Extracting Plausible Explanations of Anomalous Data

Will Bridewell and **Bruce G. Buchanan**
Department of Computer Science
University of Pittsburgh
Pittsburgh, PA 15260

Abstract

We present a perspective on theory revision that characterizes the resulting revisions as explanations of anomalous data (i.e., data that contradict a given model). Additionally, we emphasize the plausibility of these explanations as opposed to the performance of a revised model. An explanation generator implementing (part of) John Stuart Mill's Method of Induction was constructed that divides the available data into meaningful subsets to better resolve the anomalies. A domain expert judged the plausibility of the resulting explanations. We found that using relevant subsets of data can provide plausible explanations not generated when using all the data and that identifying plausible explanations can help select among equally possible revisions.

Introduction

Anomalies in science are puzzles that need to be explained. They may even be considered to be the main driving force initiating scientific inquiry. In his influential work on the philosophy of science, Thomas Kuhn (1970) wrote, "Discovery commences with the awareness of anomaly" and "closes only when the paradigm theory has been adjusted" to explain the anomaly. Darden (1991) gives a thorough analysis of the process of resolving anomalies, appealing to detailed historical descriptions. Additionally, she has, in later work, pursued a refined system for theory revision (Darden & Cook 1994; Darden 1998). Like Darden, we subscribe to Kuhn's prominent positioning of anomalies, and emphasize them in the present work.

We present an approach to anomaly resolution, a component of theory revision, that emphasizes the fruitfulness of simple hypothesis generators when applied to meaningful subsets of the data. For our purposes, we define an anomaly as a counterexample to a given model. This definition is meant to include counterexamples that result both from experimental error and from genuine mechanisms unaddressed by the model. Our goal is to resolve the discrepancies between the model and these observations that are inconsistent with that model, which are sometimes called failing negatives (Mooney & Ourston 1994). Each proposed revision to the model serves as a possible explanation of an anomaly. These explanations are created using a union of the languages that define the model and describe the observations.

Leake (1992) wrote that although "many explanations can be generated for any event . . . only some of them are plausible." However, much of the prior work in theory revision emphasizes improved theory performance regardless of a revision's plausibility (Mooney & Ourston 1994; Asker 1994; Richards & Mooney 1995; Carbonara & Sleeman 1999; Esposito *et al.* 2000). Leake's claim establishes plausible hypotheses as a subset of possible hypotheses, therefore we chose as our measure of success the plausibility of the generated explanations. In the present work, each explanation's plausibility was judged by a human expert since the true value may not be established (if at all) until much later.

Although plausibility is a more restrictive measure, we are able to avoid two major burdens that are assumed when the success of the revision process is measured solely by performance. These burdens include the verification of the genuineness of the anomaly and the selection of the best revision from numerous possibilities. The former problem is to differentiate between valid anomalies and those that may have resulted from faulty research, measurement variability, observation error, or corrupted data. In such cases, we may choose to reject the anomaly as opposed to resolving it. The latter burden arises when there are several revisions consistent with both the model and the anomaly. In the past, selection criteria have involved arbitrary selection (Mooney & Ourston 1994; Asker 1994; Esposito *et al.* 2000) and validation (Carbonara & Sleeman 1999). Systems employing arbitrary selection may use either limited generators or filtering criteria in to identify some minimal set of revisions. To avoid assuming these burdens, we recommend that theory revision be broken into three distinct steps: explanation generation, anomaly verification, and revision selection.

After an anomaly is noticed, it may be dealt with in several ways. Chin and Brewer identified eight possible reactions to anomalous data by surveying the history of science (Brewer & Chinn 1994) and running controlled psychological experiments (Chinn & Brewer 1998). Of these reactions, four involve explanation of the data: rejecting the data, reinterpreting the data (explaining away), altering the periphery of the theory (theory revision), and generating a new theory. The other four responses—ignoring the data, agreeing to explain the data at a later date, maintaining uncertainty, and excluding the data from the domain—do not lead to discovery

or learning and do not require verification of the anomaly. We take the approach of first generating explanations and then determining which of the first three responses is appropriate. (We are not attempting to generate entirely new theories.) Thus we are addressing a critical step in scientific inquiry after noticing exceptions to a model and before revising the model. Our work can also be seen as bridging a gap between data mining and model building.

