adequately described and justified. Many techniques are available, from a simple 2x2 cross tabulation of each predictor variable with the outcome to sophisticated multivariate analyses. Univariate analyses are easy to perform but do not allow for the exploration of the relationship of predictor variables with each other and with the outcome. Logistic regression analysis predict the likelihood of a binary outcome (appendicitis or no appendicitis) and tend to lead to decision rules with a higher overall accuracy of the classification of patients but possibly less than optimal sensitivity (that is less than 100% classification of abnormal patients). Chi-squared recursive partitioning is another method of analysis. This method progressively divides the patients into a subpopulation that includes only patients with a particular outcome. Rules derived by this method tend to be very sensitive.

Seventh, the decision rule must make good clinical sense, be easy to use and provide a course of action. Decision rules should demonstrate content validity, which means that most clinicians would consider the items in the rule to be clinically reasonable and would find no obvious missing items. Ease of use depends on factors such as the length of time needed to apply the rule and the simplicity of its interpretation. In the emergency department, it is unlikely that physicians would embrace a rule that requires extensive calculations or the use of a calculator. Steill and his colleagues believe that decision rules are more likely to be used if they suggest a course of action rather than merely provide a probability of outcome.
Lastly, eighth, the authors should present the accuracy of the rule based on the population from which the decision rule was derived. Sensitivity, specificity, positive predictive value and negative predictive value should be presented. Likelihood ratios can be presented if information about the posttest probability of an outcome based on the decision rule would be helpful. Receiver operator characteristic (ROC) curves give an overall estimate of the accuracy of a rule, but are of limited clinical use because the posttest probability is not provided.

Once a decision rule has been derived, McGinn and his colleagues stipulate that a rule should only be widely-adopted once they have been adequately tested in validation studies and their impact on clinical behaviour has been assessed[103]. Successful validation and impact studies move a prediction rule through a hierarchy of evidence (Level 4 to Level 1) and they recommend only implementing a rule into routine clinical practice once Level 1 evidence has been achieved. At this point, there is sufficiently large confidence in its accuracy and in its ability to improve upon current practice[104]. When rules are derived but not validated, or validated only in retrospective databases or by sample splitting within the original data set, they are of Level 4 evidence. Progression to Level 3 evidence occurs when the rule has been validated in a similar population or setting as the original derivation population. At this stage, clinicians may consider using the rule with caution if the patients in the study are similar to those in the development setting. Level 2 evidence is attained when accuracy of the rule has been demonstrated in one large prospective study with a broad spectrum of patients and clinicians or in several smaller settings that differ from one another. At this stage, the rule can be used with confidence in its accuracy. Finally, Level 1 evidence is attained after an impact analysis has been done, demonstrating a change in clinician behaviour with beneficial consequences. This validation process is very important because many statistically derived rules or guidelines fail to perform well when tested in a new population. The reason for this poor performance may be statistical (overfitting or instability in the original derived model) or may be related to differences in prevalence of disease or difference in how the decision rule was applied.
In the next section, the clinical decision rules developed for appendicitis will be presented and the adequacy of their methodology will be discussed with reference to the eight criteria presented above. When a rule has been validated, the adequacy of validation will be discussed.

2.9.2 Clinical Scoring Systems in the Diagnosis of Appendicitis

Many appendicitis scoring systems exist, however, until recently, most were developed on adult patients and have not been applied routinely in clinical practice because of a failure to achieve diagnostic accuracy in repeated validation studies[23-29]. Only scores that were either developed for pediatric patients or developed on adult patients but validated in children will be reviewed in this section.

The most widely known appendicitis score was developed by Alvarado in 1986[23]. The score, officially called the MANTRELS score, but also known as the Alvarado score, was developed by a retrospective record review of 277 patients hospitalized for abdominal pain suggestive of appendicitis from 1975 – 1976 in Philadelphia. The mean age of the study population was 25.3 years (range 4 – 80). The number of pediatric patients was not specified. Information about patient age, sex, duration of pain, symptoms, physical signs and the results of laboratory and pathology reports was abstracted from medical records. For each diagnostic indicant abstracted, the sensitivity, specificity, positive and negative predictive values and probability of disease based on a positive and negative test were calculated. A diagnostic weight was assigned to each indicant by adding the true-positive rate and true-negative rate for that item. Eight items with the highest diagnostic weights were retained after univariate association with the outcome (appendicitis) was confirmed. The retained diagnostic indicants, the weights and score value are listed in the table below:
The two components with the highest diagnostic weight (leukocytosis and tenderness in the right lower quadrant) were arbitrarily assigned a value of 2 while the remaining components were assigned a value of 1, such that the score would sum to 10. In this cohort of 277 patients, the mean score for those with appendicitis (227 patients) was 7.71 (SD 1.53) and for those without appendicitis (50 patients) was 5.24 (SD 2.02). When a cut-point of 6 was retrospectively applied to Alvarado’s cohort, such that patients with scores ≤ 5 would be observed and patients with scores ≥ 6 would be taken to the operating room, there would have been 16 (5.8%) cases of missed appendicitis and 24 (8.7%) negative appendectomies. If the cut-point was changed to 5, such that patients with scores ≤ 4 would be observed and patients with scores ≥ 5 would be taken to the operating room, the number of missed cases of appendicitis would fall to 8 (2.9%) but the number of negative appendectomies would increase to 31 (11.2%). At this cut-point, the sensitivity of the score is 92% (95%CI: 89%, 96%), specificity 52% (95%CI: 38%, 65%), positive predictive value 90% (95%CI: 85%, 93%) and negative predictive value 62% (95%CI: 47%, 75%). The prevalence of appendicitis in this cohort was 82%. Alvarado concluded that scores of 5-6 were compatible with appendicitis, scores of 7-8 indicate probable appendicitis and scores of 9-10 indicate very probable appendicitis. The diagnostic accuracy of this score does not improve the negative appendectomy and perforation rates over that of the global physician assessment,
and it may, in fact, perform more poorly in a less selective cohort of patients – that is if the score was applied to all patients with abdominal pain suggestive of appendicitis, and not just to those who were admitted to hospital with a high index of suspicion of appendicitis.

The derivation of the Alvarado score fails to fulfill four important criteria. First, the score elements were abstracted from the medical record and there is no mention of the number of missing data items. This method of data collection is considered unacceptable and places the methodology of the derivation into serious question. Second, the patient population in the derivation cohort was very selective and likely does not represent all patients presenting to that hospital with a suspicion of appendicitis. The prevalence of appendicitis in this cohort was 82%, which far exceeds the prevalence of appendicitis among patients with acute abdominal pain (estimated to be approximately 30%). Third, assessment of the reliability of the variables could not be assessed due to the retrospective nature of the design. Fourth, having three score ranges that are compatible, probable and very probable for appendicitis does not provide clear management directives for the clinician. Finally, the score was developed on the same population that it was applied to, which generally results in an overstatement of the accuracy of the score. In spite of this, the Alvarado score has been repeatedly validated in several pediatric cohorts.

