Time Constrained Dynamic Programming for Inference of Disease Age of Onset in Longitudinal Medical Information

Bradley Malin MALIN@ANDREW.CMU.EDU
School of Computer Science and Heinz School of Public Policy, Carnegie Mellon University, Pittsburgh, PA 15213

Abstract

This work addresses the issue of optimal alignments of disease states in medical information over a known time period. An inference strategy for determining disease age of onset using time constrained dynamic programming is introduced and applied to a cohort of Huntington's disease patients. The emission probabilities for diagnoses are determined using a multivariate Bernoulli model. In theory, we find the time constrained model to be more accurate than simple dynamic programming. Through this methodology, the age of onset was correctly bounded in 18% (4 of 22) of a cohort of Huntington's disease patients.

1. Introduction

One of the most sensitive data sharing situations exists in the medical community, where the information can influence an individual's credit, employment, and insurance, as well as establish stigmatizing circumstances. Different types of medical information are collected for various reasons. For example, the National Association of Health Data Organizations (NAHDO 2000) reported that 44 of 50 states have legislative mandates to gather hospital-level data on each patient visit. Many states have subsequently distributed copies to researchers, sold copies to industry, and made versions publicly available. The released copies do not include explicit identifiers, such as name or address, though they do include demographic fields such as {*5-digit ZIP,* date of birth, gender}.

Experiments have shown that matching the previous databases with publicly available information, such as population registers, can link names to 87% of individuals in the United States (216 million of 246 million people) (Sweeney 2001). Another type of medical dataset that is being compiled consists of DNA sequences. These collections are location specific, such as a particular hospital or biopharmaceutical firm. Releases of these datasets are made available, with some removal of information, so that the only demographic information associated with these databases is the knowledge of the originating institution. Yet, prior research makes it evident that there are inferences that can be made in both the DNA databases and the discharge databases to link the two, and thus append names and other information to the DNA entries (Malin & Sweeney 2000).

Previous research considers the identifiability of DNA database entries in the context of collections from multiple hospital. (Malin & Sweeney 2001) This entails having DNA databases at multiple hospitals, with overlapping entries in different databases. However, the previous work does not address the issue of DNA from a single hospital dataset or resolving ambiguous relationships in patient audit trails. To clarify, if there is only one hospital collecting and electronically storing DNA sequences[1], the construction of a patient audit trail, the record linkage technique used in the previous, can not be constructed. The re-identification cannot be performed using previous means. As a result, the DNA database entries appear to be more anonymous than in truth they are.

This paper establishes a methodology for conducting inferences that can be used to link publicly available medical information and DNA database entries. It concerns a cohort of individuals with Huntington's disease, which is a trinucleotide (three pieces of DNA) repeat disorder. It has been shown that there is an inverse relationship between the number of repeats and the age of onset (Brinkman, et. al. 1997, Timchenko & Caskey 1996). Thus, an individual's age of onset can be probabilistically deduced from a DNA database entry. Yet, the age of onset is not explicitly listed in the discharge databases, and thus must be inferred from a patient's medical record over time, henceforth referred to as a longitudinal medical profile. We conduct the inference through a combination of several techniques. Multi-variate Bernoulli is used to estimate the probability distribution of observable features being observed in a stage of the disease. Each hospital visit in a patient profile is assigned a probability of being in a certain state of the disease via naïve Bayes, using only positive observations. Then we introduce a variation of dynamic

[1] In actuality, this is an unrealistic situation. Many hospitals collect DNA and store it for additional analysis or licensed research. However, it may be the case that a particular hospital specialization in treatment of a certain disorder. In such a case, DNA entries at this hospital may be non-overlapping single occurrence database entries.

programming for determining the optimal alignment of disease states to hospital visits. The variation is significant, in that it uses a set of time constraints that are specific for the disease in question. From the calculated alignment, an age of onset bound (maximum and minimum possible) is predicted from a heuristic on the most probable alignment.

We believe that this work is a starting point to learning the relationships between DNA data and discharge data. Knowledge of how record linkage can occur between DNA and clinical information is established. Through knowledge of these relationships, privacy protection protocols for large DNA datasets can be designed. We find our approach to have 18% accuracy via leave-one-out cross-validation, on a test set of 22 individuals, for inferring an age of onset within a 5 year period. The accuracy is believed to be a residual from suboptimal estimation of emission probabilities. However, while the results are not promising, we stress the research in the constrained dynamic programming as a key contribution.

