HHS Public Access

Author manuscript

Mayo Clin Proc Digit Health. Author manuscript; available in PMC 2024 March 18.

Published in final edited form as:

Mayo Clin Proc Digit Health. 2024 March ; 2(1): 67–74. doi:10.1016/j.mcpdig.2024.01.001.

Thyroid Ultrasound Appropriateness Identification Through Natural Language Processing of Electronic Health Records

Cristian Soto Jacome, MD,
Division of Endocrinology, Diabetes, Metabolism, and Nutrition, Department of Medicine, Knowledge and Evaluation Research Unit, Mayo Clinic, Rochester, MN

Danny Segura Torres, MD,
Division of Endocrinology, Diabetes, Metabolism, and Nutrition, Department of Medicine, Knowledge and Evaluation Research Unit, Mayo Clinic, Rochester, MN

Jungwei W. Fan, PhD,
Department of Artificial, Intelligence and Informatics, Mayo Clinic, Rochester, MN

Ricardo Loor-Torres, MD,
Division of Endocrinology, Diabetes, Metabolism, and Nutrition, Department of Medicine, Knowledge and Evaluation Research Unit, Mayo Clinic, Rochester, MN

Mayra Duran, MD,
Division of Endocrinology, Diabetes, Metabolism, and Nutrition, Department of Medicine, Knowledge and Evaluation Research Unit, Mayo Clinic, Rochester, MN

Misk Al Zahidy, MS,
Division of Endocrinology, Diabetes, Metabolism, and Nutrition, Department of Medicine, Knowledge and Evaluation Research Unit, Mayo Clinic, Rochester, MN

Esteban Cabezas, MD,
Division of Endocrinology, Diabetes, Metabolism, and Nutrition, Department of Medicine, Knowledge and Evaluation Research Unit, Mayo Clinic, Rochester, MN

Mariana Borras-Osorio, MD,
Division of Endocrinology, Diabetes, Metabolism, and Nutrition, Department of Medicine, Knowledge and Evaluation Research Unit, Mayo Clinic, Rochester, MN

David Toro-Tobon, MD,
Division of Endocrinology, Diabetes, Metabolism, and Nutrition, Mayo Clinic, Rochester, MN

Yuqi Wu, PhD,
Department of Artificial, Intelligence and Informatics, Mayo Clinic, Rochester, MN

Correspondence: Address to Juan P. Brito MD, MS, Division of Endocrinology, Diabetes, Nutrition and Metabolism Mayo Clinic, 200 First Street SW, Rochester, MN 55902 (Brito.juan@mayo.edu).

POTENTIAL COMPETING INTERESTS

All authors report no competing interest

SUPPLEMENTAL ONLINE MATERIAL

Supplemental material can be found online at https://www.mcpdigitalhealth.org/. Supplemental material attached to journal articles has not been edited, and the authors take responsibility for the accuracy of all data.

Yonghui Wu, PhD,
Department of Health Outcomes & Biomedical Informatics, University of Florida, Gainesville, FL

Naykky Singh Ospina, MD, MS,
Division of Endocrinology, Department of Medicine, University of Florida, Gainesville, FL

Juan P. Brito, MD, MS
Division of Endocrinology, Diabetes, Metabolism, and Nutrition, Department of Medicine, Knowledge and Evaluation Research Unit, Mayo Clinic, Rochester, MN

Division of Endocrinology, Diabetes, Metabolism, and Nutrition, Mayo Clinic, Rochester, MN

Abstract

Objective: To address thyroid cancer overdiagnosis, we aim to develop a natural language processing (NLP) algorithm to determine the appropriateness of thyroid ultrasounds (TUS).

Patients and Methods: Between 2017 and 2021, we identified 18,000 TUS patients at Mayo Clinic and selected 628 for chart review to create a ground truth dataset based on consensus. We developed a rule-based NLP pipeline to identify TUS as appropriate TUS (aTUS) or inappropriate TUS (iTUS) using patients' clinical notes and additional meta information. In addition, we designed an abbreviated NLP pipeline (aNLP) solely focusing on labels from TUS order requisitions to facilitate deployment at other health care systems. Our dataset was split into a training set of 468 (75%) and a test set of 160 (25%), using the former for rule development and the latter for performance evaluation.

