THYROID
Volume 33, Number 8, 2023

DOI: 10.1089/thy.2023.0132

Artificial Intelligence in Thyroidology: A Narrative Review of the Current Applications, Associated Challenges, and Future Directions

David Toro-Tobon,[1] Ricardo Loor-Torres,[1] Mayra Duran,[1] Jungwei W. Fan,[2] Naykky Singh Ospina,[3] Yonghui Wu,[4] and Juan P. Brito[1]

Background: The use of artificial intelligence (AI) in health care has grown exponentially with the promise of facilitating biomedical research and enhancing diagnosis, treatment, monitoring, disease prevention, and health care delivery. We aim to examine the current state, limitations, and future directions of AI in thyroidology.
Summary: AI has been explored in thyroidology since the 1990s, and currently, there is an increasing interest in applying AI to improve the care of patients with thyroid nodules (TNODs), thyroid cancer, and functional or autoimmune thyroid disease. These applications aim to automate processes, improve the accuracy and consistency of diagnosis, personalize treatment, decrease the burden for health care professionals, improve access to specialized care in areas lacking expertise, deepen the understanding of subtle pathophysiologic patterns, and accelerate the learning curve of less experienced clinicians. There are promising results for many of these applications. Yet, most are in the validation or early clinical evaluation stages. Only a few are currently adopted for risk stratification of TNODs by ultrasound and determination of the malignant nature of indeterminate TNODs by molecular testing. Challenges of the currently available AI applications include the lack of prospective and multicenter validations and utility studies, small and low diversity of training data sets, differences in data sources, lack of explainability, unclear clinical impact, inadequate stakeholder engagement, and inability to use outside of the research setting, which might limit the value of their future adoption.
Conclusions: AI has the potential to improve many aspects of thyroidology; however, addressing the limitations affecting the suitability of AI interventions in thyroidology is a prerequisite to ensure that AI provides added value for patients with thyroid disease.

Keywords: artificial intelligence, machine learning, deep learning, thyroid

Introduction

Artificial intelligence (AI) was born in 1956 under the premise that ''Any aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.'' However, AI only began to show promising progress in the medical field in the early 2000s with the progress of computational capacity and the digitalization of health care. It was not until 2017 that the U.S. Food and Drug Administration (FDA) approved the first AI-based application for use.[1] AI has the potential to improve clinical effectiveness, access to care, and biomedical research by optimizing disease diagnosis, treatment, monitoring, prevention, and health care delivery.[2,3]

In thyroidology, the earliest published use of AI was in 1991 when researchers attempted to create diagnostic networks to interpret the thyroid function tests.[4,5] Since then, the interest in potential applications of AI has extended to almost all areas of thyroidology. In this review, we aimed to provide thyroid clinicians and researchers with a framework to

[1]Division of Endocrinology, Diabetes, Metabolism and Nutrition, Department of Medicine, Mayo Clinic, Rochester, Minnesota, USA.
[2]Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, Minnesota, USA.
[3]Division of Endocrinology, Department of Medicine, University of Florida, Gainesville, Florida, USA.
[4]Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, Florida, USA.

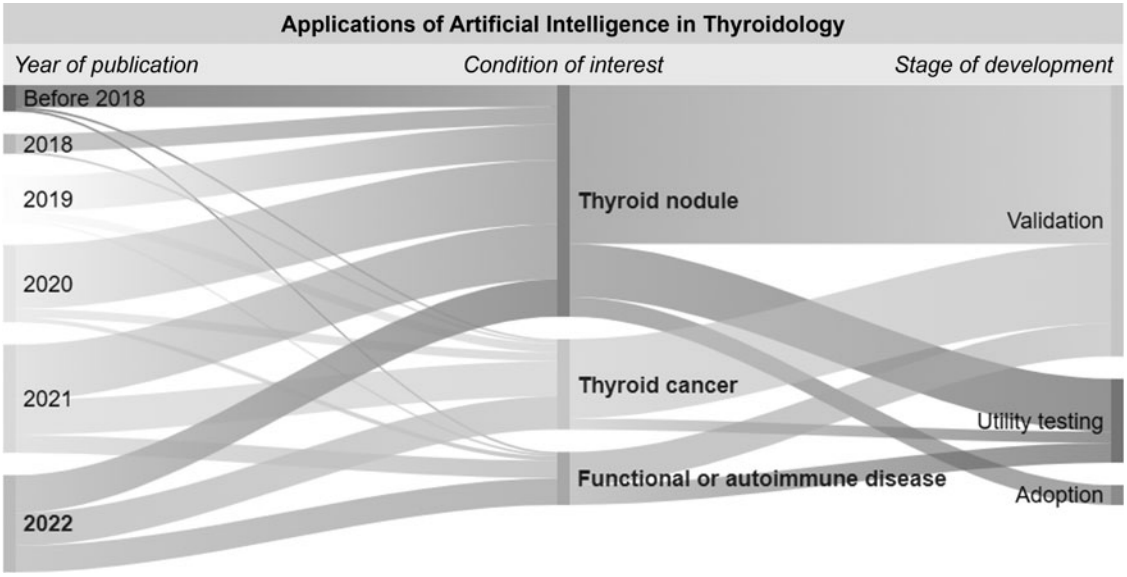

FIG. 1. Applications of AI in thyroidology grouped by year of publication, disease of interest, and stage of development. AI, artificial intelligence.

understand the latest development of AI in thyroidology by comprehensively characterizing emerging AI applications in several fields of thyroidology such as thyroid nodules (TNODs), thyroid cancer, and functional or autoimmune thyroid disease. Specifically, our framework provides a broad narrative overview of the process from conceiving to adopting AI applications, the current applications and techniques of AI in thyroidology, and the challenges that need to be addressed to facilitate future interventions as well as their added value for treating thyroid disease.

To guide this perspective review, we conducted a nonsystematic search of any study published until December 2022 that included the conception, development, validation, utility testing, or adoption of AI algorithms in thyroid conditions (Fig. 1) (Supplementary Appendix SA1).

General Concepts

AI is a computing technology capable of mimicking or surpassing human intelligence.[6] Today's AI algorithms take advantage of vast amounts of data (input) to identify complex and subtle patterns that might be difficult for humans. AI encompasses many interrelated techniques that require diverse levels of human supervision (supervised, unsupervised, semisupervised), and varying degrees of complexity for data processing.[7] In summary, *Machine learning* (*ML*) is an area of AI that allows computers to learn from data and make predictions.[8] Traditional ML models rely more on humans to identify useful features, which is a critical step to develop good predictive models (Fig. 2A). Later, *deep learning* (DL), an emerging area of ML that leverages neural network (NN) architectures, was proposed to enable machines to learn useful features (Fig. 2B).[9] DL represents a modern revamping of NNs, which were originally inspired by mimicking biological neurons' interactions.

In NN, data are combined in different ways through a series of progressive and hierarchical layers (including hidden layers) to establish relations from complex patterns.[10] In the medical domain, much detailed patient information is often captured in clinical narratives. Natural language processing (NLP) is the key AI technology that enables machines to recognize and extract information from unstructured text in the electronic medical records (EMR), thus facilitating the use of clinical text in ML models (Fig. 2C).[9]

AI Applications: From Conception to Adoption

AI innovation follows several maturing stages from conception to successful clinical adoption (Fig. 3).[11] In the conception stage, the research team defines a specific clinically relevant problem and primary outcome that reflect the current knowledge and burden of the disease of interest. This stage also involves building a multidisciplinary team with clinical and computational expertise and engaging with different stakeholders. In the data collection stage, the potential data sources and variables of interest are identified. Data are then reviewed and annotated with varying degrees of human input. In the training stage, collected data are processed and used to train the AI algorithm based on the chosen methodology. The results are periodically evaluated throughout the training stage, and the model is adjusted to improve performance.

In the validation stage, the model's performance is evaluated and further retraining or model tuning is performed where there is discordance with the initial training performance. In the utility testing phase, the model is evaluated in a real-world setting outside the research environment. The primary goal of this stage is to assess the benefits of the proposed model in real-world clinical practice, usually within the institution in which it was created. Sometimes, collaborations are established with other institutions to evaluate the model's generalizability. For a model to be considered adopted, multicenter and prospective validations, regulatory agencies' approval, and availability for use in clinical care are required. Once a model is adopted, there will be no periodic evaluation and retraining or tuning unless substandard performance occurs.