2. Computational Privacy Background

Record linkage is a useful tool in many fields, including health services, corporate, accounting, and marketing ventures (Fair & Whtridge 1997, Winkler 1995). Many datasets are released from one group to another, while anonymity of the entries is falsely assumed. Several computational systems have been designed that help render data anonymous. These systems have important relationships to this paper, as a utility that they use algorithms for discovering identifying information. These include methods that locate and remove textual information in images (Wang & Weiderhold 1998), Scrub (Sweeney 1996), which locates personally identifying information in unrestricted textual documents, and the Datafly (Sweeney 1997) and Mu-Argus (Sweeney 1998, Hunderpool & Willenborg 1996) systems, which attempt to render field-structured person-specific databases sufficiently anonymous. The previous algorithms have been paramount models for determining identifying features in static situations and situations where there is no interdependency within a dataset's tuples. However, this research does consider the relationships of multiple time points for an individual within a dataset. To understand the dataset that we work with in this research, it is helpful to understand how the data is preprocessed.

2.1 Constructing Longitudinal Medical Profiles

The state discharge databases are not provided as a compendium of individual profiles, but rather as a listing of every hospital visit in the state. Huntington's disease profiles were constructed based on unique demographic combinations of {5-digit ZIP, date of birth, gender}, the quasi identifier, associated with the hospital visit entries (Sweeney 2001). Figure 1 depicts the construction of the profiles. First, the databases were queried for visits with a Huntington's disease diagnosis. Then, the demographic combinations were used to requery the same databases to retrieve visits where the individuals were not diagnosed with the disease.

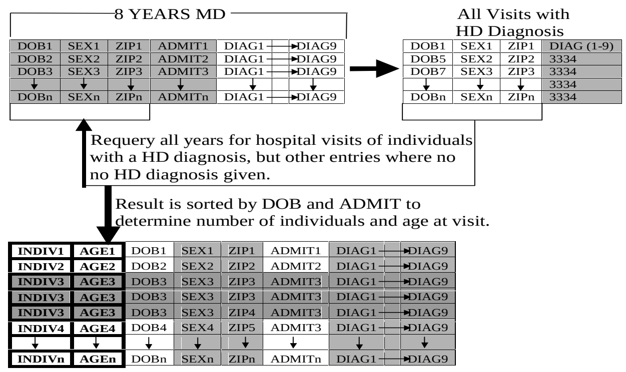

Figure 1. Constructing longitudinal medical profiles from state generated discharge databases.

The reasoning for the second querying is to retrieve hospital visits for individuals that were either 1) prior to onset of the disease or 2) visits in which they were not treated explicitly for the disease in question. A profile consists of the visits with the same triple demographic quasi-identifier.

In addition to the discharge database, a Huntington's disease age of onset registry was procured from Rush Presbyterian hospital in Chicago. The age of onset of Huntington's disease for the individual profiles was appended by linking on the quasi-identifier. Based on the age of onset and known progression of the disease, the stage of the disease was appended to each entry for a profile. For Huntington's disease, the expected length of time is approximately 5 years per stage as discussed in the clinical background (Quarrell 1999, Malin 2000). Furthermore, there is a linear progression of the disease. Once the patient reaches stage 2, remission to stage 1 is not possible.

3. Technical Approach

The methodology for inference of age of onset is a multi-tiered process. The goal is to assign a stage of the disease to use for age of onset inference. The stages are assigned through aligning the underlying stage of the disease to the observed visits in a dependent manner. Dependency of visits is based on the time elapsed from one visit to the next and a known basis on how long stages of this particular disease can persist. To generate the optimal alignment we use dynamic programming combined with the time constraints. However, in order to subject a profile to dynamic programming, we must represent each visit in probabilistic terms of being in any of the particular states of the system (the stages of the disease). The probability of each visit is determined in a naïve Bayes fashion, using only positive occurrences of observable features at each visit.

3.1 Estimation of Observable Features

For any individual, the probability that a particular hospital visit v_i from their longitudinal medical profile is in any particular stage of the disease is dependent on the observable features of that hospital visit. The features that we used for estimation were the diagnoses given at each visit, which are International Classification of Diagnoses-9 (ICD-9) codes. There are over 15000 different codes. Scaling down the number of parameters (diagnoses) was achieved by generalizing diagnosis codes. This strategy allowed for clustering of the codes into 151 code clusters. Figure 2 depicts the generalization strategy.

dx1	dx2	dx3	stage
E5001	3334	29010	2
3334	2345	29000	2

After Generalization

E50	33	29	2
33	23	29	2

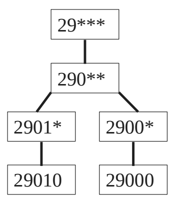

*Figure 2.*Generalization hierarchy for diagnosis (dx) coding.

