CAD: How it works, how to use it, performance

Daniele Regge\textsuperscript{a,*}, Steve Halligan\textsuperscript{b,1}

\textsuperscript{a} Institute for Cancer Research and Treatment, Candiolo-Torino, Italy
\textsuperscript{b} University College Hospital London, London, UK

\textbf{A R T I C L E  I N F O}

Article history: Received 17 April 2012 Accepted 21 April 2012

Keywords: Computer Aided Diagnosis (CAD) CT colonography (CTC) Colorectal cancer

\textbf{A B S T R A C T}

Computer-aided diagnosis (CAD) systems are software algorithms designed to assist radiologists (or other practitioners) in solving a diagnostic problem by using a visual prompt (or “CAD mark”) to direct the observer towards potential pathology. CT colonography is a recent arrival to CAD, but could represent one of its most fruitful applications in the future. In contrast to other organs, where a variety of different pathologies are equally represented, significant colorectal pathologies other than polyps and cancer are relatively uncommon. As we shall see, this simplifies the diagnostic task for artificial intelligence developers and also for radiologists who, ultimately, must make the final decision.

This review aims to present the current state-of-the-art for CAD applied to CT colonography. A brief overview of the technical essentials and of the diagnostic performance of CAD in isolation, is followed by an explanation of how CAD is used in day-to-day practice. The last section will deal with the most controversial issues affecting CAD performance in clinical practice, with a focus on the interaction between human and artificial intelligence.

© 2012 Elsevier Ireland Ltd. All rights reserved.

1. Introduction and rationale of CAD assisted CT colonography

Computer-aided diagnosis (CAD) systems are software algorithms designed to assist radiologists (or other practitioners) in solving a diagnostic problem by using a visual prompt (or “CAD mark”) to direct the observer towards potential pathology. Such algorithms may be used to help pinpoint areas of disease (computer-aided-detection, CADe), or help interpreters decide whether abnormal findings are benign or malignant (computer-aided diagnosis, CADx). Regardless of their task, the human interpreter must always make the final diagnostic decision, by deciding whether or not to “believe” the CAD marks.

CAD systems are not new to radiology. Digital platforms were developed in the 1980s with the aim of detecting and characterizing breast nodules in mammograms. More recently, CAD software has been developed to search for lung nodules in CT scans and help define their nature, and to detect prostate and breast cancer at MRI. Despite a huge research effort on the part of Academic Institutions and Industry, few CAD algorithms are used routinely in day-to-day clinical practice. Several reasons underpin this failure to implement CAD in daily practice and some are well understood while others are less obvious. Most obviously, such algorithms may be expensive to purchase and service. Furthermore, depending on the Country in question, over-restrictive licensing stipulations may restrict availability. One intuitive objection to the use of CAD systems is that they generate visual marks that the radiologist must take time to interpret. Interpretation is required because each mark is either true-positive for pathology or false-negative, and the radiologist must decide which it is. A large number of marks will likely prolong exam interpretation time. Also, most marks are false-positive and there is concern that if a disproportionate number of these are misinterpreted by the radiologist as true-positive, then the number of individuals undergoing unnecessary second level, follow-up tests will increase. Such additional exams, which are usually more invasive than first level tests, will increase procedure related costs, likely increase the rate of adverse events, and also precipitate patient anxiety. It is also possible that radiologists may be reluctant to rely on the help of artificial intelligence, which they might distrust.

CT colonography is a recent arrival to CAD, but could represent one of its most fruitful applications in the future. This is because the colorectal cancer model is relatively simple and can be epitomized as a traffic light were green is a negative test, yellow a polyp, and red a cancer. In contrast to other organs, where a variety of different pathologies are equally represented, significant colorectal...
pathologies other than polyps and cancer are relatively uncommon. As we shall see, this simplifies the diagnostic task for artificial intelligence developers and also for radiologists who, ultimately, must make the final decision. This compelling rationale probably explains why the development of CAD solutions for CT colonography has been so fast; to date, over 20 companies are producing and commercializing dedicated CAD software worldwide.

Like detection algorithms in other organs, the main objective of CAD assisted CT colonography is to identify lesions that have been missed by the reader, so called “operator errors” or “perceptual errors”. Such errors account for approximately half to three-quarters of all missed but detectable lesions [1]. Perceptual error has been evaluated extensively in breast imaging, where it is a major cause for concern. Indeed, in contrast to the colon, were the target lesion is mainly benign neoplasia, the large majority of breast lesions are malignant. Perceptual error rate is influenced by reader experience and intrinsic diagnostic threshold, working environment, exam quality, and lesion characteristics and morphology. CAD assisted CT colonography should improve polyp detection by influencing some of these factors. Intuitively, pinpointing lesions that have been overlooked because of an unfavourable location and/or atypical morphology is where most diagnostic improvement by using CAD is to be anticipated. By implication, such gains should be disproportionately beneficial for less experienced readers.

