Biomedical Data Privacy Issues and Solutions: An Interdisciplinary Graduate School Course

Tatyana Ryutov, University of Southern California

Abstract

The course was developed for graduate students interested in exploring privacy concerns in healthcare, the current law and governing regulations, and learning and applying the existing and emerging technologies to address these concerns. Biomedical data privacy is an interdisciplinary problem, and this course touches on issues in bioinformatics, computer science, law and policy, and ethics. This paper describes the design of our biomedical privacy course, the learning objectives, teaching materials and methods, the supporting learning activities, and the learning assessment and outlines plans for future improvements. A course survey was given and the student feedback was positive. A majority of students would recommend this course to other students and rated highly the remote learning experience. The course addresses emerging biomedical data privacy issues and facilitates collaboration between engineers, healthcare and legal professionals.

1. Introduction

The detail and diversity of collected healthcare and biomedical data is constantly increasing. Health information is collected through mobile devices [6], in personal domains [3], and from sensors attached on or in human bodies [13], [21].

The amount of genomic sequencing data has risen exponentially in recent years recent years [12]. Human genome sequencing has led to deeper understanding of a variety of both common and rare diseases at an unprecedented scale. Precision medicine offers tremendous potential for improving human health. This potential. This potential, however, may not materialize unless we address ethical, legal and societal challenges. Genomic data is uniquely identifiable and existing privacy laws, regulations and practices often fall short when facing the unique challenges that it presents.

The ready availability of such large volumes of detailed data has been accompanied by privacy invasions. Recent breach notification laws at the US federal and state levels have brought to the public's attention the scope and frequency of these invasions. According to [11], between 2009 and 2021, over 4,400 healthcare data breaches of 500 or more records have been reported to the United States Department of Health and Human Services Office for Civil Rights. These breaches have resulted in the loss, theft, or of over 300 million healthcare records. That equates to more than 94% of the 2021 population of the United States. The problem of preserving patient privacy has received increasing attention in the era of big data.

We developed a new interdisciplinary course in line with the most recent advancements in the privacy research. This course attracts and caters to mutual interests across different graduate programs. The goals of this class are 1) to raise awareness of the privacy-related issues in healthcare with a specific focus on genomic data; 2) to delineate the ethical, legal and societal implications of genomic medicine; 3) to present existing computational solutions that address the privacy issues of accessing and sharing genomic data, and their limitations; and 4) to introduce emerging legal and informatics solutions that would address the concerns of genomic privacy.

Recommended Preparation

Prior experience with information security, public policy, and legal frameworks is not required for this course. However, basic understanding of engineering and/or technology principles and basic programming skills are preferred. While key concepts and relevant methodology are reviewed and introduced throughout the course, students are expected to be comfortable learning about basics of human genetics, precision medicine, various cryptographic methods, and statistics. The course includes writing assignments and oral presentations.

A self-proposed **semester-long** project allows the students to select either a research-oriented or implementation-oriented direction. Project-based learning enables students to learn communication, problem solving, and critical thinking skills [20], [23]. Student can select topics their interested in, which makes them more engaged and motivated. Students can select topics their interested in, which makes them more engaged and motivated. This results in a more effective learning [15], [17].

Course Enrolment Statistics

Out of 37 enrolled students, 17 were female and 20 were male.

The table 1 shows the number of students majoring in a specific program.

Program	Number of Students
Master of Science, Healthcare Data Science	1
Master of Science, Applied Data Science	19
Master of Science, Computer Science	3
Master of Science, Cyber Security Engineering	8
Master of Science, Communication Data Science	4
Master of Science, Translational Biotechnology	1
Master of Science, Financial Engineering	1

Table 1 Course enrolment by major

2. Course Design

The **learning objective**s for the course:
After successfully completing this course, the students will be able to:

- Comprehend the significance of privacy of medical data in healthcare
- Analyze privacy laws and governing regulations
- Identify the fundamental concepts and key issues of genomic privacy
- Apply the existing privacy preserving methodologies.

Graduate students enrolled in this course approach complex biomedical data privacy issue from the following angles:

- Data Vulnerability: Demonstrate how seemingly private information, can be discovered (or exploited) using automated strategies.
- Data Protection: Select privacy protection technologies that provide formal computational guarantees of privacy in disclosed datasets.
- Technology Policy Design: Apply privacy protection technologies that complement policy regulations

The semester-long course comprises 14 lectures. The last two lectures were devoted to student project demonstrations. We host the final examination that tests basic knowledge of the key concepts and relevant methodologies on the 15th week. The exam format is a combination of short answers and essays.