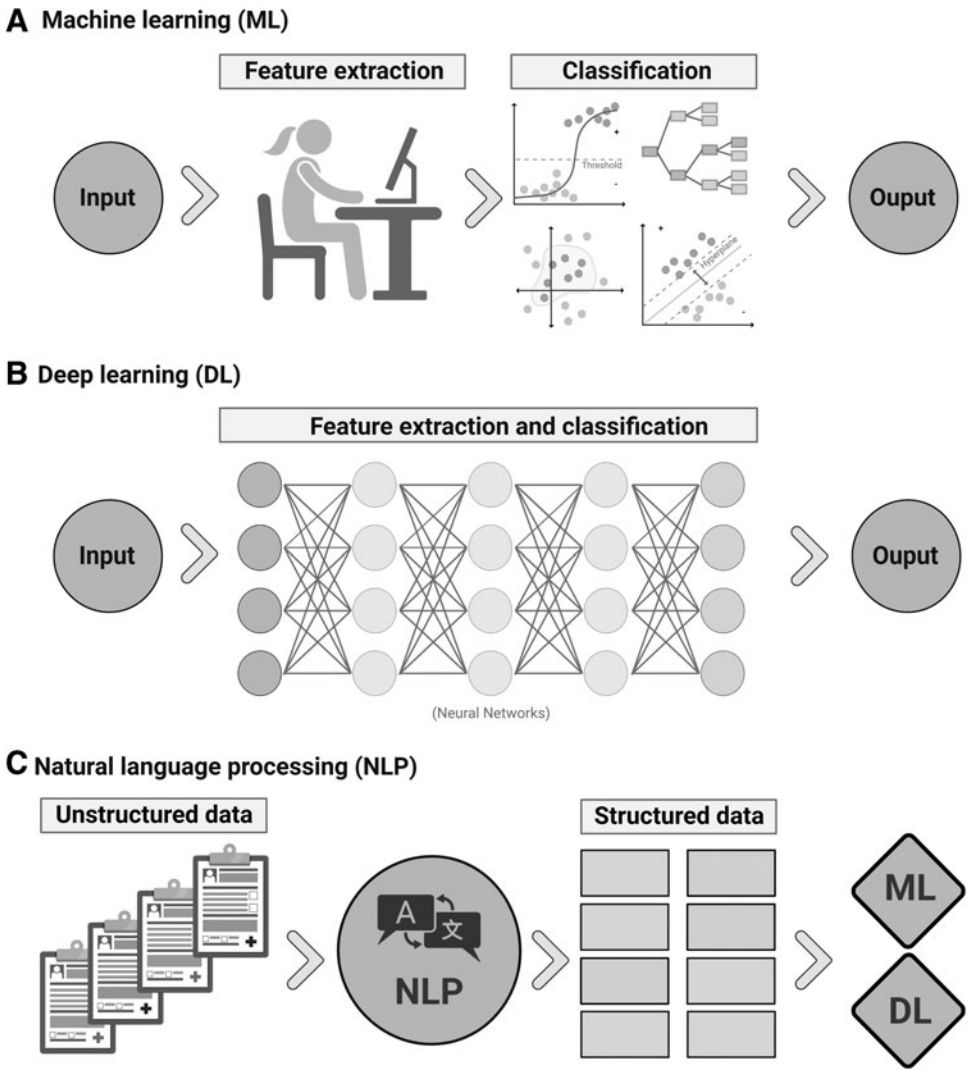

FIG. 2. Graphic representation of **(A)** machine learning (ML), **(B)** deep learning (DL), and **(C)** natural language processing (NLP).

Medical AI models should be evaluated using a combination of different metrics to prevent over- or underestimation of the results.[12] Traditional statistical measures such as sensitivity, specificity, negative predictive value, positive predictive value (also referred to as precision), true positive, true negative, false positive, and false negative are used. Other commonly used metrics include accuracy, the fraction of correctly classified outcomes among all outcomes, and the area under the curve (AUC), an aggregate measure of the model's predictive performance.[13] A model with an AUC closer to one is deemed to have excellent predictive performance, while a model with an AUC closer to 0.5 is considered to have random predictions.

Application of AI in TNODs

TNODs are the most common thyroid disease, with an estimated prevalence of up to 70% with ultrasound (US) and an overall risk of malignancy of 10%.[14] The widespread use of different imaging modalities has led to increased detection of TNODs.[15] In our review, 64% of the original studies of AI in thyroidology focused on several aspects of thyroid nodular disease, including optimizing resource utilization and providing a more accurate and personalized workup and management strategy. The available applications of AI in TNODs can be summarized based on their use of various imaging modalities and laboratory-based methods.

Thyroid US

Thyroid US is the most widely used diagnostic tool for evaluating TNODs.[16] While accessible, safe, and cost-effective, US interpretation is subject to significant inter-operator variability.[17] Over the last decade, multiple AI models have aimed at automating interpretation, improving and facilitating risk stratification of TNODs by thyroid US images, or its derived characteristics, and reducing the rate of unnecessary fine needle aspiration biopsies (FNABs),[18–20] with similar or superior performance than commonly used TNOD reporting and risk stratification systems (AUC: 0.76–0.98).[18,19,21–34]

Given the large heterogenicity in the models' design and training data sets, comparing their performance and evaluating their impact in real-world are challenging. Mature models with external validation have demonstrated mixed stand-alone performance results. Some outperformed

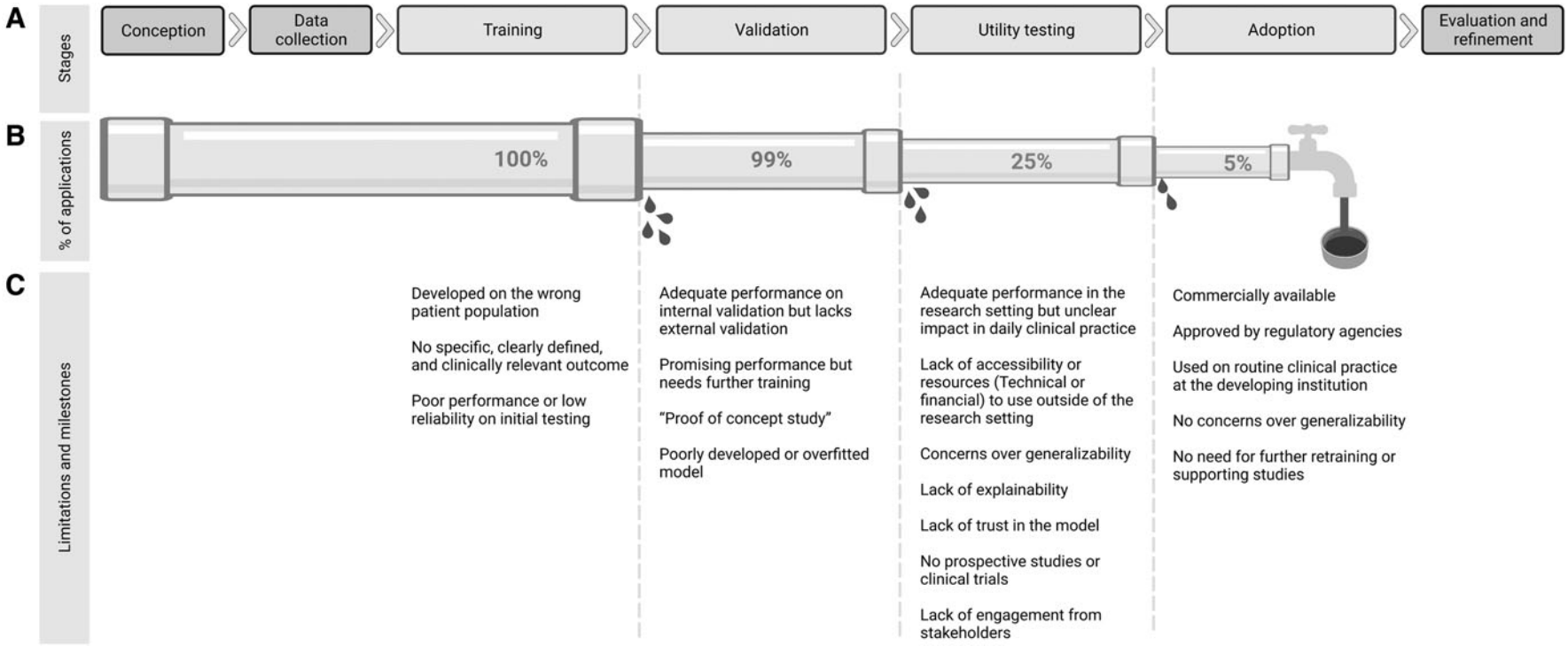

FIG. 3. (**A**) Stages from conception to successful clinical adoption on an AI model. (**B**) Percentage of applications per stage. (**C**) Limitations faced and milestones achieved at each stage.

experienced radiologists (AUC: 0.88–0.94),[35–37] others had comparable specificity but lower sensitivity than senior physicians and similar sensitivity but increased specificity than junior physicians (AUC: 0.65–0.98),[22,38–42] and others had an overall comparable performance regardless of the physicians' level of expertise (AUC: 0.82–0.96).[25,43–45] Moreover, some models increased the radiologist's performance when used as a supplementary diagnostic aid (AUC: 0.8–0.94).[38,46]

In addition, there is promising research on the incorporation of nontraditional data into the risk stratification process, such as the use of radiomics (quantitative extraction of high-dimensional image features from routine imaging) to clarify the nature of indeterminate TNODs (AUC: 0.75–0.88),[47–50] the prediction of *BRAF*^{V600E} mutation without requiring molecular testing (AUC: 0.64),[51] the use of video clips rather than static US images for TNOD characterization,[52] and the automatic incorporation of color Doppler features to enhance the TNOD risk stratification prediction (AUC: 0.89).[53] Finally, there have been preliminary studies to specifically differentiate follicular thyroid carcinoma (FTC) from adenoma (AUC: 0.80–0.96),[54,55] and to clarify the malignant versus benign nature of TNODs previously classified as indeterminate by FNAB (Accuracy: 77.4%).[56]

Despite the large number of AI models for US risk stratification of TNODs that have been developed with promising results, few of them have had sufficient external multicenter validation or prospective evaluation, only four have been approved by the FDA (S-detect, AmCAD-UT, Koios DS, MEDO-Thyroid), and none of the commercially available are being widely used.[7]

There are several limitations to the use of these AI models in real-world clinical practice. Some models are limited by suboptimal performance particularly due to the use of static images[57,58] and training data sets that have few benign nodules,[37] indeterminate nodules,[18,40,43] TIRADS 3 nodules,[26] and nonpapillary thyroid cancer (PTC) malignancies.[23,40] In addition, consistency can be affected by differences in the levels of provider expertise in semiautomatic models that require manual input,[59] and marked fluctuations in data quality due to diversity in data sources (US equipment, radiology protocols, and image segmentation methods).

Molecular testing

FNAB with cytology interpretation is the standard for preoperative diagnosis of TNODs[60]; however, the utility of FNAB can be limited by indeterminate or nondiagnostic results.[61] Thus, there has been extensive research on clarifying the benign versus malignant nature of these nodules to avoid unnecessary surgical or other invasive interventions. Over the last decade, several AI models were designed to predict the probability of malignancy of TNODs based on their molecular profile with adequate and incremental performance (AUC: 0.88–094).[62–64] Commercially available molecular diagnostic tests such as Afirma,[65–67] ThyroSeq,[68,69] Rosetta GX reveal,[70] and Thyramir[71] include AI-based classifiers and are perhaps the most widely used AI applications in thyroidology at this time.

