Management of indeterminate thyroid nodules: changing the paradigm

Koster, E.J. de

chapter 3

Non-invasive imaging biomarkers of thyroid nodules with indeterminate cytology

W.A. Noortman
E.J. de Koster
F.H.P. van Velden
L.F. de Geus-Oei
D. Vriens

Published in: Luca Giovanella, editors. Integrated Diagnostics and Theragnostics of Thyroid Diseases. Springer International Publishing; 2023. ISBN 9783031352133

Abstract

Stratified by ultrasonography and fine-needle aspiration cytology, indeterminate nodules are lesions with an intermediate risk of being malignant (approximately 25%). Diagnostic resection of half the thyroid gland provides the true nature of these lesions but at the cost of 75% of the patients being futilely operated. In case a malignancy is found, a second surgery to remove the whole thyroid gland is often necessary. Different approaches either on cytological material or using imaging have been investigated to further stratify these lesions. This chapter reviewed biomarkers obtained using conventional as well as artificial intelligence-based non-invasive imaging strategies for the differentiation of thyroid nodules with indeterminate cytology. An overview of the abilities of different tests to differentiate between benign and malignant nodules was provided, taking into account the clinical readiness and cost-effectiveness.

Introduction

After stratification by ultrasonography (US), the next step in analysis of a thyroid nodule in a nonhyperthyroid patient is by obtaining cytology. Usually this is performed by fine-needle aspiration cytology (FNAC) as this procedure is simple, safe, inexpensive and has a high accuracy. The FNAC specimens are categorised into six diagnostic categories according to the Bethesda System for the Reporting of Thyroid Cytology (See Chapter 1, Table 1) [18]. Around 20% of thyroid nodules are cytologically indeterminate, including both atypia of undetermined significance or follicular nodules of undetermined significance (Bethesda III, AUS/FLUS) with a malignancy rate of 6-18% and cytology suspicious for a follicular neoplasm (Bethesda IV, FN/SFN) or Hürthle cell neoplasm (Bethesda IV, HCN/SHCN) together having a malignancy rate of 10-40% [18, 49]. Nodules that are suspicious for malignancy upon FNAC (Bethesda V, SUSP) encounter a malignancy rate of 45-60% and can also be considered cytologically indeterminate [18].

Indeterminate thyroid cytology corresponds to histopathological follicular adenoma (FA), Hürthle cell adenoma (HCA), non-invasive follicular thyroid neoplasm with papillary-like nuclear features (NIFTP), (encapsulated) follicular variant of papillary thyroid carcinoma ((E)FVPTC), follicular thyroid carcinoma (FTC), and Hürthle cell carcinoma (HCC), but can also be seen in papillary thyroid carcinoma (PTC). Unlike histology, cytology does not provide insight into tissue structure: it does not show the capsular and/or vascular invasion that distinguishes an FTC from a benign FA. In FVPTC, the growth pattern is follicular and clearly identifying nuclear features of PTC can usually not be distinguished cytologically.

As an alternative to FNAC, the use of core needle histological biopsy has recently received increased interest [360]. Although lower nondiagnostic (Bethesda I) or indeterminate rates are published, core needle histological biopsy requires more advanced training for radiologists and histopathologists. Furthermore, this procedure is more painful for patients and has more complications including haematomas and voice changes, and therefore it has not been well-adopted.

The American Thyroid Association (ATA) specified recommendations for the clinical management of the different cytological categories [17]. Repeat FNAC for a Bethesda III nodule may oftentimes result in a Bethesda II result. For nodules that remain Bethesda III after repeat FNAC, or those with Bethesda IV or V cytology, diagnostic hemithyroidectomy is often performed [17]. As the joint malignancy rate in indeterminate nodules is approximately 25%, approximately 75% of these diagnostic surgeries result in a benign histopathological diagnosis. For these benign nodules, the diagnostic surgery can be considered unbeneficial from an oncological perspective, increasing health care consumption expenses and exposing patients to unnecessary surgical risks. In case of malignant histopathology, a completion thyroidectomy might be indicated, putting the patient at a higher risk of two-stage

surgical complications and consequently additional costs. Therefore, an additional diagnostic test or combination of tests may prevent unbeneficial diagnostic hemithyroidectomies for benign nodules by ruling out malignancy, and/or prevent two-stage surgery if malignancy can be confirmed pre-operatively.

A multimodal stepwise approach using a sensitive rule-out test and a specific rule-in test might provide the most conclusive diagnosis for indeterminate thyroid nodules [25]. The ATA guidelines state that an ideal "rule-in" test would have a positive predictive value (PPV) similar to a malignant cytologic diagnosis (i.e., 98.6%), and an ideal "rule-out" test would have a negative predictive value (NPV) similar to a benign cytologic diagnosis (i.e., 96.3%) [17, 49]. Diagnostic tests on cytological samples might include molecular tests like gene mutation panels, gene or microRNA expression profiles, immunocytochemistry, and sequencing techniques. Also, several imaging modalities may be used in the workup of indeterminate thyroid nodules in vivo, including anatomical imaging techniques such as US and magnetic resonance imaging (MRI), and molecular imaging techniques such as 2-[^{18}F]fluoro-2-deoxy-D-glucose ([^{18}F]FDG) positron emission tomography (PET) combined with computed tomography (CT) and hexakis(2-methoxy-2-methylpropylisonitrile)technetium[^{99m}Tc] ([^{99m}Tc]Tc-MIBI, also known as [^{99m}Tc]Tc-sestaMIBI) scintigraphy with single photon emission computed tomography combined with CT (SPECT/CT). Performance of any diagnostic tests are usually expressed by their sensitivity, specificity, PPV, NPV, accuracy, diagnostic odds ratio (dOR), positive likelihood ratio (LR+), negative likelihood ratio (LR-), area under the receiver operating characteristic curve (AUC), and benign call rate, when available. The reader should be aware that, albeit clinically very useful, parameters like PPV, NPV but also accuracy and benign call rate are dependent on the *a priori* risk of a patient to suffer from the disease and thus can only be compared between different cohorts if the definition and prevalence of the malignancy are similar. Most studies featured in this chapter present outcome measures of a single cohort, without validation in a separate part of the dataset or (preferably) in an external cohort. When (external) validation was performed, this is specifically mentioned.

This chapter provides an overview of biomarkers obtained using conventional as well as AI-based non-invasive imaging strategies for the differentiation of thyroid nodules with indeterminate cytology. It presents the ability of a test to differentiate between benign and malignant nodules, taking into account the clinical readiness and cost-effectiveness. This chapter presents studies with different definitions of indeterminate cytology: some include Bethesda III and IV, only; others also include Bethesda V; and some have not incorporated the Bethesda System yet. The definition of indeterminate cytology will be specifically reported, when available.

Uniting medical imaging with artificial intelligence

Unlike tissue sampling procedures, medical imaging can provide information about the entire lesion, including intra- and interlesional heterogeneity [361], thereby circumventing the shortcoming of sampling error that may occur with FNAC. Visual interpretation of images consists of (qualitative) assessment of signal intensity (e.g., density, echogenicity, radiopharmaceutical uptake, apparent diffusion coefficient), location, size, shape, deformability (elastography), border (relation with surrounding tissues), patterns or vascularity (e.g., intravenous contrast enhancement, Doppler) of lesions. Medical imaging can stratify nodules before FNAC-procedures and thereby guide the choice of sampling location. Moreover, it can provide circumstantial evidence towards the nature of the nodule, such as suspicious cervical lymphadenopathy.

Quantitative imaging

Medical images contain much more information about the biology of the lesion hidden in the myriad of voxels of both lesions and healthy tissue than can be assessed visually by a human reader [362]. (Semi-)quantitative analysis of the images provides an objective complement to visual interpretation. The use of quantitative imaging in (multidevice) studies, and to a lesser extent in clinical management, requires adequate repeatability and reproducibility [363]. Repeatability refers to the likelihood of obtaining the same result in the same patient, when examined more than once on the same system. Reproducibility refers to the ability to yield the same results when that patient is examined on different systems and/or at different imaging sites. Ultimately, quantitative imaging enables the comparison of measurements in a single subject with normative values from a healthy population and permits the monitoring of subtle changes caused by the progression or remission of disease.

PET, as no other, allows for (semi-)quantitative analysis [314]. The standardised uptake value (SUV, unit [g/mL]) expresses the ratio between the local activity concentration and the decay-corrected amount of injected radiotracer per unit of body mass. It indicates the radiotracer concentration factor in a specific region compared to homogeneous distribution of the radiotracer through the body. In case of the radiotracer [^{18}F]FDG, the SUV is generally higher in malignant than in benign lesions. Nevertheless, the SUV is not only determined by tumour biology but also by preparative, procedural and postprocedural factors. The European Association of Nuclear Medicine (EANM) established guidelines for PET tumour imaging with the aim to achieve harmonisation in multicentre settings including accreditation programs (EARL) [363].