Results: There were 449 (95.9%) patients identified as aTUS and 19 (4.06%) as iTUS in the training set; there are 155 (96.88%) patients identified as aTUS and 5 (3.12%) were iTUS in the test set. In the training set, the pipeline achieved a sensitivity of 0.99, specificity of 0.95, and positive predictive value of 1.0 for detecting aTUS. The testing cohort revealed a sensitivity of 0.96, specificity of 0.80, and positive predictive value of 0.99. Similar performance metrics were observed in the aNLP pipeline.

Conclusion: The NLP models can accurately identify the appropriateness of a thyroid ultrasound from clinical documentation and order requisition information, a critical initial step toward evaluating the drivers and outcomes of TUS use and subsequent thyroid cancer overdiagnosis.

Thyroid cancer has emerged as a growing public health concern in the United States, with a significant increase in disease rates observed over the past 3 decades.[1,2] Most cases are attributed to small papillary thyroid cancers measuring 1.5 cm or less that belong to a large reservoir of thyroid cancer in the population.[3] Despite the high incidence, mortality rates associated with thyroid cancer have remained relatively low.[4,5] This discrepancy has raised concerns about overdiagnosis, which occurs when patients are diagnosed and subsequently treated for cancers that pose no real threat to their health.[6,7] Overdiagnosis can result in unnecessary medical procedures, emotional distress, diminished quality of life, disruptions in employment, and lifelong hormone replacement therapy.[6,8] In addition, it imposes a substantial financial burden on the health care system, with estimated annual costs exceeding $1.5 billion in the United States and projected to reach $3.5 billion by 2030.[6]

Thyroid ultrasound (TUS) has been identified as a key driver of thyroid cancer overdiagnosis. The use of TUS has surged in recent years.[4,9] Research conducted among Medicare patients revealed an increase from ~200 ultrasounds per 100,000 people in 2002 to 1500 ultrasounds per 100,000 people in 2012, indicating a growth rate of 20% per year.[3] Furthermore, single-center studies suggest that 80%-90% of these ultrasounds were ordered appropriately, adhering to guideline recommendations, such as investigating symptoms related to thyroid nodules or evaluating incidental findings in other imaging modalities. The remaining 20%-10% were ordered for inappropriate reasons: screening for thyroid cancer, work-ups for hypothyroidism, patient requests, and other factors.[10,11] The factors, however, influencing the appropriate versus inappropriate ordering of TUS remain unknown. Furthermore, research on the frequency of appropriate and inappropriate TUS ordering in a larger and more diverse population has been challenging, as it often requires a laborious and costly review of medical records, limiting the size and generalizability of the analyzed sample.[12]

Natural Language Processing (NLP) is an interdisciplinary research field at the intersection of artificial intelligence and linguistics.[13] In the medical domain, NLP is the key technology to use narrative clinical text for clinical research.[14–16] For example, NLP can be applied to extract important patient information from unstructured text into a normalized and structured format suitable for analysis.[14] Rule-based NLP systems use prespecified, human-created rules to analyze and match specific patterns in the text, which is particularly useful for extracting medical concepts from clinical notes when the target terms are well-defined with enumerable patterns.[17] In addition, rule-based NLP solutions are computational-friendly and can be used as a postprocessing to fix systematic errors in machine learning-based NLP systems.[18] In this study, our goal is to create an NLP-driven algorithm for assessing the appropriateness of TUS, while also analyzing the rationale behind labeling them as appropriate (aTUS) or inappropriate (iTUS). This serves as a foundational step in developing validated, deployable algorithms capable of examining TUS appropriateness at the population level using large-scale electronic health records (EHRs). Ultimately, these algorithms will aid in further research to understand the factors driving TUS use, thereby helping to effectively tackle the issue of thyroid cancer overdiagnosis.

