Detecting Suspected Canine Hypoadrenocorticism in Free-Text Veterinary Discharge Notes Using Weak Supervision

Nathan Bollig[1,2], Chitrasen Mohanty[3], Lokeswar Sadasivuni[4]

Hypoadrenocorticism (Addison's disease) is a canine endocrine disorder that can cause significant suffering and death, yet carries a good prognosis when appropriately treated. Sometimes called "the great pretender", it often is mistaken for other conditions. In this work, we developed a tool to retrospectively identify cases of suspected canine hypoadrenocorticism in a historical corpus of discharge documents from a veterinary hospital, using a machine learning classifier trained on labels obtained through weak supervision. Our system achieves 100% recall and 11.12% precision, which demonstrates improvement over a baseline classifier that achieves 100% recall and 7.52% precision. Alternatively, training on human labels provides a 1.46x relative boost in average precision. We conclude that weak supervision is effective for training a minimally performant system that would provide utility as a case selection agent in a retrospective clinical study.

1 Introduction

In the following, we describe canine hypoadrenocorticism, the motivation for our machine learning task, and the rationale for employing weak supervision.

1.1 Canine hypoadrenocorticism

Hypoadrenocortism, or Addison's disease, is a disorder caused by the functional loss of 85-90% of the adrenal cortex, resulting in deficiency of glucocorticoid and mineralocorticoid hormones. It is most commonly caused by immune-mediated destruction of the adrenal cortex, but can be caused by insufficient secretion of adrenocorticotropic hormone (ACTH) from the pituitary gland or corticotropin-releasing hormone (CRH) from the hypothalamus due to intracranial disease [1]. Addison's disease is rare in cats and uncommon in dogs, affecting between 0.06% and 0.28% of the naturally-occurring canine population, but a higher prevalence of 1.4% to 9.4% has been observed in at-risk breeds such as bearded collies and Nova Scotia duck tolling retrievers [2]. Higher prevalence within particular canine breeds is consistent with human studies that have demonstrated a genetic predisposition for autoimmune hypoadrenocorticism [3].

Canine hypoadrenocorticism is an important disease because while it is uncommon and therefore sometimes overlooked, it is still not exceptionally rare within a population of clinically affected dogs presenting to a veterinary hospital. It is often called the "great pretender" because its clinical signs and diagnostic findings often appear similar to other diseases [1].

[1] Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI, United States of America

[2] Department of Pathobiological Sciences, School of Veterinary Medicine, University of Wisconsin-Madison, Madison, WI, United States of America

[3] Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI, United States of America

[4] Department of Statistics, University of Wisconsin-Madison, Madison, WI, United States of America

Addison's disease is characterized by reduced production of glucocorticoid and mineralocorticoid hormones, the most important of which are cortisol and aldosterone respectively. Cortisol affects nearly every tissue in the body, with some of its important functions including maintenance of blood pressure, water balance, vascular volume, vascular permeability, and maintenance of normal blood glucose levels. Because it has a role in counteracting the physiological effects of stress, clinical signs may initially only be evident in times of stress or may be acutely exacerbated by stress. On the other hand, aldosterone targets the kidney and is primarily responsible for managing sodium (Na), potassium (K), and water balance. [4]

Because of its impact on multiple physiologic systems, the clinical signs of this condition (Table 1) are often vague, nonspecific, and may be misconstrued as primary gastrointestinal (GI) disease, kidney failure, or neurologic disease. The effects can be mild but are persistent and progressive in over 50% of cases [4]. Patients can be chronically ill or can present in an acute state of cardiovascular shock, sometimes called an "Addisonian crisis".

Table 1. Potential clinical signs of canine hypoadrenocorticism.

Non-specific illness	Inappetence, anorexia, lethargy, depression, weakness, shaking, weight loss
Gastrointestinal signs	Vomiting or regurgitation, diarrhea, abdominal pain, melena, hematochezia
Hypovolemic shock	Bradycardia or tachycardia, collapse, hypothermia, weak pulses, or poor capillary refill time
Complete blood count (CBC) abnormalities	Mild to moderate normocytic, normochromic, nonregenerative anemia (PCV 20-35%); absence of a stress leukogram in a sick animal; lymphocytosis
Chemistry panel abnormalities	Low sodium, high potassium, Na:K ratio < 27:1, low chloride, azotemia, high calcium, mild acidosis, low blood glucose, low albumin, low cholesterol, increased ALT or AST

Definitive diagnosis depends on the veterinarian forming a clinical suspicion, then performing confirmatory diagnostic tests. Initial assessment of baseline cortisol is usually pursued, because normal levels of cortisol demonstrate adequate adrenocortical function and can effectively rule out Addison's disease. If the patient's resting cortisol is low, a definitive result is achieved by the ACTH stimulation test. In this test, purified porcine ACTH or synthetic ACTH is administered, and the patient's cortisol level is measured again 1 to 2 hours later. In a patient with normal adrenocortical function, ACTH should stimulate a marked increase in cortisol production. In an Addisonian patient, cortisol levels are not significantly affected because the impaired adrenal cortex is unable to respond to stimulation [5]. Following a diagnosis of Addison's disease, prognosis is good with lifelong medical therapy and lifestyle management [1].

We conclude that the veterinary clinician faces several notable challenges in addressing Addison's disease. First, while sufficiently uncommon to be overlooked, it has a nontrivial prevalence in sick, hospitalized, and breed subpopulations of canines. Second, the disease is often chronic or life-threatening, but patients have a good prognosis if properly treated, so that the opportunity cost of misdiagnosis is high. Third, the disease is difficult to diagnose because the signs are nonspecific and vague, often mimicking other diseases.

1.2 Motivation for our machine learning task

We aim to *retrospectively identify suspected cases of Addison's disease from clinical text*. A priori, this is not easy to do without clinical signs and diagnoses in a structured form, but it would enable studies of breed prevalence, diagnostic test outcomes (sensitivity, positive predictive value, etc.), and risk factors for misdiagnosis within a population of suspected Addison's patients at a particular hospital. The proposed model, functioning as a case selection agent, would enable a retrospective study that otherwise might be infeasible when clinical signs and diagnoses are stored only in free text.

Since diagnosis and treatment of Addison's disease is dependent on the veterinarian first forming a suspicion that the disease is present, there would be value also in framing this prediction task as an automated clinical decision support tool within the electronic health record (EHR). Machine learning (ML) is widely used in human clinical decision support systems [6], but at this time is far less common in veterinary systems. Recent work has shown that features derived from screening blood tests in animals can be used to train an AdaBoost model to predict hypoadrenocorticism case vs. control status with high sensitivity and specificity [7]. However, our study will focus on the area of free text input, since often in veterinary EHR systems is it infeasible to directly extract structured laboratory test features. To our knowledge, there have been no attempts to carry out this task using unstructured text input.

1.3 Weak supervision

Because veterinary EHRs often do not capture structured data and veterinary informatics researchers are not in high abundance, ML model development is highly sensitive to the bottleneck of label acquisition. Current research in the domain of veterinary text classification either depends on large, expensive datasets that are somewhat atypical in veterinary settings [8,9] or on manually assigned labels. Weak supervision provides an alternative to human-labeling, in which expert input is leveraged to programmatically generate multiple noisy labels (a process called data programming), which are then combined without direct human supervision [10]. This approach has been proposed to handle small numbers of potentially unipolar weak supervision sources, by using an algorithm based on matrix completion to estimate the parameters of a class-conditional label model [11].

Currently, there appears to be no mention of weak supervision techniques in the veterinary clinical informatics literature. The primary objective of this project is to evaluate the weak supervision approach applied to this task.