Using a multi-variate Bernoulli event model, the hospital visit is considered as a binary vector over the space of diagnoses. We model the system a similar fashion to that used of text classification (McCallum & Nigam 1998). Given a set of diagnoses *D*, each dimension of the space *t*, $t \in \{1, \ldots, |D|\}$, will correspond to diagnosis d_t from the set of generalized ICD-9 codes. We consider only the positive occurrences of diagnoses in a hospital visit. The subset of all codes, the set of *k*, us used for estimation of the probability that a patient is in a stage $s_j \in [0,3]$ of the disease at a particular visit, given the diagnoses observed. For simplicity, we assume that there exists conditional independence between the diagnoses, also known as naïve Bayes, between the diagnoses in any particular hospital visit.

$$P(s_i = s_v \mid d_{i1}, d_{i2}, \ldots, d_{ik}) = \prod_{j=1}^{k} P(d_{ij} \mid S_v) \qquad (1)$$

The probability that the stage at visit *i* is equal to any of the stages of the disease is dependent on the diagnoses given at time *i*.

As a result of assuming conditional independence, we can consider a hospital visit as a collection of multiple independent Bernoulli trials and only considering those that are non-zero. This representation lacks any indication as to the estimation of the observed diagnoses. For these experiments, we have chosen to calculate Bayes-optimal estimates for the probabilities of observing diagnoses from the set of *D* (Vapnick 1982) in a particular stage. The estimates are supplemented by a Laplacean prior, by priming each diagnosis absence and presence count by a count of one. This ensures that we avoid probabilites of zero and one in the estimation of the distribution of all stages for a particular hospital visit. We estimate the probability of a diagnosis d_t in stage s_v as the frequency with which the diagnosis shows up in the set of all hospital visits *V* (not just for one individual, but for the entire cohort). Since we are only interested in the frequency of occurrence, we let β_{it} represent an indicator function $\in \{0,1\}$ to define an absence or presence of d_i within each hospital visit. To make estimates corresponding to each stage of the disease, we can only count hospital visits that correspond to the stage of the disease being estimated. We let $P(s_j|v_i)$ be the indicator function $\in \{0,1\}$ for the current visit stage label. $P(s_j|v_i)$ is equal to 1 if the desired stage is observed and 0 otherwise:

$$P(d_t \mid s_j) = \frac{1 + \sum_{i=1}^{|V|} \beta_{it} P(s_j \mid v_i)}{2 + \sum_{i=1}^{|V|} P(s_j \mid v_i)} \qquad (2)$$

The stage probability distribution for a particular hospital visit is calculated from equation 1. However, we can not use the max of naïve Bayes with respect to stage of disease as the classifier for each hospital visit. If we were do so, the result would yield impossible alignments of stages for longitudinal medical profiles. This analysis is discussed with the use of the relative frequency estimation of visit stages (Malin 2000). As discussed before, it impossible to have remission in Huntington's disease. Consider the profile depicted in Figure 3. The optimal independent alignment would yield:

Visit	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	v15
Pr Stage0	0.081	0.031	0.054	0.256	0.020	0.139	0.032	0.022	0.016	0.054	0.033	0.095	0.033	0.018	0.279
Pr Stage1	0.585	0.660	0.510	0.279	0.270	0.096	0.115	0.469	0.261	0.510	0.085	0.013	0.085	0.305	0.141
Pr Stage2	0.242	0.194	0.318	0.197	0.235	0.003	0.413	0.217	0.630	0.318	0.320	0.093	0.320	0.136	0.024
Pr Stage3	0.093	0.114	0.117	0.268	0.474	0.761	0.440	0.291	0.093	0.117	0.561	0.799	0.561	0.540	0.556

Figure 3. A longitudinal longitudinal medical profile with probabilities of states for each visit. The depicted alignment is the result of using stage 3 in the final state, and back-recursion without time constraints, except linear progression of the disease.

v_1	v_2	…												v_{15}
1	1	1	1	3	3	3	1	1	1	3	3	3	3	3

Clearly, there is a need for a need for inter-relationship between visits.