METHODS

Data Source

We included adult patients (aged >18 years) who underwent initial TUS between January 1, 2017, and December 31, 2021, at (1) Mayo Clinic, Rochester, MN, (2) Mayo Clinic, Jacksonville, FL, (3) Mayo Clinic, Scottsdale, AZ, and (4) Midwest Mayo Clinic Healthcare System. Only patients who were granted Minnesota research authorization were included. To ensure that we captured only the patients' first TUS during the specified timeframe, we selected the initial (incident) TUS order and excluded cases with a TUS ordered before January 1, 2017. The query generated a comprehensive collection of over 18,000 patient records with at least 1 TUS; we also retrieved other pertinent variables, such as demographics, order attributes, and clinical notes. To annotate a ground truth dataset based on consensus, we randomly selected a cohort of 628 patients from this pool who had active

medical records after our institution migrated to its current EHR (EPIC system). After annotation, we divided the 628 patients into a training set of 468 (75%) patients and a test set of 160 (25%) patients. The primary goal of the project was to develop a ruled-based NLP pipeline using clinical notes. The secondary outcomes of this project were to create an abbreviated NLP pipeline using only information available on ultrasound order requisitions and to extract the reason for TUS. The study was approved by Mayo Clinic IRB # 21–002627.

Development of the NLP Pipeline

Ground Truth Dataset Based on Consensus.—We conducted a chart review of EHRs to determine the appropriateness of TUS, which serves as the ground truth dataset based on consensus to develop and evaluate NLP algorithms. Figure 1 shows the workflow for chart review. Specifically, we classified a TUS as appropriate (aTUS) if there was at least 1 predefined criterion met (Supplemental Table 1, available online at https://www.mcpdigitalhealth.org/), otherwise, TUS were classified as inappropriate (iTUS). Following the flowchart and the aTUS reference lexicon in Supplemental Table 1, 2 physicians underwent 3 rounds of iterative training sessions to achieve an inter-annotator agreement of over 80%. Subsequently, each physician independently reviewed the complete cohort of patients, classifying them as either aTUS or iTUS.

Rule-Based Automatic Classification.—An automatic classification pipeline was implemented by reproducing the exact flowchart (Figure 1) humans used to differentiate aTUS versus iTUS. The text-matching rules were developed by leveraging the natural language process tool kit, an NLP infrastructure built at Mayo Clinic. The natural language process tool kit allows iterative tuning of regular expression patterns for our use case (see Supplemental Table 2, available online at https://www.mcpdigitalhealth.org/) and handles the underlying NLP core tasks, such as sentence chunking, section identification, and negation detection. On top of the NLP component, we implemented a Python program that executes the entire decision flowchart by integrating all pertinent input variables such as encounter type, encounter date, reason for exam snippet, and the NLP extractions. The tuning took place when incorrect classification resulted from too loose or too stringent patterns (or section constraints), which then required modification of the rules to maximize the number of correctly classified cases on the training set. Independently, the test set then served to validate the generalizability of the rules.

Abbreviated NLP Pipeline

For efficiency and potentially better portability, we also explored the implementation of an abbreviated version of the full NLP pipeline (aNLP). This abbreviated pipeline relies only on information from ultrasound order requisitions, including textual descriptions of the reason for the examination and the diagnosis linked to the TUS order. For this aNLP, we adapted the system by incorporating a few modifications such as removing the thyroid in some of the NLP patterns because the anatomy was already implied by the TUS order. The adjusted flowchart for the aNLP is available as Supplemental Figure 1, (available online at https://www.mcpdigitalhealth.org). To ensure the generalizability of aNLP's performance, minor tuning was initially conducted on a smaller set of 160 patients. Subsequently, we

evaluated the algorithm on the larger set of 468 patients, which had not influenced the aNLP modifications. In other words, for aNLP development, the 160-patient set served as the training group, whereas the 468-patient set acted as the test group.