2 Methods

2.1 Data

Free-text discharge documents were obtained from the University of Wisconsin School of Veterinary Medicine. Raw data consisted of all standard-format electronic discharges from the Primary Care and Small Animal Internal Medicine departments from July 25, 2014 to February 3, 2020, for a total of 27,001 documents from 8,107 patients. Reports were in HTML format and appeared to have been created from standard templates that changed over time and were inconsistent across departments. Despite this variable structure, many reports included sections such as (1) history, (2) physical exam, (3) problem list or diagnoses, (4) diagnostic tests, and (5) treatments or medications to go home. The data also included report date, patient name, medical record number, sex, species, and the department in which they were examined.

2.2 Data and code availability

Because the report text is unstructured, no portion of raw data is publicly available, but can be made available by reasonable request to the medical records department at the University of Wisconsin School of Veterinary Medicine. All of our project code is publicly available at https://github.com/nathanbollig/vet-note-classifier.

2.3 Data pre-processing

An important initial step was to filter out all non-canine patients, leaving 17,066 documents from 5,381 patients. Common strings were removed, such as header and footer information that appeared on many documents in the corpus. Document text was changed to all lower case and punctuation was removed.

Using the BeautifulSoup HTML parser (version 4.6.3), we identified bolded section titles and mapped them to a standard list of common medical discharge *section types* by looking for specific keywords. For each bolded title that was mapped to a section type, the report text between that title and the next bolded title was defined as the corresponding section of the document. The sections extracted were (1) history, (2) physical exam, (3) diagnoses, (4) diagnostic tests, and (5) treatments. See Table S1 in Appendix A for the keywords used to map section titles to section types.

2.4 Obtaining human labels

A veterinarian (one of the authors) prepared a written set of criteria to guide human labeling based on the clinical features of this disease (Appendix B). Another author labeled documents based on these guidelines and discussion with the veterinarian. Only in a small number of cases were the guidelines in Appendix B not conclusive, and in these cases a label was decided based on a holistic assessment of the document.

A script was written to display a random discharge note from a randomly-selected patient, and labels were recorded in a spreadsheet for each note encountered. A total of 999 documents were labeled. A subset of labels was evaluated to ensure agreement between the labeler and the veterinarian.

2.5 Forming a label model

The Snorkel framework (version 0.9.3) was used to create a label model using programmatic labeling functions [12]. Given a raw document space D, a labeling function (LF) is a mapping

$$f: D \ \rightarrow \{0, 1, -1\}$$

where 1 represents a positive label, 0 a negative label, and -1 abstention. The range of each f is either $\{0, 1, -1\}$ (bipolar), $\{0, -1\}$ (unipolar negative), or $\{1, -1\}$ (unipolar positive). A label model

$$L: D \ \rightarrow \{0,1, -1\}$$

is learned from the collection of individual labelers. The outcome $L(x) = -1$ occurs when all labelers abstain on *x*.

Labeling functions were designed through an iterative process of programming, debugging, and evaluation on the tuning set. Labeling functions were organized into four categories, based on whether they were designed to match on (1) clinical signs of hypoadrenocorticism, (2) post-hoc evidence of hypoadrenocorticism, (3) laboratory findings consistent with hypoadrenocorticism, or (4) conditions that rule-out hypoadrenocorticism. Tables 2-5 presents a summary of these labeling functions. Most were keyword-based, with keywords organized into logical groups within each category. Each function was designed to evaluate either the full document text or a specific section of text extracted in pre-processing.

Table 2. Labeling functions related to clinical signs.

Group	Polarity	Keywords	Matched on
Appetite reduction	Positive	inappetance, anorexia, not eating, weight loss	Full document
GI signs 1	Positive	vomiting, diarrhea	Full document
GI signs 2	Positive	regurg	Full document
GI signs 3	Positive	melena, hematochezia	Full document
GI signs 4	Positive	abdominal pain	Full document
Non-specific signs 1	Positive	letharg, depression, depressed	Full document
Non-specific signs 2	Positive	shake, shakes, shaking, weak	Full document
Non-specific signs 3	Positive	hair loss	Full document
Signs of hypovolemic shock	Positive	bradycardia, low heart rate, tachycardia, high heart rate, collapse, hypothermia, low body temp, weak pulse, poor capillary refill, shock, hypovolem	Full document

Table 3. Labeling functions related to post-hoc evidence of suspected hypoadrenocorticism.

Group	Polarity	Keywords	Matched on
Diagnosis mentioned	Positive	hypoadrenocorticism, addisons, addison’s	Diagnoses
Diagnostic test	Positive	baseline cort, baseline cortisol, acth	Diagnostic Tests
Treatment	Positive	desoxycorticosterone pivalate, desoxycorticosterone, fludrocortisone acetate, fludrocortisone, florinef, docp	Treatment

Table 4. Labeling functions related to diagnostic findings.

Group	Polarity	Keywords	Matched on
Hyponatremia	Positive	hyponatremia, low sodium, low na, sodium low, na low	Diagnostic Tests
Hyperkalemia	Positive	hyperkalemia, high potassium, high k, potassium high, k high	Diagnostic Tests

Table 5. Labeling functions related to conditions that rule-out hypoadrenocorticism.

Group	Polarity	Keywords	Matched on
Healthy	Negative	healthy	Full document
Kidney	Negative	acute kidney, aki, akd	Full document
Parasites	Negative	parasite, worm, roundworm, toxocara, hookworm, ancyclostoma, uncinaria, whipworm, trichuris, tapeworm, cestod, taenia, echinococcus, spirometra, diphyllobothrium, mesocestoides, dipylidium, fluke, trematode, giardia	Full document
Liver	Negative	liver failure, shunt, hepatitis, cancer, lepto, hepatitis, copper, cholangio, diabetes	Full document
Pancreas	Negative	pancreatitis, pancreatic	Full document

Toxins	Negative	toxin, heavy metal, herbicide, fungicide, insecticide, rodent, aflatoxin, amanita, cycad, sago palm, algae	Full document
Effusion	Negative	effusi	Full document
Primary GI	Negative	megaesoph, dilatation, gastritis, foreign body, intuss, enteritis, colitis	Full document

We also studied composition and several enhancements to these labeling functions. Appendix C describes a composition framework and labeling function enhancements evaluated in this work.

2.6 Featurization

Although labeling functions act on raw documents in document space D, we considered final discriminative classifiers acting on a feature space $X \in \mathbb{R}^{\mathrm{nxm}}$ in which each of *n* documents was represented by *m* features. The featurization mapping $\mathcal{F}: D \rightarrow X$ converted each document to a vector of TF-IDF statistics for unigrams and bigrams, as defined below.

Term frequency – inverse document frequency

Document text was tokenized into unigrams and bigrams. Numbers, punctuation, and stop words were removed and all characters were changed to lower case. Tokens that appeared only once in the corpus or with document frequency greater than 70% were ignored.

The resulting feature space X consisted of one row for each document and one column for each unique term in the corpus. Each entry in X represented the term frequency – inverse document frequency (TF-IDF) for the corresponding document and word. To be precise, if n_{ij} represents the number of times term t_i appears in document d_j then the frequency of term t_i in document d_j is

$$\mathrm{TF} = \frac{n_{ij}}{\text{Total number of terms in } d_j}.$$

The inverse document frequency (IDF) of term t_i is

$$\mathrm{IDF} = 1 + \log \frac{1 + \text{Total number of documents}}{1 + \text{Number of documents containing } t_i}.$$