Background

Our process for generating explanations of anomalies involves applying modifications of old inductive principles. In his classic work on the scientific method, John Stuart Mill (1900) identified four methods of induction that have been rediscovered and recast numerous times[1]. Our work concerns the first two methods: the *method of agreement*

> "If two or more instances of the phenomenon under investigation have only one circumstance in common, the circumstance in which alone all the instances agree is the cause (or effect) of the given phenomenon"

and the *method of difference*

> "If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance in common save one, that one occurring only in the former; the circumstance in which alone the two instances differ is the effect, or the cause, or an indispensable part of the cause, of the phenomenon."

With the method of agreement, we ask how each anomaly is similar to identified subsets of other anomalies, while with the method of difference we ask how each anomaly differs from various subsets of nonanomalous data[2]. By casting our generators in the context of these basic questions and by identifying meaningful subsets of our working set of observations we conjecture that plausible explanations will be created.

As mentioned, we have begun to evaluate our system based on the perceived plausibility of the explanations. We take this approach to avoid passing final judgment on the quality and use of an explanation. As the anomaly under consideration may arise due to factors unrelated to our current model of the data (e.g., experimental error), we would rather maintain caution. Performing revisions based solely on an estimated performance gain can introduce arbitrary features into the model due to sample bias or other circumstantial causes. Plausibility may give us a means to intelligently select among revisions to avoid such problems.

For our evaluation, we chose the task of identifying patients with lower respiratory syndrome (LRS) from data on patients visiting hospital emergency rooms. In the event of a bioterrorist attack, the trend in emergency room visits is expected to change, and an accurate model for key ailments is required to identify such variation (Wong *et al.* 2002). In cases of widespread anthrax or SARS exposure, we must be able to accurately and effectively detect cases of LRS. To this end, a highly sensitive model with low specificity can be made into a more specific model by resolving the failing negatives (i.e., anomalies).

1. COUGH is PRESENT implies LRS is PRESENT.
2. WHEEZING is PRESENT implies LRS is PRESENT.
3. SPUTUM is PRESENT implies LRS is PRESENT.
4. A POSITIVE X-RAY for PNEUMONIA implies LRS is PRESENT.
5. DYSPNEA is PRESENT implies LRS is PRESENT.

Table 1: Overly general model of lower respiratory syndrome (LRS) used as the basis for anomaly resolution.

To evaluate our system, an expert in the field of infectious diseases established a baseline model, shown in Table 1, for the identification of LRS in emergency room patients. Anomalies to this model were analyzed and explained in the context of a selected data set. The resulting explanations were judged as plausible or implausible by the expert. Our findings served to test our hypothesis that a significant number of plausible explanations will result from examining meaningful subsets of the data and that plausibility is a useful measure for selecting revisions.

Method

We built an explanation generator to test our hypothesis that takes both data and a model. The data are previously classified feature-vectors and may contain both noisy records and missing values. The model consists of a disjunctive set of single-step classifiers or IF-THEN rules (see Tables 1 and 4 for examples). The antecedent of each classifier is a conjunction of features (attribute-value pairs joined by a relation), where the values may be either internal nodes or base values in a general-to-specific value-hierarchy. Continuous attributes are assumed to have been discretized into intervals or points during generation of the original model. The consequent of a classifier is a feature such that the value is some target class. Our conflict resolution mechanism gives priority to direct specializations (i.e., exception rules) over their more general base.

The explanation generator begins by identifying data that are incorrectly classified by the model or for which the conflict resolution mechanism fails (i.e., two or more opposing classifications are made). These exceptions to the model are termed anomalies and are judged to be in need of explanation. Thus the anomalies are false positives to a classifier in the model for which no matching exception rule exists. Within this representation, the process of anomaly resolution consists of identifying exception rules that account for each anomaly.

As an example, consider Rule 1 in Table 1: the presence of a cough implies LRS. Also, suppose that we have an ex-

[1] While we recognize that Mill's methods are overly simplistic, we believe that they provide a useful starting point for any work in explanation.

[2] We coin the distinction of *nonanomalous* (versus anomalous) data to refer to data that are wholly consistent with a specified model.

ception rule for that classifier that reads, "the presence of a cough and the presence of pharyngitis implies that there is no LRS." In a case where LRS is absent but cough is present, there are two possibilities. One possibility is that the particular patient also has pharyngitis and therefore meets the qualifications of the exception rule. Such a case would not be anomalous. However, if the patient did not have pharyngitis, since we list no other exception rules to Rule 1, the patient would be anomalous.

