Characterizing mammography reports for health analytics

Carlos Rojas
rojascc@ornl.gov

Robert Patton
pattonrm@ornl.gov

Barbara Beckerman
beckermanbg@ornl.gov

Oak Ridge National Lab
Oak Ridge, TN

ABSTRACT

As massive collections of digital health data are becoming available, the opportunities for large scale automated analysis increase. In particular, the widespread collection of detailed health information is expected to help realize a vision of evidence-based public health and patient-centric health care.

Within such a framework for large scale health analytics we describe several methods to characterize and analyze free-text mammography reports, including their temporal dimension, using information retrieval, supervised learning, and classical statistical techniques.

We present experimental results with a large collection of mostly unlabeled reports that demonstrate the validity and usefulness of the approach, since these results are consistent with the known features of the data and provide novel insights about it.

1. INTRODUCTION

Electronic health records are being promoted by the government, private companies, and the medical and research communities ([2, 12, 15]). Naturally, as massive collections of digital health data become available, the challenges and opportunities of automated analysis will appear.

In fact, in the same way that the detailed historical record of every commercial transaction allows financial entities to identify broad patterns of behavior and to estimate how financially "healthy" a particular customer is, the large scale collection of detailed health information is expected to help realize a vision of evidence-based public health and patient-centric health care, [4, 27].

That is, the analysis of health records should help to discover both the *commonalities* shared by the population as well as the *particular* features that make a certain patient unique. Moreover, taking into account the *temporal dimension* can help to detect the *appearance* of unexpected events, such as outbreaks, [10], temporal patterns between drug prescriptions and medical events, as in [20], or help to device timely and effective prevention campaigns.

Idealized patient data representation for analytics.

Patient data can relate to a wide variety of aspects (allergy information, prescription information, test results, radiology results, demographic information, various types of clinical notes, etc.) which can be available in several electronic formats (images, text, spreadsheets, etc.). Depending on how well all these data can be structured, organized, searched, filtered, and compared in an automated way, many of the principles or even specific methods of existing large scale learning and pattern discovery techniques could be applied.

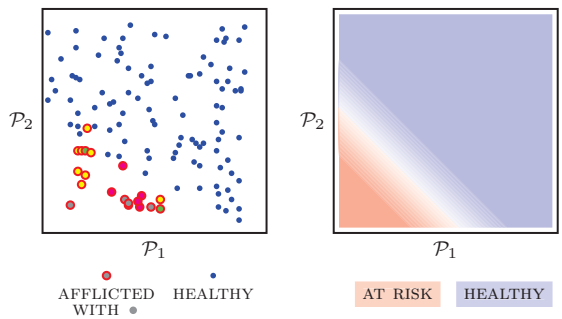
Ideally, amassing all these data about healthy and unhealthy patients would help to achieve a better understanding of the health status of large populations, and, hopefully, the development of decision systems that can facilitate the physician's work at the individual level.

In the idealized situation shown in Fig. 1a all the patient data is mapped to the space $\mathcal{P}$ with two dimensions, $\mathcal{P}_1$ and $\mathcal{P}_2$, where healthy patients appear as blue dots (•), and unhealthy patients appear as red circles (○), whose inner color indicates a particular ailment (e.g., • representing a specific heart condition). From the data mapped to this idealized space $\mathcal{P}$ one could then discover the decision boundary in Fig. 1b to determine whether a patient may be at risk of certain condition, based on the similarity of his/her data to patients already diagnosed, where the darker the color the higher the certainty of the decision.

The status of a patient, however, changes with time, and his/her associated data changes as well. Moreover, medical science is also dynamic. Thus, even in the idealized mapping of Fig. 1, the location of a patient in Fig. 1a would change with time, as would the decision boundary in Fig. 1b. Therefore, a better representa-

Copyright Notice: This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
First International Health Informatics Symposium
November 11-12, 2010
Arlington, VA
Copyright 2010 ACM X-XXXXX-XX-X/XX/XX ...$10.00.



(a) Original data (b) Decision boundary

Figure 1: An idealized mapping of patient data to a two-dimensional space: (a) Original data, (b) Decision boundary.

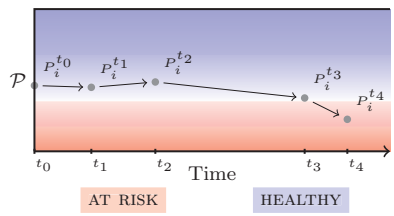


Figure 2: Idealized depiction of a patient's data trajectory.

tion for the patient data (that would take into account his/her evolution) is as a *trajectory* in time through the data space, as depicted in Fig. 2. In this idealized example, the data for patient i, P_i, moves through the one-dimensional patient space $\mathcal{P}$ at different points in time *relative* to the first time data was collected from him/her. As an instance of usage of these trajectories one could take all patients that are being treated for certain condition and analyze whether they evolve as expected.

