TIRADS management guidelines in the investigation of thyroid nodules; illustrating the concerns, costs and performance.

Tom James Cawood\textsuperscript{1}, Georgia Rose Mackay\textsuperscript{2}, Penny Jane Hunt\textsuperscript{1,2}, Donal O’Shea\textsuperscript{3}, Stephen Skehan\textsuperscript{4}, Yi Ma\textsuperscript{5}

1) Department of Endocrinology, Christchurch Hospital, Canterbury District Health Board, New Zealand
2) University of Otago, Christchurch School of Medicine, New Zealand
3) Department of Endocrinology, St Vincent’s University Hospital, Dublin, Ireland
4) Department of Radiology, St Vincent’s University Hospital, Dublin and University College Dublin, Ireland
5) Biostatistician, Department of Medical & Women’s Business Management, Canterbury District Health Board, New Zealand

Key words: TIRADS, thyroid, cancer, ultrasound

Corresponding author: Tom James Cawood
tom.cawood@cdhb.health.nz

Co-authors:
Georgia Rose Mackay
georgiamackay04@gmail.com

Penny Jane Hunt
Penny.Hunt@cdhb.health.nz

Donal O’Shea
info@dosheaendo.ie

Stephen Skehan
sske@svhg.ie

© Endocrine Society 2020. js.2019-00415. See endocrine.org/publications for Accepted Manuscript disclaimer and additional information.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs licence (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial reproduction and distribution of the work, in any medium, provided the original work is not altered or transformed in any way, and that the work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
Yi Ma
Ma.Yi@cdhb.health.nz

Funding:
Nil

Disclosures:
Nil relevant
Abstract

Ultrasound (US) risk-stratification systems for investigation of thyroid nodules may not be as useful as anticipated. We aimed to assess the performance and costs of the American College of Radiology Thyroid Image Reporting And Data System (ACR-TIRADS). We examined the data set upon which ACR-TIRADS was developed, and applied TR1 or TR2 as a rule-out test, TR5 as a rule-in test, or applied ACR-TIRADS across all nodule categories. We assessed a hypothetical clinical comparator where 1 in 10 nodules are randomly selected for FNA, assuming a pre-test probability of clinically important thyroid cancer of 5%.

The gender bias (92% female) and cancer prevalence (10%) of the data set suggest it may not accurately reflect the intended test population. Applying ACR-TIRADS across all nodule categories did not perform well, with sensitivity and specificity between 60% and 80% and overall accuracy worse than random selection (65% vs 85%). Test performance in the TR3&TR4 categories had accuracy of less than 60%. Using TR5 as a rule-in test was similar to random selection (specificity 89% vs 90%). Using TR1&TR2 as a rule-out test had excellent sensitivity (97%), but for every additional person that ACR-TIRADS correctly reassures, this requires >100 ultrasound scans, resulting in 6 unnecessary operations and significant financial cost.

Perhaps surprisingly, the performance ACR-TIRADS may often be no better than random selection. The management guidelines may be difficult to justify from a cost/benefit perspective. A prospective validation study that determines the true performance of TIRADS in the real-world is needed.
Introduction

The diagnosis or exclusion of thyroid cancer is hugely challenging. A key factor is the low pre-test probability of important thyroid cancer but a higher chance of finding thyroid cancers that are very unlikely to cause ill health during a person's lifetime.

Thyroid nodules are common, affecting around half of the population and become increasingly common with advancing age (1,2). A minority of these nodules are cancers. The prevalence of incidental thyroid cancer at autopsy is around 10% (3). The more carefully one looks for incidental asymptomatic thyroid cancers at autopsy, the more are found (4), but these do not cause unwellness during life and so there is likely to be no health benefit in diagnosing them antemortem.

Amongst thyroid nodules detected during life, the often quoted figure for malignancy prevalence is 5% (5-8), with UptoDate quoting 4% to 6.5% in nonsurgical series (9), and it is likely that only a proportion of these cancers will be clinically significant (i.e. go on to cause ill-health). It is very difficult to know the true prevalence of important, clinically consequential thyroid cancers amongst patients presenting with thyroid nodules. There are inherent problems with studies addressing the issue such as selection bias at referral centres and not all nodules having FNA. In addition, changes in nomenclature such as the recent classification change to NonInvasive Follicular Thyroid neoplasm with Papillary-like nuclear features (NIFTP) would result in a lower rate of thyroid cancer if previous studies were reported using today's pathological criteria. Data sets with a thyroid cancer prevalence higher than 5% are likely to either include a higher proportion of small clinically
inconsequential thyroid cancers or be otherwise biased and not accurately reflect the true population prevalence.

The detection rate of thyroid cancer has increased steeply with widespread utilization of ultrasound (US) and frequent incidental detection of thyroid nodules with other imaging modalities such as CT, MRI and more recently PET-CT, yet the mortality from thyroid cancer has remained static (10,11). The implication is that US has enabled increased detection of thyroid cancers that are less clinically important (11-13). The health benefit from this is debatable and the financial costs significant.

Given that a proportion of thyroid cancers are clinically inconsequential, the challenge is finding a test that can effectively rule-in or rule-out important thyroid cancer (i.e. those cancers that will go on to cause morbidity or mortality). If one accepts that the pre-test probability of a patient presenting with a thyroid nodule having an important thyroid cancer is 5%, then a clinician who tells every patient they see that they do not have important thyroid cancer will be correct 95% of the time. Any additional test has to perform exceptionally well to surpass this clinician’s 95% negative predictive performance, without generating false positive results and consequential harm.

