HIDDEN MARKOV MODELS FOR INFERENCE IN LONGITUDINAL MEDICAL INFORMATION

Bradley Malin
malin@andrew.cmu.edu
Carnegie Mellon University
School of Computer Science
Heinz School of Public Policy and Management

Abstract: Privacy of identity is an important and complex issue in the process of sharing datasets between data holding and receiving agencies. In the past, research has demonstrated that identifying features may be found and removed in text documents and entry-specific databases. However, the previous research fails to address data sets that have collected information about individuals over time. This project proposes research for exploring a method for inferring information from a collection of an individual's hospital visits over time, known as longitudinal information. Certain diseases have a DNA component that determines the age of onset of a disease, such as Huntington's disease. DNA data and clinical information are stored in separate databases, and thus they are seemingly non-relatable. If we could infer additional information that is not associated with one of the datasets and relate it to the other, then we could link entries in the two databases. Publicly available hospital discharge data contains information, such as diagnoses for an individual, which may be utilized to infer additional information about a disease with a known DNA factor. However, the age of onset is not listed in the hospital discharge data. By using known information about the disease, we will consider the disease to be time dependent, as a series of different stages of the disease. This research focuses on a hidden Markov model (HMM) for determining the unobserved stages of the individual's sequence of hospital visits over time. The time dependency of these stages is utilized to determine an underlying age of onset for each individual with Huntington's disease. We hope to demonstrate that though the identities of entries in databases seem anonymous, additional information may be inferred to compromise patient anonymity in a sensitive dataset, such as a DNA database.

1 Introduction

1.1 Problem

This proposal deals with the issue of data privacy. One of the most sensitive data sharing situations exists in the medical community, where information may influence an individual's credit, employment, insurance, as well as bring about stigmatizing circumstances. Different types of medical information are collected for various reasons. For example, the National Association of Health Data Organizations (NAHDO) reported that 44 of 50 states have legislative mandates to gather hospital-level data on each patient visit. [Nah00] 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 been shown that matching such databases with population registers can link names to 87% of individuals in the United States (216 million of 246 million people) [Swe01]. Another type of medical dataset that is being compiles is DNA sequences. These collections are location specific for many monetary and legislative reasons. When releases of these datasets are made available, the only demographic information associated with these databases is the knowledge of the originating institution. Yet, there are inferences that can be made in both datasets to link the seemingly autonomous entries. This research looks into the one of the types of inferences that can be made to link publicly available medical information and DNA database entries. This problem allows for innovative study in the field of data privacy for several reasons:

1. **Longitudinal study.** Unique combinations of information associated with each hospital visit, such as demographics permit the construction of individual profiles over time. These profiles may range over many years of information. Medical information about an individual over time is referred to longitudinal medical information. It will be referred to as such for the remainder of this proposal.
2. **Hidden information.** The linking of datasets may be accomplished by crossing of the information via common attributes. If no common attribute exists, then one may be inferred from the existing data. Such "hidden" information may be derived from data associated with a specific hospital visit or from a longitudinal profile.
2. **Novel time series model.** The profiles permit a time series model and analysis of the profiles.

This particular study focuses on determining "hidden" information that can be inferred from publicly available hospital discharge data that can be associated with DNA datasets. We will use Huntington's disease as the main study since it has a well-defined clinical literature. [Qua99] Furthermore, the relationship between DNA mutations and the age of onset of Huntington's disease is well known. [BM97]

1.2 Deliverables

The core deliverable of this work is an algorithm for modeling diseases in longitudinal information resulting from known DNA mutations. The model will show that relationships can be established between autonomous databases, and one of the databases has been re-identified, linkages can infringe on individual's privacy. The research, while oriented toward demonstrating privacy problems, has usefulness in other areas as well. Directly stemming from this research will be practical medical techniques. First, there will be the ability to infer additional clinical information from a medical record for a more detailed analysis of the history of a disease. Second, is the ability for molecular biology researchers to predict the sequence of DNA that corresponds to clinical information.