3.2 Time Constraints

To prevent impossible stage alignments for longitudinal medical profiles, we must utilize knowledge about the disease that we are considering. Huntington's disease consists of 3 stages; the early, middle, and late stages, and a prestage[2]. The stages occur in a linear progression towards death. Clinically, they persist for an average of 5 years each (not including the prestage), though there is some variability in the duration of any particular stage. Currently, there exists no real treatment or cure for Huntington's disease, and as such, it is impossible for an individual to proceed into remission or a previous stage. Based on this knowledge, we can construct a set of rules governing the time-dependency of the disease:

	Constraint
1	Final stage can not be Stage 0
2	4 years ≤ length (Stage 1 || 2) ≤ 7 years
3	If an individual skips over one stage: <table><tr><th>Visit</th><th>v1</th><th>v2</th><th>v3</th><th>v4</th><th>v5</th></tr><tr><th>Stage</th><td>0</td><td>1</td><td>3</td><td>3</td><td>3</td></tr></table> Then a minimum of 4 years between visits is necessary[3]

Table 1: Time constraints of Huntington's disease progression.

Rule 1 ensures, that at some point in the underlying state sequence, an onset of the disease does occur. Though this may seem like a trivial concept, it is necessary to rule out a disease stage sequence of all asymptomatic states (the prestage), despite the possibility that this stage may have the highest probability at each visit given the observed features. Rule 2 helps to provide some margin of error for alignments at stage transition boundaries. Though, we know that there is an approximate five-year span to each stage, there are many occurrences of shorter and longer durations of the disease (Quarell 1999). We do not explicitly provide any type of probability of transition within the allotted margin, but simply allow for a transition. The support for Rule 3 is based in the transition of a disease. If an individual skips from stage i to stage $i+2$, then the minimum number of years for the stage between the two stages must be accounted for. This must be considered in conjunction with Rule 1 to ensure that the patient has not spent a less than sufficient quantity of time in the stage i.

By adding time constraints, it is obvious that we are modeling more than a simple linear progression. Furthermore, the naïve Bayes estimation of the stage distribution at a particular visit is a poor classifier of the visit, due to the neglect of time constraints.

3.3 Time Constrained Dynamic Programming

It is necessary to model the unknown stages of a longitudinal medical profile (for Huntington's disease) with visits dependent on each other. The dependency from one visit to another is based on time. The time elapsed from one visit to another is an adequate feature for constraining the classification of individual's profile. Due to these constraints, it is no longer possible to tell what state the system (or profile) is currently in, simply by using a naïve Bayes classifier. Since visits are inter-dependent, we are interested in the sequence of underlying states, rather than the prediction of a single hospital visit in particular. In the speech recognition community, determining the underlying states from a sequence of observable features over time is known as decoding (Rabiner 1989). The most common decoding algorithm is known as the Viterbi algorithm (Durbin, et. al. 1998). It is necessary to have an understanding of this algorithm, since it is the basis for the incorporation of the time constraints into a time series model.

Consider a longitudinal medical profile with four hospital visits, each visit a year apart from the previous. At each of these visits, there is a set of diagnoses observed. We consider the sets of diagnoses as D_1, D_2, D_3, *and D4*. Many different state sequences could have given rise to the observed sets of diagnoses. For example, the state sequences could have been generated from the underlying states (1,1,1,1), (2,2,2,2), (2,3,3,3), or (1,1,2,2), though there would be many more possible alignments of the sequence. The difference between these alignments is that they would emit the diagnoses at each visit with different probabilities. We are interested in the path through the sequence alignment that provides the highest probability. In other words, we are interested in:

$$\pi^* = \arg\max_{\pi} P(D, \pi) \tag{3}$$

We can generate the probability $v_k(i)$, which is the most probable path ending in state k with observation i. The most probable path can be found recursively through posterior decoding of the alignment. Posterior decoding can be reasoned as determining what the most probable path is by considering the final state to be fixed as k. We define the time- constrained dynamic programming (TCDP) algorithm, which can be defined in a three-step process: initialization, back-recursion, and iteration. First,

[2] In medical terms, this is known as the asymptomatic period. With a DNA based diseases, the patient has been diagnosed through genetic testing, however, there has yet to be onset of disease.

[3] An extension of this rule is easily derived for a skip of 2 stages. In such an event, a minimum of 8 years is necessary between visits.

we define, an array A whose size is equal to the length L of a longitudinal medical profile.

Step	**Process**
Initialization	$p(L) = p(D_L\|S_k)$; k = max k $A_k(L)=s_k$
Back-recursion (i begins at L-1)	$p(i) = \max p(D_L\|S_k)t_i$; until I = 1 $Sk(i) \geq S_k(i+1)$ $A_k(i)= s_k$
Iteration	k-1, until k=1;

Table 2: TCDP algorithm.

