A Bayesian Classifier for Differentiating Benign versus Malignant Thyroid Nodules using Sonographic Features

Yueyi I. Liu, PhD[1, 2], Aya Kamaya, MD[1], Terry S. Desser, MD[1]
Daniel L. Rubin, MD, MS[1, 2],
[1]Department of Radiology, Stanford University, Stanford, CA; [2]Stanford Medical Informatics, Stanford University, Stanford, CA

Abstract

Thyroid nodules are a common, yet challenging clinical problem. The vast majority of these nodules are benign; however, deciding which nodule should undergo biopsy is difficult because the imaging appearance of benign and malignant thyroid nodules overlap. High resolution ultrasound is the primary imaging modality for evaluating thyroid nodules. Many sonographic features have been studied individually as predictors for thyroid malignancy. There has been little work to create predictive models that combine multiple predictors, both imaging features and demographic factors. We have created a Bayesian classifier to predict whether a thyroid nodule is benign or malignant using sonographic and demographic findings. Our classifier performed similar to or slightly better than experienced radiologists when evaluated using 41 thyroid nodules with known pathologic diagnosis. This classifier could be helpful in providing practitioners an objective basis for deciding whether to biopsy suspicious thyroid nodules.

Introduction

Thyroid nodules are extremely common—found in 4-8% of adults by palpation, 10-41% by ultrasound, and 50% at autopsy[1,2]. High resolution ultrasound is the primary imaging modality for evaluating these nodules. Current management guidelines from the American Thyroid Association recommends diagnostic thyroid ultrasound for all patients with thyroid nodules[3]. Furthermore, potentially malignant nodules should undergo ultrasound-guided fine needle aspiration (FNA) to achieve tissue diagnosis.

In contrast to the high prevalence of thyroid nodules, thyroid cancer is rare. Fewer than 7% of all nodules are malignant[4]. The American Cancer Society estimates that 33,550 new cases of thyroid cancer will be diagnosed in 2007[5]. The vast majority, approximately 88% of these cancers, will be papillary thyroid carcinoma[6]. These are usually slow growing cancers with excellent prognosis. Other histological types of thyroid cancer include follicular (5-10%), medullary(3-5%), anaplastic (1-2%), lymphoma (1-2%), and thyroid metastasis from other cancers (<1%).

Even though only a small fraction of all nodules are malignant and thyroid cancers generally have good prognosis, the morbidity and mortality rates increase in relation to the stage of the disease. Therefore, in order to diagnose thyroid cancers early on, but spare other patients from the risks of unnecessary fine needle aspirations, it is essential to have a strategy to determine which nodules should undergo FNA.

Many sonographic features have been described and studied as potential predictors of thyroid malignancy. These include size, multiplicity, echogenicity, presence of microcalcifications, margin, contour, shape, architecture, and vascularity. For example, microcalcifications are present in 26-59% of all thyroid cancers, and hypoechogenicity is present in 26-87% of all thyroid cancers (see [7] for review). Moreover, it is clear that no single feature can distinguish benign from malignant nodules. Most of the articles published focus on the sensitivity, specificity, and positive predictive values of individual features of thyroid cancer. Eight classic patterns highly suggestive of benign or malignant nodules have been described[8]. For example, a solid hypoechoic nodule with microcalcifications is highly suggestive of papillary thyroid carcinoma. However, less than half of all thyroid nodules fit into one of the classic patterns suggestive of benignity or malignancy. To prevent the cost (both risk of complications to patients and financial cost) of biopsying benign nodules, we desire a robust model to estimate the probability of a given thyroid nodule as being malignant (i.e., needing an FNA) given the multiplicity of sonographic features that are mutually informative of the underlying disease. This problem lends itself well to a Bayesian classifier.

A Bayesian classifier, which is a form of Bayesian network, consists of nodes (representing variables) and edges connecting the nodes. The relationships among the variables are represented by the direct acyclic graph. Bayesian classifiers have been used in many areas of medicine. For example, Burnside et al. built a Bayesian classifier to predict breast cancer risk based on mammography findings[9]. Kline et al. created a classifier to identify a low-risk subset of patients suspected of having a venous thromboembolism using clinical data that are readily available[10].

