Introduction

Cholangiocarcinoma (CC) is a malignant tumour arising from epithelial lining of the biliary system that accounts 3% of all gastrointestinal tumours. According to its location, it can be divided into intrahepatic CC (ICC), which accounts for 20–25%, hilar CC (50–60%), and distal extrahepatic CC (20–25%) (1,2).

It is reported that the incidence rate of all forms of CC is demonstrating an increasing trend (3). However, coding misclassification of Klatskin tumour as ICC may have resulted in a skewed incidence rate by overestimating ICC by 13% and underestimating extrahepatic CC by 15% (3,4). In 2011, the Institute of Medicine (US) revised the 21-year-old definition of clinical practice guidelines (CPGs) (Institute of Medicine, 1990) as follows, “Clinical practice guidelines are statements that include recommendations...
intended to optimize patient care that are informed by a systematic review of the evidence and an assessment of the benefits and harms of alternative care options” (5).

The Appraisal of Guidelines for Research and Evaluation (AGREE) II instrument is the latest of more than 40 tools for the appraisal of CPGs (6,7). The AGREE instrument and its further refinements is the only CPG appraisal tool that has been developed and validated internationally, formally endorsed by several organizations including the WHO Advisory Committee on Health Research, and used by many guideline development groups (8,9). Detailed information is available on the AGREE web site (www.agreetrust.org).

The aim of the present study was to evaluate the quality of current CC guidelines, with a primary focus on resection, using the AGREE II instrument. The present study focused on the methodological analysis and did not analyse the recommended practices.

Methods

Study selection and data review extraction

A systematic review of the literature in Cochrane, PubMed, Embase, and Google Scholar (including studies of the last 20 years) was conducted to identify guidelines using Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) criteria (10). References of retrieved articles were also searched manually for further guidelines. After independent evaluation of the CPGs by PG and RS, the following data were extracted: country of origin, year of publication, development and/or revision committee, evaluation measures, and funding sources. Evidenced-based CPGs in the English language pertaining to the resection of CC were included (Figure 1, Tables 1 and 2).

Appraisal of guidelines

The AGREE II tool comprises 23 items divided into six domains: scope and purpose, stakeholder involvement, rigor of development, clarity of presentation, applicability, and editorial independence. For further details regarding the criteria used to describe and evaluate the 6 domains and the 23 consisted items, please see Figure S1: AGREE reporting checklist. After undergoing online training (www.agreetrust.org) to ensure appraisal standardisation, four appraisers (PG, AA, RS, KR), as recommended by the AGREE
II consortium, evaluated the guidelines, independently using the AGREE II tool (September 2013 version). As per the AGREE II manual, discrepancies of more than 2 standard deviations (SDs) were resolved through dialogue. Authors had the ability to change their entry after group discussion. Domain scores were calculated with the following formula:

\[
\frac{(\text{Obtained score} - \text{Minimum possible score})}{(\text{Maximum possible score} - \text{Minimum possible score})} \times 100\%
\]

