Aaron C. Miller*, Alan T. Arakkal, Scott H. Koeneman, Joseph E. Cavanaugh and Philip M. Polgreen

A clinically-guided unsupervised clustering approach to recommend symptoms of disease associated with diagnostic opportunities

https://doi.org/10.1515/dx-2022-0044
Received May 9, 2022; accepted August 26, 2022;
published online September 21, 2022

Abstract

Objectives: A first step in studying diagnostic delays is to select the signs, symptoms and alternative diseases that represent missed diagnostic opportunities. Because this step is labor intensive requiring exhaustive literature reviews, we developed machine learning approaches to mine administrative data sources and recommend conditions for consideration. We propose a methodological approach to find diagnostic codes that exhibit known patterns of diagnostic delays and apply this to the diseases of tuberculosis and appendicitis.

Methods: We used the IBM MarketScan Research Databases, and consider the initial symptoms of cough before tuberculosis and abdominal pain before appendicitis. We analyze diagnosis codes during healthcare visits before the index diagnosis, and use k-means clustering to recommend conditions that exhibit similar trends to the initial symptoms provided. We evaluate the clinical plausibility of the recommended conditions and the corresponding number of possible diagnostic delays based on these diseases.

Results: For both diseases of interest, the clustering approach suggested a large number of clinically-plausible conditions to consider (e.g., fever, hemoptysis, and pneumonia before tuberculosis). The recommended conditions had a high degree of precision in terms of clinical plausibility: >70% for tuberculosis and >90% for appendicitis. Including these additional clinically-plausible conditions resulted in more than twice the number of possible diagnostic delays identified.

Conclusions: Our approach can mine administrative datasets to detect patterns of diagnostic delay and help investigators avoid under-identifying potential missed diagnostic opportunities. In addition, the methods we describe can be used to discover less-common presentations of diseases that are frequently misdiagnosed.

Keywords: administrative data; diagnostic delay; machine learning.

Introduction

Diagnostic errors are an important cause of avoidable harms and increased healthcare costs [1, 2]. Yet, diagnostic errors represent a challenging area of research that relies on a range of methods (e.g., chart reviews or surveys) and measures (e.g., mortality, costs, or malpractice-claims) [3–5]. Delays in diagnosing a disease represent an important type of diagnostic error [6]. To study diagnostic delays, a growing body of research has utilized large administrative datasets, such as insurance claims or hospital discharge records generated for institutional or billing purposes. Such data have advantages for studying the diagnostic process. First, they tend to be less costly and easier to analyze compared to surveys or chart reviews that require additional collection and processing. Second, they often represent heterogeneous patient populations covering wide geographic areas. Third, these data often contain longitudinal information spanning multiple institutions and settings. Thus, patients who receive fragmented care can be studied across disconnected health systems.

Studying diagnostic delays, whether through chart review or using administrative records, requires researchers to define the criteria for identifying delays. Specifically, the types of antecedent healthcare visits, defined as visits that precede the index disease diagnosis (i.e. the initial diagnosis of the underlying disease) that represent a missed diagnostic opportunity. For example, visits where a patient presented with fever or cough in the weeks prior to a tuberculosis

***Corresponding author: Aaron C. Miller**, Department of Internal Medicine, Roy J. and Lucille A. Carver College of Medicine, University of Iowa, Iowa City, IA, 52242, USA, Phone: (319) 335-3053, E-mail: aaron-miller@uiowa.edu
Alan T. Arakkal, Scott H. Koeneman and Joseph E. Cavanaugh, Department of Biostatistics, College of Public Health, University of Iowa, Iowa City, IA, USA
Philip M. Polgreen, Department of Internal Medicine, Carver College of Medicine, University of Iowa, Iowa City, IA, USA

diagnosis may be considered as *antecedent conditions* signaling a potential missed diagnostic opportunity. Numerous studies have used this approach to study diagnostic delays with different administrative data sources [7–24]. However, a challenge for defining antecedent conditions associated with a disease of interest is identifying which symptoms to evaluate. Including too few symptoms may underestimate the number of missed diagnostic opportunities. Failing to include rare symptoms may also lead to systematic biases in research design that excludes patient populations with atypical disease presentations. Such patients might also be the most susceptible to diagnostic delays [25].

In addition to considering symptoms, patients may be mistakenly diagnosed with diseases that share similar symptoms to the index disease (e.g., pneumonia instead of tuberculosis) [9, 10]; such visits also represent missed diagnostic opportunities and should be considered in the evaluation process. Furthermore, some of the most important missed opportunities to consider may occur when patients present with less common symptoms of disease. Alternatively, missed opportunities may occur because of atypical disease manifestations. If such visits are missed by clinicians, researchers may also fail to consider such antecedent visits as a potential missed opportunity. Because this selection process is intensive (e.g., exhaustive literature reviews) and may miss atypical manifestations, automated approaches may help to identify potential missed opportunities.

The goal of this paper is to demonstrate an exploratory machine-learning approach that can be used to aid the process of selecting the criteria (symptoms and symptomatically similar diseases) used to identify potential missed opportunities in subsequent study designs. Machine learning techniques are used in many applications as "recommender systems" to aid human selection; examples include online shopping and streaming services [26]. In the diagnostic delay context, pattern recognition techniques may detect visits appearing to match common trends of known diagnostic opportunities and suggest similar antecedent conditions for consideration. In this paper we present a methodological framework for using administrative data and unsupervised clustering approaches (i.e., machine learning techniques where patterns are uncovered from data without a defined or labelled outcome) to identify sets of diagnosis codes that may capture potential diagnostic opportunities. We apply our approach to the diseases of tuberculosis and appendicitis to demonstrate how antecedent signs and symptoms of these diseases can be discovered. We then evaluate the antecedent conditions recommended in terms of their clinical plausibility.

Methods

Temporal visit trends prior to diagnosis

A common observation in studies of diagnostic opportunities is an increasing trend of visits with symptoms of a given disease (e.g., chest pain before AMI) prior to the index diagnosis/visit. This increasing trend has been found for multiple diseases including stroke [8, 27], AMI [7, 22, 27], tuberculosis [9, 10], endemic fungal infections [16, 18, 19], and many others [11, 17, 20, 21]. Figure 1 depicts an example of this trend for tuberculosis, where visits for cough and fever increase prior to an index diagnosis. When nearing the index date of a given disease diagnosis, one expects to observe more healthcare encounters where patients present with signs and symptoms of the disease. Such symptomatic visits ultimately lead to diagnosis, and many of these encounters represent diagnostic opportunities.