CT also allows for quantification, as attenuation coefficients of tissues are linearly transformed to Hounsfield units (HU), where a value of 0 represents the attenuation coefficient of distilled water and a value of -1000 represents the attenuation coefficient of air. However, in practice, deviations in this linearity occur. Increasing the tube voltage and with that the photon energy generally decreases the probability of interactions, i.e., attenuation and, therefore, increases penetration. Also, different scanners deliver different tube currents or photons to the subjects for a given milliamperage × seconds (mAs), as a consequence of differences in beam filtration, variances in tube potential, and rotation times [364]. Consequently, a fixed milliamperage yields different exposures, resulting in noise differences and inconsistencies in HU measurements. Other critical factors include spatial and temporal resolution, reconstruction kernel, subject positioning within the CT scanner bore, breath-holding techniques, and the (frequency of) monitoring of the CT scanner calibrations (i.e., quality control procedures). No central accreditation programmes have been ventured yet, but harmonisation has been attempted in specific applications [364].

Quantitative analysis of MRI is even more complex, due to the relative scale of the so-called weighted images. Image contrast is affected by factors intrinsic to the tissue, specific to the examination, and dependent on the hardware. Also, conventional MRI techniques lack biological specificity, i.e., different physiological and pathological substrates can produce similar changes in image contrast. MRI studies can be quantified by obtaining parametric maps of meaningful physical or chemical variables (e.g., apparent diffusion coefficient, ADC) that can be measured in physical units (mm^2/s for ADC) and compared between tissue regions and among subjects. Like for CT, only local initiatives aim to harmonise images [365, 366].

Conventional US is qualitative in nature, but quantitative US can provide specific numbers related to tissue features that can increase the specificity of image findings [367]. Qualitative bright mode (B-mode) US displays a morphological representation of the tissue, obtained from the radiofrequency data. Quantitative US, on the other hand, processes the raw radiofrequency data from tissue backscatters to characterise and distinguish phenotypic changes at a cellular level. Other US techniques like spectral-based parameterisation, elastography, shear wave imaging, flow estimation, and envelope statistics can also be performed quantitatively. However, most clinical devices do not incorporate quantitative US yet.

Artificial intelligence

Recent developments in computer science have led to advanced artificial intelligence (AI) approaches, capable of capturing the information concealed in the image in the interest of lesion or disease detection, classification and diagnosis, segmentation, image reconstruction and quantification [368]. An important breakthrough within AI was the advancement of machine

learning, the ability of a system to extract information from raw data and to learn from experience. Decision trees, random forests, and support-vector machines are well-known examples of machine learning algorithms. More recently, deep learning, which is a subset of machine learning that uses a (convolutional) neural network structure loosely inspired by the human brain, emerged, providing even more sophisticated algorithms (Figure 1) [11]. Growing amounts of data and the availability of powerful computational hardware have empowered AI, allowing computers to better represent and interpret complex data [369].

The development and, to a lesser extent, use of AI in oncology are rapidly emerging, also in thyroid cancer. Applications vary from detection of abnormalities, lesions characterisation and the prediction of treatment response [370-373]. Whereas the first AI algorithms performed simple tasks with subhuman performance, more recent algorithms sometimes surpass humans in task-specific applications.

As a result, tasks that, until a couple of years ago, could only be performed by humans, can now be executed by AI algorithms. In addition, AI algorithms have the potential to reduce variation, improve efficiency and prevent avoidable medical errors, when integrated in clinical practice as tools to assist clinicians [374]. Quantitative assessment by an algorithm reduces subjectivity that comes with visual assessment, because of the education and experience of a human reader, thereby preventing inter- and intraobserver variability [369]. In addition, a human reader can consider only a few variables at a time, quickly approaching the information processing capacity [375]. In contrast to qualitative assessment by a human reader, AI algorithms evaluate a large number of complex quantitative variables together, consistently, fast and efficiently. A major challenge of AI, however, is that the quality of a model highly depends on the input data, which is also referred to as 'garbage

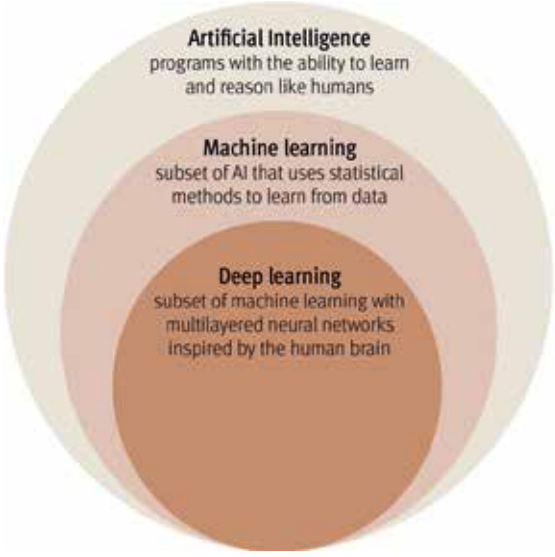

Figure 1. Differences between artificial intelligence, machine learning, and deep learning

in, garbage out'. Furthermore, AI algorithms are often considered as black boxes, since they usually lack an easy and intuitive interpretation that can be interpreted in the domain of biology or radiology [376, 377]. Explainable AI (XAI) is developed to facilitate the interpretation of data in the context of a specific application and to retrace the results on demand [376]. Moreover, AI methodology is often heterogeneous and not unambiguously reported, complicating validation of the model. Model validation is a crucial step towards clinical translation, verifying whether the model is predictive for the general target population or just for a particular subset of patients. Models must be validated using an independent test set, preferably using data from a different institution. Currently, a lack of this external validation is still one of the major limitations of AI, while replication might be of even more scientific value than original discoveries [378].

Since 2012, AI analysis of a large number of quantitative variables derived from medical images has been studied in the field of radiomics [379]. Radiomics consist of the conversion of (parts of) medical images into a high-dimensional set of quantitative features and the subsequent mining of this dataset for potential information useful for the quantification or monitoring of tumour or disease characteristics in clinical practice. The field of radiomics includes the extraction of predefined, handcrafted intensity (i.e., first order), shape and texture features combined with statistical methods or machine learning algorithms for modelling; and more recent deep learning algorithms that both learn features from raw data and perform modelling (Figure 2) [380]. To create a holistic model, in addition to the imaging features, clinical characteristics or other -omics data, like genomics, proteomics or metabolomics, are also incorporated [381]. Radiomic analysis aims to find stable and clinically relevant image-derived biomarkers for tumour characterisation, prognostic stratification and response prediction, thereby contributing to precision medicine. In this chapter, the umbrella term radiomics encompasses a broad spectrum of image analysis methods, ranging from simple AI-based methods to sophisticated deep learning algorithms.

The promises of radiomics were high. Hypothesising that medical images contained much more information than could be assessed by the human eye, radiomics was expected to contribute to medical decision making on a large scale and even to provide new insights in disease processes [362]. Yet, as for any new technology, many (technical and statistical) challenges have to be faced before reaching the goal of large-scaled implementation in clinical practice. Radiomic features are sensitive to technical variations in the different steps of the radiomic pipeline (Figure 2), hampering the reproducibility, validation and clinical translation of radiomic research. These technical variations should be as small as possible in order to attribute differences in feature values to tumour biology instead of technical variation.

Image acquisition and reconstruction largely contribute to data inhomogeneity. Radiomic analysis often consists of retrospective analysis of standard-of-care images and reanalysis of previously published cohorts, where scanners and scan protocols may vary widely between different