**Table 1** Characteristics of the studies

<table>
<thead>
<tr>
<th>Guideline, year</th>
<th>Developers</th>
<th>Content</th>
<th>Guidelines review</th>
<th>Scoring system</th>
<th>Evidence based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asian-Pacific Consensus, 2012</td>
<td>Asian-Pacific consensus</td>
<td>Perihilar CC, E, D&amp;T</td>
<td>A modified Delphi process</td>
<td>Levels of evidence, Systematic literature review, and consensus panel</td>
<td></td>
</tr>
<tr>
<td>Brazilian, 2015</td>
<td>Brazilian Gastrointestinal Tumor Group</td>
<td>ICC, Hilar CC, D&amp;T, FU</td>
<td>NR</td>
<td>Level of evidence, GRADE, Systematic literature review, expert panel</td>
<td></td>
</tr>
<tr>
<td>BSG, 2012</td>
<td>BSG</td>
<td>CC, E, D&amp;T, FU</td>
<td>BSG peer group review</td>
<td>Oxford + GRADE, Systematic literature review</td>
<td></td>
</tr>
<tr>
<td>Chinese Consensus, 2014</td>
<td>A consensus of surgical specialists from China</td>
<td>CC, E, D&amp;T, FU</td>
<td>NR</td>
<td>NR, NR</td>
<td></td>
</tr>
<tr>
<td>EASL, 2014</td>
<td>European association for the study of the liver</td>
<td>ICC, E, D&amp;T</td>
<td>EASL reviewers</td>
<td>Oxford + GRADE, Systematic literature review and expert panel</td>
<td></td>
</tr>
<tr>
<td>ESMO, 2016</td>
<td>ESMO Guidelines Committee</td>
<td>Biliary cancer, E, D&amp;T, FU</td>
<td>NR</td>
<td>Levels of evidence, GRADE, ESMO-Guidelines-Mетодology Expert panel</td>
<td></td>
</tr>
<tr>
<td>AHPBA Consensus for Hilar CC, 2014</td>
<td>AHPBA Committee</td>
<td>Hilar CC, E, D&amp;T</td>
<td>NR</td>
<td>NR, Expert panel</td>
<td></td>
</tr>
<tr>
<td>AHPBA Consensus for ICC, 2014</td>
<td>AHPBA Committee</td>
<td>ICC, E, D&amp;T</td>
<td>NR</td>
<td>NR, Expert panel</td>
<td></td>
</tr>
<tr>
<td>Italian Guidelines, 2010</td>
<td>AIGO, AIOM, AIRO</td>
<td>CC, E, D + T</td>
<td>NR</td>
<td>Expert panel</td>
<td></td>
</tr>
<tr>
<td>NCCN, 2015</td>
<td>National Comprehensive Cancer Network</td>
<td>CC, GBC, HCC</td>
<td>Members of NCCN</td>
<td>Oxford + GRADE, Systematic literature review, expert panel</td>
<td></td>
</tr>
<tr>
<td>Palliative Surgery for Hilar CC, 2008</td>
<td>University of Mainz, Germany</td>
<td>CC, D&amp;T</td>
<td>NR</td>
<td>NR, Expert panel</td>
<td></td>
</tr>
<tr>
<td>SEOM, 2015</td>
<td>Sociedad Española de Oncología Medica</td>
<td>CC, GBC, E, D&amp;T</td>
<td>NR</td>
<td>NR, Expert panel</td>
<td></td>
</tr>
</tbody>
</table>

AHPBA, American Hepato-Pancreato-Biliary Association; SEOM, Sociedad Española de Oncología Medica; ESMO, European Society of Medical Oncology; EASL, European Association for the Study of the Liver; NCCN, National Comprehensive Cancer Network; AIGO, Italian Association of Hospital Gastroenterology; AIOM, Italian Association of Medical Oncology; AIRO, Association of Oncological Radiotherapy; BSG, British Society of Gastroenterology; CC, cholangiocarcinoma; ICC, intrahepatic cholangiocarcinoma; GBC, gallbladder cancer; E, epidemiology; D&T, diagnosis & treatment; FU, follow-up; NR, non-reported.

**Results**

Search results yielded a total of 13 guidelines, eight of which were produced by multi-national organizations (Figure 1, Table 1) (11-23). Overall, the guidelines scored poorly: the median overall score was just 43%, with the highest overall score of 82% given to the British Society of Gastroenterology (BSG) guidelines, followed by 79% for the Asia-Pacific and European Association for the Study of the Liver (EASL) guidelines. Of the 13 guidelines, eight
scored under 50%. Median scores were particularly low in the following domains: II: stakeholder involvement (39%); III: rigor of development (30%); and V: applicability (13%). Domain IV: clarity of presentation and domain I: scope & purpose were the highest scoring at 76% and 65%, respectively (Figure 2).

**Domain I: scope & purpose**

In this domain, questions pertain to the aims and objectives of the guidelines and the target users and population. Generally, the majority of the guidelines performed well, and the median score was 65%. The BSG scored the highest (81%), whilst the SEOM (Spanish Society of
Medical Oncology) scored the lowest at 15% (Figure 3).