Bond and his colleagues were the first to prospectively validate the Alvarado score in a pediatric cohort[24]. They studied 187 children with abdominal pain of less than one week’s duration and who had all of the score elements collected by the treating physician. One hundred and sixteen (61%) of children had pathology confirmed appendicitis with a very high perforation rate (60% of the appendectomies performed) and a low negative appendectomy rate (5.6%). In this population, there were 2 cases of missed appendicitis. If the MANTRELS score was retrospectively applied to this population, using a cut-point of 5, there would have been no cases of missed appendicitis, but 46 (23%) negative appendectomies. With a cut-point of 7, there would have been 12 (10%) cases of missed appendicitis and 21 (15%) negative appendectomies. The authors concluded that the score did not adequately separate patients with and without
appendicitis, however, using any cut-point to guide operative management would have substantially improved upon the perforation rate in their clinical practice. On this group of patients with a very high prevalence of appendicitis (61%) and an unusually high perforation rate, it would be expected that the MANTRELS score would have been able to better discriminate between those patients with and without appendicitis. It is possible that the varied clinical experience of the enrolling physicians (resulting in less standardized interpretations of physical findings) contributed to this poor performance. Subsequent validation studies in children by Macklin et al. (1997), Impellizzeri et al. (2002) and Schneider et al. (2007) used the same methodology, criteria for patient inclusion and cut-points, and showed similar performance of the score[26, 105, 106]. The prevalence of appendicitis in Schneider’s cohort (34%) was almost half that of the populations of Macklin, Impellizzeri and Bond. Referring to McGinn et al’s hierarchy of evidence, with the inconsistent validation of the Alvarado score in pediatric populations, it has not progressed past Level 4 evidence.

The Alvarado score has been criticized for using the sum of the true-positive rate and true-negative rate to determine the diagnostic weight of each score item. Since Alvarado defined accuracy as the weighted average of the sensitivity and specificity using weights determined by the prevalence of appendicitis, it is problematic when the sensitivity and specificity of a score item are not equal, as the diagnostic weight is directly proportional to disease prevalence[107]. In populations with high disease prevalence, items with high sensitivity are disproportionately rewarded with high diagnostic weights while in populations with low disease prevalence, items with high specificity receive higher diagnostic weights. Peter and Hedges advocate that the likelihood ratio would be a better discriminator between those with and without disease and a clinical score using this method might be more generalizable to other populations. To the best of our knowledge, this proposal for an alternative methodology did not receive any further comments by other investigators and a score based on likelihood ratios was never formally created or validated. Despite this and other methodologic criticisms and the mixed results in the validation studies, the Alvarado score is the most widely used score in pediatric clinical practice.
In 2000, Dado and his colleagues retrospectively applied the modified Lindberg score to 197 pediatric patients operated on for appendicitis[108]. This scoring system was prospectively developed by Fenyo (1987) on a cohort of 259 adult patients using Bayesian methodology[109].

The variables with their score values are listed in the following table:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Indicator</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>Male</td>
<td>+ 8</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>- 8</td>
</tr>
<tr>
<td>WBC (x 1,000 / mm$^3$)</td>
<td>&lt; 12.0</td>
<td>- 15</td>
</tr>
<tr>
<td></td>
<td>12.1 – 20.0</td>
<td>+ 2</td>
</tr>
<tr>
<td>Duration of pain (hours)</td>
<td>&lt; 24</td>
<td>+ 3</td>
</tr>
<tr>
<td></td>
<td>24 – 48</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>&gt; 48</td>
<td>- 12</td>
</tr>
<tr>
<td>Progression of pain</td>
<td>Yes</td>
<td>+ 3</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>- 4</td>
</tr>
<tr>
<td>Temperature ≥ 37.5</td>
<td>Yes</td>
<td>+ 7</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>- 4</td>
</tr>
<tr>
<td>Vomiting</td>
<td>Yes</td>
<td>+ 7</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>- 5</td>
</tr>
<tr>
<td>Migration of pain</td>
<td>Yes</td>
<td>+ 4</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>- 11</td>
</tr>
<tr>
<td>Rebound tenderness</td>
<td>Yes</td>
<td>+ 5</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>- 10</td>
</tr>
<tr>
<td>Rigidity</td>
<td>Yes</td>
<td>+ 15</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>- 4</td>
</tr>
<tr>
<td>Tenderness outside RLQ</td>
<td>Yes</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>+ 4</td>
</tr>
<tr>
<td>Constant</td>
<td></td>
<td>-10</td>
</tr>
</tbody>
</table>
Using abstracted data from the medical record, scores for each patient were calculated. Beginning with a baseline value of –10, the total value for the score is obtained by summing the values assigned to the presence or absence of patient characteristics. Fenyo suggested that patients with scores \( \geq 2 \) should be taken directly to the operating room for appendectomy, patients with scores between –3 and –18 should be admitted for 24 hours of observation and patients with scores \( \leq -19 \) should be discharged home. In Dado’s cohort, 100% of patients were operated on (because of the retrospective methodology used), 23% had a negative appendectomy and 6% had perforated appendicitis. If the modified Lindberg score had been applied to this cohort of 197 patients, 44 (22%) patients would be discharged home, 52 (26%) patients would have been admitted for observation and 101 (52%) patients would have been taken directly to the operating room for an appendectomy. There would have been 15 (34%) patients with missed appendicitis in the group that was discharged home but only 4 (4%) negative appendectomies. In the group of patients that were observed in hospital, 73% had appendicitis. It is difficult to postulate what the disposition of these patients would have been after undergoing their period of observation. The authors report that the score had a sensitivity of 86% with a specificity of 87%. However, in this calculation, they only included patients who were either discharged home or were taken to the operating room, omitting the cohort of patients who were admitted for observation. If we assume that all of the patients who were admitted for observation were operated on, this would increase the sensitivity to 90% but the specificity would decrease to 61%. Conversely, if we assumed that all patients under observation were discharged home without appendectomies, the sensitivity would decrease to 65% and the specificity would increase to 91%. The true estimate of diagnostic accuracy most likely lies somewhere between these two extremes. An additional concern with this study is that the study population contains only children who were operated on for appendicitis which compromises the generalizability of the results. A more relevant validation cohort would be all children presenting with acute abdominal pain suggestive of appendicitis, because this is the cohort of patients that a scoring system would be applied to in clinical practice. The modified Lindberg scoring system has not been validated in other pediatric populations and has not progressed past Level 4 evidence.
In 2002, Samuel developed the first appendicitis score that was unique to children[30]. It was prospectively developed on 1170 children aged 4 to 15 years (mean 10.3 years) who were referred to the pediatric surgical service with abdominal pain suggestive of appendicitis. A standardized data collection form containing demographic data (age and sex), presence of symptoms (anorexia, nausea and vomiting, a history of migration of pain to the right lower quadrant), physical signs (right iliac fossa pain on palpation, right lower quadrant pain with hopping, tenderness with cough or percussion to the RLQ, and temperature elevation), laboratory investigations (WBC, differential, urinalysis) and pathology results from any appendectomies performed was completed on all patients by a pediatric surgeon. Using identical methodology to Alvarado, Samuel calculated the joint probability (the sum of probabilities of disease when the test was positive and negative) for each data item collected. This joint probability represented the diagnostic weight. For each item, diagnostic weights were compared for those with appendicitis to those without. Stepwise multiple linear regression of all parameters was used to create a model for predictors of appendicitis. Eight items (3 signs, 3 symptoms and 2 laboratory investigations) were retained and comprise the Pediatric Appendicitis Score (PAS). Score elements, diagnostic weights and values are in the following table:
<table>
<thead>
<tr>
<th>Diagnostic Indicants</th>
<th>Diagnostic Weight No Appendicitis Group</th>
<th>Score Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>cough or percussion or hop tenderness</td>
<td>0.96</td>
<td>2</td>
</tr>
<tr>
<td>Anorexia</td>
<td>0.88</td>
<td>1</td>
</tr>
<tr>
<td>Pyrexia</td>
<td>0.87</td>
<td>1</td>
</tr>
<tr>
<td>Nausea / emesis</td>
<td>0.86</td>
<td>1</td>
</tr>
<tr>
<td>Tenderness in RLQ</td>
<td>0.84</td>
<td>2</td>
</tr>
<tr>
<td>Leukocytosis &gt; 10,000</td>
<td>0.81</td>
<td>1</td>
</tr>
<tr>
<td>Polymorphonuclear neutrophilia</td>
<td>0.80</td>
<td>1</td>
</tr>
<tr>
<td>Migration of pain</td>
<td>0.80</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