Grading

The grading weights are 40% for homework assignments, 25% for semester project, 25% for the final examination and 10% for class participation.

The grading weights for the semester project are 5% for the project proposal, 5% for the project progress report, 20% for the project presentation and 70% for the final project report. The instructor evaluates the progress and provides structural suggestions to facilitate students' learning and project formulation. Each student presents a PowerPoint and live or recorded demos for their project. Student presentations are assessed based on the peer review: each student completes a brief survey providing their thoughts and reactions to the presentations.

Course content

The course first introduces students to the data privacy definitions and privacy frameworks. We start with the concept of and privacy as a right [27] and a tort [22] and then discuss a modern view on privacy as a contextual integrity [18]**.** We then move onto legal aspects of privacy, in particular we examine privacy law and policy in the United States [9], [10] and the European Union [8]. We discuss the limitations of legal and regulatory frameworks to protect genomic privacy.

We cover a tutorial on Ethical Theory and Biomedical ethics principles [1], [4]. Students are introduced to the concepts of ethical dilemma, ethical analysis and professional code of ethics (e.g., National Society of Genetic Counselors Code of Ethics). Special attention is given to ethical issues in Genomics, in particular we examine ethical issues related to incidental and secondary findings in genomic testing **[5].** Students are encouraged to practice ethical analysis by selecting a controversial topic of interest related to biomedical context and conducting an analysis of the ethical dilemma.

We next introduce the fundamentals of genomics, including the basics of human genome, genetic variations and disorders, and disorders, clinical genomics. Students are then introduced to sequencing technologies, Next-Generation sequencing, variant calling, annotation and interpretation and standard sequence formats. We next explore case studies on resolving individual identity from aggregated genetics data, and performing linkage attacks. Students perform hands-on exercises, including DNA analysis using Genome browser, determining membership and identifying individuals included in a specific dataset.

Next, we cover the basics of information security, such as information security and privacy goals, security policy, authentication and identification, access control models, monitoring and auditing. Students then are introduced to practical tools that support the security and privacy goals. Students are encouraged to experiment with the Role based access control (RBAC), and Attribute based access control (ABAC) in Microsoft Azure environment. We then move onto the basics of cryptography, including symmetric and public key cryptography, message integrity, digital signatures, etc. Next, the students learn the emerging cryptographic tools for privacy protection such as homomorphic encryption and blockchain technology.

We next explore different approaches to de-identification of unstructured data, including natural language processing [24], [2], [26] and Machine Learning. Moving on, we discuss privacy-preserving record linkage, formal models of anonymization, including k-anonymity, l-diversity, and t-closeness [7], [25], [14], [16]. Lastly, we introduce students to Differential Privacy [19] and emerging technologies to protect Genomic data privacy.

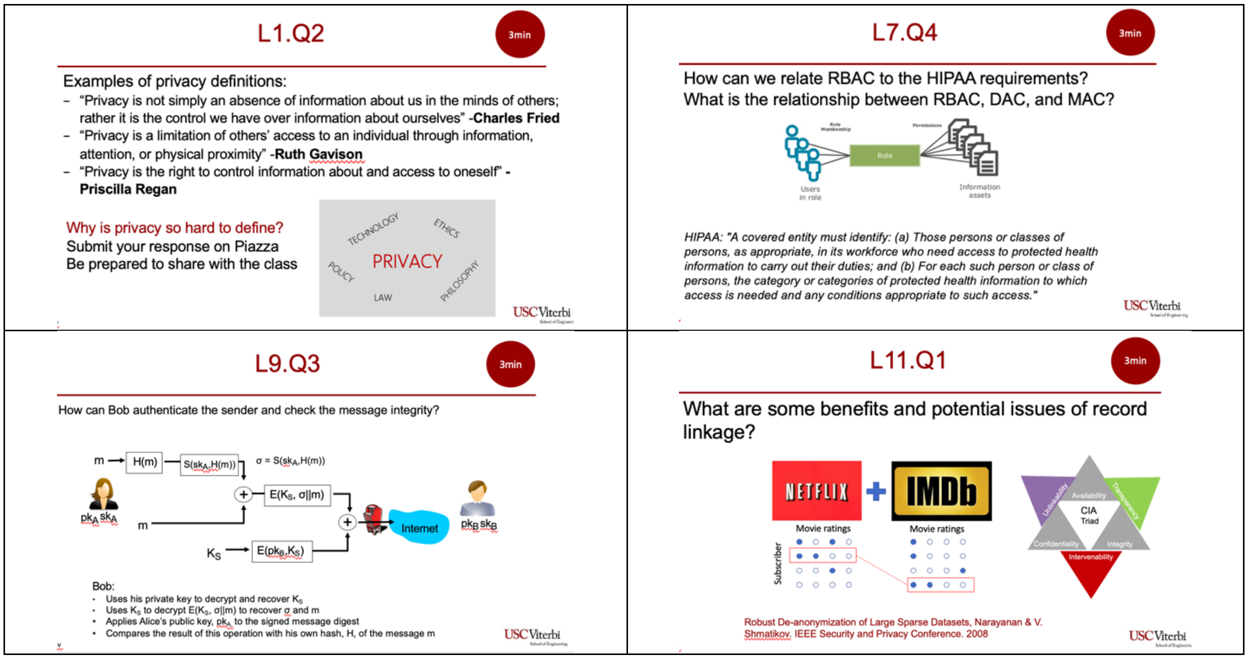
There is no primary textbook for this course. Reading assignments are selected from academic literature, various periodicals and other sources.