**Domain II: stakeholder involvement**

An integral part of the AGREE II scoring checklist is the involvement of the relevant stakeholders in the guideline production process. Overall, the median score was 39% with the Asia-Pacific, Chinese, and BSG guidelines yielding the highest scores, at 65%, 63%, and 61%, respectively. The SEOM guidelines scored the lowest at 8% (Figure 3).

**Domain III: rigor of development**

This was one of the lowest scoring domains. The overall score across all 13 guidelines was 30%, with the Asia-Pacific guidelines scoring the highest (81%) as they clearly laid out the methodology of the development of the guidelines from the evidence found in the literature (Figure 4). The Hilar guidelines had the lowest score at only 12%.

**Domain IV: clarity of presentation**

Domain IV was the highest scoring domain with a median score of 76%. Among all of the guidelines, the Asia-Pacific and Japanese guidelines both scored 90%. The Hilar consensus statement received the lowest score of 44% (Figure 4).

**Domain V: applicability**

Scores in this domain were the lowest of all, with the median score at just 13%. All of the guidelines scored less than 50%, with the highest scoring guideline being the Asia-Pacific guideline at 47% (Figure 5). The two lowest scoring guidelines scored a mere 1% (hilar and ICC-intrahepatic CC guidelines).

**Domain VI: editorial independence**

Scores in domain VI were generally reasonable at 56%. The highest score (92%) was achieved by the European Society of Medical Oncology (ESMO) (Figure 5). One guideline scored 0% (guidelines for palliative surgery of CC).

**Recommendation for use**

None of the 13 guidelines was recommended universally for use without modification. The Asia-Pacific and EASL guidelines scored a definitive ‘Yes’ by three appraisers and ‘Yes, with modification’ by the fourth appraiser. Five of the 13 guidelines (Hilar, ICC, Japanese, Palliative, and SEOM) scored a unanimous ‘No’ by the appraisers (Figure 6).

**Score discrepancies**

Discrepancies in scoring across appraisers were low overall;...
therefore, no further rounds of re-scoring or discussion were required to resolve issues.

**Discussion**

The present study is the first appraisal of the current CC guidelines using the validated AGREE II instrument. Overall, the quality of the guidelines as assessed by the AGREE II evaluation checklist was mediocre at best, with a median total score of only 43%. The BSG (82%), EASL (79%), and Asia-Pacific (79%) guidelines scored the highest overall score. Generally, the guidelines scored poorly in the domains of applicability, rigour of development, and stakeholder involvement, at 13%, 30%, and 39%, respectively. The highest scores were observed in the domains of clarity of presentation and scope and purpose, at...
76% and 65%, respectively.

These findings highlight that careful attention and further developmental work on existing guidelines is required, particularly in the areas of clinical implementation and the involvement of patients/advocacy groups. Another area in which the guidelines are performing particularly poorly is their rigour of development. Often, the guidelines did not stipulate how they arrived at their recommendations. This is likely surprising given both the importance of transparency in guideline development and existence of validated systems for evaluating the scientific literature. One such system is the GRADE system, which has been specifically developed for the evaluation of evidence and classification of recommendations for guideline development (24). Nonetheless, only seven of 13 guidelines used the GRADE system to evaluate the quality of evidence and categorise the strength of recommendations, namely the Brazilian, NCCN, EASL, BSG, and Japanese guidelines (12,15,17,20,22).

Two others, the ESMO and the Asia-Pacific guidelines used alternative systems. The ESMO guidelines used the US Public Health Service grading system, and the Asia-Pacific guidelines categorised the evidence and classified the recommendations using a voting system based on a modification of the Canadian Task Force on the Periodic Health Examination (11,19). Four of the guidelines did not use any system at all (13,14,17,18). Consequently, the median score for rigour of development was only 38%. Guidelines that used a system to evaluate the evidence (i.e., the GRADE system) naturally scored higher, whilst those that did not employ any discernible system, such as the Hilar guidelines, scored poorly (12%).