Tenderness elicited by cough, percussion and hopping were highly correlated (r=0.9), and were therefore combined into a single diagnostic variable. The three items combined were highly specific (100%) and predictive of disease (100%) with a high diagnostic index (0.96), so were assigned a score value of 2. Tenderness in the right lower quadrant was also assigned a score value of 2. It was highly specific (100%) and predictive of disease (100%) but had a poor negative predictive value (57%). The authors do not specify why this variable was assigned a higher score over other variables with higher diagnostic weights. The remainder of the diagnostic variables were assigned a value of 1 for a total score value of 10. It should be noted that more specific definitions of pyrexia and polymorphonuclear neutrophilia are not provided.

Of the 1,170 patients evaluated, 734 (63%) had pathology-proven appendicitis. Twenty-nine percent of patients had perforated or gangrenous appendicitis while only 3% had negative appendectomies. Using a cut-point of 5, if Samuel's score had been applied to this population,
such that all children with scores \( \leq 5 \) are discharged home without any investigations and children with scores \( \geq 6 \) are taken to the operating room for an appendectomy, there would have been no cases of missed appendicitis and the negative appendectomy rate would have been 1.6%. At this cut-point, the sensitivity of the score is 100% with a specificity of 92%. Although these values are likely overestimated as they were obtained from the derivation dataset, at the time this was published, the performance of this diagnostic score was superior to any previous reports in the pediatric literature. Also uniquely, definitive management decisions can be taken using these cutoffs because there is no "gray zone" where the diagnosis is considered as uncertain.

The derivation of the PAS adhered to most methodologic standards, but failed to fully satisfy four of the criteria. First, the study population is described as all patients with abdominal pain suggestive of appendicitis, however, the details of how and where they presented are not given. Given that the author is a surgeon, it is assumed, but not known definitively, that all patients in the study population were referred to the surgical service to rule out appendicitis and were not "all comers" presenting to the emergency room with abdominal pain suspicious of appendicitis. The high prevalence of appendicitis (63%) further supports this assumption. This unknown, but presumed element, makes it difficult for the reader to understand the generalizability of the results. Second, some predictors are not fully defined. The duration of time that historical items (eg. nausea/emesis and anorexia) could be present and qualify as being related to the symptoms is not specified. For example, if a child had an episode of nausea 4 days ago and is now presenting with new onset abdominal pain, does this symptom qualify for inclusion in the score? Further, absolute definitions for pyrexia and polymorphonuclear neutrophilia were not provided. There are numerous temperature thresholds that could be interpreted as "pyrexia" and a wide variation in percentages that are considered to qualify as "neutrophilia". The lack of definitions is especially problematic as it does not allow for exact replication of the score in future studies and could contribute to the production of discrepant results in these studies. Third, the reliability of the predictor variables was not assessed. The score contains many subjective pieces of
information from the history and physical examination and without assessment of the intra or inter-observer reliability, the reproducibility of the findings is unknown.

Samuel performed a small validation study after the completion of the derivation study and the results of the 66 patients assessed are provided as an addendum to the original article. Unfortunately, the accuracy in the derivation study was not replicated in the validation sample. He reports that the score had a sensitivity of 100%, specificity of 87%, positive predictive value of 90% and negative predictive value of 100%, however, only the children with PAS $\leq 5$ and PAS $\geq 8$ are included in this calculation. There is no mention of the disposition of the patients with scores between 6 and 7. Samuel does acknowledge that the score elements overlap with other pediatric diseases, and therefore this scoring system cannot provide 100% diagnostic certainty, however, it provides clinicians with a rational approach to a problem. Despite the fact that the diagnosis is uncertain between 6 and 7, this "gray zone" is smaller than that presented by Alvarado and may help clinicians narrow the group of patients on whom to conduct further investigations.

The PAS has been further validated in one cohort of pediatric patients. Schneider and her colleagues prospectively validated the PAS in a convenience sample of 588 patients who had a surgical consultation for possible appendicitis[105]. Pediatric emergency medicine physicians, who were blind to the scoring system, completed data collection forms on all enrolled patients. The authors defined pyrexia as temperature $\geq 37.3$ °C and polymorphonuclear neutrophilia as a neutrophils $\geq 75%$. Data elements were collected prior to surgical evaluation and prior to obtaining imaging investigations. Final diagnoses were obtained from pathology reports if the patient had an appendectomy and by a follow-up phone call 2 weeks after discharge if the patient did not have surgery. The study population had a mean age of 11.9 years. Thirty-four percent of patients had appendicitis and 18% had perforated appendicitis. The authors do not comment on the negative appendectomy rate. Using a cut-point of 5, as suggested by Samuel, the score had a sensitivity of 82% (95%CI: 77%, 87%), specificity of 65% (95% CI: 60%, 70%), negative
predictive value of 88% (95% CI: 84%, 91%) and positive predictive value of 54% (95% CI: 48%, 60%). This indicates that there would have been 36 (12%) cases of missed appendicitis and 136 (45%) patients with a negative appendectomy. We are unable to calculate the performance of the score at other cut-points as sufficient data are not provided. The authors concluded that although the concept of using a clinical scoring system to guide patient care is powerful, the PAS did not perform well enough to recommend incorporation into routine clinical practice. The results from this study are disappointing because they do not replicate the accuracy Samuel showed in his derivation study, however, are not entirely unexpected. As discussed above, the accuracy in Samuel’s study was likely overstated as his estimates were obtained from application of the rule to the original derivation set.