This course was co-created with a collaborator from the USC Keck School of Medicine, who gave lectures on genomic basics, data processing and technologies. We invited a guest lecturer from the industry, a Software Engineer from Google's Ads Data Hub, who gave a lecture on Differential Privacy and emerging privacy preserving technologies in bioinformatics.

Engagement

This course was conducted entirely online, using a combination of synchronous and asynchronous methods. The remote learning format of this interdisciplinary course was selected in order to eliminate the need for students from different schools to travel between campuses. For example, USC Health Sciences Campus is located seven miles from the USC University Park Campus.

To make this on-line class more engaging, each lecture contains about six "Discussion questions" (administered using Google forms) that the students need to answer to receive a class participation score. The students are asked to solve problems, formulate and answer questions, and debate during class. The students have a few minutes to work on each question. Subsequently, the solutions are discussed by the entire class. Table 2 shows some examples of the discussion questions.



L1.Q2

3min

Examples of privacy definitions:

- "Privacy is not simply an absence of information about us in the minds of others; rather it is the control we have over information about ourselves" -**Charles Fried**
- "Privacy is a limitation of others' access to an individual through information, attention, or physical proximity" -**Ruth Gavison**
- "Privacy is the right to control information about and access to oneself" -**Priscilla Regan**

Why is privacy so hard to define?
Submit your response on Piazza
Be prepared to share with the class

USC Viterbi

L7.Q4

3min

How can we relate RBAC to the HIPAA requirements?
What is the relationship between RBAC, DAC, and MAC?

HIPAA: "A covered entity must identify: (a) Those persons or classes of persons, as appropriate, in its workforce who need access to protected health information to carry out their duties; and (b) For each such person or class of persons, the category or categories of protected health information to which access is needed and any conditions appropriate to such access."

USC Viterbi

L9.Q3

3min

How can Bob authenticate the sender and check the message integrity?

Bob:

- Uses his private key to decrypt and recover K_S
- Uses K_S to decrypt $E(K_S, \sigma||m)$ to recover σ and m
- Applies Alice's public key, pk_A to the signed message digest
- Compares the result of this operation with his own hash, H, of the message m

USC Viterbi

L11.Q1

3min

What are some benefits and potential issues of record linkage?

Robust De-anonymization of Large Sparse Datasets, Narayanan & V. Shmatikov, IEEE Security and Privacy Conference. 2008

USC Viterbi

Table 2 Examples of Discussion questions

To increase engagement, each student prepares a short presentation on a topic they are interested in, related to but not included in the class material. They assume the role of a teacher which allows them to enhance their understanding of the material and share it with others.

3. Student Projects

Data privacy is a complex, multi-faceted topic, and the course approaches biomedical data privacy issues from three angles:

1. Ethical angle: ethical issues involved in collecting and sharing biomedical data.
2. Legal angle: the most important privacy laws and governing regulations.
3. Technological angle: existing and emerging privacy preserving technologies.

The students have two options for the project format: write a research paper or design and implement a prototype of a privacy-preserving solution.

