Is ultrasonography or computed tomography more accurate in diagnosing adults with suspected appendicitis? A systematic review

Ibrahim Hadadi
Student ID: 200866167

Submitted in accordance with the requirements for the degree of
MSc Medical Imaging
Module Code MEDP5330M

University of Leeds
Division of Biomedical Imaging

“I confirm that this submission is my own work and that I have not presented anyone else’s work as my own, and that full and appropriate acknowledgement has been given where reference has been made to the work of others. I have read and understood the University’s published rules on plagiarism and also any rules specified at School/Faculty level. I understand that if I commit plagiarism I can be expelled from the University and that it is my responsibility to be aware of the University’s regulations on plagiarism and their importance.”

September 2017

Word count: 5226
Abstract: 246
Abstract

**Objective:** The study objective was to conduct a systematic review to compare the accuracy of ultrasonography (US) and computed tomography (CT) in the diagnosis of acute appendicitis in adults. The current reference standard for diagnosis of appendicitis was histopathology and surgery.

**Methods:** A search, restricted to studies published from 2010 onwards, was performed of Medline, Embase and Google Scholar in July 2017. Backwards and forwards reference list checking was employed to identify studies not recovered during the electronic database search. The title and abstract of each study were scanned and the full text for each included study was evaluated. The methodological quality of the included studies was assessed using the Quality Assessment of Diagnostic Accuracy Studies 2 tool. The study results were synthesised narratively and weighted averages were calculated.

**Results:** Only six of the 83 identified studies were included. According to the weighted averages, CT (95%) was more sensitive than US (75%) in diagnosing acute appendicitis. However, minor variations were observed between US and CT regarding the weighted averages obtained for specificity (82% versus 87%), positive predictive value (PPV) (92% versus 95%) and negative predictive value (NPV) (55% versus 61%).

**Conclusion:** Although CT was superior to US in diagnosing adults with suspected acute appendicitis, the two modalities were comparable in terms of specificity, PPV and NPV. Thus, the use of CT is recommended as a conditional modality. The patients should be referred to CT if the results are negative.

**Keywords:** appendicitis, adults, accuracy, ultrasonography, computed tomography
Introduction

Appendicitis is inflammation of the appendix. It is one of the most critical acute abdominal conditions in paediatric and adult patients admitted to hospitals, with a reported prevalence of 8% in males and 7% in females (Reich et al., 2011). Globally, over 250,000 diagnoses of appendicitis are made annually (Sternbach and Rosen, 1995). Appendicitis is difficult to diagnose as the patient does not present with obvious pathognomonic signs and symptoms. In addition, a differential diagnosis of appendicitis is often a clinical challenge because the signs and symptoms mimic those of other abdominal conditions, such as diverticulitis, pelvic inflammatory disease, duodenal ulcer, and gall bladder-related disease (Sternbach and Rosen, 1995).

According to Gaitini (2011), 80% of acute appendicitis diagnoses are made based on the physical examination and laboratory tests. Diagnostic imaging is required to enhance the accuracy of the diagnosis, minimise the development of complications (such as perforation and peritonitis) and reduce the negative appendectomy rate (Hernanz-Schulman, 2010).

The two common diagnostic imaging modalities, ultrasonography (US) and computed tomography (CT), are invaluable in expediting the identification and contemporary treatment of appendicitis and in reducing associated morbidity and mortality (Atema et al., 2015). CT has been shown to be more accurate than US in diagnosing appendicitis in adults, with reported sensitivity and specificity of 95–98% and 94–99%, respectively, compared to 74–86% and 81–97%, respectively, for US (Terasawa et al., 2004; Gaitini et al., 2008). However, the significance of US as a valuable diagnostic modality, particularly with regard to reducing the negative appendectomy rate, has been demonstrated in several studies. For example, according to Khanzada et al. (2009) and Saeed et al. (2009), the sensitivity and specificity of US were 85-98% and 88-98%, respectively. This indicates that US has high diagnostic capability that can be applied to patients with suspected appendicitis presenting at the emergency department.

Although CT is superior with regard to diagnostic accuracy, there are three limitations to the use of abdominal CT. The first is that the use of the scan involves high ionising radiation dose, which is considered a risk factor for the development of cancer (Brenner and Hall, 2007). The second is that it is not readily available in medical centres, mostly in developing countries, owing to its high cost (Howell et al., 2010). Finally, when the contrast medium is administered orally, the length of stay in emergency department for patients will increase. On the other hand, administering contrast medium intravenously (IV) may cause an allergic reaction or nephrotoxicity (Berg et al., 2006).

