Revisiting the Informed Consent Principle for Data Release

Ilaria Venturini
Sapienze Universit di Roma, Roma, Italy

Keywords: Consent for Data Release, Personalized Privacy, Privacy Threats.

Abstract: In this paper the Informed Consent Principle for sensitive information management is revisited under the perspective of ubiquitous and embedded computing and of personalized privacy. We claim that an Informed Fair Consent Principle, here proposed, takes into account privacy threats, pointed out in the paper, which may stem from some granted consent. Potential implementations of the fair personalized privacy are shortly discussed.

1 INTRODUCTION

Privacy has several meanings and does not always involve information. Here the focus is on information privacy, that is on the right of information owners to determine when, how and to what extent information about them is collected, stored, shared, processed, communicated to others and disseminated (Chen et al., 2009). It is nowadays well known that privacy, as well as security, has to face new challenges at every technological innovation which facilitates the dissemination of information (De Capitani and Samarati, 2006). Besides technological novel open issues which arise because of the technological evolution, also traditional notions may need to be revisited because of novel technologies. This is the case for the consent notion, which is crucial for the privacy preservation. Privacy is context dependent at least in a twofold sense. It depends on local law, though privacy cannot be limited to compliance to law regulations (sometimes fragmented if not inconsistent (Applebaum et al., 1987)), and on the application domain. Throughout this paper, the sanitary context is the considered application domain. Health related information is sensitive. Privacy preservation is a traditional issue in the sanitary context since the formulation of the *Hippocratic Oath*. Nowadays, the amount of socio-sanitary and clinical data is dramatically increasing also because of ubiquitous and embedded computing. Worn and implanted devices, as utilized for instance in remote patient monitoring via sensors and RFID tag-reader technology (Solanas and Castella-Roca, 2008), generate increasing quantities of data that will be made available to a growing number of applications in research, clinical audit, administration and marketing. Patients are the humans who can be considered as mobile embedded systems, due to the smart low-resource devices (digital assistants, health monitoring devices, pacemakers, etc.) they carry and on which they depend more or less seriously. Such devices communicate with back-end systems and data are gathered in databases or in centralized clinical data repositories (Fung et al., 2010).

Socio-sanitary systems have been set up which may connect patients, caregivers, sanitary administrators, healthcare organizations, pharmacies, etc. Throughout any regional setting, segments of information are maintained by autonomous, networked clinical information sources having differing internal structures, database schemata and vocabularies to describe the notions they use. Such gathered data are often shared with users who need them for data mining tasks. Patient records are often released to an external medical center for information sharing. Health data sets are often released for, e.g., various statistics, research purposes, classifiers and education. In some cases publishers know in advance the kind of mining work that need to be performed on the released data, but in other cases they do not. In pervasive and cloud computing environments, the patient control on personal data looses its potential effectiveness once personal information is managed on servers located in a Country different from that where the patient lives. Thus the consents granted by patients have to be analyzed more deeply for what concerns the authorizations they may determine on information management.

Contributions and Structure of the Paper

The Informed Consent Principle for sensitive infor-

Venturini I..
Revisiting the Informed Consent Principle for Data Release.
DOI: 10.5220/0004340000920097
In *Proceedings of the 3rd International Conference on Pervasive Embedded Computing and Communication Systems* (PECCS-2013), pages 92-97
ISBN: 978-989-8565-43-3

mation release is revisited. A privacy threat is singled out which derives from consents granted by patients and which has a social privacy impact. A Fair Consent Principle is proposed to overcome the underlined privacy problem and its effectiveness is discussed.
The paper is structured into five sections besides Introduction and Conclusions. In Section 2, the informed consent for personal data publication is considered, exemplifying in the healthcare framework. In Section 3, the problem of the patient consent that concerns other individuals is addressed in the personalized privacy perspective. In Section 4, a Fair Consent Principle is proposed. In Section 5 fair personalized privacy is discussed taking into account both privacy protection and data utility issues. In Section 6 related work is mentioned.