1.3 Practical Impact

The large compilations of medical information are valuable resources for researchers and a source of monetary income for the data holding institutions. Currently, the privacy of an individual may be

compromised by re-identifying the individuals in a database. This may be done by creditors, insurance agencies, or employers. All of these companies would be using legal means to establish a more complete profile on their clients or employees. Though it is illegal to use this information in such a manner, it is difficult to prove that such illegal discrimination has occurred. By demonstrating where compromising of privacy can occur in such datasets, the privacy community can begin research on how to prevent these inferences from occurring. The data should be made available for research purposes. Why else collect it? It is necessary to ensure the privacy of individuals in future releases of the collected information.

2 Prior Research

Record linkage is a useful tool in many fields, including health services, corporate, and marketing ventures [FW97, Win95]. 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 are important because they use algorithms for determining identifying information. These include methods that locate and remove textual information in images [WW98], Scrub [Swe96], which locates personally identifying information in unrestricted textual documents, and the Datafly [Swe97] and Mu-Argus [Swe98, HW96] 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.

So, where does this research differ from that of the previous? The data that the above programs work on are static in that single hospital visits, or banking records, or whatever the data being studied are single event occurrences. However, this research will determine identifying features from a longitudinal dataset. This increase in dimensionality permits additional inferences about the individuals being studied. In the current techniques, identifying information is searched for based on combinations of demographics, which can be linked to other datasets, using the same fields. Yet, this research uses information that is not directly observable, such as how certain combinations of diagnoses can represent a certain stage of disease.

This research is not just localized to identification of information, but it must be tied to inference models for medical data. There have been many prediction models used with medical data. Most similar to this research is the applied work of partially observable Markov decision processes (POMDPs) [Sin92, Pee99]. The algorithm has been shown to have merit with management of patients with ischemic heart disease [HF98]. The POMDPs utilizes the believed state of the disease and the treatments rendered over time. Thus, future treatments are predicted based on observations seen and past actions performed.

Our prior research has demonstrated that publicly available hospital discharge data permits the formation of genetic population subsets. [MS00] The subsets are based on international coding diagnoses (ICD-9) and gene mutation to disease presentation derivatives. From these subsets, longitudinal profiles may be constructed of individuals over several years of information [Mal00]. Time series analyses with defined stages have been studied in speech recognition and text segmentation by many researchers. Most notably has been the research in hidden Markov models (HMMs) [Rab90, Yam98], which has been utilized in speech recognition. HMMs are useful for predicting patterns and relating associated states to one another. A variation on HMMs has been useful for incorporating expert systems with a number of independent, modular knowledge sources to communicate in a common environment, as in HEARSAY-II [EHLR80].

The methods of this project will incorporate the variations of HMMs to design new algorithms for inference. Building upon the above-mentioned architecture, this research will utilize constraint-based structures to define time-series relationships with completely obscured states, unlike the POMDP research. In addition, hidden information will be inferred from a rule-based model. This research is novel in that inference of hidden medical data has not been attempted from text based clinical information alone.

3 Approach

The main premise of this project is to design and implement methods for inferring information from longitudinal medical information. The approach is based on studying observable features, such as diagnoses given at particular hospital visits, of publicly available clinical information to determine additional information about the medical condition of an individual. Additional information, such as the age of onset of a disease, is a feature that permits a relationship between hospital discharge data and DNA data to be established.

3.1 Prior Work: Construction of Longitudinal Medical Profiles

To determine the feasibility of analyzing longitudinal profiles, research was conducted to determine if features of the profile could be used as attributes for inference. The discharge databases are not provided as a compendium of individual profiles, but 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} associated with the hospital visit entries [Swe01]. 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 demographics. 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. [Qua99] Furthermore, there is a linear progression of the disease. Once the patient reaches stage 2, remission to stage 1 is not possible.