Although the scenarios from Figs. 1 and 2 (which encompass all conditions integrally) may be completely unattainable in reality, they are already in place for many specific cases. For example, to determine blindness and low vision conditions the patient data space is given by all the possible results of a visual acuity test, and the decision boundary has been determined by the World Health Organization' International Classification of Diseases[1]. Thus, what can be done is to systematically deal with one or few conditions at a time.

In this spirit, we present a methodological approach to characterize clinical notes, specifically, mammography reports, including their temporal dimension. We also discuss how to use these characterizations, and provide examples of the insights that can be drawn from them.

[1] `http://apps.who.int/classifications/apps/icd/icd10online/index.htm?navi.htm+ka00`

Problem definition.

Given a large amount of mammography reports, some of them labeled with diagnostic information (i.e., NORMAL or SUSPICIOUS findings), we want to characterize the *reports* and the *patient's report history,* as a first step towards large scale data analytics.

Our solution.

Our approach is to obtain a representation of the data that allows to understand important features of the full data set without being overwhelming, and to validate it using the labeled data.

We represent a *report* using either the full text, or as a reduced vocabulary based on the clinical descriptions. We use both types of representations independently to calculate a measure of distance between every pair of reports. These distances are then used to map the reports to two- and one-dimensional spaces.

Also, we map the sequences to *trajectories* in a two-dimensional space: the first dimension is the time between mammographies (relative to each patient's first) and the second dimension is the one-dimensional mapping of the report.

We represent a *patient's report history* as a sequence of reports. We use a well-known measure between time series to calculate the distance between sequences using the text representations.

Paper contributions.

In the context of large scale health analytics, we discuss several alternatives to characterize mammography reports with varying complexity that *do not* incorporate domain knowledge, including novel ideas to take temporal information into account.

We also use a combination of information retrieval, supervised learning, and classical statistical techniques to gain a better understanding of the mammography reports, which, to the best of our knowledge, has not been done before.

Paper organization.

The paper is organized as follows. Section 2 reviews relevant research work and concepts mostly about mammography data and vector space representation for documents. Section 3 describes formally the different characterizations we defined for the data. In Section 4 we detail the variety of experiments carried out to explore the usefulness of the characterizations, and discuss the results in Section 5. Section 5 concludes the paper.

2. BACKGROUND

Mammography data.

The data from a mammography exam usually consists of four images (two projections, caudal and lateral, per side, right and left) and a radiology report containing the observations and professional opinion of the radiology expert[s] about the mammogram.

Computer-aided analysis of mammography data for the detection and diagnosis of breast cancer has been an

important subject of interdisciplinary research over the past decades (e.g., [9, 28]). From the computer science perspective the problem has been framed as an image classification task. Mammograms are to be classified as normal or abnormal, with the type of abnormality to be identified as well. The ultimate goal for such computer systems is to reduce the amount of human work needed to perform the analysis, while maintaining or improving the detection rates.

Until recently, text reports were written in medical terminology but were largely free text. Currently the controlled BI-RADS vocabulary[2] is used. An important feature of the pre-BI-RADS text is the prevalence of negation phrases (e.g., "no strongly suspicious forms", "no malignancy"), which can have several variations with the same basic meaning (e.g.,"no mammographic finding suspicious", "no strongly suspicious forms", "no strongly suspicious features"). These negation features are pervasive through all types of clinical notes, as reported in [8, 18].

There has been research with free-text mammography reports to automatically generate the BI-RADS descriptors ([7, 19]); to discover significant phrase patterns, such as the basic template that unifies the variations of a group of negation phrases (e.g., "no * suspicious"), called *s-grams* [21]; to extract temporal properties associated to medical events, [11]; to identify findings suspicious for breast cancer [13]; in ad-hoc classification, [3].

Analysis of the temporal dimension of the mammography images has also being researched, although a lesser degree. In the clinical side, the authors in [24] report that "Comparison with prior mammograms significantly improves overall performance [of breast cancer detection in screening]." Computer-aided methods have been researched to detect masses and to analyze their change, [16, 29].

To our knowledge, there is no prior work in characterizing the temporal properties of mammography reports.

Vector space representation for documents.

In text mining and information retrieval the basic document representation is a "bag-of-words", i.e., the list of words that appear in the document without taking into account order. Very common words (either collection-specific or function words such as articles and prepositions) are usually ignored; they are called *stop words*. In order to reduce the size of the vocabulary, words may undergo *stemming* to obtain their common roots; for instance, all *write*, *writer*, and *writing* would map to *writ*.

The bag-of-words representation can be transformed to a vector in an $m-$dimensional space where m is the number of distinct words in all documents under consideration, i.e., the vocabulary $V = \{w_i\}$ with $|V| = m$. In this *vector space* representation the particular document document i in the collection D is $d_i = \{w_{i,1}, w_{i,2}, \ldots, w_{i,m}\}$, where $w_{i,j}$ is the *weight* of the word w_j in the document d_i, [26]. Most documents only have a small subset of all the words in V, and, therefore, only the non-zero weights are actually considered. The weighting schema usually reflects a balance between how frequent the word is in the document vs. how frequent it is in the collection. Words that appear in many documents are not helpful in characterizing them, the same way that describing an individual dog as "four-legged" would not help to identify it, nor to distinguish it from other dogs.