This section summarizes 37 student course projects.

Technological angle

The majority of the students selected technological aspect for their projects. The projects in this category included? the following:

1. Comparative studies using simulations.

- Monte Carlo simulation using SAS toto evaluate the re-identification risk of a Healthcare database.
- Evaluation experiments to compare performance of various Differentially Private algorithms for Genome-wide association studies.

2. Implemented system prototypes. Some examples:

- A clinical data de-identification tool with the BERT model and NER technique.
- A RBAC-based access control system for protecting personal health records.
- Tokenization and De-identification of Health Care Data.
- Healthcare apps and data privacy improvement.
- Machine Learning for encrypted healthcare data using partially homomorphic encryption.
- Block chain database for storing Medical Records.
- HIPAA-compliant reference architecture for Virtual Reality and Augmented Reality.

3. Research papers. Notable examples:

- Evaluation of Access Control Mechanisms (RBAC and ABAC) to meet Healthcare industry standards.
- Health data privacy and protection for Internet of Things.
- Comparison of access control frameworks (LiMDAC framework, Greenshield's DSMIN framework, and Saini's blockchain solution) for distributed biomedical data systems.
- Analysis of security requirements, concerns, and current solutions for medical devices and medical networks.
- Design of privacy technologies for personal health records (PHR).
- Comparison of existing solutions for genomic privacy for aggregated data.
- Exploration of health information protection with bio-identification technologies (fingerprint and facial recognition).
- Analysis of existing data protection models in healthcare wearable technology.

Legislation and policy

Projects in this category were entirely research papers exploring specific privacy-related issues and policies, identifying limitations of the current legal frameworks, and proposing possible development directions. Some examples of projects in this category**:**

- How personal biometric data being protected in the United States.
- Privacy issues related to visual assistance technologies (VAT).
- Should genomic research results be returned to research participants?
- Privacy risks and solutions in commercial DNA testing.
- Current state, regulations and privacy-preserving approaches in Telehealth & Telemedicine.
- Legal principles of biomedical data privacy in the post-pandemic age.
- The risks and legal Protection of patients' privacy during COVID-19 pandemic, a comparative study of the U.S. and China approaches.

Ethics

Several projects explored ethical implications of biomedical data privacy. These projects were research papers aiming to understand the issues and the perspectives of different sides. The projects in this category include the following:

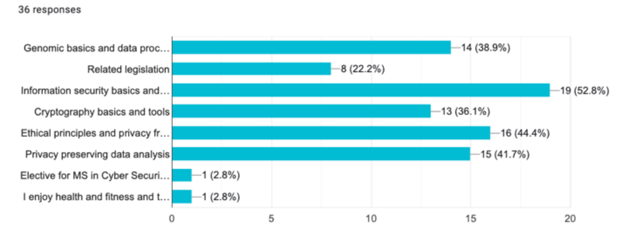
- The ethics of genomic data privacy.
- Ethical use of artificial Intelligence in healthcare.
- Implications of capturing the user's fingerprints or facial biometric information.

- Ethical reflection on the development strategy of biobanks in the era of Big Data.

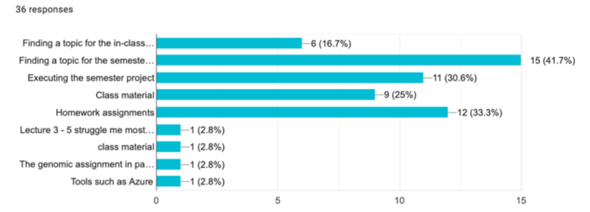
4. Course Evaluation and Conclusions

Table 3 shows our survey results to evaluate the course. A majority of students gave high ratings for recommending this course to other students and the remote learning experience. The survey results indicate positive feedback for the course.

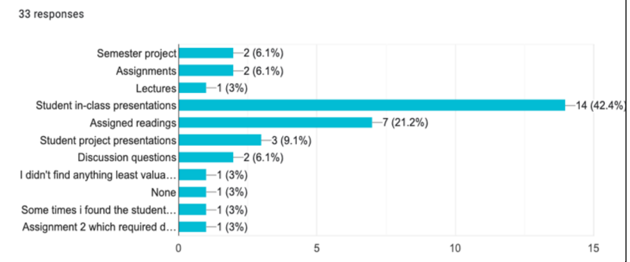
What is your initial learning motivation for taking this class? Can chose more than one.