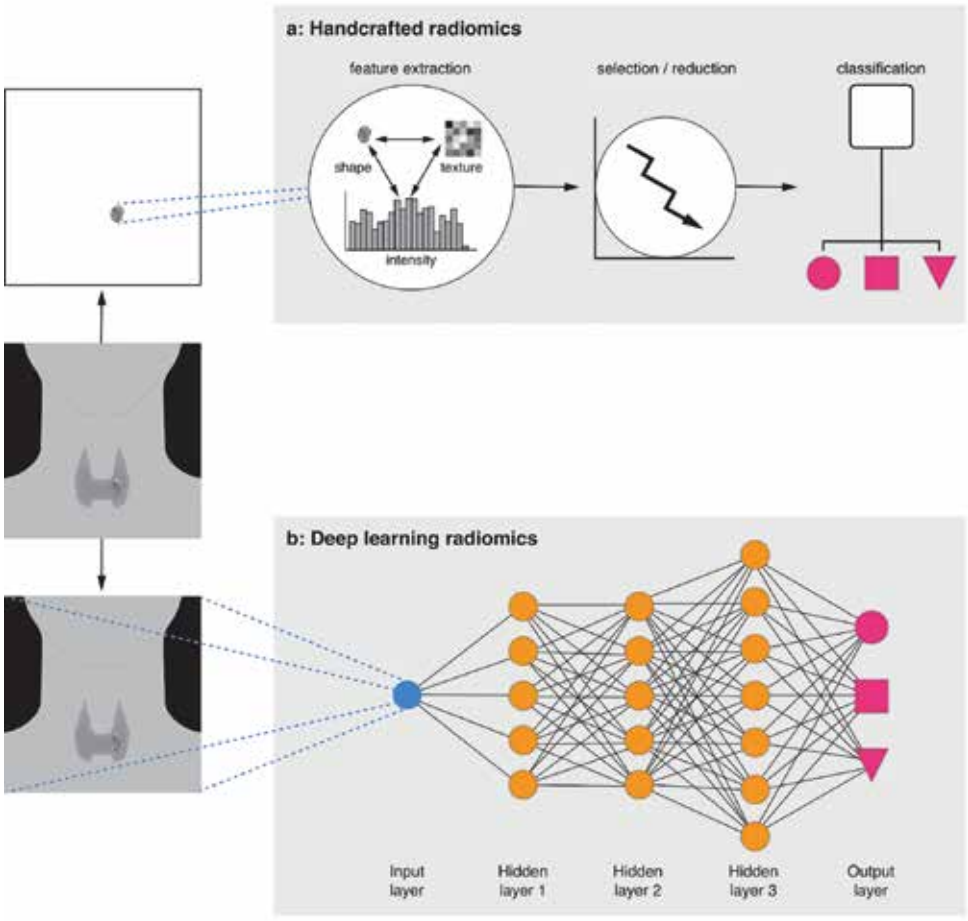

Figure 2. Handcrafted and deep learning radiomic pipeline.
A) In the handcrafted pipeline predefined features are extracted from a manually or (semi-) automatically defined volume of interest (VOI). Feature selection or dimension reduction is performed and these features are consecutively introduced in a statistical or machine learning model. B) Deep learning radiomics does not require VOI delineation, but processes the images in their raw form. The deep learning architecture consists of several hidden layers including convolutional and pooling layers, that extract increasingly complex features and perform feature selection and classification.

manufacturers and medical centres. Also, volume of interest segmentation should be performed in a standardised manner, preferably (semi-)automatically using an algorithm to reduce inter- and intraobserver variability [382]. In addition, a lack of standardisation in definition and extraction of radiomic features introduced variation. The Image Biomarker Standardisation Initiative (IBSI) made an effort to harmonise this by providing common nomenclature, mathematical definitions, benchmarks for image processing and feature extraction, and reporting guidelines [383, 384]. Similarly, repeatability and reproducibility studies have been performed to identify features that show minimal variations at different time points, under different conditions and with different feature definitions [385, 386].

Besides overcoming technical variations, another challenge of radiomics lies in the large number of features (generally over 100 features per lesion) compared to the number of subjects in a study (varying from several tens to hundreds in typical PET and CT studies, respectively). In contrast to traditional biomarker research, which is hypothesis-driven, radiomic research is of explorative nature. In explorative or data-driven research, a biological rationale of a feature representing certain disease characteristics lacks [387]. Therefore, many features are investigated, under the assumption that some features show association with underlying biology. Simultaneously, because of variation in scan protocols, it is challenging to find sufficiently large homogeneous datasets. When the number of data points (patients or scans) are small compared to the number of features, overfitting occurs, negatively impacting the generalisation performance of the radiomic model [388]. Overfitting means that the model is specifically adjusted to the training, or input, dataset, solely reflecting its noise and random fluctuations, and, consequently, it cannot be applied to other datasets, i.e., it is not generalisable. Therefore, before modelling, the number of features should be reduced using feature selection (supervised by outcome) or dimensionality reduction (unsupervised) [389]. In the modelling step, an AI algorithm may be used to fit a function to the input data and compares it with the desired output (e.g. tumour phenotype) minimising a cost-function [390]. Several (integrated) algorithms for both feature selection/dimensionality reduction and modelling are available, but no consensus on which one to use for radiomic analysis exists. The choice of the algorithm has been shown to affect the prediction performance of the radiomic model and depends on the nature of the data [361]. Many radiomic studies employ multiple AI algorithms, which comes with the risk of multiple testing and thus increasing the false-discovery rate. Multiple-modelling strategies can be justified when comprehensively documented to ensure reproducibility, and when extensively (and externally) validated [391]. In addition to external validation of the radiomic model, another strategy that contributes to clinical translation is the comparison of the performance of a radiomic model with the performance of current approaches, e.g., blood biomarkers or visual interpretation. Also, false discoveries can be minimised by, among other things, validation of the results using sham data, i.e., randomly shuffling outcome labels or using radiomic features from healthy tissue, test-retest studies, and by studying the biological rationale, or semantics, of the radiomic features in the model [392, 393].

Modalities

Ultrasonography

US is an anatomical as well as functional imaging technique that uses pulses of high-frequency (2-15 MHz) sound emitted by a transducer to capture tissue characteristics in real-time. The pulses are reflected by the tissue and returned to the transducer. The amplitude and time of the echo represent the reflection properties of specific tissue, which form the images. Conventional B-mode (for brightness) US displays the acoustic impedance of a two-dimensional cross-section of tissue, but other types capture blood flow, tissue motion, the presence of specific molecules, or the stiffness of tissue. Drawbacks of US are its limited field of view, its dependency on skilled operators, and its interobserver variability.

Conventional (B-mode) ultrasonography

US is an important step in the initial workup of thyroid nodules for its non-invasiveness, cost-effectiveness and global availability. A large body of literature has investigated the role of US in the stratification of thyroid nodules. Two meta-analyses demonstrated that, in otherwise unselected nodules, US features like composition, hypoechogenicity, microcalcification, irregular margins (i.e., infiltrative or microlobular margins), and a taller-than-wide shape are suspicious for thyroid malignancy [229, 230]. The current ATA guidelines provide a decision tree based on nodule size and other US features with an incremental suspicion for malignancy. These well-known US features are mainly characteristic of PTC, the most prevalent thyroid malignancy. FVPTC and FTC may exhibit other characteristics and may be less easily diagnosed using this decision tree [17, 231, 232]. In an unselected population, no US feature alone is sensitive or specific enough to accurately identify malignancies, but combinations of features might provide new insights [229].

The use of US in thyroid nodules with indeterminate cytology is less widely studied. Both previously mentioned meta-analyses briefly discussed its value in indeterminate nodules (Bethesda System was not taken into account) [229, 230]. As FTC has a higher prevalence in indeterminate nodules, US using the classic characteristics is less accurate in indeterminate nodules than in unselected thyroid nodules, generally demonstrating limited sensitivity. Only solid nodules, in contrast to partially cystic nodules, demonstrated sensitivities above 90% (range: 46% to 100%) [25]. The features taller-than-wide shape, presence of irregular margins, presence of microcalcifications, and nodule diameter larger than 4 cm were promising, with specificities ranging from 72% to 99%, 65% to 100%, 36% to 100%, and 69 to 94%, respectively [25]. Remonti et al. presented an increased central vascularisation as the best predictor for malignancy, with a specificity of 96%, but other studies showed extremely poor specificities, ranging from 0% to 100% [230].

TI-RADS

Since 2009, several US-based risk stratification systems to identify nodules that warrant biopsy or sonographic follow-up have been proposed. Following the BI-RADS classification system that is widely used in breast imaging, the American College of Radiology (ACR) presented the TI-RADS (for Thyroid Imaging, Reporting and Data System). TI-RADS aims to (1) provide recommendations for reporting incidental thyroid nodules, (2) develop a set of standard terms (lexicon) for US reporting, and (3) propose a TI-RADS risk stratification system on the basis of the lexicon [394]. The ACR TI-RADS scores the composition, echogenicity, shape, margin, and echogenic foci of a thyroid nodule, all consisting of 0 up to 3 points. The total number of points determines whether a nodule is considered benign (TR1, 0 points), not suspicious (TR2, 2 points), mildly suspicious (TR3, 3 points), moderately suspicious (TR4, 4 to 6 points), or highly suspicious (TR5, ≥7 points) and also guides the decision to perform FNAC or follow-up: no FNAC or follow-up (TR1-2), FNAC if nodule maximum diameter (ø)≥2.5 cm and follow-up if ø≥1.5 cm (TR3), FNAC if ø≥1.5 cm and follow-up if ø≥1.0 cm (TR4) or FNAC if ø≥1 cm and follow-up if ø≥0.5 cm (TR5).