Test training then validation data sets

In order to develop a medical test a typical process is to generate a hypothesis from which a prototype is produced. If it performs well enough then the test is applied to a training set of data to better establish performance characteristics. Once the test is considered to be performing adequately, then it would be tested on a validation data set. The test may cycle
back between being used on training and validation data sets to allow for improvements and re-testing. The true test performance can only be established once the optimized test has been applied to one or more validation data sets and compared to the existing gold standard test. These final validation sets must fairly represent the population upon which the test is intended to be applied, as the prevalence of the condition in the test population will critically influence the test performance, particularly the positive predictive value (PPV) and negative predictive value (NPV). Such validation data sets need to be unbiased. In a clinical setting this would typically be an unselected sample of the test population, for example a consecutive series of all patients with a thyroid nodule presenting to a clinic, ideally across multiple centres. The gold test standard would need to be applied for comparison.

In the case of thyroid nodules there are further challenges. A study that looked at all nodules in consecutive patients (e.g. perhaps FNA of every nodule >10mm) would be required in order to get an accurate measure of the cancer prevalence in those nodules that might not typically get FNA. However, there are ethical issues with this, as well as the problem of over-diagnosis of small clinically inconsequential thyroid cancer.

Whilst the details of the design of the final validation study can be debated, the need for a well-designed validation study in order to determine the test characteristics in the real-world setting is a basic requirement of any new test.

Methods

This study aimed to assess the performance and costs of the American College of Radiology (ACR) Thyroid Image Reporting And Data System (TIRADS), by firstly looking for any important
issues in the methodology of its development, and then illustrating the performance of TIRADS for the initial decision for or against FNA, compared to an imagined clinical comparator of a group in which 1 in 10 nodules were randomly selected for FNA. This paper has only examined the ACR TIRADS system, noting that other similar systems exist such as Korean TIRADS (14) and EU TIRADS (15). We refer to ACR-TIRADS where data or comments are specifically related to “ACR TIRADS” and use the term “TIRADS” either for brevity or when comments may be applicable to other TIRADS systems.

Methodological concerns with ACR TIRADS data set

The main source data set for the ACR TIRADS recommendations was large and consisted of US images and FNA results of over 3400 nodules (16). This data set was a sub-set of data obtained for a previous study and there are no clear details of the inclusion and exclusion criteria, including criteria for FNA. The data set was 92% female and the prevalence of cancerous thyroid nodules was 10.3% (typical of the rate found on histology at autopsy, and double the 5% rate of malignancy in thyroid nodules typically quoted in the most relevant literature). Based on the methodology used to acquire the data set, the gender bias and cancer rate in the data set, it is unlikely to be a fair reflection of the population upon which the test is intended to be applied, and so cannot be considered a true validation set. Therefore taking results from this data set and assuming they would apply to the real-world population raises concerns.

Additional issues with the ACR TIRADS data set and guidelines

There are a number of additional issues that should be taken into account when examining the ACR TIRADS data set and resultant management recommendations.
Firstly, it should be noted that 10% of FNA or histology results were excluded due to “non-diagnostic findings” (16). These patients are not further considered in the ACR TIRADS guidelines. The actual number of inconclusive FNA results in the real-world validation set has not been established (as that study has not been done), but the typical rate is 30% (by this we mean non-diagnostic (i.e. insufficient cells), or indeterminate, (i.e. AUS / FLUS / follicular neoplasm / suspicious for follicular neoplasm (Bethesda I, III, IV)). For example, a previous meta-analysis of over 25,000 FNAs showed 33% were in these groups (17). After repeat US-guided FNA some patients achieve a cytological diagnosis, but typically two thirds remain indeterminate (18), accounting for approximately 20% of initial FNAs e.g. 10-30% (12) 31% (19) 22% (20). It is this proportion of patients that often go on to diagnostic hemithyroidectomies, from which approximately 20% are cancers (12,17,21), meaning the majority (80%) end up with ultimately unnecessary operations. The financial costs and surgical morbidity in this group must be taken into account when considering the cost/benefit repercussions of a test that includes US imaging for thyroid cancer.

Secondly, the proportion of patients in the different ACR TIRADS (TR) categories may, or may not, reflect the real-world population (Table 1). If one assumes that they do, then it is important to note that 25% of patients make up TR1 & TR2 and only 16% of patients make up TR5. Therefore 60% of patients are in the middle groups (TR3 & TR4), where the US features are less discriminatory. The consequences of these proportions are highly impactful when considering the real-world performance of ACR-TIRADS.
Thirdly, when moving on from the main study in which ACR TIRADS was developed (16) to the ACR TIRADS white paper recommendations (22), the TIRADS model changed by the addition of a fifth US characteristic (taller than wide), plus the addition of size cut-offs. Therefore, the rates of cancer in each ACR TIRADS category in the data set where they used four US characteristics can no longer be assumed to be the case using the five US characteristics plus the introduction of size cut-offs. Methodologically, the change in the ACR-TIRADS model should now undergo a new study using a new training data set (to avoid replicating any bias), before then undergoing a validation study.

**ACR TIRADS to rule-out or rule-in thyroid cancer**

Putting aside any potential methodological concerns with ACR TIRADS, it may be helpful to illustrate how TIRADS might work if one assumed that the data set used was a fair approximation to the real-world population. As the data set prevalence of thyroid cancer was 10%, compared to the generally accepted lower real-world prevalence of 5%, one can reasonably assume that the actual cancer rate in the ACR TIRADS categories in the real world would likely be half that quoted from the ACR TIRADS data set, and we illustrate below.