The vector space representation facilitates the computation of similarity between documents. A common measure is the *cosine similarity*:

$$\begin{aligned} sim\,(d_i, d_j) = \cos\theta &= \frac{d_i \cdot d_j}{\|d_i\| \times \|d_i\|} \\ &= \frac{\sum_{k=1}^{m} w_{i,k} \times w_{j,k}}{\sqrt{\sum_{k=1}^{m} w_{i,k}^2} \times \sqrt{\sum_{k=1}^{m} w_{j,k}^2}} \end{aligned}$$

where θ is the angle between the two vectors. Since the weights are non-negative, $sim\,(d_i, d_j)$ is always between 0 and 1.

Dynamic time warping distance.

Dynamic time warping (DTW) is a technique to measure the distance between time series, which has been widely used in speech recognition and data mining, e.g.,[5, 30, 25]. DTW matches points in the series that do not necessarily correspond in time, but that minimize the total distance, using dynamic programming. Naturally, a measure of distance between *pairs* of points must be specified.

DTW is particularly useful when the time series have different lengths and a measure such as Euclidean distance cannot be applied. Fig. 3 shows an example of such matching. In this case, the total distance is the sum of the lengths of the segments that connect the two time series (in red).

Multidimensional Scaling and Principal Component Analysis.

Multidimensional scaling [6], *MDS*, is a statistical technique that allows to reconstruct points in an Euclidean space of a user-specified number of dimensions, based solely in their distance information. MDS works as an iterative process that attempts to minimize the differences between the original distances and the distances in the reconstructed space.

Principal component analysis [14], *PCA*, operates directly with the points in an Euclidean space and transforms them into a lower dimensionality. In PCA's output, the new first dimension captures the projection of the data with the greatest variance, the new second dimension captures the second greatest variance, and so on.

3. CHARACTERIZING MAMMOGRAPHY REPORTS FOR ANALYTICS

[2]Breast Imaging Reporting and Data System, developed by the American College of Radiology.

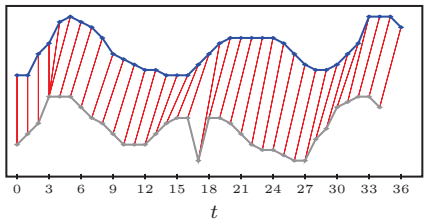


Figure 3: Dynamic Time Warping example of matching between two time series.

To represent the mammography report for patient i we explored several alternatives. They are listed in Table 1, and explained below.

Vector Space with TF.

This representation uses the vector space model with the weight being simply the number of times the word appears in the report, i.e., the *term frequency TF*. This representation was obtained using *rainbow* [17], an application for information retrieval and text classification. Default stop words were removed, except for the words `no` and `not`, but *no stemming* was performed.

Thus, in this representation most of the information of the raw text is preserved. For the patient i, its report at time t_j in this representation is noted $P_i^{t_j}$.

Vector Space with TF-ICF.

This representation uses *term frequency - inverse corpus frequency* weighting (*TF-ICF*). In this weighting scheme, the weight $w_{i,j}$ of the word j in the document i is given by:

$$w_{i,j} = \log(1 + n_{ij}) \times \log\left(\frac{N+1}{n_j+1}\right)$$

where n_{ij} is the term frequency of the word j in document i, N is the number of words in a reference corpus, and n_j is the overall number of times the word j appears in the reference corpus. The reference corpus is TREC-5 ([1]), a collection of about a quarter million of news feeds documents. TF-ICF has been shown to be adequate for typical text analysis tasks [23]. Default stop words were removed *including* the words `no` and `not`, and stemming was performed, using the Porter stemming algorithm,[22].

Thus, this representation does not take into account collection specific weights, but performs vocabulary reduction via stemming, and removes the potentially useful negation indicators. For the patient i, its report at time t_j in this representation is noted $\mathbb{P}_i^{t_j}$. The complete patient history is noted $\overrightarrow{\mathbb{P}_i} = \{\mathbb{P}_i^{t_0}, \mathbb{P}_i^{t_1}, \ldots, \mathbb{P}_i^{t_{|\overrightarrow{\mathbb{P}_i}|-1}}\}$, and the space of all reports is noted $\mathbb{P}$.

Vector Space with s-grams.

The s-grams unify variations from a group of phrases. For example, "`no mammographic finding suspicious`", "`no strongly suspicious forms`", "`no strongly suspicious features`" can be unified as "`no suspicious`". More precisely, s-grams or *skip grams*, are word pairs in their respective sentence order that allow for arbitrary gaps between the words [21].