2 INFORMED CONSENT

Since the *Hippocratic Oath* doctors are obligated to shield their patients from releasing information about their diseases. The doctor-centric approach has been long-standing. Events in the half of 20th century, especially during World War II, gave rise to the ethical standard known as *Informed Consent*, thus moving towards a patient-centric approach. New principles that created protections and rights for patients were incorporated into the *Declaration of Helsinki* in 1964, which has been updated several times since its creation. Current guiding principles are embodied in various international guidelines. For what concerns the patient consent for data publication, protection of patients' rights to privacy, which was agreed on by the International Committee of Medical Journal Editors (ICMJE), was published in the British Medical Journal in November 1995. The statement about the informed consent, which is now included in the ICMJE Uniform Requirements, follows:

'Patients have a right to privacy that should not be infringed without informed consent. Identifying information should not be published in written descriptions, photographs, and pedigrees unless the information is essential for scientific purposes and the patient (or parent or guardian) gives written informed consent for publication. Informed consent for this purpose requires that the patient be shown the manuscript to be published.'

Since then consent is understood to be informed. General principles for security and privacy protection have been formulated for privacy-compliant policies in order to protect and manage private information that reside in databases, data repositories, etc. Various Informed Consent Principle formulations depend on geographic locations and national laws. A concise formulation that is usually in privacy-aware information systems states:

'Personal information that has been collected should have the consent of the information donor.'

Consent from patients should be obtained before publishing personal information about them, whether or not patients can be identified, at a greater reason if there is any doubt about achieving complete anonymity. The various aspects that have been addressed in the published literature concerning patient consent for healthcare suggest that a consent freely given by someone with the mental ability to do so is an ongoing process (Royal College, 2011). Publication without consent is acceptable in some cases, e.g., if the patient is long dead and has no living relatives or if the patient is unable to give consent. Emergency situations may require forcing the privacy rules of the privacy policy in use. Anyway, disclosures without consent must always be legal and justifiable as necessary (for instance, in case there is a public interest or a serious unlawful act or a legal process).

3 PERSONALIZED PRIVACY

Raw data gathered from embedded devices are organized as records (*microdata*) grouped as row tables, with columns as attributes. There are attribute classes: *Identifiers* (as complete names, social security numbers, passport numbers) that explicitly identify individuals; *Quasi-identifiers* (as postal codes, gender, age) that identify individuals once they are combined with other information; *Sensitive Attributes* (as disease, disability state, financial state, religion) and *Non-Sensitive Attributes*. Table 1 (a modified version of Table 3 in (Xiao and Tao, 2006)) is a raw data table.
In Table 1, 'sta.pneumonia', 'stre.pneumonia' and 'as.hemoph. A' stand for 'staphylococcus pneumonia', 'streptococcus pneumonia' and 'asymptomatic hemophilia A', respectively.

For the disclosure perspective, microdata undergo some anonymity process. Specifically, for achieving de-identification, *Identifiers* are replaced by numerical values more or less randomized. However, nowadays it is well known that by linking *Quasi-identifiers* in de-identified tables with some external data (somehow available) may determine re-identification. Then also *Quasi-identifiers* are somehow anonymized. *Sensitive Attributes* are unmodified.

Of the two disclosure types, namely *identity dis-*

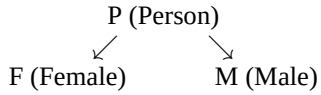
Table 1: Raw data table.

Name	*Age*	*Sex*	*Zipcode*	*Diagnosis*
1 Andy	5	M	120	sta.pneumonia
2 Bill	9	M	140	stre.pneumonia
3 Ken	6	M	180	sta.pneumonia
4 Nash	8	M	190	as.hemoph. A
5 Joe	12	M	220	sta.pneumonia
6 Sam	19	M	240	stre.pneumonia
7 Linda	21	F	580	as.hemoph. A
8 Jane	26	F	360	stre.pneumonia
9 Sarah	28	F	370	as.hemoph. A
10 Mary	56	F	330	sta.pneumonia

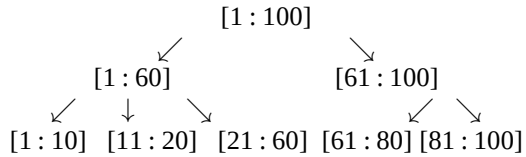
Table 2: Personalized Table 1.