US has conventionally been used to diagnose suspected acute appendicitis, and unlike CT, is not limited by issues of safety concerns, an extended hospital duration and prohibitive financial expense. This is because it is a relatively safe procedure, does not require patient preparation and is readily available (Pinto et al., 2013). Unlike CT, one of the main
disadvantages of US is that the accuracy of the latter primarily relates to the expertise of the person performing it. It has been argued that sensitivity comparable to that achieved using CT, i.e., as high as 90%, can be realised when US is carried out by a skilled US operator (Van Randen et al., 2011). Furthermore, the accuracy of ultrasonography can be lower than that of CT in certain populations, including obese patients, women of reproductive age, and in cases where there is variation in the position of the appendix (Pinto et al., 2013; Van Randen et al., 2009). It has been found that, it is challenging to perform a US scan in obese patients as fatty tissues reduce the scanning resolution (Old et al., 2005; Sauvain et al., 2016). This is also a clinical challenge in women as the appendix is located in the right lower quadrant of the abdomen and thus the appendicitis could be confused with other pelvic conditions, such as the rupture of an ovarian cyst or twisting of the right ovary (Kruszka and Kruszka, 2010).

A précis of studies on the accuracy of US and CT in diagnosing adults with suspected acute appendicitis has not been included in a recent systematic review. The last systematic review that was relevant to this topic was conducted by Weston et al. in 2005.

Therefore, the current study objective was to conduct a systematic review of studies on the diagnostic accuracy of US and CT in diagnosing acute appendicitis in adults. This was accomplished by:

- Critically evaluating the review methods used in the studies.
- Carrying out a search process of relevant studies for inclusion in the review.
- Summarising and grouping the findings of the included studies.
- Evaluating the practical and academic implications of the findings.

**Method**

When conducting a systematic review, the recommendation is that reviewers should identify and justify the following elements; search strategy, study eligibility criteria, study selection process, data extraction process, study quality assessment and data synthesis process (Gough et al., 2012; Liberati et al., 2009; Petticrew & Roberts, 2008; Tacconelli, 2010). Therefore, these elements are taken into consideration in this study.

**Search Strategy**

Medline and Embase databases were searched using Ovid. Since many studies may not be indexed in main databases such as Medline and Embase (Khan et al., 2011; Petticrew and Roberts, 2008), Google Scholar was searched to retrieve such studies, thereby, this reduces the risk of missing relevant studies (Lefebvre et al., 2008; Popay et al., 2006). Backwards and forwards reference searching was used to identify other studies not identified in the electronic databases. The reference list of each included study was screened to identify other relevant research (backwards reference searching) and Google Scholar was utilised to identify any new relevant research in which the included studies had been cited (forwards reference searching). Medical subject headings (MeSH) were applied to cover all search
terms relating to “adults”, “ultrasonography”, “computed tomography”, and “appendicitis” (Table 1). The Boolean operators (OR) and (AND) were also used to minimise the chance of irrelevant articles being put forward and to expand the search.

Table 1: The search terms used that were applied to electronic databases to identify relevant studies for inclusion in the current review

<table>
<thead>
<tr>
<th>Search Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>exp ultrasonography OR ultraso*.mp.</td>
</tr>
<tr>
<td>AND</td>
</tr>
<tr>
<td>exp tomography, X-ray computed OR “computed tomography” .mp.</td>
</tr>
<tr>
<td>AND</td>
</tr>
<tr>
<td>exp diagnostic accuracy OR “sensitivity and specificity” .mp.</td>
</tr>
<tr>
<td>AND</td>
</tr>
<tr>
<td>exp appendicitis OR “appendicitis” .mp.</td>
</tr>
<tr>
<td>AND</td>
</tr>
<tr>
<td>exp adult</td>
</tr>
<tr>
<td>AND</td>
</tr>
<tr>
<td>exp adolescent</td>
</tr>
</tbody>
</table>

The inclusion and exclusion criteria were based on the well-known acronym “PICTR” (participants [adolescents and adults], index tests [US], the comparator [CT], the targeted condition [appendicitis], and the references [histopathology or a surgical procedure]), in order to correlate them with the question posed in the systematic review (Bossuyt and Leeflang, 2008).

Studies were included if the main focus was on adult patients with suspected appendicitis, and those in which adolescents aged ≥ 12 years and adults were evaluated (with the majority of the sample being adults). A decision was made to primarily assess adult populations because in children, preference is given to the use of US rather than CT (Pinto et al., 2013). In addition, generally, the accuracy of US is determined in research, without the application of CT, to avoid exposing paediatric populations to ionising radiation. Thus, the establishment of the most optimum diagnostic modality for use in adults remains a point of contention (Pinto et al., 2013). A final inclusion criterion was studies that were published from 2010 onwards to ensure the inclusion of current imaging modalities in the studies. This is because the resolution of radiological modalities have improved in the last ten years (Carroll et al., 2013). Thus, the diagnostic accuracy of US and CT would be varied depending on which type of modality was employed during the diagnosis of appendicitis.

Exclusion criteria were studies in which other modalities (such as magnetic resonance imaging) were used, in which paediatric populations were evaluated, and in which the
diagnostic ability of US and CT was examined in relation to diseases other than appendicitis. Studies for which sufficient data extraction was not possible, for instance, a case report paper, were also excluded, as were those that were reported in a language other than English.