We are here imagining the consequence of 100 patients presenting to the thyroid clinic with either a symptomatic thyroid nodule (e.g. a nodule apparent to the patient from being palpable or visible) or an incidentally found thyroid nodule. In order to show the best possible performance of ACR TIRADS we are comparing it to clinical practice in the absence of TIRADS or other US thyroid nodule stratification tools, and based on a pre-test probability of thyroid cancer in a nodule being 5%, where 1 in 10 nodules are randomly selected for FNA. We chose a 1 in 10 FNA rate to reflect that roughly 5% of thyroid nodules are palpable and so would
likely go forward for FNA and we considered that a similar number would be selected for FNA based on clinical grounds such as other risk factors or the patient wishes. Furthermore, we are presuming other clinical factors (palpability, size, number, symptoms, age, gender, prior radiation exposure, family history, etc.) add no diagnostic value above random selection. We realise that such factors may increase an individual’s pre-test probability of cancer and clinical decision-making would change accordingly (e.g. proceeding directly to FNA), but we are here ascribe no additional diagnostic value to avoid over-estimating the performance of the clinical comparator. To further enhance the performance of TIRADS we presume that patients present with only one TR category of thyroid nodules. This allows a patient with a TR1 or TR2 nodule to be reassured that they, as a person, have a low risk of thyroid cancer, rather than a person with a mixture of nodules (not just TR1 or TR2) not being able to be reassured. This assumption is obviously not valid and favors TIRADS management guidelines, but we think it is helpful for clarity and illustrative purposes.

We first estimate the performance of ACR TIRADS guidelines’ recommended approach to the initial decision to perform FNA, by using TR1 or TR2 as a rule-out test, or using TR5 as a rule-in test, since applying TIRADS at the extremes of pre-test cancer risk (TR1 & TR2 for lowest risk, and TR5 for highest risk), is most likely to perform best. For this we do not take in to account nodule size, as size is not a factor in the ACR TIRADS guidelines for initial FNA in the TR1 and TR2 categories (where FNA is not recommended irrespective of size) or in the TR5 category (except in TR5 nodules of ≥0.5cm to <1.0cm, in which case US follow-up is recommended rather than FNA). Secondly, we then apply TIRADS across all five nodule categories to give an idea how TIRADS is likely to perform overall. For this we do take into account the nodule size cut-offs but note that for the TR3 and TR4 categories ACR TIRADS
does not detail how it chose the size cut-offs of 2.5cm and 1.5cm, respectively. These cut-offs are somewhat arbitrary, with conflicting data as to what degree, if any, size is a discriminatory factor. Many studies have not found a clear size/malignancy correlation, and where it has been found, the magnitude of the effect is modest. There are even data showing a negative correlation between size and malignancy (23). Perhaps the most relevant positive study is from Korea which found in TR4 group the cancer rate was no different between nodules measuring between 1-2cm (22.3%) and those 2-3cm (23.5%), but the rate did increase above 3cm (40%) (24). In the TR3 category there was a gradual difference in cancer rate in those 1-2cm (6.5%), and those 2-3cm (8.4%) and those >3cm (11.3%). To illustrate the effect of the size cut-offs we have given two examples, one where the size cut-offs are not discriminatory and the cancer rate is the same above and below the size cut-off, and the second example where the cancer risk of the nodule doubles once the size goes above the cut-off. The two examples provide a range of performance within which the real test performance is likely to be, with the second example likely to provide TIRADS with a more favourable test performance than in the real-world. For the calculations we assume an approximate size distribution where a third of TR3 nodules are ≥25mm and half of TR4 nodules are ≥15mm. We have also assumed that all nodules are at least 10mm and so the TRS nodule size cut-off of 5mm does not apply.

For TIRADS to add clinical value, it would have to clearly out-perform the comparator (random selection), particularly as we have made some assumptions that favor TIRADS performance. We have also estimated the likely costs associated with using the ACR TIRADS guidelines, though for simplicity have not included the costs of molecular testing for indeterminate
nODULES (which is not readily available in the New Zealand public health system) nor any US follow-up and associated costs.

Results
We have detailed the data set used for the development of ACR TIRADS(16) in Table 1, plus noted the likely cancer rates in the real world if one assumes that the data set cancer prevalence (10.3%) is double that in the population upon which the test is intended to be used (pre-test probability of 5%).

Using TIRADS as a rule-out cancer test would be the finding that a nodule is TR1 or TR2 and hence has a low risk of cancer, compared to being TR3-5. Whereas using TIRADS as a rule-in cancer test would be the finding that a nodule is TR5, with a sufficiently high chance of cancer that further investigations are required, compared to being TR1-4.

The summary of test performance of random selection, ACR TIRADS as a rule-out test, ACR TIRADS as a rule-in test, and ACR TIRADS applied across all TIRADS categories are detailed in Table 2, and the full data, definitions and calculations are given in supplementary data (25).

ACR TIRADS as a rule-out test
We found better sensitivity, PPV and NPV with TIRADS compared to random selection (97% vs 1%, 13% vs 1%, and 99% vs 95%, respectively), whereas specificity and accuracy were worse with TIRADS compared to random selection (27% vs 90%, and 34% vs 85%, respectively), Table 2 (25).
ACR TIRADS as a rule-in test

We found sensitivity and PPV with TIRADS was poor, but was better than random selection (sensitivity 53% vs 1%, and PPV 34% vs 1%) whereas specificity, NPV and accuracy was no better with TIRADS compared to random selection (specificity 89% vs 90%, NPV 94% vs 95%, and accuracy 85% vs 85%), Table 2 (25).

ACR TIRADS across all nodule categories

ACR TIRADS performed poorly when applied across all five TR categories, with specificity lower than with random selection (63% vs 90%). PPV was poor (20%), NPV was no better than random selection, and accuracy was worse than random selection (65% vs 85%). Sensitivity of ACR TIRADS was better than random selection, between 74% to 81% (depending on whether the size cut-offs add value) compared to 1% with random selection. However, most of the sensitivity benefit is due to the performance in the TR1 and TR2 categories, with sensitivity in just the TR3 and TR4 categories being only 46% to 62%, depending on whether the size cut-offs add value (data not shown).