In this representation every report is the list of the s-grams, i.e., the weights are either 0 or 1. Thus, this representation is extremely compact but, at least in principle, captures the relevant clinical descriptors from the reports. For the patient i its report at time t_j in s-gram representation is noted $p_i^{t_j}$. The complete patient history is noted $\overrightarrow{p_i} = \{p_i^{t_0}, p_i^{t_1}, \ldots, p_i^{t_{|\overrightarrow{p_i}|-1}}\}$, and the space of all reports is noted p.

Two dimensional Euclidean Space mapped from p and $\mathbb{P}$.

From the s-gram and the TF-ICF representations we computed the cosine similarity $\cos\theta$ between every pair of selected patient records representations, converted them to a *distance* measure by computing $(1 - \cos\theta)$, and obtained a symmetric matrix of distances D. We applied multidimensional scaling to the matrix D to obtain a two dimensional space. For the patient i its report at time t_j in this space is noted $\mathcal{P}_i^{t_j}$, and the space of all mapped reports is $\mathcal{P}$. When needed, to distinguish whether $\mathcal{P}$ is obtained from the s-gram or the TF-ICF representation we note it as $p \to \mathcal{P}$ and $\mathbb{P} \to \mathcal{P}$, respectively.

One dimensional Euclidean Space mapped from $\mathcal{P}$.

This representation is simply the first component of the results of principal component analysis when applied to the patient reports in their $\mathcal{P}_i^{t_j}$ representation, obtained from the TF-ICF model (i.e., from $\mathbb{P} \to \mathcal{P}$), and it is noted $\rho_i^{t_j}$. As before, the space of all mapped reports is ρ.

Two dimensional Euclidean Space mapped from $\overrightarrow{\mathbb{P}_i}$.

Thanks to the temporal information, the patient history $\overrightarrow{\mathbb{P}_i}$ can be considered a time series. Thus, for all pairs of patients we computed the distance between their trajectories using DTW (with $(1 - cos\theta)$ as the distance measure between matching points) and generated the distance matrix $\mathcal{D}$. As before, we applied MDS to the matrix $\mathcal{D}$ to obtain a two dimensional space. The representation of the full report history for patient i in this space is noted $\mathcal{T}_i$, and the complete space is $\mathcal{T}$.

Timestamps representation.

Timestamps for a patient are *relative* to the first exam, and measured in fractions of years. That is, by convention, $t_0 = 0$, and every subsequent t_{j-1} is the time interval between the first exam and the $j-$th exam. For example, if a patient has her *first* exam on January 1st, 1983 and her *third* exam on November 1st, 1983, then $t_2 = 304/365 = 0.83$. This allows to "align" patient histories regardless of the actual dates of their exams.

4. EXPERIMENTAL RESULTS

Dataset description.

For Patient i		Type of Representation
Report at time t_j	Complete History	
$P_i^{t_j}$		Vector Space with TF
$\mathbb{P}_i^{t_j}$	$\overrightarrow{\mathbb{P}_i}$	Vector Space with TF-ICF
$p_i^{t_j}$	$\overrightarrow{p_i}$	Vector Space with s-grams
$\mathcal{P}_i^{t_j}$	$\overrightarrow{\mathcal{P}_i}$	2D Euclidean Space mapped from $\mathbb{P}_i^{t_j}$ or $p_i^{t_j}$
$\rho_i^{t_j}$	$\overrightarrow{\rho_i}$	1D Euclidean Space mapped from $\mathcal{P}_i^{t_j}$
	$\mathcal{T}_i$	2D Euclidean Space mapped from $\overrightarrow{\mathbb{P}_i}$

Table 1: Summary of representation notation. See text for details.

```
Y
***********************************
RADIOLOGY CONSULTATION XX-XX:16:99
BHIS #:  XXXXXXXXX OUTPATIENT 6346144
XXXXX, XXXX X.
XXX XXX XXX 81 FEMALE
Clinical data:  Routine screening.  Patient is currently
asymptomatic.  She is on hormonal therapy.  XXXX, MD Screen-
ing Mammogram (XXXX/Other Covered Commercial Ins) -- Exam
#XXX on XX/XX/XX COMPARISON: XX/XX/XX FINDINGS: Standard
films of both breasts were obtained.  [...]
```

Figure 4: Fragment of a report.

Our data is a collection of $57,794$ mammography reports, $19,905$ of them with an assigned set of mammography images, and $12,372$ patients,. There are 90 unique reports labeled as NORMAL, and 96 unique reports labeled as SUSPICIOUS (note that there is not necessarily a diagnosis). Timestamps span over 14 years. The data was collected and anonymized by researchers at the University of Chicago, and prepared and indexed by the authors. Fig. 4 shows a fragment of a report.

Besides the raw text for each report we extracted the s-grams, and also associated the anonymized patient identifier, the date of the mammography, and a short description of the type of exam (e.g., "`Screening Mammogram`").