**Quality Assessment of the Included Studies**

The quality of the methodology of the included studies was assessed using the Quality Assessment of Diagnostic Accuracy Studies (QUADAS)-2, which is divided into two parts. The risk of bias in relation to patient selection, the index test, reference standard, and flow and timing was evaluated using the first part. Bias risk in terms of patient selection was considered to be low if consecutively presenting patients with suspected appendicitis were randomly assigned to different treatment groups. Also, it is important to avoid inappropriate exclusions (Whiting et al., 2011).

In relation to the index test, it needs to ensure that the way in which the index test is conducted and interpreted is not influenced by knowledge of the reference standard results. Thus, the US radiologists had to be blinded to the reference standard findings (histopathology and surgery), with a parallel scenario in terms of the comparator (CT radiologists) to make certain that the risk of bias in both the US and CT interpretations was low. Accordingly, a low bias risk is considered to apply if the index test is conducted and interpreted prior to receiving the results for the reference standard (Whiting et al., 2011). According to Leeflang et al., (2008), the threshold of the index test should be pre-specified to prevent an estimate of its diagnostic accuracy being influenced.

Ideally, an interpretation of the reference standard should take place without prior knowledge of the index test and comparator results (Whiting et al., 2011). Furthermore, the accuracy of the index test is usually based on the assumption that the sensitivity of the reference standard is 100% in accurately identifying the targeted condition. This premise is supported by Biesheuvel et al. (2007) who reported that specific disagreements over the interpretation of the index test and reference standard results are usually caused by the incorrect diagnosis of the index test.

A low risk of bias is attributed to the flow and timing of the tests if all the patients are exposed to the same reference standard (Rutjes et al., 2003). Also, it is important to ensure a short time interval between conducting the index test and performing the reference standard test. Unlike chronic diseases, the diagnosis of acute infectious conditions, such as appendicitis, needs to occur relatively rapidly owing to the urgency of the situation (Whiting et al., 2011).

The flow and timing of the index (US) and comparator (CT) tests was taken into consideration in the current systematic review. The interval between the diagnostic examinations should be short for patients undergoing both US and CT. This is because a greater degree of bias is likely to apply to a longer interval. A greater interval might also
increase the precision of accuracy of the comparator in the event of rapid deterioration of the disease as the signs would be more obvious in such a situation and thus expedite easier identification. Finally, all participants should be included in the flow and timing analysis as the patients who were recruited in the study should be the same ones included in the $2 \times 2$ table of results in order to reduce any potential bias (Macaskill et al., 2010).

The second part of QUADAS-2 assesses concerns about the applicability of patient selection, and that of the index test and reference standard used. According to Reitsma et al. (2009), the demographic features or degree of severity of the condition in the patients selected for inclusion in the study should be similar to that of targeted patients in a real-life clinical scenario to enhance applicability. Differences in the technology or application used, or the interpretation of the index test can impact on the estimation of its diagnostic accuracy. Thus, the technology, application and interpretation used for the index test should be the same as that used for the standard reference test specified in the review question (Stengel et al., 2005). Finally, a low level of applicability can be attributed to the reference standard test if it is unable to demonstrate suitable ability in terms of accurately classifying the targeted condition specified in the review question (Whiting et al., 2011).

**Data Extraction Process**

Before extracting data, a data extraction form was developed in this review. This form was also piloted by 2 studies that were included in this review. The form extracted data regarding the study characteristics (e.g. study design, author, and publication year), population characteristics (e.g. age, gender, and sample size), index test (US) and comparator (CT) characteristics (e.g. machine type, technique, and sonographer experience), and reference standard characteristics (histopathology and surgery).

**Data Synthesis**

At the beginning, the search process were explained and its findings were presented in a flowchart. Then, the characteristics of the included studies were summarised in tables. Finally, the findings of included studies were summarised narratively, and weighted averages of each measure (i.e. sensitivity, specificity, positive predictive value, and negative predictive value) were calculated. The weighted average was used because it is an accurate statistical technique for combining data from studies with different sample sizes through giving more weight findings of studies with larger samples (McDaniel, 2011; Tacconelli, 2010). Further, calculating the weighted average enables the reviewer to easily compare the diagnostic accuracy of US and CT, thereby, the research question can be answered.

**Results**

**Identification of the Studies**

The search process commenced on 8 July 2017 and ended on 15 July 2017. Eighty-three studies were identified following a search of the pre-specified electronic databases. Seventeen studies were immediately excluded owing to duplication. The titles and abstracts
of the remaining 66 studies were screened and 58 studies were excluded. Fifteen of the studies were reviews articles, paediatric populations were included in 13 of them, 22 studies did not meet either the US or CT criteria, five studies did not meet the outcome criteria (diagnostic accuracy), and alternative imaging modalities, such as MRI, were used in three studies.