Another area in which guidelines appear to perform universally poorly is stakeholder involvement, as has been extensively documented in the literature. The median score for this domain was just 39%, with the SEOM guidelines scoring only 8%. In the age of patient-centred care, patient autonomy, and informed consent, this lack of engagement with patients, patient advocacy groups, and the general public is concerning. In many instances, even other professionals involved in the treatment of patients with the condition were not consulted and the entire guideline was written by members of a single specialty. Although such guidelines can produce specialised recommendations from one aspect of care, i.e., surgical or oncological, they may miss other aspects not immediately within the purview of their specialty, which contradicts the ethos of holistic and multi-disciplinary care. Furthermore, very little guidance is provided by the guidelines. Of the 13 presented in this study, only two (Brazilian and ESMO) gave any recommendations for the follow-up and long-term management of these patients.

Another critical element of guidelines for the management of biliary tree pathology lies in the difference between benign and malignant disease and how they should be managed. For example, immunoglobulin G4 cholangiopathy is a multisystem inflammatory disorder that may present with intrahepatic biliary strictures in 51% of cases and proximal extrahepatic ducts in 49%; this disease should always be included in the differential diagnosis of biliary strictures (25). The BSG guidelines alone stressed the importance of differentiating between benign and malignant strictures. Such oversights highlight the limitations of the current guidelines in their scope and depth. In addition, the Japanese guidelines strongly recommend biopsy or cytology before surgery in order to differentiate malignant from benign strictures. Nakayama et al. reported that 10% of suspected and operable CCs were benign strictures (26). The remaining guidelines do not give recommendations regarding preoperative biopsy.
Moreover, staging laparoscopy is a useful tool to avoid unnecessary operations, though only the Asia-Pacific guidelines recommend the use of staging laparoscopy. In terms of ICC, the NCCN guidelines recommend colonoscopy and gastroscopy to rule out metastases from an asymptomatic gastrointestinal tumour. None of the other guidelines recommend it.

Most of the guidelines used the 7th AJCC staging system. It is reported that the main limitations of the AJCC are the definition of resectability and prediction of survival (27). The Blumgart staging system, which is based on the extent of biliary duct involvement by the tumour, the presence or absence of portal involvement, and the presence or absence of lobar atrophy, can define resectability and predict survival more accurately than the AJCC and Bismuth-Corlette staging systems (Table 2) (27). However, none of the guidelines proposed the above staging system.

The guidelines define the following as risk factors for CC: primary sclerosing cholangitis, parasitic infestations, hepatolithiasis, choledochal cysts, pancreatobiliary maljunction (PBM), toxins, and hepatitis B (HBV) and C (HCV) infections. However, only the BSG guidelines recommend the surveillance of patients with primary sclerosing cholangitis. In cases of PBM and choledochal cysts, the Japanese guidelines recommend cholecystectomy and extrahaepatic common bile duct excision to prevent cancer development.

ICC and hepatocellular carcinoma (HCC) share many risk factors such as HBV, HCV, and cirrhosis. Guidelines on HCC recommend surveillance every six months with tumour markers and imaging modalities for high-risk patients (28). However, no such recommendation for surveillance of CC exists in any of the current guidelines.

Currently, there is no solid evidence to support standard lymph node dissection in patients with CC (29). However, NCCN guidelines suggest lymph node dissection to achieve better prognosis.

Expert consensus on ICC recommends that lymph node dissection be a standard part of the surgical management of patients with ICC.

It is reported that ≥7 lymph nodes are sufficient for the prognostic staging of hilar CC (30).

Another area in which guidelines scored particularly poorly was advice and guidance on how to implement the recommendations. This is an issue that has plagued many a guideline across a variety of medical sub-specialities, as well documented in the literature (28). Unfortunately, the CC guidelines are no exception. Very few guidelines documented the human and material resources required for the implementation of the recommendations or gave clear instructions as to how the recommendations could be put into action, leaving the readership with no real sense of direction or where to start.

There are even further shortcomings in the current CC guidelines. Perhaps most striking is the quality of the evidence on which many of the guidelines are based. Unfortunately, there is currently a distinct lack of randomised control trial data for many of the recommendations in place. A large proportion of the key points in many of the guidelines are based on observational data of potentially questionable reliability. If guidelines are to improve further, bold and rigorous studies within the boundaries of ethical consideration are required to further our understanding of the management of CC.