Table 2 shows the 10 most frequent exam descriptions; unsurprisingly, screening mammograms are the most frequent. Table 3 shows all the exam descriptions for the frequency for NORMAL reports; they add up to 89 instead of 90 because of a missing exam description. Table 4 shows the 10 most frequent exam descriptions for SUSPICIOUS reports; they add up to more than 96 because of multiple exam descriptions.

We used 137 distinct s-grams, obtained with the technique described in [21]. Table 5 shows the 10 most frequent s-grams; unsurprisingly, the most common are negation phrases.

Exam Description	Frequency
Screening Mammogram	40796
Diagnostic Mammogram	4848
Mammo-Limited compression views	4088
Dedicated Mammogram	2782
Breast Ultrasound Limited Study	1983
Bilateral Mammogram	990
Mammography Comparison	920
Right Unilateral Mammogram	488
Left Unilateral Mammogram	440
Needle Biopsy Procedure - Breast	417

Table 2: The 10 most frequent values for the exam description.

Exam Description	Frequency
Screening Mammogram	64
Bilateral Mammogram	11
Diagnostic Mammogram	7
Dedicated Mammogram	6
Mammo-Limited compression views	1

Table 3: The frequent values for the exam description for NORMAL reports.

Exam Description	Frequency
Mammo-Limited compression views	38
Breast Ultrasound Limited Study	35
Diagnostic Mammogram	16
Needle Biopsy Procedure - Breast	8
Screening Mammogram	7
Dedicated Mammogram	6
Left Stereo Core Needle Breast [...]	6
Bilateral Mammogram	5
Right Unilateral Mammogram	4
Breast Ultrasound	2

Table 4: The 10 most frequent values for the exam description for SUSPICIOUS reports.

S-gram	Frequency
no malignancy	32959
no suspicious	25867
clustered microcalcifications	25736
no masses	25438
no clustered	23622
no microcalcifications	21026
no architectural	20903
no radiographic	20659
no change	18876
no features	17348

Table 5: The 10 most frequent s-grams.

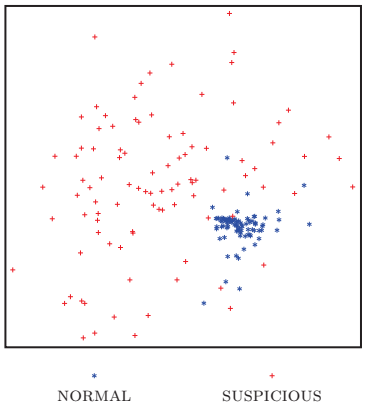


Figure 5: NORMAL and SUSPICIOUS reports mapped to an Euclidean two dimensional space from $\mathbb{P}_i^{t_j}$.

Supervised learning.

Since there are labeled reports we can explore how hard it is to distinguish the reports from a healthy patient from those of a patient who is potentially ill.

Thus, we performed experiments using the Naive Bayes text classifier as implemented in *rainbow*, with the TF and the s-gram representations. We use 10 runs using 60% of the labeled data as training and 40% as testing. The average accuracy for the TF representation was 95.4 ($\sigma = 2.8$) and for the more compact s-gram representation was 95.7 ($\sigma = 2.4$). That is, both representations have almost the same, and very high, performance.

We also performed experiments using a nearest neighbor classifier with the 2D Euclidean Space $\mathcal{P}$ mapped from the TC-ICF and the s-gram spaces ($\mathbb{P}$ and p, respectively). Again with 10 runs and a 60%/40% training/testing split we obtained an average accuracy of 89.45 ($\sigma = 2.76$) and 91.21 ($\sigma = 1.46$), for the TC-ICF and s-gram representations, respectively. Therefore, at least from the labeled data, even the transformed Euclidean space still contains enough information to distinguish the labels for most of the reports.

Reports.

The reports in their two dimensional Euclidean space representation $\mathcal{P}$ can be visualized, as shown in Fig. 5; this particular visualization comes from the mapping of the reports in their TF-ICF space, $\mathbb{P}$. Thus, at least in this small sample of labeled data, the visualization reveals a space close to the ideals of Fig. 1, although reversed: instead of a large region of NORMAL and small region of SUSPICIOUS, the NORMAL reports appear clustered together, while the SUSPICIOUS reports are scattered. This can be explained by considering that normal reports tend to *enumerate all the suspicious features that are not present* and, therefore, are similar to each other, while the suspicious reports *describe in detail the few suspicious features that do appear* and, therefore, are different from all other reports. This is also consistent with the differences in the variety of exam types for both classes (See Tables 3 and 4)

Adding more reports, however, clouds the picture slightly. Fig. 6b shows 8,000 reports mapped to $\mathcal{P}$ from $\mathbb{P}$, including those labeled, which are shown by themselves in Fig. 6a. Although there is still a core of normality, it is more mangled with the scattered suspiciousness. In fact, when a nearest neighbor classifier is applied to this space, the average accuracy decreases to 84.46 with a much larger $\sigma = 5.18$ (as before, we ran the classifier 10 times with 60%/40% different training/testing splits.)