The full text of the remaining eight studies was perused, and three further studies were excluded as they did not meet the inclusion criteria. Sufficient information about the accuracy of the comparator (CT) was not provided in the study by Aranda-Narváez et al. (2013). Paediatric populations were included in the study by Mallin et al. (2015). Van Randen et al. (2010) had published another study in 2011 and this had been included in the systematic review. Thus, it was determined that the inclusion of both studies might introduce bias. As the same data set used in both. A further study was identified following forwards and backwards reference list checking. Thus, six studies were finally included. The study selection process is outlined in Figure 1.
Records identified through a database search and Google Scholar ($n = 83$)

Records remaining after the removal of duplicated studies ($n = 66$)

Records remaining after screening the titles and abstracts ($n = 8$)

Records were excluded for various reasons ($n = 58$):
- Population ($n = 13$)
- Index test ($n = 18$)
- Comparator ($n = 4$)
- Outcome ($n = 5$)
- Irrelevant modality ($n = 3$)
- Reviews ($n = 15$)

Records remaining after perusal of the full-text citations ($n = 5$)

Full-text citations excluded ($n = 3$)
- Population ($n = 1$)
- Comparator ($n = 1$)
- Duplicated Authors ($n = 1$)

Citations included in the quantitative synthesis of reports ($n = 6$)

Citation included using a reference list check ($n = 1$)

**Figure 1:** A flow diagram of the search process applied in the study
Results of the Quality Assessment of the Included Studies

QUADAS-2 was used to assess the risk of bias in the included studies and concerns about their applicability in terms of the four domain of patient selection, the index test, the reference standard, and the flow and timing of the tests. The scores obtained for each domain, as per QUADAS-2, for each of the studies are shown in Table 2.

Table 2: The results of the Quality Assessment of Diagnostic Accuracy Studies-2

<table>
<thead>
<tr>
<th>References</th>
<th>Risk of bias</th>
<th>Applicability concerns</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Patient</td>
<td>Index test</td>
</tr>
<tr>
<td>Reich et al., 2011</td>
<td>☹</td>
<td>☹</td>
</tr>
<tr>
<td>Van Randen et al., 2011</td>
<td>☹</td>
<td>☹</td>
</tr>
<tr>
<td>Elghany and Ali, 2011</td>
<td>☹</td>
<td>☹</td>
</tr>
<tr>
<td>Ozkan et al., 2014</td>
<td>☹</td>
<td>☹</td>
</tr>
<tr>
<td>Muzaffar and Bhatti, 2014</td>
<td>☹</td>
<td>☹</td>
</tr>
<tr>
<td>Uzunosmanoğlu et al., 2017</td>
<td>☹</td>
<td>☹</td>
</tr>
</tbody>
</table>

☹ = low risk, ☹ = high risk, ? = unclear

The proportion of studies that scored low, high and unclear in each domain of the risk-of-bias and applicability assessments are shown in Figures 2 and 3, respectively.

Figure 2: The proportion of studies that scored low, high and unclear in each domain of the risk-of-bias assessment.
Figure 3: The proportion of studies that scored low, high and unclear in each domain of the applicability assessment

Moderate results were obtained following the quality assessment of the studies with regard to risk bias and applicability using QUADAS-2. The risk of bias in terms of patient selection was high overall since only two of the six included studies (Elghany and Ali, 2011; Van Randen et al., 2011) used consecutive allocation or randomisation when selecting the patient sample, and avoided inappropriate exclusions.

A low risk of bias in terms of how the index test was performed and how its results were interpreted was attributed to four studies; Elghany and Ali, 2011; Reich et al., 2011; Van Randen et al., 2011 and Uzunosmanoğlu et al., 2017.

A low risk of bias was found in relation to the reference standard since a known and accurate reference standard was used in all the studies, and the results of the reference standard test were interpreted without prior knowledge of the index test results. Lastly, the risk of bias with respect to the flow and timing of the tests was low for four of the studies (Elghany and Ali, 2011; Ozkan et al., 2014; Reich et al., 2011; and Uzunosmanoğlu et al., 2017) because these were similar for all the patients, and an appropriate interval was observed between performing the index and reference standard tests.

The applicability of the studies was considered to be moderate. The applicability of patient selection was high in four of the studies (Elghany and Ali, 2011; Ozkan et al., 2014; Uzunosmanoğlu et al., 2017; and Van Randen et al., 2011) because suitable patients with appropriate characteristics were included as per the requirements outlined in the research question. The index test characteristics and interpretation of the test results were performed according to necessary requirements in this review in syntax error two of the studies.
(Elghany and Ali, 2011 and Van Randen et al., 2011), indicative of high applicability. The latter was also found for the reference standard in all the studies as it was capable of accurately identifying the targeted medical condition in a way that met the requirements posed in the research question.

**Characteristics of the Included Studies**

Of the six included studies, three were published in 2011, a further two were published in 2014 and the remaining one was published in 2017, see table 3. Two studies were conducted in Turkey and the remaining four (one each) were performed in the USA and Israel, the Netherlands, Saudi Arabia and Egypt. Two were a consecutive case series, two were retrospective studies, one was prospective, and the study design for the last one was not described. A total of 857 participants were recruited in the six included studies. The average sample size was 143 subjects, ranging from 60–284 participants per individual study. The age of the participants was ≥ 18 years in four of the studies, while both adolescent patients and adults were included in the remaining two studies. Specifically, patients aged ≥ 16 years were included in the study by Elghany and Ali (2011) and those aged ≥ 12 years were the subjects in the research conducted by Muzaffar and Bhatti (2014). However, the majority of the participants were adults in both these studies.