This review aims to present the current state-of-the-art for CAD applied to CT colonography. A brief overview of the technical essentials and of the diagnostic performance of CAD in isolation, is followed by an explanation of how CAD is used in day-to-day practice. The last section will deal with the most controversial issues affecting CAD performance in clinical practice, with a focus on the interaction between human and artificial intelligence.

2. Inside the black box

Radiologists do not need to know how CAD algorithms work in order to employ a system in their reporting. However, a superficial understanding of what happens inside the “black box” does provide some insight as to why things occasionally don’t happen according to plan (e.g. too many CAD candidates, excessive processing times, incomplete segmentation) and as to the type of lesion that might be missed by a specific CAD platform.

Typically, CAD development schemes require three sequential steps. The first step is colonic segmentation, whereby voxels located at the interface between the luminal content and mucosa are isolated and extracted from surrounding CT data (Fig. 1a and b). Segmentation is usually performed by defining local density thresholds, a process that discriminates voxels originating in the colon wall from those of surrounding tissues [2]. While segmentation is usually an automatic process, some CAD systems have tools available to remove structures that have been erroneously segmented with the colon (most frequently small bowel, stomach and lung bases), or to reconnect colonic segments separated by spasm, etc., so that evaluation of the entire colon can be performed (Fig. 1c). Since it is important to detect lesions that are submerged by fluid, segmentation should include electronic cleansing of tagged fluid [3], i.e. residual fluid is eliminated by the software.

The second step in the development of a CAD algorithm is the identification of polyps, or “candidates”. This process is based on the analysis of specific morphological features of the extracted surface, in particular those that indicate a polypoid structure protruding from the luminal surface. There are several ways to approach this; the most efficient is to use shape and curvature indexes. The shape index examines a group of contiguous voxels and assigns to each a score from −1 to +1, according to their morphology at a specific point. Dome-shaped structures, such as polyps, are assigned

Fig. 1. Example of colon segmentation. (a) The software places seeds automatically within the colon lumen. The seeds expand until they reach the bowel wall. All extra data is removed after the interface between lumen and bowel wall has been identified. (b) CT images are then stacked and the surface between slices is generated by interpolation. The colon is then smoothed, artificially coloured and light is projected on its surface to improve rendering. (c) The image includes two small bowel loops (i.e. the yellow and blue transparent structures on the upper left side) that have been segmented with the colon, and can be removed manually.
the highest values. The curvature index describes the degree of local surface “slope”. Low values are assigned to flat surfaces and, alternatively, high values are assigned to steep slopes. Polyps usually have intermediate curvature values. A combination of the two indexes allows discrimination between polyps and folds or dips in the endoluminal surface.

This extraction phase generates a large number of candidate polyps, typically between 50 and 100 for each CT series, most of which are thickened haustral folds or fecal residues. Even though most candidates are easy to dismiss by an experienced radiologist, review of such a large number is a laborious and time-consuming task. So, the third step in CAD development is to reduce the number of polyp candidates presented to the reader by selecting only those that have the highest probability of being true positive polyps. This process is performed by assessing multiple different characteristics for each candidate (e.g. volumetric texture analysis, CT attenuation, random orthogonal shape section, optical flow, etc.). A discrete value is then assigned to each candidate based on the calculated features and a statistical classifier is subsequently applied to select CAD candidates that have the highest probability of being true positives. Ultimately, polyp candidates selected by the classifier are highlighted to the viewer of the CAD interface with a marker, or a different color from that of the surrounding normal mucosa. A list of individual CAD candidates is usually available in addition, for both the prone and supine scans separately (Fig. 2).