Our previous research utilized the frequency of combinations of diagnoses per stage to determine the best stage. [Mal00]. Assignment of the stage was based on the diagnosis combination with the highest relative frequency from any of the combinations of diagnoses that occur in the specific hospital visit. Stage assignment assumed independency of hospital visits, thus there were occurrences of inconsistencies in the predictions. For example, there were situations where an individual proceeded from stage 2 in one hospital visit to stage 1 in a later one. A type of "smoothing" of the predicted stages was performed by using the predictions with the least variance in the frequency of the diagnosis combination. A heuristic was employed to bound the predicted age of onset. The number of years corresponding to an observed stage transition 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). At this visit, the next stage up threshold is assumed or observed and the correlated number of years was subtracted from the current age of the individual. Leave one out cross-validation of the method showed of the research showed that 82% of the test cases were correctly bounded in an age range of five years or less. However, leave-one-out cross-validation showed that approximately 30% of the cases were correctly predicted. Thus, while we believe that there is a relationship between the diagnoses and the stage of disease, relative frequency with a rough heuristic is not the best methodology.

3.2 Constrained Hidden Markov Model

Rather than incorporate a method for "smoothing" the seemingly incoherent stage assignments in profiles, a probabilistic algorithm would be more useful. Thus, the hidden Markov model (HMM) would allow for incorporation of different stages of a disease emitting the same observable features, the diagnoses. The depiction of this model is shown in Figure 1. The ability for the disease to transition from one stage to the next or remain in the current stage is denoted by a_{ij}, where i is the current stage and j is the next stage observed. For the disease that we are considering, let us assume that we know how many N states there are. The states are denoted as stages of the disease. Of these, N is equal to 4 different stages in Huntington's disease, which will be the disease model of choice for this research.

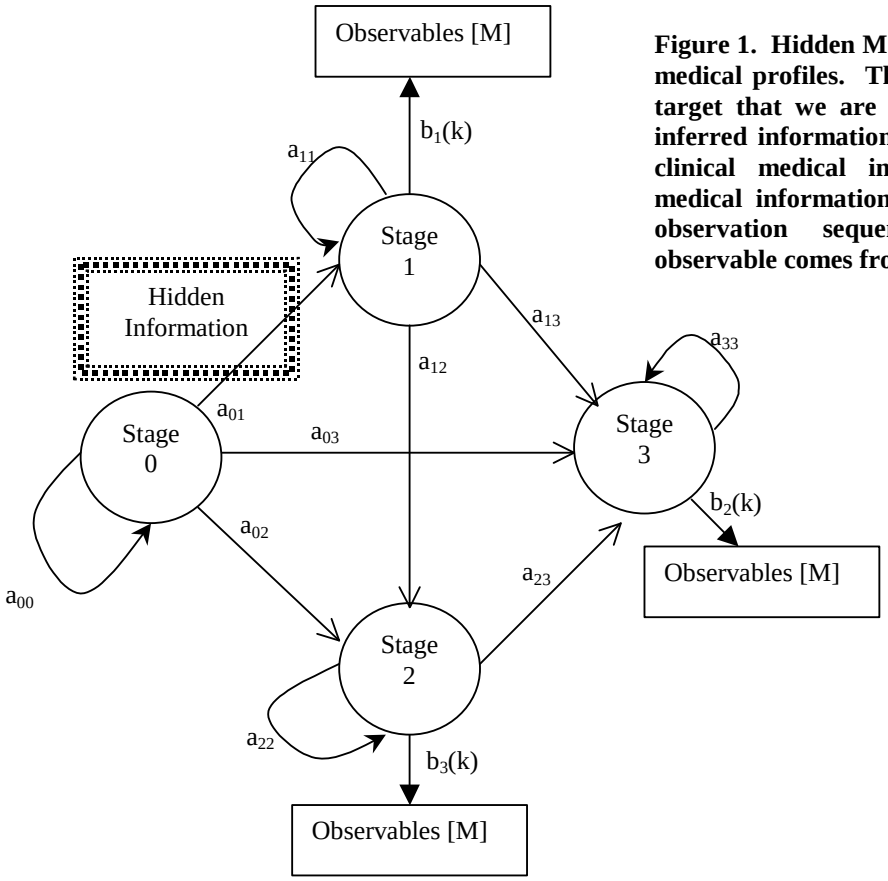

Figure 1. Hidden Markov model for longitudinal medical profiles. The hidden information is the target that we are interested in. This type of inferred information is an identifier that relates clinical medical information to DNA/genetic medical information. To infer this target, the observation sequence (which stage each observable comes from) must be determined.