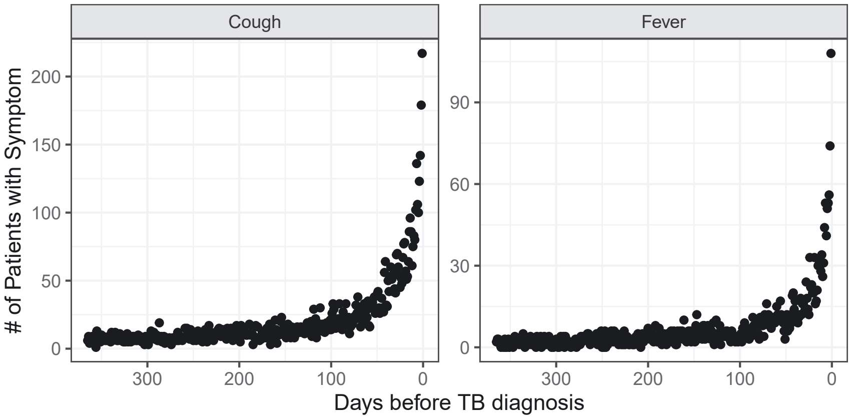

Figure 1: Symptomatic visits for cough and fever prior to tuberculosis diagnosis. The number of patients with a healthcare visit for a given symptom are presented each day prior to the index tuberculosis diagnosis. Similar increasing patterns of symptomatic visits have been found across a wide range of diseases.

The goal of our methodological approach is to identify the set of antecedent conditions suggesting a given disease is present (i.e., diagnosis codes for symptoms or other diseases with similar symptoms) and where encounters for such conditions may represent diagnostic opportunities. Thus, our objective is to find conditions that satisfy the increasing trends depicted in Figure 1. We do so in a clinically-guided fashion where a single *focal condition* known to be associated with the disease of interest (e.g., cough prior to tuberculosis) is selected by an expert, then other conditions exhibiting similar trends recommended by the clustering approach are evaluated by the expert.

Data processing and curve fitting

For this study, we utilized data from the IBM MarketScan Research Databases from 2001 to 2015. These data represent longitudinal health insurance claims in the United States covering inpatient and outpatient settings. We identified all patients diagnosed with the diseases of interest below. We computed the number of healthcare visits every day for each of the top 500 most common ICD-9-CM diagnosis codes during the year before the index diagnosis. To remove unrelated observation effects and create comparable model fits across codes we applied two data transformations. First, we converted daily counts of healthcare visits to relative frequencies by dividing the daily counts by the number of patients with a healthcare visit each day prior to diagnosis. Supplemental Figure 1 depicts how this can eliminate unrelated conditions appearing to increase due to patient observation. Second, we normalized the relative frequencies by subtracting the minimum and dividing by the range for each code.

Next, we estimated the temporal trends in the daily relative frequencies leading up to diagnosis using a piecewise linear regression model, where separate linear trends are estimated before and after a certain number of days prior to diagnosis (i.e., change-point). We enforced a continuity requirement on the model so the fitted trend is continuous before and after the change-point. We used the Akaike Information Criterion to select the optimal change-point. We apply this procedure to fit separate models for each of the top ICD-9-CM codes. The estimated parameters of these models are used in the clustering process. Each resulting model contains 4 parameters: a change-point, intercept, and two slope parameters (before and after the change-point).

Guided unsupervised clustering

Unsupervised clustering is a category of machine learning techniques that can be used to identify natural clusters of data that share similar characteristics and a relatively common approach is the *k*-means algorithm. Figure 2 provides a visualization that demonstrates the theoretical basis for how k-means clustering is applied to visit trends prior to diagnosis.

For each of the top ICD-9-CM codes in consideration, we applied the *k*-means clustering algorithm to the estimated parameters from the fitted models. Thus, we use the clustering algorithm to find clusters of ICD-9-CM codes that have trend parameters most similar to one another. We use the *kmeans* function as part of the *stats* package in the R programming language, with the default Hartigan and Wong algorithm [28]. We set an initial number of random cluster centers of 50 and a maximum number of iterations of 100. After identifying clusters for a given value *k*, we isolate the cluster containing the focal symptom of each disease, described below. Finally, we use both expert review and prior literature to determine which codes in the focal cluster are clinically plausible signs the disease may be present.

In order to provide a reproducible example, we have developed scripts for the R programming language that can be used to replicate all of the general approaches described in this analysis. This code can be found at https://github.com/aarmiller/diagnosis_cluster, along with synthetic data to demonstrate each stage of this analysis.

Example applications

To demonstrate the feasibility of our clustering approach, we selected two diseases with differing characteristics in trends of diagnostic opportunities – one where diagnostic delays occur over multiple weeks (tuberculosis) and the other where delays last multiple days (appendicitis). For each disease, we select a single well-known and common symptom to initiate the search process and select the focal cluster. We validated each case study populations by requiring one inpatient or >2 outpatient diagnoses to identify the index diagnosis.

Tuberculosis: Tuberculosis is a highly infectious communicable disease that is a major cause of morbidity and mortality, and is associated with over 1.4 million annual deaths worldwide [29]. Missed opportunities to diagnose tuberculosis are relatively common, often lasting up to a few months [9, 10, 30, 31], and have important public health implications [32]. Supplemental Table 1 provides the list of ICD-codes used to identify index cases of tuberculosis. The primary symptom associated with pulmonary tuberculosis selected as our focal condition is cough, ICD-9-CM code 786.2. Because the number of index cases of tuberculosis were relatively low and the delay duration can last multiple months, we aggregated daily counts of each ICD-9 code at a weekly level. We consider potential missed opportunities based on antecedent conditions within 90 days prior to the index diagnosis.

Appendicitis: Acute appendicitis is an inflammation of the appendix and one of the most common causes of emergency abdominal surgery [33, 34]. Most diagnostic delays occur within days, or at most a few weeks, of the index diagnosis [23, 35]. Delays in the diagnosis and treatment of appendicitis can be costly, potentially leading to perforated appendicitis [36, 37]. The primary symptom associated with appendicitis is abdominal pain, for which we used the ICD-9-CM code 789.00 (abdominal pain, unspecified site) as the focal condition. We consider potential missed opportunities based on antecedent conditions within 21 days prior to the index diagnosis.