Name	*Age*	*Sex*	*Zipcode*	*Diagnosis*
1	[1 : 10]	M	120	sta.pneumo.
2	[1 : 10]	M	[101 : 200]	stre.pneumo.
3	[1 : 10]	M	[151 : 200]	sta.pneumo.
4	[1 : 10]	M	[151 : 200]	as.hemoph. A
5	[11 : 20]	M	[201 : 250]	sta.pneumo.
6	[11 : 20]	M	[201 : 250]	stre.pneumo.
7	21	F	580	as.hemoph. A
8	[21 : 60]	F	[351 : 400]	stre.pneumo.
9	[21 : 60]	F	[351 : 400]	as.hemoph. A
10	56	P	330	sta.pneumo.

closure (i.e., discovering that a certain individual is included in a target table) and *sensitive attribute disclosure* (i.e., associating a sensitive attribute with individuals), we here address the second one. Since privacy is not an all-or-nothing notion, it can be weak or strong and it makes sense to have privacy levels. Privacy is essentially personalized, as it has been pointed out in (Xiao and Tao, 2006). A universal protection level may result as insufficient for some parties and excessive for others. What will cause offence strongly depends on individuals. To avoid undesired protection, a patient may be allowed to require a privacy protection under a taxonomy. In Table 2 with anonymized *Identifiers*, personalized generalizations have been performed on *Quasi-identifiers* and on *Diagnosis* attribute values. The taxonomy tree, labeled at nodes, for *Sex* is:



and for *Age* is:



The unmodified (with respect to Table 1) *Diagnosis* attribute values in Table 2 mean that all patients have given their consent to allow releasing the true diagnosis for their diseases.

But 'asymptomatic hemophilia A' potentially allows association with more than one identifiable individual (not necessarily in the same table). An additional external knowledge about for instance Sarah's relatives allows the data recipient to know that they have the same disease, more or less seriously, Sarah has. Analogously, releasing data on an infectious disease does potentially compromise the privacy of individuals who live in close contact with the patient. Thus some patient's privacy options on sensitive attribute values that involve more than one individual may compromise the privacy of the involved individuals. The attacker model here is *semi-honest*, i.e. weak for what concerns computation capabilities but with the possibility of gathering external knowledge. Therefore, the attacker (may be the data recipient) might associate (with a probability depending on the disease nature) a diagnosis in the observed table with individuals outside the table, by exploiting only external knowledge. Thus the attacker is supposed to not spend work and time to circumvent privacy protection mechanisms in order to perform strong attacks.

We term as *multicarrier* every attribute value that involves more than one carrier individual. Otherwise the attribute value is termed as *singlecarrier* (e.g., every 'pneumonia'). Actually, 'asymptomatic hemophilia A' entails that patient's relatives (e.g., any parent, son, brother, sister) have with high probability the same disease, more or less seriously. All 'dominant autosomal diseases' (e.g., Huntington's disease, Neurofibromatosis, a form of hypercholesterolemia and a form of nanism) and 'autosomal recessive disorders' (e.g., thalassemias) are other examples of multicarrier diseases.

A privacy-aware socio-sanitary system might be unfair towards individuals having a crucial relationship with the information donor but who may even not know that they have a particular genetic trait.

4 INFORMED FAIR CONSENT

We say that a consent is *fair* if it does not potentially compromise the privacy of others. The Fair Consent Principle statement we are going to propose is an extension of the concise Informed Consent Principle in Section 2.

'Personal information that has been collected should have the consent of the information donor and should not compromise privacy of whoever is potentially identifiable by the donor's consent.'

Notice that 'potentially identifiable by the donor's consent' is a difficult to quantify privacy threat. Nonetheless, it increases with increasing ubiquitous embedded computing, cloud computing. Once biological scientific results (as *DNA*) have added the *biological family* notion to the traditional *legal family*, a patient with a multicarrier disease may not know to have genetic relatives (e.g., biological fathers, biological parents) somewhere and thus may give informed consent to publish personal genetic data. In most cases, it may be impossible or impractical to obtain consent from them. Such relatives could be more and more easily identifiable once the patient's data can be linked to data in remote sites gathered from embedded devices. Moreover, seeking information about the medical history of a biological family may reveal information that was not intended for disclosure about unexpected individuals.