While the process described above can be initiated by software code alone, the best CAD systems are “trained” by using large numbers of CT colonography datasets containing endoscopically
validated polyps. Essentially, by using such databases software developers can fine-tune the CAD algorithm by exposing it to real polyps. All of the processes described thus far can be conceptualized a CAD “development”. However, when the CAD software appears to reach adequate diagnostic performance, it is then tested using a CT colonography database composed of cases that the software has not “seen” before. The “stand-alone” performance of a CAD algorithm excludes the radiologist interaction, and merely reports the ability of the software to detect true polyps (sensitivity) and reports the number of false-positive detections per series, which is equivalent to 1-specificity. Since 2002, several researchers have reported the stand-alone sensitivity and specificity of various academic and commercial CAD systems [4,5]. As CAD schemes evolve, performances improve in tandem. Recently, Lawrence [6] reported per-polyp CAD sensitivities in isolation, regardless of histology, of 90% and 96% at 6- and 10-mm-diameter thresholds, respectively, at a mean false-positive rate of 4.7 per series. The diagnostic performance of CAD systems can also be represented by Free-response Receiver Operating Characteristic (FROC) curves [7]. FROC curves combine sensitivity and specificity into a single metric which is then plotted at a range of different diagnostic thresholds. Specifically, sensitivity is plotted against 1-specificity (i.e. the false-positive rate) as the classifier’s threshold value for a true positive finding is lowered progressively (i.e. as the system becomes progressively more sensitive). FROC curves allow different CAD systems to be compared and are used by researchers to select the most appropriate single operating point for the software, which is inevitably a compromise between sensitivity and specificity. Although the stand-alone performance of an algorithm provides some information regarding the diagnostic capabilities of the software, this alone cannot determine the diagnostic performance of CAD in clinical practice. This is because it is necessary to incorporate the human reader and understand how their interaction with artificial intelligence affects diagnostic performance.

3. How to use CAD

A CAD reading paradigm is a description of how interpreting a CTC exam with CAD is achieved. The main difference between different paradigms is in the timing with which CAD candidates are presented to the reader. The performance of different paradigms and their interaction with the human reader for CTC have been assessed recently by researchers, and findings are summarized below.

1) “First reader” CAD. In the first reader paradigm, the CAD algorithm is activated immediately when the reader interprets a case. A list of CAD marks are presented to the interpreter for review. The reader reviews all CAD candidates sequentially and reports those that he/she believes are true lesions, if any. Interpretation is restricted to the CAD marks and because a full, unassisted review of the entire endoluminal surface is not required, reading time is reduced greatly [8]. Shortening interpretation time may be important if large number of cases must be read sequentially, which could be the case in the future if CTC screening is implemented widely. However, in the process, sensitivity and positive predictive value should not be affected negatively, so it is important that the stand-alone detection characteristics of the system are outstanding. Preliminary data regarding use of the first read paradigm in mass screening programmes demonstrate that accuracy of CTC is sustained when compared to unassisted reading, but reading time is shorter [9]. In a symptomatic environment, where disease prevalence is higher, the first-read paradigm could possibly be approved if the stand-alone sensitivity was 100%. If per-polyp sensitivity falls much below this for larger polyps, the first read paradigm should probably not be recommended. This is especially the case because CAD schemes in isolation are less efficient when identifying lesions that do not protrude into the colon lumen, such as infiltrating flat cancers. Therefore clinically relevant lesions could be missed if radiologists do not review the entire CTC dataset [10].

2) “Second-reader” CAD. In second-read mode, the radiologist first performs an initial unassisted read of the entire CTC dataset, using a primary 2D or 3D read, according to individual preference and exactly mimicking day-to-day practice without CAD. After this unassisted read, CAD is activated and any candidates are checked for additional findings missed on the initial inspection. As for the first-read paradigm, readers must decide whether to accept or reject each CAD candidate; however, they should not reverse decisions taken during their unassisted reading. Intuitively, second-read CAD should be most beneficial when adopted by less experienced readers; radiologists relatively unfamiliar with CTC have lower detection rates and their gain when using CAD should be larger than for experienced readers [11]. Baker et al. [12] have shown that, for polyps 6 mm or larger, second-read CAD improves detection of readers with some or no experience in CTC reporting from 81% to 91%. These authors also found that CAD second-read increases detection of intermediate size polyps. Recently, Halligan et al. [13] demonstrated that second-read CAD also benefits the experienced CTC reader, finding that mean per-polyp sensitivity for polyps 6 mm or larger increased on average by 9%. As expected, additional interpretation of CAD candidates extends overall CTC reading time by a few minutes [12,14].

3) “Concurrent” CAD. In the concurrent-reader paradigm, CAD is applied immediately prior to review, like the first-read paradigm. However, the reader does not restrict himself to the CAD marks but also interprets the whole dataset. This paradigm thus combines elements of both the first- and second-read paradigms; CTC interpretation is performed once, with the CAD candidates on the screen from the beginning. Readers may choose to turn CAD marking on and off any time they wish. Halligan et al. [11] have shown that concurrent read increases the per-polyp detection of inexperienced radiologists on average by 9.1% (i.e. from 17% to 26%) while reducing reading time by slightly over 2 min. The benefit they observed was mainly due to increased detection of medium and small-sized polyps. However, the magnitude of benefit was not as pronounced as for second-read CAD, when the two paradigms were compared directly on the same datasets and readers [13]. The reasons underpinning the lower diagnostic performance of concurrent reading are not yet well understood. It is possible that, by focusing their attention on CAD marks, readers reduce their assessment of the remaining portions of colorectal mucosa as opposed to what happens during unassisted reading [14]. In summary, according to available evidence, concurrent reading is less effective than the second-read paradigm and its routine use is not recommended in clinical practice at the present time. The same comment applies to first-read CAD.