**Table 3:** The characteristics of the study design and the populations of the included studies

<table>
<thead>
<tr>
<th>References</th>
<th>Study Design</th>
<th>Origin</th>
<th>Sample Size</th>
<th>Age Mean (In Years)</th>
<th>Age Range (In Years)</th>
<th>Females (%)</th>
<th>US + CT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reich et al., 2011</td>
<td>Retrospective study</td>
<td>USA and Israel</td>
<td>276</td>
<td>30 (US)</td>
<td>18-99</td>
<td>54</td>
<td>No</td>
</tr>
<tr>
<td>Van Randen et al., 2011</td>
<td>A consecutive case series</td>
<td>The Netherlands</td>
<td>284</td>
<td>47</td>
<td>19–94</td>
<td>43</td>
<td>Yes</td>
</tr>
<tr>
<td>Elghany and Ali, 2011</td>
<td>A consecutive case series</td>
<td>Egypt</td>
<td>63</td>
<td>38</td>
<td>16-81</td>
<td>56</td>
<td>Yes</td>
</tr>
<tr>
<td>Ozkan et al., 2014</td>
<td>Retrospective study</td>
<td>Turkey</td>
<td>74</td>
<td>36</td>
<td>18–54</td>
<td>30</td>
<td>Yes</td>
</tr>
<tr>
<td>Muzaffar and Bhatti, 2014</td>
<td>Not stated</td>
<td>Saudi Arabia</td>
<td>100</td>
<td>25</td>
<td>12-53</td>
<td>38</td>
<td>No</td>
</tr>
<tr>
<td>Uzunosmanoğlu et al., 2017</td>
<td>Prospective study</td>
<td>Turkey</td>
<td>60</td>
<td>30</td>
<td>18–65</td>
<td>55</td>
<td>Yes</td>
</tr>
</tbody>
</table>

CT: computed tomography, US: ultrasonography

Graded compression was the US technique used in three of the included studies, general abdominal US was performed in one study and the type of US technique used was not identified in the remaining studies. Linear and convex transducers were used in two studies,
linear transducers only were utilised in two, and the type of transducer was not described in the remaining two studies. While the experience and expertise of the sonographer was reportedly high in one study, it ranged from low to high in the other three studies, and was not documented in the remaining two studies. Contrast-enhanced CT was employed in four studies, and both contrast-enhanced and unenhanced CT was used in the other two studies. Type of US and CT machines were identified only in two studies. The reference standard was histopathology for four of the studies, and both histopathology and surgery for the remaining two studies (Table 4).

Table 4: The characteristics of the index tests, comparators and reference standards of the included studies

<table>
<thead>
<tr>
<th>References</th>
<th>US Technique</th>
<th>Transducer</th>
<th>Sonographer’s Experience</th>
<th>US Machine</th>
<th>CT Technique</th>
<th>CT Machine</th>
<th>Reference Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reich et al., 2011</td>
<td>Graded compression</td>
<td>Linear</td>
<td>High</td>
<td>Phillips HDT 5000</td>
<td>Contrast-enhanced</td>
<td>SOMATON Sensation or Definition, Siemens</td>
<td>Histopathology</td>
</tr>
<tr>
<td>Van Randen et al., 2011</td>
<td>General abdominal</td>
<td>Linear and convex</td>
<td>Low and high</td>
<td>Not stated</td>
<td>Contrast-enhanced and unenhanced</td>
<td>Not stated</td>
<td>Histopathology</td>
</tr>
<tr>
<td>Elghany and Ali, 2011</td>
<td>Graded compression</td>
<td>Linear and convex</td>
<td>Low and high</td>
<td>Power Vision 8000 Toshiba</td>
<td>Contrast-enhanced</td>
<td>Somatom Plus Volume Zoom Siemens</td>
<td>Surgery and Histopathology</td>
</tr>
<tr>
<td>Ozkan et al., 2014</td>
<td>Not stated</td>
<td>Not stated</td>
<td>Low and high</td>
<td>Not stated</td>
<td>Contrast-enhanced</td>
<td>Not stated</td>
<td>Surgery and Histopathology</td>
</tr>
<tr>
<td>Muzaffar and Bhatti, 2014</td>
<td>Graded compression</td>
<td>Linear</td>
<td>Not stated</td>
<td>Not stated</td>
<td>Contrast-enhanced and unenhanced</td>
<td>Not stated</td>
<td>Histopathology</td>
</tr>
<tr>
<td>Uzunosmanoğlu et al., 2017</td>
<td>Not stated</td>
<td>Not stated</td>
<td>Not stated</td>
<td>Not stated</td>
<td>Contrast-enhanced</td>
<td>Not stated</td>
<td>Histopathology</td>
</tr>
</tbody>
</table>