Our study has several limitations, some of which can be attributed to the very nature of the AGREE system. For example, it is debatable whether every domain in the AGREE system should carry the same weight in terms of scores as some of the others. The checklist has been criticised for its assumption that all domains are equally important for determining high-quality guidelines. Other criticisms of the checklist are that although assessors fill in their respective scores independently of each other, there is still the possibility of bias (positive or negative) and a certain level of subjectivity in the scoring. For example, certain guidelines from reputable international bodies may hold a favourable stance in the assessor’s mind before the assessor even begins to undertake the scoring. Conversely, a guideline from a less well established body may not score as highly due to a lack of ‘prestige’. A further limitation of our study is that we only selected guidelines in the English language for reasons of practicality; as such, we may have missed potentially high-quality guidelines presented in other languages.

Of note, The Asia-Pacific guidelines performs particularly well because their guideline is well set out, well written, thorough and easy to read. The evidence they base their recommendation on is no different to that found in many of the other guidelines, however what the Asia-Pacific guideline does well that others do not is the following:

(I) They clearly explain in detail the quality of the evidence they base their recommendations on, the degree of consensus and their methodology as to how they came about making the recommendations.

(II) They also rate their own recommendations in terms of quality and how strongly they recommend...
a particular practice.

(III) The guideline is well presented, thorough in terms of addressing all the relevant aspects of investigation, treatment and outcome.

As a result of the above, the Asia-Pacific consistently scores well throughout most of the Domains.

Recently, Idrees et al. (31) evaluated the impact of the centralisation of care and compliance with NCCN guidelines for resected CCs on long-term survival. It was reported that over time, in the USA, compliance with NCCN guidelines increased. In particular, for the period of 2004–2007, the compliance was 30%, which increased to 46% in 2011–2013. Of note, five-year overall survival was 45% in the patients who received NCCN-compliant surgical management, as compared to 40% in those who did not receive surgical care according to the guidelines. Interestingly, it was reported that the centralisation of care contributed only 8% of the improvement in survival, while compliance with guidelines improved survival by 17% (31).

Conclusions

The quality of the current guidelines for CC is generally poor or based on relatively low-quality evidence. It is imperative that future updated guidelines rely on high-quality trial data and take a multi-disciplinary approach by including patients and advocacy groups in the formulation of recommendations. Furthermore, a clear plan as to how to put the recommendations into practice (in both resource-rich and poor regions of the world) is desperately needed.

Acknowledgments

None.

Footnote

Conflicts of Interest: The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References


**AGREE Reporting Checklist**

**Annex A**

**AGREE Reporting Checklist**

**Objectives**

Describe the clinical problem or health topic.

**Population**

Describe the different options for managing the condition or health issue.

**Target Users**

Describe the different options for managing the condition or health issue.

**Inclusion and Exclusion Criteria**

Provide an explicit statement that all group members have declared whether they have any competing interests.

**Evidence Selection Criteria**

Provide explicit criteria for how evidence was selected for the guideline.

**Applicability**

Describe the strengths and limitations of the external review.

**Methodology for the Updating Procedure**

Describe the methods used to update the guideline, including the current version and the method for aggregating the body of evidence.

**Implementation**

Describe the process for updating the guideline.

**Decision-Making Process**

Provide a description of the decision-making process.

**Reporting均**

Describe the content and structure of the full version of the guideline.

**Annex B**

**AGREE Reporting Checklist**

**Objectives**

Describe the clinical problem or health topic.

**Population**

Describe the different options for managing the condition or health issue.

**Target Users**

Describe the different options for managing the condition or health issue.

**Inclusion and Exclusion Criteria**

Provide an explicit statement that all group members have declared whether they have any competing interests.

**Evidence Selection Criteria**

Provide explicit criteria for how evidence was selected for the guideline.

**Applicability**

Describe the strengths and limitations of the external review.

**Methodology for the Updating Procedure**

Describe the methods used to update the guideline, including the current version and the method for aggregating the body of evidence.

**Implementation**

Describe the process for updating the guideline.

**Decision-Making Process**

Provide a description of the decision-making process.

**Reporting均**

Describe the content and structure of the full version of the guideline.