The Journal of Clinical Endocrinology & Metabolism, 2021, Vol. XX, No. XX, 1–11
https://doi.org/10.1210/clinem/dgab435
Clinical Research Article

Clinical Research Article

Predicting Malignancy in Pediatric Thyroid Nodules: Early Experience With Machine Learning for Clinical Decision Support

Lebohang Radebe,[1,2] Daniëlle C. M. van der Kaay,[3] Jonathan D. Wasserman,[4,5,*] and Anna Goldenberg[1,2,6,7,*]

[1]Genetics and Genome Biology Program, The Hospital for Sick Children, Toronto, Ontario M5G OA4, Canada; [2]Department of Computer Science, University of Toronto, Toronto, Ontario M5T 3A1, Canada; [3]Erasmus Medical Center - Sophia Children's Hospital, Rotterdam 3000 CB, the Netherlands; [4]Division of Endocrinology, The Hospital for Sick Children, Toronto, Ontario M5G 1X8, Canada; [5]Department of Paediatrics, Faculty of Medicine, The University of Toronto, Ontario M5S 1A8, Canada; [6]Vector Institute, Toronto, Ontario M5G 1M1, Canada; and [7]CIFAR, Toronto, Ontario, M5G 1M1, Canada

ORCiD numbers: 0000-0003-3184-9165 (L. Radebe); 0000-0001-7088-8146 (J. D. Wasserman); 0000-0002-2416-833X (A. Goldenberg).

*J.D.W. and A.G. contributed equally to this work.

Abbreviations: AUROC, area under the receiver operator curve; CV, cross-validation; FNAB, fine-needle aspiration biopsy; FN, false negative; FNR, false-negative rate; FP, false positive; FPR, false-positive rate; ML, machine learning; MTNS, McGill Thyroid Nodule Score; PTC, papillary thyroid carcinoma; RF, random forest.

Received: 29 January 2021; Editorial Decision: 10 June 2021; First Published Online: 23 June 2021; Corrected and Typeset: 8 October 2021.

Abstract

Objective: To develop a machine learning tool to integrate clinical data for the prediction of non-benign thyroid cytology and histology.

Context: Papillary thyroid carcinoma is the most common endocrine malignancy. Since most nodules are benign, the challenge for the clinician is to identify those most likely to harbor malignancy while limiting exposure to surgical risks among those with benign nodules.

Methods: Random forests (augmented to select features based on our clinical measure of interest), in conjunction with interpretable rule sets, were used on demographic, ultrasound, and biopsy data of thyroid nodules from children younger than 18 years at a tertiary pediatric hospital. Accuracy, false-positive rate (FPR), false-negative rate (FNR), and area under the receiver operator curve (AUROC) are reported.

Results: Our models predict nonbenign cytology and malignant histology better than historical outcomes. Specifically, we expect a 68.04% improvement in the FPR, 11.90% increase in accuracy, and 24.85% increase in AUROC for biopsy predictions in 67 patients (28 with benign and 39 with nonbenign histology). We expect a 23.22% decrease in FPR,

ISSN Print 0021-972X ISSN Online 1945-7197
Printed in USA

32.19% increase in accuracy, and 3.84% decrease in AUROC for surgery prediction in 53 patients (42 with benign and 11 with nonbenign histology). This improvement comes at the expense of the FNR, for which we expect 10.27% with malignancy would be discouraged from performing biopsy, and 11.67% from surgery. Given the small number of patients, these improvements are estimates and are not tested on an independent test set.

Conclusion: This work presents a first attempt at developing an interpretable machine learning based clinical tool to aid clinicians. Future work will involve sourcing more data and developing probabilistic estimates for predictions.

Key Word: thyroid nodule malignancy

Papillary thyroid carcinoma (PTC) is the most common endocrine malignancy. PTC is also the most common malignancy overall in young women aged 15 to 29 years. The annual incidence is rising worldwide, with a female predominance emerging in early adolescence (1). Although thyroid nodules in children are significantly more likely to be malignant than in adults, even in children, roughly 70% of nodules are benign (2, 3). The challenge to the clinician is therefore to stratify those patients most likely to harbor malignancy and to prioritize surgery for these patients, while limiting exposure to surgical risks among those with benign nodules.

Clinical evaluation, ultrasound, and fine-needle aspiration cytology are fundamental modalities in assessing the likelihood that a nodule is malignant, yet each is associated with a substantial proportion of indeterminate results and each lacks the ability to accurately exclude malignancy in a substantial proportion of nodules (4-7). While some clinical features, including rock-hard texture and presence of cervical adenopathy, are closely associated with malignancy, there are no absolute predictors of benign disease. Gannon and colleagues (5) reviewed a large cohort of pediatric thyroid nodules and determined that ultrasound alone cannot satisfactorily identify sonographic features to adequately exclude malignancy. Similarly, ultrasound cannot be used to refine malignancy risk for cytologically indeterminate pediatric nodules (8). Finally, in a recent meta-analysis of pediatric cytopathology series, 5% to 28% of fine-needle aspiration biopsy (FNAB) specimens were nondiagnostic and 3.3% to 38% were cytologically indeterminate (7).