The above algorithm defines an array for an alignment with a final state corresponding to each viable end state. In the back-recursion step, we multiply the probability of the observable features given a state by an indicator function t_{kl}. This indicator function is 1 if the presence of stage at time i is a legal stage for the alignment given the time constraints previously defined, and 0 otherwise. As equation 3 denotes, the most probable state sequence will be the max of the probability of the path defined for each A_k. The max probability can be found by walking forward through the path and multiplying the probabilities for the diagnosis set given the stage at $A_k(i)$.

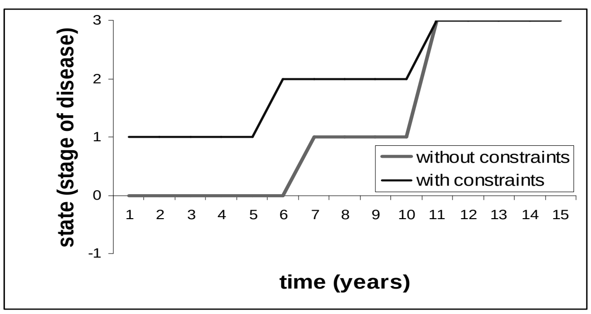

Figure 4: Comparison of dynamic programming with and without time constraints.

Consider the following scenario. A profile consists of 15 visits, each with a time elapse of one year. The probability distributions are listed as in figure 3. The optimal alignment using a the Viterbi algorithm would be:

v_1	v_2	…												v_{15}
0	0	0	0	0	0	1	1	1	1	3	3	3	3	3

However, using TCDP, the alignment is:

v_1	v_2	…												v_{15}
1	1	1	1	1	2	2	2	2	2	3	3	3	3	3

The TCDP alignment seems to fit the data much better than simple dynamic programming. This can be more easily seen a graphical representation, as in figure 4. Though, both of the alignments finish is stage 3, the time constraints provide a more realistic transition through states of the disease.

3.4 Age of Onset

Prediction of the age of onset uses the aligned medical profile to estimate a range to the age of onset. We employ a heuristic based on the expected progression of the disease. The heuristic relates to the expected length of time that an individual resides in a stage of the disease. For Huntington's disease, the expected length of time is approximately 5 years per stage as discussed in the clinical background. (Quarrell 1999) Thus, to predict the age of onset, the stage with the most visits is utilized[4]. In the event that there are multiple stages each with the max number of visits, the stage closer to the age of onset is used as basis for the estimate.

The estimation procedure uses the first visit where a stage transition has occurred, or may have occurred. Some individuals reside in only one stage throughout a profile. In such a situation the first visit is taken to be the threshold between the stage observed and the previous stage. The number of years corresponding to the threshold is subtracted from the current age of the individual at this visit. The second bound to the age range is created from the last observed visit in the stage. A visit may be both the first and last observation. In this situation, the largest age of onset bound (five years) is guaranteed. At this visit, the next stage up threshold is assumed or observed and the correlated number of years are subtracted from the current age of the individual.

4. Results

Data used for this study included hospital discharge data (called health data) for the state of Illinois from 1990 through 1997. In many states, such datasets are publicly available. (Sweeney 1997) There are approximately 1.3 million hospital discharges per year in the health data, which reportedly corresponds to hospital compliance of 99+% for all discharges occurring in Illinois hospitals. (NAHDO 1998) Diagnosis codes (ICD-9), procedure codes, patient demographics and hospital identity are among the information stored for each visit. Previous research has demonstrated that publicly available discharge data permits the formation of genetic population subsets. (Malin & Sweeney 2000)

4.1 Matching Stages in Profiles

In the dataset, we could perform record linkage to append the age of onset for 22 individuals in an initial cohort of over 400 patients diagnosed with Huntington's disease.

[4] Though not shown, we find that the most prevalent stage provides a more accurate representation of the age bounds, than using the complete alignment.

The number of hospital visits for the 22 individuals totaled 69 hospital visits. The mean number of hospital visits per profile was approximately 12 visits, with a standard deviation of 1.8896. The mean observation time of a total profile was approximately 1.5 years with a standard deviation of 3.4128 years. Due to the small sample size, we employed leave-one out cross validation as explained in Mitchell (1997). This entailed, estimating the emission probabilities at each stage based on the frequencies of the features in all profiles except for the profile that we were currently predicting. We circumvented the occurrences of zero and one probabilities diagnoses in a certain stage with a Laplacean prior as explained above.