Performance evaluation

To evaluate our approach, we compare different values for *k* from 2 to 25. For each of the resulting focal clusters we describe [1]: the size of the cluster [2]; the number of conditions that were clinically plausible, and [3]; the *clinically-plausible precision*, defined as the percent of conditions in a given cluster considered to be clinically plausible for diagnostic opportunities. Conditions were labelled as clinically plausible if they were either [1] a known sign or symptom of the disease or [2] a disease with similar symptoms (e.g., pneumonia instead of TB). To evaluate the stability of clusters for different values of *k*, we repeated our clustering approach 10,000 times for each *k* and report the 0.05 to 0.95 quantiles of resulting cluster measures. In addition, we compute the number of *potential* missed opportunities that may be captured using the cluster-

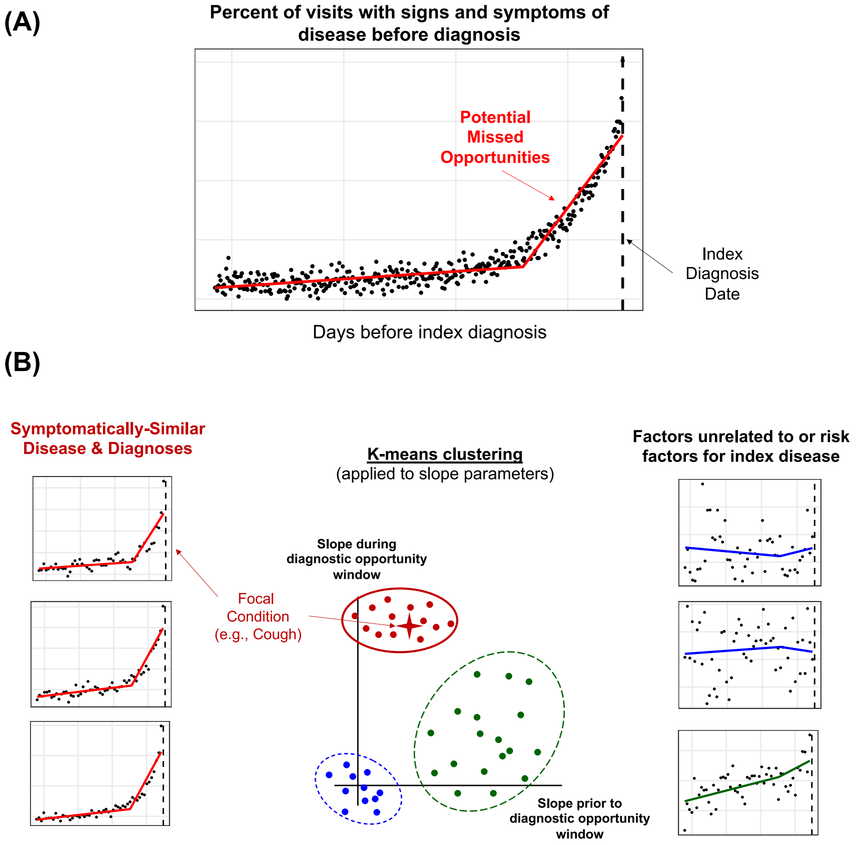

Figure 2: Visual depiction of the theoretical reasoning behind the clustering algorithm. Figure A depicts the empirical pattern of healthcare visits for related symptoms before the index disease diagnosis (data pictured correspond to tuberculosis). There is an increase in symptomatic healthcare visits before diagnosis. The trend is estimated with two curves: The first segment (flatter) captures the period where clinical disease is unlikely to be present, the second segment (steeper) captures symptoms of clinical disease and potential missed opportunities. Figure B depicts how the k-means clustering algorithm is applied to the trends for each potential antecedent condition (note, this is an oversimplified depiction where only the two slope parameters are used to identify k=3 clusters). The central plot depicts examples of clusters of conditions identified based on the slope parameters. The plots on either side of the clustering graph depict examples of trends that might fit the patterns of conditions shown in the corresponding cluster colors. The cluster containing the focal condition, such as cough (highlighted in red) has a slope near zero in the first period, and a large positive slope in the period right before diagnosis. The other two clusters contain slopes that are near zero in both periods (blue) or have a lesser slope in the second period (green) compared to those in the focal cluster.

based list of antecedent-conditions vs. the single focal symptom suggested by expert review (note, we refer to these as "potential" since the presence of symptoms during visits prior to diagnosis does not necessarily imply a missed opportunity).

Results

Example 1: Tuberculosis

We identified 19,423 patients with an index diagnosis of tuberculosis between 2001 and 2015. In total, these patients had 325,039 healthcare visit dates in the year prior to the index tuberculosis diagnosis. Figure 3 depicts the raw counts and relative frequency of visits for the focal symptom of cough prior to the index diagnosis. Figure 3 also depicts the fitted curve using the piecewise linear model, for which the objective of the clustering approach was to identify antecedent conditions with similar model parameters.

When applying the clustering approach, it is necessary to select a number of clusters k from which to obtain recommendations from the focal cluster. Supplemental Figure 2 provides a summary of results in terms of the overall size, number of clinically-plausible conditions, and clinically-plausible precision of the resulting focal clusters, across different values of k and clustering trials. We selected k=11 as a type of kink-point in the tradeoff between cluster size and clinically-plausible precision. For k=11, the level of clinically-plausible precision achieves a near maximum at 70.8%, suggesting a potential reviewer

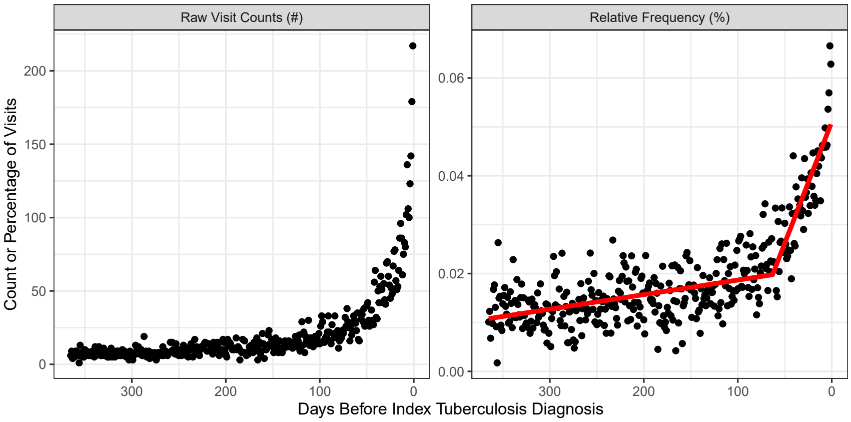

Figure 3: Visits for focal condition of cough prior to tuberculosis. The left figure depicts the raw counts. The right figure depicts counts converted to relative visit frequency along with the linear change-point model used to fit the trend and derive parameter estimates used for the cluster analysis.

would have few unrelated conditions to exclude. This focal cluster also results in a reasonably large number of clinically-plausible conditions (on average 34.1 conditions) recommended to a reviewer.