Trajectories.

Using the temporal dimension of the reports and their one dimensional space ρ we can visualize the patient's histories as trajectories. Thus, we mapped *all* reports from patients *with a labeled report* to their TC-ICF representation in $\mathbb{P}$, obtained their two dimensional representation in $\mathcal{P}$ and, finally, obtained their one dimensional representation in ρ. In total, there were 716 reports. Figure 7 shows the last two reports in the trajectories of patients having NORMAL (Fig. 7a) and SUSPICIOUS (Fig. 7b) reports.

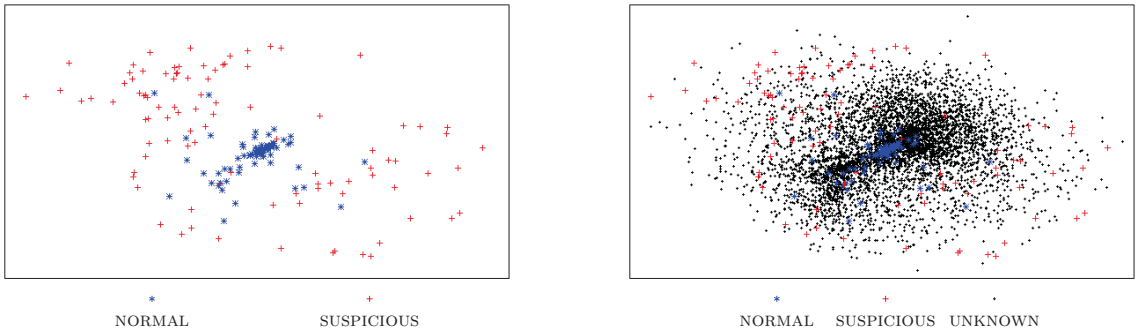
Both sets of reports appear to span the same period of time, i.e., about 14 years. From the NORMAL reports it appears that the patients are taking a little longer than the recommended year for subsequent exams. On the other hand, the last two points in the trajectories of patients with SUSPICIOUS reports appear much closer in time, suggesting that the report *before* the SUSPICIOUS one generates a sense of urgency and the need for another exam. In fact, from Table 7b, several SUSPICIOUS reports refer to ultrasounds and biopsies procedures, which would have been done *after* observing suspicious features in a regular, screening mammogram.

Patients' histories.

The full patients' histories mapped to the two dimensional space $\mathcal{T}$ are shown in Figs. 8a (only patients' histories with a labeled report) and 8b (all mapped patients' histories). There are 7,444 patient histories comprising 34,103 reports. Only 126 patients with labeled reports actually have a history, 66 and 60 of them with a NORMAL and a SUSPICIOUS report, respectively.

From Fig. 8a it appears that patients' histories with reports labeled as NORMAL do not cluster together, as well as their reports do. In fact, they are almost as scattered as the patients' histories with SUSPICIOUS reports. There is still some degree of separation between the two classes, however. Thus, when a nearest neighbor classifier is applied to this space, the average accuracy is 77.6 ($\sigma = 2.8$), for 10 runs with 60%/40% different training/testing splits. Note that this result (from patients' histories) does not necessarily compares with the ones above (from reports.)

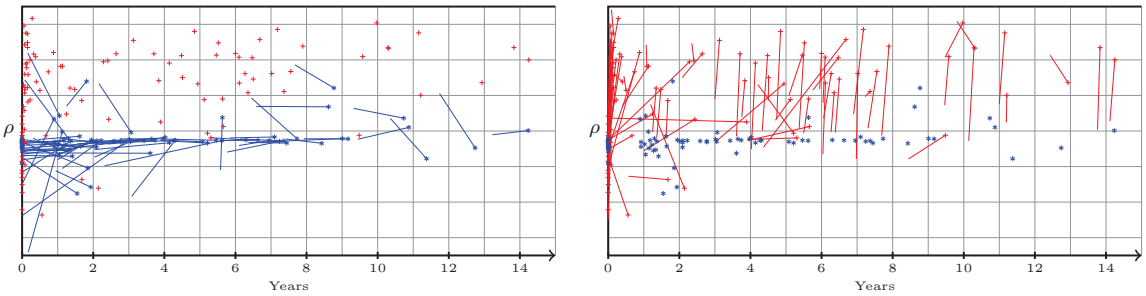
The 'layers' of points in Fig. 8b are a consequence of the DTW distance and the great differences between patients histories' lengths. Since DTW can match the same point in one time series to several points in the



(a) Only labeled reports

(b) All mapped reports

Figure 6: 8,000 reports mapped to the Euclidean two dimensional space $\mathcal{P}$, including those labeled.



(a) NORMAL.

(b) SUSPICIOUS.

Figure 7: Last two reports in the trajectory of patients with labeled reports.

other (See Fig. 3), it effectively penalizes pairs of patients' histories having very different lengths; thus, in Fig. 8b, their mapped points would tend to be separated, while those with same length would 'chain together'. This helps to explain the scattering pattern of the patients' histories with labeled reports.