Performance Evaluation

To gauge the pipelines performance of classifying aTUS versus iTUS, we computed the 2 × 2 confusion matrix. The following accuracy metrics were calculated against the ground truth dataset based on consensus annotations: Sensitivity ([SN], true positive rate), Specificity ([SP], true negative rate), Positive Predictive Value ([PPV], the likelihood that a positive test result indicates the actual presence of the condition), Accuracy ([ACC], the overall correctness determined by the sum of true positives and true negatives divided by the total), and F1-measure (F1), which is the harmonic mean of PPV and SN.

Reasons for aTUS and iTUS

During ground truth dataset based on consensus development, reviewers extracted text excerpts to justify the reason for TUS. Similarly, our NLP pipeline likewise generated excerpts from clinical notes as evidence for TUS orders. We grouped this evidence into broader categories, as described in Table 1, and assessed agreement on true positives (cases identified as appropriate by both NLP and team consensus) and true negatives (cases identified as inappropriate by both). The κ agreement between reviewers and NLP experts was found to be 0.83.

RESULTS

System Performance

Per annotation, domain experts identified 449 (95.94%) aTUS and 19 (4.06%) iTUS from the training set of 468 patients and identified 155 (96.88%) aTUS and 5 (3.12%) iTUS from the test set of 160 patients. In the training cohort, the NLP pipeline had an SN of 0.99, SP of 0.95, PPV of 1.0, ACC of 0.99, and F1 of 0.99 for detecting aTUS. In the testing cohort, there was a SN of 0.96, SP of 0.80, PPV of 0.99, ACC of 0.96, and F1 of 0.97. For the aNLP, it achieved an SN of 0.97, SP of 0.89, PPV of 0.99, ACC of 0.97, and F1 of 0.98 upon the tuning, and an SN of 0.97, SP of 0.89, PPV of 1.0, ACC of 0.97, and F1 of 0.98 on the set for validation (Table 2).

Reasons for aTUS

The NLP pipeline detected multiple reasons for appropriateness within each case. Specifically, we found 638 reasons among 446 true positive cases in the training set and 200 reasons among 149 true positives in the testing set. Altogether, this amounted to 838 appropriate reasons, averaging 1.4 reasons per case. The most prevalent reason was evaluating known thyroid nodular disease, cited 416 times (50%). Often, this arose from incidental nodule discoveries in unrelated imaging studies. The second most common reason cited 180 times (21%), was identifying a nodule during a routine physical exam. The distribution of these appropriate reasons is illustrated in Figure 2.

Reasons for iTUS

We identified 22 inappropriate reasons within both the training and testing set cohorts. The most common reason detected was thyroid dysfunction (12, 54.54%), followed by screening for thyroid cancer or thyroid nodule (8, 36.36%), and no specified reason (2, 9%).

DISCUSSION

We have successfully developed rule-based algorithms that use unstructured data from EHRs to determine the appropriateness of TUS. The pipeline reported a high level of accuracy, with a PPV of 0.99 in the testing cohort.

Implication for Research

The development of this algorithm represents an important milestone in thyroid cancer overdiagnosis research, providing a valuable tool to explore the frequency and factors influencing appropriate versus inappropriate TUS usage through large-scale EHRs. Through analyzing EHR data, researchers may explore the appropriateness of TUSs on a larger scale. Researchers can identify the drivers behind TUS use by delving into patient characteristics, health care provider practices, and clinical scenarios. This knowledge is essential for developing and evaluating targeted interventions and policy changes to promote appropriate use and reduce unnecessary ultrasounds.

In addition, the model holds the potential to facilitate comparative analysis across different health care systems, shedding light on potential disparities, best practices, and areas in need of improvement. Moreover, the algorithm's output enables researchers to map the reasons for ordering TUSs. Although we have highlighted several appropriate reasons for ordering ultrasounds, we have also observed that many of these appropriate reasons could be targets for future interventions.[5] For instance, we have noticed that many patients undergo appropriate TUS after a nodule is discovered during a routine physical examination. However, most of these patients are asymptomatic, and the routine physical examination in asymptomatic individuals remains controversial. The recent US Task Force recommendations advise against routine thyroid physical examinations in asymptomatic individuals.[19] Thus, mapping the reasons for ultrasounds provides researchers with additional information to address the primary drivers of overdiagnosis in cases that lead to unnecessary diagnoses of thyroid cancer.