CT: computed tomography, US: ultrasonography

Finding of the Included Studies

A comparison was made of the sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) obtained for the diagnostic accuracy of US and CT (Table 5). The sensitivity of US and CT in the included studies was 69–93% and 84–100%,
respectively. The specificity of US and CT was not determined in the study by Reich et al. (2011). The specificity of US and CT in the remaining studies ranged from 47–96% and from 63–100%, respectively. The PPV ranged between 82–100% for US and from 88–100% for CT. The NPV for US and CT was 27–81% and 26–100%, respectively. In order to compare the overall diagnostic accuracy of US and CT, the weighted average of each statistical measure (i.e., sensitivity, specificity, PPV and NPV) was calculated (Table 6). The weighted average results were discussed in the discussion section.
**Table 5:** Findings of included studies

<table>
<thead>
<tr>
<th>References</th>
<th>Modality</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reich et al., 2011</td>
<td>US</td>
<td>69.0</td>
<td>not stated</td>
<td>94.5</td>
<td>not stated</td>
</tr>
<tr>
<td></td>
<td>CT</td>
<td>100.0</td>
<td>not stated</td>
<td>100.0</td>
<td>not stated</td>
</tr>
<tr>
<td>Van Randen et al., 2011</td>
<td>US</td>
<td>76.0</td>
<td>95.0</td>
<td>not stated</td>
<td>not stated</td>
</tr>
<tr>
<td></td>
<td>CT</td>
<td>94.0</td>
<td>95.0</td>
<td>not stated</td>
<td>not stated</td>
</tr>
<tr>
<td>Elghany and Ali, 2011</td>
<td>US</td>
<td>78.3</td>
<td>96.3</td>
<td>100.0</td>
<td>81.0</td>
</tr>
<tr>
<td></td>
<td>CT</td>
<td>97.3</td>
<td>96.3</td>
<td>97.3</td>
<td>26.3</td>
</tr>
<tr>
<td>Ozkan et al., 2014</td>
<td>US</td>
<td>71.2</td>
<td>47.0</td>
<td>82.2</td>
<td>31.8</td>
</tr>
<tr>
<td></td>
<td>CT</td>
<td>97.2</td>
<td>62.5</td>
<td>92.1</td>
<td>83.3</td>
</tr>
<tr>
<td>Muzaffar and Bhatti, 2014</td>
<td>US</td>
<td>75.0</td>
<td>60.0</td>
<td>92.3</td>
<td>27.3</td>
</tr>
<tr>
<td></td>
<td>CT</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Uzunosmanoğlu et al., 2017</td>
<td>US</td>
<td>93.0</td>
<td>57.0</td>
<td>87.0</td>
<td>72.0</td>
</tr>
<tr>
<td></td>
<td>CT</td>
<td>84.0</td>
<td>64.0</td>
<td>88.0</td>
<td>56.0</td>
</tr>
</tbody>
</table>

CT: computed tomography, US: ultrasonography, NPV: negative predictive value, PPV: positive predictive value

*: All the figures are reported as percentages (%).

**Table 6:** The weighted averages of each statistical measure for ultrasonography and computed tomography.

<table>
<thead>
<tr>
<th>Diagnostic Modalities</th>
<th>Weighted Averages</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sensitivity</td>
</tr>
<tr>
<td><strong>US</strong></td>
<td>75.15</td>
</tr>
<tr>
<td><strong>CT</strong></td>
<td>94.77</td>
</tr>
</tbody>
</table>

CT: computed tomography, NPV: negative predictive value, PPV: positive predictive value, US: ultrasonography

*: All the figures are reported as percentages (%).
Discussion

CT was found to be more sensitive than US in diagnosing acute appendicitis in all the studies in this review, with the exception of the research conducted by Uzunosmanoğlu et al. (2017). However, this was the smallest sample size used in all the studies and the participants were not consecutively recruited. Details were also not provided on the experience of the sonographer, and the type of transducer and US technique used. The specificity of CT and US was comparable in two of the studies (Reich et al., 2011; Van Randen et al., 2011), while that of CT was superior in the remaining three (Muzaffar and Bhatti, 2014; Ozkan et al., 2014; Uzunosmanoğlu et al., 2017). The PPV was higher using CT than US in three of the studies (Muzaffar and Bhatti, 2014; Ozkan et al., 2014; Reich et al., 2011), but was virtually the same in two (Elghany and Ali, 2011; Uzunosmanoğlu et al., 2017). The NPV was higher using US than CT in two studies (Elghany and Ali, 2011; Uzunosmanoğlu et al., 2017), but by contrast was comparably lower in another two studies (Muzaffar and Bhatti, 2014; Ozkan et al., 2014).

The weighted average for sensitivity for the diagnostic modalities was substantially different (75% for US versus 95% for CT). This indicates lower diagnostic accuracy when identifying appendicitis in adult patients using US rather than CT. Specifically, a quarter of adult participants with appendicitis could be misdiagnosed using US, compared with 5% of patients using CT. This could reduce confidence in reliance on the US results only, leading to increased demand for the use of CT in adult patients. The diagnostic accuracy of US, unlike that of CT, can be affected by various factors, including the type of US technique used and the experience of the sonographer (Terasawa et al., 2004). However, from point of a positive view that the US was accurately diagnosed 75% of patients, thus it can help in reducing ionising radiation and only 25% of the patients can be referred to CT scan.