Applying the scoring system to a cohort with a lower pre-test probability of disease may, at least in part, be responsible for this inferior performance. In Samuel’s cohort, the long mean duration of appendicitis (2.3 days) and a high perforation rate (29%) may indicate that only patients with a high probability of disease are referred to surgery for evaluation of appendicitis. This is also reflected in the high prevalence of appendicitis (63%) in this cohort. Although all of the patients in Schneider’s cohort also had surgical consultations, the criteria for consultation were likely more lenient as the patients had a much lower prevalence of disease (34%). Seven hundred and fifty-five patients were recruited in the study, representing 92% of all eligible patients, however datasets were not complete for 167 patients, leaving 588 patients for analysis. Information about the similarity between the patients who were recruited and included in the analysis (588 patients), recruited but not included in the analysis (167 patients) and not recruited (66 patients) is not provided. If these patients were systematically different, the study population would have been subject to a selection bias. Another potential limitation is that the authors excluded all patients who did not have any follow-up. Follow-up was done in one of several ways: they first attempted to contact the child’s parent or guardian and if they were not available, the patient’s pediatrician was contacted. If both the parent and pediatrician were unable to be contacted, the medical record was reviewed to determine final outcome. This method of follow-up is problematic – the
child’s pediatrician may not be aware of final outcome if the patient sought care at an alternative location and the medical record may not contain complete information if a patient was seen at an alternative hospital for further evaluation or for an appendectomy. In this sense, follow-up was not complete in this study. The PAS was not able to progress past Level 4 evidence on the basis of the results from this validation study.

A second appendicitis score for childhood appendicitis was developed by Lintula and his colleagues in 2005[110]. Between December 1999 and October 2000, they prospectively collected 35 clinical data items on 131 children aged 4 – 15 years presenting to the emergency department with suspected appendicitis. Using univariate analysis, 15 predictors were found to have no prognostic significance for appendicitis. Backward stepwise logistic regression was performed on the remaining 19 variables and 9 variables were retained in the final model. Using the coefficients from the logistic regression model, each predictor variable was assigned a score. The variables with their score values are listed in the following table:

<table>
<thead>
<tr>
<th>Predictor Variable</th>
<th>Score (32)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (male)</td>
<td>2</td>
</tr>
<tr>
<td>Intensity of pain (severe)</td>
<td>2</td>
</tr>
<tr>
<td>Relocation of pain</td>
<td>4</td>
</tr>
<tr>
<td>Vomiting</td>
<td>2</td>
</tr>
<tr>
<td>Pain in RLQ</td>
<td>4</td>
</tr>
<tr>
<td>Fever (≥ 37.5°C)</td>
<td>3</td>
</tr>
<tr>
<td>Guarding</td>
<td>4</td>
</tr>
<tr>
<td>Bowel sounds (absent, tinkling, high pitched)</td>
<td>4</td>
</tr>
<tr>
<td>Rebound tenderness</td>
<td>7</td>
</tr>
</tbody>
</table>
In the derivation cohort, 43 (33%) of patients had appendicitis and the authors report that patients with scores of 15 or less could be discharged home, while patients with scores of 21 or more could be taken to the operating room for an appendectomy. Patients with scores between 16 and 20 had an intermediate probability of disease and should be observed. Using these parameters, the score had a sensitivity of 100% (95% CI: 87.5%, 100%) and specificity of 100% (95% CI: 95.6%, 100%).

This score, developed by Lintula and his colleagues, fails to fulfill two important criteria for the derivation of a clinical decision rule. First, the reliability of the predictor variables is unknown as there is not assessment of inter or intra-observer reliability. This is important given that many of the items included in the score are subjective elements of the clinical examination (eg. intensity of pain, quality of bowel sounds). Second, the number of subjects with appendicitis (43 patients) in the derivation sample is inadequate given the number of predictor variables in the scoring system. Using the rule of thumb of 10 patients with a positive outcome for every predictor variable in the decision rule, there should have been 90 children with appendicitis. This is a serious fault as the score may appear to perform better than it actually does due to overfitting of the data.

The same authors subsequently validated the score in 109 pediatric patients and reported these results in the same manuscript as the results of the derivation study. Using the cut-points found in the derivation study, the sensitivity of the score was 85.0% (95% CI: 67.5%, 94.1%) and specificity was 94.9% (95% CI: 87.7%, 98.0%). The confidence intervals around the sensitivity of the score are wide and do not provide the confidence needed to ensure that no patients with appendicitis would be missed using this score. This score has not been externally validated in other settings by other investigators. Again, this score has not progressed past Level 4 evidence and cannot be recommended for implementation into clinical practice.
Since validation studies of clinical scoring systems have not produced consistent results, some investigators have suggested combining diagnostic modalities to explore whether diagnostic performance can be improved. McKay and Shepherd (2007) conducted a retrospective review of a mixed adult and pediatric population of patients who had a CT scan to rule out a diagnosis of appendicitis between February and December 2004 to see if the Alvarado score could help determine which patients should be imaged[111]. The number of pediatric patients is not specified. They report that the Alvarado score is 96% specific when scores of \( \leq 3 \) are used to discharge patients home without further investigations and 75% specific when scores \( \geq 7 \) are used to decide the need for appendectomy without imaging. This correlates with a negative appendectomy rate of 22%. The authors recommend using CT to help direct care for patients with scores of 4, 5 or 6. With the addition of CT scanning to this group of equivocal patients, the sensitivity and specificity were 90% and 95% respectively. When CT scanning was used as the sole diagnostic tool for the entire population, the sensitivity was 94% and specificity was 97%. Although the use of the Alvarado score reduces the amount of radiation patients are exposed to, the sensitivity of this method is poor and should not be recommended. Further, the patients in this study population were likely already subject to some pre-selection by the treating physicians as this group does not likely represent all patients with abdominal pain suggestive of appendicitis. If the pre-test probability of appendicitis in the group of patients who had a CT was higher than the group of patients who had suspected appendicitis but did not have a CT, the performance of the Alvarado score in this study may be overstated. Conversely, it may be understated if the pre-test probability for the patients in the CT scan group was lower.

**Summary of Literature Review**

The importance of diagnosing appendicitis early enough that perforation is avoided while minimizing the number of negative appendectomies that are performed is widely recognized. The diagnosis is a challenge in pediatrics because of the patients inability to clearly express their symptoms, overlap of signs and symptoms with other common childhood illnesses and varied clinical presentations. The search of the perfect diagnostic tool has been unsatisfactory.
Laboratory investigations are a helpful adjunct to diagnosis, but on their own perform poorly. Imaging with ultrasound performs very well in the hands of experienced operators however has important limitations, leading to misclassification of cases, when commonly found patient characteristics such as obesity, pain or an unusual location of the appendix are present. In teaching hospitals where residents often perform ultrasounds, especially after hours, the results from accuracy studies has not been consistent. The only highly sensitive and specific tool is CT scan, but because of its risk associated with lifetime risk of mortality from cancer, it cannot be used freely as a diagnostic aid. Clinical scoring systems have received a lot of attention in the literature because they are cost-effective, time-efficient tools that present no risk to the patient. They systematically quantify the individual contributions that various components make towards diagnosis, prognosis or likely response to treatment. Although many scores exist for appendicitis, most have been developed for adult patients and none have been routinely applied in clinical care because of inconsistent results in pediatric validation studies. The PAS was the first score developed uniquely for pediatric patients. In 2002, Samuel reported this 8 item score to be highly sensitive and specific. The PAS was replicated in a second study with a lower prevalence of appendicitis but did not demonstrate equal accuracy to Samuel. Since the PAS is the most promising diagnostic score for appendicitis that is specific to children, further validation attempts are warranted in methodologically sound studies using all children with abdominal pain suspicious of appendicitis who are evaluated by emergency physicians, in order to ascertain whether the PAS is useful in decreasing the negative appendectomy rates while minimizing missed or delayed diagnoses. If successful, this would allow the PAS to progress to Level 2 evidence.
Chapter 3

Materials and Methods

Study design
We conducted a prospective observational cohort study in the emergency department of the Montreal Children’s Hospital between November 2003 and July 2005. This urban, tertiary care pediatric teaching hospital has an annual emergency department census of 65,000 visits per year. The hospital serves a population of 3 million in the greater Montreal area and, as a part of the Government of Quebec’s new system of integrated university health networks, is a referral center for the Outaouais, Abitibi-Temiscamingue, Montérégie, James Bay and Nunavik. This referral area represents approximately 23% of the population of Quebec. The study received approval from the Research Ethics Board at the Montreal Children’s Hospital (Appendix A). Informed written consent was obtained from all parents or legal guardians and assent was obtained from children 7 years or older (Appendix B).