5 FAIR PERSONALIZED PRIVACY

National and international privacy legislation norms require implementable notions in order to be effective. Often principles are not a practice also because of implementation difficulties. The question whether the Fair Consent Principle can be effective in particular for what concerns the tradeoff between privacy protection and released data utility naturally arises.

A strong privacy preservation may determine that released data become useless. Actually, whereas adopting the strongest security technology which is at disposal is never dangerous (although it may have a high technological cost), the strongest privacy technological solution might be dangerous in some contexts. The contrast between privacy protection and data utility has received much attention in the last years. An acceptable trade-off between releasing information while preserving privacy is a major issue (Fung et al., 2010). Since medico-sanitary data accept a low information loss level, hiding or removing information from them may dramatically compromise their utility. To support data-mining tasks and to protect sensitive personal information, some well known anonymity methods have been proposed. Namely, suppression of *Identifiers* has been combined with generalization or perturbation on *Quasi-identifiers*. Specifically, generalizations (e.g., k-anonymity, l-diversity, t-closeness) essentially consist of grouping data into broader classes; data perturbation essentially consists of adding noise (e.g., numerical rounding, attribute random swapping, partially suppressing records). Several well-known papers have been published on merits and drawbacks of those privacy protection methods. In (Brickell, 2008) it is claimed, on the basis of some experimental results, that even a modest enhancing in privacy protection determines almost complete loss of data mining utility.

In the multicarrier attribute case, a social and ethical criterion may underline that the consent of just one element of a group of individuals is not enough. We here follow neither of the two extreme positions, namely on the one hand the impractical personalized policy which could force us to complete the obtained consent with that of all the remaining components of the group, and on the other hand the analogously strict criterion which could force us to suppress a record with a partial consent. Other solutions are possible which may be preferred taking into account the application domains. Two of them are:

1) generalizing also on *Sensitive Attributes* under a taxonomy;
2) releasing the actual sensitive values together with the maximal anonymity for all *Quasi-identifiers*.
We are going to shortly discuss such potential solutions.

1) If generalizations occur inside a taxonomy, we claim that real-world taxonomies in use have to be preferred. In the sanitary context, *Medical Taxonomies* are more or less comprehensive hierarchical medical coding systems which encode disease diagnoses into *medical codes*. Such a coding helps accessing to health records according to diagnoses, procedures for use in clinical care, health statistics reporting, research and education. A review of medical coding systems is in (Cimino, 1996). No accepted standard exists for coding patient information. Major coding schemes are usually compatible with the archetypal medical coding system, namely ICD (International Classification of Disease). Since medical knowledge grows with new terms to add and old ones to discard, ICD is revised at year intervals. ICD-9 and ICD-10 are the ninth and tenth editions, respectively. ICD-10-CM (Clinical Modification of ICD-10), will replace ICD-9-CM on October 2013.

After the Health Insurance Portability and Accountability Act of 1996 (HIPAA), using the Health Care Common Procedure Coding System (HCPCS) became mandatory. Only the governing body responsible for maintaining the code set has the authority to provide advice on how to apply a code set.

As in Figure 1, medical codes are numbers such that, for a finer granularity, digits are added in decimal place. Usually, decimal from .0 to .7 codes a more specific core term, .8 codes another category and .9 'unspecified'. In any case, a label is composed of the

286 Coagulation Defects
286.0 Non-specific Hemophilia A
286.01 Asymptomatic Hemophilia A
286.02 Symptomatic Hemophilia A
286.1 Non-specific Hemophilia B
286.2 Non-specific Hemophilia C
286.3 congenital deficiency of other clotting factors
286.4 Von Willebrand's disease
286.5 Intrinsic Anticoagulants
286.52 Acquired hemophilia
286.53 Antiphospholipid antibody with hemorrhages
286.59 Other hemorrhagic disorders
286.6 Defibrination syndrome
286.7 Acquired coagulation factor deficiency
286.9 Other and unspecified coagulation defects

Figure 1: Part of ICD-9-CM customized 'Blood Diseases'.

diagnosis term and of its medical code.