The main goal of this research was to estimate the age of onset of the disease. However, in order to do so, we needed to know the sequence of underlying disease states for each profile. The time-constrained dynamic programming approach was able to match underlying stages in 32% of the total profiles, or 7 of 22 profiles. Of the profiles that actually had matches within them, 38% of the progression was correctly predicted. Examples of partially matched and unmatched profiles can be seen in figures 5a and 5b.

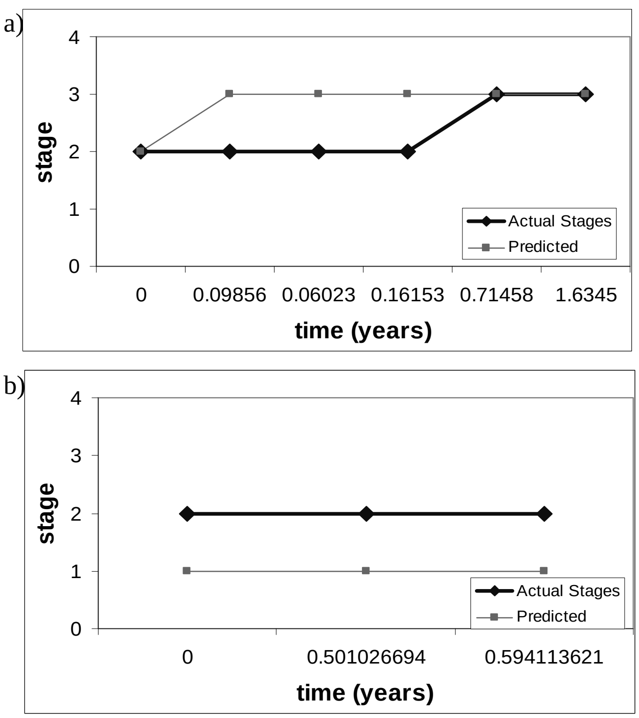

Figure 5. Samples of a) a partially stage-matched and an b) unmatched longitudinal medical profile.

The underlying marginal distribution of stages in hospital visits for the 22 individuals was {4,33,19,13} for the stages {0,1,2,3}. The most prevalent stage was stage 1, which accounted for approximately 48% of the dataset. If we were to use the most prevalent stage as our classifier, then it is clear that this would have a higher accuracy in terms of inferring the stage of longitudinal medical profiles. Using the most prevalent stage, we achieve matches in 45% of the profiles. Though, this may seem true, it must take into account that stage inference is the penultimate goal of this research. Knowing the stages of the underlying disease, is important, however, the main goal is to determine the age of onset. This is where privacy may be compromised, and the thus must be accounted for.

4.2 Age of Onset Inference

Using the heuristic as described above, we attempted to infer the age of onset for the profiles. Through this heuristic, the largest age of onset bound (max age minus min age prediction) that could be derived was a period of five years. The average size of the bound was 3.77 with a standard deviation of 1.51. Approximately 18% (4 of 22 profiles) of the profiles were correctly bounded, as can be seen in Figure 6.

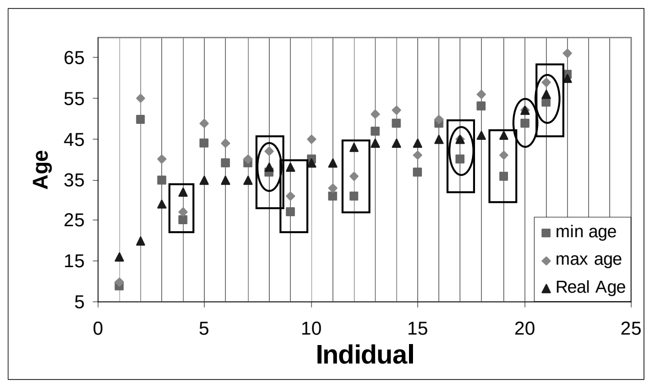

Figure 6. Age of onset bounds for a Huntington's disease cohort of size 22. The correctly bounded profiles are within ovals. Profiles within rectangles were correctly bounded by the uniform classifier.

Though the accuracy of our methodology is not high, it is better than random. This can be justified by considering the possible range for disease age for onset. Let us consider that the majority of Huntington's disease onset ages are distributed over the ages of 5 to 70. Thus, there exists an age of onset range of 65 years. If one was to randomly assign ages of onset, we would expect accuracy to of onset age prediction to be the inverse of the range. However, the methodology that we use has a much higher accuracy than the random expectation.