Importantly, despite the high performance of these models in validation studies, their clinical utility might be limited by patient selection bias, and significant interinstitutional variations in performance metrics due to their specific pretest malignancy probability of cytologically indeterminate TNODs.[72]

Cytology

Cytological categorization of FNAB is user dependent, and higher quality improves care.[73] AI-based systems have been developed to improve FNAB interpretation, and prediction of malignancy risk, by automating the analysis of cytology images and identifying subtle cytology patterns. Early explorations on digital cytomorphologic evaluation through ML and DL have shown promising results in the automatic classification of TNOD's cytology (AUC: 0.75–0.93),[74–79] including indeterminate cytology TNODs (AUC: 0.75–0.96).[75,80] These models achieved comparable to superior performance than a pathologist when used stand-alone (Precision: 0.87), and improved pathologists' accuracy when used as a supplementary diagnostic aid (Precision: 0.81–0.88).[81]

In addition, there have been preliminary explorations of models to automatically identify region of interest (ROI) on whole slide cytology images to expedite the cytopathologists' review process with adequate concordance compared with manual ROI identification,[82] to predict *BRAF-RAS* gene expression and identify follicular-patterned thyroid neoplasms based on automatic evaluation of cytologic patterns (AUC: 0.98–0.99),[83,84] and to use adjuvant NLP-extracted features (demographic, US and biochemical characteristics) to improve cytologic classification of indeterminate TNODs (AUC: 0.85).[85]

The performance and generalizability in most of these models are limited by small training data sets with a modest amount of indeterminate or borderline TNODs (such as samples that express characteristics from various categories).

Other diagnostics

There are other early exploratory initiatives to further improve several aspects of the diagnosis and management of TNODs, including the use of NLP for automatic identification and workup tracking of thyroid incidentalomas on computed tomography (CT) reports (AUC: 0.99),[86] the use of CT or magnetic resonance imaging (MRI) images for TNOD risk stratification (AUC: 0.85–0.87),[87–91] the incorporation of demographic, ultrasonographic, biochemical, and cytologic characteristics into an AI-based decision tree model aimed at decreasing the false-negative rate of TNODs that undergo FNAB (Accuracy: 95.5%),[92] the automatic analysis of intraoperative TNOD's frozen sections (Precision: 16.7–96.7%),[93] and the prediction of TNOD's volume reduction response with radiofrequency ablation (Accuracy: 85.1%).[94]

The performance of these models might be limited by variations in documentation style and completeness on NLP extracted data, inadequate quality of CT or MRI images, imbalanced cohorts of benign versus malignant TNODs, and limited amount of non-PTC histologies in training data sets.

Application of AI in Thyroid Cancer

In recent years, the incidence of thyroid cancer has been on the rise, particularly among women in the United States.[95,96] As such, there is an urgent need to develop better tools for

risk stratification, recurrence prediction, and response to therapies. While the previous section described the applications for identification, evaluation, and risk stratification of TNODs, this section further expands on the additional aspects of risk stratification and management in patients with established thyroid cancer, including prediction of nodal and distant metastases, recurrence, prognosis, and treatment.

Preoperative risk stratification

Several models incorporated clinical, biochemical, anatomical, pathology, and US features to predict the presence of cervical lymph node metastasis (LNM) in patients with PTC (AUC: 0.67–0.91) with comparable performance to radiologists' interpretation of neck US.[97–101] Lee et al. evaluated the automatic detection of cervical LNM on the CT neck of patients with PTC with promising results (AUC: 0.95).[102] Their model performed similarly to experienced radiologists but superior to junior radiologists and trainees.[103] Other models used neck CT or US radiomics to automatically identify cervical LNM in patients with PTC with superior performance than experienced radiologists (AUC: 0.70–0.93).[104,105]

Finally, researchers used demographic and clinicopathological variables (including demographics, histology, and staging) to predict distant metastasis in patients with PTC or FTC (AUC: 0.85–0.91).[106–108] Despite promising results, these models' utility in clinical practice has not been demonstrated, and only two are readily accessible as online tools outside of the research setting.[99,102]

In addition, US and CT are commonly used to diagnose extrathyroidal extension (ETE) preoperatively, but their sensitivity and specificity are limited.[109] US radiomic models performed better than regular US interpretation in predicting ETE given the ability to capture risk factors not usually accessible to the human eye (such as PTC density and enhanced tissue heterogeneity; AUC: 0.83),[110] and CT radiomic models performed similar to experienced radiologists (AUC: 0.75).[111] In addition, CT or MRI radiomics, and US-based models have showed better performance in predicting thyroid capsule invasion (AUC: 0.82),[112] and preoperatively predicting advanced or aggressive PTC (AUC: 0.85–0.96).[113–115]

In general, these models are still in the training or utility testing stages and their generalizability is limited by small training data sets, inadequate data set heterogeneity with very modest presence of non-PTC histologies, low rate of events of interest, lack incorporation of additional clinical data such as molecular or biochemical markers into the predictive models, and scarce testing on images with diverse quality and segmentation techniques, or obtained with different diagnostic equipment.

Prognosis and recurrence risk

Several models use the patient's age and specific malignant disease characteristics (tumor size, metastatic involvement, and nodular disease) to predict the staging of well-differentiated thyroid cancer with similar performance to the 8th American Joint Committee on Cancer (AJCC) staging system (AUC: 0.85–0.98).[116–118] Furthermore, there are models that used clinicopathological, biochemical, and molecular data to predict the risk of thyroid cancer recurrence (Accuracy: 95.7%).[119,120] A similar model by Kim et al., in addition, used radiation and systemic therapy data to predict survival in patients with distal metastasis status post-thyroidectomy.[119] These models are limited by selection bias and missing or nonstandardized data due to the retrospective nature of the single-center training databases, insufficient long-term survival or recurrence data, and lack of external or prospective validation.

Treatment

Researchers have evaluated the use of AI to predict treatment responses and potential complications. For example, Lubin et al. used ML to identify clinical factors (including tumor focality, preoperative staging, and biochemical markers) predictive of radioactive iodine (RAI) failure.[121] Similarly, Seib et al. used preoperative patient and malignant disease characteristics to predict postoperative complications (such as hypocalcemia, recurrent laryngeal nerve injury, or hematoma; AUC: 0.72).[122] Lastly, Liu et al. used data from quality-of-life questionnaires, sociodemographic and clinical characteristics to predict reduction of quality of life in patients with thyroid cancer 3 months after thyroidectomy (AUC: 0.89).[123] The generalizability of these models could be limited by their small training data sets, and lack of accountability for important risk factors such as prediagnostic psychological health when assessing post-surgical quality of life.

Other studies have evaluated the use of AI to improve the efficacy and safety of the different treatment strategies. Gong et al. developed a model for real-time identification and measurement of the recurrent laryngeal nerve using computer vision during thyroidectomy (Precision: 75.6%).[124] Their model showed promising results and demonstrated feasibility to augment intraoperative decision-making. In addition, another early model from Lin et al. aimed to optimize radiotherapy precision in patients with metastatic thyroid cancer through the use of positron emission tomography CT (PET-CT) with encouraging results.[125] Due to the small training data sets, these models could have suboptimal performance with anatomical variations, inconsistent image quality, and the presence of indeterminate or challenging diagnostic findings.

Application of AI in Autoimmune and Functional Disease

Researchers have also explored the use of AI to understand the intricate aspects of hypo- and hyperthyroidism pathophysiology, automate diagnostic workflow, and enhance current diagnostic and therapeutic approaches.

Hypothyroidism

Two computer-assisted diagnosis (CAD) models trained using DL have been evaluated for the automatic diagnosis of Hashimoto's thyroiditis (HT) from thyroid US data analysis.[126,127] One of these models achieved excellent accuracy (AUC: 0.94), demonstrating consistency on external validation, and had higher performance than radiologists regardless of their level of expertise.[127] These models are subject to patient selection bias, and their performance upon presence

of confounders such as TNODs has not been tested. In addition, some models have been trained to predict thyroid dysfunction and patient-specific thyrotropin (TSH) levels based on demographics, and clinical and biochemical data with mixed performance (AUC: 0.61–0.87).[128,129] These models are limited by retrospective training data sets with missing data, imbalanced patient subpopulations, and low rates of events of interest.

Thyrotoxicosis

Several ML models have been developed to aid in the diagnosis of hyperthyroidism. Some models performed well in classifying common thyroid scintigraphy uptake patterns and differentiating between entities such as Graves' disease (GD) and subacute thyroiditis (ST; Accuracy: 87.7–99.3%).[130,131] An NN model from Ma et al. outstandingly distinguished between GD, ST, and HT using single-photon emission CT (SPECT) images (Accuracy: 99–99.6%).[132] In addition, some models were developed to predict the presence and the etiology of thyrotoxicosis based on patient characteristics and biochemical data from the EMR.[133,134] Other researchers have used AI to personalize treatment for GD by predicting patients' responses to nonsurgical therapies such as antithyroid drugs (ATDs) and RAI. Orunesu et al. trained an NN model that used baseline patient characteristics to predict outcomes after discontinuation of ATDs with high sensitivity and specificity (84.6% and 77.2%).[135] Duan et al. used similar characteristics to predict the chance of post-RAI hypothyroidism with suboptimal performance (AUC: 0.72).[136]

The generalizability of some of these models is limited due to considerable discrepancies in their training and validation performances. In addition, training data sets were small, contained missing data on key variables such as TRAb levels, and were subject to subpopulation imbalances or low rate of events of interest.

Furthermore, AI has emerged as an alternative to deepen the understanding of the complex pathophysiology of different diseases, including hyperthyroidism. While single genes have been associated with GD, Shen et al. trained an ML model to identify different multigene associations that could be involved in the pathogenesis of GD (AUC: 0.9).[137] Although this model does not reach current clinical practice, its outcomes could facilitate future targeted therapeutic development and strategies for early disease identification and genetic counseling.