Cost estimates of ACR TIRADS as a rule-out test

For a rule-out test, sensitivity is the more important test metric. A negative result with a highly sensitive test is valuable for ruling out the disease. Therefore using TIRADS categories TR1 or TR2 as a rule-out test should perform very well, with sensitivity of the rule-out test being 97%. However, in the data set only 25% of all nodules were categorised as TR1 or TR2 and these nodules harboured only 1% of all thyroid cancers (9 of 343). So just using ACR TIRADS as a rule-out test could be expected to leave 99% of undiagnosed cancers amongst the remaining 75% of the population, in whom the investigation and management remains unresolved.
If one assumes that in the real world 25% of the patients have a TR1 or TR2 nodule, applying TIRADS changes the pre-test 5% probability of cancer to a post-test risk of 1%, so the absolute risk reduction is 4%. Therefore, for every 25 patients scanned (100/4 = 25) and found to be either TR1 or TR2, one additional person would be correctly reassured that they do not have thyroid cancer. However, given that TR1 and TR2 make up only 25% of the nodules, then in order to find 25 nodules that are TR1 or TR2, you would need to do 100 scans. So the number needed to scan (NNS) for each additional person correctly reassured is 100 (NNS = 100). This is likely an underestimate of the number of scans needed, given that not all nodules that are TR1 or TR2 will have purely TR1 or TR2 nodules on their scan. For those that also have one or more TR3, TR4 or TR5 nodules on their scan, they cannot have thyroid cancer ruled-out by TIRADS as the possibility that their non-TR1/TR2 nodules may be cancerous is still unresolved.

If you do 100 (or more) US scans on patients with a thyroid nodule, and apply the ACR TIRADS management guidelines for FNA, this results in costs and morbidity from the resultant FNAs and the indeterminate results that are then considered for diagnostic hemithyroidectomy. The costs depend on the threshold for doing FNA. If you assume that FNA is done as per reasonable application of TIRADS recommendations (in all patients with TR5 nodules, half of patients with TR4 nodules and a third of patients with TR3 nodules) and the proportion of patients in the real world have roughly similar proportion of TR nodules as the data set used, then 100 US scans would result in FNAs of about half of all patients scanned (of data set, 16% were TR5, 37% were TR4 and 23% were TR3, so FNA number from 100 scans = 16+(0.5x37)+(0.3x23) = 42). Given the need to do more than 100 US scans to find 25 patients with just TR1 or TR2 nodules, this would result in at least 50 FNAs being done.
For every 100 FNAs performed, about 30 are inconclusive, with most (e.g. 20% of the original 100) remaining indeterminate after repeat FNA and requiring diagnostic hemithyroidectomy. Ultimately most of these turn out to be benign (80%), so for every 100 FNAs, you end up with 16 (100 x 0.2 x 0.8) unnecessary operations being performed.

Therefore, compared to randomly selecting 1 in 10 nodules for FNA, using ACR TIRADS to correctly rule out thyroid cancer in one additional patient would require more than 100 US scans (NNS > 100) to find 25 TR1 and TR2 patients, triggering at least 40 additional FNAs and resulting in approximately six additional unnecessary diagnostic hemithyroidectomies at significant economic and personal costs. The financial cost depends on the health system involved but as an example, in New Zealand where healthcare costs are modest by international standards in the developed world, compared to randomly selecting 1 in 10 nodules for FNA, using ACR TIRADS would result in approximately NZ$140,000 spent for every additional patient correctly reassured that they do not have thyroid cancer(25).

Cost estimates of ACR TIRADS as a rule-in test

The more important test metric for diagnosing a disease is the specificity, where a positive test helps rule-in the disease. The specificity of TIRADS is high (89%) but, perhaps surprisingly, is similar to randomly selecting of 1 in 10 nodules for FNA (90%).

If one decides to FNA every TR5 nodule, from the original ACR TIRADS data set, 34% were found to be cancerous but note that this data set likely has double the prevalence of thyroid cancer compared to the real-world population.
TR5 in the data set made up 16% of nodules, in which they found half of the thyroid cancers (183/343). This equates to two to three cancers if one assumes a thyroid cancer prevalence of 5% in the real world. To find 16 TR5 nodules requires 100 people to be scanned (assuming for illustrative purposes one nodule per scan). So, for 100 scans, if FNA is done on all TR5 nodules, this will find half of the cancers and so will miss half of the cancers.

Performing FNA on TR5 nodules is a relatively effective way of case finding thyroid cancers. However, if the concern is that this might miss too many thyroid cancers then this could be compared to the range of alternatives, i.e. doing no tests or doing many more FNAs. If a clinician does no tests and no FNAs then that clinician will miss all thyroid cancers (5 people per 100). Thus the absolute risk of missing important cancer goes from 5% (with no FNAs) to 2.5% using TIRADS and FNA of all TR5, so NNS = 100/2.5 = 40. Alternatively, if you randomly FNA 1 in 10 nodules then you will miss 4.5 thyroid cancers (4-5 people per 100). Thus the absolute risk of missing important cancer goes from 4.5% to 2.5%, so NNS = 100/2= 50.

Compared to randomly doing FNA on 1 in 10 nodules, using ACR TIRADS and doing FNA on all TR5 requires NNS of 50 to find one additional cancer. This comes at the cost of missing as many cancers as you find, spread amongst 84% of the population, and doing 1 additional unnecessary operation (16x0.2x0.8 = 2.6, minus the 1.6 unnecessary operations resulting from random selection of 1 in 10 patients for FNA(25)), plus the financial costs involved. The cost of seeing 100 patients and only doing FNA on TR5 is at least NZ$100,000 (compared to $60,000 for seeing all patients and randomly doing FNA on 1 in 10 patients), so being at least NZ$20,000 per cancer found if the prevalence of thyroid cancer in the population is 5% (25).
The optimal investigation and management of the 84% of the population harboring the remaining 50% of cancer remains unresolved.