In total 88 conditions were identified as being clinically plausible across all clustering trials. The top 25 most frequently identified conditions are outlined in Table 1. This table also provides the frequency of each condition appearing in the focal cluster for *k*=11. As seen in Table 1, many of the commonly associated antecedent conditions, such as fever, hemoptysis, pneumonia or other pulmonary symptoms (pneumothorax, lung abscess, mass or neoplasm), were consistently identified across most focal clusters. Each of these top 25 conditions was selected in every focal cluster for k=11. Figure 4 depicts visit count trends for 9 of the top 25 antecedent conditions in the focal cluster.

To evaluate the potential effectiveness of our approach for suggesting criteria to identify potential diagnostic delays, we evaluated the number of visits and patients that would be captured by different sets of antecedent conditions within 90 days of the index diagnosis. Supplemental Figure 4 depicts the range in potential missed opportunities and patients identified using the sets of clinically plausible antecedent conditions recommended by each trial. The number of potential missed opportunities using the focal symptom of “cough” was 4,382 healthcare visits from 2,842 patients. Expanding to the set of conditions identified in the kink-point cluster (k=11) resulted in 31,162 visits from 9,078 patients representing a potential missed opportunity. Using the entire set of 88 plausible antecedent conditions recovered from our cluster-based approach resulted in 49,063 visits from 11,386 patients representing a potential missed opportunity. Thus, the prevalence of potential diagnostic delays identified ranged from 14.6% of patients, using cough alone, to 58.6% using all clinically plausible conditions suggested by the clustering algorithm.

Example 2: Appendicitis

We identified 572,836 patients with an index diagnosis of appendicitis between 2001 and 2015 that had over 4.5 million healthcare visit days in the year prior to the index diagnosis. Figure 5 depicts the raw count, relative frequency and fitted curves of visits for the focal symptom of “unspecified abdominal pain” prior to the index appendicitis diagnosis.

Supplemental Figure 5 depicts the results of our cluster analysis across different values of k and clustering trials. For values *k*≥4 there was a consistently high level of clinically-plausible precision >90%. We selected the value k=4 as the kink-point in the tradeoff between clinically-plausible precision (92.6%) and cluster size (54 total conditions and 50 that were clinically plausible).

A total of 63 clinically plausible conditions were identified across trials; Table 2 presents the top 25 conditions. The clustering approach identified many diagnoses known to be associated with symptoms of appendicitis, including specific sites for abdominal pain, vomiting, fever, nausea, leukocytosis, intestinal infection, or symptomatically-similar diseases such as gastritis, and pancreatitis. All of the top 43 clinically plausible conditions appeared in every focal cluster for k=4. Figure 6 depicts visit count trends for 9 of the antecedent conditions selected from the 43 conditions in the focal cluster.

We computed the number of potential missed opportunities within 21 days prior to the index appendicitis diagnosis using different sets of antecedent conditions.

Table 1: Top 25 clinically-plausible antecedent conditions for tuberculosis recommended by the "cough" focal cluster.

ICD-9-CM code	Percent of all focal clusters that contained code	Percent of focal clusters for k=11 that contained code	Description
786.2	100	100	Cough
780.6	100	100	Fever and other physiologic disturbances of temperature regulation
786.30	100	100	Hemoptysis, unspecified
482.9	99.96	100	Bacterial pneumonia, unspecified
423.9	99.92	100	Unspecified disease of pericardium
784.2	99.11	100	Swelling, mass, or lump in head and neck
785.6	99.11	100	Enlargement of lymph nodes
793.1	99.11	100	Lung field
786.09	99.11	100	Other respiratory abnormalities
486	99.11	100	Pneumonia, organism unspecified
786.3	99.11	100	Hemoptysis
793.11	99.09	100	Solitary pulmonary nodule
518.81	99.08	100	Acute respiratory failure
793.19	99.05	100	Other nonspecific abnormal finding of lung field
518.82	98.71	100	Other pulmonary insufficiency, not elsewhere classified
235.7	98.69	100	Neoplasm of uncertain behavior of trachea, bronchus, and lung
511.9	98.67	100	Unspecified pleural effusion
482.89	91.91	100	Pneumonia due to other specified bacteria
239.1	91.85	100	Neoplasm of unspecified nature of respiratory system
780.60	85.92	100	Fever, unspecified
512.8	84.41	100	Other pneumothorax and air leak
518.89	84.39	100	Other diseases of lung, not elsewhere classified
786.6	84.39	100	Swelling, mass, or lump in chest
485	83.13	100	Bronchopneumonia, organism unspecified
513.0	80.9	100	Abscess of lung

Conditions are ordered by the percent of focal clusters containing the antecedent condition appeared, across 10,000 trials and values of *k* from 2 through 25 (for a total of 240,000 different clusters). See Supplemental Table 2 for the 63 remaining conditions that were identified as clinically plausible.

Supplemental Figure 7 depicts the resulting number of potential missed opportunities and patients identified using the antecedent conditions for different values of k. The focal symptom of unspecified abdominal pain identified 49,371 potential missed opportunities from 41,596 patients. Expanding to the additional set of 50 antecedent conditions identified in the cluster k=4 resulted in 137,003 visits from 98,111 patients representing a potential missed opportunity. Using the entire set of plausible antecedent conditions recovered from our cluster-based approach resulted in 142,359 visits from 101,013 patients representing a potential missed opportunity. Thus, the potential prevalence of identified diagnostic delays among our study population ranged from 7.3% of patients, using abdominal pain alone, to 17.6%, using all clinically plausible conditions suggested by the clustering algorithm.