Next Steps and Limitations

Before implementing this algorithm in other health care systems, validating its generalizability is crucial, particularly in settings with more frequent inappropriate TUSs. This could improve the model's ability to differentiate between appropriate and inappropriate TUSs. In this project, we opted to report the model's performance for detecting aTUS over iTUS because of concerns stemming from the limited iTUS sample size.

Although our study leverages a substantial dataset from tertiary care centers and the Midwest Mayo Clinic Health Care System, it is important to acknowledge that the sample

may not fully represent the broader demographic characteristics and clinical practices of rural and suburban clinics. In addition, while acknowledging the data imbalance with a higher proportion of appropriate TUSs, it is important to consider that this distribution may also reflect the true prevalence of appropriate versus inappropriate ultrasound ordering in clinical practice. These limitations suggest caution in generalizing our findings to these settings and encourage future research with more balanced datasets. On contrary, it is also crucial to recognize that the efficacy of our algorithm in enhancing neck ultrasound order appropriateness hinges on the precision and uniformity of documentation practices within health care settings. For this reason, subsequent research, including the validation of this NLP approach in different health care systems, will enable a thorough reevaluation of the pipeline's performance in light of varying documentation practices.

However, the validation process has known portability challenges, primarily due to the incompatible EHR syntax or semantics (eg, section header naming conventions) across different health care systems.[20] The rule-based approach used in developing this algorithm closely mimics the human decision-making process for determining ultrasound appropriateness.[18] Our approach involved recognizing that certain sections within medical notes provide higher quality and more informative data than others. For example, during the consensus annotation process, we found that the impression report and plan sections of the ordering provider's note contained clear descriptions of the reasons for ordering a TUS. Thus, our workflow focused on exploring these subsections to assess appropriateness. It is possible that other health care systems may not have similar subsection mapping capabilities within their clinical notes, potentially affecting the pipeline's performance. To overcome potential portability issues, we developed an abbreviated version of the NLP pipeline that only requires information available in the TUS requisition form. We believe the latter is more likely to be consistent across different health care systems (Supplemental Table 3, available online at https://www.mcpdigitalhealth.org/). Moreover, the use of advanced language models, such as BERT, for encoding clinical notes and ultrasound order requisitions presents a promising alternative to enhance the consistency and interpretability of data, potentially improving the portability and applicability of our NLP algorithms across different health care platforms.[21] As such, future implementation of validated algorithms will be individualized based on the data type available. Furthermore, it is important to note that Mayo Clinic transitioned to the Epic EHR system in 2017. Considering the relevance of clinical note architecture in accurately deploying our rule-based approach, we decided to develop and test our algorithm solely on cases after the implementation of this system. This choicerestrictsthefuturedeploymentofthecurrentrule-basedalgorithmtoaspecifictimeinterval and may limit our analysis of ultrasounds conducted before 2017.

CONCLUSION

The development of our rule-based NLP pipeline for determining the appropriateness of TUS using EHR data is a promising first milestone in understanding the appropriateness of thyroid ultrasound. However, it is crucial to consider the validation process, challenges related to algorithm portability, and the limitations of our algorithm's scope due to the transition to Epic for future implementation and analysis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding:

Naykky Singh Ospina is supported by the National Cancer Institute of the National Institutes of Health under Award Number K08CA248972. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Juan P. Brito, Cristian Soto Jacome and Jungwei W. Fan are supported by the National Cancer Institute of the National Institutes of Health under Award Number R37CA272473. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Yonghui Wu is supported by Patient-Centered Outcomes Research Institute® (PCORI) under Award Number ME-2018C3–14754 and National Institute on Aging under Award Number R56AG069880. The content is solely the responsibility of the authors and does not necessarily represent the official views of the PCORI and National Institutes of Health.