The other half of the cancers that are missed by only doing FNA of TR5 nodules will mainly be in the TR3 and TR4 groups (that make up 60% of the population), and these groups will have a 3% to 8% chance of cancer, depending upon whether the population prevalence of thyroid cancer in those being tested is 5% or 10%. If the proportions of patients in the different TR groups in the ACR TIRADs data set is similar to the real-world population, then the prevalence of thyroid cancer in the TR3 and TR4 groups is lower than in the overall population of patients with thyroid nodules. The performance of any diagnostic test in this group has to be truly exceptional to outperform random selection and accurately rule-in or rule-out thyroid cancer in the TR3 or TR4 groups. TIRADS does not perform to this high standard. Following ACR TIRADS management guidelines would likely result in approximately half of the TR3 & TR4 patients getting FNAs \((0.5 \times 37) + (0.3 \times 23) = 25\), of total 60), finding up to one cancer, and result in four diagnostic hemithyroidectomies for benign nodules \(25 \times 0.2 \times 0.8 = 4\). The more FNAs done in the TR3 and TR4 groups, the more indeterminate FNAs and the more financial costs and unnecessary operations.

Given that ACR TIRADS test performance is at its worst in the TR3 and TR4 groups, then the costs-effectiveness of TIRADS will also be at its worst in these groups, in particular due to the false positive TIRADS results.

Of note, we have not taken into account any of the benefits, costs or harms associated with the proposed US follow-up of nodules, as recommended by ACR-TIRADS. It is perhaps worth
noting that the US follow-up is mainly recommended for the smaller TR3 and TR4 nodules, and the prevalence of thyroid cancer in these groups in a real-world population with overall cancer risk of 5% is low, likely <3%. The chance of finding a consequential thyroid cancer during follow-up is correspondingly low. It is also relevant to note that the change in nodule appearance over time is poorly predictive of malignancy. At best, only a minority of the 3% of cancers would show on follow-up imaging features suspicious for thyroid cancer that correctly predict malignancy. Some cancers would not show suspicious changes thus US features would be falsely reassuring. The vast majority of nodules followed-up would be benign (>97%), and so the majority of FNAs triggered by US follow-up would either be benign, indeterminate or false positive, resulting in more potential for harm (16 unnecessary operations for every 100 FNAs).

Discussion

The cost effective diagnosis or exclusion of consequential thyroid cancer is an everyday problem faced by all thyroid clinicians. The challenge of appropriately balancing the risks of missing an important cancer versus the chance of causing harm and incurring significant costs from over-investigation is major. Those working in this field would gratefully welcome a diagnostic modality that can improve the currently uncertainty.

The TIRADS reporting algorithm is a significant advance with clearly defined objective sonographic features that are simple to apply in practice. The ACR-TIRADS guidelines also provide easy-to-follow management recommendations which have understandably generated momentum. Unfortunately, the collective enthusiasm for welcoming something that appears to provide certainty has perhaps led to important flaws in the development of
the models being overlooked. ACR TIRADS has not been applied to a true validation set upon which it is intended to be used, and therefore needs to be considered with caution when applying it to the real-world situation. The current ACR TIRADS system changed from that assessed during training, with the addition of the taller-than-wide criterion and size criteria, which further questions the assumption that the test should perform in the real world as it did on a the initial training data set.

The low pretest probability of important thyroid cancer and the clouding effect of small clinically inconsequential thyroid cancers makes the development of an effective real-world test incredibly difficult. Any test will struggle to outperform educated guessing to rule out clinically important thyroid cancer. The NNS for ACR TIRADS is such that it is hard to justify its use for ruling out thyroid cancer (NNS>100), at least on a cost/benefit basis. Using ACR-TIRADS as a rule-in test to identify a higher risk group that should have FNA is arguably a more effective application. Quite where the cutoff should be is debatable, but any cutoff below TR5 will have diminishing returns and increasing harms. A TR5 cutoff would have NNS of 50 per additional cancer found compared to random FNA of 1 in 10 nodules, and probably a higher NNS if one believes that clinical factors can increase FNA hit rate above the random FNA hit rate.

Whilst we somewhat provocatively used random selection as a clinical comparator, we do not mean to suggest that clinicians work in this way. Clinicians should be using all available data to arrive at an educated estimate of each individual patient’s pre-test probability of having clinically significant thyroid cancer and use their clinical judgement to help advise each patient of their best options. This approach likely performs better than randomly selecting 1 in 10
nodules for FNA, but we intentionally made assumptions that would favor the performance of ACR TIRADS to illustrate that if a poor clinical comparator cannot clearly be beaten, then the clinical value that such new systems bring is correspondingly poor.

This study has many limitations. It is limited by only being an illustrative example that does not take clinical factors into account such as prior radiation exposure and clinical features. The ACR TIRADS management flow-chart also does not take into account these clinical factors. It would be unfair to add these clinical factors to only the TIRADS arm or only to the clinical comparator arm, and they would cancel out if added to both arms, hence they were omitted. As noted above, we intentionally chose the clinical comparator to be relatively poor and not a fair reflection of real-world practice, to make it clearer to what degree ACR TIRADS adds value.

The ACR TIRADS white paper (22) very appropriately notes that the recommendations are intended to serve as guidance and that professional judgement should be applied to every case including taking into account factors such as a patient’s cancer risk, anxiety, co-morbidities and life expectancy. However, the ACR TIRADS flow chart with its sharp cut-offs conveys a degree of certainty which may not be valid and may be hard for the clinician to resist. If a guideline indicates that FNA is recommended, it can be difficult to oppose this based on other factors. Such guidelines do not detail the absolute risk of finding or missing a cancer, nor the often excellent outcome of the treatment of thyroid cancer, nor the potential for unnecessary operations. Such data should be included in guidelines, particularly if clinicians wish to provide evidence-based guidance and to obtain truly informed consent for any action that may have negative consequences.
Another clear limitation of this study is that we only examined the ACR TIRADS system. Other similar systems are in use internationally, e.g. Korean-TIRADS(14) and EU-TIRADS(15). These appear to share the same basic flaw as the ACR-TIRADS, in that the data sets of nodules used for their development is not likely to represent the population upon which it is intended for use, at least with regard to pre-test probability of malignancy (e.g. malignancy rate 12% for Korean TIRADS(26), and 18% and 31% for EU TIRADS categories 4 & 5 (27,28)). Attempts to compare the different TIRADS systems on data sets that are also not reflective of the intended test population are similarly flawed (e.g. malignancy rates of 41% (29)). Other limitations include the various assumptions we have made and the fact that we applied ACR TIRADS to the same data set upon which is was developed. However, these assumptions have intentionally been made to favor the expected performance of ACR-TIRADS, and so in real life ACR-TIRADS can be expected to perform less well than we have illustrated.