Discussion

In this study we proposed a machine learning approach to recommend symptoms or antecedent health conditions that may indicate potential missed diagnostic opportunities. Starting with a single well-described symptom (e.g., cough for tuberculosis) we used unsupervised k-means clustering to identify other conditions that exhibited similar visit patterns prior to the actual diagnosis. Our findings demonstrated that a large number of symptoms and conditions could be identified with a high degree of precision in terms of clinical plausibility. Moreover, inclusion of these additional conditions resulted in more than twice the number of potential diagnostic opportunities and patients being identified with a potential diagnostic delay compared to using a single common symptom.

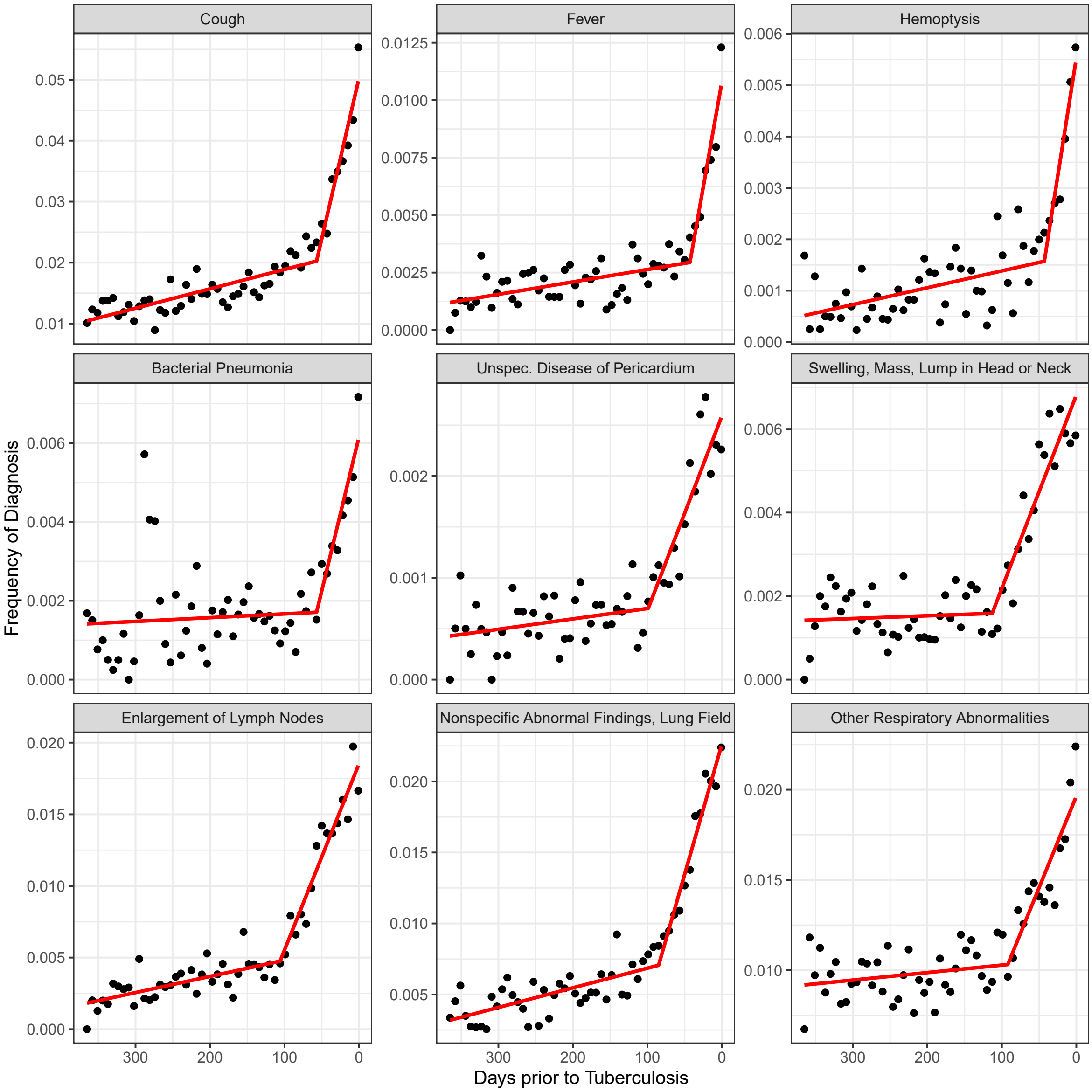

Figure 4: Examples of trends in top antecedent conditions selected in the “cough” focal cluster prior to tuberculosis. The black dots depict 7-day average counts of visits with the given diagnosis relative to visit frequency. The linear piecewise model used to fit the trend and derive parameter estimates for the cluster analysis is depicted by the red line (see Supplemental Figure 3 for the remaining top 25 conditions).

Mining administrative data sources is a promising approach to study the diagnostic process, but methods for many of these approaches are still in the early stages of development [38]. The size of these datasets may offer the potential to make discoveries in the diagnostic process that inform the study of diagnostic errors. One particular issue that arises in the study of diagnostic errors is the need to use clinical expertise to define the criteria for healthcare visits that represent potential missed opportunities to diagnose a disease. We demonstrated a relatively simple machine-learning approach that can scan hundreds or thousands of diagnostic codes to identify those with similar patterns to known symptoms of disease. We show our approach can effectively recover a large set of clinically plausible conditions that could be used to detect diagnostic delays. Indeed, we found that the additional conditions recommended by our clustering-based approach significantly increased the potential number of diagnostic opportunities identified, compared to a single symptom. While such criteria must ultimately be refined based on clinical expertise, this exploratory process may significantly aid in the study of diagnostic delays. Our

