HIDDEN MARKOV MODELS FOR INFERENCE IN LONGITUDINAL MEDICAL INFORMATION

Bradley Malin
malin@andrew.cmu.edu
Heinz School of Public Policy and Management
Center for Automated Learning and Discovery
Advisor: Latanya Sweeney

Abstract: Privacy of identity is an important and complex issue in the process of sharing datasets between data between holding and receiving agencies. In the past, research has shown that identifying features may be found and removed in text documents and entry-specific databases. This application proposes research on defining aspects of information that can be inferred from longitudinal medical data. The research focuses on a hidden Markov model (HMM) for determining the hidden states of genetic disease over time. This model extends initial research I have completed with the same topic. However, it is the goal of this research to formalize the model of hidden information inference from data that is time dependent.

The model focuses on combining m-estimate frequency and variance metrics previously determined for initial research of this dataset. The combination of the statistics will allow for a more readily comparable analysis of emittance probabilities of observables for the HMM. In the initial research, it was not optimally determined which states were the correct emitting sources of the observed variables. This model will take into account all possible states for each set of observed features. The optimal sequence will be determined to maximize the probability of observed features.

Initial research without the model demonstrated that the hidden information for a Huntington's disease, the age of onset of the disease, could be error bound. The hidden information is a linker between clinical medical and DNA information. Thus, it is desirable to minimize the range of prediction for the hidden information of disease. By modeling the disease in a Markov model, I hope to portray a more accurate representation of the disease than the initial finding used. By using a more formal model, the hidden information may be bounded by tighter confidence intervals. Once the methodology for predicting hidden information is complete and accurate, privacy research will be able to be initiated, thus determining algorithms for removal or distortion of information to maintain privacy of individuals in the dataset while maximizing research uses of the information. The goal is to determine the identifying information, even when it is not explicitly presented.

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 stigmatizing circumstances. Data sets may be shared from both clinical and DNA databases in a deidentified manner. However, research on demographics and quasi-identifiers have shown that this method of anonymization is extremely inefficient. In many cases, individuals, in a deidentified data set, can be reidentified through a cross with another datasets [Swe01]. This problem allows for innovative study in the field of data privacy for several reasons:

1. **Longitudinal study.** Combinations of demographics associated with each hospital visit permit the construction of individual profiles. These profiles may range over many years of information.
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 analysis of the datasets. Yet this study allows for the incorporation of new constraints and models based on Markov models and automatons.

This particular study focuses on determining "hidden" information inferred from clinical information that may be associated with DNA datasets. For not only does a DNA entry harbor information on the "hidden" information in the disclosed dataset, but it also contains probabilistic information corresponding to a multitude of other health aspects and demographics, as well as other direct family members.

1.2 Deliverables

The core deliverable of this work is an algorithm for modeling diseases resulting from known DNA mutations in longitudinal clinical datasets. Directly stemming from this research will be the ability for researchers:

- *to infer additional clinical information that may be missing*, for more detailed analysis and relational study
- *to predict the sequence of DNA that corresponds to the hidden clinical*,

proving that reidentified clinical datasets can be linked to seemingly anonymous DNA information may.

1.3 Practical Impact

Results of this research will benefit both researchers and data proprietors since:

- *researchers will be able to construct and use more complete information sets*

yet more importantly,

- *the privacy community will be able to determine where privacy may be compromised in DNA datasets*

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 detection algorithms for identifying data within Scrub use local knowledge sources, share results and compete based on the certainty of their findings. The knowledge sources consist of lists of information such as area codes and first names and helper methods that determine if words sound like medical terms. The Datafly system maintains anonymity in released data by automatically substituting, generalizing, and suppressing information via an iterative method. The decisions for what information to distort are made at the attribute and entry-specific level, so that the final result is a subset of the parent database. The ambiguity of the entries in the resulting subsets is controllable by the holder of the information. Mu-Argus, is similar to the Datafly program in that it protects data by enforcing the ambiguity rule (also known as *k* mapping). However, instead of removing entire rows of with outlier information, the cells with the outliers are blanked out.

The previous selections have been paramount models for determining identifying features in static situations. The data that the programs work on are localized in that single hospital visits, or banking records, or whatever the data being studied are single event occurrences. The dataset that this research will determine identifying features on is a longitudinal, or time-based, dataset. This increases in dimensionality permits additional inferences about the individuals being studied.

In addition to identification algorithms, 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.

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, the blackboard architecture 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 and blackboard architecture 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 of publicly available clinical information to determine additional information about the medical condition of an individual. The additional information that is considered in this project is a feature that permits a relationship between clinical and DNA data to be established. The framework designed is generalizable to non-genetic diseases that have differential presentation over time.

3.1 Prior Work

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. Publicly available hospital discharge databases were partitioned into longitudinal profiles of the genetic-based ailment Huntington's disease. The profiles were constructed based on identifiable demographics associated with the entries [Swe01]. From these profiles, diagnoses were used as the observable information to infer hidden information. The hidden information was appended to the profiles from a database listing the age of onset of Huntington's disease for the individuals of the profiles. 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. Via frequency and variance metrics, combinations of diagnoses were determined as to the probability of associated a combination with specific stages of the disease. The details of the metrics may be found in the previous work [Mal00]. The diagnoses combinations and metrics were stored in a lookup table.