Selection of participants
All children between the ages of 4 and 18 years with less than 3 days of abdominal pain and in whom the emergency physician considered a diagnosis of appendicitis were eligible for enrollment in the study. Physicians consider a diagnosis of appendicitis based on the assimilation of information obtained from their clinical examination (which includes a complete history and physical examination) and results from additional testing (some may receive laboratory or imaging investigations). At our institution, it is usual practice to obtain a CBC on all children with suspected appendicitis. Children were excluded if they were non-verbal, had a previous appendectomy or had chronic abdominal pathology (eg. inflammatory bowel disease, a history of complex abdominal surgery, or significant congenital abdominal anomalies) that may have interfered with the assessment of the abdomen. Patients were either approached by the part-time emergency department research assistant or by the treating physician for participation in the study. The research assistant was trained in obtaining informed consent.
Methods of measurement and outcome measures

After obtaining informed consent, the emergency physician completed a one-page data collection form. The data collection form contained information about the physician’s level of training (R2, R3, R4, R5, fellow or staff), patient age, gender, date and time of the examination, the date and time of the onset of symptoms, and each of the eight PAS components. The signs and symptoms components of the PAS were documented as present or absent and the numeric results from the CBC (WBC, hemoglobin, hematocrit and percentage of neutrophils) were recorded. The data collection form is included in Appendix C. All data collection forms were completed prior to obtaining any imaging investigations or surgical consultation. All physicians in medical specialties at the second year resident level or higher could enroll patients. Pediatric Emergency Medicine attending physicians and fellows were introduced to the data collection form and study definitions prior to the start of the study. Residents from other specialties working in the emergency department for shorter periods received this same introduction during an orientation session on the first day of their rotation. Physicians were not informed of the weights used in the scoring system and there was no indication of score values on the data collection form. Participation in the study did not influence the child’s treatment in the emergency department or their disposition. Decisions for laboratory or imaging investigations, surgical consultation and disposition from the emergency department were left to the discretion of the treating physician. When possible, a second physician performed an independent assessment of the child and completed a second identical data collection form, to test the interobserver reliability of the score as a whole. It was recommended that this second evaluation be done immediately following the first, however the time of the evaluations was not documented.

Patients who were discharged directly home from the emergency department were contacted (by the principal investigator (MB) or the research assistant) by telephone at one month to verify final outcome. Although most children with abdominal pain caused by appendicitis will progress to perforation within 72 hours of the onset of symptoms, a delay of one month was chosen to
ascertain final outcome for study logistics. It was not felt that this delay would compromise the quality of the information we received during follow-up. Patients or parents were asked about symptom resolution, or if they/their child had an appendectomy at the Montreal Children’s Hospital or elsewhere since their emergency department visit. Responses to these questions were recorded on a paper spreadsheet and later entered into the electronic database. If a patient underwent an appendectomy at the study site or elsewhere, the medical record was obtained and the pathology was reviewed. Appendicitis was defined as appendectomy with positive histology. A negative appendectomy was defined as an appendectomy with negative histology. Missed appendicitis was defined as a child who was discharged home from our emergency department but within one week had an appendectomy with positive histology.

For the purpose of the analysis, the patients were separated into two groups – those with histology confirmed appendicitis and those without appendicitis. The later group included children who underwent appendectomy but who had negative histology. The number and results of imaging investigations with CT or US and the results from the pathology reports were abstracted from the medical records. The components of the PAS were used exactly as described by Samuel, however, he did not provide definitions for polymorphonuclear neutrophilia or pyrexia. Based on previous studies, we defined polymorphonuclear neutrophilia as $\geq 75\%$ neutrophils on complete blood count and pyrexia as $> 38^\circ C$ (oral or rectal). The PAS was calculated for each patient in each group (MB) at the time of data entry. The frequency distribution of PAS scores was calculated for each group.

**Primary data analysis**

Data was entered (MB) on a monthly basis into a Microsoft Access database. For each patient, the age, gender, birth date, date and time of the onset of symptoms, date and time of the emergency department visit, disposition from emergency room (discharge home, discharged but returned and had appendectomy or appendectomy performed at initial visit) and the results of the surgical pathology for patients who had appendectomies were entered. In addition, the presence
or absence of each PAS component (observed tenderness with cough or percussion, observed tenderness with hopping, history of anorexia, history of vomiting, history of migration of the pain to the right lower quadrant, observed maximal tenderness in the right lower quadrant on physical examination, fever in the emergency department, the white blood cell count and the percentage of neutrophils) was entered into the database. Information about the level of physician training was not entered as there was a prohibitively large amount of missing data for this variable. The score was calculated as detailed by Samuel[30]. The sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), positive and negative likelihood ratios (LR +ve and LR -ve respectively) with 95% confidence intervals (CIs) were calculated for each value of the score (1 – 10) to determine how well the score could predict appendicitis at each point. A receiver operator characteristic curve (ROC) was created to assess the overall performance of the score. Cohen’s $\kappa$ coefficient was calculated to measure interobserver agreement on the subset of patients evaluated by 2 physicians[112]. A reliable estimate was considered to have a $\kappa$ higher than 0.6. Data was analyzed using SAS (version 9.1, SAS Institute Inc., Cary, NC) statistical software.
Chapter 4

Results

A convenience sample of 275 patients was enrolled between November 2003 and July 2005. Twenty-nine patients were excluded for failure to meet inclusion criteria (14 patients with abdominal pain > 72 hours or age < 4 years) and for not having a CBC (15 patients). Of the remaining 246 patients, the mean age was 10.9 years (IQR 8.1,13.7 years) and no patients had missing data for any of the score components. Ninety-five children were taken to the operating room for appendectomies. Of these children, 83 (34% of the total population) had pathology proven appendicitis, 14 (16.9%) had perforated appendicitis and 12 (12.6%) had negative appendectomies. One hundred percent of patients were contacted by telephone at one month to verify final outcome. There were no cases of missed appendicitis (patients who were discharged home from the emergency room who subsequently underwent an appendectomy). The patients with and without appendicitis were similar, with the exception of mean PAS score, and their characteristics are described in Table 1. The PAS was calculated for each patient in each group and the distribution of scores is shown in Figure 1.