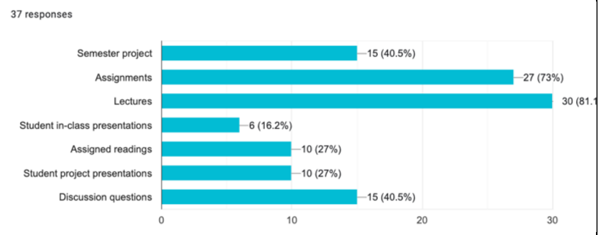
In this class, what did you struggle with the most? Can chose more than one.



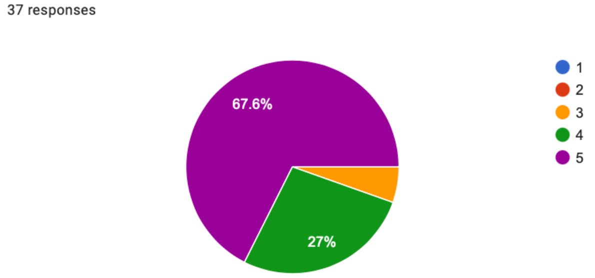
What class components/activities do you find least valuable? Can chose more than one.



What class components/activities do you find most valuable? Can chose more than one.



On a scale of 1(low) -5 (high), how would you recommend this class to your friends?



On a scale of 1(low)-5 (high), rate your **remote** learning experience.

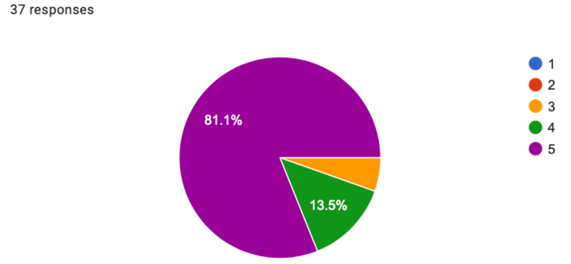


Table 3 Course survey results

This timely course addresses current biomedical data privacy issues and serves as a bridge to foster collaboration between engineers, healthcare and legal professionals.

From a teaching perspective, the two main challenges were:

- Balancing theoretical lecture materials and applied demonstrations.
- Creating appropriate hands-on and programming assignments for students with different backgrounds.

According to the feedback, the remote learning format and the combination of synchronous and asynchronous modality worked very well for graduate students. Some examples of the students' remarks:

“As a working professional, I enjoy the flexibility of the program”.

“It is more flexible than in-class learning. I can go back to watch recording repeatedly if there's anything I miss or didn't understand”.

“Replay is extremely helpful, especially on the biomedical part. It has a lot of professional words which I cannot catch at the first time.”

When asked about negative experiences related to remote learning, the students pointed out the lack of collaboration between students. For example:

"I did not have opportunity to discuss with the classmates since we don't know each other".
"Personally, I like this teaching method. Maybe sometimes I can't engage in the class and discuss with other students in remote learning".
"It is not easy to interact with classmates in remote classes".

Finding a topic for the semester project, completing homework assignments, and executing the semester project were ranked as the most challenging aspects of the class. Some students suggested to schedule the progress report earlier in the semester to increase the time between the progress and final reports. They felt that the relatively short interval between the reports resulted in some repeated content that was presented earlier. A number of students thought they could benefit from more hands-on and coding assignments. Some students noted that some assigned readings were difficult to follow and suggested to include relevant video content.

Based on the students' feedback, we plan to implement the following changes to our teaching practices:

- Develop larger and more varied sets of assignments for students with different backgrounds.
- Promote student interaction to compensate for the lack of in-person setting. For example, projects can be conducted in groups.
- Provide more project guidance. One of the main difficulties noted by the students was developing projects from scratch. We will provide sample projects and more initial discussions.
- Include more visual aids, such as live demonstrations and videos. We will select relevant healthcare related visual materials.

This paper summarizes the design of our interdisciplinary course for graduate students interested in biomedical privacy. The paper discusses the evaluation results based on the students' feedback, offers lessons learned from our semester-long experience of teaching this interdisciplinary class and outlines plans for future improvements. This timely course addresses current biomedical data privacy issues and serves as a bridge to foster collaboration between engineers, healthcare and legal professionals.

Acknowledgements

We would like to thank Dr. David W. Craig and Dr. Xiaowu Gai for their contributions to the development of this interdisciplinary class.