Minor variations, of between 3% and 6%, were observed for US and CT, respectively, in the weighted averages respectively obtained for specificity (82% versus 87%), PPV (92% versus 95%) and NPV (55% versus 61%). This suggests that US and CT have a similar ability to correctly identify patients without appendicitis. The current study finding supports those of Poortman et al. (2009) and Lamèris et al. (2009), who recommended that US should be used as a first line technique for diagnosing adults with suspected acute appendicitis and that CT should only be used in cases when US diagnosis is in doubt.

In general, the findings of this review are comparable with findings of two previous systematic reviews carried out by Terasawa et al. (2004) and Weston et al. (2005). In other words, this review and those reviews found that CT was more accurate than US in diagnosing acute appendicitis in adults. To be more precise, Terasawa et al. (2004) found that the overall sensitivity of US was lower than that of CT (86% versus 94%), with corresponding figures for specificity of 81% versus 95%. In the second review (Weston et al., 2005), the overall sensitivity of US was lower than that of CT (88% versus 96%, respectively), whereas the overall specificity achieved using US and CT was almost comparable (92% versus 94%,
respectively). The findings for PPV and NPV were not documented in these reviews (Terasawa et al., 2004; Weston et al., 2005).

The slight differences in reported sensitivity and specificity between the current review and the two previous reviews could be attributed to the fact that only surgery was used as a reference standard for all studies included in this review, while follow-up was the case for the previous reviews when reporting on negative results. According to Giljaca et al. (2017) and Weston et al. (2005), the sensitivity and specificity of US and CT may be overestimated by using the follow-up as a reference standard. This overestimation may be attributed to the fact that the appendicitis becomes more complicated over time and, thereby, it will be easier to be diagnosed.

There were several limitations to the current systematic review. Firstly, the findings cannot be generalised to paediatric populations because the focus was on adult participants. Secondly, the search was only applicable to studies that were published from 2010 onwards. This reduced the number of studies that could potentially be included. However, this period was chosen to ensure the inclusion of recent studies in which the most technologically advanced diagnostic techniques, such as US and CT, were utilised. This was a research project that would contribute to the acquisition of an academic degree. Thus, the timeline for the project was also restricted (to three months). Finally, the results of the included studies could not be combined by calculating overall sensitivity and specificity, as has been performed in other reviews, because most of the studies included in the current review did not report sufficient information with respect to true positive, false positive, true negative and false negative data of diagnosed patients (The two-by-two table). However, the weighted average for the statistical measures were calculated.

The findings in this review can be translated to practical and research recommendations. In respect to the practical recommendation, healthcare settings should use US as a first line technique and use CT only for diagnosing the cases that it was difficult to diagnose them using the US. This is in line with the general recommendation in the literature which is avoiding the routine use of CT because its high ionising radiation dose (Brenner and Hall, 2007), and contrast medium consequences such as allergy and longer stay in the emergency department (Berg et al., 2006). With reference to research recommendation, researchers should take into consideration the methodological quality of their studies. Specifically, it was demonstrated in this study that the patient selection process was flawed in the majority of the included studies (Table 2), this may led to over- or underestimation of the diagnostic accuracy of US and CT (Weston et al., 2005). Hence, future studies should use random sample or consecutive sampling technique for enrolling participants, and they should avoid case-control design and excluding participants inappropriately.
Conclusion

A systematic review of the literature was conducted to determine and compare the accuracy of US and CT in diagnosing appendicitis in adults. Generally, CT was superior to US in terms of sensitivity. However, US and CT were comparable in terms of specificity, PPV and NPV, with a difference ranging from 3–6%. These results do not support the routine use of CT for diagnosing adults and adolescents with suspected acute appendicitis. Therefore, it can be concluded that CT could be used as a complimentary technique for US.

References


Appendices

Appendix 1. Search Strategy using Medline and Embase databases

Database: Ovid MEDLINE(R) <1996 to July Week 1 2017>

Search Strategy:

-------------------------------------------------------------------------------
1 exp "Sensitivity and Specificity"/ (477510)
2 Diagnostic accuracy.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] (25519)
3 exp Adolescent/ (1117130)
4 exp Adult/ (4211123)
5 exp Ultrasonography/ (289279)
6 ultraso*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] (291692)
7 exp Tomography, X-Ray Computed/ (291794)
8 “computed Tomography”.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] (149496)
9 exp Appendicitis/ (7776)
10 “appendicitis”.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] (9580)
11 1 or 2 (488862)
12 5 or 6 (380714)
13 7 or 8 (344885)
14 9 or 10 (9580)
15 3 and 4 and 11 and 12 and 13 and 14 (56)
16 limit 15 to yr="2010 -Current" (24)

Database: Embase <1996 to 2017 Week 29>

Search Strategy:

1 exp "Sensitivity and Specificity"/ (272714)
2 diagnostic accuracy.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word] (208974)
3 exp Adolescent/ (952114)
4 exp Adult/ (4936306)
5 exp Ultrasonography/ (527710)
Appendix 2. Excluded Studies after reading titles and abstracts