In several recent pediatric case series, including one at our own institution, the malignancy rate among surgically managed patients with thyroid nodules was 51% to 64% (2, 7, 9). This defines a high rate of surgery for benign disease and highlights the inherent limitations of current interpretation of the preoperative diagnostic modalities to accurately identify patients with high risk of malignancy for surgery.

In an effort to refine preoperative risk assessment, the McGill Thyroid Nodule Score (MTNS) was developed to integrate clinical, sonographic, and cytologic data (10, 11). It uses expert-defined weighted parameters to generate a likelihood score. This has recently been applied to 2 small pediatric series and appears to show some promise (12, 13). These works adopt an expert-opinion approach to selecting included variables and relative weighting, rather than a top-down derivation of the predictive score based on the primary data.

Using a retrospectively acquired data set of 198 pediatric patients, we sought to derive a computational model to integrate clinical, radiological, and cytological features to refine prediction of malignancy, so as to stratify those patients most likely to harbor malignancy and to identify those that may qualify for expectant management.

Machine learning (ML) algorithms, a subfield of artificial intelligence, have become increasingly popular in medicine because of their ability to learn highly complex patterns from data. Popular applications include disease identification, for example, through predicting cancer using radiological images (14); understanding the human genome, for example, through recognizing patterns in DNA sequences and understanding the mechanisms of gene expression using large, complex genetic, and genomic data sets (15); and improving patient health outcomes, for example, through optimizing caregiver workflows and resources or predicting patients most at risk for adverse events (16). These applications have produced state-of-the-art results, frequently surpassing clinical intuition.

Random forest (RF) is a type of ML method that offers interpretability while retaining predictive power of other ML algorithms (17). RFs have been applied to medical data sets for which prediction power as well as interpretability are often necessary. For example, RFs were used to predict early rejection in kidney transplantation patients using features including: demographic data (eg, sex, age), number of years on dialysis, cytometry crossmatch, and total number of human leukocyte antigen mismatches between donor and recipient (18). Using a small sample size of just 80 patients, RFs were able to not only predict which patients were more likely to suffer rejection with an accuracy of 85%, but they were able to identify key

risk factors associated with acute rejection. With respect to thyroid cancer diagnosis, RFs were used to classify nodules as benign vs malignant using tissue microarray data of 100 benign and 105 malignant thyroid lesions (19). In this instance, the RFs achieved an accuracy of 91% and were used not only for prediction but for understanding which variables were important for distinguishing between benign and malignant nodules. In another study, an RF classifier performed better than a radiologist at predicting malignancy of thyroid nodules using 11 sonographic features extracted from 2064 adult thyroid nodules (20). In this case, the RF classifier achieved an area under the curve of 0.938 compared to the radiologist's area under the curve of 0.843.

While the advantages of applying RFs in the medical realm are clear, our work pushes their application a step further. This paper presents interpretable and easily implementable diagnostic aids for physicians who want to predict the need for thyroid biopsies and surgery for pediatric patients. To our knowledge, our work is the first to approach this task using demographic, ultrasound, and biopsy data already collected from patient records. Furthermore, our computational pipeline selects only the most statistically important variables for the construction of the final model, a model that transforms the RF into an even more interpretable rule set (21). Each rule is presented to the physician with an associated accuracy—namely, the number of patients used to construct that rule, and the respective rates of benign and nonbenign histology. The innovation of the present rule set is that there does not exist a single decision tree that could capture the complexity of the proposed decision-making process yet the clarity of the proposed model is on par with a decision tree. These historical data are then used to generate predictions about future patients based on their similarity to previous patients, thus demystifying the model. Additionally, we can refine and strengthen the confidence of rules as more data are observed in the clinic, thereby ensuring that the resulting model is the state of the art as it continues to be used in the clinic. Ultimately, we envision broad clinical adoption of such models to aid in the determination of the need for biopsy and surgery, resulting in substantial reduction of the burden of unnecessary thyroidectomy and improved quality of life for the population of patients with thyroid nodules.

Materials and Methods

Experimental Design

This study was approved by the SickKids Research Ethics Board. Consent was waived for retrospective chart review. Medical records were reviewed for 198 consecutive patients with thyroid masses at a single tertiary care institution. As an inherent limitation of retrospective studies, complete data were available for a minority of cases. Exclusion criteria included inadequate quality of original ultrasound (primarily for older studies) or absence of data recorded in the medical chart. In addition, historically, a significant proportion of patients proceeded directly to thyroidectomy without prior biopsy. This was predicated on a recognition of high rates of malignancy when compared to adult nodules and lack of data supporting use of FNAB in children. This practice has since been superseded at our institution, but is still reflected in this historical data set. In total, 55 out of 198 patients (27.78%) proceeded directly to thyroidectomy, and therefore all were excluded from analysis. Where available, preoperative ultrasounds were prospectively reviewed using prespecified criteria by 3 radiologists, blinded to surgical histology and clinical outcome, as previously described (4). Of the remaining 143 patients, preoperative ultrasound studies of adequate diagnostic quality were available for 140 patients (69 patients with malignant nodules and 71 benign) and thus were potentially eligible to be included in our study subject to missingness. Final diagnosis was determined based on surgical histology. For nonoperative cases, a minimum 2-year follow-up without disease progression was established as the criterion for likely benign disease.