A ROC curve (Figure 2) was constructed for the PAS performance in our population and yielded an area under the curve of 0.895. Although this indicates that the PAS is a highly accurate test, a single cut-point could not be used to direct care in our population as the combination of false positives and negatives at each point was unacceptably high. The best cut-point, found at the upper left hand corner on the ROC curve, was the same as Samuel’s, where children with scores of 6 or more are taken to the operating room for an appendectomy and children with scores of 5 or less are discharged home with a diagnosis of “no appendicitis”. At this point, the score is very sensitive [92.8% (95%CI: 85.1%, 96.6%)] but not very specific [69.3% (95%CI: 61.9%, 75.9%)]. There would be 50 (37.6%) negative appendectomies, but only 6 (7.2%) cases of missed appendicitis. The sensitivity, specificity, negative predictive value (NPV), positive predictive value
(PPV), likelihood ratios for positive (LR +ve) and negative (LR –ve) tests, potential negative appendectomy and missed appendicitis rate were calculated for each score value and can be found in Tables 2 a – j. No other single cut-point offered an improved sensitivity and specificity. However, the score’s performance did improve when 2 thresholds were used – one to decide who could be discharged home and one to decide who should have an appendectomy – to direct patient care.

If a score of 4 or less was used to discharge patients home without further investigations, 2 patients (2.4%) with appendicitis would have been discharged home. At this cut-point, the score had a sensitivity of 97.6 (95% CI: 91.6%, 99.3%) with a NPV of 97.7% (95% CI: 92.0%, 99.4%). If a score of 8 or more was used to determine the need for appendectomy, 8 children (8.8%) would have undergone a negative appendectomy. At this cut-point, the score had a specificity of 95.1% (95% CI: 90.6%, 97.5%) and a PPV of 85.2% (95% CI: 73.4%, 92.3%). Patients with scores between 5 and 7 have an uncertain diagnosis, and the score is unable to adequately distinguish between those who have appendicitis and those who don’t. These children require further imaging investigations to determine the etiology of their abdominal pain. If this decision-making process had been applied to our population, 37 (41%) imaging investigations would have been avoided because patients would have either been discharged home with a low-likelihood of appendicitis or taken directly to the operating room for a high-suspicion of appendicitis without any further investigations. The distribution of imaging investigations according to score category (scores \( \leq 4 \), scores 5 - 7, scores \( \geq 8 \)) is described in Table 3.

Interobserver scores were obtained for 37 (14.6%) of the total 246 patients. The \( \kappa \) coefficient was 0.65, indicating substantial agreement beyond chance between the raters. Ninety-two percent of the pairs correlated within 2 points (Figure 3). Five patients’ management would have changed on the basis of these scores. Four of these patients would have undergone further imaging instead of being discharged home directly. None of these patients had appendicitis. One patient
would have been taken directly to the operating room instead of undergoing further imaging. This patient had pathology confirmed appendicitis.
Tables

Table 1. Study subject characteristics

<table>
<thead>
<tr>
<th>Patient Characteristics</th>
<th>Appendicitis (n = 83)</th>
<th>No Appendicitis (n = 163)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean age (IQR)</td>
<td>11.6 years (9.0, 14.2)</td>
<td>10.5 years (7.6, 14.5)</td>
</tr>
<tr>
<td>Sex (female)</td>
<td>35%</td>
<td>43%</td>
</tr>
<tr>
<td>Mean symptom duration</td>
<td>1.5 days</td>
<td>1.6 days</td>
</tr>
<tr>
<td>Mean PAS (SD)</td>
<td>7.5 (1.2)</td>
<td>4.3 (1.5)</td>
</tr>
<tr>
<td>Imaging investigations</td>
<td>37 (46%)</td>
<td>54 (33%)</td>
</tr>
</tbody>
</table>

Table 2. Sensitivity, Specificity, PPV and NPV for PAS score 1 – 10

These tables show the sensitivity, specificity, PPV and NPV for each score value.

The tables are set up as follows:

<table>
<thead>
<tr>
<th>Positive Test</th>
<th>Appendixitis</th>
<th>No Appendixitis</th>
<th>Positive Predictive Value (PPV) = a / (a + b) (95% CI)</th>
<th>Likelihood Ratio for positive test (LR +ve) = sens / (1-spec) (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative Test</td>
<td></td>
<td></td>
<td>Negative Predictive Value (NPV) = d / (c + d) (95% CI)</td>
<td>Likelihood Ratio for negative test (LR -ve) = (1-sens) / spec (95% CI)</td>
</tr>
</tbody>
</table>

- Sensitivity = a / (a + c) (95% CI)
- Specificity = b / (b + d) (95% CI)

Potential negative appendectomy rate = b / (a+c+b)

Potential missed appendicitis rate = c / (a+c)

= Number of cases of missed appendicitis

= Number of negative appendectomies
a) PAS = 1

Positive Test = PAS ≥ 1; Negative Test = PAS ≤ 0

<table>
<thead>
<tr>
<th>Appendicitis</th>
<th>No Appendicitis</th>
<th>PPV = 34.7% (29.0%, 41.0%)</th>
<th>LR +ve = 1.03 (099, 1.08)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAS ≥ 1</td>
<td>83</td>
<td>156</td>
<td></td>
</tr>
<tr>
<td>PAS ≤ 0</td>
<td>0</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

Sensitivity=100% (95.6%, 100%) Specificity=4.3% (2.1%, 8.6%)

Potential negative appendectomy rate = 65.3%
Potential missed appendicitis rate = 0%

b) PAS = 2

Positive Test = PAS ≥ 2; Negative Test = PAS ≤ 1

<table>
<thead>
<tr>
<th>Appendicitis</th>
<th>No Appendicitis</th>
<th>PPV = 36.1% (30.1%, 42.5%)</th>
<th>LR +ve = 1.10 (1.04, 1.16)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAS ≥ 2</td>
<td>83</td>
<td>147</td>
<td></td>
</tr>
<tr>
<td>PAS ≤ 1</td>
<td>0</td>
<td>16</td>
<td></td>
</tr>
</tbody>
</table>

Sensitivity=100% (95.6%, 100%) Specificity=9.0% (8.5%, 9.4%)

Potential negative appendectomy rate = 63.9%
Potential missed appendicitis rate = 0%
c) **PAS = 3**

Positive Test = PAS ≥ 3; Negative Test = PAS ≤ 2

<table>
<thead>
<tr>
<th>Appendicitis</th>
<th>No Appendicitis</th>
<th>PPV = 39.0%</th>
<th>LR +ve = 1.24</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAS ≥ 3</td>
<td>83</td>
<td>(32.7%, 45.7%)</td>
<td>(1.15, 1.35)</td>
</tr>
<tr>
<td>PAS ≤ 2</td>
<td>0</td>
<td>NPV = 100%</td>
<td>LR –ve = 0</td>
</tr>
<tr>
<td></td>
<td>130</td>
<td>(89.6%, 100%)</td>
<td></td>
</tr>
</tbody>
</table>

Sensitivity = 100% (95.6%, 100%)  
Specificity = 20.2% (14.8%, 27.0%)

Potential negative appendectomy rate = 61.0%
Potential missed appendicitis rate = 0%