If we were to employ the uniform classifier again (the use of the most prevalent stage), there are several problems. First, it is not sensitive to time dependency. Second, it will always bound on the first stage. Yet, despite these problems, the accuracy is still greater than that of the methodology presented in this paper. The accuracy of the uniform classifier was approximately 32%. The correctly bounded profiles are shown on also on Figure 6. However, such results should be taken with some cynicism. First of all, the dynamic when aligning profiles that span longer periods of time. The low accuracy of the algorithm on this test set, may simply be the artifact of the small total time elapsed for each patient profile. Second, the incorrect bounding may be the result of poor estimation of the observable feature distribution. Both of these possibilities are examined in more detail below. It should also be noted, that the uniform classifier does not correctly bound all of the same samples as the dynamic programming algorithm. This is further proof, that the methodology presented in this paper has the ability, given the correct emission probabilities, to be a useful inference technique.

5. Related Work

Dynamic programming (DP) has been useful in many fields, most notably speech recognition and nucleic acid sequence analysis (Rabiner 1989, Bucher, et. al. 1996). Until this research, DP had not been useful for privacy research or longitudinal medical data analysis. Of the latter, there has been some work with longitudinal medical data (Hauskrecht & Fraser 1998). However, most of the medical informatics community has focused on prediction of future events based on previous observerations. Many researchers are interested in how to augment current treatments through use of partially observable Markov decision processes (POMDP), another type of time dependent analysis. We believe that, this research is notable for medical informatics in that it uses known information to add important constraints to the modeling of time series medical data.

Though this research is important for medical informatics research, it is mostly the result of a need in the privacy community for an understanding of the relationships between clinical and DNA database entries. Privacy researchers have concentrated their efforts in using record linkage techniques for determining if relationships exist between seemingly autonomous datasets. Most privacy researchers, tend to concentrate on readily available identifiers in datasets or combinations of several attributes, such as {date of birth, gender, and zip code} (Sweeney 2001).

While the previous yields highly accurate results for some datasets, there is one issue that seems to have several issues that have been left unaddressed. First, what happens when the attributes between two datasets do not match each other? In recent work, we have demonstrated that one can utilize the audit trails of patients in different hospitals to establish record linkages (Malin & Sweeney 2001). The relationship between the age of onset of disease and mutation size is not considered. Furthermore, our previous research assumes that we can track DNA entries across multiple hospitals.

Thus, we prove through this research and the previous that different datasets do not necessary have equal attributes in order to be matched. Yet, this begs a second question that has been shirked by the privacy community. Can record linkage occur when two datasets have uncommon attributes, and one of the datasets is localized to a single location. This is precisely the situation that this current research attends to. While the attributes do not explicitly relate to each other, we can infer additional information from each dataset to generate a new common attribute. In the Huntington's disease situation, the age of onset can be inferred from the size of the DNA repeat (Brinkman). Additionally, we show that there exist a methodology to infer the age of onset from clinical datasets. The main contribution of this research is the incorporation of time dependency into a time series analysis model to establish this inference.

6. Summary

Given medical data over a known time period, this research demonstrates that additional information can be inferred from the dataset. We utilized dynamic programming with time constraints to establish this inference. Inference of the information, which was the age of onset of Huntington's disease, can be used to link clinical datasets to DNA datasets. We found 18% accuracy of bounding the age of onset within a five year time period, which is greater than random prediction of disease's age of onset. We believe that this research has profound impact in the privacy community. Due to our results, we recommend that care must be taken before the release of population-based DNA datasets for public (or outside the originating institution) research and scrutiny.

The accuracy of our results leaves much room for improvement. It seems that a majority of the error in the age of onset bounding is the result of poor estimation of the emission probabilities for the observable features. This research made the assumption that diagnoses at each hospital visit are conditionally independent. Yet, it is a highly plausible situation that, indeed, this is not the case. Diagnosis 1 and diagnoses 2 may possess a uniform distribution over all stages, and yet the probabilities of seeing the two together in any particular stage may grievously deviate from the conditional independence assumption. If this is true, then it would be much easier to determine the underlying state, due to a more accurate observable feature. Second, clinical studies have demonstrated (Quarrell 1999) that certain diagnoses have a dependence of the disease stage. While, we attempted to capture this property in our estimation of the observable probabilities, it might be more useful to not use a flat prior. It could be of great use to design a knowledge base to boost the priors and incorporate a more semantical structure to the relationships of observable features to the underlying state of the system. Third, our estimation of emission probabilities was greatly hindered by the size of our cohort. It is necessary to have a large dataset to determine if our error is the result of poor estimation of the emission probabilities. A larger dataset is readily available, though it is devoid of the age of onset target information. It might be a feasible solution to attempt a learning scheme in labeled and unlabelled data to boost the available information of observable features.