Thyroid eye disease

Accurate evaluation and severity assessment are important for treating thyroid eye disease (TED) with emerging therapies. However, current tools such as clinical activity score, vision, inflammation, strabismus, and appearance, and European Group of Graves' Orbitopathy classifications have limitations, leading to misclassifications or missed diagnoses.[138,139] As a solution, AI-based alternatives are being explored to improve diagnosis, severity assessment, and monitoring of TED progression and treatment response.[140–143] A model by Song et al. diagnosed TED through the screening of orbital CT images with outstanding results (AUC: 0.91).[140] Wen et al. trained a model that used specific local and remote brain functional connectivity abnormalities on functional brain MRI to diagnose TED with an accuracy of 78.5%.[142] In addition, this model provided further insight into the mechanisms of cognitive and visual symptoms of patients with TED. Likewise, a model by Huang et al. identified several features of TED based on the analysis of patients' facial images with promising performance (AUC: 0.6–0.93).[144]

Similarly, Lin et al. estimated disease activity on TED patients using orbital MRI images with higher performance than clinicians' assessment (AUC: 0.92).[141] For TED management, Hu et al. predicted response to systemic glucocorticoids through the integration of orbital MRI radiomics and the disease duration, with adequate performance (AUC: 0.85–0.91).[143] The performance of these models is limited by the small data sets with imbalanced subpopulations of interest, and variations in images quality or radiology protocols. In addition, they lack utility testing on real clinical practice.

Discussion

Challenges and future directions

Notwithstanding the advantages the reviewed AI applications may offer, their implementation in real-world clinical practice is challenging due to several limitations. Figure 3 displays the current stage of development of the reviewed applications based on the available published data as well as the general challenges faced, and milestones achieved on each stage.

As already described, each model might have specific limitations depending on its aim and design. In addition, general challenges must be addressed to improve the successful adoption of AI in thyroidology (Table 1). The performance of an AI model is intrinsically tied to the quality of the data. Small, homogeneous, single-center, biased, or retrospective data sets may impair the algorithm's performance, particularly when applied to external institutions or real-world scenarios. Therefore, it is crucial to use larger, more diverse, multicentric, and prospective data sets to ensure the performance and generalizability of the AI model.[145–148] In addition, models that use demographic variables can be subject to discriminatory bias when trained with databases that reflect historical health inequities in underrepresented groups (differences by race, gender, and socioeconomic backgrounds).[148]

Thus, training data sets should be representative of the general population, and conscious efforts should be made to identify and correct performance variations when the model is applied to subpopulations with diverse demographics. Furthermore, a lack of interpretability of the model's reasoning process (the ''black box'' effect) could prevent high-performing algorithms from being adopted or achieving their highest potential. For instance, even if a model can predict the risk of thyroid cancer recurrence with high accuracy, it might not be widely trusted and adopted by the medical community if there is not some insight into the reasoning and weight of variables behind the model's prediction.

Incorporating known pathophysiology into the model can provide valuable context for clinicians and provide them with some degree of self-explainability. For instance, providing the key features that contribute to the model's prediction can help increase the model's transparency and trustworthiness, promote the discovery of new or subtle clinical patterns, and facilitate continuous improvement by identifying factors that impact the precision of the algorithm.[149,150]

Table 1. Challenges and Future Directions

Challenge	*Future directions*
Inconsistent or inadequate performance due to suboptimal quality of training data sets	Train models on larger and more representative data sets (including higher diversity of conditions of interest, disease phenotypes, underrepresented population groups, types of practice, and data sources)
	Conduct multicentric and prospective studies
Lack of explainability (''the black-box effect'')	Account for known pathophysiology when training the model
	Train self-explainable models
	Design tools that provide some insight into the model's reasoning process
Inconsistent or suboptimal generalizability due to differences in data sources	Understand differences in documentation protocols
	Design NLP tools to leverage nonstructured data from the EMR
	Create standardization protocols to account for the ever-changing nature of health care data
	Validate the models with different equipment or data sources
Overenthusiastic model promises	Define clear, specific, clinically relevant, and actionable outcomes
	Evaluate the impact on routine clinical care
Lack of stakeholder engagement	Estimate financial and expert resources needed to facilitate suitability of the AI application
	Create multidisciplinary research and clinical implementation teams
	Account for stakeholders' expectations and fears
Lack of incorporation into routine clinical practice	Consider the logistics needed to incorporate the model into routine health care workflow (i.e., user-friendly interfaces, and compatibility with the EMR or stand-alone use as a CDA)
	Streamline the application to prevent worsening the health care providers' burden
	Select and implement applications that reflect the institution-specific gap in level of expertise or deficit in clinical workflow
	Account for availability of resources and compatibility of the application with legacy infrastructure
Inappropriate regulations	Establish cross-sector collaborations to design regulations that guarantee the safety of the applications while accounting for the dynamic learning and progressive improvement of AI with continued use
	Evaluate compliance with ethical standards
Unclear performance in time	Consider performing periodic performance reassessment
	Retrain or refine the model when there is suboptimal performance
	Analyze feedback from clinicians and users, and adjust the application accordingly
Lack of patients' acceptance and trust	Promote patient engagement through transparency and education through the different stages of model development

AI, artificial intelligence; CDA, clinical diagnostic aid; EMR, electronic medical record; NLP, natural language processing.

Moreover, differences in data sources from variations in diagnostic equipment, the structure in which clinical data are documented in the EMR, and institutional practice changes over time can affect the data quality and the performance of the model.[151] Thus, it is important to evaluate the performance of the models when using different data sources and retrain the model when needed. Furthermore, there should be standardization protocols that accommodate fluctuations originating from the ever-changing nature of health care data and account for different diagnostic and laboratory equipment. This would improve the models' consistency and facilitate the development of high-quality data sets that could be used in further model training.[145–148]

In addition to addressing the current limitations, future research of AI in thyroidology could expand to other areas that are yet to be explored, especially nonimage-based models, applications aimed at generating new knowledge rather than simply automating processes, and predictive models to facilitate ''precision medicine.'' Examples of current research interest include the use of physiologic data captured through wearable devices to facilitate diagnosis, follow up or medication management, the use of NLP models that leverage large volumes of unstructured data from the EMR for augmenting or enriching uncoded or not well-coded data elements, multimodal approaches that incorporate different types of inputs, and more complex pathophysiologic models that include genomics, proteomics, and metabolomics.

Implementation of AI applications in the real world

AI applications should ensure that the model possesses clearly defined, clinically relevant, and actionable outcomes that reflect the burden of the problem being addressed. For instance, while a model could prove high performance in diagnosing HT from US data, this might not have added benefit when applied to current real-world practice. Thus, before full adoption into routine clinical practice, external validation, prospective and multicenter clinical trials, and correlation between the model's accuracy and its real-world clinical efficacy are imperative.[148,152] In addition, incorporating these AI models into the routine clinical workflow is limited by the availability of expert resources, lack of real-time access to the models outside of the research setting, difficult incorporation of the models into the EMR, and high costs associated with software or hardware acquisition.[22]

Thus, stakeholder engagement through the distinct phases of development and deployment is fundamental for successful adoption. Implementing any AI strategy requires careful

Table 2. Considerations for Critical Appraisal of the Literature

Category	*Considerations*
Methodology	Is the aim of the study reflective of the burden or need being addressed?
	Is there a comprehensive description of the study design, including type of study, data source, inclusion and exclusion criteria for data set cases, and AI techniques used?
	What is the size of the data set?
	Is the data set representative of the general disease and population of interest (i.e., heterogeneous data set)? Or is it reflective of specific disease phenotypes (i.e., homogeneous data set)?
	What is the quality of the data set? (i.e., Was the data set created by human annotators or automatically by an AI model? Are the data extracted from free-text documentation in the EMR or directly identified from more reliable sources such as diagnostic images or pathology results? How were the data processed?)
	Are there clear descriptions on how the data set was used for training, testing, and validation? (i.e., Was the data set split 80% for training and 20% for testing, was cross-validation used, or was external validation performed?)
	Are the performance metrics clearly specified?
	Was data collection compliant with patient privacy and confidentiality regulations?
Results	Are the results reflective of the model's aim?
	Were the results compared with a gold standard?
	Are there particular scenarios, disease phenotypes, or patient subpopulations in which the model cannot be used or in which the performance is suboptimal?
	Does the model provide some degree of explainability? (i.e., Does the model report the weight of different factors behind its prediction?)
	Is the performance reproducible and generalizable? (i.e., Is there external, prospective, or multicenter validation?)
Utility and implementation	Is there any description of the model's utility or role in current real-world clinical practice? (i.e., Does the model add any benefit to the care or outcome of patients with thyroid disease?).
	Is the model's stage of development explicitly stated?
	Is the model available for testing or use outside of the research setting?
	Are the limitations of the model appropriately characterized?
	Are future directions for research or clinical implementation described?
	If used in real-world clinical practice, is the model approved by appropriate regulatory agencies?

understanding of the available financial and expert resources and a multidisciplinary approach that accounts for the expectations and fears of all the affected stakeholders.[2,153]

Furthermore, cross-sector collaborations and coordinated efforts between domain experts (i.e., researchers and clinical experts), technology experts (i.e., technology firms and AI vendors), law makers, regulatory agencies, health care decision makers (i.e., health system and insurance executives), and patients are fundamental for the advancement of AI in thyroidology and the successful incorporation of AI models into routine clinical care.[2] Finally, current and future AI models for thyroid conditions should consider appropriate regulatory frameworks that ensure AI interventions' safe and ethical use while accounting for their dynamic learning and progressive improvement over time.[154]

Critical appraisal of AI literature

Despite the current large body of AI literature and its expected exponential growth over the next couple of years, there are not currently validated and widely accepted critical appraisal tools. Several authors have characterized different frameworks to understand AI reports.[155–158] Table 2 summarizes important considerations for thyroid clinicians and researchers when analyzing an AI article in thyroidology.