Our first generator for exception rules (i.e., explanations) is based on the method of agreement. Strict application of this method requires a researcher to either record all features of the instances under investigation or ensure that the subset of recorded features includes all features causally affecting the phenomenon under study. Unfortunately, meeting either of these requirements verges on impossibility. The former option requires complete knowledge of the universe. In the case of the latter option an observer can never be certain, barring the recording of all features (whatever that may mean), that she has identified all features causally contributing to the phenomenon under study. Therefore, to make progress in the process of induction, we assume that the researcher records those features constituting the known and suspected causes of the studied phenomenon (or more accurately, that the researcher has selected the features that have a high prior probability of influencing the outcome of her experiment). With this assumption in mind, we set "anomalous" as the phenomenon and ask the questions in Table 2.

A1 How is this anomaly similar to other anomalous data?

A2 How is this anomaly similar to other anomalies with the same classification?

A3 How is this anomaly similar to other anomalies sharing the same faulty classifier?

A4 How is this anomaly similar to other anomalies with the same classification that share the same faulty classifier?

Table 2: Questions to be answered using the Method of Agreement

Each of the four questions presented in Table 2 requires the existence of multiple anomalies, and in all but the first case, the anomalies under comparison must also have aspects in addition to anomalousness in common. Question A1 mirrors the standard question, "How is this new observation similar to prior observations of the same class?" Here, the class under consideration is "anomaly." Question A2 limits the domain of the first question to those anomalies that are of the same actual class as the anomaly under question (e.g., a group that consists of all anomalies that are observed to be negative for LRS). Question A3 implies that a particular classifier may be at fault (by being too general) for the existence of all the anomalies. That is, we might only be interested in those anomalies where the primary classifier requires wheezing to be present. Finally, Question A4 significantly restricts the observations we are willing to consider to the intersection of the data satisfying both Questions A2 and A3. Thus we would want only those examples classified by Rule 2 (wheezing is present implies LRS is present) that are also negative for LRS.

```
Method-of-Agreement(anomalies):
  pool = shared-features(anomalies)
  explanations = [ ]
  for each a in anomalies
    /* consider each rule that incorrectly classifies the
       anomaly */
    for each i in (incorrect-classifiers(a))
      e = new explanation(antecedent(i), class(a))
      push(necessary-features(pool, anomalies),
           antecedent(e))
      while (overly-general(e) and features-left(pool, e))
        /* add a feature that best separates the anomaly from
           nonanomalous data still covered by the current
           explanation*/
        push(sufficient-feature(pool – antecedent(e),
                                anomalies)
             antecedent(e))
      push(e, explanations)
  return explanations
```

Figure 1: Algorithm for the method of agreement generator

To answer these questions, we define agreement or similarity as a function of the features shared among the anomalies[3]. Therefore, the agreement generator presented in Figure 1 begins by identifying the pool of features common among the current group of anomalies, which may be any of the groups defined by Table 2. The generator next explores each incorrect-classifier for each anomaly, creating a root explanation from the antecedent of that classifier and the observed class of the current anomaly. Since the root explanation shares an antecedent with a classifier from the model, there will likely be anomalies (i.e., counterexamples) to that explanation among the nonanomalous data. The first step toward refinement of this explanation requires the generator to identify all features that uniquely bar at least one such false positive. These necessary features are added to the antecedent of the explanation. If the explanation still admits false positives, unused features are selected from the pool based on the number of false positives that they exclude. That is, those features that eliminate the most false positives are appended first. The addition of sufficient features continues until either the explanation no longer creates new anomalies or the pool of features is empty. In the latter case, the explanation is discarded.

Our final two generators implement the method of difference. Here the ambiguities are identical to those in the method of agreement. We retain the assumption that the researcher has recorded the causal features while introducing a second assumption. The original statement of the method of difference requires our examples to have every feature in common save one. Due to the difficulty of meeting this

[3]Although we use a function that emphasizes matching features, others may wish to define similarity using a different function.

requirement, we assume that any subset of features may in itself be considered a single "circumstance" or feature. The existence of feature interactions (cancellations, feedback, etc.) renders this step both reasonable and necessary.