Finally, we believe that though dynamic programming helps to provide an accurate model of time dependency. However, it is apparent that a more generative learning model is necessary. There are many time-dependent diseases that could be modeled in a similar manner to that of Huntington's disease. Yet many of them including, different types of cancer and neural-based diseases have a progression, as well as remission. The use of dynamic programming, similar to that of the Viterbi algorithm, suggests that a hidden Markov model would be the best type of computational learning model for this system.

Acknowledgements

The author thanks the State of Illinois for the use of their data. Further thanks are extended to Latanya Sweeney and Sebastian Thrun for insightful comments during the research and evaluation period of this work.

References

Brinkman, R.R., Mezei, M.M., Theilmann, J., Almqvist, E., & Hayden, M.R. (1997) The likelihood of being affected with Huntington disease by a particular age, for a specific CAG size. *Am. J. Hum. Genet. 60. 1202-1210.*

Bucher, P., Karplus, K., Moeri, N. & Hofmann, K. (1996). A sequence similarity search algorithm based on a probabilistic interpretation of an alignment scoring system. In States, D.J., Agrawal, P., Gaasterland, T., Hunter, L., & Smith, R.L. (Eds.), *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology.44-51*. AAAI Press.

"Data release overview." (1998). *State of Illinois Health Care Cost Containment.* Springfield.

Durbin, R.., Eddy, S., Krogh, A., & Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge. Cambridge University Press.

Fair, M., & Whtridge, P. (1997). Record Linkage Tutorial. *Record Linkage Workshop. Arlington, VA.*

Hauskrecht, M., & Fraser, H. (1998). Modeling Treatment of Ischemic Heart Disease with Partially Observable Markov Decision Processes. *Proc AMIA Symp 538-542.*

Hundepool, A., & Willenborg, L. (1996). μ- and τ-argus: software for statistical disclosure control. *Third International Seminar on Statistical Confidentiality*.

McCallum, A., & Nigam, K. (1998). A comparison of event models for naïve Bayes text classification. *AAAI-98 Workshop on "Learning for Text Categorization"*.

Malin, B. (2000). Inferring disease age of onset through computerized medical record systems: A Huntington's disease approach. *Carnegie Mellon University.* Unpublished.

Malin, B.. & Sweeney, L. (2000). Determining the identifiability of DNA database entries. *Proc AMIA Symp*. 537-541.

Malin, B. & Sweeney, L. (*under submission* 2001). Re-identification of DNA through an automated linkage process.

Mitchell TM. (1997). *Machine Learning.(pg. 235)* WCB McGraw-Hill. New York.

National Association of Health Data Organizations. (2000). *NAHDO Inventory of State-wide Hospital Discharge Data Activities* (Falls Church: National Association of Health Data Organizations).

Quarrell, O. (1999). *Huntington's Disease: The Facts*. New York: Oxford University Press.

Rabiner L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE. 77(2).257—286.*

Sweeney, L. (2001). Simple Demographics Often Identify People Uniquely. *Heinz School Working Paper*. Carnegie Mellon University.

Sweeney, L. (1998). Three computation systems for disclosing medical data in the year 1999. Medinfo. 9 (Pt 2). 1124-1129.

Sweeney L. (1997). Guaranteeing anonymity when sharing medical data, the datafly system. Proceedings, JAMIA. Washington, DC. Hanley & Belfus, Inc..

Sweeney L. (1996). Replacing personally-identifying information in medical records, the scrub system. In: Cimino, JJ, ed. Proceedings, JAMIA. Washington, DC (pgs 333-337) Hanley & Belfus, Inc.

Timchenko, L.T., & Caskey, C.T. (1996). Trinucleotide repeat disorders in humans: discussions of mechanisms and medical issues. *FASEB J. 14.1589-97.*

Vapnick, V. (1982). Estimation of dependencies based on statistical data. New York. Springer Publisher.

Wang, J.Z., & Wiederhold, G. (1998). System for efficient and secure distribution of medical images on the Internet. *Proc AMIA Symp. 907-11*.

Winkler, W.E. (1995). Matching and Record Linkage. In *Business Survey Methods*. New York. John Wiley and Sons.