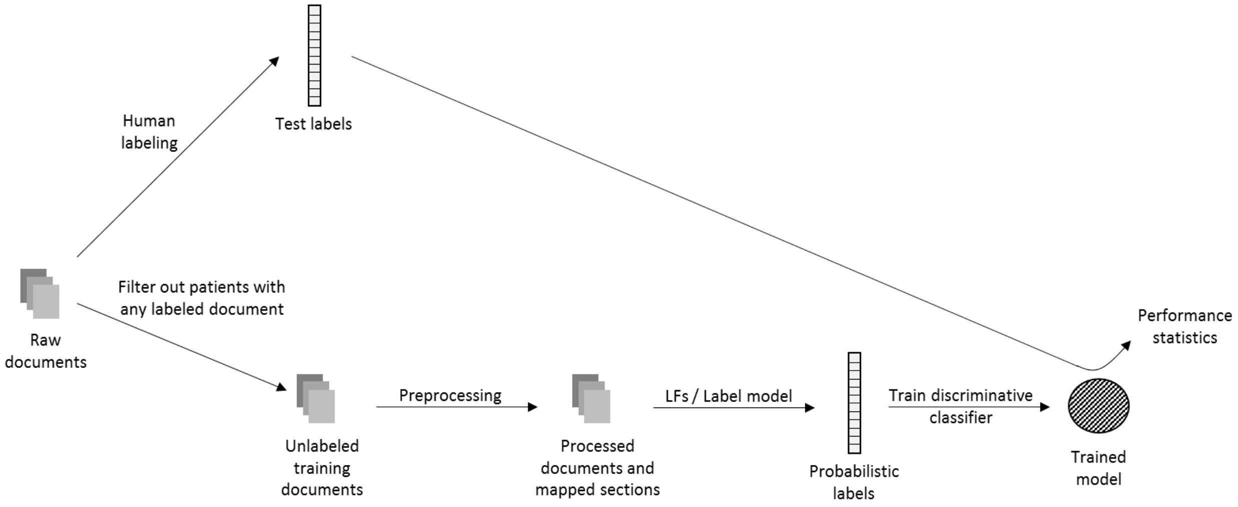
This creates a weight inversely correlated to the term’s frequency in the corpus. The TF-IDF score for term t_i in document d_j is the product

$$\text{TF-IDF} = \mathrm{TF} \cdot \mathrm{IDF}.$$

2.7 Main experiments

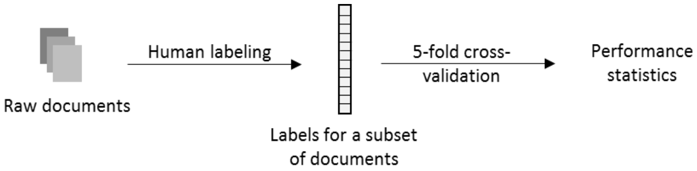
We evaluated a final discriminative classifier trained on the labels generated by the label model. Figure 1 summarizes this weak supervision pipeline. Human-labeled documents were reserved as a test set, and from the unlabeled documents all patients with a labeled document in the test set were then removed. The remaining unlabeled training documents were subjected to preprocessing and labeled by labeling functions. A label model was trained and used to generate probabilistic labels for documents in the training set. Documents on which all LFs abstained were filtered out and the remaining binary class-labeled documents were used to train a final discriminative classifier.

Figure 1. Summary of the weak supervision pipeline.


Human labeling
Test labels
Raw documents
Filter out patients with any labeled document
Unlabeled training documents
Preprocessing
Processed documents and mapped sections
LFs / Label model
Probabilistic labels
Train discriminative classifier
Trained model
Performance statistics

We also evaluated a discriminative classifier trained on human labels (without weak supervision) using stratified 5-fold cross-validation (Figure 2).

Figure 2. Training on human labels.



For each of these two main experiments, we evaluated several machine learning algorithms in scikit-learn [13], as summarized in Table 6. Hyperparameters were optimized by grid search using F1 scores computed via internal 5-fold cross-validation. This was done independently on each training fold for the 5-fold cross-validation experiment. Feature importance was assessed using mean decrease in Gini impurity for random forests and coefficient size for logistic regression.

Table 6. Learning algorithms used for main experiments. When multiple hyperparameters are listed, all were considered in grid search.

Learning algorithm	**Hyperparameters**
Random Forest	Number of trees: 100 Maximum depth: None, 12, 15, 17 Minimum samples to split: 2, 3 Minimum samples for leaf: 5, 7, 9
Logistic Regression	Regression penalty: L1, L2 Cost: 0.1, 1, 10, 100, 1000

2.8 Statistical measures of performance

Table 7 defines performance metrics utilized in this work. For metrics that require selecting a probabilistic threshold for the predicted outcome (such as the F1 score), the value 0.5 was used. The average precision (AP)

$$AP = \sum_i (R_i - R_{i-1}) P_i$$

is a metric that considers all possible thresholds, i.e. R_i and P_i are the recall and precision at the *i*th threshold, respectively. Like the area under the precision-recall curve, AP is a single number that summarizes the entire curve.

Confidence intervals were computed using bootstrapping as in [14,15] and summarized in Table 8 (adapted from [16] with permission from the author). Statistical significance was assessed relative to a significance level of 5%.

In our experiments, we employed an indiscriminately positive classifier as a null model for baseline comparison of precision, recall, F1, and AP metrics. In precision-recall analysis, this is a more appropriate baseline than a random classifier [17].

Table 7. Performance metrics. TP – total true positives; TN – total true negatives; FP – total false positives; FN – total false negatives; R_i and P_i are the recall and precision at the *i*th threshold.

Statistic	**Definition**
Accuracy	(TP + TN) / total
Recall	TP / (TP + FN)
Precision	TP / (TP + FP)
F1 score	2 precision · recall / (precision + recall)
Average precision	$AP = \sum_i (R_i - R_{i-1}) P_i$

Table 8. Determining Confidence Intervals.

Step 1: Cross-Validation (if applicable)	Pool test set predictions across cross-validation folds, if applicable.
Step 2: Bootstrapping	Repeat 2000 times: • Sample with replacement from the pooled predictions to create a bootstrapped set of predicted labels equal in size to the set of pooled predictions. • Calculate the performance metric for the classifier using this bootstrapped set.
Step 3: Confidence Interval Calculation	Determine the 2.5 and 97.5 percentile of the distribution of F1 scores computed in Step 2.

2.9 Error analysis

For selected experiments, we conducted an error analysis to provide a human interpretation of model output. Model predictions were joined with the original test data and provided to a human reviewer as a spreadsheet.

3 Results

Among 999 labeled documents, 823 were unique. There were 143 documents with duplicate labels, of which 139 (97.2%) had no conflicts. Within this group, there were 135 documents with all negative labels and 4 documents with all positive labels.

The tuning set consisted of 427 unique documents, and a set of 478 unique documents were used for generating our main experimental results. Due to random selection of examples for labeling, 82 documents in the latter set were also in the tuning set. Class prevalence was 0.061 and 0.075 in the tuning and test sets, respectively.

When using tuning set labels, there were 14,227 documents (4,956 patients) in the training set of the label model, and 13,319 documents (4,816 patients) in the training set of the final classifier after filtering out documents on which all labeling functions abstained. When using test set labels, there were 14,054 documents (4,903 patients) in the label model training set and 13,163 documents (4,764 patients) in the training set of the final classifier.

3.1 Label model performance and error analysis

In total, 42.3% (202/478) of probabilistic labels were correct, using 0.5 as the decision threshold. The label model achieved recall (97.2%), precision (11.3%), and F1 score (20.2%) comparable to final classifier performance as shown in the next subsection. A manual analysis of label model predictions was performed.

There was one false negative prediction. In this case, the patient was diagnosed with large bowel diarrhea due to suspected dietary indiscretion. Although Addison's disease was technically a possible cause of the illness, and had not been ruled out, the likelihood of Addison's disease was in fact low in this case. The label model assigned a probabilistic label of 0.44, close to the decision threshold.

Among incorrect probabilistic labels, 99.6% (275/276) were false positives, with 50.5% of them were over 0.9. There were some indications that label magnitudes were consistent with clinical judgement at the extremes, i.e. close to 0.5 and very close to 1. Among the 20 documents with false positive labels >0.9999995, there are 4 that mention Addison's disease as a possible cause or a baseline cortisol having been performed, 3 that mention hyperadrenocorticism (for which sometimes ACTH stimulation is performed), and 5 that mention signs of Addison's disease without any definitive rule-out conditions present. About 25% (5/20) of these most confident predictions are cases where reasonable clinicians might differ in how they would assign a binary label. Among the 13 documents with false positive labels <0.537, there are 10 with all signs not consistent with Addison's disease, 2 having GI signs with low risk of Addison's disease (but no identified rule-out), and 1 with GI signs and a positive finding of Giardia. We suspect a different clinician could have assigned a positive binary label in only up to 15.4% (2/13) of these least confident positive predictions.