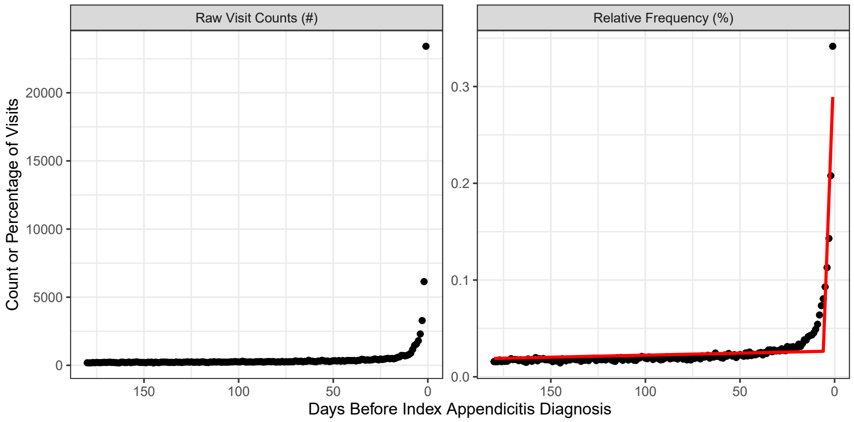

Figure 5: Visits for focal condition of unspecified abdominal pain prior to appendicitis. The left figure depicts the raw counts. The right figure depicts counts relative to visit frequency along with the linear change-point model fit to the data. Note: Counts are only depicted over the 180 days prior to diagnosis to better illustrate the trend. The fitted curves rely on all 365 days prior to diagnosis.

Table 2: Top 25 clinically-plausible antecedent conditions for appendicitis recommended by the "unspecified abdominal pain" focal cluster.

ICD-9-CM code	Percent of all focal clusters that contained code	Percent of focal clusters for k=4 that contained code	Description
789.00	100	100	Abdominal pain, unspecified site
789.05	100	100	Abdominal pain, periumbilic
789.07	100	100	Abdominal pain, generalized
787.03	99.99	100	Vomiting alone
789.03	99.99	100	Abdominal pain, right lower quadrant
288.60	99.99	100	Leukocytosis, unspecified
535.50	99.99	100	Unspecified gastritis and gastroduodenitis, without mention of hemorrhage
789.06	99.99	100	Abdominal pain, epigastric
577.0	86.6	100	Acute pancreatitis
593.2	86.6	100	Cyst of kidney, acquired
646.83	86.6	100	Other specified complications of pregnancy, antepartum condition or complication
648.93	86.6	100	Other current conditions classifiable elsewhere of mother, antepartum condition or complication
V72.6	86.6	100	Laboratory examination
796.9	84.77	100	Other nonspecific abnormal findings
789.09	84.28	100	Abdominal pain, other specified site
558.9	62.81	100	Other and unspecified noninfectious gastroenteritis and colitis
780.60	62.81	100	Fever, unspecified
535.00	62.81	100	Acute gastritis, without mention of hemorrhage
787.01	60.06	100	Nausea with vomiting
787.02	59.63	100	Nausea alone
276.51	59.63	100	Dehydration
088	59.57	100	Intestinal infection due to other organism, not elsewhere classified
789.01	56.88	100	Abdominal pain, right upper quadrant
564.00	47.72	100	Constipation, unspecified
780.6	47.72	100	Fever and other physiologic disturbances of temperature regulation

Conditions are ordered by the percent of focal clusters containing the antecedent condition, across 10,000 trials and values of *k* from 2 through 25 (for a total of 240,000 different clusters). See Supplemental Table 4 for the remaining 38 conditions that were deemed clinically plausible.

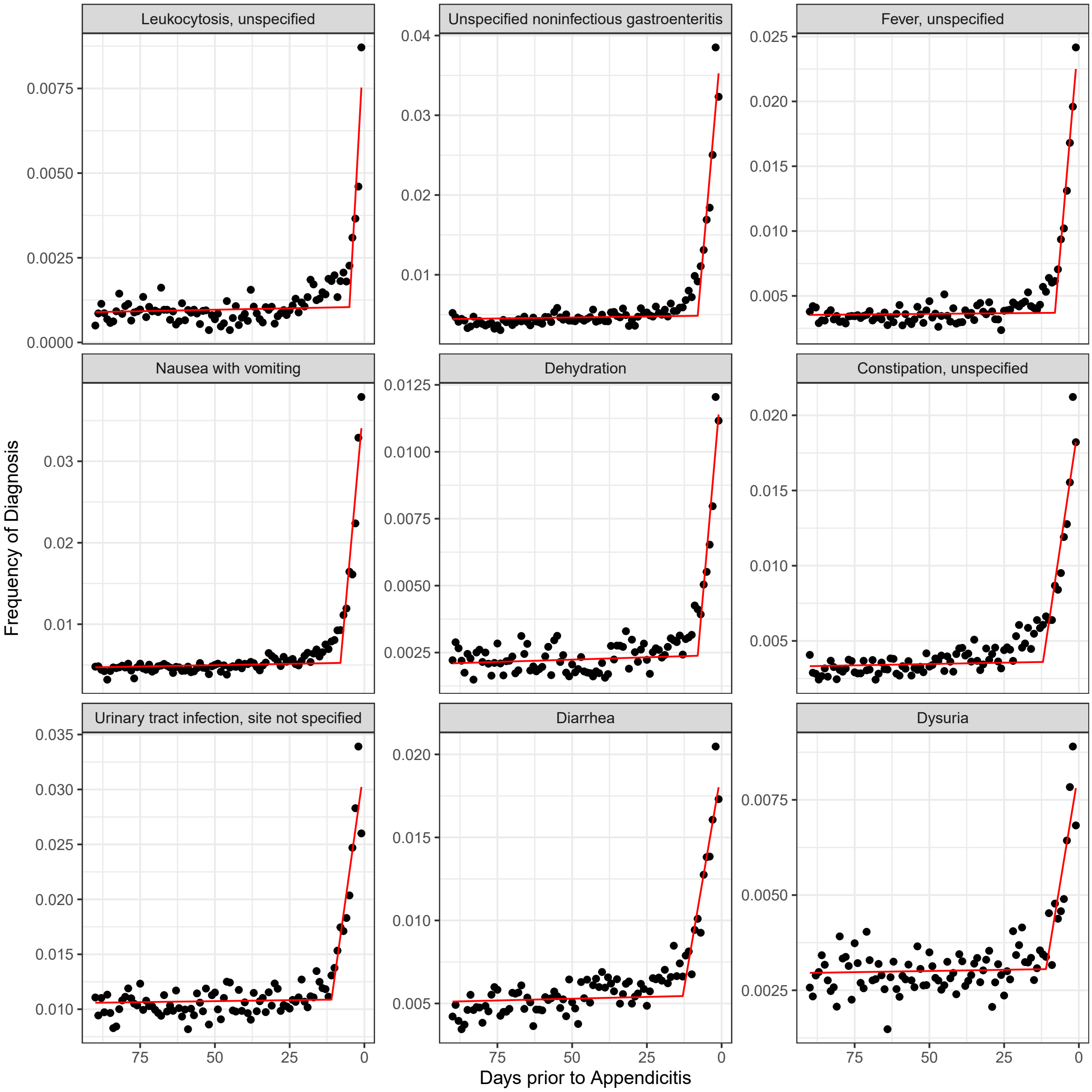

Figure 6: Examples of selected trends in antecedent conditions contained in the “abdominal pain” focal cluster prior to appendicitis. The black dots depict 7-day average counts of visits with the given diagnosis relative to visit frequency. The linear piecewise model used to fit the trend and derive parameter estimates for the cluster analysis is depicted by the red line (see Supplemental Figure 6 for the remaining top 25 conditions).