---

d) **PAS = 4**

Positive Test = PAS ≥ 4; Negative Test = PAS ≤ 3

<table>
<thead>
<tr>
<th>Appendicitis</th>
<th>No Appendicitis</th>
<th>PPV = 42.9%</th>
<th>LR +ve = 1.48</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAS ≥ 4</td>
<td>82</td>
<td>(36.1%, 50.0%)</td>
<td>(1.33, 1.66)</td>
</tr>
<tr>
<td>PAS ≤ 3</td>
<td>1</td>
<td>NPV = 98.2%</td>
<td>LR –ve = 0.04</td>
</tr>
<tr>
<td></td>
<td>109</td>
<td>(90.4%, 99.7)</td>
<td>(0, 0.13)</td>
</tr>
</tbody>
</table>

Sensitivity = 98.8% (93.5%, 99.8%)  
Specificity = 33.1% (26.4%, 40.7%)

Potential negative appendectomy rate = 57.8%
Potential missed appendicitis rate = 1.2%
e) **PAS = 5**

Positive Test = PAS ≥ 5;  Negative Test = PAS ≤ 4

<table>
<thead>
<tr>
<th></th>
<th>Appendixitis</th>
<th>No Appendixitis</th>
<th>PPV = 50.9% (43.2%, 58.6%)</th>
<th>LR +ve = 2.04 (1.75, 2.43)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAS ≥ 5</td>
<td>81</td>
<td>78</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PAS ≤ 4</td>
<td>2</td>
<td>85</td>
<td>NPV = 97.7% (92.0%, 99.4%)</td>
<td>LR –ve = 0.05 (0, 0.13)</td>
</tr>
</tbody>
</table>

Potential negative appendectomy rate = 48.4%
Potential missed appendicitis rate = 2.4%

f) **PAS = 6**

Positive Test = PAS ≥ 6;  Negative Test = PAS ≤ 5

<table>
<thead>
<tr>
<th></th>
<th>Appendixitis</th>
<th>No Appendixitis</th>
<th>PPV = 60.6% (51.9%, 68.7%)</th>
<th>LR +ve = 3.02 (2.42, 3.91)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAS ≥ 6</td>
<td>77</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PAS ≤ 5</td>
<td>6</td>
<td>113</td>
<td>NPV = 95.0% (89.4%, 97.9%)</td>
<td>LR –ve = 0.10 (0.04, 0.20)</td>
</tr>
</tbody>
</table>

Potential negative appendectomy rate = 37.6%
Potential missed/perforated appendicitis rate = 7.2%


**g) PAS = 7**

Positive Test = PAS ≥ 7; Negative Test = PAS ≤ 6

<table>
<thead>
<tr>
<th></th>
<th>Appendicitis</th>
<th>No Appendicitis</th>
<th>PPV</th>
<th>LR +ve</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAS ≥ 7</td>
<td>61</td>
<td>24</td>
<td>71.8% (61.4%, 80.2%)</td>
<td>5.00 (3.49, 7.69)</td>
</tr>
<tr>
<td>PAS ≤ 6</td>
<td>22</td>
<td>139</td>
<td>86.3% (80.2%, 90.8%)</td>
<td>0.31 (0.20, 0.43)</td>
</tr>
</tbody>
</table>

- Sensitivity = 73.5% (63.1%, 81.8%)
- Specificity = 85.3% (79.0%, 89.9%)

Potential negative appendectomy rate = 22.4%
Potential missed appendicitis rate = 26.5%

**h) PAS = 8**

Positive Test = PAS ≥ 8; Negative Test = PAS ≤ 7

<table>
<thead>
<tr>
<th></th>
<th>Appendicitis</th>
<th>No Appendicitis</th>
<th>PPV</th>
<th>LR +ve</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAS ≥ 8</td>
<td>46</td>
<td>8</td>
<td>85.2% (73.4%, 92.3%)</td>
<td>11.31 (6.09, 26.25)</td>
</tr>
<tr>
<td>PAS ≤ 7</td>
<td>37</td>
<td>155</td>
<td>80.7% (74.6%, 85.7%)</td>
<td>0.47 (0.36, 0.58)</td>
</tr>
</tbody>
</table>

- Sensitivity = 55.4% (44.7%, 65.6%)
- Specificity = 95.1% (90.6%, 97.5%)

Potential negative appendectomy rate = 8.8%
Potential missed appendicitis rate = 44.6%
i) **PAS = 9**

Positive Test = PAS ≥ 9; Negative Test = PAS ≤ 8

<table>
<thead>
<tr>
<th>Appendicitis</th>
<th>No Appendicitis</th>
<th>PPV = 96.3% (81.7%, 99.3%)</th>
<th>LR +ve = 52.17 (13.27, 2008.76)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAS ≥ 9</td>
<td>26</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>PAS ≤ 8</td>
<td>57</td>
<td>162</td>
<td></td>
</tr>
</tbody>
</table>

Sensitivity=31.3% (22.4%, 42.0%) Specificity=99.4% (96.6%, 99.9%)

Potential negative appendectomy rate = 1.2%

Potential missed appendicitis rate = 68.7%

---

j) **PAS = 10**

Positive Test = PAS ≥ 10; Negative Test = PAS ≤ 9

<table>
<thead>
<tr>
<th>Appendicitis</th>
<th>No Appendicitis</th>
<th>PPV = 100% (56.5%, 100%)</th>
<th>LR +ve = ∞</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAS ≥ 10</td>
<td>5</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>PAS ≤ 9</td>
<td>78</td>
<td>163</td>
<td></td>
</tr>
</tbody>
</table>

Sensitivity=6.0% (2.6%, 13.3%) Specificity=100% (97.7%, 100%)

Potential negative appendectomy rate = 0%

Potential missed appendicitis rate = 94.0%
Table 3. Imaging investigations separated by score category

<table>
<thead>
<tr>
<th>Score 0 – 4 n=81</th>
<th>Imaging</th>
<th>No Imaging</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 – 4</td>
<td>2</td>
<td>19</td>
</tr>
<tr>
<td>5 – 7</td>
<td>22</td>
<td>33</td>
</tr>
<tr>
<td>8 – 10</td>
<td>14</td>
<td>2</td>
</tr>
</tbody>
</table>

= Imaging investigations that would have been avoided if children with scores ≤ 4 were discharged home with a low likelihood of appendicitis and children with scores ≥ 8 were taken directly to the operating room without any further investigations.
Figures

Figure 1. Distribution of scores in patients with and without appendicitis

![Figure 1](attachment:figure1.png)

Figure 2. Receiver Operator Characteristic Curve

![Figure 2](attachment:figure2.png)

Area under the curve 0.895 (95%CI: 0.855, 0.934)
Figure 3. Distribution of the correlation of inter-observer scores

Correlation of Interobserver Scores

- Perfect Agreement
- Correlation within 1 point
- Correlation within 2 point
- Correlation within 3 point
Chapter 5

Discussion

In this prospective validation study of the PAS using a convenience sample of children aged 4 to 17 years (mean 10.9 years) with abdominal pain suggestive of appendicitis, we were unable to reproduce the accuracy of the PAS estimated by Samuel in his derivation study in 2002. Despite a relatively good estimate of overall accuracy by the ROC curve (AUC = 0.895) in our population, using the single cut-point proposed by Samuel to decide if patients should be discharged home (scores ≤ 5) or undergo an appendectomy (scores ≥ 6) resulted in an unacceptably high rate of negative appendectomies (37.6%). The sensitivity and specificity at this cut-point were 92.8% (95% CI: 85.1%, 96.6%) and 69.3% (95% CI: 61.9%, 75.9%) respectively. These results are similar to those found by Schneider and her colleagues. To our knowledge, our study is the first to examine the inter-observer reliability of the score and showed good agreement.