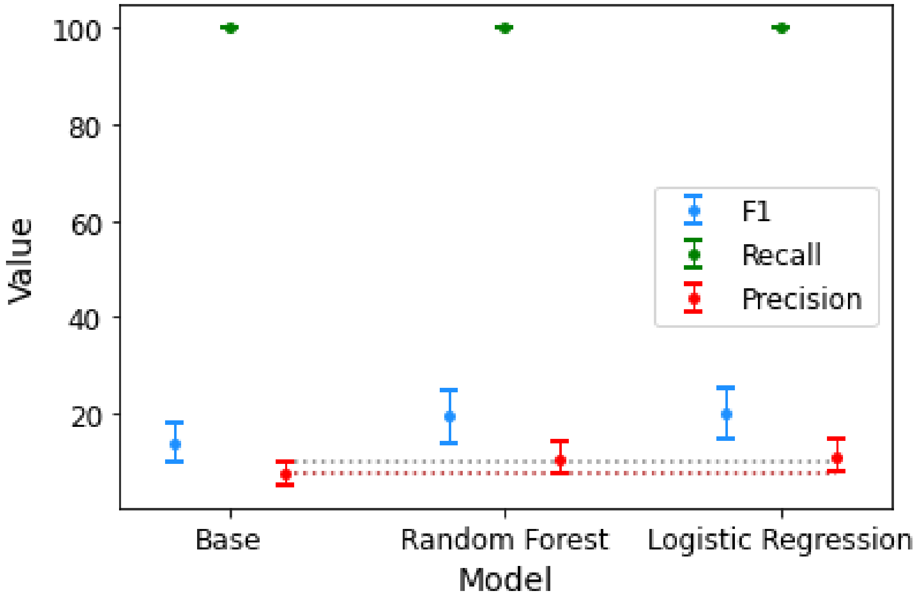
For most false positive labels, however, the label value was not reliably correlated with the confidence of a human reviewer. In particular, the number of rule-out criteria met (as measured by the number of rule-out labelers that matched on a given document) did not correlate strongly with the magnitude of the label value (Spearman's correlation 0.24, $p < 10^{-4}$). A negative correlation would be expected if the label magnitudes were consistent with clinical judgement across the full range of label values.

3.2 Experiment 1

The weak supervision pipeline was evaluated on the test set using random forest and logistic regression to learn a final discriminative classifier on binary class labels derived from the label model. Figure 3 shows the performance of the final ML classifiers relative to a baseline null model that made indiscriminately positive predictions. Both ML classifiers achieved 100% recall and precision higher than the baseline precision of 7.52%, which was equal to the positive class prevalence. Logistic regression, for example, achieved a precision of 11.12%, with a lower 95% confidence interval bound of 7.94%.

Full data is presented in Table S2 in Appendix A. The optimal hyperparameters selected by grid search are detailed in Table S3 in Appendix A.

Figure 3. Final classifier performance in the weak supervision pipeline. F1 scores, recall, and precision for a random forest and logistic regression model trained using weak supervision, compared to an indiscriminately-positive baseline classifier. Horizontal dotted lines are drawn at the value of baseline precision and the upper limit of its confidence interval. Error bars represent 95% confidence intervals estimated by bootstrapping.



Precision-recall curves and feature importance charts are presented in Figures 4 and 5, respectively. The top features appear clinically relevant. For example, terms like “vomiting”, “lethargy”, “diarrhea”, “regurgitation”, “anorexia”, “collapse”, “weight loss”, “weakness”, “shaking”, and “inappetence” describe clinical signs of hypoadrenocorticism. Terms like “leptospirosis injection”, “vaccine given”, “pancreatitis”, “effusion”, and “healthy” may reduce suspicion of disease.

Figure 4. Precision-recall curves for final classifiers in the weak supervision pipeline. PR curves for random forest (left) and logistic regression (right) classifiers.

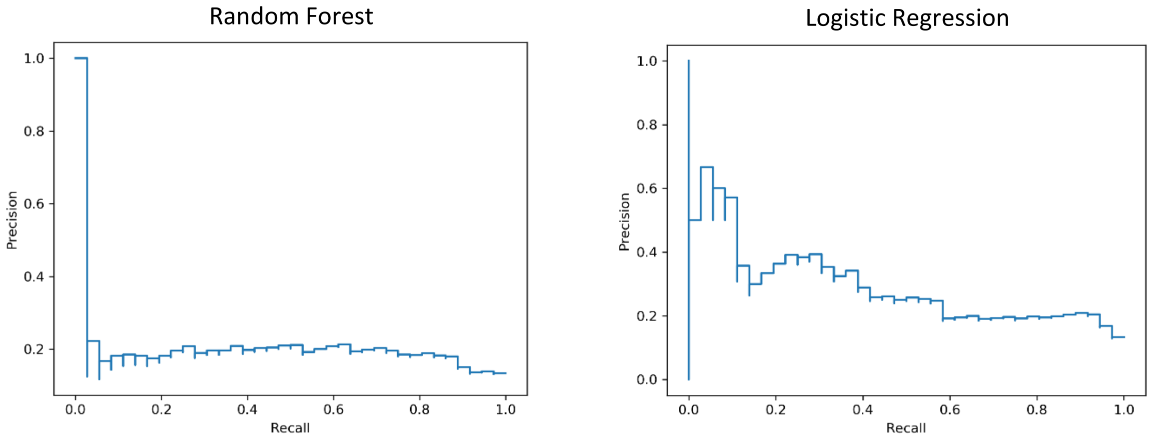
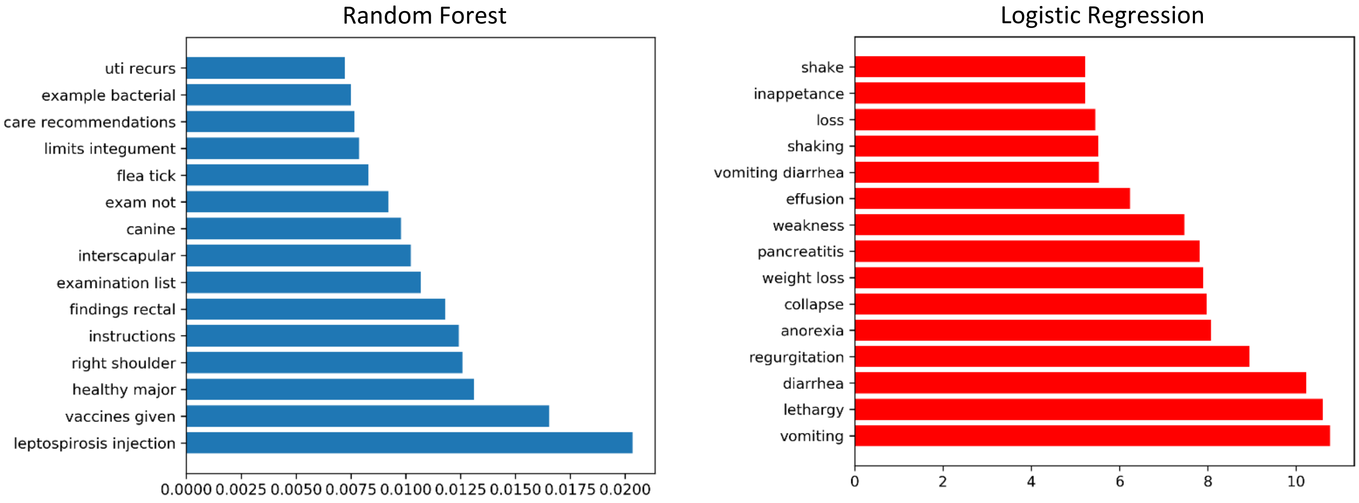


Figure 5. Most important features for final classifiers in the weak supervision pipeline. Feature importance is measured as mean decrease in Gini impurity for random forest (left) and magnitude of regression coefficient in logistic regression (right).