The key next step for any of the TIRADS systems, and for any similar proposed test system including artificial intelligence (30-32), is to perform a well-designed prospective validation study to measure the test performance in the population upon which it is intended for use. Such a study should also measure any unintended harm, such as financial costs and unnecessary operations, and compare this to any current or gold standard practice against which it is proposed to add value. It should also be on an “intention-to-test” basis and include the outcome for all those with indeterminate FNAs. Until a well-designed validation study is completed, the performance of TIRADS in the real-world is unknown. The figures that TIRADS provides, such as cancer prevalence in certain groups of patients, or consequent management guidelines, only apply to populations that are similar to their data set.
It is interesting to see the wealth of data used to support TIRADS as being an effective and validated tool. Many of these papers share the same fundamental problem of not applying the test prospectively to the population upon which it is intended for use. Instead it has been applied on retrospective data sets, with cancer rates far above 5%, rather than on consecutive unselected patients presenting with a thyroid nodule (33). It has been retrospectively applied to thyroidectomy specimens, which is clearly not representative of the patient presenting with a thyroid nodule (34-36), and has even been used on the same data set used for TIRADS development, clearly introducing obvious bias (32,37). These publications erroneously add weight to the belief that TIRADS is a proven and superior model for the investigation of thyroid nodules.

A recent meta-analysis comparing different risk stratification systems, included 13,000 nodules, mainly from retrospective studies, had a prevalence of cancer of 29%, and even in that setting the test performance of TIRADS was disappointing (e.g. sensitivity 74%, specificity 64%, PPV 43%, NPV 84%), and similar to our estimated values of TIRADS test performance (38).

Whilst our findings have illustrated some of the shortcomings of ACR TIRADS guidelines, we are not able to provide the ideal alternative. In a cost conscious public health system, one could argue that after selecting out those patients that clearly raise concern for a high risk of cancer (from history including risk factors, examination, existing imaging, etc.) the clinician could reasonably inform an asymptomatic patient that they have a 95% chance of their nodule being benign. If a patient was happy taking this small risk (and particularly if the
patient has significant co-morbidities), then it would be reasonable to do no further tests, including no US, and instead do some safety netting by advising the patient to return if symptoms changed (e.g. subsequent clinically apparent nodule enlargement). If a patient presented with symptoms (e.g. concerns about a palpable nodule) and/or was not happy accepting a 5% pre-test probability of thyroid cancer then further investigations could be offered, noting that US cannot reliably rule in or rule out thyroid cancer for the majority of patients, and that doing any testing comes with unintended risks. The chance of finding cancer is 1 in 20, whereas the chance of testing resulting in an unnecessary operation is around 1 in 7. Those wishing to continue down the investigative route could then have US, using TIRADS or ATA guidelines or other measures to offer some relative risk-stratification. Until TIRADS is subjected to a true validation study, we do not feel that a clinician can currently accurately predict what a TIRADS classification actually means, nor what the most appropriate management thereafter should be.

Conclusion
The findings that ACR TIRADS has methodologically concerns, is not yet truly validated, often performs no better than random selection, and drives significant costs and potential harm, are very unsettling but result from a rational and scientific assessment of the foundational basis of the ACR TIRADS system. TIRADS can be welcomed as an objective way to classify thyroid nodules into groups of differing (but as yet unquantifiable) relative risk of thyroid cancer. However, the consequent management guidelines are difficult to justify at least on a cost basis for a rule-out test, though ACR TIRADS may provide more value as a rule-in test for a group of patients with higher cancer risk. A robust validation study is required before the performance and cost-benefit outcomes of any of the TIRADS systems can be known. There

23
remains the need for a highly-performing diagnostic modality for clinically important thyroid cancers.
References