In addition to ACR TI-RADS, the European Thyroid Association and the Korean Society of Thyroid Radiology/Korean Thyroid Association developed similar US risk stratification systems; the EU-TI-RADS and K-TI-RADS, respectively [395, 396]. Also, the ATA and the American Association of Clinical Endocrinologists/American College of Endocrinology/Associazione Medici Endocrinologi propose US risk stratification systems [17, 397]. An international survey investigating the utilisation of all five aforementioned risk stratification systems with 875 respondents in 52 countries demonstrated that almost one third of respondents used more than one risk stratification system in their practice, potentially leading to confusion [398]. Grani et al. compared the risk stratification systems in 477 patients and found that the systems vary widely in their ability to reduce the number of unnecessary thyroid nodule FNACs (17.1 up to 53.4%) [399]. The ACR TI-RADS outperformed the others, classifying more than half of the biopsies as unnecessary with a false-negative rate of 2.2%. The remainder of this chapter focuses on the ACR TI-RADS.
Over recent years, TI-RADS has become fully incorporated in the management of thyroid nodules [17]. As FNAC may be more systematically withheld for patients with a presumed benign nodule with TI-RADS 1, 2, and most 3, the patient population that is selected for additional diagnostic tests has potentially changed. Many of the studies on additional diagnostics have not incorporated TI-RADS yet, and it is currently unclear to say what effect the introduction of TI-RADS may have on the diagnostic accuracy and therapeutic yield of other tests. A different population, or reference class, with a larger proportion of malignant nodules, impacts the PPV and NPV, but also the sensitivity and specificity.

Stratification of cytologically indeterminate nodules according to the risk of malignancy as determined by combining the TI-RADS and Bethesda system might be of interest, notwithstanding a limited body of evidence comprising of small retrospective cohorts. Larcher de Almeida et al. investigated

the risk of malignancy in indeterminate thyroid nodules by combining the ATA classification with cytological subcategorisation (nuclear atypia, architectural atypia, oncocytic atypia) [400]. They found that the risk of malignancy reached almost 80% when both nuclear atypia and ATA-based high-risk US features are present. The presence of these cytological features also increased the risk of malignancy in the ATA-based intermediate-risk category. Architectural atypia and oncocytic patterns were not independently related to higher cancer risk. Moreover, a recent meta-analysis by Staibano et al. including 17 studies investigated sonographic risk criteria (ACR TI-RADS, EU TI-RADS, K-TI-RADS, or ATA) for further prognostication of Bethesda III and IV nodules [401]. In both Bethesda III and Bethesda IV nodules separately, ATA had the highest pooled specificity of 90% and 94% (sensitivity of 52% and 15%), while K-TI-RADS had the highest pooled sensitivity of 78% and 91% (specificity of 53% and 40%), respectively. EU-TI-RADS does not contribute to the clinical management of patients with cytologically indeterminate Hürthle cell nodules, particularly those classified as Bethesda IV [402]. These results underline the combination of cytological subcategorisation and US risk stratification in the management of indeterminate nodules. A conservative approach is proposed in nodules with low-risk US suspicion and Bethesda III, while additional diagnostics and surgery should be considered for nodules with high-risk US suspicion and Bethesda IV or V [276, 403].

AI has been investigated for the optimisation of the ACR TI-RADS risk stratification. Wildman-Tobriner et al. developed AI TI-RADS as a simplification of ACR TI-RADS in unselected nodules, where six features were assigned zero points, using a genetic algorithm inspired by natural selection and its genetic underpinnings [404]. The model was trained using 1,325 nodules and validated using 100 nodules, resulting in similar AUCs for ACR TI-RADS and AI TI-RADS of 91% and 93%, respectively. Specificity of AI TI-RADS (65%) was higher than that of ACR TI-RADS (47%).

US radiomics is gaining interest in thyroid nodules. Yoon et al. built a US radiomic score for the differentiation of benign and malignant lesions, retrospectively including 155 nodules with Bethesda III and IV indeterminate cytology [405]. Seven hundred thirty radiomic features were extracted from a square region of interest delineated on a representative 2D image of the initial US. A radiomic score incorporating fifteen radiomic features combined with clinical variables (nodule size, gender, age, Bethesda category) performed significantly better than a model composed of clinical variables only with cross-validated AUCs of 84% and 58%, respectively. Major limitations of this study are the use of clinical US images instead of quantitative images and the choice of a representative image by a human reader. Although inherent to US imaging, bias is introduced by the implicit radiologist input in the selection of the 2D slices as described as the Clever Hans effect by Wallis et al. in a widely used MRI dataset [406].

In addition, in unselected nodules, US radiomic analysis has been extensively studied for the differentiation of benign and malignant nodules. A recent meta-analysis by Cleere et al. including 75 studies found a pooled sensitivity of 87% and a pooled specificity of 84%, which indicates that,

for some patients, the use of radiomics could possibly circumvent the need for FNAC and surgical resection [407]. For deep learning radiomics using convolutional neural networks (CNN), the pooled sensitivity and specificity were 85% and 82%, significantly lower than for studies using non-CNN (sensitivity: 90%, specificity: 88%), which might be due to a larger required sample size for a deep learning radiomic study (at least 800) compared to a handcrafted radiomic study (around 100) [392]. The meta-analysis only touches upon the heterogeneous methodology of included studies, stating the broad spectrum of analysis methods and interobserver variability of US. Notwithstanding, radiomic features extracted from US images are impacted by the slice variability and pre-processing [408]. To improve feature repeatability, the use of intensity standardisation with outlier removal applied to the region of interest and a fixed bin size grey-level discretisation could be performed and these and other pre-processing steps should be extensively documented [408]. When standardisation of the radiomic methodology is performed and US radiomics is validated in large prospective cohorts, it has the potential to become a non-invasive and cost-effective diagnostic tool in (cytologically indeterminate) thyroid nodules.

Elastosonography

One of the key features during palpation of thyroid nodules is the degree of firmness; malignant nodules tend to be firmer than benign ones. Palpation, however, is highly subjective and depends on the size and location of the nodule and on the skill of the practitioner. Elastosonography, a dynamic US technique that is used to evaluate the biomechanical viscoelastic properties of tissue, provides a quantitative method to measure tissue firmness or elasticity. Lyshchik et al. were the first to practice elastosonography for the evaluation of the elasticity of thyroid nodules, measuring the tissue distortion while applying a standardised dosed external force by the US transducer [281]. Elastosonography methodology is diverse, but it follows the principle of estimating displacements fields in tissue using correlation techniques that track the echo delays in waveforms recorded before and after the quasistatic compression. Qualitative evaluation of the thyroid elasticity is performed by repeated manual compression (also known as strain elastosonography), taking into account the amount of compression and different zones of interest (i.e., healthy tissue should be included in the measurement, which might be complicated in the presence of thyroid diseases or large nodules) [409]. Alternatively, and circumventing the problem that the mechanical compression force applied to the tissue cannot be measured accurately and thus the absolute tissue strain cannot be calculated, shear wave elastosonography has been developed. This technique evaluates tissue stiffness through focused pulses of US instead of mechanical compression [410]. This acoustic force causes horizontal displacements in the tissue, which are called shear waves. These shear waves contain quantitative data about the elastic properties of the tissue that can be measured in propagation speeds of these sheer wave (m/s) or nodule stiffness (kilopascals). It has the advantages of being more objective, having a higher reproducibility, and having decreased operator dependence.

A colour-coded image superimposed on the greyscale B-mode US images is generated, with colours in the red spectrum representing soft tissues and colours in the blue spectrum representing firm tissues. (Semi)quantitative analysis uses numerical values that correspond to the deformation ratios (strain) or stiffness (sheer wave), scored according to several systems.

A meta-analysis by Nell et al. based on 20 qualitative elastosonography studies concluded that elastosonography could accurately diagnose benign nodules with both a pooled sensitivity and a NPV of 99%, thereby safely dismissing FNAC, on condition that only completely soft nodules are classified as benign (benign call rate 14%) [286]. The role of elastosonography in the preoperative workup of cytologically indeterminate thyroid nodules is limited. Qualitative US based on colour-scales has insufficient sensitivity and specificity and semiquantitative methods lack validation. A meta-analysis including 20 studies on both qualitative and quantitative techniques and a total of 1.734 indeterminate thyroid nodules reported an overall pooled sensitivity and specificity of 77% and 87%, respectively, with similar diagnostic accuracies for real-time, shear wave and strain ratio elastosonography [411]. The power of the available evidence is negatively impacted by methodological heterogeneity in imaging techniques, image processing, and elasticity scoring methods across studies. Yet, the suggested rule-out capacity of qualitative elastosonography when only completely soft nodules are included is worth validating in indeterminate thyroid nodules, for its easy implementation and potential low costs. Elastosonography can be performed during regular thyroid US with the same equipment, prolonging the procedure only five minutes. To the best of our knowledge, no cost-effectiveness studies have been carried out in indeterminate thyroid nodules, possibly limited by the heterogeneous methodology.