Predicting the Need for Biopsy and Surgery

Following retrospective chart analysis, we dichotomized the cohort into those with thyroid malignancy (based on surgical histology) or with presumed benign disease (based on either histology or, for patients managed nonoperatively, absence of clinical progression after a minimum of 2 years' follow-up). The objective of this analysis was to test the hypothesis that a computational approach could use preoperative variables to predict nodules unlikely to be benign, thus "needing" surgery. We then applied ML algorithms to identify those factors most predictive of malignant histology.

In a subsequent iteration, we "blinded" the algorithms to cytopathology and applied a similar approach to the prediction of cytopathology, based exclusively on clinical parameters and sonographic features. The rationale for doing so was to identify whether a computational model could achieve a satisfactory prediction of nodules unlikely to have benign cytology. Those nodules predicted by the model to have Bethesda 3 to 6 cytopathology were categorized as "nonbenign" cytology. Derivation of test performance is detailed later. We assessed this data set using the MTNS, as modified for pediatrics by Canfarotta et al (12). Inasmuch as the present series did not assess nodules

for interval growth, this variable was excluded from the MTNS calculations.

Statistical Analysis

In our study, we had one patient with new-onset Hashimoto thyroiditis who presented with a thyroid mass and was contemporaneously found to have a thyrotropin level of 89.60 (when the range without this patient was [0.01-17.00]), and one person with a purely cystic composition of the nodule. These are too few patients (outliers) for an algorithm to draw conclusions. From a modeling perspective, outliers are excluded because it is dangerous to assume their presentation is a consistent pattern across all patients and therefore build this pattern into the model if we have only 1 or 2 patients who fit the criteria. Given that our data set is small, withholding a separate test set was not feasible. Thus, we used k-fold cross-validation (CV) to estimate the performance we can expect on a withheld test set. CV works by partitioning patients into different groups, using k-1 folds for training and retaining one for testing, and repeating for all folds. To address the class imbalance for the surgery prediction model, the majority class was downsampled to create a class imbalance ratio of at most 40:60.

Discovering Complex Patterns in the Data

We built an RF approach to discover complex patterns between variables in the training data: The algorithm creates sets of rules that are applied sequentially in order to iteratively split a cohort of patients into smaller and smaller groups based on sets of similar characteristics (17). These complex patterns are used to group patients of similar histology so that the final rule applied separates benign from malignant patients. Individual trees in the forest are not interpretable in isolation, and the trees may seem counterintuitive when viewed individually. Therefore, our final models are interpretable rule sets based on the RFs, described later. A new test patient is classified as having likely benign or likely nonbenign histology by feeding the information through the rule sets. This limitation of RFs is addressed later.

Creating Models for a Clinical Setting

Given the need for use in a clinical setting, we created an easy-to-use, interpretable model. We did this by using only the most predictive variables to reduce the size and complexity of the model while retaining predictive accuracy. In our clinical context, failing to detect a malignant nodule—a false negative (FN)—is worse than performing a biopsy or surgery on what turns out to be a benign nodule—a false positive (FP); the former could result in a patient who has cancer not receiving adequate intervention. Thus, we redefine importance in the standard RF mean decrease in accuracy importance feature algorithm: instead of treating FNs and FPs as equally bad, an important predictor prioritizes decreasing FNs first, followed by FPs. In this way, we first minimized false-negative rate (FNR)—the proportion of people for whom the model fails to detect a malignant nodule but who actually have malignant nodules (FN/FN + true positive), followed by the false-positive rate (FPR)—the proportion of people who incorrectly undergo a biopsy or incorrectly undergo surgery out of all people who have benign nodules (FP/FP + true negative).

Our procedure consists of 3 steps. First, individual feature performance was calculated in the same way as RF using out-of-bag observations (17), however, using FNR and FPR as importance metrics. Second, we ranked features by sorting them based on the largest decrease in FNR followed by FPR. The feature with the smallest decrease was deemed least important (17). If there was a tie in the lowest FNR and FPR, we randomly selected a feature to remove. Third, having established an importance measure, we then selected the optimal subset of features for prediction using backwards feature elimination (22) and the mean decrease in FNR/FPR criteria described earlier.

We started with the full set of demographic, ultrasound, and biopsy predictors and iteratively removed the least informative feature at each step. The optimal number of features was determined by looking at the size of the predictor set that resulted in the lowest FNR followed by lowest FPR during the backwards feature elimination process; if there was a tie, the model with the smallest number of predictors was chosen. After performing CV to estimate the performance we can expect on a withheld test set, we reperformed the whole procedure on the entire data set and the model with the optimal number of important predictors was selected as the final model.

Applicability to Clinical Settings

Given that we are building a model that will be used in the clinical domain, interpreting the exact role of individual features for prediction can be difficult for each individual patient. Thus, following model building and assessment of feature importance, we extracted a set of rules to represent the forest. These rules have the advantage of being shorter and interpretable by clinicians. We used inTrees, a package in R, to perform this function (21). inTrees extracts rules from RFs based on minimizing the error rate and maximizing the frequency of observations that were used to construct the rule. Thus, we use it to transform our RF into rule sets.

From a clinical perspective, determining the likelihood or increase in odds of having malignant histology is preferable to strictly assigning binary labels. One challenge with rules sets is that they assign hard labels to each rule. In addition, given the sample size, accurate probabilities cannot be derived for our rule set. In lieu of probabilities, the relative frequencies of the number of patients classified by each rule will be used as an indication of the accuracy of each rule. This means that when a patient is classified as likely benign histology by our final rule set, the clinician will be able to see the number of historical patients who were classified using this rule. In this way, the rule set acts as a diagnostic aid to clinicians, not as a black box that merely puts out labels.