3.3 Experiment 2

We evaluated discriminative classifiers on human labels using stratified 5-fold cross-validation, and compared performance to the final classifiers in the weak supervision pipeline using average precision (Figure 6). Both models trained on human labels perform better than their counterparts trained using weak supervision. In the case of logistic regression, human label training results in an AP of 45.7%, with 32.6% the lower limit of the confidence interval. The AP of a logistic regression model trained on weak supervision labels is 31.3%. This suggests that logistic regression trained on human labels slightly outperforms logistic regression trained using weak supervision.

The full data and optimal hyperparameters for this experiment are presented in Tables S4-S5 and Figure S1 in Appendix A. The precision-recall curves are presented in Figure 7.

Figure 6. Comparison of weak supervision and training on human labels. Random forest and logistic regression models are compared between experiment 1 (training with weak supervision) and experiment 2 (training on human labels). Error bars represent 95% confidence intervals estimated by bootstrapping.

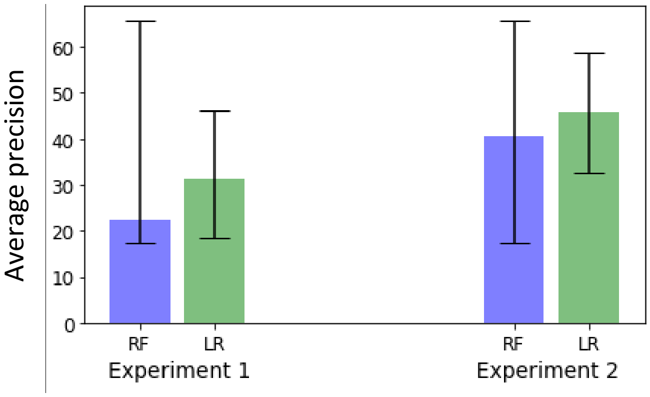
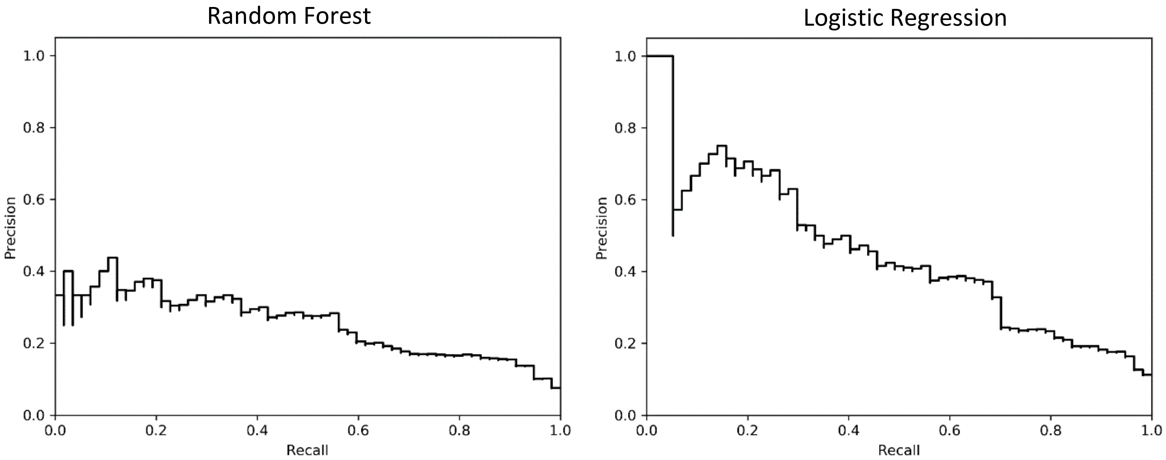


Figure 7. Precision-recall curves. PR curves for random forest (left) and logistic regression (right) classifiers, using pooled predictions across 5-fold cross-validation.



4 Discussion

We explored using weak supervision to train a machine learning classifier to detect suspected cases of canine hypoadrenocorticism in text discharge notes at a veterinary hospital. Our results showed that with weak supervision we can achieve close to 100% recall with precision about 1.48x higher than the positive class prevalence. If this model were used to filter negative predictions out of the corpus, it would therefore concentrate the number of positives by ~1.5x without resulting in a loss of positives. A retrospective clinical study requires document review by human experts and often extensive manual effort to select cases. Given a large, entirely free-text corpus, an automated case selection system with close to 100% recall and non-trivial precision supports this research endeavor. However, this system is not suitable as a clinical decision support tool that would proactively monitor patient information and notify clinicians of suspected cases. This would require high recall and much higher precision than what we have achieved.

Our comparison of discriminative classifiers trained on weak supervision vs. human labels showed that training with human labels provides a performance boost of 1.46x relative to weak supervision. Therefore, weak supervision was effective but not as effective as training on human labels.

However, this comparison is best made when the amount of human effort for both approaches is held constant, and this study did not attempt to quantify human effort. We estimate that designing labeling functions took about 15-20 hours, and labeling took about 20-25 hours. If the amount of human input was different between these two approaches, then this complicates an analysis of their comparative merits.

We will now explore several aspects of our system, focusing on reasons for its low overall precision.

4.1 The task and context

The task is objectively hard. As explained in the introduction, Addison's disease is commonly overlooked by clinicians because it is often misconstrued as something else. In addition, the training signal is a binary label, and in a clinician's mind the suspicion of Addison's disease is not a binary signal. In practice, clinicians often construct (perhaps intuitively or implicitly) a ranked list of possible causes for the patient's illness, and Addison's disease may be higher or lower on that list. Finally, a clinician's level of suspicion depends on a combination of situation-specific factors that may not always be reflected in the discharge text.

4.2 Labeling functions and label model design

Given that label model test performance is comparable to the final classifier, we believe the label model most strongly limits the performance of this system. Therefore, we suspect that the labeling functions (Tables 2-5) are the most culpable components of our pipeline. Indeed, they are quite primitive. For example, one labeling function assigns a positive label to any document containing the word "diarrhea", but this is a crudely inaccurate representation of the target outcome. It would be more accurate to specify that a document containing the word "diarrhea" should be marked as positive only when no rule-out conditions are present. In this example, confidence in a positive prediction should be higher if there are laboratory test results consistent with the disease. The confidence should be even higher if there is post-hoc evidence of the disease having been suspected by the clinician.

This logic was explored using the composition functions outlined in Appendix C. This approach dramatically reduced the quality of the label model, likely because of the introduction of additional dependencies among labelers. The solution to this problem may be to specify a dependency graph among labelers [11], rather than to allow the algorithm to learn the graph from scratch. Unfortunately, we found near the end of our project that the current version of Snorkel does not yet include this feature [18]. Therefore, we conclude our problem would likely benefit from composite labeling functions built using the primitives we studied in this paper, perhaps as outlined in Appendix C, but we were unable to test this hypothesis.

With respect to the labeling functions presented in Tables 2-5, adding additional functions reduced performance of our label model. In the suggested composition framework, however, the system may benefit from additional labelers. We imagine that additional labelers could focus on a more robust collection of rule-out conditions and expanded recognition of laboratory test results. For example, there are several existing systems for extracting numeric laboratory test results from free text [19,20]. This technology might have utility in a veterinary context, although accurate interpretation of numeric results would necessarily depend on using canine reference ranges. This direction seems intriguing, but it remains unclear if it would enhance the quality of the generated labels because Addisonian patients do not always exhibit the same patterns of laboratory abnormalities, and some of the textbook anomalies are known to occur in only a small number of patients with this disease [4].