References

[1] Andrade, G., "Medical ethics and the trolley Problem", 2019.

[2] Berman JJ. "Concept-match medical data scrubbing: how pathology text can be used in research", 2003.

[3] Chen M, Gonzalez, S Vasilakos A et al., Body area networks: a survey. Mobile Netw Appl 2011; 16:171–93.

[4] Tom L. Beauchamp and James F. Childress, "Principles of Biomedical Ethics", 2008.

[5] Christenhusz, G. M., Devriendt, K., & Dierickx, K. "To tell or not to tell? A systematic review of ethical reflections on incidental findings arising in genetics contexts." European Journal of Human Genetics, 21(3): 248-255, 2013.

[6] Estrin D Sim I .,"Open mHealth architecture: an engine for health care innovation". Science, 330:759–60, 2010.

[7] Friedman, R. Wolff and A. Schuster, "Providing k-Anonymity in Data Mining", Int J Very Large Databases, 17:789–804. doi: 10.1007/s00778-006-0039-5, 2008.

[8] The General Data Protection Regulation https://gdpr-info.eu

[9] The Genetic Information Nondiscrimination Act of 2008.

[10] U.S. Department of Health and Human Services Summary of the Privacy Rule of the Health Information Portability and Accountability Act (HIPAA).

[11] https://www.hipaajournal.com/healthcare-data-breach-statistics/

[12] Lathe, W., Williams, J., Mangan, M. & Karolchik, D., Genomic Data Resources: Challenges and Promises. Nature Education 1(3):2, 2008.

[13] N. Li, T. Li, S. Venkatasubramanian, "t-Closeness: privacy beyond k-anonymity and l-diversity". Proceedings of the 23rd IEEE International Conference on Data Engineering, 106-115, 2007.

[14] Li M, Lou, W, Ren K., "Data security and privacy in wireless body area networks". IEEE Wireless Commun, 17:51–8, 2010.

[15] Lord, S. M., Chen, J. C., Nottis, K., Stefanou, C., Prince, M., & Stolk, J., "Role of faculty in promoting lifelong learning: Characterizing classroom environments". Education Engineering, 14(16), 381-386. doi:10.1109/EDUCON.2010.5492553; 2010.

[16] Machanavajjhala, A., Gehrke, J., Kifer, D.: "ℓ-diversity: Privacy beyond k-anonymity". In: ICDE, pp. 24–35, 2006.

[17] Mega, C., Ronconi, L., De Beni, R.,"What makes a good student? How emotions, self-regulated learning, and motivation contribute to academic achievement". Journal of Educational Psychology, 106(1), 121-131. doi:10.1037/a0033546, 2014.

[18] Nissenbaum, H., "Privacy as contextual integrity". Washington Law Review, 79, 119–157, 2004.

[19] K. Nissim, et al. "Differential privacy: a primer for a non-technical audience". White paper of the Privacy Tools for Sharing Research Data Project, Harvard University, 2017.

[20] Pengyue Guo and Nadira Saab and Lysanne S. Post and Wilfried F. Admiraal "A review of project-based learning in higher education: Student outcomes and measures", International Journal of Educational Research, 102:101586, DOI:10.1016/j.ijer.2020.101586, 2020.

[21] M. H. P. Jr. Hanson, A. Barth, K. Ringgenberg, B. Calhoun, J. Aylor and J. Lach, "Body Area Sensor Networks: Challenges and Opportunities," IEEE Computer, pp. 58-65, 2009.

[22] William L. Prosser. Privacy. California Law Review, 1960.

[23] Roessingh, H., Chambers, W., "Project-based learning and pedagogy in teacher preparation: Staking out the theoretical mid-ground". International Journal of Teaching and Learning in in Higher Education, 23(1), 60-71, 2011.

[24] P. Ruch, et al. “Medical document anonymization with a semantic lexicon.” Proceedings of the American Medical Informatics Association Annual Fall Symposium (AMIA), 729-733, 2000.

[25] Sweeney L. “Achieving k-anonymity privacy protection using generalization and suppression”, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5):571-588, 2002.

[26] Taira R, Bui A, Kangarloo H. “Identification of patient name references within medical documents”. In Proceedings of the 2002 AMIA Annual Fall Symposium, 757-761, 2002.

[27] S. Warren and L. Brandeis. “The right to privacy.” Harvard Law Review, V. IV, No. 5., 1890.