Abbreviations and Acronyms:

ACC	accuracy
aTUS	appropriate thyroid ultrasound
aNLP	abbreviated version of the full NLP pipeline
EHRs	electronic health records
NLP	natural language processing
NLPTK	natural language process tool kit
PPV	positive predictive value
SN	sensitivity
SP	specificity
sNLP	NLP pipeline
TUS	thyroid ultrasound

REFERENCES

1. Lim H, Devesa SS, Sosa JA, Check D, Kitahara CM. Trends in thyroid cancer incidence and mortality in the United States, 1974–2013. JAMA. 2017;317(13):1338–1348. [PubMed: 28362912]
2. Megwalu UC, Moon PK. Thyroid Cancer Incidence and Mortality Trends in the United States: 2000–2018. Thyroid. 2022;32(5):560–570. [PubMed: 35132899]
3. Haymart MR, Banerjee M, Reyes-Gastelum D, Caoili E, Norton EC. Thyroid ultrasound and the increase in diagnosis of low-risk thyroid cancer. J Clin Endocrinol Metab. 2019;104(3):785–792. [PubMed: 30329071]
4. Udelsman R, Zhang Y. The epidemic of thyroid cancer in the United States: the role of endocrinologists and ultrasounds. Thyroid. 2014;24(3):472–479. [PubMed: 23937391]
5. Davies L, Welch HG. Current thyroid cancer trends in the United States. JAMA Otolaryngol Head Neck Surg. 2014;140(4):317–322. [PubMed: 24557566]
6. Nguyen BM, Lin KW, Mishori R. Public health implications of overscreening for carotid artery stenosis, prediabetes, and thyroid cancer. Public Health Rev. 2018;39(1):18. [PubMed: 29988604]

7. Jegerlehner S, Bulliard JL, Aujesky D, et al. Overdiagnosis and overtreatment of thyroid cancer: A population-based temporal trend study. PLoS One. 2017;12(6):e0179387. [PubMed: 28614405]
8. Moynihan R, Doust J, Henry D. Preventing overdiagnosis: how to stop harming the healthy. BMJ. 2012;344(7859):e3502. [PubMed: 22645185]
9. Lincango-Naranjo E, Solis-Pazmino P, El Kawkgi O, et al. Triggers of thyroid cancer diagnosis: a systematic review and meta-analysis. Endocrine. 2021;72(3):644–659. [PubMed: 33512656]
10. Davenport C, Alderson J, Yu IG, et al. A review of the propriety of thyroid ultrasound referrals and their follow-up burden. Endocrine. 2019;65(3):595–600. [PubMed: 30955175]
11. Landry BA, Barnes D, Keough V, et al. Do family physicians request ultrasound scans appropriately? Can Fam Physician. 2011;57(8):e299–e304. [PubMed: 21841093]
12. Edwards MK, Iñiguez-Ariza NM, Singh Ospina N, Lincango-Naranjo E, Maraka S, Brito JP. Inappropriate use of thyroid ultrasound: a systematic review and meta-analysis. Endocrine. 2021;74(2):263–269. [PubMed: 34379311]
13. Juhn Y, Liu H. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. J Allergy Clin Immunol. 2020;145(2):463–469. [PubMed: 31883846]
14. Sezgin E, Hussain SA, Rust S, Huang Y. Extracting medical information from free-text and unstructured patient-generated health data using natural language processing methods: feasibility study with real-world data. JMIR Form Res. 2023;7:e43014. [PubMed: 36881467]
15. Miller DD, Brown EW. Artificial intelligence in medical practice: the question to the answer? Am J Med. 2018;131(2):129–133. [PubMed: 29126825]
16. Toro-Tobon D, Loor-Torres R, Duran M, et al. Artificial intelligence in thyroidology: a narrative review of the current applications, associated challenges, and future directions. Thyroid. 2023;33(8):903–917. [PubMed: 37279303]
17. Kang N, Singh B, Afzal Z, van Mulligen EM, Kors JA. Using rule-based natural language processing to improve disease normalization in biomedical text. J Am Med Inform Assoc. 2013;20(5):876–881. [PubMed: 23043124]
18. Gunter D, Puac-Polanco P, Miguel O, et al. Rule-based natural language processing for automation of stroke data extraction: a validation study. Neuroradiology. 2022;64(12):2357–2362. [PubMed: 35913525]
19. United States Preventive Services Task Force, Bibbins-Domingo K, Grossman DC, et al. Screening for thyroid cancer: US Preventive Services Task Force recommendation statement. JAMA. 2017;317(18):1882–1887. [PubMed: 28492905]
20. Wyles CC, Fu S, Odum SL, et al. External validation of natural language processing algorithms to extract common data elements in THA operative notes. J Arthroplasty. 2023;38(10):2081–2084. [PubMed: 36280160]
21. Lian R, Hsiao V, Hwang J, et al. Predicting health-related quality of life change using natural language processing in thyroid cancer. Intell Based Med. 2023;7.