4.3 Effects of class skew

Based on human labeling, we estimate that the class prevalence in our data set is about 7%, which represents significant class skew. Class skew may be related to the observed poor precision, and we hypothesize that it would have the strongest impact on the label model.

Recall that documents on which the label model abstained (i.e. all labelers abstained) were removed from the training set of the final classifier. If we accept the generated probabilistic labels as correct and a decision threshold of 0.5, this had the effect of setting the class prevalence in the final training set to 0.67. However, we have shown that most label model errors are false positives, and over 50% of false positive labels were over 0.9. Furthermore, our error analysis showed that false positive label noise does not appear to carry a predictive signal, except at the extremes of values very close to 0.5 or 1. Therefore, the true class prevalence in the final training set is likely less than 0.67, and it is difficult to know its true value. The imprecision introduced by the label model would obscure any attempt to enforce a balanced class distribution with adaptive sampling.

4.4 Difficulties in human label acquisition

Another weakness of our approach is that human labels were assigned by the authors. Although one of the authors was a veterinarian, it would have been more ideal to get expert labels from a clinician who regularly works with these cases, such as a veterinary internal medicine specialist. We expect that if the accuracy of human-assigned labels were to have suffered, it would likely be caused by more false negative labels than false positives, a result consistent with mistakes in clinical practice. If the human labels (our "ground truth" labels) were to contain false negatives, then this would have several effects: (1) lower observed class prevalence in the test set, (2) reduced validity of test set statistics, and (3) reduced performance of the models trained on human labels. Without additional validation of the human labels, we are unable to determine if these consequences have impacted our results.

Our difficulty in recruiting an expert labeler exemplifies the high cost of human label acquisition. Weak supervision helps to reduce this demand for expert input, and we hope that it will continue to stimulate ongoing research in this domain. Another advantage of weak supervision is that the style of expert input required for data programming may appear more appealing to human experts than the type of repetitive work necessary to label individual documents.

4.5 Future directions

We briefly turn to a discussion of enhancements and future directions to extend this work. Aside from implementing the composition framework described in Appendix C, another valuable consideration would be alternative featurization schemes. Featurizing documents as vectors in an embedding space using an approach like doc2vec [21] may provide a better representation than TF-IDF vectors.

It is also a common approach in weak supervision to train final classifiers using probabilistic labels rather than class labels, i.e. using models that can be trained with noise-aware loss. Given that label noise in our experiments did not appear to be meaningful, we considered this to be lower priority in the first version of our system. In the future, if the quality of the label model output were to be improved, then it would become important to consider final classifiers that can extract potential signals from the label noise. Preliminary results (not presented in this paper) suggested that training a logistic regression classifier with noise aware loss did not change the precision of the system, given the current label model limitations.

In addition to weak supervision, another strategy for making use of a limited expert labeling budget is active learning. This involves using the current model to select unlabeled data instances that will be labeled by a human expert and then incorporated into the training set for training an updated model. The simplest approach is uncertainty-based sampling, in which selection of data is governed by the current model's prediction uncertainty [22]. This approach has been shown to improve model performance in a variety of biomedical prediction tasks [23–26]. However, active learning may introduce sampling bias that degrades performance in some learning tasks [27]. In our system, it would be possible to use the probabilistic output of the label model to direct sampling for labeling by a human oracle, by selecting documents with a label close to 0.5. Treating this oracle as a strongly-weighted supervision source may increase overall accuracy of the label model output. This approach depends, of course, on having a sufficiently high-quality label model output to direct active sampling.

5 Conclusion

Information extraction from unstructured medical text is especially important in veterinary medicine, given the reduced emphasis on structured coding in some veterinary EHR systems (relative to human EHRs). Machine learning-based text classification models can help impute categorical structure, but they present the challenge of acquiring enough training data from human experts. In this work, we focused on recognizing suspected canine hypoadrenocorticism from medical discharge text, and we evaluated weak supervision methods in this domain. Since the uncommon and elusive nature of this disease makes it an important subject for retrospective clinical studies, we proposed using this text classification system as a case selection agent, and we showed that our tool would provide some utility in this capacity. We discussed factors that limit precision and proposed ideas for further work that will advance information extraction in veterinary clinical informatics.

6 Acknowledgements

Special thanks to Dr. Lauren Trepanier and Dr. Elizabeth Alvarez at the University of Wisconsin School of Veterinary Medicine (SVM) for their help in motivating this project idea, and the support of hospital director Dr. Ruthanne Chun for coordinating our request for medical records. Many thanks to Theodoros Rekatsinas and Mark Craven for technical input and support as well, and to Rebecca White at the SVM for extracting the data.

References

1. Van Lanen K, Sande A. Canine hypoadrenocorticism: pathogenesis, diagnosis, and treatment. Top Companion Anim Med. 2014;29: 88–95. doi:10.1053/j.tcam.2014.10.001

2. Decôme M, Blais M-C. Prevalence and clinical features of hypoadrenocorticism in Great Pyrenees dogs in a referred population: 11 cases. Can Vet J. 2017;58: 1093–1099.

3. Mitchell AL, Macarthur KDR, Gan EH, Baggott LE, Wolff ASB, Skinningsrud B, et al. Association of Autoimmune Addison's Disease with Alleles of STAT4 and GATA3 in European Cohorts. PLOS ONE. 2014;9: e88991. doi:10.1371/journal.pone.0088991

4. Klein SC, Peterson ME. Canine hypoadrenocorticism: Part I. Can Vet J. 2010;51: 63–69.

5. Klein SC, Peterson ME. Canine hypoadrenocorticism: part II. Can Vet J Rev Veterinaire Can. 2010;51: 179–184.

6. Kruse CS, Goswamy R, Raval YJ, Marawi S. Challenges and Opportunities of Big Data in Health Care: A Systematic Review. JMIR Med Inform. 2016;4: e38. doi:10.2196/medinform.5359

7. Reagan KL, Reagan BA, Gilor C. Machine learning algorithm as a diagnostic tool for hypoadrenocorticism in dogs. Domest Anim Endocrinol. 2020;72: 106396. doi:10.1016/j.domaniend.2019.106396

8. Nie A, Zehnder A, Page RL, Zhang Y, Pineda AL, Rivas MA, et al. DeepTag: inferring diagnoses from veterinary clinical notes. Npj Digit Med. 2018;1: 60. doi:10.1038/s41746-018-0067-8

9. Zhang Y, Nie A, Zehnder A, Page RL, Zou J. VetTag: improving automated veterinary diagnosis coding via large-scale language modeling. Npj Digit Med. 2019;2: 1–8. doi:10.1038/s41746-019-0113-1

10. Ratner A, De Sa C, Wu S, Selsam D, Ré C. Data Programming: Creating Large Training Sets, Quickly. ArXiv160507723 Cs Stat. 2017 [cited 10 Apr 2020]. Available: http://arxiv.org/abs/1605.07723

11. Ratner A, Hancock B, Dunnmon J, Sala F, Pandey S, Ré C. Training Complex Models with Multi-Task Weak Supervision. ArXiv181002840 Cs Stat. 2018 [cited 10 Apr 2020]. Available: http://arxiv.org/abs/1810.02840

12. Ratner A, Bach SH, Ehrenberg H, Fries J, Wu S, Ré C. Snorkel: Rapid Training Data Creation with Weak Supervision. Proc VLDB Endow. 2017;11: 269–282. doi:10.14778/3157794.3157797

13. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12: 2825–2830.