The underlying model is a discrete Markov process. Once disease onsets, an individual proceeds from one stage into the next until death, though an individual does not necessarily achieve the final stage before death. The stages are the actual progression of the disease, however, this is not necessarily the observed case. It would be an easier situation to model, if we could monitor an individual daily. Yet, we do not have a continuous system and the individual is not under constant monitoring. Rather, we are only able to observe an individual when they make a visit to a hospital. This being the case, the hidden Markov model will account for the skipping the expected next stage. The depiction in Figure 1 is actually a rough picture of how the hidden Markov model will be set up.

The observables are distinct symbols V drawn from a total set of M observables. The distinct symbols may combinations of symbols or individual symbols from the set. For instance the generalized diagnosis code {33*} (* corresponds to infinite size wildcard) is an observation from the total set M, while {33*, 12*} is also considered an observation from the set. Additional diagnoses do not represent additional observations. Rather the subset of observations from the total set correspond to an observable V_i. The total set of observables is common to all stages. Thus, when a subset of observables are used, there is a relationship between all stages.

Each state j has its own emittance probability distribution of observables $B = \{b_j(k)\}$, where

$$b_j(k) = P[v_k \text{ at } t | q_t = S_j] \quad ; \quad 1 \le j \le N, \quad 1 \le k \le M$$

and q_t is equal to the state at time t. The probability distribution may be represented in a relative manner. Instead of using an absolute probability distribution, one can use a relational aspect of the information, which is the main reason why relative frequency was used in the prior research. This would allow for a

depiction of which observables were more likely to have been emitted by which stage. Determining the metric by which the emittance probabilities are drawn from is discussed in the next section.

The initial state distribution $\pi = \{\pi_i\}$ is defined as

$$\pi_i = \Pr[q_1 = S_i] \quad ; \quad 1 \leq j \leq N$$

Thus the initialization of the model will be dependent on determining which state was the most probable at the first time point of the model. The most probable state is determined based on the observables and their relative emittance probabilities as stated above, yet, it should be noted what the HMM will achieve. Given the observation of a sequence $O = O_1O_2 \ldots O_T$, (as in Figure 2) the model uses the associated probability distributions to predicted the most probable corresponding to the state sequence $Q = q_1q_2 \ldots q_n$, where q are states at given time points 1 through n. It is not the HMM that predicts the hidden information. Rather the HMM allows for the most optimal stage progression patterns to be established.