approach may also be applied to study designs involving non-administrative data sources; for example, by first identifying the set of conditions to search for during a retrospective chart review.

In addition to recommending antecedent conditions that might indicate diagnostic opportunities, similar techniques may also be integrated into other aspects of the diagnostic process, such as the development of diagnostic training exercises, clinical decision support systems, or trigger rules designed to flag potential diagnostic errors. Our approach could also be used to increase the current understanding of the natural history of diseases, by describing the frequency and temporal ordering of different symptoms prior to diagnosis. For example, this approach could be applied to discover which conditions appear weeks before a diagnosis vs. days before.

We demonstrated a simple k-means application using only limited clinical feedback (i.e., to select the focal cluster and evaluate clinical plausibility of recommended conditions). However, there are numerous ways to extend our methodology and more thoroughly integrate clinical expertise. For example, other types of event codes can be considered (e.g., procedure, medication), different curve fitting and change-point techniques can be applied, other information can be integrated into the feature space, and different clustering algorithms can be used (e.g., semi-supervised approaches). In addition, greater clinical feedback may be integrated into the analysis process. Recommendations can be made beyond a single focal cluster using multiple initial conditions suggested by experts. Sequential and iterative approaches may be used to expand, grow, or combine clusters with the feedback of expert reviewers, and unrelated data points or clusters may be excluded. Clusters might also be labeled by clinical reviewers as representing other aspects of disease, such as risk factors or triggering events (e.g., alcohol use or infection prior to stroke). Each of these represent possible future extensions of this approach.

There are a number of limitations to consider when using our approach and administrative data to study diagnostic delays. First, administrative data are generated for billing purposes; patterns that emerge may be the biproduct of the administrative data generating process and may omit information in the clinical record. Clinical expertise is critical for evaluation, and results should generally be regarded as exploratory or hypothesis generating. Second, different approaches may yield dramatically different results depending on the model fitting and clustering approaches used. Third, as with any machine-learning based recommendation system, computational resource costs may need to be considered and more advanced techniques may require additional computing resources. Fourth, these methods, especially the curve fitting approaches, require a sufficiently large number of observations. Smaller datasets may lack sufficient observations to obtain stable results across granular codes, and aggregation (e.g., using Clinical Classification Software codes) may need to be applied to identify related code sets. Finally, we used data from the United States and our results may not generalize to other locations.

Cluster-based approaches, coupled with large administrative data sources, may help discover patterns in the diagnostic process. The approach we presented provides an easy-to-implement recommender system that can allow future investigators to mine large databases for potential signals of disease and better study the diagnostic process. There remain a wide range of extensions to the proposed methodological framework. Future investigations should explore how this framework and other machine-learning based approaches may aid the discovery process for studying diagnostic delays.

Research funding: This work was supported by the Agency for Healthcare Research and Quality grant number 5R01HS027375.

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Competing interests: Authors state no conflict of interest.

Informed consent: Not applicable.

Ethical approval: The University of Iowa Institutional Review Board deems this study to be human studies exempt.

References

1. Singh H, Graber ML. Improving diagnosis in health care-the next imperative for patient safety. N Engl J Med 2015;373:2493–5.
2. Newman-Toker DE, Schaffer AC, Yu-Moe CW, Nassery N, Saber Tehrani AS, Clemens GD, et al. Serious misdiagnosis-related harms in malpractice claims: the “Big Three” - vascular events, infections, and cancers. Diagnosis (Berl) 2019;6:227–40.
3. Zwaan L, Singh H. The challenges in defining and measuring diagnostic error. Diagnosis (Berl) 2015;2:97–103.
4. Graber ML. The incidence of diagnostic error in medicine. BMJ Qual Saf 2013;22(2 Suppl):ii21–7.
5. Singh H, Sittig DF. Advancing the science of measurement of diagnostic errors in healthcare: the Safer Dx framework. BMJ Qual Saf 2015;24:103–10.
6. National Academies of Sciences E, Medicine. Improving diagnosis in health care. Washington, DC: National Academies Press; 2015.
7. Moy E, Barrett M, Coffey R, Hines AL, Newman-Toker DE. Missed diagnoses of acute myocardial infarction in the emergency department: variation by patient and facility characteristics. Diagnosis 2015;2:29–40.
8. Newman-Toker DE, Moy E, Valente E, Coffey R, Hines AL. Missed diagnosis of stroke in the emergency department: a cross-sectional analysis of a large population-based sample. Diagnosis (Berl) 2014;1:155–66.
9. Miller AC, Arakkal AT, Koeneman S, Cavanaugh JE, Gerke AK, Hornick DB, et al. Incidence, duration and risk factors associated with delayed and missed diagnostic opportunities related to tuberculosis: a population-based longitudinal study. BMJ Open 2021;11:e045605.
10. Miller AC, Polgreen LA, Cavanaugh JE, Hornick DB, Polgreen PM. Missed opportunities to diagnose tuberculosis are common among hospitalized patients and patients seen in emergency departments. Open Forum Infect Dis 2015;2:ofv171.
11. Hester LL, Gifkins DM, Bellow KM, Vermeulen J, Schecter JM, Strony J, et al. Diagnostic delay and characterization of the clinical prodrome in AL amyloidosis among 1523 US adults diagnosed between 2001 and 2019. Eur J Haematol 2021;107:428–35.
12. Salazar AS, Keller MR, Olsen MA, Nickel KB, George IA, Larson L, et al. Potential missed opportunities for diagnosis of cryptococcosis and the association with mortality: a cohort study. E Clin Med 2020;27:100563.
13. Surrey E, Soliman AM, Trenz H, Blauer-Peterson C, Sluis A. Impact of endometriosis diagnostic delays on healthcare resource utilization and costs. Adv Ther 2020;37:1087–99.
14. Benedict K, Lyman M, Jackson BR. Possible misdiagnosis, inappropriate empiric treatment, and opportunities for increased diagnostic testing for patients with vulvovaginal candidiasis-United States, 2018. Plos One 2022;17:e0267866.