14. Gao S, Young MT, Qiu JX, Yoon H-J, Christian JB, Fearn PA, et al. Hierarchical attention networks for information extraction from cancer pathology reports. J Am Med Inform Assoc. 2018;25: 321–330. doi:10.1093/jamia/ocx131

15. DiCiccio TJ, Efron B. Bootstrap Confidence Intervals. Stat Sci. 1996;11: 189–212.

16. Bollig N, Clarke L, Elsmo E, Craven M. Machine learning for syndromic surveillance using veterinary necropsy reports. PLOS ONE. 2020;15: e0228105. doi:10.1371/journal.pone.0228105

17. Flach P, Kull M. Precision-Recall-Gain Curves: PR Analysis Done Right. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R, editors. Advances in Neural Information Processing Systems 28. Curran Associates, Inc.; 2015. pp. 838–846. Available: http://papers.nips.cc/paper/5867-precision-recall-gain-curves-pr-analysis-done-right.pdf

18. Documentation unclear about LabelModel strategy · Issue #1462 · snorkel-team/snorkel. In: GitHub [Internet]. [cited 17 Apr 2020]. Available: https://github.com/snorkel-team/snorkel/issues/1462

19. Hao T, Liu H, Weng C. Valx: A system for extracting and structuring numeric lab test comparison statements from text. Methods Inf Med. 2016;55: 266–275. doi:10.3414/ME15-01-0112

20. Liu S, Wang L, Ihrke D, Chaudhary V, Tao C, Weng C, et al. Correlating Lab Test Results in Clinical Notes with Structured Lab Data: A Case Study in HbA1c and Glucose. AMIA Summits Transl Sci Proc. 2017;2017: 221–228.

21. Le QV, Mikolov T. Distributed Representations of Sentences and Documents. ArXiv14054053 Cs. 2014 [cited 24 Apr 2020]. Available: http://arxiv.org/abs/1405.4053

22. Settles B. Active Learning. Synth Lect Artif Intell Mach Learn. 2012;6: 1–114. doi:10.2200/S00429ED1V01Y201207AIM018

23. Chen Y, Lasko TA, Mei Q, Denny JC, Xu H. A study of active learning methods for named entity recognition in clinical text. J Biomed Inform. 2015;58: 11–18. doi:10.1016/j.jbi.2015.09.010

24. Fong A, Howe JL, Adams KT, Ratwani RM. Using Active Learning to Identify Health Information Technology Related Patient Safety Events. Appl Clin Inform. 2017;8: 35–46. doi:10.4338/ACI-2016-09-CR-0148

25. Kholghi M, De Vine L, Sitbon L, Zuccon G, Nguyen A. Clinical information extraction using small data: An active learning approach based on sequence representations and word embeddings. J Assoc Inf Sci Technol. 2017;68: 2543–2556. doi:10.1002/asi.23936

26. Zhang H-T, Huang M-L, Zhu X-Y. A Unified Active Learning Framework for Biomedical Relation Extraction. J Comput Sci Technol. 2012;27: 1302–1313. doi:http://dx.doi.org.ezproxy.library.wisc.edu/10.1007/s11390-012-1306-0

27. Varghese A, Hong T, Hunter C, Agyeman-Badu G, Cawley M. Active learning in automated text classification: a case study exploring bias in predicted model performance metrics. 2019. doi:10.1007/s10669-019-09717-3

Appendix A. Supplemental Data

Table S1. Keywords used to map section titles to sections.

Standard Section	Keywords in HTML section title
History	history, concern
Physical Exam	physical
Diagnoses	suspect, diagnosis, diagnose, problem list, problem
Diagnostic Tests	test, diagnostic, procedure
Treatments	treatment, summary, to go home, comment, recommendation, instructions, plan, follow, medication

Table S2. Optimal hyperparameters selected for final classifiers in experiment 1 (weak supervision).

Learning algorithm	Hyperparameters
Random Forest	Maximum depth: None Minimum samples to split: 2 Minimum samples for leaf: 5
Logistic Regression	Regression penalty: L2 Cost: 10

Table S3. Test set performance for final classifiers in experiment 1 (weak supervision).

	Baseline	Random Forest	Logistic Regression
F1	14.00% (9.94%, 17.91%)	19.43% (14.05%, 25.00%)	19.89% (14.68%, 25.56%)
Recall	100.0% (100.0%, 100.0%)	100.0% (100.0%, 100.0%)	100.0% (100.0%, 100.0%)
Precision	7.52% (5.23%, 9.83%)	10.68% (7.60%, 14.08%)	11.12% (7.94%, 14.58%)
Accuracy	7.58% (5.23%, 9.83%)	37.29% (33.26%, 41.63%)	39.83% (35.77%, 44.14%)
Average Precision	N/A	22.27% (13.33%, 32.44%)	31.33% (18.59%, 46.02%)

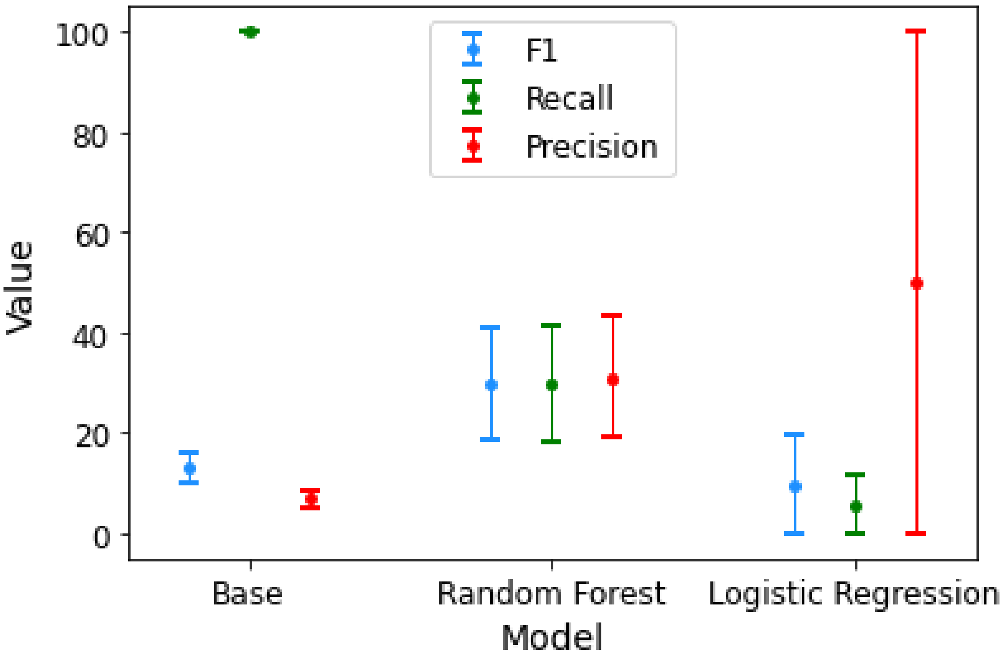
Table S4. Optimal hyperparameters selected in experiment 2. Values are presented for each of 5 folds.

Learning algorithm	Hyperparameters (on each fold)
Random Forest	Maximum depth: (12,12, None, None, None) Minimum samples to split: (2, 2, 2, 2, 2) Minimum samples for leaf: (9, 9, 9, 9, 9)
Logistic Regression	Regression penalty: (L2, L2, L2, L2, L2) Cost: (1000, 1000, 1000, 1000, 100)

Table S5. Performance of aggregate test predictions in experiment 2. Note that the prediction threshold was not adjusted to favor maximal recall.

	Baseline	Random Forest	Logistic Regression
F1	12.98% (9.95%, 15.92%)	29.96% (18.46%, 40.91%)	9.342% (0.000%, 19.725%)
Recall	100.0% (100.0%, 100.0%)	29.84% (18.18%, 41.67%)	5.29% (0.00%, 11.67%)
Precision	6.97% (5.24%, 8.77%)	30.74% (19.14%, 43.34%)	50.06% (0.00%, 100.0%)
Accuracy	6.94% (5.36%, 8.77%)	90.51% (88.43%, 92.45%)	93.06% (91.23%, 94.76%)
Average Precision	N/A	26.96% (18.09%, 37.74%)	45.70% (32.63%, 58.63%)

Figure S1. Summary of model performance in experiment 2. F1 scores, recall, and precision for a random forest and logistic regression model trained using weak supervision, compared to an indiscriminately-positive baseline classifier. Error bars represent 95% confidence intervals estimated by bootstrapping. Note that the prediction threshold was not adjusted to favor maximal recall.