In Table 3, ICD-9-CM medical codes replace their corresponding diseases. The codes for blood diseases can be seen in Figure 1. The codes in records n.4, n.7 and n.9 have been generalized by replacing them with the most general code in Figure 1.

Table 3: Table 2 modified by using medical codes for *Diagnosis*, with a generalized code for 'asymptomatic hemophilia A'.

Name	*Age*	*Sex*	*Zipcode*	*Diagnosis*
1	[1 – 10]	M	[101 – 200]	482.4
2	[1 – 10]	M	[101 – 200]	482.3
3	[1 – 10]	M	[151 – 200]	482.4
4	[1 – 10]	M	[151 – 200]	286
5	[11 – 20]	M	[201 – 250]	482.4
6	[11 – 20]	M	[201 – 250]	482.3
7	21	F	580	286
8	[21 – 60]	F	[351 – 400]	482.3
9	[21 – 60]	F	[351 – 400]	286
10	56	P	330	482.4

A drawback for sensitive attribute generalization is that it may determine an information loss that might turns out as unacceptable for data mining applications in some contexts. Moreover, generalizing sensitive attributes gives rise to the so called 'divergence' drawback. Specifically, as 'Coagulation Defects' includes several variants of blood disease, the certain knowledge about 'asymptomatic hemophilia A' becomes an uncertain knowledge about all diseases concerning blood coagulation. The gain in privacy protection is questionable. Nevertheless, such a solution might be preferred in contexts where uncertainty is acceptable for the data mining to be performed.

2) Releasing the true sensitive attribute values and maximal anonymity for all the *Quasi-identifiers* could balance each other, as in Table 4, under the assumed

Table 4: Non-generalized *Diagnosis* multicarrier values while maximally generalized corresponding *Quasi – identifiers* values.

Name	*Age*	*Sex*	*Zipcode*	*Diagnosis*
1	[1 – 10]	M	[101 – 200]	482.4
2	[1 – 10]	M	[101 – 200]	482.3
3	[1 – 10]	M	[151 – 200]	482.4
4	[1 – 100]	P	[100 – 1000]	286.01
5	[11 – 20]	M	[201 – 250]	482.4
6	[11 – 20]	M	[201 – 250]	482.3
7	[1 – 100]	P	[100 – 1000]	286.01
8	[21 – 60]	F	[351 – 400]	482.3
9	[1 – 100]	P	[100 – 1000]	286.01
10	56	P	330	482.4

taxonomies for *Sex* and *Age*, and an analogous taxonomy for *Zipcode*.

The intended aim of such a solution (may be interesting for some statistical analysis) was to make anonymous individuals with a multicarrier disease. The fact that complete anonymity is never yielded here may be tolerated because the assumed attacker is *semi-honest*.

6 RELATED WORK

Problems related to the informed consent have been underlined especially for care delivering, as in (Applebaum et al., 1987). In the personalized privacy approach, generalizations under taxonomies often utilize simple ad hoc defined taxonomies, as in (Poovammal and Ponnavaikko, 2009), where patients are allowed to choose among three alternative options for data release (the true diagnosis, a less informative diagnosis, no diagnosis at all). In (Bertino et al., 2005) outsourced medical data are addressed for the identity disclosure at a non-personalized privacy setting, together with data ownership protection via digital watermarking. The binning algorithms there proposed for the information disclosure control on *Quasi-identifiers* exploit a *k-anonymity* that does not require all generalization nodes to be at the same level in domain hierarchy trees representing taxonomies. The information loss constraint is quantified as *usage metrics* for a maximal allowable information loss.
The multicarrier notion formulated in this paper is in (Chen et al., 2007) as a kind of external knowledge exploitable to discover sensitive data. *Multi-owner privacy* in (Li et al., 2010) and (Ren et al., 2011) and the *multiparty privacy* in, e.g., (Chen and Liu, 2009))cover privacy problems which arise when some parties jointly are actively involved in a computation task, usually on the web.