Results

Predicting the Need for Biopsy

Of the 140 potentially eligible patients in the entire cohort, only 67 had no missing values for all entries, and thus had sufficient data for biopsy prediction. Of these patients, 28 had benign histology and 39 had nonbenign. Table 1 shows historical patient outcomes. If a nodule was suspicious (based on clinical suspicion and/or sonographic features), biopsy was performed. If nodules were strongly suspected of being benign, biopsy was deferred. Outcome was then determined by histology and/or clinical follow-up as defined in "Materials and Methods."

Of the 5 nodules without biopsy (and with adequate follow-up), none progressed to malignancy, suggesting adequate stratification of very low-risk nodules. We acknowledge a significant negative bias included here, as most patients with very low-risk presentation (for example unambiguous colloid cysts) may not have had ongoing follow-up at the tertiary site, and would have been excluded from this analysis. Thus this category is certainly underrepresented.

Table 1. Historical cytologic outcomes based on predictions made on clinical impressions for all patients

		Prediction based on clinical impression	
		Benign	Nonbenign/ Uncertain
Cytology or 2-year follow-up	Benign	5	23
	Nonbenign	0	39

If a nodule was suspicious and therefore deemed nonbenign or uncertain (based on clinical suspicion and/or sonographic features), biopsy was performed. If nodules were strongly suspected of being benign, biopsy was deferred. Outcome was then determined by histology and/or clinical follow-up as defined in "Materials and Methods."

Among patients who were felt to merit FNAB, 24 had benign cytology (Bethesda 2) and 39 had indeterminate or malignant cytology. We asked whether a computational model could better identify those patients who would ultimately go on to have benign cytology, based solely on clinical variables and ultrasound, thereby avoiding biopsy altogether.

Table 2 describes the outcome of these models derived from the retrospective data. Both models, RF and rule set, reduced the biopsy rate at the expense of the FN (nodules for which biopsy was not recommended, but that would turn out to be malignant) rate.

The rule set is the final derived model. Also shown are the test characteristics of the RF model. Our model was able to identify the need for biopsy with an accuracy of 77.57% (± 5.07% SD) compared with the historical accuracy of 65.67%, an increase of 11.90%. For all patients who have nonbenign cytology, we expect 10.27% (± 6.78% SD) would be incorrectly identified as not needing a biopsy compared to a historical rate of 0.00%. This means our prediction would perform worse than the clinical impression alone by 10.27%. For those with benign histology, we expect 14.10% (± 5.43% SD) would be incorrectly identified as needing a biopsy compared to a historical rate of 82.14%. This means our prediction would perform better than the historical rate by 68.04%. The area under the receiver operator curve (AUROC) is 83.78 (± 4.46% SD) compared to the historical practice of 58.93, indicating a 24.85% increase in performance. In summary, using the rule set model to identify nodules for biopsy would reduce the number of biopsies performed unnecessarily (for benign nodules) while maintaining a low miss rate for nodules with nonbenign cytology. Our final biopsy rule set, trained using all the historical data, is presented in Table 3. Comparison of historical outcomes, with current practice (based on sonographic risk assessment) (23) and the rule set is summarized in Table 4. We also asked whether the MTNS, as modified for pediatrics (12), could discriminate between benign and nonbenign cytology. The MTNS, however, relies heavily on the *results* of cytology to generate a score to predict malignancy, thus it cannot be used in its current incarnation to ascertain the *need* for biopsy. We analyzed the data from our series using the MTNS, after excluding the cytology scores and these are presented in Supplementary Figure 1 (24).

Modeling Histology Among Patients With Benign, Insufficient or Indeterminate Biopsy Results

We also asked whether an ML model could predict malignant histology among nodules with nonmalignant cytology (Bethesda 1-5). The rationale for this was that any patient

Table 2. Machine learning prediction of benign and nonbenign cytology

	Accuracy, % (± SD)	False-negative rate, % (± SD)	False-positive rate, % (± SD)	Area under receiver operator curve, % (± SD)
Historical practice (clinical formulation)	65.67	0.00	82.14	58.93
Random forest	83.55 ± 1.58	12.50 ± 4.79	21.43 ± 9.22	83.04 ± 2.48
Rule set	77.57 ± 5.07	10.27 ± 6.78	14.10 ± 5.43	83.78 ± 4.46

Compares the results of historical practice to random forest classifier and the simplified rule set using 4 measures of performance.