Conclusions

The integration of AI into thyroid-related research and daily clinical practice marks the beginning of a new era in thyroidology. AI has the potential to improve the consistency and accuracy of diagnosis, decrease health care professionals' workload, predict response to therapy, and facilitate the development of clinical decision support systems. In addition, AI can increase access to specialized care, identify subtle risk patterns, and promote personalized care. However, several limitations preclude most current models from being used in real-world clinical practice. To fully realize the potential of AI in thyroidology, rigorous methodological planning and suitability testing are necessary to identify and address obstacles and increase the likelihood of successful adoption of AI interventions.

Acknowledgment

Figures were created with BioRender.com

Authors' Contributions

D.T.T.: conceptualization, methodology, literature review, and article preparation (initial draft, review, and editing). R.L.T. and M.D.: literature review and article preparation (initial draft and review). J.P.B.: conceptualization, methodology, and article preparation (review and editing). J.W.F., N.S.O., and Y.W.: article preparation (review and editing).

Disclaimer

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the Patient-Centered Outcomes Research Institute.

Author Disclosure Statement

None of the authors has any relevant disclosures.

Funding Information

J.P.B. and N.S.O. were supported by the National Cancer Institute of the National Institutes of Health under Award Numbers R37CA272473 and K08CA248972, respectively. Y.W. was supported by the National Institute on Aging of the National Institutes of Health under Award Number R56AG069880, and the Patient-Centered Outcomes Research Institute under Award number ME-2018C3-14754.

Supplementary Material

Supplementary Appendix SA1

References

1. Kaul V, Enslin S, Gross SA. History of artificial intelligence in medicine. Gastrointest Endosc 2020;92(4):807–812; doi: 10.1016/j.gie.2020.06.040
2. Chen M, Decary M. Artificial intelligence in healthcare: An essential guide for health leaders. Healthc Manage Forum 2020;33(1):10–18; doi: 10.1177/0840470419873123
3. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018;2(10):719–731; doi: 10.1038/s41551-018-0305-z
4. Bolinger RE, Hopfensperger KJ, Preston DF. Application of a virtual neurode in a model thyroid diagnostic network. Proc Annu Symp Comput Appl Med Care 1991; 310–314.
5. Forsström J, Nuutila P, Irjala K. Using the ID3 algorithm to find discrepant diagnoses from laboratory databases of thyroid patients. Med Decis Making 1991;11(3):171–175; doi: 10.1177/0272989x9101100305
6. Jung J-Y, Park D. Chapter 1—Are AI Models Explainable, Interpretable, and Understandable? In: Human-Centered Artificial Intelligence. (Nam CS, Jung J-Y, Lee S. eds.) Academic Press: Cambridge, MA; 2022; pp. 3–16.
7. Tessler FN, Thomas J. Artificial intelligence for evaluation of thyroid nodules: A primer. Thyroid 2022;33(2): 150–158; doi: 10.1089/thy.2022.0560
8. Vieira S, Lopez Pinaya WH, Mechelli A. Chapter 1—Introduction to Machine Learning. In: Machine Learning. (Mechelli A, Vieira S. eds.) Academic Press: Cambridge, MA; 2020; pp. 1–20.
9. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. Nat Med 2019;25(1):24–29; doi: 10.1038/s41591-018-0316-z
10. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–444; doi: 10.1038/nature14539
11. Mincu D, Roy S. Developing robust benchmarks for driving forward AI innovation in healthcare. Nat Mach Intell 2022;4(11):916–921; doi: 10.1038/s42256-022-00559-4
12. Hicks SA, Strümke I, Thambawita V, et al. On evaluation metrics for medical applications of artificial intelligence. Sci Rep 2022;12(1):5979; doi: 10.1038/s41598-022-09954-8
13. Parasa S, Repici A, Berzin T, et al. Framework and metrics for the clinical use and implementation of artificial intelligence algorithms into endoscopy practice: Recommendations from the American Society for Gastrointestinal Endoscopy Artificial Intelligence Task Force. Gastrointest Endosc 2023;97(5):815.e1–824.e1; doi: 10.1016/j.gie.2022.10.016
14. Durante C, Grani G, Lamartina L, et al. The diagnosis and management of thyroid nodules: A review. JAMA 2018; 319(9):914–924; doi: 10.1001/jama.2018.0898
15. Brito JP, Al Nofal A, Montori VM, et al. The impact of subclinical disease and mechanism of detection on the rise in thyroid cancer incidence: A population-based study in Olmsted County, Minnesota during 1935 through 2012. Thyroid 2015;25(9):999–1007; doi: 10.1089/thy.2014.0594
16. Grani G, Sponziello M, Pecce V, et al. Contemporary thyroid nodule evaluation and management. J Clin Endocrinol Metab 2020;105(9):2869–2883; doi: 10.1210/clinem/dgaa322
17. Persichetti A, Di Stasio E, Coccaro C, et al. Inter- and intraobserver agreement in the assessment of thyroid nodule ultrasound features and classification systems: A blinded multicenter study. Thyroid 2020;30(2):237–242; doi: 10.1089/thy.2019.0360
18. Zhao CK, Ren TT, Yin YF, et al. A comparative analysis of two machine learning-based diagnostic patterns with thyroid imaging reporting and data system for thyroid nodules: Diagnostic performance and unnecessary biopsy rate. Thyroid 2021;31(3):470–481; doi: 10.1089/thy.2020.0305
19. Liu Y, Li X, Yan C, et al. Comparison of diagnostic accuracy and utility of artificial intelligence-optimized ACR TI-RADS and original ACR TI-RADS: A multi-center validation study based on 2061 thyroid nodules. Eur Radiol 2022;32(11):7733–7742; doi: 10.1007/s00330-022-08827-y
20. Stib MT, Pan I, Merck D, et al. Thyroid nodule malignancy risk stratification using a convolutional neural network. Ultrasound Q 2020;36(2):164–172; doi: 10.1097/ruq.0000000000000501
21. Thomas J, Haertling T. AIBx, artificial intelligence model to risk stratify thyroid nodules. Thyroid 2020;30(6):878–884; doi: 10.1089/thy.2019.0752
22. Zhu J, Zhang S, Yu R, et al. An efficient deep convolutional neural network model for visual localization and automatic diagnosis of thyroid nodules on ultrasound images. Quant Imaging Med Surg 2021;11(4):1368–1380; doi: 10.21037/qims-20-538
23. Wang L, Yang S, Yang S, et al. Automatic thyroid nodule recognition and diagnosis in ultrasound imaging with the YOLOv2 neural network. World J Surg Oncol 2019;17(1): 12; doi: 10.1186/s12957-019-1558-z
24. Watkins L, O'Neill G, Young D, et al. Comparison of British Thyroid Association, American College of Radiology TIRADS and Artificial Intelligence TIRADS with histological correlation: Diagnostic performance for predicting thyroid malignancy and unnecessary fine needle aspiration rate. Br J Radiol 2021;94(1123):20201444; doi: 10.1259/bjr.20201444
25. Kim YJ, Choi Y, Hur SJ, et al. Deep convolutional neural network for classification of thyroid nodules on ultrasound: Comparison of the diagnostic performance with that of radiologists. Eur J Radiol 2022;152:110335; doi: 10.1016/j.ejrad.2022.110335
26. Wu GG, Lv WZ, Yin R, et al. Deep learning based on ACR TI-RADS can improve the differential diagnosis of thyroid nodules. Front Oncol 2021;11:575166; doi: 10.3389/fonc.2021.575166