Given that sonographic features predictive of malignancy have been extensively studied and the sensitivity and specificity of these features for malignancy are readily available, we hypothesize that a Bayesian classifier can be created using data from literature supplemented by expert knowledge to model thyroid nodules. To our knowledge, this has not been reported in the literature to date.

Materials and Methods

In order to construct our Bayesian classifier and perform inference, we used the Netica development environment (http://www.norsys.com). We created a Bayesian network (BN) comprising a node for disease and nodes for the observed sonographic findings and patient demographics. We represented the pathology of thyroid nodules as a disease node with two states (benign vs. malignant). All sonographic features known to be predictors of malignancy were included in the BN. Given that age and gender also significantly influence the probability of a nodule being malignant, we included these demographic features in our as well. The structure of our BN is shown in Figure 1. The pretest probabilities of thyroid malignancies by age and gender were derived from the SEER database (http://seer.cancer.gov/faststats/sites.php?site=Thyroid+Cancer). We discretized age to <50 (which accounts for 75% of the population) and >=50 (25% of the population). Female and male each account for half of the population. Table 1 shows the conditional probability table for malignant nodules given age and gender.

The conditional probability tables for the sonographic features are constructed using literature and expert knowledge. We performed an extensive literature review to identify all papers that discuss sonographic features as predictors of thyroid malignancy. Sensitivity and specificity of these features, as well as the number of nodules used to derive these numbers, were gathered. We then calculated the average of the sensitivity and specificity for each feature weighted by the number of nodules in each study. We also obtained the sensitivity and specificity for each feature from a radiologist specializing in thyroid ultrasound. For most of the features, the weighted average of the sensitivity and specificity from literature agree very well with expert opinion. When they do not agree or when it is not possible to calculate a weighted average for a feature, we chose to follow the expert opinion. The details of the eight sonographic features are shown in table 2.

To evaluate our classifier, we randomly selected 21 benign thyroid nodules and 20 malignant nodules from 37 patients who underwent ultrasound guided FNA in 2007 and early 2008. All final diagnoses were determined by pathology. The 20 malignant nodules included 18 papillary thyroid carcinomas, one lymphoma, and one poorly differentiated carcinoma. Follicular lesions were not included since FNA, the test used to establish the final diagnosis in our study, cannot distinguish benign follicular adenomas from malignant follicular carcinomas.

We also compared the performance of our classifier to that of two radiologists specializing in thyroid ultrasound, one with five years of experience (radiologist 1), the other with 20 years of experience (radiologist 2). They each rated each nodule on a scale of 1 to 5 (1-benign, 2-probably benign, 3-not sure, 4-probably malignant, 5-malignant). Receiver operating characteristic (ROC) curves were generated using the ROCKIT 1.1B software (http://www-radiology.uchicago.edu/krl/KRL_ROC/software_index6.htm). One radiologist was aware of the case mix (number of benign versus malignant nodules), and the other was not.

Results

We evaluated our model using 41 thyroid nodules from 37 patients. ROC curves of our classifier and the radiologists' predictions are shown in Figure 2. The area under the curve (Az) value of our model is 0.851 (95% confidence interval (CI): 0.745-0.939), which is similar to or slightly better than those of the radiologists (0.846 (CI: 0.678-0.943) for radiologist 1 and 0.719 (CI: 0.543-0.854) for radiologist 2).

To evaluate the potential of our classifier, we investigated which nodules should be biopsied by looking at sensitivity and specificity at different decision thresholds. If we set sensitivity to be 100% (i.e., all malignant nodules are biopsied), 16 of the 21 benign nodules will also be biopsied, resulting in a specificity value of 33.3%. If we lower sensitivity to 80% (i.e., 4 of the 20 malignant nodules will be missed), specificity increases to 76%, i.e., only 5 of the 21 benign nodules will be biopsied.

We will illustrate the behavior of our Bayesian classifier using several cases.