After construction of the lookup table, a predicted stage was assigned to the entries of each profile via an iterative algorithm searching through the diagnosis combinations. The assigned stages were then used to predict the age of onset target of the disease. This was achieved by corrected what seemed to be erroneous errors in the longitudinal profile. For instance, the disease in question has a linear forward moving progression. Once the patient reaches stage 2, remission to stage 1 is not possible. Thus, the profiles with such occurrences were corrected for. 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. [Qua99] Thus, to predict the age of onset, the stage with the most visits is utilized. 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 threshold (transformation of stage) has occurred, or may have occurred. Some individuals reside in one stage only 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). 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. The results of the research showed that 82% of the test cases were correctly bounded in an age range of five years or less.

Though the prior research was a success, there were several drawbacks. First of all, the age of onset range was estimated while neglecting possibly a majority of the hospital visits made by an individual. Second, upon manual inspection, the correction of some profiles seemed incoherent with the expected results. Thus, to account for this problem, this research project will focus on incorporating methods for preventing the previous difficulties.

3.2 Constrained Hidden Markov Model

Rather than incorporate a method for removing seemingly incoherent stage assignments in profiles, use of 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 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 S 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.

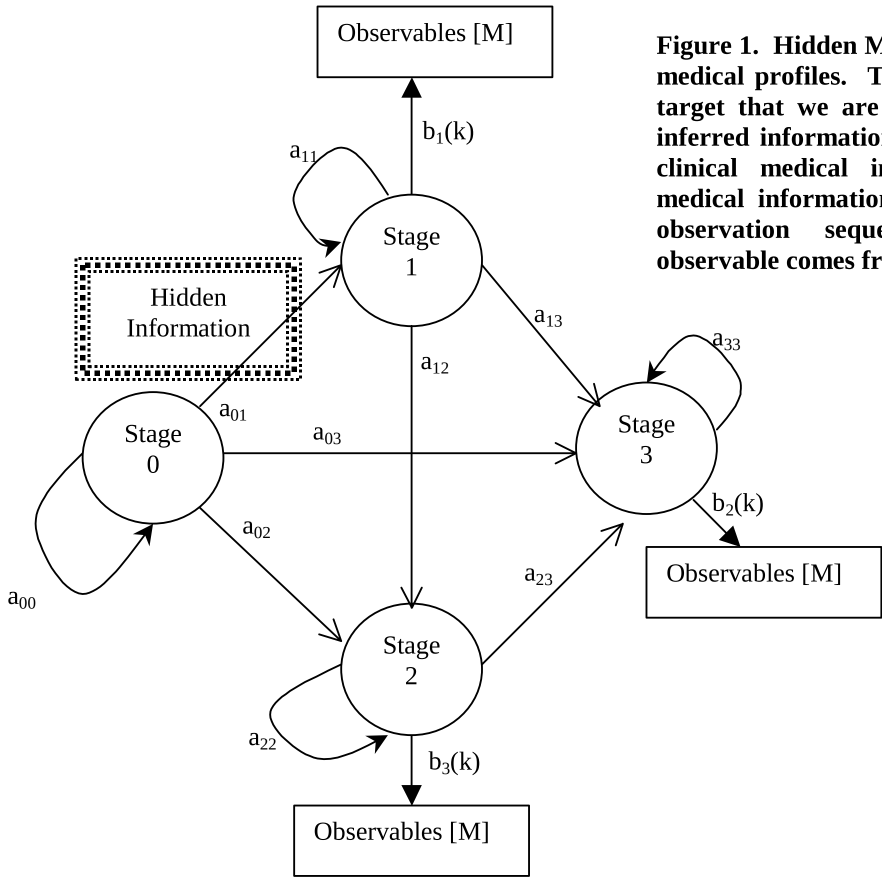

Figure 1. Hidden Markov model for longitudinal medical profiles. The hidden information is the target that we are interested in. This type of inferred information is an identifier that relates clinical medical information to DNA/genetic medical information. To infer this target, the observation sequence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

The underlying model is a discrete Markov process, however, unlike a typical Markov model, a_{ij} may be represented as some other function than a probability. Instead, a_{ij} can be a set of rules determined from observed or expert data. At this time, it is unknown if a_{ij} will be represented as a discrete probability or by a set of rules. In actuality, a_{ij} may be a joint probability based on the probabilities of the observables from all stages. Furthermore, the stages of the disease are never actually observed.

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

$$b_j(k) = P[v_k \ \ at\ t|q_t = S_j] \qquad ; \qquad \begin{array}{l} 1 \le j \le N \\ 1 \le k \le M \end{array}$$

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

$$\pi_i = \Pr\,[q_1 = S_i] \qquad ; \qquad 1 \le j \le N$$

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 $O = O_1O_2. . .O_T$, (as in Figure 2) the model uses the associated probability distributions to predicted the most probable corresponding to the state sequence $Q = q_1q_2. . . q_n$, 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.