Appendix B. Instructions for manual labeling

The following articles provide the background for the labeling criteria summarized below.

- S. C. Klein and M. E. Peterson, “Canine hypoadrenocorticism: Part I,” Can Vet J, vol. 51, no. 1, pp. 63–69, Jan. 2010. ("Clinical signs and physical examination findings" and "Laboratory abnormalities")
- K. Van Lanen and A. Sande, “Canine hypoadrenocorticism: pathogenesis, diagnosis, and treatment,” Top Companion Anim Med, vol. 29, no. 4, pp. 88–95, Dec. 2014, doi: 10.1053/j.tcam.2014.10.001. ("Clinical findings" and "Laboratory findings")

The easiest way to confirm a positive label is the presence of post-hoc evidence of clinical suspicion of Addison's Disease. This would be one of:

- Patient has a diagnosis or problem of "hypoadrenocorticism" or "Addison's Disease" mentioned in the note. A positive label should be assigned even if the note indicates suspicion without a definitive diagnosis.
- Patient was given a "baseline cortisol" or "ACTH stimulation" ("ACTH stim") test, regardless of its outcome.
- If the patient has been, is, or will be prescribed one of the following drugs, they likely have a diagnosis of Addison's Disease:
    - Desoxycorticosterone pivalate (DOCP)
    - Fludrocortisone acetate (florinef)

In other cases, a clinician should be suspecting the disease if the following conditions are met:

- Patient has signs of GI or nonspecific disease, or hypovolemic shock, indicated by any of the following:
    - Inappetance
    - Anorexia
    - Vomiting or regurgitation
    - Diarrhea
    - Melana and hematochezia
    - Abdominal pain
    - Weight loss
    - Lethargy
    - Depression
    - Weakness
    - Shaking
    - Hair loss
    - Hypovolemic shock: bradycardia or tachycardia, collapse, hypothermia, weak pulses, or poor capillary refill time
- The patient has not been diagnosed with one of these diseases:
    - Kidney failure (acute kidney injury, AKI, acute kidney disease, AKD)
    - GI parasites (worms)
    - Primary liver disease (not just evidence of liver damage on chemistry panel)
    - Acute pancreatitis
- The patient does **not** have a history of toxin ingested noted.
- In most but not all cases, the patient will have some abnormal results on a screening chemistry panel or CBC. If the other conditions are met but none of the following are present (or the

patient has an opposing or contradictory result to these), this warrants further review. These are loosely in order from more common to less common findings.

- Low sodium (Na), i.e. hyponatremia
- High potassium (K), i.e. hyperkalemia
- Low chloride (Cl), i.e. hypochloremia
- Azotemia: increased BUN and/or creatinine
- Mild to moderate normocytic, normochromic, nonregenerative anemia (PCV 20-35%)
- High calcium, i.e. hypercalcemia
- Absence of a stress leukogram in a sick animal (normal CBC is considered abnormal here)
- High lymphocytes, i.e. lymphocytosis
- Mild acidosis indicated by low bicarbonate (HCO3) or low pH
- Low blood glucose, i.e. hypoglycemia
- Low albumin, i.e. hypoalbuminemia
- Low cholesterol, i.e. hypocholesterolemia

Appendix C. Proposed labeling function composition framework and further enhancements

Tables 2-5 in the paper describe the labeling functions that we used in our analysis. As explained in the Discussion, these labeling functions are quite crude, and composition seemed necessary to develop functions that are more consistent with clinical logic. This section will outline the enhanced and composite functions built from the labelers in Tables 2-5. Please see the Discussion for full explanation of why these enhancements were not used in the final analysis.

Composition

Define each labeling function as $f: D \rightarrow \{0, 1, -1\}$, where D represents raw document space, 1 represents a positive label, 0 a negative label, and -1 abstention.

The primitive labelers were grouped into four categories:

1. Clinical signs of hypoadrenocorticism
2. Post-hoc evidence of hypoadrenocorticism
3. Laboratory findings consistent with hypoadrenocorticism
4. Conditions that rule-out hypoadrenocorticism

Let C_1, C_2, C_3, and C_4 represent the disjoint sets of labeling functions corresponding to the above 4 categories.

Define the following selection function

$$s_{\text{rule\_out}}(x) = \begin{cases} 1 \text{ if } f_{\text{prim}}(x) = 0 \text{ for any } f_{\text{prim}} \in C_4 \\ 0 \text{ otherwise} \end{cases}$$

which indicates 1 iff the input x satisfies any rule-out condition. Similarly let

$$s_{\text{any\_sign}}(x) = \begin{cases} 1 \text{ if } f_{\text{prim}}(x) = 1 \text{ for any } f_{\text{prim}} \in C_1 \\ 0 \text{ otherwise} \end{cases}$$

which indicates 1 iff the input x presents any clinical sign of hypoadrenocorticism. Finally let

$$s_{\text{any\_lab}}(x) = \begin{cases} 1 \text{ if } f_{\text{prim}}(x) = 1 \text{ for any } f_{\text{prim}} \in C_3 \\ 0 \text{ otherwise} \end{cases}.$$

Primitive labelers in C_2 and C_4 adequately reflect clinical logic on their own. However, the other two categories require composition of primitive labelers. For each function $f_{\text{prim}} \in C_1$, we built a function f such that

$$f(x) = \begin{cases} 1 \text{ if } f_{\text{prim}} = 1 \text{ and } s_{\text{rule\_out}} = 0 \\ 0 \text{ otherwise} \end{cases}$$

and a function g such that

$$g(x) = \begin{cases} 0 \text{ if } f(x) = 0 \\ 1 \text{ if } f(x) = 1 \text{ and } s_{\text{any\_lab}} = 1. \\ -1 \text{ otherwise} \end{cases}$$

There is one f and g for each primitive labeler in C_1. Note that each f encodes the following clinical reasoning: "suspect disease if a clinical sign is present and no rule-out conditions are present". The function g encodes: "suspect disease if a clinical sign is present and no rule-out conditions are present and at least one laboratory rest result is present". For each $f_{\text{prim}} \in C_1$, we propose including the corresponding f and g (and not f_{prim}) in the final set of labelers. The labeler dependency graph would need to reflect that each g can only be 1 when the corresponding f is one. All f and g are zero when a rule-out condition is present.

We propose similar logic for the third category. For each function $f_{\text{prim}} \in C_3$, we built a function f such that

$$f(x) = \begin{cases} 1 \text{ if } f_{\text{prim}} = 1 \text{ and } s_{\text{rule\_out}} = 0 \\ 0 \text{ otherwise} \end{cases}$$

and a function g such that

$$g(x) = \begin{cases} 0 \text{ if } f(x) = 0 \\ 1 \text{ if } f(x) = 1 \text{ and } s_{\text{any\_sign}} = 1. \\ -1 \text{ otherwise} \end{cases}$$

Now these f encode the following clinical reasoning: "suspect disease if a laboratory result is present and no rule-out conditions are present". The function g encodes: "suspect disease if a laboratory result is present and no rule-out conditions are present and at least one clinical sign is present". As above, for each $f_{\text{prim}} \in C_3$ we propose including the corresponding f and g (and not f_{prim}) in the final set of labelers. The labeler dependency graph would be similarly affected.

Other labeler enhancements

We also considered two additional enhancements.

- Labelers that respond with a label when a keyword is absent from a list of keywords.
- Labelers that abstain when they match on a keyword near a word that indicates negation, such as "not", "no", "absent", or "none".

Although our results indicated these enhancements did not improve the accuracy of label model output, it is possible that a successful implementation of the proposed composition framework (with a correctly specified labeler dependency graph) would benefit from these enhancements.