D1 How is the anomaly different from all nonanomalous examples?

D2 How is the anomaly different from nonanomalous examples of the same class?

D3 How is the anomaly different from nonanomalous examples of the incorrectly assigned class?

D4 How is the anomaly different from nonanomalous examples of the incorrectly assigned class that share the same faulty classifier?

Table 3: Questions to be answered using the Method of Difference

Once clear in our assumptions, we can begin to answer the questions in Table 3. The first question asks which features of the anomaly might serve to separate it from all of the correctly classified data. The second question refines the first query, asking which features might be responsible for the incorrect classification. So, for an anomaly to Rule 2 (wheezing is present implies LRS is present), we ask how that case differs from all other cases that are negative for LRS. Question D3 functions as a standard specialization operator in that it asks how to carve the observation space in a way that the anomaly is excluded from the incorrectly assigned class. That is, it reverses Question D2 by asking how the anomaly to Rule 2 differs from cases that are positive for LRS. Finally, Question D4 concentrates on precisely those members of the assigned class that are most similar, in important respects, to the anomaly. Or, in particular, how does the anomaly differ from cases correctly predicted by Rule 2?

Analogous to our first generator, we define a difference as a feature, or set of features, of the anomaly not shared by any nonanomalous data in the current subset. The algorithm for the basic method of difference generator can be found in Figure 2. This generator begins by searching for features unique to each anomaly with respect to the current subset of nonanomalous data. These features, called separators, serve to separate the anomaly from those data. Next, each incorrect classifier associated with a particular anomaly spawns a root explanation as in the agreement generator. Since we require only a single separator to isolate the anomaly from the nonanomalous data, each root explanation is specialized with each separator. For each anomaly, this process yields a number of explanations equal to the number of incorrect classifiers for that anomaly multiplied by the number of separators. As a final step, if the explanation creates any new anomalies among the complete set of nonanomalous data, it is discarded.

The second generator based on the method of difference is also presented in Figure 2 and can be understood as identifying the branch of a decision tree that classifies an anomaly. As with the other generators, a root explanation is created from each incorrect classifier to an anomaly. As an extension of the basic-difference generator, features are added to the antecedent of the explanation until the anomaly is perfectly separated from the nonanomalous data. Selecting the next specialization involves identifying the feature that best isolates the anomaly. For example, if the addition of "cough is absent" places the anomaly in a group with three members of the nonanomalous subset and the addition of "wheezing is absent" places the anomaly in a group with five, then the feature representing "cough is absent" will be added to the rule's antecedent. In the case of a tie, an arbitrary feature is chosen. This "decision branch" approach identifies only a single possible revision. This revision will be discarded if the rule is too general in the context of all the nonanomalous data. While this approach could be without the basic-difference generator, we anticipate that explanations requiring a single separator will correspond well with plausible explanations. Therefore we find it important to identify them all.

```
Method-of-Difference-Basic(anomalies, data):
 explanations = [ ]
 for all a in anomalies
  /* consider each feature that differentiates the anomaly
     from the nonanomalous data */
  for all s in separators(a, data)
   /* consider each rule that incorrectly classifies the
      anomaly */
   for all i in incorrect-classifiers(a)
    e = new explanation(antecedent(i), class(a))
    push(s, antecedent(e))
    unless(overly-general(e))
     push(e, explanations)
 return explanations

Method-of-Difference-DB(anomalies, data):
 explanations = [ ]
 for all a in anomalies
  /* consider each rule that incorrectly classifies the
     anomaly */
  for all i in incorrect-classifiers(a)
   e = new explanation(antecedent(i), class(a))
   while(overly-general(e) and
         features-left(all-features(), e))
    /* add the feature that best distinguishes between the
       anomaly and the nonanomalous data to the current
       explanation */
    push(best-separator(a, e, data), antecedent(e))
   unless(overly-general(e))
    push(e, explanations)
 return explanations
```

Figure 2: Algorithms for the basic-difference generator and the decision-branch generator.

Once the generation process is finished, we are left with a list of records. Each explanation consists of its method of generation, the anomalies that it explains, the base rule associated with those anomalies, and its exception rule. For the purposes of the current work, we are mostly interested

in the last two fields. Therefore, we extract the base and exception rule pairs, as shown in Table 4, and remove all duplicates. These rules are then presented as the final output.