15. Chase DM, Neighbors J, Perhanidis J, Monk BJ. Gastrointestinal symptoms and diagnosis preceding ovarian cancer diagnosis: effects on treatment allocation and potential diagnostic delay. Gynecol Oncol 2021;161:832–7.
16. Miller AC, Arakkal AT, Koeneman SH, Cavanaugh JE, Thompson GR, Baddley JW, et al. Frequency and duration of, and risk factors for, diagnostic delays associated with histoplasmosis. Journal of Fungi 2022;8:438.
17. Miller AC, Koeneman SH, Arakkal AT, Cavanaugh JE, Polgreen PM. Incidence, duration, and risk factors associated with missed opportunities to diagnose herpes simplex encephalitis: a population-based longitudinal study. Open Forum Infect Dis 2021;8:ofab400.
18. Benedict K, Kobayashi M, Garg S, Chiller T, Jackson BR. Symptoms in blastomycosis, coccidioidomycosis, and histoplasmosis versus other respiratory illnesses in commercially insured adult outpatients, United States, 2016-2017. Clin Infect Dis 2020;73: e4336–44.
19. Benedict K, Beer KD, Jackson BR. Histoplasmosis-related healthcare use, diagnosis, and treatment in a commercially insured population, United States. Clin Infect Dis 2020;70:1003–10.
20. Nassery N, Horberg MA, Rubenstein KB, Certa JM, Watson E, Somasundaram B, et al. Antecedent treat-and-release diagnoses prior to sepsis hospitalization among adult emergency department patients: a look-back analysis employing insurance claims data using Symptom-Disease Pair Analysis of Diagnostic Error (SPADE) methodology. Diagnosis (Berl) 2021;8:469–78.
21. Gold JAW, Jackson BR, Benedict K. Possible diagnostic delays and missed prevention opportunities in pneumocystis pneumonia patients without HIV: analysis of commercial insurance claims data-United States, 2011-2015. Open Forum Infect Dis 2020;7:ofaa255.
22. Sharp AL, Baecker A, Nassery N, Park S, Hassoon A, Lee MS, et al. Missed acute myocardial infarction in the emergency department-standardizing measurement of misdiagnosis-related harms using the SPADE method. Diagnosis (Berl) 2021;8:177–86.
23. Mahajan P, Basu T, Pai CW, Singh H, Petersen N, Bellolio MF, et al. Factors associated with potentially missed diagnosis of appendicitis in the emergency department. JAMA Netw Open 2020;3:e200612.
24. Liberman AL, Newman-Toker DE. Symptom-Disease Pair Analysis of Diagnostic Error (SPADE): a conceptual framework and methodological approach for unearthing misdiagnosis-related harms using big data. BMJ Qual Saf 2018;27:557–66.
25. Bjerager M, Palshof T, Dahl R, Vedsted P, Olesen F. Delay in diagnosis of lung cancer in general practice. Br J Gen Pract 2006; 56:863–8.
26. Park DH, Kim HK, Choi IY, Kim JK. A literature review and classification of recommender systems research. Expert Syst Appl 2012;39:10059–72.
27. Waxman DA, Kanzaria HK, Schriger DL. Unrecognized cardiovascular emergencies among medicare patients. JAMA Intern Med 2018;178:477–84.
28. Hartigan JA, Wong MA. Algorithm as 136: a K-means clustering algorithm. J R Stat Soc Series C 1979;28:100–8.
29. Chakaya J, Khan M, Ntoumi F, Aklillu E, Fatima R, Mwaba P, et al. Global tuberculosis report 2020 - reflections on the global TB burden, treatment and prevention efforts. Int J Infect Dis 2021; 113(1 Suppl):S7–12.
30. Wallace RM, Kammerer JS, Iademarco MF, Althomsons SP, Winston CA, Navin TR. Increasing proportions of advanced pulmonary tuberculosis reported in the United States: are delays in diagnosis on the rise? Am J Respir Crit Care Med 2009;180:1016–22.
31. Loutet MG, Sinclair C, Whitehead N, Cosgrove C, Lalor MK, Thomas HL. Delay from symptom onset to treatment start among tuberculosis patients in England, 2012-2015. Epidemiol Infect 2018;146:1511–8.
32. Mindra G, Wortham JM, Haddad MB, Powell KM. Tuberculosis outbreaks in the United States, 2009-2015. Public Health Rep 2017;132:157–63.
33. Buckius MT, McGrath B, Monk J, Grim R, Bell T, Ahuja V. Changing epidemiology of acute appendicitis in the United States: study period 1993-2008. J Surg Res 2012;175:185–90.
34. Pittman-Waller VA, Myers JG, Stewart RM, Dent DL, Page CP, Gray GA, et al. Appendicitis: why so complicated? Analysis of 5755 consecutive appendectomies. Am Surg 2000;66:548–54.
35. Choi JY, Ryoo E, Jo JH, Hann T, Kim SM. Risk factors of delayed diagnosis of acute appendicitis in children: for early detection of acute appendicitis. Korean J Pediatr 2016;59:368–73.
36. Papandria D, Goldstein SD, Rhee D, Salazar JH, Arlikar J, Gorgy A, et al. Risk of perforation increases with delay in recognition and surgery for acute appendicitis. J Surg Res 2013; 184:723–9.
37. Glerum KM, Selbst SM, Parikh PD, Zonfrillo MR. Pediatric malpractice claims in the emergency department and urgent care settings from 2001 to 2015. Pediatr Emerg Care 2021;37: e376–9.
38. Singh H, Bradford A, Goeschel C. Operational measurement of diagnostic safety: state of the science. Diagnosis 2021;8:51–65.

Supplementary Material: The online version of this article offers supplementary material (https://doi.org/10.1515/dx-2022-0044).