27. Li X, Zhang S, Zhang Q, et al. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: A retrospective, multicohort, diagnostic study. Lancet Oncol 2019;20(2):193–201; doi: 10.1016/S1470-2045(18)30762-9
28. Swan KZ, Thomas J, Nielsen VE, et al. External validation of AIBx, an artificial intelligence model for risk stratification, in thyroid nodules. Eur Thyroid J 2022; 11(2):e210129; doi: 10.1530/etj-21-0129
29. Buda M, Wildman-Tobriner B, Hoang JK, et al. Management of thyroid nodules seen on US images: Deep learning may match performance of radiologists. Radiology 2019; 292(3):695–701; doi: 10.1148/radiol.2019181343
30. Bai Z, Chang L, Yu R, et al. Thyroid nodules risk stratification through deep learning based on ultrasound images. Med Phys 2020;47(12):6355–6365; doi: 10.1002/mp.14543
31. Wildman-Tobriner B, Buda M, Hoang JK, et al. Using artificial intelligence to revise ACR TI-RADS risk stratification of thyroid nodules: Diagnostic accuracy and utility. Radiology 2019;292(1):112–119; doi: 10.1148/radiol.2019182128
32. Gomes Ataide EJ, Ponugoti N, Illanes A, et al. Thyroid nodule classification for physician decision support using machine learning-evaluated geometric and morphological features. Sensors 2020;20(21):6110.
33. Wu H, Deng Z, Zhang B, et al. Classifier model based on machine learning algorithms: Application to differential diagnosis of suspicious thyroid nodules via sonography. Am J Roentgenol 2016;207(4):859–864; doi: 10.2214/AJR.15.15813
34. Chen D, Hu J, Zhu M, et al. Diagnosis of thyroid nodules for ultrasonographic characteristics indicative of malignancy using random forest. BioData Min 2020;13:14; doi: 10.1186/s13040-020-00223-w
35. Zhang B, Tian J, Pei S, et al. Machine learning-assisted system for thyroid nodule diagnosis. Thyroid 2019;29(6): 858–867; doi: 10.1089/thy.2018.0380
36. Sun C, Zhang Y, Chang Q, et al. Evaluation of a deep learning-based computer-aided diagnosis system for distinguishing benign from malignant thyroid nodules in ultrasound images. Med Phys 2020;47(9):3952–3960; doi: 10.1002/mp.14301
37. Zhao Z, Yang C, Wang Q, et al. A deep learning-based method for detecting and classifying the ultrasound images of suspicious thyroid nodules. Med Phys 2021; 48(12):7959–7970; doi: 10.1002/mp.15319
38. Zhang Y, Wu Q, Chen Y, et al. A clinical assessment of an ultrasound computer-aided diagnosis system in differentiating thyroid nodules with radiologists of different diagnostic experience. Front Oncol 2020;10:557169; doi: 10.3389/fonc.2020.557169
39. Xia S, Yao J, Zhou W, et al. A computer-aided diagnosing system in the evaluation of thyroid nodules—Experience in a specialized thyroid center. World J Surg Oncol 2019; 17(1):210; doi: 10.1186/s12957-019-1752-z
40. Chen Y, Gao Z, He Y, et al. An artificial intelligence model based on ACR TI-RADS characteristics for US diagnosis of thyroid nodules. Radiology 2022;303(3):613–619; doi: 10.1148/radiol.211455
41. Liu T, Guo Q, Lian C, et al. Automated detection and classification of thyroid nodules in ultrasound images using clinical-knowledge-guided convolutional neural networks. Med Image Anal 2019;58:101555; doi: 10.1016/j.media.2019.101555
42. Barczyński M, Stopa-Barczyńska M, Wojtczak B, et al. Clinical validation of S-Detect™ mode in semi-automated ultrasound classification of thyroid lesions in surgical office. Gland Surg 2020;9(Suppl 2):S77–S85; doi: 10.21037/gs.2019.12.23
43. Park VY, Han K, Seong YK, et al. Diagnosis of thyroid nodules: Performance of a deep learning convolutional neural network model vs. radiologists. Sci Rep 2019;9(1): 17843; doi: 10.1038/s41598-019-54434-1
44. Zhou H, Jin Y, Dai L, et al. Differential diagnosis of benign and malignant thyroid nodules using deep learning radiomics of thyroid ultrasound images. Eur J Radiol 2020;127:108992; doi: 10.1016/j.ejrad.2020.108992
45. Zhu YC, Jin PF, Bao J, et al. Thyroid ultrasound image classification using a convolutional neural network. Ann Transl Med 2021;9(20):1526; doi: 10.21037/atm-21-4328
46. Wei X, Gao M, Yu R, et al. Ensemble deep learning model for multicenter classification of thyroid nodules on ultrasound images. Med Sci Monit 2020;26:e926096; doi: 10.12659/msm.926096
47. Keutgen XM, Li H, Memeh K, et al. A machine-learning algorithm for distinguishing malignant from benign indeterminate thyroid nodules using ultrasound radiomic features. J Med Imaging (Bellingham) 2022;9(3):034501; doi: 10.1117/1.Jmi.9.3.034501
48. Gild ML, Chan M, Gajera J, et al. Risk stratification of indeterminate thyroid nodules using ultrasound and machine learning algorithms. Clin Endocrinol (Oxf) 2022; 96(4):646–652; doi: 10.1111/cen.14612
49. Wang S, Xu J, Tahmasebi A, et al. Incorporation of a machine learning algorithm with object detection within the thyroid imaging reporting and data system improves the diagnosis of genetic risk. Front Oncol 2020;10: 591846; doi: 10.3389/fonc.2020.591846
50. Daniels K, Gummadi S, Zhu Z, et al. Machine learning by ultrasonography for genetic risk stratification of thyroid nodules. JAMA Otolaryngol Head Neck Surg 2020; 146(1):36–41; doi: 10.1001/jamaoto.2019.3073
51. Yoon J, Lee E, Koo JS, et al. Artificial intelligence to predict the BRAFV600E mutation in patients with thyroid cancer. PLoS One 2020;15(11):e0242806; doi: 10.1371/journal.pone.0242806
52. Luo H, Ma L, Wu X, et al. Deep learning-based ultrasonic dynamic video detection and segmentation of thyroid gland and its surrounding cervical soft tissues. Med Phys 2022;49(1):382–392; doi: 10.1002/mp.15332
53. Zhu YC, Du H, Jiang Q, et al. Machine learning assisted doppler features for enhancing thyroid cancer diagnosis: A multi-cohort study. J Ultrasound Med 2022;41(8):1961–1974; doi: 10.1002/jum.15873
54. Seo JK, Kim YJ, Kim KG, et al. Differentiation of the follicular neoplasm on the gray-scale US by image selection subsampling along with the marginal outline using convolutional neural network. Biomed Res Int 2017;2017: 3098293; doi: 10.1155/2017/3098293
55. Yang B, Yan M, Yan Z, et al. Segmentation and classification of thyroid follicular neoplasm using cascaded convolutional neural network. Phys Med Biol 2020; 65(24):245040; doi: 10.1088/1361-6560/abc6f2
56. Chen L, Chen M, Li Q, et al. Machine learning-assisted diagnostic system for indeterminate thyroid nodules. Ultrasound Med Biol 2022;48(8):1547–1554; doi: 10.1016/j.ultrasmedbio.2022.03.020

57. Wang J, Jiang J, Zhang D, et al. An integrated AI model to improve diagnostic accuracy of ultrasound and output known risk features in suspicious thyroid nodules. Eur Radiol 2022; 32(3):2120–2129; doi: 10.1007/s00330-021-08298-7
58. Koh J, Lee E, Han K, et al. Diagnosis of thyroid nodules on ultrasonography by a deep convolutional neural network. Sci Rep 2020;10(1):15245; doi: 10.1038/s41598-020-72270-6
59. Jeong EY, Kim HL, Ha EJ, et al. Computer-aided diagnosis system for thyroid nodules on ultrasonography: Diagnostic performance and reproducibility based on the experience level of operators. Eur Radiol 2019;29(4): 1978–1985; doi: 10.1007/s00330-018-5772-9
60. Haugen BR, Alexander EK, Bible KC, et al. 2015 American Thyroid Association Management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: The American Thyroid Association guidelines task force on thyroid nodules and differentiated thyroid cancer. Thyroid 2016;26(1):1–133; doi: 10.1089/thy.2015.0020
61. Cibas ES, Ali SZ. The Bethesda system for reporting thyroid cytopathology. Am J Clin Pathol 2009;132(5): 658–665; doi: 10.1309/ajcpphlwmi3jv4la
62. Tomei S, Marchetti I, Zavaglia K, et al. A molecular computational model improves the preoperative diagnosis of thyroid nodules. BMC Cancer 2012;12(1):396; doi: 10.1186/1471-2407-12-396
63. Pankratz DG, Hu Z, Kim SY, et al. Analytical performance of a gene expression classifier for medullary thyroid carcinoma. Thyroid 2016;26(11):1573–1580; doi: 10.1089/thy.2016.0262
64. Hao Y, Duh Q-Y, Kloos RT, et al. Identification of Hürthle cell cancers: Solving a clinical challenge with genomic sequencing and a trio of machine learning algorithms. BMC Syst Biol 2019;13(2):27; doi: 10.1186/s12918-019-0693-z
65. Diggans J, Kim SY, Hu Z, et al. Machine learning from concept to clinic: Reliable detection of BRAF V600E DNA mutations in thyroid nodules using high-dimensional RNA expression data. Pac Symp Biocomput 2015;371–382.
66. Nasr CE, Andrioli M, Endo M, et al. Real-world performance of the Afirma Genomic Sequencing Classifier (GSC)—A meta-analysis. J Clin Endocrinol Metab 2023; 108(6):1526–1532; doi: 10.1210/clinem/dgac688
67. Patel KN, Angell TE, Babiarz J, et al. Performance of a genomic sequencing classifier for the preoperative diagnosis of cytologically indeterminate thyroid nodules. JAMA Surg 2018;153(9):817–824; doi: 10.1001/jamasurg.2018.1153
68. Steward DL, Carty SE, Sippel RS, et al. Performance of a multigene genomic classifier in thyroid nodules with indeterminate cytology: A prospective blinded multicenter study. JAMA Oncol 2019;5(2):204–212; doi: 10.1001/jamaoncol.2018.4616
69. Skaugen JM, Taneja C, Liu JB, et al. Performance of a multigene genomic classifier in thyroid nodules with suspicious for malignancy cytology. Thyroid 2022;32(12): 1500–1508; doi: 10.1089/thy.2022.0282
70. Lithwick-Yanai G, Dromi N, Shtabsky A, et al. Multicentre validation of a microRNA-based assay for diagnosing indeterminate thyroid nodules utilising fine needle aspirate smears. J Clin Pathol 2017;70(6):500–507; doi: 10.1136/jclinpath-2016-204089
71. Ablordeppey KK, Timmaraju VA, Song-Yang JW, et al. Development and analytical validation of an expanded mutation detection panel for next-generation sequencing of thyroid nodule aspirates. J Mol Diagn 2020;22(3):355–367; doi: 10.1016/j.jmoldx.2019.11.003
72. Khan TM, Zeiger MA. Thyroid nodule molecular testing: Is it ready for prime time? Front Endocrinol (Lausanne) 2020;11:590128; doi: 10.3389/fendo.2020.590128
73. Sakorafas GH. Thyroid nodules; interpretation and importance of fine-needle aspiration (FNA) for the clinician—Practical considerations. Surg Oncol 2010;19(4):e130–e139; doi: 10.1016/j.suronc.2010.06.003
74. Ren Y, He Y, Cong L. Application value of a deep convolutional neural network model for cytological assessment of thyroid nodules. J Healthc Eng 2021;2021: 6076135; doi: 10.1155/2021/6076135
75. Elliott Range DD, Dov D, Kovalsky SZ, et al. Application of a machine learning algorithm to predict malignancy in thyroid cytopathology. Cancer Cytopathol 2020;128(4): 287–295; doi: 10.1002/cncy.22238
76. Gopinath B, Shanthi N. Computer-aided diagnosis system for classifying benign and malignant thyroid nodules in multi-stained FNAB cytological images. Australas Phys Eng Sci Med 2013;36(2):219–230; doi: 10.1007/s13246-013-0199-8
77. Böhland M, Tharun L, Scherr T, et al. Machine learning methods for automated classification of tumors with papillary thyroid carcinoma-like nuclei: A quantitative analysis. PLoS One 2021;16(9):e0257635; doi: 10.1371/journal.pone.0257635
78. Fragopoulos C, Pouliakis A, Meristoudis C, et al. Radial basis function artificial neural network for the investigation of thyroid cytological lesions. J Thyroid Res 2020; 2020:5464787; doi: 10.1155/2020/5464787
79. Guan Q, Wang Y, Ping B, et al. Deep convolutional neural network VGG-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: A pilot study. J Cancer 2019;10(20):4876–4882; doi: 10.7150/jca.28769
80. Yao K, Jing X, Cheng J, et al. A study of thyroid fine needle aspiration of follicular adenoma in the ''atypia of undetermined significance'' Bethesda category using digital image analysis. J Pathol Inform 2022;13:100004; doi: 10.1016/j.jpi.2022.100004
81. Dov D, Kovalsky SZ, Assaad S, et al. Weakly supervised instance learning for thyroid malignancy prediction from whole slide cytopathology images. Med Image Anal 2021; 67:101814; doi: 10.1016/j.media.2020.101814
82. Dov D, Kovalsky SZ, Feng Q, et al. Use of machine learning-based software for the screening of thyroid cytopathology whole slide images. Arch Pathol Lab Med 2022;146(7):872–878; doi: 10.5858/arpa.2020-0712-OA
83. Dolezal JM, Trzcinska A, Liao C-Y, et al. Deep learning prediction of BRAF-RAS gene expression signature identifies noninvasive follicular thyroid neoplasms with papillary-like nuclear features. Mod Pathol 2021;34(5): 862–874; doi: 10.1038/s41379-020-00724-3
84. Anand D, Yashashwi K, Kumar N, et al. Weakly supervised learning on unannotated H&E-stained slides predicts BRAF mutation in thyroid cancer with high accuracy. J Pathol 2021;255(3):232–242; doi: 10.1002/path.5773
85. Luong G, Idarraga AJ, Hsiao V, et al. Risk stratifying indeterminate thyroid nodules with machine learning. J Surg Res 2022;270:214–220; doi: 10.1016/j.jss.2021.09.015