Computed tomography

CT is a 3D anatomical imaging technique that reflects X-ray attenuation by different tissues. CT scanners use a rotating X-ray tube and an oppositely placed row of detectors placed in the gantry to measure X-ray photon attenuations, which are reconstructed into tomographic images. Contrast-enhancement by iodine-based intravenous contrast may be performed to highlight structures such as blood vessels that otherwise would be difficult to distinguish from their surroundings on native-phase CTs, to obtain functional (perfusion) information about tissues and to improve soft-tissue contrast. However, the usage of iodinated intravenous contrast media is relatively contra-indicated as a post-thyroidectomy radioiodine (^{131}I) ablation dose might be indicated in patients suffering from differentiated thyroid cancer. The effectiveness of radioiodine therapy might decrease by recent use of high doses of iodine and an interval of at least one month between iodinated contrast and radioiodine is recommended [412].

CT has not been investigated in thyroid nodules specifically with indeterminate cytology. In unselected nodules, some studies have been performed. Lee et al. found no significant differences between benign and malignant lesions in number of lesions, lesion size, presence of calcifications, lesion consistency, and lesion attenuation on CT in a dataset of 109 nodules (100 benign, 9 malignant) [413]. Another study in PTC found that CT was inferior to US for the evaluation of thyroid nodules [414]. More recently, AI has been investigated for CT lesion characterisation. Peng et al. investigated first order features for the identification of malignant nodules (N=50), benign nodules (N=84), and healthy controls (N=150), resulting in a sensitivity, specificity, PPV, NPV, and accuracy of 82%, 93%, 92%, 85%, 95%, and 88%, respectively [415]. It should be noted that results have not been validated using a test set. Li et al. developed a deep learning model for automatic recognition and classification of thyroid nodules on iodine contrast-enhanced CTs [416]. The model was trained in a dataset of 786 nodules (543 benign and 243 malignant) and validated in a test set of 137 nodules (103 benign and 34 malignant), resulting in an accuracy of 85%. There is a large class imbalance between benign and malignant nodules, which might have affected the accuracy, but authors state that this was corrected for using class weights.

The role of CT in the preoperative differentiation of thyroid nodules is limited compared to other imaging techniques. Yet, since CT is an important source of thyroid incidentalomas (incidence: 15% [417]), computer aided detection systems to automatically recognise and classify thyroid incidentalomas on CT might be of interest.

MRI

Magnetic resonance imaging (MRI) is a 3D anatomical as well as functional imaging technique based on nuclear magnetic resonance [418]. MRI scanners use strong magnetic fields, magnetic field gradients, and radiofrequency waves to generate images of the organs in the body, with improved soft-tissue contrast compared to (contrast enhanced) CT. Protons (hydrogen atoms) in body tissue that contain water, give off a signal that can be processed into an image. First, a pulse of electromagnetic radiation is used to excite nuclei of atoms in the magnetic field with exactly the right resonance frequency. The excited nuclear spins of the hydrogen nucleus undergo relaxation to the ground state while emitting radiofrequency waves, which are measured with a receiving coil. The contrast between different tissues is determined by the speed at which the nuclear spin of excited nuclei returns to the ground state. Since different tissues have different hydrogen densities, details of the anatomy can then be observed.

Different tissue properties can be measured using different pulse sequences of pulsed magnetic field gradients, radiofrequency pulses, intervals between delivery of successive pulses, between pulse delivery and receipt of the echo signal, etc. Intravenous contrast, mostly by paramagnetic

substances containing gadolinium, enhance relaxation of the excited nuclear spins and thus add information about tissue perfusion. MRI-imaging is less widely available, more complex, lengthy and costly than US and CT, but provides unsurpassable soft-tissue contrast without the use of ionizing radiation.

The classic spin and gradient echo sequences resulting in T1-, T2-, proton density- and susceptibility-weighted sequences seem to have limited classification value in indeterminate thyroid nodules. Effective T2-mapping (T2*-mapping) was explored by Shi et al. in 28 patients with thyroid nodules of different cytological subclasses, subjected to (therapeutic and diagnostic) surgery, describing 100% specificity and 84-90% sensitivity to distinguish malignant and benign thyroid nodules [419]. The used of dynamic gadolinium contrast-enhanced MRI has found conflicting results [420, 421]. A much larger body of evidence has been found for diffusion-weighted MRI (DWI) and proton-magnetic resonance spectroscopy (MRS).

Diffusion weighted magnetic resonance imaging

DWI is a specific form of MR imaging that is sensitive to the random Brownian motion of water molecules within a voxel of tissue. The easier water molecules diffuse and move around in a region, the higher the isotropic signal will be at higher degrees of diffusion weighting (b-value). Apparent Diffusion Constant (ADC [mm^2/s]) imaging results from a series of conventional DWI-sequences with different b-values. The change in signal is proportional to the rate of diffusion. An ADC-image thus is an MRI-image that more specifically shows diffusion than conventional DWI, by eliminating the T2-weighting that is otherwise inherent to conventional DWI. Contrary to DWI, the standard greyscale of ADC-images is to represent a smaller magnitude of diffusion as darker. Generally, highly cellular tissues or those with cellular swelling exhibit lower ADC-values.

The use of DWI in cytologically indeterminate thyroid nodules is limited and methodology varies largely. Nakahira et al. evaluated the role of the ADC in 42 nodules, including 15 (36%) with indeterminate cytology (Bethesda System was not taken into account) [319]. The final diagnosis was confirmed by surgery and mean ADCs (acquired with b-values of 0 and 1,000 s/mm^2) were compared between benign and malignant nodules (all with indeterminate cytology). Malignant nodules showed significantly lower ADCs than benign nodules. For all nodules, a cut-off value for malignant nodules of 1.60×10^{-3} mm^2/s yielded a sensitivity, specificity, and accuracy of 95%, 83%, and 88%, respectively. It was concluded that ADC measurements could potentially quantitatively differentiate between benign and malignant thyroid nodules, even those of indeterminate cytology [319]. Chung et al. investigated the value of histogram analysis of ADC maps in the differentiation of follicular thyroid carcinoma from follicular adenoma in 17 Bethesda III and IV indeterminate nodules on US-guided core needle biopsy [422]. Histogram parameters were derived from ADC values (acquired with b-values of 0 and 800 s/mm^2) obtained from the entire tumour volume and compared with the histopathological diagnosis. It was found that 10^{th}, 25^{th}, and 50^{th} percentiles of the ADC values

were all significantly lower in follicular adenoma than in follicular thyroid carcinoma. ROC curve analysis revealed that the 25th percentile resulted in the highest AUC of 87%, with an optimal cut-off value of 0.353 × 10^{-3} mm^2/s. A lower ADC value in follicular adenoma compared to follicular carcinoma seems contradictory with results of Nakahira et al., where lower ADCs were found in malignant nodules. The probable reason for this is that Nakahira et al. predominantly included PTC with histological characteristics of calcifications and desmoplastic reactions, which cause restriction of free water movement in the cellular environment and reduce ADC values, whereas follicular neoplasms including Hürthle cell nodules are known for their varying colloid tissue involvement and thus histologically contain more fluid. Thus, DW-MRI would inaccurately provide a more benign image [328, 422].

DWI has been more extensively investigated in the differentiation between benign and malignant unselected thyroid nodules. A 2014 meta-analysis of Wu et al. summarised seven studies with 358 subjects and presented a pooled sensitivity of 91%; a specificity of 93%, a LR+ of 12.24; a LR- of 0.99; a diagnostic OR of 123.78; and an AUC of the summary ROC of 94% [423]. In 2016, a meta-analysis by Chen et al. summarised 15 studies with 765 lesions and presented a pooled sensitivity of 90%; a specificity of 95%; a LR+ of 16.49; a LR- of 0.11; a diagnostic OR of 150.73; and an AUC of the summary ROC of 95% [329]. Most studies showed a significantly lower mean ADC value in malignant lesions compared to benign lesions, because of larger nuclei, denser stroma and higher cell counts, all of which led to increased cellularity and reduced extracellular space. However, no absolute cut-off was found. This could be attributed to heterogeneous methodology such as varying b-values and differences in ADC measurements. Other explanations could be a diversity in patient population or components with high diffusivity in malignant lesions, like cystic components, central necrosis, or intratumoural haemorrhage.

DWI seems a promising non-invasive, non-radiative and accurate technique for the pre-operative differentiation of (cytologically indeterminate) thyroid nodules. Nevertheless, while the worldwide availability of MRI scanners is growing, MRI is still considered an expensive technique in terms of hardware, overhead costs, and the relatively long scan duration. Large-scale trials are necessary to assess and validate its clinical value, to establish harmonisation in methodology, to determine cut-off values, and to study cost-effectiveness, specifically in FNAC indeterminate thyroid nodules.

Magnetic resonance spectroscopy

Magnetic resonance spectroscopy (MRS) is an analytical method used for the in vivo chemical characterisation of tissue, measuring the presence and concentration of various metabolites. Magnetic resonance principles are used to detect various nuclei, such as hydrogen-1 (^1H), which all can provide valuable metabolic and physiological information [424]. ^1H-MRS is able to capture the metabolic profile of a lesion, by determination of the relative concentrations and physical properties of a variety of biochemicals. These include several low molecular weight metabolites such as choline, creatine, glutamate, lactate, and different amino acids. Spectroscopy uses the chemical

shift of a nucleus to observe, identify and quantify biologically important compounds in tissue. An anatomical MR image is acquired, on which a volume of interest is selected, and the MR spectrum is acquired. As protons in water are far more abundant than the metabolites of interest (10^4:1), the water signal should be suppressed during MRS-pulse sequences.