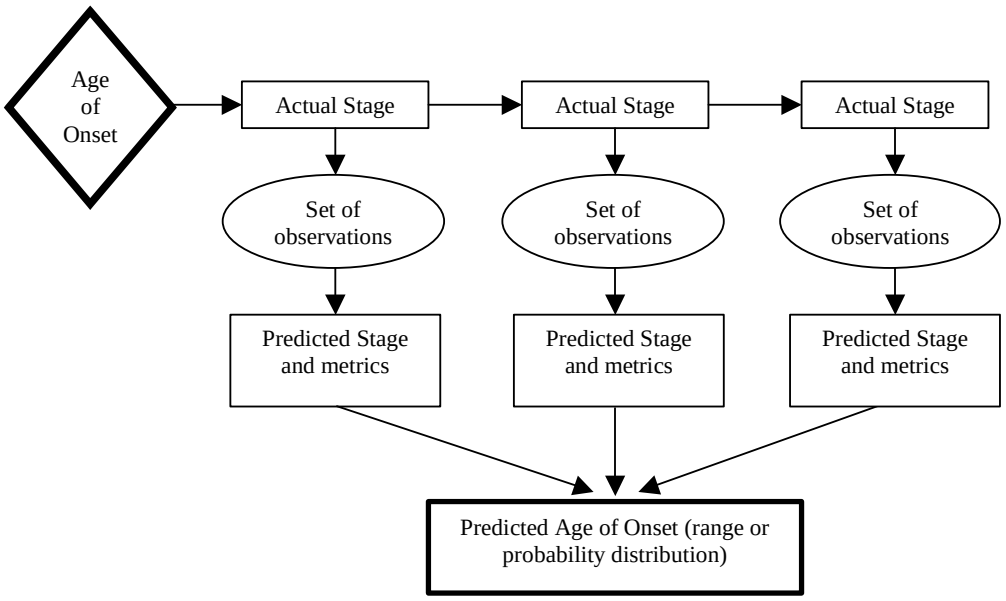

Figure 2. Model of presentation and inference strategy. Observables are collected at various time points. The actual state of the disease is unknown to the individual making the observations. The state may be inferred from the collected observations over time. The actual states can be further related to some aspect of the disease that is not directly observable. This emit source may be an age of onset of the disease for Huntington's disease or the clinical type of disease for cystic fibrosis.

The states at each HMM observation are not independent of each other. There are constraints on the model. The constraints are dependent on the disease that is being represented by the system. For Huntington's disease, the main constraint has to do with the progression and treatment of the disease. Huntington's disease is a disorder without major treatments. Once the individual has reached onset age of the disease, the patient will pass through linearly from stage 1 to stage 3, provided death does not intervene. Due to such constraints, the Viterbi algorithm is an ideal choice for modeling the progression from state to state of model. [Rab90]

3.3 Determining the Emittance Probability Distribution

In previous research, the probability for observables in each state was determined from an m-estimate frequency analysis and a variance measure. The theory of m-estimate is described in Mitchell [Mit97]. The application and in-depth explanation of how this frequency estimate was attempted can be found in

the prior work [Mal00]. The frequency estimate was determined by scoring the number of times an observable, in this situation a subset of diagnoses *V*, is detected in each possible stage of the disease. Associated with each of the frequency estimates are variance measures for each stage.

M-frequency estimation is one possible way to proceed. However, there are many other techniques for estimating the probability distribution. It should be noted, that to get an accurate probability distribution, we must assume conditional independence of the diagnoses at each hospital visit. It is not that the stage of the disease causes diagnosis A, which in turn causes diagnosis B. Instead, both diagnosis A and diagnosis B are caused by the stage of the disease. This assumption of conditional independence permits all diagnosis at each visit to be used in the prediction of the hidden stage. No stage is represented as the definitive stage. The training set for the model is relatively small, only 22 profiles. Out of these, the known stages were calculated from given ages of onset (hidden target). An example of diagnosis frequency is listed in Table 1. If we use the conditional independence assumption, then we may end up with probabilities of zero confounding our estimates of a certain stage. To prevent this from occurring, each diagnosis will be primed with a Laplacean prior of 1 positive observation for each diagnosis.

	Actual Number of Observations			
Observable	Stage 0	Stage 1	Stage 2	Stage 3
V_1	0	1	0	3
V_2	0	2	6	3
V_3	0	0	4	0

Table 1. Actual number of observations for listings in table 1. The underlying distributions of the sample were: Stage 0: 3 ; Stage 1: 38 ; State 2: 20 ; Stage 3: 8

The lower the number of the frequency metric, the less probable that this observable is a predictor of that stage. Similarly, the greater the variance with relation to the underlying marginal distribution of stages (the known hidden stages at each hospital set in the training set) denotes that a certain diagnosis is less probable for classification to a specific stage. Thus, while an observable may have been exhibiting the most probable frequency in stage 1, it may have the least amount of confidence in this stage. Such an occurrence could be due to a situation such as an observation being observed 1 in 3 occurrences of stage 0, as well as 1 of 19 occurrences in another stage. The HMM must be represented with a defined method for determining emittance probabilities.