Case 1. A nodule in a 44 year old woman. The nodule was hypoechoic, of mixed solid and cystic component, with intrinsic vascularity, smooth margins, and ring down artifact. It was no taller than it was wide, no capsular invasion or microcalcification was seen. This was proven to be a benign nodule on pathology. Ring down artifact is almost exclusively associated with benign nodules [11]. Our classifier predicted this nodule to be benign with a probability of 99.99%.

Case 2. A nodule in a 62 year old man. The nodule was solid, hypoechoic, taller than it was wide in shape, with ill-defined margins and intrinsic vascularity. Both microcalcifications and capsular invasion were present. No ring down artifact was present. This was pathologically proven to be a malignant nodule, concordant with multiple sonographic features associated with malignancy (solid, taller than wide in shape, ill-defined margins, microcalcification, capsular invasion). Our classifier predicted this nodule to be malignant with a probability of 99.96%.

Case 3. A nodule in a 44 year old woman. The nodule was solid, hypoechoic, with smooth margins and intrinsic vascularity. It was not taller than wide in shape, had no microcacification, capsular invasion, or ring down artifact. Several of the sonographic features are associated with benign nodules (smooth margin and no microcalcification), yet other features are associated with malignancy (solid, hypoechoic nodule, intrinsic vascularity). This nodule turned out to be malignant on pathology. According to our classifier, the posterior probability of the nodule being malignant is 49.7%. The two radiologists both rated this nodule as 2 (probably benign).

Discussion

We built a Bayesian network for thyroid nodule classification. Though not truly a naïve Bayes classifier as the node Thyroid Nodule does have parents, our classifier is very similar to a Naïve Bayes classifier in that it assumes that all the sonographic features are conditionally independent. This is a strong assumption, and could be too-simplistic. However, in the case of thyroid nodule evaluation, the domain experts believed the assumption that each feature is an independent feature is a good assumption. Furthermore, despite the independence assumption, naïve Bayes classifiers have been previously shown to perform well in classification tasks. Recently, Zhang et al. offered some theoretical reasons behind the surprisingly good performance of naïve Bayes classifiers[12].

In our initial evaluation of the BN using 41 nodules from 37 patients, we showed that our classifier performed similarly or slightly better than expert radiologists. One of the radiologists (radiologist 2) evaluated the ultrasound images completely blinded. The other radiologist (radiologist 1), though unaware of the final diagnosis of each nodule, was familiar with the cases by enumerating the sonographic features of each nodule. Hence, radiologist 1 was likely biased by awareness of the prior probability of a nodule being malignant in our test cases. Therefore, it is not surprising that radiologist 1 had slightly better accuracy than radiologist 2 (figure 2). We plan to undertake a more thorough evaluation in which only cases which have never been seen by either radiologist will be used to reduce potential bias.

We obtained the conditional probability of our Bayesian network from two sources--both the weighted average of sensitivity and specificity reported in the literature and expert opinion. For most of the features, these two sources agree very well. When the two sources differed, we chose the expert opinion over that of the literature. This is because in these situations, the terminology used in the literature to describe the same feature is always variable. For example, irregular margins are sometimes described as “microlobulated”, “macrolobulated”, or “blurred”. There are also subtle differences in the definition of these terms. For one particular feature (vascularity), there is so much difference in the classification of vascularity into different stages and types that it is not possible to calculate a weighted average for the sensitivity/specificity. This strongly suggests the need for a controlled terminology in reporting the ultrasound imaging features of thyroid nodules. Controlled terminology is well established in other radiology domains such as breast imaging, where BI-RADS, a controlled terminology, is used to describe mammogram features. In our project, we made the first step towards creating a set of mutually exclusive but collectively exhaustive descriptors for common ultrasound features of thyroid nodules that could ultimately establish a controlled terminology for thyroid nodule evaluation.