7 CONCLUSIONS

The claim of this paper was to discuss the informed consent for data release and to point out a social threat that may stem from some granted patient consents. A fair consent is proposed to enhance personalized privacy toward a fair personalized privacy, hopefully improving privacy protection in social services. If a privacy-aware system implements fair personalization, privacy assurance (that provides how much a party can trust a system as able to protect privacy) may be enhanced. Patients who have not enough trust in a health system's privacy protection capability might suppress some relevant information. This could lead to a poor care treatment and to an increased sanitary risk (e.g., if an infectious disease is omitted). We also claim that several implementations of the Fair Consent Principle are possible and that an acceptable tradeoff between privacy protection and data utility can be yield if implementations are tailored to the data mining requests.

Further Work. This work, as a position paper, leaves room for some developments, as for instance specific implementations, models and policies for a fair privacy. Let us mention just two scenarios. The one concerns fair consent policies for health and rights of donor-conceived children in an ubiquitous computing environment with weak control on the gathered data (due to possible data mining in countries with different laws). Such policies involve both technology and law, technology for effective solutions and for resilience to malicious managements. Another scenario concerns the fair consent model for so called dynamic data sets gathered from patients monitored via embedded devices during also their movements.

REFERENCES

Applebaum, P. S., Lidz, C. W. and Meisel, A. (1987). *Informed Consent: Legal Theory and Clinical Practice*. Fair Lawn, NJ: Oxford University Press.

Bertino, E., Ooi, B. C., Yang, Y. and Deng, R. H.(1987). Privacy and Owneship Preserving of Outsourced Medical Data. In *Proceedings of the 21st International Conference on Data Engeneering (ICDE 2005)*, IEEE Conference Publications, 5-8 April 2005.

Brickell, J. and Shmatikov, V. (2008). The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing. In *Proceedings of KDD'08*.

Chen, B.C., Lefevre, K. and Ramakrishnan, R. (2007). Privacy Skyline: Privacy with multidimensional adversal knowledge. In *Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB2007)*.

Chen, B.C., Kifer, D., Lefevre, K. and Machanavajjhala, A. (2009). Privacy-Preserving Data Publishing. In *Foundations and Trends in Databases*, Vol.2, Issues 1-2.

Chen, K. and Liu, L. (2009). Privacy-preserving Multiparty Collaborative Mining with Geometric Data Perturbation. In *IEEE Transactions on Parallel and Distributed Computing*, Vol.XX, No. XX. IEEE.

Cimino, J.J. (1996). Review Paper: Coding Systems in Health care. In *Methods of Information in Medicine*,35, 273-84.

De Capitani di Vimercati, S. and Samarati, P. (2006). Privacy in the electronic society. In *Proceedings of the International Conference on Information Systems Security (ICISS 2006)*, Kolkata, India.

Fung, B. C. M., Wang, K., Chen, R. and Yu, Ph. S. (2010). Privacy-preserving data publishing: A survey of recent developments. In *ACM Comput. Surv.*, 42(4), ACM, 1-53.

Li, M., Yu, S., Ren, K. and Lou, W. (2010). Securing Personal Health Records in Cloud Computing: Patient-Centric and Fine-Grained Data Access Control in Multi-owner Settings. In *SECURCOMM 2010, Security and Privacy in Communication Networks, 6th Int. ICST Conference*.

Poovammal, E. and Ponnavaikko, M. (2009). Categorical Grading Based Personalized Privacy. Preservation Against Attacks. *World Congress on Engineering*, Vol.1, 22-31.

Ren, Y., Cheng, E., Peng, Z., Huang, X. and Song, W. (2011). A privacy policy conflict detection method for multi-owner privacy data protection. In *Journal Electronic Commerce Research*, Vol.11, Issue 1, 22-31.

Royal College of Nursing Research Society (2011). *Informed consent in health and social care research. RCN guidance for nurses*, Second edition. Fair Lawn, NJ: Oxford University Press.

Solanas A. and Castella-Roca J. (2008). RFID Technology for the Healthcare Sector. *Recent Patents on Electrical Engineering*, vol.1, 22-31.

Xiao, X. and Tao, Y. (2006). Personalized Privacy Preservation. *SIGMOD*.