Evaluation

Our source data were extracted by a medical expert from 282 emergency department reports. Thirty-nine variables covering signs, symptoms, findings, and diagnoses related to respiratory syndrome were defined and given one of either three or five values depending on the nature of the attribute. For the current research, these values were recast as either "present" or "absent." A value of "present" signifies that the attribute was explicitly mentioned as present in the report. A value of "absent" signifies that the attribute was either explicitly mentioned as absent or was not mentioned. The expert then assessed whether the patient described in the report was positive or negative for LRS.

Original Rule:
DYSPNEA is PRESENT implies LRS is PRESENT

Exception Rule:
COUGH is ABSENT and
X_RAY_PULMONARY_EDEMA is PRESENT and
DYSPNEA is PRESENT implies LRS is ABSENT

Table 4: An example explanation stating that an x-ray that is positive for pulmonary edema coupled with an absence of cough precludes lower respiratory syndrome (LRS) even though dyspnea is present.

The expert in infectious diseases initially provided a simple, highly sensitive model listed in Table 1 for defining LRS. Although the rules in this model may be probabilistic, we treat them as categorical to simplify the determination of inconsistency. Along with the model, 190 observations (selected uniformly at random from the total 282) were given to the explanation generator. Thirty-one data were identified as anomalies to the model, and of these, nine were false positives to two rules. When an anomaly violates multiple rules, an explanation is required for each violation, which gives a total of forty violations in need of resolution. The anomalies, along with the rules they violated, were sent to the three explanation generators. The resulting explanations were organized by their top-level generator (i.e., agreement, basic difference, and decision-branch difference) and duplicates by the same generator were removed. A subset of the remaining explanations was selected for plausibility evaluation. The agreement generator produced three explanations, all of which were selected for evaluation. The basic-difference generator identified seven explanations and three were selected uniformly at random. The decision-branch generator extracted 42 unique explanations, 21 of which were extracted uniformly at random. One of these 21 explanations was identical to an explanation chosen from the simple-difference generator, leaving a total of 26 explanations for evaluation.

These explanations were paired with the original, overly general rules from the model and given to the expert who volunteered the original model along with a version of the following task instructions.

> "For each rule + explanation pair, indicate whether the explanation is plausible as an exception rule to the original rule from the model. If your answer is yes, then judge whether there is a subset of the antecedent for the explanation that is more plausible. If your original answer is no, then judge whether there is a subset of the antecedent that would render the explanation plausible."

When asked for a clarification of the term "plausible" the expert was asked to use the following question as a guide.

> "Would you feel comfortable making a decision about the case based on this explanation where the decision may be to further investigate the anomaly using this explanation as a guide or to use this explanation to explain away the anomaly."

All answers were given in a "yes" or "no" format.

Results

Required Data Set for the Generation Method	Explanations	
	Plausible	Implausible
All Data (A1 and D1)	3	4
Subset (A2–A4, D2–D4)	7	5
Either of the Above	5	2

Table 5: The number of plausible and implausible explanations that require all the data for a particular generation method, a subset of that data, or when either option is sufficient.

Although we chose our explanations from unique collections, some explanations were generated multiple times for different anomalies and by multiple generators. The results of our experiment are presented in Table 5. Of the 26 explanations, 15 were judged to be plausible, four of which would have been more plausible given a simpler antecedent. Eight of the plausible explanations were generated by viewing either all anomalous data (method of agreement) or all nonanomalous data (method of difference). The remaining seven plausible explanations were only generated when a selected subset of examples, corresponding to one of the questions from A2–A4 and D2–D4, was used. Five of the 11 explanations judged as implausible were generated solely by viewing subsets. All generated explanations could be inserted into the original model without creating new anomalies within the data set. That is, they were all valid choices for theory revision.

Examining the individual explanations, we found that all three explanations generated by the method of agreement required a limited subset of the anomalies. That is, there were no features that all the anomalies held in common; therefore smaller groups of anomalies were required to identify meaningful relationships. These three explanations were all judged plausible with no subset of the antecedent listed as more plausible. The first of these explanations resolved all

23 anomalies to Rule 5 in Table 1, while the second resolved all four anomalies to Rule 3, and the third resolved all three anomalies to Rule 2. This left ten anomalies to Rule 1, five of which were resolved by plausible explanations in the evaluated subset.