A potential benefit of our BN is that it may be useful to radiologists to decide when to biopsy ultrasound nodules. As the prior probability of malignant nodules is low, many thyroid nodules currently undergoing biopsy are benign. If radiologists could use a tool such as our BN to objectively calculate the probability of malignancy, then they could make their decisions in a normative fashion. As a result this would potentially reduce the number of unnecessary thyroid biopsies and improve their positive predictive value. In fact, such improvement has been shown in applying BNs in mammography[13]. Since thyroid carcinomas are generally slow growing with good prognosis, it may be acceptable to have sensitivity lower than 100% in order to achieve a relatively high specificity to prevent unnecessary biopsies. At sensitivity of 80%, our classifier is quite specific –

only five benign nodules will be recommended for biopsy. Sensitivity of 80% means that four of the 20 malignant nodules would have been missed. All four nodules were papillary thyroid carcinomas of early stages (two of T1 and two of T2, none with nodal or distal metastasis).

We will refine our model in the future by evaluating the impact of the independence assumption. We also plan to build a web site so that our classifier can be easily accessed by other radiologists.

References

1. Mortensen JD, Woolner LB, Bennett WA. Gross and microscopic findings in clinically normal thyroid glands. J Clin Endocrinol Metab 1955; 15:1270-1280.
2. Wiest PW, Hartshorne MF, Inskip PD, et al. Thyroid palpation versus high-resolution thyroid ultrasonography in the detection of nodules. J Ultrasound Med 1998; 17:487-496.
3. Cooper DS, Doherty GM, Haugen BR, et al. Management guidelines for patients with thyroid nodules and differentiated thyroid cancer. Thyroid 2006; 16:109-142.
4. Papini E, Guglielmi R, Bianchini A, et al. Risk of malignancy in nonpalpable thyroid nodules: predictive value of ultrasound and color-Doppler features. J Clin Endocrinol Metab 2002; 87:1941-1946.
5. Jemal A, Siegel R, Ward E, et al. Cancer statistics, 2007. CA Cancer J Clin 2007; 57:43-66.
6. Davies L, Welch HG. Increasing incidence of thyroid cancer in the United States, 1973-2002. JAMA 2006; 295:2164-2167.
7. Frates MC, Benson CB, Charboneau JW, et al. Management of thyroid nodules detected at US: Society of Radiologists in Ultrasound consensus conference statement. Radiology 2005; 237:794-800.
8. Reading CC, Charboneau JW, Hay ID, Sebo TJ. Sonography of thyroid nodules: a "classic pattern" diagnostic approach. Ultrasound Q 2005; 21:157-165.
9. Burnside E, Rubin D, Shachter R. A Bayesian network for mammography. Proc AMIA Symp 2000:106-110.
10. Kline JA, Novobilski AJ, Kabrhel C, Richman PB, Courtney DM. Derivation and validation of a Bayesian network to predict pretest probability of venous thromboembolism. Ann Emerg Med 2005; 45:282-290.
11. Ahuja A, Chick W, King W, Metreweli C. Clinical significance of the comet-tail artifact in thyroid ultrasound. J Clin Ultrasound 1996; 24:129-133.
12. Zhang H. The Optimality of Naive Bayes. FLAIRS Conference 2004.
13. Burnside ES, Ochsner JE, Fowler KJ, et al. Use of microcalcification descriptors in BI-RADS 4th edition to stratify risk of malignancy. Radiology 2007; 242:388-395.

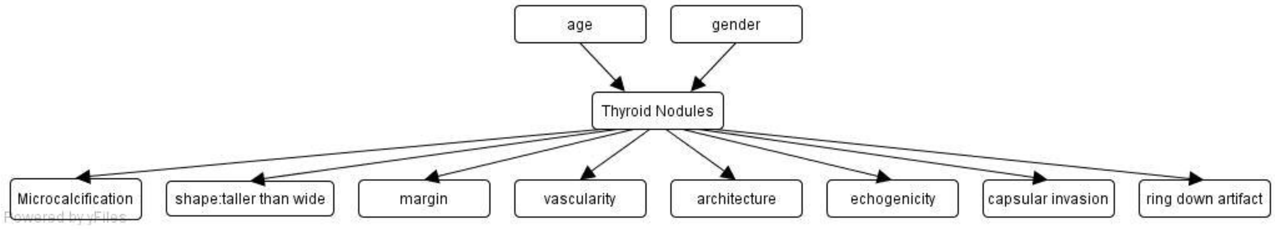

Figure 1. Our Bayesian classifier for thyroid nodules.