<table>
<thead>
<tr>
<th>Study</th>
<th>Reason of Exclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>Result</td>
</tr>
<tr>
<td>--------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Shrestha et al., 2015, Mallin et al., 2015, Lahaye et al., 2015, Yoshimi, 2011</td>
<td>They did not meet comparator criterion</td>
</tr>
<tr>
<td>Stephanie et al., 2014, Drake et al., 2012, LeBedis et al., 2015, Park et al., 2014, O’Malley et al., 2016,</td>
<td>Outcome condition</td>
</tr>
<tr>
<td>Imler et al., 2017, Dillman et al., 2016, Leeuwenburgh et al., 2014</td>
<td>Case studies and articles using irrelevant imaging modality such as MRI</td>
</tr>
<tr>
<td>Wong et al., 2016, Srinivasan et al., 2015, Saito et al., 2013, Aspelund et al., 2014, Pacharn et al., 2010, Aranda-Narváez et al., 2013, Mittal et al., 2013, Abo et al., 2011, Westerland et al., 2016, Karabulut et al., 2014, Krishnamoorthi et al., 2011, Elikashvili et al., 2014, Bachur et al., 2012, Nah et al., 2017, Lee et al., 2016, Park et al., 2013</td>
<td>They involved paediatric population</td>
</tr>
</tbody>
</table>

**Appendix 3. QUADAS-2 checklist**

**Domain 1: Patient selection**

**A. Risk of bias**

Describe methods of patient selection:

- **Was a consecutive or random sample of patients enrolled?** Yes/No/Unclear
- **Was a case-control design avoided?** Yes/No/Unclear
**Domain 1: Study Selection**

<table>
<thead>
<tr>
<th>Question</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Did the study avoid inappropriate exclusions?</td>
<td>Yes/No/Unclear</td>
</tr>
<tr>
<td>Could the selection of patients have introduced bias?</td>
<td>RISK: LOW/HIGH/UNCLEAR</td>
</tr>
</tbody>
</table>

**B. Concerns regarding applicability**

Describe included patients (prior testing, presentation, intended use of index test and setting):

<table>
<thead>
<tr>
<th>Question</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is there concern that the included patients do not match the review question?</td>
<td>CONCERN: LOW/HIGH/UNCLEAR</td>
</tr>
</tbody>
</table>

**Domain 2: Index test(s) (if more than 1 index test was used, please complete for each test)**

**A. Risk of bias**

Describe the index test and how it was conducted and interpreted:

<table>
<thead>
<tr>
<th>Question</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Were the index test results interpreted without knowledge of the results of the reference standard?</td>
<td>Yes/No/Unclear</td>
</tr>
<tr>
<td>If a threshold was used, was it pre-specified?</td>
<td>Yes/No/Unclear</td>
</tr>
<tr>
<td>Could the conduct or interpretation of the index test have introduced bias?</td>
<td>RISK: LOW/HIGH/UNCLEAR</td>
</tr>
</tbody>
</table>

**B. Concerns regarding applicability**

Is there concern that the index test, its conduct, or interpretation differ from the review question? | CONCERN: LOW/HIGH/UNCLEAR |

**Domain 3: Reference standard**

**A. Risk of bias**

Describe the reference standard and how it was conducted and interpreted:

<table>
<thead>
<tr>
<th>Question</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is the reference standard likely to correctly classify the target condition?</td>
<td>Yes/No/Unclear</td>
</tr>
<tr>
<td>Were the reference standard results interpreted without knowledge of the results of the index test?</td>
<td>Yes/No/Unclear</td>
</tr>
<tr>
<td>Could the reference standard, its conduct, or its interpretation have introduced bias?</td>
<td>RISK: LOW/HIGH/UNCLEAR</td>
</tr>
</tbody>
</table>

**B. Concerns regarding applicability**

Is there concern that the target condition as defined by the reference standard does not match the review question? | CONCERN: LOW/HIGH/UNCLEAR |

**Domain 4: Flow and timing**

**A. Risk of bias**
Describe any patients who did not receive the index test(s) and/or reference standard or who were excluded from the 2x2 table (refer to flow diagram):

Describe the time interval and any interventions between index test(s) and reference standard:

- Was there an appropriate interval between index test(s) and reference standard? Yes/No/Unclear
- Did all patients receive a reference standard? Yes/No/Unclear
- Did patients receive the same reference standard? Yes/No/Unclear
- Were all patients included in the analysis? Yes/No/Unclear

Could the patient flow have introduced bias? RISK: LOW/HIGH/UNCLEAR
### Appendix 4. Quality assessment results of included studies

<table>
<thead>
<tr>
<th>Question</th>
<th>Study NO.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reich et al., 2011</td>
<td>No</td>
<td>Yes</td>
<td>Uncle ar</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Van Randen et al., 2011</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Elghany and Ali, 2011</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ozkan et al., 2014</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Uncle ar</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Muzaffar and Bhatti, 2014</td>
<td>Uncle ar</td>
<td>Yes</td>
<td>Uncle ar</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Uzunosmanoğlu et al., 2017</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>