3.4 Inferring the Hidden Target and Evaluation

The main goal of this research is to optimize the model for determining the states of a disease with a time dependent presentation. However, this project is part of a larger methodology for inferring an information linker between clinical medical information and DNA medical information. This part of the research can be achieved before optimal HMM parameters are determined. However, the predicted hidden information is a bounded maximum and minimum range for the age of onset. In previous research, prediction of the age of onset used the smoothed longitudinal medical profile to estimate a range to the age of onset. It utilized 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. This age of onset is predicted from the most prevalent stage, the one with the most entries in a profile. Future research will use all entries, however, for this research, the heuristic will continue to be utilized.

Basically, what the previous means is that the stage with the most observations will be considered the most accurate. As the most stage, the first observation of that stage will be considered the beginning of that stage. The final observation of the used stage will be considered to be the final observation of that stage. The expectation number of years from the age of onset, as explained in the literature, will be subtracted from these bounds to make up the range for the target.

It is difficult to compare the accuracy of the results from this research with other techniques. The model from this project is novel and does not propose a new look on an old problem. As a result, cross-

validation will be used as analysis for determining the accuracy of the research methodology. Furthermore, as a result of the small sample size, leave-one-out cross-validation will most likely be the method of choice.

4 Work Plan

The multi-tiered aspect of this project allows for a relatively coherent break down of time allocation. There are approximately seven to eight weeks of time available for research. The work plan has been broken down into two week periods. The tentative work plan will proceed as follows:

- *Weeks 1 and 2:*
  - Before the Markov model can be implemented, the statistical aspects of the emittance probabilities must be attended to. This entails determining the optimal method(s) of combining the frequency with the variance estimates for each observable at each stage.
  - The dataset that I am currently working with is relatively small. Thus, I will be attempting to procure more robust sets or information to create more profiles from several institutions, including Indiana University, Rush Presbyterian Hospital in Chicago, and University of Chicago.
- *Weeks 3 and 4:* At this point, the emittance and initial estimates of transition probabilities associated with the discretized Markov model will be defined.
- *Weeks 5 and 6:* Automation of the algorithm will be written as a JDBC (Java Database Connectivity) application. The application will allow for re-esimation of the transition probabilities of the smaller states within stages via optimization of the Viterbi algorithm
- *Weeks 7 and 8:* The previous weeks will yield results of different methodologies and parameters of estimation. Thus, this period will be used for analyzing the results of the inference strategies. Cross-validation and other model analysis methods will be invoked. It will be determined which methodology minimizes the bounding on the prediction of the hidden target.

5 Summary

The health care industry spent an estimated $10 to $15 billion on information technology in 1996 [Mun96]. The goal of this investment is to facilitate creation of an electronic patient record, which will increase the availability of patient information for swifter delivery of care and insurance in the medical realm. Nonetheless, there is a secondary use that natural stems from the collection of mass amounts of information. The data will be released for research purposes in various fields from genotype (one's genes)-phenotype (actual observed features) relation studies, pharmacology, basic molecular studies, insurance quality control, demographics, economical issues, and many other groups.

The medical information of an individual is highly sensitive, and many would prefer that their identities were not associated with information that has the ability to affect employment, insurance, and general daily situations due to stigmatism. Such issues and others are addressed by many recent articles [CR2000, HR00]. Research has been performed in the fields of computer science, statistics, and public policy to combat the erosion of privacy in the medical field. This research will demonstrate that there is information that may be inferred from longitudinal datasets. The previous research to date has concentrated only on an entry-by-entry analysis of datasets. Preliminary research that I have performed on this topic reveals that information may be inferred from such time series analysis, but this research will formally outline and define a model by which identifying features may be found in medical information that relates to genetic information, not just a single disease specific model. Furthermore, privacy is a desired aspect of information beyond the medical community. This research is capable of being extrapolated to other fields, such as marketing, governmental, and corporation information.