Table 1. Probability of malignant nodule given age and gender.

	Age <50, Male	age<50, Female	Age>=50, Male	Age>=50, Female
P(malignant nodule)	0.2	0.12	0.3	0.2

Table 2. Definitions and parameters for the sonographic features in our Bayesian classifier.

Sonographic Feature	Definition	States	P(feature| malignant)	P(feature| benign)
Microcalcification	Punctuate echogenic foci without acoustic shadowing or associated comet-tail artifact	Present	0.5	0.1
		Absent	0.5	0.9
Shape: Taller than wide	AP dimension > transverse dimension	Present	0.25	0.05
		Absent	0.75	0.95
Margin	Margin of the thyroid nodule	Smooth	0.15	0.5
		Irregular	0.35	0.3
		Ill_defined	0.50	0.2
Capsular invasion	Extension of a nodule beyond the thyroid capsule	Present	0.2	0.0005
		Absent	0.8	0.9995
Architecture	Composition of the thyroid nodule	Solid	0.82	0.5
		Almost_solid (<25% cystic)	0.10	0.17
		Mixed (25-75% cystic)	0.05	0.17
		Cystic (>75% cystic)	0.03	0.16
Echogenicity	The echogenicity of the thyroid nodule relative to surrounding thyroid parenchyma	Hypoechoic	0.85	0.5
		Isoechoic	0.1	0.25
		Hyperechoic	0.05	0.25
Ring down artifact	Punctuate echogenic foci associated with comet-tail artifact	Present	0.00005	0.08
		Absent	0.99995	0.92
Vascularity	Whether flow is seen on color Doppler interrogation	Intrinsic	0.55	0.4
		Perinodular	0.44	0.4
		avascular	0.01	0.2

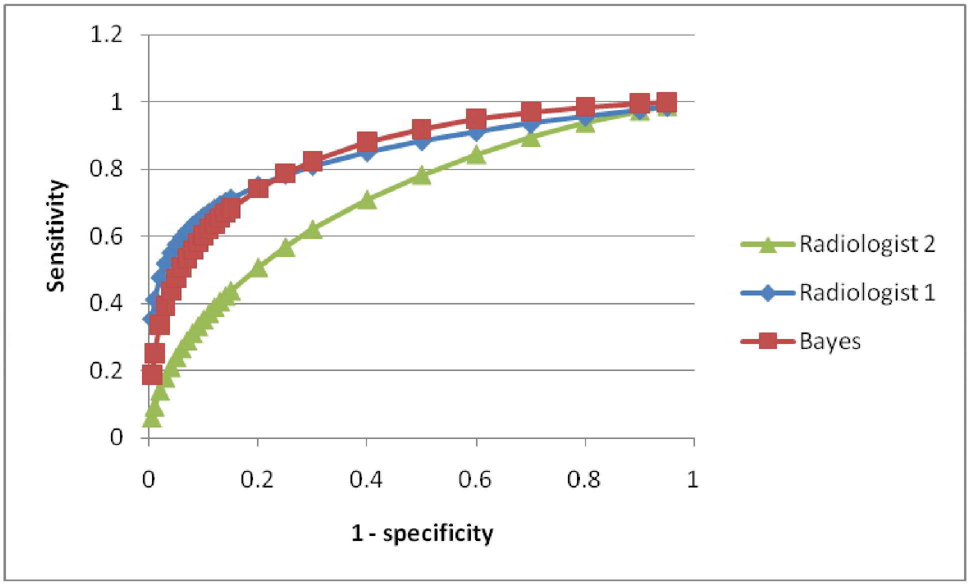

Figure 2. ROC curves for the Bayesian classifier and two radiologists.