The use of MRS specifically in indeterminate thyroid nodules was rather limited. Therefore, we focused on MRS in the differentiation of thyroid carcinoma in general. The use of magnetic resonance principles in thyroid cancer is in fact not new, but originates from ex vivo proteomic and metabolomic research [425]. MRS of cytology and biopsy specimens was attempted to overcome the limitations of FNAC [426, 427]. Also, ex vivo operative specimens have been analysed, for the identification of the morphologic features of malignancies in the first place; with the advancement of the technology followed by the differentiation between benign and malignant neoplasms [428]. Ex vivo spectra showed lower content of lipids and higher concentrations of amino acids in malignant compared to benign nodules [429].

The first in vivo study by King et al. succeeded in discriminating thyroid carcinomas from normal thyroid tissue based on the 1.5 Tesla ^{1}H-MRS spectra [430]. In their cohort of eight patients (three anaplastic carcinomas, two papillary carcinomas, one follicular carcinoma) and five healthy controls, they found that choline-to-creatine ratio seemed a useful marker for the pre-operative differentiation. This was confirmed by other studies, showing that a choline-peak was rather specific for malignancies [431, 432]. It should be noted that these studies considered the absolute choline peak, without the creatine reference. Creatine is considered a convenient internal standard, for its relatively constant level in metabolically active tissues. More recently, the choline-to-creatine ratio was further evaluated by Aghaghazvini et al. in a cohort of 9 malignant (7 papillary, 2 follicular) and 23 benign nodules using 3 Tesla ^{1}H-MRS [433]. At an echo time of 136 ms, a choline-to-creatine ratio of 2.5 corresponded best with histopathology with a sensitivity, specificity, PPV and NPV of 75%, 100%, 100%, and 92%, respectively.

Whereas the MRS choline-to-creatine ratio seems a promising biomarker for the differentiation of thyroid nodules, all presented studies were performed in small cohorts and with varying methodology. To the best of our knowledge, only four papers on in vivo MRS in thyroid nodules were published in almost two decades, which might indicate limited clinical interest or limited feasibility. In contrast, the field of MRS is emerging and in recent years, the use of MRS in clinical practice has increased, because of the installation of human MRI systems with high field strengths (≥7 Tesla). Higher field strengths result in spectral dispersion, i.e., a larger frequency between peaks, improving the resolution, which allows more accurate quantification of tissue compounds in smaller lesions [434]. Future studies should validate these preliminary findings and also cast light on cost-effectiveness.

Multiparametric MRI

Both DWI and ^{1}H-MRS seem promising in the evaluation of indeterminate thyroid nodules, but a multiparametric approach has not been extensively studied. Aydin et al. were the first to describe multiparametric MRI, being DW-MRI and ^{1}H-MRS, in unselected thyroid nodules [435]. Other approaches combined ADC values and descriptions of the time-signal intensity curves of dynamic contrast enhanced MRI, finding an accuracy of 91% [436], used ADC values and T1-weighted and T2-weighted tumour-to-non tumour ratios with accuracies over 90% [437] or compared DWI with proton transfer imaging and found DWI to be superior [438]. Wang et al. investigated conventional MRI, DWI, and DCE in a retrospective cohort of 181 consecutive subjects (148 benign and 111 unstratified malignant nodules, confirmed by pathological results) [439]. The multivariable analysis revealed that ADC value, irregular shape, ring sign in the delayed phase, and cystic degeneration were independent predictors of malignancy, with an AUC, sensitivity, and specificity of these variables combined of 99%, 97%, and 95%, respectively. Song et al. investigated intravoxel incoherent motion MRI and DCE in 38 unstratified nodules and found that parameters were significantly different between benign and malignant nodules [440].

Multiparametric radiomics has also been investigated in a dataset of 120 PTC patients to predict aggressiveness based on 1393 radiomic features extracted from T1-weighted, T2-weighted, and ADC-images. The dataset was split into a training (N=96) and test set (N=24) and machine learning was performed for feature selection and classification, resulting in an AUC in the test set of 92% (compared to 56% for clinical characteristics alone) [441]. Another T1, T2, and ADC radiomics approach in 132 PTC (92 training, 40 test) used a machine learning algorithm to detect extrathyroidal extension, resulting in an AUC in the test set of 87% [442].

[^{99m}Tc]Tc-MIBI scintigraphy

Scintigraphy is a 2D functional imaging technique that in vivo localises gamma-emitting isotopes such as technetium-99m using a gamma camera. Scintigraphy with the radiopharmaceutical [^{99m}Tc]Tc-MIBI reflects perfusion and the number of active mitochondria in cells [293]. It is primarily known for its use in myocardial perfusion imaging, the evaluation of hyperparathyroidism, and molecular breast imaging. [^{99m}Tc]Tc-MIBI scintigraphy has been investigated for the differentiation between benign and malignant nodules based on the uptake and by assessing an eventual increase in uptake within the nodule over time. [^{99m}Tc]Tc-MIBI is more suitable than [^{99m}Tc]pertechnetate or radioisotopes of iodine (^{123}I, ^{124}I (PET), ^{131}I). Iodine radioisotopes are often used to assess thyroid nodule functioning (“hot” or “cold”), which are unspecific and ineffective for the further stratification of cytologically indeterminate thyroid nodules. Whereas malignant nodules are almost solely cold, as cell dedifferentiation results in a decrease of the sodium-iodide symporter and thereby

lower [^{99m}Tc]pertechnetate or radioiodine uptake, benign nodules can be hot as well as cold, while far outnumbering carcinomas. [^{99m}Tc]Tc-MIBI uptake is independent of iodine trapping and organification in the thyrocytes.

Increased [^{99m}Tc]Tc-MIBI uptake and late retention are often observed in malignant nodules [351]. A meta-analysis from 2013 by Treglia et al. showed 82% sensitivity and 63% specificity for [^{99m}Tc]Tc-MIBI scintigraphy in clinically suspicious, hypofunctioning, cytologically unselected thyroid nodules [293]. Three studies examined [^{99m}Tc]Tc-MIBI scintigraphy in cytologically indeterminate nodules (Bethesda III/IV/pre-Bethesda) [351, 443-447], based on an early image between 10 and 20 minutes post injection and a delayed image between 60 to 120 minutes post injection. The uptake in the nodule and retention on the delayed image were assessed and compared with physiological washout in normal thyroid tissue. Nodules with increased uptake on early images that persisted or increased on the delayed images were suspicious for malignancy. The sensitivities and specificities for visual interpretation ranged from 56% to 96% and from 20% to 95%, respectively. Semi-quantitative analysis was performed using the retention index (RI, percentage MIBI uptake reduction in a nodule between the early and the late image, corrected for uptake in the contralateral lobe) and the wash-out index (WO_{Ind}, percentage MIBI uptake reduction in a nodule between the early and the late image, corrected for uptake in tissue outside the thyroid). At the cut-off determined, for the RI, the sensitivities were 100% and the specificities ranged from 57% to 90% [443, 446]. For the WO_{Ind}, sensitivities were 100% and specificities ranged from 89% to 100% [351, 446, 447]. A recent retrospective multicentre study by Schenke et al. including 365 hypofunctioning Bethesda III and IV nodules in 12 European centres concluded that negative [^{99m}Tc]Tc-MIBI result on visual evaluation is an effective tool to rule-out thyroid malignancy in 18% of negative nodules [447]. Semi-quantitative image analysis may considerably improve the overall diagnostic performance at an optimal WO_{Ind} cut-off of -19%, with a sensitivity, specificity, PPV, NPV, accuracy, and benign call rate of 100%, 89%, 82%, 100%, 93%, and 61% respectively. These findings cannot be extrapolated to all patients with indeterminate cytology, since preselection of intermediate- or high-risk nodules by EU-TI-RADS and the exclusion of hyperfunctioning nodules, probably by thyroid scintigraphy, is required.

Planar gamma camera imaging is globally widely available. Also, the tracer [^{99m}Tc]Tc-MIBI can be easily complexed using MIBI-kits and an on-site molybdenum-technetium generator. The average costs of [^{99m}Tc]Tc-MIBI scintigraphy range from €119 to €500 in Europe and from $669 to $1156 in the United States [448]. A cost-effectiveness study from 2014 found that [^{99m}Tc]Tc-MIBI-based management was more cost-effective than Afirma® gene expression classifier testing from a German perspective [448], but modelled costs for [^{99m}Tc]Tc-MIBI scintigraphy and thyroid surgery were likely underestimated and performance parameters were extrapolated from unselected nodules. Disadvantages of [^{99m}Tc]Tc-MIBI scintigraphy include the limited spatial resolution of the gamma camera, which limits the test sensitivity in lesions smaller than 30 mm, and the radiation burden is 2 to 6 millisievert for an adult male (20-30% higher for females) [449].