4. Limits of knowledge of CAD performance and controversies

CAD improves detection of colorectal lesions by directing the reader to suspicious areas of colorectal mucosa. For both inexperienced readers and experts, the diagnostic performance of CAD is
best when used as the second-reader, even though overall reporting time is increased.

However, because CAD schemes have been tested mostly in retrospective cohorts and in a laboratory environment, little is known regarding its performance in day-to-day clinical practice when used by representative radiologists. An Italian cross-sectional multicentre study has recently shed some new light on the utility of CAD in clinical practice. The study, recruiting 18 experienced readers from 10 Academic centres, concluded that second-read CAD significantly improves detection of polyps between 6 and 9 mm, but observed no benefit for larger lesions [15]. In order to establish firmly if there is a role for CAD in clinical practice, and for screening, further evidence will be required from large-scale, preferably randomized trials, retracing the path of recently published studies evaluating the utility of breast mammography CAD schemes [16].

Little is known regarding how inexperienced readers – i.e. “beginners” and readers with limited experience with CT colonography – interact with CAD in clinical practice. When compared to an unassisted read, second-read CAD increases the false-positive rate of non-experts in 43% of cases (i.e. 3 of 7 readers) [12]. However, it has been shown that training reduces the number of false-positive diagnoses by readers [17]. Therefore, there is compelling evidence that the level of reader experience affects their ultimate performance when using CAD. Exam-related factors, such as the number of CAD prompts [18] and the type of bowel preparation [10], may also influence readers’ false-positive rates. On this point, results are somewhat controversial. For example, a study by Taylor et al. [19] found that a relatively high number of false-positives does not significantly distract readers from their true-positive detections, but does increase interpretation time. Certainly, since reading time is proportional to the number of CAD marks, interpretation time will be prolonged when the number of marks is high.

Since research findings remain inconsistent and controversial, it is generally agreed that radiologists should adopt CAD only after they have been adequately trained in unassisted CT colonography interpretation. However, since the definition of “adequately
trained" is itself still a matter of debate, the suggested timing for CAD adoption is uncertain. At the time of writing, the authors believe that in terms of clinical utility, there is sufficient current evidence to suggest that the additional polyps detected by CAD outweigh any additional false-positive diagnoses. CAD is less likely to be clinically useful in situations where there are overwhelming numbers of false positive prompts, such as occurs in a poorly prepared colon for example.

Finally, readers should be aware that it is possible to erroneously reject true positive CAD prompts. In most studies, the sensitivity of CAD in isolation has been superior to that achieved by radiologists assisted by the same CAD algorithm [5]. It is clear that diminished sensitivity is due to the fact that readers ignore a percentage of true-positive CAD marks [11,20]. The reasons underpinning this apparently illogical behaviour are uncertain. Taylor et al. investigated this phenomenon by comparing two groups of polyps that were correctly detected by CAD; one group where CAD detection resulted in an additional correct detection by readers, and a second group for which the CAD mark did not improve sensitivity [20]. They found that lesions misclassified as false-negatives by readers were significantly larger and irregular. The authors concluded that difficulty with characterization – i.e. due to an irregular and/or flat morphology – was the main determinant why radiologists rejected true positive CAD prompts. Other reasons that may cause readers to reject correctly annotated polyps are: polyps near or on folds, lesions coated or submerged by tagged fluid, a high number of CAD annotations (Fig. 3), and a location near other polyps.

5. Concluding remarks

CAD for CT colonography has progressed a long way over the last decade, and systems with excellent detection characteristics have been developed, validated, and are commercially available. While such systems do not obviate the need for specific training in interpretation of CT colonography, they do provide gains in sensitivity when searching for polyps without a disproportionate drop in specificity. These gains are largely dependent on the expertise of the reader, but a diagnostic benefit of varying degrees seems to be available to readers of all abilities when CAD is used as stipulated by regulatory authorities. The next decade will reveal whether these systems are adopted by users and so become integrated into day-to-day practice, or whether they remain confined to the laboratory.

References