1. Tan GH, Gharib H. Thyroid incidentalomas: management approaches to nonpalpable nodules discovered incidentally on thyroid imaging. Ann Intern Med. 1997;126(3):226-231.
2. Jiang H, Tian Y, Yan W, Kong Y, Wang H, Wang A, Dou J, Liang P, Mu Y. The Prevalence of Thyroid Nodules and an Analysis of Related Lifestyle Factors in Beijing Communities. 2016;13(4):442.
3. Furuya-Kanamori L, Bell KJL, Clark J, Glasziou P, Doi SAR. Prevalence of Differentiated Thyroid Cancer in Autopsy Studies Over Six Decades: A Meta-Analysis. J Clin Oncol. 2016;34(30):3672-3679.
4. Harach HR, Franssila KO, Wasenius VM. Occult papillary carcinoma of the thyroid. A "normal" finding in Finland. A systematic autopsy study. Cancer. 1985;56(3):531-538.
5. Hegedus L. Clinical practice. The thyroid nodule. N Engl J Med. 2004;351(17):1764-1771.
6. Bessey LJ, Lai NB, Coorough NE, Chen H, Sipple RS. The incidence of thyroid cancer by fine needle aspiration varies by age and gender. J Surg Res. 2013;184(2):761-765.
7. Lin JD, Chao TC, Huang BY, Chen ST, Chang HY, Hsueh C. Thyroid cancer in the thyroid nodules evaluated by ultrasonography and fine-needle aspiration cytology. Thyroid. 2005;15(7):708-717.
8. Bongiovanni M, Crippa S, Baloch Z, Piana S, Spitale A, Pagni F, Mazzucchelli L, Di Bella C, Faquin W. Comparison of 5-tiered and 6-tiered diagnostic systems for the reporting of thyroid cytopathology: a multi-institutional study. Cancer Cytopathol. 2012;120(2):117-125.
9. Ross DS. Diagnostic approach to and treatment of thyroid nodules. Uptodate. 2019.
10. Davies L, Welch HG. Current thyroid cancer trends in the United States. JAMA Otolaryngol Head Neck Surg. 2014;140(4):317-322.
11. Park S, Oh CM, Cho H, Lee JY, Jung KW, Jun JK, Won YJ, Kong HJ, Choi KS, Lee YJ, Lee JS. Association between screening and the thyroid cancer "epidemic" in South Korea: evidence from a nationwide study. BMJ. 2016;355:i5745.
12. Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE, Pacini F, Randolph GW, Sawka AM, Schlumberger M, Schuff KG, Sherman SI, Sosa JA, Steward DL, Tuttle RM, Wartofsky L. 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer. Thyroid. 2016;26(1):1-133.
13. Haymart MR, Banerjee M, Reyes-Gastelum D, Caoili E, Norton EC. Thyroid Ultrasound and the Increase in Diagnosis of Low-Risk Thyroid Cancer. J Clin Endocrinol Metab. 2019;104(3):785-792.
14. Shin JH, Baek JH, Chung J, Ha EJ, Kim JH, Lee YH, Lim HK, Moon WJ, Na DG, Park JS, Choi YJ, Hahn SY, Jeon SJ, Jung SL, Kim DW, Kim EK, Kwak JY, Lee CY, Lee HJ, Lee JH, Lee JH, Lee KH, Park SW, Sung JY, Korean Society of Thyroid R, Korean Society of R. Ultrasonography Diagnosis and Imaging-Based Management of Thyroid Nodules: Revised Korean Society of Thyroid Radiology Consensus Statement and Recommendations. Korean J Radiol. 2016;17(3):370-395.
15. Russ G, Bonnema SJ, Erdogan MF, Durante C, Ngu R, Leenhardt L. European Thyroid Association Guidelines for Ultrasound Malignancy Risk Stratification of Thyroid Nodules in Adults: The EU-TIRADS. Eur Thyroid J. 2017;6(5):225-237.
16. Middleton WD, Teefey SA, Reading CC, Langer JE, Beland MD, Szabunio MM, Desser TS. Multinstitutional Analysis of Thyroid Nodule Risk Stratification Using the American College of Radiology Thyroid Imaging Reporting and Data System. Am J Roentgenol. 2017;208:1331-1341.
17. Bongiovanni M, Spitali A, Faquin WC, Mazzucchelli L, Baloch ZW. The Bethesda System for Reporting Thyroid Cytopathology: a meta-analysis. Acta Cytol. 2012;56(4):333-339.
18. Allen L, Al Afif A, Rigby MH, Bullock MJ, Trites J, Taylor SM, Hart RD. The role of repeat fine needle aspiration in managing indeterminate thyroid nodules. J Otolaryngol Head Neck Surg. 2019;48(1):16.
19. Nayar R, Ivanovic M. The indeterminate thyroid fine-needle aspiration: experience from an academic center using terminology similar to that proposed in the 2007 National Cancer Institute Thyroid Fine Needle Aspiration State of the Science Conference. Cancer. 2009;117(3):195-202.
20. Anderson TJ, Atalay MK, Grand DJ, Baird GL, Cronan JJ, Beland MD. Management of nodules with initially nondiagnostic results of thyroid fine-needle aspiration: can we avoid repeat biopsy? Radiology. 2014;272(3):777-784.
21. Cibas ES, Ali SZ, Conference NCITFSotS. The Bethesda System For Reporting Thyroid Cytopathology. Am J Clin Pathol. 2009;132(5):658-665.
22. Tessler FN, Middleton WD, Grant EG, Hoang JK, Berland LL, Teefey SA, Cronan JJ, Beland MD, Dessar TS, Frates MC, Hammers LW, Hamper UM, Langer JE, Reading CC, Scoutt LM, Stavros AT. ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White Paper of the ACR TI-RADS Committee. J Am Coll Radiol. 2017;14(5):587-595.
23. Cavallo A, Johnson DN, White MG, Siddiqui S, Antic T, Mathew M, Grogan RH, Angelos P, Kaplan EL, Cipriani NA. Thyroid Nodule Size at Ultrasound as a Predictor of Malignancy and Final Pathologic Size. Thyroid. 2017;27(5):641-650.
24. Hong MJ, Na DG, Baek JH, Sung JY, Kim JH. Impact of Nodule Size on Malignancy Risk Differs according to the Ultrasonography Pattern of Thyroid Nodules. Korean J Radiol. 2018;19(3):534-541.
25. Cawood T, Mackay GR, Hunt PJ, O’Shea D, Skehan S, Ma Y. TIRADS management guidelines in the investigation of thyroid nodules; an illustration of the concerns, costs and performance. Figshare Digital Repository. 2020;Deposited January 2020. https://doi.org/10.6084/m9.figshare.11640168.v.
26. Na DG, Kim JH, Kim DS, Kim SJ. Thyroid nodules with minimal cystic changes have a low risk of malignancy. Ultrasonography. 2016;35(2):153-158.
27. Russ G, Bigorgne C, Royer B, Rouxel A, Bienvenu-Perrard M. [The Thyroid Imaging Reporting and Data System (TIRADS) for ultrasound of the thyroid]. J Radiol. 2011;92(7-8):701-713.
28. Yoon JH, Lee HS, Kim EK, Moon HJ, Kwak JY. Malignancy Risk Stratification of Thyroid Nodules: Comparison between the Thyroid Imaging Reporting and Data System and the 2014 American Thyroid Association Management Guidelines. Radiology. 2016;278(3):917-924.
29. Xu T, Wu Y, Wu RX, Zhang YZ, Gu JY, Ye XH, Tang W, Xu SH, Liu C, Wu XH. Validation and comparison of three newly-released Thyroid Imaging Reporting and Data Systems for cancer risk determination. Endocrine. 2019;64(2):299-307.