5. DISCUSSION OF RESULTS

The experimental results indicate that the information in the mammography reports is basically preserved through the transformations induced by the various representations, since both the visualizations and the supervised learning results consistently match to what is known from the data. More importantly, as was the case for the different temporal dynamics that surfaced in Fig. 7, *novel* insights about the data were gained.

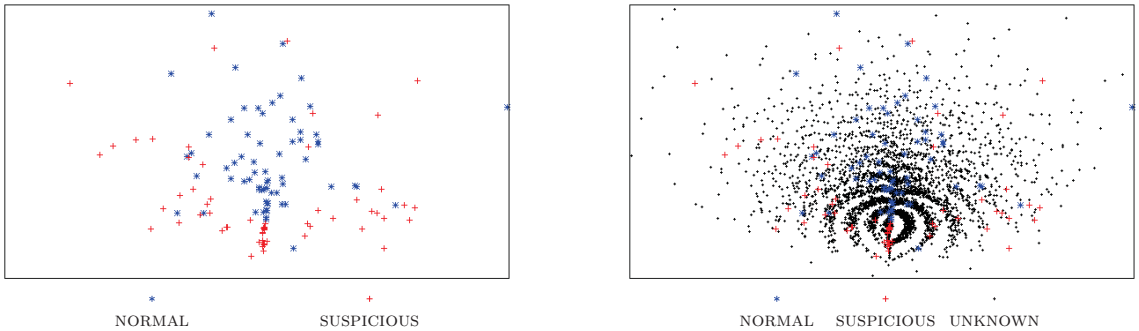
It is important to highlight that, except arguably for the s-grams representation, *no domain knowledge* was incorporated (e.g., no BI-RADS metadata), and not even information extraction techniques were used. Therefore, the representations are, in a sense, rough. Moreover, dynamic time warping is probably not the best way to measure the distance between patients' histories, since it matches *all* points rather than finding best *subsequences*. Even worse, DTW assumes that the timestamps are uniformly separated (i.e., that the time series are *sampled* at a constant rate) which is not the case for clinical notes, as Fig. 7 shows clearly.

The fact that *despite* these limitations the different representations still keep the gist of the original data can be attributed to the richness of the raw reports. In fact, a lay person is probably able to distinguish a suspicious report from a normal one since, after all, a radiology expert has provided a very precise description of what the mammography *means*. The interesting implication is that vast collections of free-text clinical notes may not require an excessive amount of work to be useful.

6. CONCLUSIONS

We have discussed an idealized scenario for large scale health analytics and presented several ways to characterize mammography reports that fit this scenario in broad terms. We took advantage of these characterizations to analyze a large collection of reports, including their temporal dimension, using methods from information retrieval, supervised learning, and classical statistical techniques.

The experimental results demonstrate the validity and the usefulness of the approach, since they both conformed to what was expected from the data and helped to get novel insights about it.



(a) Only patients' histories with labeled reports

(b) All mapped patients' histories

Figure 8: 7,444 patients' histories mapped to the Euclidean two dimensional space $\mathcal{T}$, including those with labeled reports.

An important goal towards the future is to discover *why* certain patients clinical notes or histories are mapped together. That is, to *identify* the commonalities that make them appear together which, hopefully, would correspond to specific clinical aspects.

Since we have discussed a notion of time and space for the patient data, questions about the *rate* of change and the *direction* of the trajectory naturally arise. For example, one would be interested in finding the trajectories that *contain subsequences* that are closer to a given patient's history, since that could help to determine the likely path of the patient's evolution.

Thus, we look forward to continue research in several aspects: exploring of the clinical significance of the characterizations; the inclusion of domain knowledge information in the representations; more appropriate ways to measure and compare the temporal properties of the data.

7. ACKNOWLEDGMENTS

We thank Robert M. Nishikawa, Ph.D., Department of Radiology, University of Chicago, for providing the large dataset of unstructured mammography reports.

Prepared by Oak Ridge National Laboratory, P. O. Box 2008, Oak Ridge, Tennessee, 37831-6285, managed by UT-Battelle, LLC, for the U.S. Department of Energy Under contract DE-AC05-00OR22725. Research partially sponsored by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, LDRD #5327.

8. REFERENCES

[1] TREC-5, 1999. http://trec.nist.gov.

[2] North Carolina Medical Journal. Special Issue on Data and Health Policy, March/April 2008.

[3] D. B. Aronow, F. Fangfang, and W. B. Croft. Ad hoc classification of radiology reports. *J Am Med Inform Assoc*, 6(5):393–411, 1999.

[4] R. Bakalar. IBM's Vision for the Future in Patient-Centric Global Health Care: IBM's Vision of How Advanced Health Analytics and Automated Health Information Infrastructure Will Transform Anatomic Pathology Services. *Archives of Pathology & Laboratory Medicine*, 132(5):766–771, 2008.