Table 3. Example of a final rule set to determine indication for biopsy

Rule No.	Rule	Decision	Historical No. of patients that correctly satisfy specific rule	Historical No. of patients that correctly satisfy all rules
1	Composition of nodule is entirely solid LNs appear normal Tumor is unifocal	Likely nonbenign—recommend biopsy	12/12	12/12
2	Nodule > 50% cystic Margin is regular LNs appear normal	Likely benign-defer biopsy	7/7	19/19
3	Composition of nodule is entirely solid Hypoechoic halo is either absent OR complete but not partial Margin is irregular/microlobulated/ spiculated	Likely nonbenign–recommend biopsy	4/4	23/23
4	Composition of nodule is entirely solid Margin is indistinct Tumor is unifocal OR multifocal (unilaterally)	Likely nonbenign–recommend biopsy	9/10	32/33
5[a]	Composition of nodule is mixed solid/ cystic < 50% cyst Hypoechoic halo is absent OR complete (ie, not absent) Tumor is unifocal OR multifocal (bilaterally)	Likely benign–defer biopsy	10/11	42/44
6	Composition of nodule is entirely solid LNs are enlarged but normal appearing OR are suspicious for metastasis	Likely nonbenign—recommend biopsy	6/7	48/51
7[a]	LNs are not visualized or are visualized contralateral to primary tumor (but not ipsilaterally) Tumor is multifocal (unilaterally)	Likely benign—defer biopsy	3/4	51/55
8	Composition of nodule is entirely solid Margin is indistinct Tumor is multifocal and bilateral	Likely nonbenign—recommend biopsy	3/4	54/59
9	Otherwise	Likely nonbenign—recommend biopsy	3/8	57/67

Abbreviation: LN, lymph node.
[a]These misclassification are the least acceptable type of error—patients classified as likely benign when they were not.

with malignant (Bethesda 6) cytology would de facto merit surgery. The converse, at least in children, is not necessarily true, in that the FNR of benign fine-needle aspiration cytology (malignant histology in a nodule with benign biopsy) is higher than in adults (6,7). Additionally, sampling error may lead to missed malignancy among large nodules greater than 3 cm. As such, we elected to include nodules with benign (Bethesda 2) cytology in our analysis. Included in these data were 3 clinically "FN" nodules with benign cytology, which were ultimately demonstrated to harbor malignancy, based on surgical histology.

Of the 140 potentially eligible patients, 53 had benign, insufficient, or indeterminate biopsy results and sufficient data to build a model. Of these, 42 had benign histology and 11 were malignant. Table 5 shows patient outcomes according to historical practice. The malignancy rate in those

with nonmalignant cytology who underwent surgery was 11 out of 40 (27.5%). Stated otherwise, 72.5% of patients underwent potentially avoidable surgery for benign disease, had more accurate preoperative stratification been available. We therefore set out to determine whether a predictive model could reduce this rate. The results of our final rule set are included in Table 6. Our model predicted malignancy with an accuracy of 77.47% (± 2.71% SD) compared with the historical accuracy of 45.28%, an increase of 32.19%. If this model were to replace current practice, we expect 11.67% (± 1.32% SD) would be triaged to nonoperative management, compared to a historical practice of 0.00%.

Table 4. Comparison of decision making for the need for biopsy according to historical practice, current practice and our biopsy rule set model

Cytology or 2-year follow-up[a]	Predicted to not need/need biopsy		
	Historical data set	All patients evaluated according to current practice[b]	Rule set model
Benign	5[a]/23	13/15	20/8
Nonbenign	0/39	3/36	2/37

[a]Only patients with minimum 2-year follow-up were included.
[b]Decision to pursue biopsy based on clinical and sonographic features.

Table 5. Historical outcomes based on predictions made on clinical impressions for patients with nonmalignant cytology (Bethesda 1-5)

		Prediction based on clinical impression	
		Suspected benign—managed nonoperatively	Uncertain (cannot exclude malignancy) —underwent surgery
Histology	**Benign**	13	29
	Malignant	0	11

If a nodule was uncertain (based on clinical suspicion and/or sonographic features and/or biopsy results), surgery was performed. If nodules were strongly suspected of being benign, surgery was deferred. Outcome was then determined by histology and/or clinical follow-up as defined in "Materials and Methods".

While at face value this seems a high "miss" rate, it must be interpreted in the context of the typically indolent nature of papillary thyroid carcinoma in children, for whom opportunity for surgical salvage with excellent outcomes exists. Ongoing follow-up of such patients would still afford the opportunity for surgical cure with progression of underlying disease, while nonprogressive disease could be monitored indefinitely.

This model would still endorse "unnecessary" surgical management in 45.83% of patients; however, this reflects a reduction of 23.22% over historical practice. In other words, this model would spare 1 patient in 4 unnecessary surgery. The AUROC is 61.64 (± 10.28% SD) compared to a historical value of 65.48%, indicating a 3.84% decrease in performance. Our final surgery rule set, trained using all the historical data, is presented in Table 7. We compare the historical data with predictions based on the "modified" MTNS, current practice based on American Thyroid Association criteria (23) and the rule set model in Table 8.

Discussion

This approach represents a first-pass effort at applying an ML solution to identifying those patients and those nodules that would most benefit from biopsy and from surgical intervention. While clearly not appropriate for clinical decision-making at present, these analyses clearly demonstrate an opportunity for applying computational approaches to retrospective data to refine clinical decision-making.

These models are presently limited by an unacceptable "miss rate." Ongoing refinement of the models and larger multi-institutional data sets may help reduce this rate to one that approaches a clinically acceptable rate.

At no point would we envision an ML approach to supersede clinical intuition, experience, and data integration. Rather, this would be an ancillary tool to help refine and supplement diagnostic modalities, which themselves are fraught with imperfect predictive capacity, as evidenced by the high operative rates for benign disease.