The explanations generated through the method of difference generally applied to only a single anomaly, although the five that overlapped multiple anomalies were all judged as plausible.

Discussion

The primary result of our experimentation confirms the first half of our hypothesis. That is, applying Mill's Method to meaningful subsets of data generated a relatively large number (80% of the total) of plausible hypotheses. A reasonably large number of these hypotheses (47%) were uniquely identified by the subset approach. When examining subsets of anomalies (method of agreement) or of nonanomalous data (method of difference) we generated 12 of the 15 plausible explanations. For comparison, only eight of those 15 (53%) were generated using the full sets of examples. This result suggests that the subsets that we have defined are both fruitful, allowing us to locate a large number of hypotheses, and helpful, locating hypotheses that may not be found with a more general set of data. A possible corollary of this finding is that the quantity of data used during anomaly resolution may not be as important as the appropriateness of that data to the task.

The second major result indicates that plausibility could serve as a useful means of selecting model revisions. Since none of the explanations would lead to additional anomalies in the original data set, we are free to select arbitrarily from that set to achieve model repair. However, because 42% of the explanations were implausible, there is a high likelihood of making a poor choice. If we could automatically identify the plausible revisions, it seems reasonable that we could minimize the need for future alterations of the model.

Although this work represents only one step in the process of scientific inquiry, it is an important step. Moreover, in the context of data mining and machine learning, finding plausible explanations for why an observation is an exception to general rules in a model builds confidence in the previously accepted model. In future work, we would like to step outside the boundaries of categorical rules and models to determine when an observation or set of observations should be considered anomalous within the context of statistical models and how those anomalies can be resolved.

Acknowledgments

References

Asker, L. 1994. Improving the accuracy of incorrect domain theories. In Cohen, W., and Hirsh, H., eds., *Proceedings of the Eleventh International Conference on Machine Learning*, 19–27. New Brunswick, NJ: Morgan Kaufmann Publishers.

Brewer, W. F., and Chinn, C. A. 1994. Scientists' responses to anomalous data: Evidence from psychology, history, and philosophy of science. In *The Proceedings of the Biennial Meeting of the Philosophy of Science Association*, volume 1, 304–313.

Carbonara, L., and Sleeman, D. H. 1999. Effective and efficient knowledge base refinement. *Machine Learning* 37(2):143–181.

Chinn, C. A., and Brewer, W. F. 1998. An empirical test of a taxonomy of responses to anomalous data in science. *Journal of Research in Science Teaching* 35(6):623–654.

Darden, L., and Cook, M. 1994. Reasoning strategies in molecular biology: Abstractions, scans and anomalies. In *The Proceedings of the Biennial Meeting of the Philosophy of Science Association*, volume 2, 179–191.

Darden, L. 1991. *Theory Change in Science: Strategies from Mendelian Genetics*. Oxford University Press.

Darden, L. 1998. Anomaly-driven theory redesign: Computational philosophy of science experiments. In Bynum, T. W., and Moor, J., eds., *Digital Phoenix: How Computers are Changing Philosophy*. Oxford, UK: Blackwell. 62–78.

Esposito, F.; Semeraro, G.; Fanizzi, N.; and Ferilli, S. 2000. Multistrategy theory revision: Induction and abduction in inthelex. *Machine Learning* 38(1–2):133–156.

Kuhn, T. S. 1970. *The Structure of Scientific Revolutions*. Chicago, IL: The University of Chicago Press, second edition.

Leake, D. 1992. *Evaluating Explanations: A Content Theory*. Lawrence Erlbaum Associates, Inc.

Mill, J. S. 1900. *A System of Logic Ratiocinative and Inductive Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation*. Longmans, Green, and Co., 8th edition.

Mooney, R. J., and Ourston, D. 1994. A multistrategy approach to theory refinement. In *Machine Learning: A Multistrategy Approach*. San Mateo, CA: Morgan Kaufman. 141–164.

Richards, B. L., and Mooney, R. J. 1995. Automated refinement of first-order horn-clause domain theories. *Machine Learning* 19(2):95–131.

Wong, W.-K.; Moore, A.; Cooper, G.; and Wagner, M. 2002. Rule-based anomaly pattern detection for detecting disease outbreaks. In *The Proceedings of the 18th National Conference on Artificial Intelligence*, 217–223.