86. Canton SP, Dadashzadeh E, Yip L, et al. Automatic detection of thyroid and adrenal incidentals using radiology reports and deep learning. J Surg Res 2021;266:192–200; doi: 10.1016/j.jss.2021.03.060
87. Li W, Cheng S, Qian K, et al. Automatic recognition and classification system of thyroid nodules in CT images based on CNN. Comput Intell Neurosci 2021;2021: 5540186; doi: 10.1155/2021/5540186
88. Li Z, Zhang H, Chen W, et al. Contrast-enhanced CT-based radiomics for the differentiation of nodular goiter from papillary thyroid carcinoma in thyroid nodules. Cancer Manag Res 2022;14:1131–1140; doi: 10.2147/cmar.S353877
89. Zhang X, Lee VCS, Rong J, et al. Multi-channel convolutional neural network architectures for thyroid cancer detection. PLoS One 2022;17(1):e0262128; doi: 10.1371/journal.pone.0262128
90. Sharafeldeen A, Elsharkawy M, Khaled R, et al. Texture and shape analysis of diffusion-weighted imaging for thyroid nodules classification using machine learning. Med Phys 2022;49(2):988–999: doi: 10.1002/mp.15399
91. Naglah A, Khalifa F, Khaled R, et al. Novel MRI-based CAD system for early detection of thyroid cancer using multi-input CNN. Sensors (Basel) 2021;21(11):3878; doi: 10.3390/s21113878
92. Idarraga AJ, Luong G, Hsiao V, et al. False negative rates in benign thyroid nodule diagnosis: Machine learning for detecting malignancy. J Surg Res 2021;268:562–569; doi: 10.1016/j.jss.2021.06.076
93. Li Y, Chen P, Li Z, et al. Rule-based automatic diagnosis of thyroid nodules from intraoperative frozen sections using deep learning. Artif Intell Med 2020;108:101918; doi: 10.1016/j.artmed.2020.101918
94. Negro R, Rucco M, Creanza A, et al. Machine learning prediction of radiofrequency thermal ablation efficacy: A new option to optimize thyroid nodule selection. Eur Thyroid J 2020;9(4):205–212; doi: 10.1159/000504882
95. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. CA Cancer J Clin 2015;65(1):5–29; doi: 10.3322/caac.21254
96. Megwalu UC, Moon PK. Thyroid cancer incidence and mortality trends in the United States: 2000–2018. Thyroid 2022;32(5):560–570; doi: 10.1089/thy.2021.0662
97. Liu W, Wang S, Xia X, et al. A proposed heterogeneous ensemble algorithm model for predicting central lymph node metastasis in papillary thyroid cancer. Int J Gen Med 2022;15:4717–4732; doi: 10.2147/ijgm.S365725
98. Zhu W, Huang X, Qi Q, et al. Artificial neural network-based ultrasound radiomics can predict large-volume lymph node metastasis in clinical N0 papillary thyroid carcinoma patients. J Oncol 2022;2022:7133972; doi: 10.1155/2022/7133972
99. Zhu J, Zheng J, Li L, et al. Application of machine learning algorithms to predict central lymph node metastasis in T1–T2, non-invasive, and clinically node negative papillary thyroid carcinoma. Front Med 2021;8:635771; doi: 10.3389/fmed.2021.635771
100. Zou Y, Shi Y, Liu J, et al. A comparative analysis of six machine learning models based on ultrasound to distinguish the possibility of central cervical lymph node metastasis in patients with papillary thyroid carcinoma. Front Oncol 2021;11:656127; doi: 10.3389/fonc.2021.656127
101. Wu Y, Rao K, Liu J, et al. Machine learning algorithms for the prediction of central lymph node metastasis in patients with papillary thyroid cancer. Front Endocrinol (Lausanne) 2020;11:577537; doi: 10.3389/fendo.2020.577537
102. Lee JH, Ha EJ, Kim JH. Application of deep learning to the diagnosis of cervical lymph node metastasis from thyroid cancer with CT. Eur Radiol 2019;29(10):5452–5457; doi: 10.1007/s00330-019-06098-8
103. Lee JH, Ha EJ, Kim D, et al. Application of deep learning to the diagnosis of cervical lymph node metastasis from thyroid cancer with CT: External validation and clinical utility for resident training. Eur Radiol 2020;30(6):3066–3072; doi: 10.1007/s00330-019-06652-4
104. Li J, Wu X, Mao N, et al. Computed tomography-based radiomics model to predict central cervical lymph node metastases in papillary thyroid carcinoma: A multicenter study. Front Endocrinol (Lausanne) 2021;12:741698; doi: 10.3389/fendo.2021.741698
105. Yu J, Deng Y, Liu T, et al. Lymph node metastasis prediction of papillary thyroid carcinoma based on transfer learning radiomics. Nat Commun 2020;11(1):4807; doi: 10.1038/s41467-020-18497-3
106. Liu W-C, Li Z-Q, Luo Z-W, et al. Machine learning for the prediction of bone metastasis in patients with newly diagnosed thyroid cancer. Cancer Med 2021;10(8):2802–2811; doi: 10.1002/cam4.3776
107. Liu W, Wang S, Ye Z, et al. Prediction of lung metastases in thyroid cancer using machine learning based on SEER database. Cancer Med 2022;11(12):2503–2515; doi: 10.1002/cam4.4617
108. Mao Y, Lan H, Lin W, et al. Machine learning algorithms are comparable to conventional regression models in predicting distant metastasis of follicular thyroid carcinoma. Clin Endocrinol (Oxf) 2023;98(1):98–109; doi: 10.1111/cen.14693
109. Seo YL, Yoon DY, Lim KJ, et al. Locally advanced thyroid cancer: Can CT help in prediction of extrathyroidal invasion to adjacent structures? AJR Am J Roentgenol 2010; 195(3):W240–W244; doi: 10.2214/ajr.09.3965
110. Wang X, Agyekum EA, Ren Y, et al. A radiomic nomogram for the ultrasound-based evaluation of extrathyroidal extension in papillary thyroid carcinoma. Front Oncol 2021;11:625646; doi: 10.3389/fonc.2021.625646
111. Yu P, Wu X, Li J, et al. Extrathyroidal extension prediction of papillary thyroid cancer with computed tomography based radiomics nomogram: A multicenter study. Front Endocrinol (Lausanne) 2022;13:874396; doi: 10.3389/fendo.2022.874396
112. Wu X, Yu P, Jia C, et al. Radiomics analysis of computed tomography for prediction of thyroid capsule invasion in papillary thyroid carcinoma: A multi-classifier and two-center study. Front Endocrinol (Lausanne) 2022;13: 849065; doi: 10.3389/fendo.2022.849065
113. Dai Z, Wei R, Wang H, et al. Multimodality MRI-based radiomics for aggressiveness prediction in papillary thyroid cancer. BMC Med Imaging 2022;22(1):54; doi: 10.1186/s12880-022-00779-5
114. Edwards K, Halicek M, Little JV, et al. Multiparametric radiomics for predicting the aggressiveness of papillary thyroid carcinoma using hyperspectral images. Proc SPIE Int Soc Opt Eng 2021;11597:1159728; doi: 10.1117/12.2582147
115. Cordes M, Götz TI, Lang EW, et al. Advanced thyroid carcinomas: Neural network analysis of ultrasonographic characteristics. Thyroid Res 2021;14(1):16; doi: 10.1186/s13044-021-00107-z