Table 6. Machine learning prediction for benign vs nonbenign histology

	Accuracy, % (± SD)	False-negative rate, % (± SD)	False-positive rate, % (± SD)	Area under receiver operator curve, % (± SD)
Historical practice	45.28	0.00	69.05	65.48
Random forest classifier	83.24 ± 4.33	29.17 ± 17.18	14.09 ± 8.79	78.37 ± 4.96
Rule set	77.47 ± 2.71	11.67 ± 1.32	45.83 ± 20.83	61.64 ± 10.28

Compares the results of historical practice to random forest classifier and the simplified rule set using 4 measures of performance.

Table 7. Final surgery decisional rule set (after biopsy)

Rule No.	Rule	Decision	Historical No. of patients that correctly satisfy specific rule	Historical No. of patients that correctly satisfy all rules
1	Margin is regular Cytology is benign or inadequate (Bethesda 1 or 2)	Likely benign–defer surgery	27/27	27/27
2	There are no echogenic foci Solid component is hypoechoic or markedly hypoechoic LNs are not visualized or are visualized contralateral to the primary tumor (but not ipsilaterally) Cytology is benign (Bethesda 2)	Likely benign-defer surgery	5/5	32/32
3	LNs are not visualized or are visualized contralateral to primary tumor (but not ipsilaterally) Cytology is inadequate (Bethesda 1)	Likely benign-defer surgery	4/4	36/36
4	Solid component is hypoechoic or markedly hypoechoic LNs are not visualized or are visualized contralateral to the primary tumor (but not ipsilaterally) Cytology is indeterminate (Bethesda 3-5)	Likely nonbenign–consider surgery	6/7	42/43
5	Solid component is isoechoic, hyperechoic or mixed echogenicity Hypoechoic halo is absent Margin is irregular/microlobulated/spiculated OR indistinct Cytology is benign or indeterminate (Bethesda 2-5)	Likely nonbenign–consider surgery	3/5	45/48
6	Otherwise	Likely nonbenign–consider surgery	2/5	47/53

Abbreviation: LN, lymph node.

Table 8. Comparison of decision making for the need for surgery according to historical practice, Modified McGill Thyroid Nodule Score, cytology alone, and our surgery rule set model

Histology or 2-year follow-up	Predicted to not need/need surgery			
	All patients evaluated according to historical practice	**Modified McGill Thyroid Nodule Score ≥ 8/ ≥ 9**	Cytology alone	Rule set model
Benign	13/29	36/6 39/3	40/14	36/6
Nonbenign	0/11	3/33 4/32	2/22	0/11

Improvement in Prediction

Our models predict nonbenign cytology and malignant histology better than historical practice, with lower FPRs (electing for biopsy or surgery in the context of benign nodules), higher accuracy, and higher AUROC rates. Specifically, our biopsy predictions see improvement across all 3 measures, with notable improvements in the FPR; we expect a 68.04% improvement in the FPR, 11.90% increase in accuracy, and 24.85% increase in the AUROC. This indicates that our biopsy model comprises a simple set of rules is expected to outperform historical practice in determining the need for biopsy. In addition, there are also

significant refinements in identifying those patients with indeterminate or inadequate cytology most likely to harbor malignancy, with a notable increase in accuracy; we expect an 23.22% decrease in FPR, 32.19% increase in accuracy, and 3.84% decrease in the AUROC.

For most complex problems in the medical realm, researchers and physicians alike accept trade-offs between minimizing the FNRs and FPRs. For both our models, the improvements in FPR, accuracy, and AUROC came at the expense of increases in FNR, the most clinically unacceptable type of error. Specifically, for patients with nonbenign cytology, we expect 10.27% would be discouraged from performing biopsy, and 11.67% of those with malignancy would be incorrectly steered away from surgery if relying on the prediction models alone.

With relatively small cohort sizes, our models are limited in their ability to distinguish between benign and malignant disease. Given this limitation, we are encouraged by these preliminary results as they demonstrate learning is taking place and point to future directions of research. Thus, while the increase in the FNR is problematic, we expect our models to improve with increasing sample size.

Generalizability and Interpretability

To anticipate how well these models will perform, it is important to consider the concept of model overfit. Overfitting is a commonly encountered obstacle in ML that affects the ability of models to generalize to unseen data. Specifically, lack of generalizability occurs when a model performs so well that it learns patterns in the training data that are not present in the overall population. Instead, these patterns are nuances of a particular set of patients. In our case, this would mean our models perform well on our cohort, but would fail to predict well on 2 types of unseen patients: those that will be seen at our same institution in the future, or those from other institutions. Given the rarity of pediatric thyroid cancer, our models are trained on a relatively small number of patients. Since the sample size is so small and RF does well extracting patterns from the data, we want to mitigate the likelihood that RF is overfitting. Simplifying the RF to the rule set accomplishes 2 things: First, the model is more likely to generalize to withheld data and second, the model is more interpretable.