[5] D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *KDD Workshop*, pages 359–370, 1994.

[6] I. Borg and P. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer, 1996.

[7] B. Burnside, H. Strasberg, and D. Rubin. Automated indexing of mammography reports using linear least squares fit. In *Proc. of the 14th International Congress and Exhibition on Computer Assisted Radiology and Surgery*, pages 449–454, 2000.

[8] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. Evaluation of negation phrases in narrative clinical reports. *Proc AMIA Symp*, pages 105–109, 2001.

[9] M. Giger. Computer-aided diagnosis of breast lesions in medical images. *Computing in Science Engineering*, 2(5):39 –45, sep/oct 2000.

[10] R. Heffernan, F. Mostashari, D. Das, A. Karpati, M. Kulldorff, and D. Weiss. Syndromic surveillance in public health practice, New York City. *Emerg Infect Dis*, 10(5):858–64, 2004.

[11] R. G. Henk Harkema, Andrea Setzer and M. Hepple. Mining and modelling temporal clinical data. In *Proceedings of the UK e-Science All Hands Meeting*, 2005.

[12] C. Howell. Stimulus package contains $19 billion for health care technology spending and adoption of electronic health records. Wisconsin Technology Network news, February 19 2009. (Retrieved April 29, 2010, at http://wistechnology.com/articles/5523/).

[13] N. L. Jain and C. Friedman. Identification of findings suspicious for breast cancer based on natural language processing of mammogram

reports. *Proc AMIA Annu Fall Symp*, pages 829–833, 1997.

[14] I. Jolliffe. *Principal Component Analysis*. Springer, 2002.

[15] S. Lohr. Tech companies push to digitize patients' records. New York Times, September 10 2009.

[16] F. Ma, M. Bajger, and M. Bottema. Temporal analysis of mammograms based on graph matching. *Digital Mammography*, pages 158–165, 2010.

[17] A. K. McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/ mccallum/bow, 1996.

[18] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*, pages 128–144, 2008.

[19] H. Nassif, R. Woodsz, E. Burnsidey, M. Ayvacix, J. Shavlik, and D. Page. Information Extraction for Clinical Data Mining:A Mammography Case Study. In *ICDM - DDDM09 Workshop*, 2009.

[20] G. Norn, J. Hopstadius, A. Bate, K. Star, and I. Edwards. Temporal pattern discovery in longitudinal electronic patient records. *Data Mining and Knowledge Discovery*, 20:1–27, 2010.

[21] R. M. Patton, T. E. Potok, B. G. Beckerman, and J. N. Treadwell. A genetic algorithm for learning significant phrase patterns in radiology reports. In *GECCO '09: Proceedings of the 11th Annual Conference Companion on Genetic and Evolutionary Computation Conference*, pages 2665–2670, New York, NY, USA, 2009. ACM.

[22] M. F. Porter. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137, 1980.

[23] J. W. Reed, Y. Jiao, T. E. Potok, B. A. Klump, M. T. Elmore, and A. R. Hurson. Tf-icf: A new term weighting scheme for clustering dynamic data streams. In *ICMLA '06: Proceedings of the 5th International Conference on Machine Learning and Applications*, pages 258–263, Washington, DC, USA, 2006. IEEE Computer Society.

[24] A. Roelofs, N. Karssemeijer, N. Wedekind, C. Beck, S. van Woudenberg, P. Snoeren, J. Hendriks, M. Rosselli del Turco, N. Bjurstam, H. Junkermann, et al. Importance of Comparison of Current and Prior Mammograms in Breast Cancer Screening1. *Radiology*, 242(1):70, 2007.

[25] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 26(1):43 – 49, feb 1978.

[26] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513 – 523, 1988.

[27] J. Studnicki, J. W. Fisher, and C. N. Eichelberger. NC-CATCH: North Carolina Comprehensive Assessment for Tracking Community Health. [2], pages 122–126.

[28] J. Tang, R. Rangayyan, J. Xu, I. El Naqa, and Y. Yang. Computer-aided detection and diagnosis of breast cancer with mammography: Recent advances. *Information Technology in Biomedicine, IEEE Transactions on*, 13(2):236 –251, march 2009.

[29] S. Timp, C. Varela, and N. Karssemeijer. Temporal change analysis for characterization of mass lesions in mammography. *Medical Imaging, IEEE Transactions on*, 26(7):945 –953, july 2007.

[30] B.-K. Yi, H. V. Jagadish, and C. Faloutsos. Efficient retrieval of similar time sequences under time warping. In *ICDE '98: Proceedings of the Fourteenth International Conference on Data Engineering*, pages 201–208, Washington, DC, USA, 1998. IEEE Computer Society.

APPENDIX

A. PREFERRED REVIEW APPROACH

- Primary and secondary focus of the paper: Computing, Medicine.
- Main three topics covered in the paper:
  - Large-scale longitudinal mining of medical records
  - Computational support for patient-centered and evidence-based care
  - Innovative applications in electronic health records