[^{18}F]FDG PET/CT

PET/CT is a 3D functional imaging technique that in vivo localises and quantifies positron-emitting isotopes such as fluorine-18. PET imaging with the non-metabolisable glucose analogue [^{18}F]FDG reflects glucose metabolism [314]. After intravenous injection, [^{18}F]FDG is, like D-glucose, taken up in eucaryotic cells by the membrane-bound sodium-dependent glucose transporters' (GLUT) family. In the cytosol it is phosphorylated to [^{18}F]FDG-6-phosphate by members of the hexokinase family. As phosphoglucose isomerase, the enzyme responsible for the second step in the glycolytic pathway, does not interact with deoxyglucose, [^{18}F]FDG cannot be degraded further. Moreover, as most mammalian cells lack the enzyme to dephosphorylate [^{18}F]FDG-6-phosphate, it accumulates in the cells, the rate dependent on perfusion, GLUT-capacity and hexokinase activity. Many pathological conditions cause regional alterations in glucose metabolism in tissues, through which [^{18}F]FDG PET/CT is an important tool in detection and staging of cancer and active inflammations. [^{18}F]PET/CT is an imaging technique using ionising radiation with high sensitivity, but limited specificity due to other causes of [^{18}F]FDG-uptake: coincidental findings may lead to further invasive diagnostics. [^{18}F]FDG PET/CT is an important source of thyroid incidentalomas (i.e., unexpected incidental findings during an [^{18}F]FDG PET/CT for other indications), with a pooled incidence of around 2.5% [450] and a rate of malignancy of around 30% [451]. These incidentalomas require additional workup by FNAC when their diameter exceeds 1 cm [452].

[^{18}F]FDG PET/CT has a limited role in the management of thyroid cancer. Only when radioiodine refractory disease is suspected, [^{18}F]FDG PET/CT plays an important role in disease monitoring [17]. In radioiodine refractory disease, differentiated thyroid carcinomas lost the capacity to concentrate radioiodine, but still have measurable thyroglobulin serum values as sign of vital residual disease. [^{18}F]FDG PET/CT is also utilised for the initial staging of poorly differentiated or invasive Hürthle cell carcinoma. Similarly, [^{18}F]FDG PET/CT plays an important role in staging of undifferentiated forms of thyroid cancer such as anaplastic thyroid cancer.

[^{18}F]FDG PET/CT is mentioned but not routinely advised in the current ATA guidelines for the management of thyroid nodules with indeterminate cytology, despite a growing body of evidence. The first prospective study by Kresnik et al. dates from 2003 and evaluated the usefulness of [^{18}F]FDG PET/CT in the preoperative assessment of 43 suspicious thyroid nodules with suggestive cytologic results (pre-Bethesda) [453]. They found that thyroid carcinomas, in contrast to most benign thyroid nodules, demonstrate significantly increased glucose metabolism; at a cut-off value of the SUV of 2 g/mL, a 100% sensitivity, 63% specificity, and 100% negative predictive value was reached. However, the study Kresnik et al. did not represent the general population, because the study was performed in an area of endemic goiter and patients with papillary carcinoma were selected as positive control group. Subsequently, De Geus-Oei et al. investigated [^{18}F]FDG PET/CT in 44 patients

with indeterminate cytology, defined as inconclusive fine-needle aspiration biopsy (pre-Bethesda), who subsequently underwent diagnostic hemithyroidectomy [37]. They demonstrated that a negative [^{18}F]FDG PET/CT could theoretically reduce the number of futile hemithyroidectomies by 66% at a NPV of 100%. A subsequent meta-analysis from 2011 by Vriens et al., including six studies, presented a pooled sensitivity of 95% and a pooled specificity of 48%, resulting in a NPV and PPV of 96% and 39% respectively (benign call rate: 37%) [304]. In 2017, a review by De Koster et al. reported sensitivities and specificities of [^{18}F]FDG PET/CT to detect thyroid carcinoma in indeterminate thyroid nodules ranging from 77% to 100% and 33% to 64%, respectively [25].

These findings were recently validated in a recent multicentric diagnostic randomised controlled trial that assessed the impact of [^{18}F]FDG PET/CT in the management of thyroid nodules with double-read Bethesda III or IV cytology to rule out malignancy, avoid futile diagnostic surgeries, and improve patient outcomes (EfFECTS trial) [454]. De Koster et al. randomised 132 patients with an indeterminate nodule who were scheduled for diagnostic surgery and underwent an [^{18}F]FDG PET/CT scan into a PET/CT-driven arm or a diagnostic surgery arm. In the PET/CT-driven arm, diagnostic surgery was advised in visually [^{18}F]FDG-positive nodules and active surveillance in [^{18}F]FDG-negative nodules. In the diagnostic surgery arm, all patients were advised to continue with the scheduled diagnostic surgery. Patient management was considered unbeneficial (i.e., diagnostic surgery for benign nodules or active surveillance for malignant/borderline nodules) in 42% of patients in the [^{18}F]FDG PET/CT-driven arm and 83% in the diagnostic surgery arm. No wrongful active surveillance for malignant/borderline nodules was reported. As such, [^{18}F]FDG PET/CT-driven management avoided 40% of diagnostic surgeries for benign nodules. Therapeutic yield was the highest (48% reduction in diagnostic surgeries) when only non-Hürthle cell nodules were considered, as nearly all Hürthle cell nodules were [^{18}F]FDG-positive on visual interpretation.

Several studies have reported the quantitative assessment of [^{18}F]FDG PET/CT images using the SUV of the indeterminate thyroid nodule, with a higher SUV_{max} reported in thyroid malignancies than in benign lesions [40, 309, 444, 455-458]. Nevertheless, major variations in SUV cut-offs and diagnostic accuracy are found between studies. Deandreis et al. and Rosario et al. respectively included 56 indeterminate nodules (pre-Bethesda) and 63 Bethesda III/IV nodules and showed that a SUV_{max} cut-off of at least 5 g/mL was 91% specific to detect thyroid carcinoma, NIFTP, and FT-UMP [309, 456]. This was substantiated by Piccardo et al. in 111 indeterminate nodules, but no AUCs or corresponding sensitivity and specificity were reported [40]. Contrarily, Merten et al. demonstrated that a cut-off of 5 g/mL was only 41% specific but 80% sensitive in their study in 51 Bethesda IV nodules [458]. Pathak et al. reported a SUV_{max} cut-off of 3.25 g/mL best differentiated 42 non-Hürthle cell nodules with 79% sensitivity and 83% specificity [457]. An additional analysis of the EfFECTS trial dataset assessed the added value of SUV metrics, SUV-ratios (node to contralateral normal thyroid) and radiomics for the preoperative differentiation [455]. None of these previous studies used ROC curve analysis to determine SUV cut-offs that corresponded to optimal test sensitivity, i.e.,

a NPV similar to a benign cytologic diagnosis (i.e., 96.3%) as per the ATA recommendations for a useful rule-out test [17]. De Koster et al. performed quantitative analysis and ROC curve analysis of the EfFECTS dataset, including 123 patients who underwent [^{18}F]FDG PET/CT according to the EANM guidelines [455]. Quantitative [^{18}F]FDG PET/CT assessment ruled out malignancy in indeterminate thyroid nodules, optimising the rule-out ability when distinctive SUV cut-offs were applied to Hürthle and non-Hürthle cell nodules. In non-Hürthle cell nodules, malignancy could be ruled out at a SUV_{max} cut-off of 2.1 g/mL (similar to visual interpretation) with a sensitivity of 96% and benign call rate of 18%. In Hürthle cell nodules, a higher cut-off at 5.2 g/mL could rule out malignancy with a sensitivity of 100% and benign call rate of 17%. As such, quantitative analysis appears advantageous over visual analysis in Hürthle cell nodules. Consequently, [^{18}F]FDG PET/CT may be a reliable rule-out test for both non-Hürthle and Hürthle cell nodules, although external validation of these SUV thresholds is required before implementation in clinical practice.