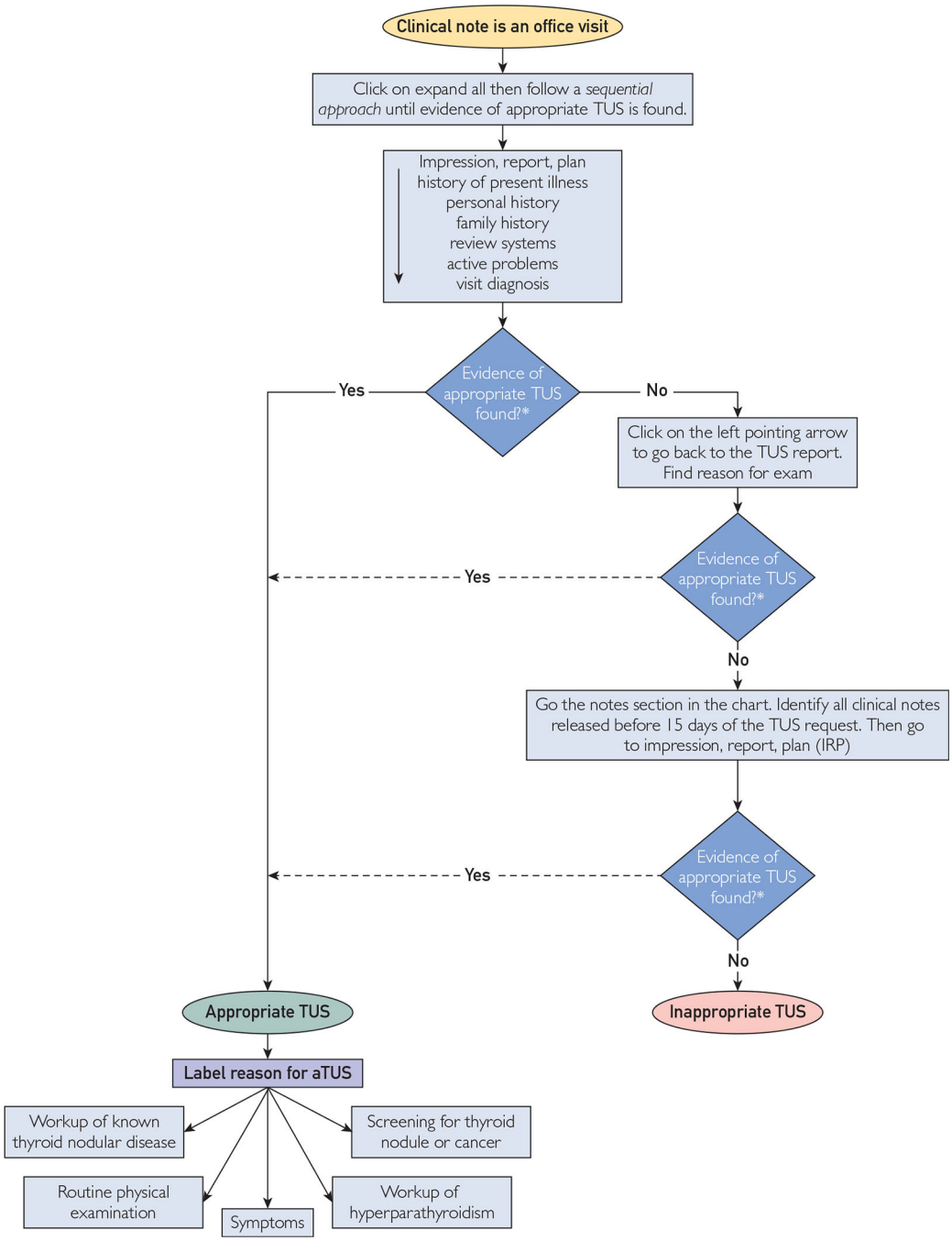

FIGURE 1.