Given these findings, we proposed two cut-points to determine which patients should be discharged home, undergo an appendectomy or undergo further imaging investigations to determine the etiology of their abdominal pain. The first cut-point (score ≤ 4) had a high sensitivity [97.6% (95% CI: 91.6%, 99.3%)] and NPV [97.7% (95% CI: 92.0%, 99.4%)] while the second cut-point (score ≥ 8) had a high specificity [95.1% (95% CI: 90.6%, 97.5%)] and PPV [85.2% (95% CI: 73.4%, 92.3%)]. If this decision-making strategy was applied to our population and compared to usual care at the Montreal Children’s Hospital, the negative appendectomy rate would decrease from 12.6% to 8.8% and imaging investigations would decrease by 41%, however, the missed appendicitis rate would increase from 0% to 2.4%. Minimizing the number of imaging investigations is particularly important in the pediatric population for two reasons. First, when using ultrasound to diagnose appendicitis, definitive management decisions may be delayed because the results of the study may not help with the diagnosis. Second, CT scans have the potential to harm the patient by increasing their lifetime mortality risk from cancer.
These results are based on the assumption that the children with scores between 5 and 7 receive an accurate diagnosis with further imaging investigations.

The utility of such a score may be even more important in settings that do not have the pediatric expertise found in Children’s Hospitals. By standardizing the collection, assimilation and interpretation of patient information, the PAS has the potential to have a positive impact on patient outcomes while minimizing the number of imaging investigations ordered.

The study must be interpreted in light of the following limitations.

First, we did not track missed but eligible patients. We used a convenience sample of children who were enrolled in the study at the physician’s discretion. Although physicians were encouraged to enroll every child in whom they were considering a diagnosis of appendicitis, due to the busy work environment in the emergency department this often does not happen. There may have been an over-representation of equivocal cases in our sample if physicians tended to enroll children when the diagnosis was uncertain as opposed to when they were more certain of disposition (home or operating room), decreasing the sensitivity and specificity of the score. This could have contributed to the discrepancy between our results and Samuel’s results, as his cohort contained patients with a higher probability of disease. The prevalence of disease in Samuel’s population was 63% compared to 34% in ours. If, on the other hand, there was an over-representation of patients with a more certain diagnosis of appendicitis in our population, then the values we obtained for sensitivity and specificity would be an exaggeration of the true estimates.

Second, we allowed all medical physicians at the second year resident level or higher to enroll patients and were unable to determine the effect of training level on the performance of the score due to prohibitively large amounts of missing data for this variable. We initially chose this level of training because it was felt that after 2 years of post-graduate training (equivalent to a family
medicine residency) physicians should have the clinical skills necessary to make an accurate physical assessment. However, since the assessment of children is a skill that can take years to become comfortable with and perfect, it is possible that residents who are not in pediatric specialties were unable to provide as accurate an assessment of the physical signs as an attending staff physician in pediatric emergency medicine. There are 4 points attributed to the physical assessment in the PAS, and therefore any inaccuracies in this regard could substantially influence the score results. However, this most likely did not occur, because we know from the inter-observer reliability analysis that the agreement was good with 92% of assessments correlating within 2 points.

Third, the inter-observer scores were completed by several different pairs of physicians throughout the study period. Not having consistent raters could have affected the accuracy of our estimate of agreement. We are unable to say in which direction this might have occurred.

Fourth, we assessed the reliability of the score as a whole but did not assess the reliability of the score components. This would have given us insight into the signs that had the greatest degree of inter-observer reliability, and would have been particularly important given that we used physicians with varied experience and clinical skills.

Several factors may explain why the PAS did not perform as well in our population as it did in Samuel’s using a single cut-point. First, we may have used different definitions for pyrexia and neutrophilia. We created definitions for these terms as they were not provided in Samuel’s manuscript. It is possible that using discrepant thresholds for these values could have influenced the performance of the score. Second, our population was less selective than Samuel’s, containing all children presenting to the emergency department with abdominal pain suggestive of appendicitis, not just those with a high suspicion of the disease. Eliminating this selection bias would have decreased the positive and negative predictive values of the score. Finally, a factor related to the derivation of the rule may also have contributed to our discrepant results. The PAS
was created using step-wise regression to guide the inclusion of variables into the model. Failure of clinical prediction rules to achieve comparable performance in future validation studies is a well-recognized problem in models developed using these methods[113]. Using p-values to guide inclusion or exclusion of independent variables in a model tends to find strong evidence for weak or non-existent effects, and these techniques ignore model uncertainty, leading to poor generalizability. Although we have no proof to say that this occurred in the case of the PAS, it is possible that using an alternate model selection method, such as Bayes Factors, could have diminished these concerns, producing a more generalizable model with less overfitting to the data.

Using McGinn et al’s hierarchy of evidence (Level 4 to Level 1)[103], the PAS has not progressed past Level 4 evidence. Schneider and we attempted to validate the PAS in the hands of medical physicians and in a less selective population of children. It is important for a rule to show stability in different populations and in the hands of different clinicians if it is to be able to progress through the hierarchy of evidence. This failure to show equivalent or acceptable performance in a different population of children and in the hands of medical physicians instead of surgeons does not allow it to progress to Level 2 evidence. Unfortunately, the score has not yet been validated in a population that is similar to the one in which Samuel derived the score, thus, we do not know whether Level 3 evidence is attainable or not.

An accurate, rigorously developed scoring system for childhood appendicitis would be extremely valuable to clinicians. Using Samuel’s single cut-point to guide care, the PAS has failed validation in two independent cohorts of children. We believe it is unrealistic to expect that a score with a single cut-point will be able to accurately categorize all patients with and without appendicitis, given the extremely varied presentation of childhood appendicitis. However, using a score to decide who should be discharged home directly, who should receive imaging because of equivocal findings and who should be taken directly for an appendectomy is more realistic and almost as useful. We cannot recommend the adoption of our proposed cut-points into clinical
practice without further investigation and validation, however, we believe this strategy shows promise and should be prospectively evaluated in a subsequent cohort of children. If this scoring system proves to be accurate in a validation study of all children presenting to an emergency department with abdominal pain suggestive of appendicitis when evaluated by experienced pediatric medical physicians, it will achieve Level 3 evidence. This implies that experienced pediatric clinicians can have confidence in applying the rule to a population of children that is similar to that of the validation cohort. Important subsequent steps would be to validate the scoring system in a variety of clinical settings by clinicians of varying skills followed by an assessment of the impact of the implementation of the scoring system on clinical practice, patient outcomes and/or cost effectiveness. Finally, assessing the incremental value of each diagnostic technique (e.g. physical examination, PAS score, ultrasound, CT) using multi-variable logistic regression would provide new and interesting information for clinicians.
Bibliography


104. Moore, C., T.G. McGinn, and W. Ho, *adapted from Mount Sinai School of Medicine Home page*.