The improvement in the metrics measured demonstrates the improvement in generalizability we anticipated when converting from RFs to rule sets. Specifically, predicting malignant histology using the rule set results, instead of the RFs, results in a substantial improvement in the FNR of 17.5%—the metric we are most concerned with limiting. Thus, for patients with malignant histology, the rule set performs better at predicting malignancy. With respect to other metrics, the overall accuracy and AUROC decreased by 5.77% and 3.84% respectively due to the improvements in FNR coming at the expense of inappropriately including patients with benign histology. While the FPR increased by 31.74%, this new metric remains 23.22% better than historical performance, indicating an overall increase in the quality of the predictions in terms of the metrics we are most concerned with. In terms of predicting nonbenign cytology, by converting from RFs to rule sets, FNR and FPR saw improvements in performance by 2.23% and 7.33%, with marginal improvements in the AUROC of 0.74%, indicating a slight increase in performance when converting from the RF to rule set.

Second, rule sets are more straightforward to follow and actionable in the context of clinical care than are RFs. As a list of sequential rules that split the patients into groups based on their characteristics, the rule sets can be examined on 2 levels. First, the rules themselves are based on the relationship between the different features extracted from ultrasound and biopsy results. Thus, combinations of features can be examined to gain deeper insight into how these features are related on a biological level. Second, each rule is presented to the clinician with the number of patients from the training data who were predicted using this rule, providing the clinician an indication of the rule's historical accuracy. This historical data can then be used to make predictions on the likelihood of benign or nonbenign histology based on the similarity of a particular patient with the previous patients, thereby demystifying the model and providing an interpretable way of predicting the need for biopsies and surgeries.

Limitations and Next Steps

We acknowledge 2 major limitations related to the small sample sizes used to build these models. First, these models may be learning patterns from a cohort of patients that are fundamentally different from patients at other institutions or patients that will be seen in the future. When discussing generalizability, we addressed how it can be improved by using a rule set instead of an RF; in this approach, predictions are improved by changing the type of model we are using. While using a different model addresses the issue of overfitting, it does not fix biases that are learned because of differences between cohorts of patients. These differences can be overcome only by training on data that are representative of all types of pediatric thyroid patients, and thus would require more training data. Thus validation on a large external data set will be an important subsequent step in the refinement of this ML approach.

Second, we want to provide clinicians a probabilistic estimate of whether a risk-benefit favors biopsy and surgery: Either a likelihood estimate or change in odds of a patient benefitting from a biopsy or surgery are more clinically useful compared

with hard predictions of likely benign vs likely malignant histology. When discussing interpretability, we addressed how we overcame this challenge: Each rule has an associated relative frequency of patients from the training data who were captured by that rule, providing the clinician an indication of the rule's historical accuracy. While these relative frequencies are helpful as they provide a clinician historical evidence as to the accuracy of a particular rule, generating a probabilistic interpretation would still be ideal because it is the more clinically relevant measure. However, this probabilistic interpretation is hindered by the small sample size because currently a consistent probability estimate cannot be generated.

Given the rarity of pediatric thyroid cancer, sourcing more data can be a challenge. Thus, future work will involve collecting more data to refit the models to improve the prediction accuracy, test these models on an external data set, and generate consistent probability estimates for the rule set.

Conclusion

This study summarizes initial experiences using an ML approach to integrate clinical and sonographic data to model cytologic outcomes and to integrate clinical, sonographic, and cytologic data to model likelihood of malignancy. In routine practice, clinicians integrate these data routinely to identify those patients most appropriate for biopsy and/or surgery; however, this "gestalt" approach is limited by the experience of the clinician and the completeness of the available data. While the present retrospective study did not generate a model adequate to replace existing practice, largely because of the limitation in cohort size with sufficient data points, the improved accuracy and AUROC for identifying biopsy and surgical candidates are encouraging. This serves as a proof of principle that systematic data ascertainment and mathematical modeling may eventually facilitate the development of a powerful tool to help guide clinical decision-making and to avoid unnecessary interventions. Expansion of the training data sets using additional pediatric and adult data may accomplish this goal.

Acknowledgments

Financial Support: This work was supported in part by the Garron Family Cancer Center. AG and LR were supported by Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, Canadian Institute for Advanced Research and Genome Canada. The authors are grateful to the Rare Disease Foundation for financial support of this project.

Author Contributions: J.D.W. and A.G. conceived the study and supervised the work. D.C.M.v.d.K. performed chart review and data collection. L.R. processed the experimental data, performed the data analysis and generated models and rule sets, drafted the manuscript, and designed the figures. All authors discussed the results and contributed to the final manuscript.

Additional Information

Correspondence: Anna Goldenberg, PhD, Division of Endocrinology, The Hospital for Sick Children, 686 Bay St, Toronto, ON, M5G OA4, Canada. Email: anna.goldenberg@sickkids.ca.

Disclosures: The authors have nothing to disclose.

Data Availability: Restrictions apply to the availability of some or all data generated or analyzed during this study to preserve patient confidentiality or because they were used under license. The corresponding author will on request detail the restrictions and any conditions under which access to some data may be provided.

Appendix

Package Specifications

All models were built using R version (version 3.3.2) [RRID:SCR_001905] (25). The RF models were built using RandomForest package (version 4.6-12) [RRID:SCR_015718] (26). The number of features (mtry) was set to the default value of the square root of the number of features. Given the small sample sizes and to improve interpretability, a small number of trees (ntrees parameter) 3, 5, and 7 trees were tested through CV, and ntrees was set to 5 for both models. The subset of rules was extracted using the inTrees (version 1.1) package [RRID:SCR_017299] (27).