Decision flow chart to classify thyroid ultrasound appropriateness by human and NLP abstraction. NLP, natural language process.

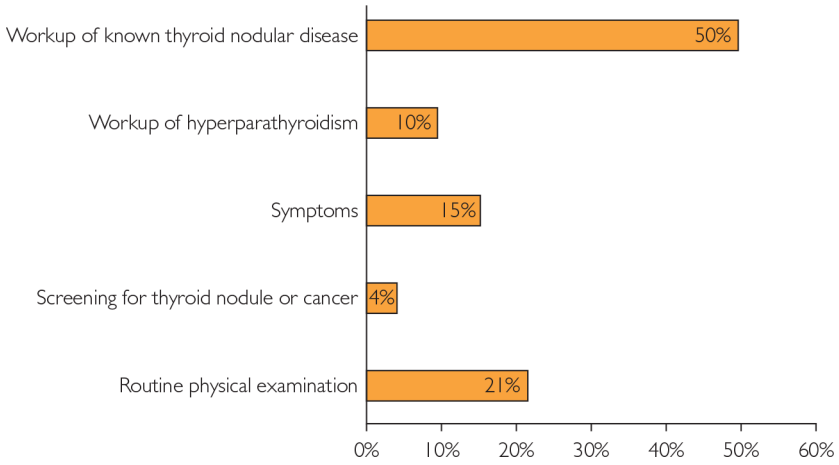

FIGURE 2.
Reasons for aTUS. aTUS, appropriate thyroid ultrasound.

TABLE 1.

Classification Framework for Appropriate Thyroid Ultrasound Use

Categories	Description
Symptoms	TUS was requested because of the presence of symptoms associated with thyroid nodules or masses. Examples of such symptoms could include dysphonia.
Work-up of known thyroid nodular disease	TUS was ordered as a follow-up investigation after thyroid nodes were incidentally detected during a previous imaging test conducted for reasons unrelated to thyroid disorders or symptoms. An example would be when a computed tomography scan was performed for posterior neck pain, and incidental thyroid or neck abnormalities were discovered. This category also includes when TUS was ordered for the work-up of previously known thyroid nodular disease.
Screening for thyroid nodule or cancer	TUS was used in asymptomatic patients, specifically aiming to detect thyroid nodules or thyroid cancer in patients with personal history of hereditary syndromes associated with thyroid cancer or a family history of thyroid cancer.
Routine physical examination	TUS was ordered as part of the work-up of a thyroid finding in physical examination (eg, thyroid nodule or thyroid enlargement)
Work-up of hyperparathyroidism	TUS was ordered as part of the preoperative work-up of hyperparathyroidism

Abbreviations: TUS, thyroid ultrasound.

TABLE 2.

Standard and Abbreviated NLP Models' Performance When Compared With Ground Truth Data Based on Consensus

Machine (sNLP) Ground truth	**aTUS**	**iTUS**	**Total**
Training set (n=468)			
aTUS	446	3	449
iTUS	1	18	19
Total	447	21	468
Machine (sNLP)			
Ground truth			
Testing set (n= 160)			
aTUS	149	6	155
iTUS	1	4	5
Total	150	10	160
Machine (aNLP)			
Ground truth			
Training set (n=160)			
aTUS	151	4	155
iTUS	1	4	5
Total	152	8	160
Machine (aNLP)			
Ground truth			
Testing set (n=468)			
aTUS	435	14	449
iTUS	2	17	19
Total	437	31	468

Abbreviations: sNLP, NLP pipeline; aNLP, abbreviated NLP pipeline.