Two recent publications investigated [^{18}F]FDG PET/CT radiomics in cytologically indeterminate thyroid nodules for the classification of malignancies [455, 459]. Giovanella et al. published the first retrospective study in 78 Bethesda III/IV patients (65 non-Hürthle nodules), suggesting a multiparametric model including cytological classification and two radiomic features [459]. The included features were the autocorrelation of the grey level cooccurrence matrix, a feature that describes the fineness of a texture, and the sphericity of the nodule shape, indicating a taller than wide shape. The cross-validated models with the two radiomic features resulted in AUCs of 73% and 73% for all nodules and in a subgroup of non-Hürthle cell nodules, respectively. In non-Hürthle cell nodules, a model with both the radiomic features and the cytological classification resulted in an AUC of 82%. A secondary analysis of the EfFECTS dataset performed additional radiomic analysis in [^{18}F]FDG-positive scans only [455]. The authors found that radiomic analysis did not contribute to the additional differentiation of [^{18}F]FDG-positive nodules. Both studies concluded that radiomic analysis alone on [^{18}F]FDG PET/CT seems of no added value in the management of indeterminate thyroid nodules. However, implemented in the multiparametric model of two radiomic features and the cytological classification that Giovanella et al. proposed, clinical application of radiomics seems feasible, although validation is required.

The availability of PET/CT scanners and tracers is increasing but varies worldwide. Transport distances are limited due to the short half-life of ^{18}F (~110 minutes), which is produced in cyclotrons. The radiation exposure of an [^{18}F]FDG PET/CT scan is mainly accounted for by the [^{18}F]FDG dosage, which amounts about 3.5 millisievert for an administered activity of 185 MBq [363]. The radiation exposure of CT largely varies, but can be less than 0.5 millisievert for a low-dose CT of the neck region only. Costs for an investigation are generally higher than for the other modalities described, because of the costs of PET hardware and the production and transportation of radiopharmaceuticals. Two studies assessed the cost-effectiveness of an [^{18}F]FDG PET/CT-driven management as compared to diagnostic surgery in all Bethesda III/IV patients. A 2014 cost-effectiveness model by Vriens et al.

showed that [^{18}F]FDG PET/CT decreased the number of futile surgeries by 47%, thereby reducing the expected 5-year direct medical costs per patient by €822 (from €8,804 to €7,983) as compared to surgical treatment while maintaining health-related quality of life (HRQoL). This study also concluded that, from a European perspective, [^{18}F]FDG PET/CT would be cost-effective over molecular testing [53]. Another cost-effectiveness study performed by the same group was recently conducted using the observed health care consumption and HRQoL data of the EfFECTS trial, which had found a similar reduction in futile surgeries [460]. This study assessed all societal costs over a lifelong horizon, and found that an [^{18}F]FDG PET/CT-driven management reduced the lifelong societal costs by almost €10,000 as compared to diagnostic surgery, with similar HRQoL for both strategies. While diagnostic surgery for a nodule with benign histopathology resulted in more cognitive impairment and physical problems including cosmetic complaints, the reassurance of a negative [^{18}F]FDG PET/CT resulted in sustained HRQoL throughout the first year of active surveillance [461].

Combined approaches

Every currently known engagement point from the genotype to the phenotype of the tumour is being explored. Combined, the various research fields encompass an extensive range of investigative methods. Individually they usually focus on one or two methods only, making one-to-one comparison of these diagnostics difficult. The 2015 American Thyroid Association (ATA) guidelines suggested several additional tests, but a definitive answer or complete overview of all available tests is still lacking [17]. Alongside higher-level expert discussions and lobbying of MedTech companies, clinical endocrinologists and thyroid surgeons ponder about the best solution for their individual patients. Their choices depend on the characteristics of their patient populations, availability and costs of a certain test, and personal preference. In any case, a useful additional test should be accurate, accessible, affordable, and affect patient management. A multimodal stepwise approach using a sensitive rule-out test and a specific rule-in test might provide the most conclusive diagnosis, e.g., in a specific test a relatively higher threshold value may be recommended to minimise missing malignancy in screening, while when appended to another diagnostic test, a relatively lower threshold value may be recommended to reduce false-positive results. Nevertheless, research into combined approaches is limited.

Piccardo et al. compared [^{18}F]FDG PET/CT, multiparametric US (including elastosonography), and [^{99m}Tc]Tc-MIBI scintigraphy in 87 nodules with indeterminate cytology (according the Società Italiana di Anatomia e Citologia Patologica-International Academy of Pathology classification published in 2010), wherefrom 18 nodules were found to be malignant in histopathology. Separately, [^{18}F]FDG PET/CT outperformed qualitative multiparametric US and [^{99m}Tc]Tc-MIBI scintigraphy for the detection of thyroid malignancy. Also, combined approaches were evaluated, demonstrating that (1) a negative [^{18}F]FDG PET/CT correctly predicted benign findings on histopathology, (2) a positive

[^{18}F]FDG PET/CT was significantly associated with malignancy when qualitative [^{99m}Tc]Tc-MIBI scans were rated as negative, and (3) the association of a positive [^{18}F]FDG PET/CT combined with a positive multiparametric US was significantly more specific than [^{18}F]FDG PET/CT alone in identifying differentiated thyroid cancer.

A combined approach by Trimboli et al. investigated whether [^{18}F]FDG PET/CT could play a role in the stratification of nodules with an intermediate risk upon EU-TI-RADS in 93 unselected nodules with EU-TI-RADS 4 and 5, including 38 nodules with indeterminate cytology [462]. They found that thyroid lesions classified as EU-TI-RADS 4 and with no [^{18}F]FDG uptake could be excluded from further examination. Another study by Piccardo et al. also investigated [^{18}F]FDG PET/CT, EU-TI-RADS, and the Italian consensus for the classification and reporting of thyroid cytology (ICCRTC) to distinguish differentiated thyroid cancers and FNs from nodular hyperplasias in 201 Bethesda III and IV thyroid nodules [40]. On multivariate analysis, [^{18}F]FDG PET/CT (OR 9.04), ICCRTC (OR 7.57), and EU-TI-RADS (OR 4.41) were all independent risk factors associated with differentiated thyroid carcinomas and FNs. These studies conclude that [^{18}F]FDG PET/CT could serve as a reliable rule-out test in case of nodules with intermediate risk upon US stratification.

Future perspectives

Medical imaging plays an important role in the preoperative workup of cytologically indeterminate thyroid nodules. A comprehensive overview of imaging biomarkers exemplified in this chapter can be found in Table 2. Most biomarkers used in the clinical work-up are visual interpretation or basic quantitative metrics. AI applications and radiomic methodologies, on the other hand, are less well established, but are currently developed on a large scale. Extensive external validation should be performed in order to achieve implementation of AI-derived imaging biomarkers in clinical practice. Many of the imaging biomarkers have either an adequate rule-in or rule-out capacity, but no single biomarker seems to serve both purposes well. A multimodal stepwise approach using a sensitive rule-out test and a specific rule-in test complementing each other might provide the most conclusive diagnosis for indeterminate thyroid nodules [25].

It should be noted that test performance of a test depends on the patient population. With the introduction of US risk stratification systems, FNAC might be withheld more often for patients with a presumed benign nodule, thereby potentially changing the composition of patient population and increasing its associated risk of malignancy. The proportion of Hürthle cell nodules is additionally crucial, as these nodules should be considered a separate entity with varying diagnostic yield of the different imaging modalities [454]. In addition, the prevalence of malignancy and the performance, costs, and feasibility of the imaging techniques might vary globally. Clinical utility should be examined in local implementation studies.

Table 2. Overview of imaging biomarkers in the management of cytologically indeterminate thyroid nodules exemplified in this chapter

Technique	Sensitivity	Specificity	Benign call rate (given a prevalence of 26%)	Advantages	Drawbacks	Cost-effectiveness
US	ATA: 52%* [401] ACR TI-RADS: 70%* [401] EU-TI-RADS: 38%* [401] K-TI-RADS: 78%* [401] *In Bethesda III nodules	ATA: 90%* [401] ACR TI-RADS: 60%* [401] EU-TI-RADS: 81%* [401] K-TI-RADS: 53%* [401] *In Bethesda III nodules	ATA: 79% ACR TI-RADS: 52% EU TI-RADS: 76% K-RADS: 45%	Global availability Low costs No ionising radiation Possibility of US-guided FNAC	Operator dependency Limited prospective clinical validation	Presumed, but unpublished
CT	NA	NA	NA	NA	Not investigated in indeterminate nodules	NA
MRI	97%* [439] *In unselected nodules, sensitivity and specificity in indeterminate nodules unknown	95%* [439]	NA	No ionising radiation	High costs Limited evidence No methodological consensus Research ongoing Limited (but increasing) availability of (high-field) MRI scanners	Currently unknown
[^{99m}Tc]Tc-MIBI scintigraphy (WO_{ind})	100%* [447] *Requires preselection of hypo-functioning nodules	89%* [447]	66%*	More widely available and lower costs than PET	Ionising radiation Limited sensitivity in lesions smaller than 30 mm	Unclear
[^{18}F]FDG PET/CT	94% [454]	40% [454]	31% [454]	High NPV High benign call rate Effective reducing futile diagnostic lobectomies	High costs Ionising radiation Limited but increasing availability of scanners and radiotracers Incidental findings (low specificity)	Reduced the lifelong societal costs by almost €10,000 as compared to diagnostic surgery [460]