30. Zhang B, Tian J, Pei S, Chen Y, He X, Dong Y, Zhang L, Mo X, Huang W, Cong S, Zhang S. Machine Learning-Assisted System for Thyroid Nodule Diagnosis. Thyroid. 2019.

31. Wang L, Yang S, Yang S, Zhao C, Tian G, Gao Y, Chen Y, Lu Y. Automatic thyroid nodule recognition and diagnosis in ultrasound imaging with the YOLOv2 neural network. World J Surg Oncol. 2019;17(1):12.

32. Wildman-Tobriner B, Buda M, Hoang JK, Middleton WD, Thayer D, Short RG, Tessler FN, Mazurowski MA. Using Artificial Intelligence to Revise ACR TI-RADS Risk Stratification of Thyroid Nodules: Diagnostic Accuracy and Utility. Radiology. 2019:182128.

33. Trimboli P, Ng R, Royer B, Giovanella L, Bigorgne C, Simo R, Carroll P, Russ G. A multicenter validation study for the EU-TIRADS using histological diagnosis as a gold standard. Clin Endocrinol (Oxf). 2019.

34. Gao L, Xi X, Jiang Y, Yang X, Wang Y, Zhu S, Lai X, Zhang X, Zhao R, Zhang B. Comparison among TIRADS (ACR TI-RADS and KWAK-TI-RADS) and 2015 ATA Guidelines in the diagnostic efficiency of thyroid nodules. Endocrine. 2019;64(1):90-96.

35. Horvath E, Silva CF, Majlis S, Rodriguez I, Skoknic V, Castro A, Rojas H, Niedmann JP, Madrid A, Capdeville F, Whittle C, Rossi R, Domínguez M, Tala H. Prospective validation of the ultrasound based TIRADS (Thyroid Imaging Reporting And Data System) classification: results in surgically resected thyroid nodules. Eur Radiol. 2017;27(6):2619-2628.

36. Ha SM, Baek JH, Na DG, Suh CH, Chung SR, Choi YJ, Lee JH. Diagnostic Performance of Practice Guidelines for Thyroid Nodules: Thyroid Nodule Size versus Biopsy Rates. Radiology. 2019;291(1):92-99.

37. Middleton WD, Teefey SA, Reading CC, Langer JE, Beland MD, Szabunio MM, Desser TS. Comparison of Performance Characteristics of American College of Radiology TIRADS, Korean Society of Thyroid Radiology TIRADS, and American Thyroid Association Guidelines. AJR Am J Roentgenol. 2018;210(5):1148-1154.

38. Castellana M, Castellana C, Treglia G, Giorgino F, Giovanella L, Russ G, Trimboli P. Performance of five ultrasound risk stratification systems in selecting thyroid nodules for FNA. A meta-analysis. J Clin Endocrinol Metab. 2019.
Table 1. Data set used for development of ACR TIRADS(16), and used for this paper. The possible cancer rate column is a crude, unvalidated estimate, calculated by proportionately reducing the cancer rates by 10.3% : 5% to reflect the likely difference in the cancer rate in the data set used (10.3%) and in the population presenting with a thyroid nodule (5%). These figures cannot be known for any population until a real-world validation study has been performed on that population.

| TIRADS category | Number of nodules | % of total nodules | Number of cancerous nodules | Number of benign nodules | Cancer prevalence in that TR category (overall cancer rate in the data set was 10.3%) | Possible cancer prevalence in that TR category if overall cancer rate in test population is 5% |
|-----------------|------------------|--------------------|-----------------------------|-------------------------|-----------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------|
| TR1             | 299              | 9%                 | 1                           | 298                     | 0.3%                                                                              | 0.2%                                                                                         |
| TR2             | 548              | 16%                | 8                           | 540                     | 1.5%                                                                              | 0.7%                                                                                         |
| TR3             | 775              | 23%                | 37                          | 738                     | 4.8%                                                                              | 2.4%                                                                                         |
| TR4             | 1251             | 37%                | 114                         | 1137                    | 9.1%                                                                              | 4.6%                                                                                         |
| TR5             | 534              | 16%                | 183                         | 351                     | 34.3%                                                                             | 17.1%                                                                                         |
| Total           | 3407             | 100%               | 343                         | 3064                    |                                                                                   |                                                                                              |
### Table 2. Summary test performance of random selection of 1 in 10 nodules for FNA, compared to ACR-TIRADS. Full data including 95% confidence intervals are given in the supplementary data (25)

|                       | Random Selection | ACR TIRADS as a rule-out test | ACR TIRADS as a rule-in test | ACR TIRADS, assuming TR3 and TR4 size cut-offs double the cancer rate | ACR TIRADS assuming TR3 and TR4 size cut-offs make no difference to cancer rate |
|-----------------------|------------------|-------------------------------|-----------------------------|---------------------------------------------------------------------|--------------------------------------------------------------------------------|
| **Sensitivity**       | 1%               | 97%                           | 53%                         | 81%                                                                 | 74%                                                                            |
| **Specificity**       | 90%              | 27%                           | 89%                         | 63%                                                                 | 62%                                                                            |
| **Positive Predictive**| **Value**       | 1%                            | 13%                         | 34%                                                                 | 20%                                                                            |
| **Negative Predictive**| **Value**       | 95%                           | 99%                         | 94%                                                                 | 97%                                                                            |
| **Accuracy**          | 85%              | 34%                           | 85%                         | 65%                                                                 | 63%                                                                            |