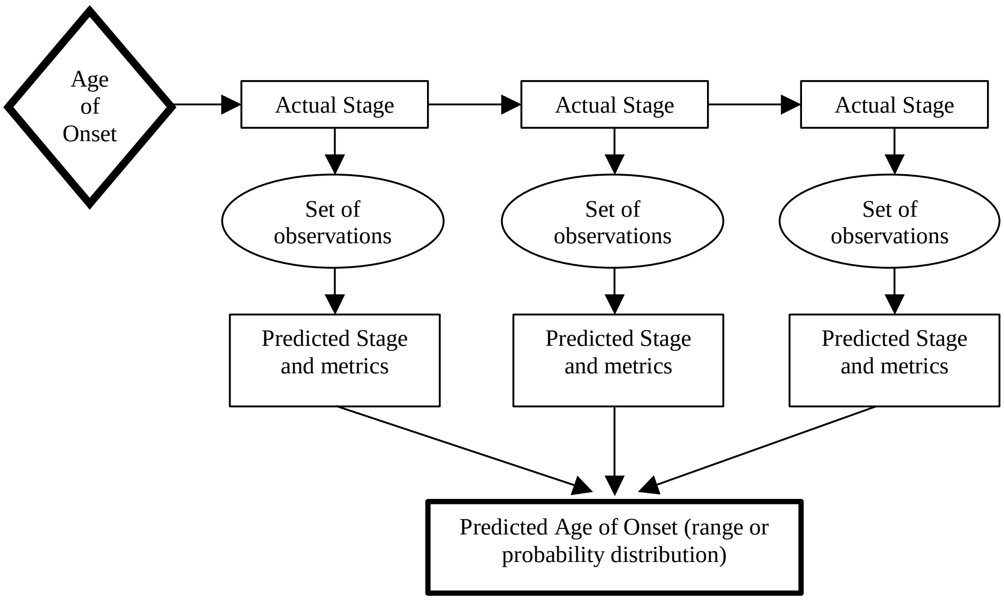

Figure 2. Model of presentation and inference strategy. Observables are collected at various time points. The actual state of the disease is unknown to the individual making the observations. The state may be inferred from the collected observations over time. The actual states can be further related 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. Death may occur at any time during the progression of the disease, since the death of a Huntington's disease patient is due to confounding complications not directly due to the disease itself. For example, Huntington's disease causes the degeneration of control of the nervous system and rigidity of the muscles, but an individual can not die from muscle atrophy. However, an individual can choke on their food or fall down stairs. Yet, since there is no treatment, there is no remission of the disease observed.

3.3 Determining the Emittance Probability Distribution

The probability for observables in each state is 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 is achieved can be found in the prior work [Mal00]. The frequency estimate is 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.

No stage is represented as the definitive stage. Rather, each stage has a probability of being the emitting stage based on a lookup table of metrics associated with each diagnosis combination. 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 a frequency metric and the associated variances are listed in Table 1 and Table 2.

	Frequency Metrics				Variance Metrics			
Observable	Stage 0	Stage 1	Stage 2	Stage 3	Stage 0	Stage 1	Stage 2	Stage 3
V_1	0	-2	0	0.667	0	0.0263	0	0.375
V_2	0	-3.5	0.166	-1.666	0	0.05263	0.3	0.375
V_3	0	0	1	0	0	0	0.2	0

Table 1. Sample frequency and variance estimates.

	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 2. 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. The lower the number of variance metric, the greater the variance of the observable. Thus, while an observable may have be exhibit 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. Therefore, the first step of this research will be to define a combined metric of frequency and variance.

3.4 Inferring the Hidden Target

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 once the before optimal HMM parameters are determined. However, the predicted hidden information is predicted in a bounded state, with maximum and minimum range. 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 uses 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.

4 Work Plan

The multi-tiered aspect of this project allow 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. This will require meeting with several statistical experts that have expertise in the use of statistics in large datasets, as well as medical informatics. The professors that I will be looking to for help are Steve Roerhig and Rema Padman.
- 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. These institutions are areas of data warehousing and expertise in the field of Huntington's disease, including Indiana University, Rush Presbyterian Hospital in Chicago, and University of Chicago.

- ***Weeks 3 and 4****:* At this point, this the actual probabilities associated with Markov model transition probabilities will be defined. It is currently unknown if these will be discrete numbers, patterns of observations over time, or some other undefined method.
- ***Weeks 5 and 6****:* Automation of the algorithm will occur after definition of the model. The algorithm is likely to be written as a JDBC (Java Database Connectively) application. This will permit control of the coding and personal methodology. Though, the automation process may not be limited to JDBC if at first this proves too difficult. There are other available programs that may be determined to be beneficial and more efficient than starting from scratch.
- ***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.
  - Other models may be utilized at this point in time. It may be more realistic to attempt a model that has some temporal dependence on treatments as well as diagnoses. Therefore, POMDP models may be beneficial at this point in time. These results will be compared with that of the HMMs.

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 the electronic patient record, which will increase the availability of patient information for swifter delivery of care and insurance in the medical realm. However, there is a secondary use that natural stems from the collection of mass amounts of information. The data will be handed out 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. However, 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

[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.

[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.