116. Yang CQ, Gardiner L, Wang H, et al. Creating prognostic systems for well-differentiated thyroid cancer using machine learning. Front Endocrinol (Lausanne) 2019;10:288; doi: 10.3389/fendo.2019.00288
117. Mao Y, Huang Y, Xu L, et al. Surgical methods and social factors are associated with long-term survival in follicular thyroid carcinoma: Construction and validation of a prognostic model based on machine learning algorithms. Front Oncol 2022;12:816427; doi: 10.3389/fonc.2022.816427
118. Mourad M, Moubayed S, Dezube A, et al. Machine learning and feature selection applied to SEER data to reliably assess thyroid cancer prognosis. Sci Rep 2020; 10(1):5176; doi: 10.1038/s41598-020-62023-w
119. Kim SY, Kim YI, Kim HJ, et al. New approach of prediction of recurrence in thyroid cancer patients using machine learning. Medicine (Baltimore) 2021;100(42): e27493; doi: 10.1097/md.0000000000027493
120. Park YM, Lee B-J. Machine learning-based prediction model using clinico-pathologic factors for papillary thyroid carcinoma recurrence. Sci Rep 2021;11(1):4948; doi: 10.1038/s41598-021-84504-2
121. Lubin DJ, Tsetse C, Khorasani MS, et al. Clinical predictors of I-131 therapy failure in differentiated thyroid cancer by machine learning: A single-center experience. World J Nucl Med 2021;20(3):253–259; doi: 10.4103/wjnm.WJNM_104_20
122. Seib CD, Roose JP, Hubbard AE, et al. Ensemble machine learning for the prediction of patient-level outcomes following thyroidectomy. Am J Surg 2021;222(2):347–353; doi: 10.1016/j.amjsurg.2020.11.055
123. Liu YH, Jin J, Liu YJ. Machine learning-based random forest for predicting decreased quality of life in thyroid cancer patients after thyroidectomy. Support Care Cancer 2022;30(3):2507–2513; doi: 10.1007/s00520-021-06657-0
124. Gong J, Holsinger FC, Noel JE, et al. Using deep learning to identify the recurrent laryngeal nerve during thyroidectomy. Sci Rep 2021;11(1):14306; doi: 10.1038/s41598-021-93202-y
125. Lin Q, Qi Q, Hou S, et al. Application of Pet-CT fusion deep learning imaging in precise radiotherapy of thyroid cancer. J Healthc Eng 2021;2021:2456429; doi: 10.1155/2021/2456429
126. Zhang Q, Zhang S, Pan Y, et al. Deep learning to diagnose Hashimoto's thyroiditis from sonographic images. Nat Commun 2022;13(1):3759; doi: 10.1038/s41467-022-31449-3
127. Zhao W, Kang Q, Qian F, et al. Convolutional neural network-based computer-assisted diagnosis of Hashimoto's thyroiditis on ultrasound. J Clin Endocrinol Metab 2022; 107(4):953–963; doi: 10.1210/clinem/dgab870
128. Cheng X, Li S, Deng L, et al. Predicting elevated TSH levels in the physical examination population with a machine learning model. Front Endocrinol (Lausanne) 2022; 13:839829; doi: 10.3389/fendo.2022.839829
129. Santhanam P, Nath T, Mohammad FK, et al. Artificial intelligence may offer insight into factors determining individual TSH level. PLoS One 2020;15(5):e0233336; doi: 10.1371/journal.pone.0233336
130. Yang P, Pi Y, He T, et al. Automatic differentiation of thyroid scintigram by deep convolutional neural network: A dual center study. BMC Med Imaging 2021;21(1):179; doi: 10.1186/s12880-021-00710-4
131. Qiao T, Liu S, Cui Z, et al. Deep learning for intelligent diagnosis in thyroid scintigraphy. J Int Med Res 2021; 49(1):300060520982842; doi: 10.1177/03000605 20982842
132. Ma L, Ma C, Liu Y, et al. Thyroid diagnosis from SPECT images using convolutional neural network with optimization. Comput Intell Neurosci 2019;2019:6212759; doi: 10.1155/2019/6212759
133. Hu M, Asami C, Iwakura H, et al. Development and preliminary validation of a machine learning system for thyroid dysfunction diagnosis based on routine laboratory tests. Commun Med (Lond) 2022;2:9; doi: 10.1038/s43856-022-00071-1
134. Kim J, Baek H-S, Ha J, et al. Differential diagnosis of thyrotoxicosis by machine learning models with laboratory findings. Diagnostics 2022;12(6):1468.
135. Orunesu E, Bagnasco M, Salmaso C, et al. Use of an artificial neural network to predict Graves' disease outcome within 2 years of drug withdrawal. Eur J Clin Invest 2004; 34(3):210–217; doi: 10.1111/j.1365-2362.2004.01318.x
136. Duan L, Zhang HY, Lv M, et al. Machine learning identifies baseline clinical features that predict early hypothyroidism in patients with Graves' disease after radioiodine therapy. Endocr Connect 2022;11(5):e220119; doi:10.1530/ec-22-0119
137. Shen F, Cai W, Gan X, et al. Prediction of genetic factors of hyperthyroidism based on gene interaction network. Front Cell Dev Biol 2021;9:700355; doi: 10.3389/fcell.2021.700355
138. Dolman PJ. Grading severity and activity in thyroid eye disease. Ophthalmic Plast Reconstr Surg 2018;34(4S): S34–S40; doi: 10.1097/iop.0000000000001150
139. Bartalena L, Kahaly GJ, Baldeschi L, et al. The 2021 European Group on Graves' orbitopathy (EUGOGO) clinical practice guidelines for the medical management of Graves' orbitopathy. Eur J Endocrinol 2021;185(4):G43–G67; doi: 10.1530/eje-21-0479
140. Song X, Liu Z, Li L, et al. Artificial intelligence CT screening model for thyroid-associated ophthalmopathy and tests under clinical conditions. Int J Comput Assist Radiol Surg 2021;16(2):323–330; doi: 10.1007/s11548-020-02281-1
141. Lin C, Song X, Li L, et al. Detection of active and inactive phases of thyroid-associated ophthalmopathy using deep convolutional neural network. BMC Ophthalmol 2021; 21(1):39; doi: 10.1186/s12886-020-01783-5
142. Wen Z, Wan X, Qi CX, et al. Local-to-remote brain functional connectivity in patients with thyroid-associated ophthalmopathy and assessment of its predictive value using machine learning. Int J Gen Med 2022;15:4273–4283; doi: 10.2147/ijgm.S353649
143. Hu H, Chen L, Zhang JL, et al. T(2)-weighted MR imaging-derived radiomics for pretreatment determination of therapeutic response to glucocorticoid in patients with thyroid-associated ophthalmopathy: comparison with semiquantitative evaluation. J Magn Reson Imaging 2022; 56(3):862–872; doi: 10.1002/jmri.28088
144. Huang X, Ju L, Li J, et al. An intelligent diagnostic system for thyroid-associated ophthalmopathy based on facial images. Front Med (Lausanne) 2022;9:920716; doi: 10.3389/fmed.2022.920716
145. Sessions V, Valtorta M. The Effects of Data Quality on Machine Learning Algorithms. MIT International Conference on Information Quality: Cambridge, MA; 2006.

146. Chen H, Ji Y. Adversarial training for improving model robustness? Look at both prediction and interpretation. arXiv preprint arXiv:220312709; 2022.
147. Chapman WW, Nadkarni PM, Hirschman L, et al. Overcoming barriers to NLP for clinical text: The role of shared tasks and the need for additional creative solutions. J Am Med Inform Assoc 2011;18(5):540–543; doi: 10.1136/amiajnl-2011-000465
148. Kelly CJ, Karthikesalingam A, Suleyman M, et al. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019;17(1):195; doi: 10.1186/s12916-019-1426-2
149. Amann J, Blasimme A, Vayena E, et al. Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. BMC Med Inform Decis Mak 2020;20(1):310; doi: 10.1186/s12911-020-01332-6
150. Combi C, Amico B, Bellazzi R, et al. A manifesto on explainability for artificial intelligence in medicine. Artif Intell Med 2022;133:102423; doi: 10.1016/j.artmed.2022.102423
151. Nestor B, McDermott M, Chauhan G, et al. Rethinking clinical prediction: Why machine learning must consider year of care and feature aggregation. 2018. arXiv preprint arXiv:181112583 2018.
152. Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. NPJ Digit Med 2018;1(1):40; doi: 10.1038/s41746-018-0048-y
153. Scott IA, Carter SM, Coiera E. Exploring stakeholder attitudes towards AI in clinical practice. BMJ Health Care Inform 2021;28(1):e100450; doi: 10.1136/bmjhci-2021-100450
154. U.S. Food & Drug Administration. Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD). 2019. Available from: https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf [Last accessed: June 19, 2023].
155. Faes L, Liu X, Wagner SK, et al. A clinician's guide to artificial intelligence: How to critically appraise machine learning studies. Transl Vis Sci Technol 2020;9(2):7; doi: 10.1167/tvst.9.2.7
156. Pham N, Hill V, Rauschecker A, et al. Critical appraisal of artificial intelligence–enabled imaging tools using the levels of evidence system. Am J Neuroradiol 2023;44(5):E21–E28; doi: 10.3174/ajnr.A7850
157. van Smeden M, Heinze G, Van Calster B, et al. Critical appraisal of artificial intelligence-based prediction models for cardiovascular disease Eur Heart J 2022;43(31):2921–2930; doi: 10.1093/eurheartj/ehac238
158. Kocak B, Kus EA, Kilickesmez O. How to read and review papers on machine learning and artificial intelligence in radiology: A survival guide to key methodological concepts. Eur Radiol 2021;31(4):1819–1830; doi: 10.1007/s00330-020-07324-4

Address correspondence to:
Juan P. Brito, MD
Division of Endocrinology, Diabetes, Nutrition and Metabolism
Department of Medicine
Mayo Clinic
200 First Street SW
Rochester, MN 55902
USA

E-mail: brito.juan@mayo.edu