References

1. Vergamini LB, Frazier AL, Abrantes FL, Ribeiro KB, Rodriguez-Galindo C. Increase in the incidence of differentiated thyroid carcinoma in children, adolescents, and young adults: a population-based study. *J Pediatr.* 2014;**164**(6):1481-1485.
2. Gupta A, Ly S, Castroneves LA, et al. A standardized assessment of thyroid nodules in children confirms higher cancer prevalence than in adults. *J Clin Endocrinol Metab.* 2013;**98**(8):3238-3245.
3. Al Nofal A, Gionfriddo MR, Javed A, et al. Accuracy of thyroid nodule sonography for the detection of thyroid cancer in children: systematic review and meta-analysis. *Clin Endocrinol (Oxf).* 2016;**84**(3):423-430.
4. Martinez-Rios C, Daneman A, Bajno L, van der Kaay DCM, Moineddin R, Wasserman JD. Utility of adult-based ultrasound malignancy risk stratifications in pediatric thyroid nodules. *Pediatr Radiol.* 2018;**48**(1):74-84.
5. Gannon AW, Langer JE, Bellah R, et al. Diagnostic accuracy of ultrasound with color flow Doppler in children with thyroid nodules. *J Clin Endocrinol Metab.* 2018;**103**(5):1958-1965.
6. Cherella CE, Angell TE, Richman DM, et al. Differences in thyroid nodule cytology and malignancy risk between children and adults. *Thyroid.* 2019;**29**(8):1097-1104.
7. Amirazodi E, Propst EJ, Chung CT, Parra DA, Wasserman JD. Pediatric thyroid FNA biopsy: outcomes and impact on management over 24 years at a tertiary care center. *Cancer Cytopathol.* 2016;**124**(11):801-810.

8. Richman DM, Cherella CE, Smith JR, et al. Clinical utility of sonographic features in indeterminate pediatric thyroid nodules. *Eur J Endocrinol*. 2021;**184**(5):657-665.
9. Canadian Pediatric Thyroid Nodule (CaPTN) Study Group. The Canadian Pediatric Thyroid Nodule Study: an evaluation of current management practices. *J Pediatr Surg*. 2008;**43**(5):826-830.
10. Sands NB, Karls S, Amir A, et al. McGill Thyroid Nodule Score (MTNS):"rating the risk," a novel predictive scheme for cancer risk determination. *J Otolaryngol Head Neck Surg*. 2011;**40**(Suppl 1):S1-S13.
11. Scheffler P, Forest VI, Leboeuf R, et al. Serum thyroglobulin improves the sensitivity of the McGill Thyroid Nodule Score for well-differentiated thyroid cancer. *Thyroid*. 2014;**24**(5):852-857.
12. Canfarotta M, Moote D, Finck C, et al. McGill Thyroid Nodule Score in differentiating benign and malignant pediatric thyroid nodules: a pilot study. *Otolaryngol Head Neck Surg*. 2017;**157**(4):589-595.
13. Creo A, Alahdab F, Al Nofal A, Thomas K, Kolbe A, Pittock S. Diagnostic accuracy of the McGill Thyroid Nodule Score in paediatric patients. *Clin Endocrinol (Oxf)*. 2019;**90**(1):200-207.
14. Giger ML. Machine learning in medical imaging. *J Am Coll Radiol*. 2018;**15**(3 Pt B):512-520.
15. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet*. 2015;**16**(6):321-332.
16. Bates DW, Heitmueller A, Kakad M, Saria S. Why policymakers should care about "big data" in healthcare. *Health Policy Technol*. 2018;**7**(2):211-216.
17. Breiman L. Random forests. *Mach Learn*. 2001;**45**(1):5-32.
18. Shaikhina T, Lowe D, Daga S, Briggs D, Higgins R, Khovanova N. Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. *Biomed Signal Process Control*. 2019;**52**:456-462.
19. Wiseman SM, Melck A, Masoudi H, et al. Molecular phenotyping of thyroid tumors identifies a marker panel for differentiated thyroid cancer diagnosis. *Ann Surg Oncol*. 2008;**15**(10):2811-2826.
20. Zhang B, Tian J, Pei S, et al. Machine learning-assisted system for thyroid nodule diagnosis. *Thyroid*. 2019;**29**(6):858-867.
21. Deng H. Interpreting tree ensembles with inTrees. *Int J Data Sci Anal*. 2019;**7**(4):277-287.
22. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res*. 2003;**3**(Mar):1157-1182.
23. Francis GL, Waguespack SG, Bauer AJ, et al.; American Thyroid Association Guidelines Task Force. Management guidelines for children with thyroid nodules and differentiated thyroid cancer. *Thyroid*. 2015;**25**(7):716-759.
24. Radebe L, Goldenberg A, Wasserman J. Supplementary Figure 1—application of the modified (pediatric) McGill Thyroid Nodule Score (MTNS). 2021. https://figshare.com/articles/figure/Supplementary_Figure_1_-_Application_of_the_modified_Pediatric_McGill_Thyroid_Nodule_Score_MTNS_/1455597.
25. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2014. http://www.R-project.org/
26. Liaw A, Wiener M. Classification and regression by randomForest. *R News* 2002;**2**(3):18-22. https://cran.r-project.org/web/packages/randomForest/.
27. Houtao D. Interpreting Tree Ensembles with inTrees. Technical Report, 2014. https://cran.r-project.org/web/packages/inTrees/.