6 References

[BM97] Brinkman RR, Mezei MM, Theilman J, Almqvist E, and Hayden MR. The likelihood of being affected with Huntington's disease by a particular age, for a specific CAG size. *Am. J. Hum. Genet.* 1997; 60: 1202-1210.

[CR00] Who knows your medical secretes? *Consumer Reports*. August 2000: 22-26

[EHLR80] Erman LD, Hayes-Roth F, Lesser VR, Redd DR. The HEARSAY-II speech-understanding system: Integrating knowledge to resolve uncertainty. *Computing Surveys*. 1980; 12 (2): 213-253.

[FW97] Fair M and Whtridge P. Record Linkage Tutorial. *Record Linkage Workshop*. Arlington, VA. March 20, 1997.

[HF98] Hauskrecht M and Fraser H. Modeling Treatment of Ischemic Heart Disease with Partially Observable Markov Decision Processes. *Proc AMIA Symp* Nov 1998. 538-542.

[HR00] Hall MA and Rich SS. Laws Restricting Health Insurers' Use of Genetic Information: Impact on Genetic Discrimination. Am J Hum Genet. 2000; 66: 293-307.

[HW96] Hundepool A and Willenborg L. μ- and τ-argus: software for statistical disclosure control. *Third International Seminar on Statistical Confidentiality*. Bled: 1996.

[KLC98] Kaelbling LP, Littman ML, Cassandra AR. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*. 1998; 101: 99-134.

[Mal00] Malin B. Inferring Disease Age of Onset Through Computerized Medical Record Systems: A Huntington's Disease Approach. *Carnegie Mellon University*. Unpublished.

[Mit97] Mitchell TM. Machine Learning. WCB McGraw-Hill. New York. 1997.

[Mun96] Munro N. Infotech Reshapes Health Care Marketplace. *Washington Technology*. Aug 8, 1.

[MS00] Malin BA and Sweeney L. Determining the Identifiability of DNA Database Entries. *Proc AMIA Symp*. Nov 2000; 537-541.

[Nah00] National Association of Health Data Organizations, *NAHDO Inventory of State-wide Hospital Discharge Data Activities* (Falls Church: National Association of Health Data Organizations, May 2000).

[Qua99] Quarrell O. Huntington's Disease: The Facts. Oxford University Press. New York. 1999.

[Pee99] Peek NB. Explicit temporal models for decision-theoretic planning of clinical management. Artif Intell Med. 1999 Feb; 15(2):135-54.

[Rab90] Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE*. 1990.

[Sin92] Singh SP. Transfer of learning by composing solutions of elemental sequential tasks. *Machine Learning*. 1992; 8: 323-329.

[Swe00] Sweeney L. Foundations of Privacy Protection from a Computer Science Prospective. *Proc. Joint Stat. Meet.* Indianapolis, IN. August 2000.

[Swe01] Sweeney L. Simple Demographics Often Identify People Uniquely. *Heinz School Working Paper*. Carnegie Mellon University. 2001.

[Swe96] Sweeney L. Replacing personally-identifying information in medical records, the scrub system. In: Cimino, JJ, ed. Proceedings, JAMIA. Washington, DC: Hanley & Belfus, Inc., 1996:333-337.

[Swe97] Sweeney L. Guaranteeing anonymity when sharing medical data, the datafly system. *Proceedings, JAMIA*. Washington, DC: Hanley & Belfus, Inc., 1997.

[Swe98] Sweeney L. Three computation systems for disclosing medical data in the year 1999. *Medinfo*. 1998; 9 Pt 2: 1124-1129.

[Win95] Winkler WE. Matching and Record Linkage. In *Business Survey Methods*. John Wiley and Sons. 1995

[WW98] Wang JZ , Wiederhold G. System for efficient and secure distribution of medical images on the Internet. *Proc AMIA Symp*. 1998; 907-11.

[Yam98] Yamron J, Carp I, Gillick L, Lowe S, and van Mulbregt P. A hidden Markov model approach to text segmentation and event tracking. In *Proc ICASSP '98*. IEEE. 1998.