Protecting Doctors' Identity in Drug Prescription Analysis (Draft Version)

Václav Matyáš Jr.
University of Cambridge Computer Laboratory
Vaclav.Matyas@cl.cam.ac.uk

Abstract: This paper describes work undertaken to assure the privacy of doctors in a system that enables some other parties to analyse prescription information at a reasonably detailed level. Our task was to determine what "reasonably" should mean. This paper outlines risks to doctor privacy that we identified during field work and data analysis. We then describe the measures chosen to safeguard doctors' identity when they do not explicitly consent to its being known. Although we concentrate on the technical measures, some relevant organisational and contractual issues are also mentioned.

1. Introduction

The research exercise outlined in this paper was undertaken to develop reliable measures that would protect the identity of doctors while allowing analysis of their prescriptions at a reasonably detailed level. The work was undertaken for and on behalf of IMS Health, which plans to introduce a product named *Xponent* that will provide data on the prescribing patterns of individuals by geographic regions. This information will be used primarily by pharmaceutical companies to provide them with a better understanding of their customers' behaviour. This understanding will help each company to provide a better service to the prescribers through medical sales representatives. The most obvious improvement will be that each prescriber will be more likely to receive pertinent information from a representative. The pharmaceutical company will also benefit from an improvement in promotional efficiency by avoiding unwanted calls and mailings. IMS Health also intends to share this information resource with healthcare professional bodies to assist in improving patient care.

Our research focussed on presenting the data to the end-users in such a way that they cannot find out a doctor's identity unless she explicitly consents to it being known. It did not involve evaluation of the processes and measures involved in the data collection and initial processing. Thus we have taken care only of some specific issues otherwise involved in clinical system security as discussed, e.g., in [1], but rather of those relevant to the problem of inference controls [7]. Let us review the overall information flow:

1. Pharmacies participating in this project supply the data on a weekly basis. The communication modules are provided by the data collector and are integrated into the pharmacy software by its suppliers. *No patient identifying information is ever involved.* The data is processed before the transfer through two batches – one batch includes prescription information with scrambled doctor references and

another batch contains the scrambled doctor reference and doctor information. A trusted third party is involved in the process so the data collector is not able to identify doctors who did not agree to being identified. The first batch is encrypted with a public encryption key of the data collector (IMS Health) and the second batch is encrypted with a public key of the third party; both batches are sent to the respective parties. Only the third party, possessing the corresponding private key for decryption, can then recover the second batch data in the readable form. Analogously only the data collector can decrypt the first batch of data.

2. The third party then decrypts all its batches and links together all references of individual doctors, as different references for the same doctor are provided by different pharmacies. Doctor identification information is only revealed for doctors who consented. For all other doctors only unidentifiable pseudonyms will be provided. These references (identities or pseudonyms) are then sent to the data collector over a secured link.
3. The data collector receives this data, links it with the prescription information (using the doctor references), validates it and prepares it for end-user distribution. The data will be provided for particular districts, called *cells* in this paper.

The format in which the data is available to the end-users is crucial to protecting the non-consenting doctors' identity. This format has been the major concern of our research. One can easily imagine situations where releasing certain information could, e.g., lead to identification of doctors, such as prescription of rare drugs, absence from a practice leading to a decrease in prescriptions. Here the usual database inference risk [4, 5, 6, 7] is not much of a threat compared to other kinds of indirect (non-automated) inference and so-called social engineering.

Doctor	Drug 1	Drug 2	Drug 3	Drug 4	Drug 5	Drug 6
Smith	13	3	15	34	12	19
Jones	8	16	25	28	27	20
Dr 1	5	11	13	15	10	25
Dr 2	16	15	**49**	23	**3**	15
Dr 3	26	7	25	19	27	19
Dr 4	**35**	7	11	24	21	**2**
Others	56	25	71	64	39	22

Table 1 – Naïve view at the doctors' prescriptions in a cell.

2. Risks

It would not be secure to provide the data "as is" (Table 1). An experienced pharmaceutical representative might identify a doctor who uses a particular drug only rarely or another drug frequently. Doctors' prescribing trends are also available in historical perspective. This causes a high risk of revealing the identity of a doctor who has a temporary decrease of prescriptions due to holiday or illness.

On the other hand, it is worth noting that no single end-user of the data will be allowed to obtain the data for a large part of the drug market, let alone for the entire market.

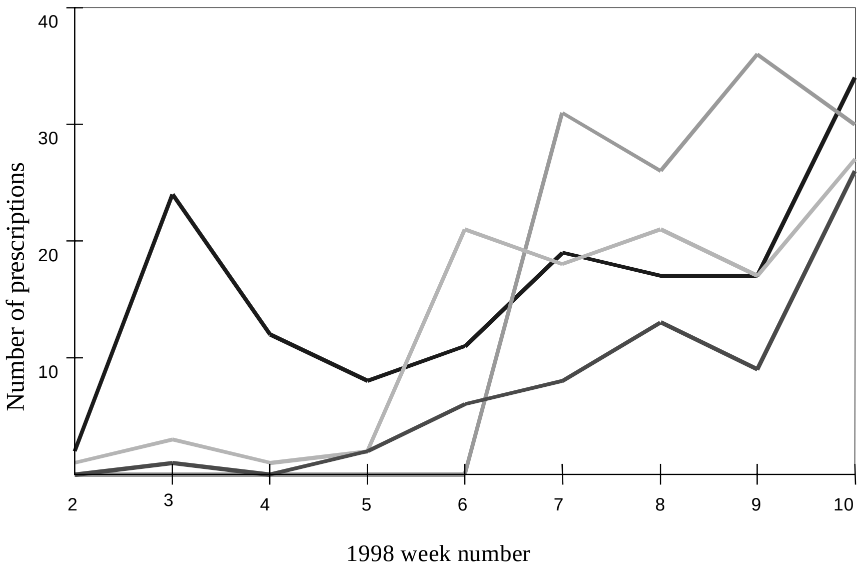

Figure 1 – Naïve view at the history trends for four doctors throughout nine weeks.

The following sections outline the conditions for setting up the regions for doctor cells, as well as structure of the cells. The formats of data available through the cell reports and solutions for providing the history data (e.g., Figure 1) are outlined briefly. We then overview the primary measures determining what data will be involved in the final stage of processing. Rather conservative measures have been suggested by the author for application in the first year of the project are described here. Throughout this period, both the number of consenting doctors and participating pharmacies (see Figure 2) should stabilise, whereas at present they are volatile. Once we have more experience with a large exercise of this type, it should be possible to adjust the measures accordingly. The paper is concluded by some additional technical and also non-technical suggestions for future research.

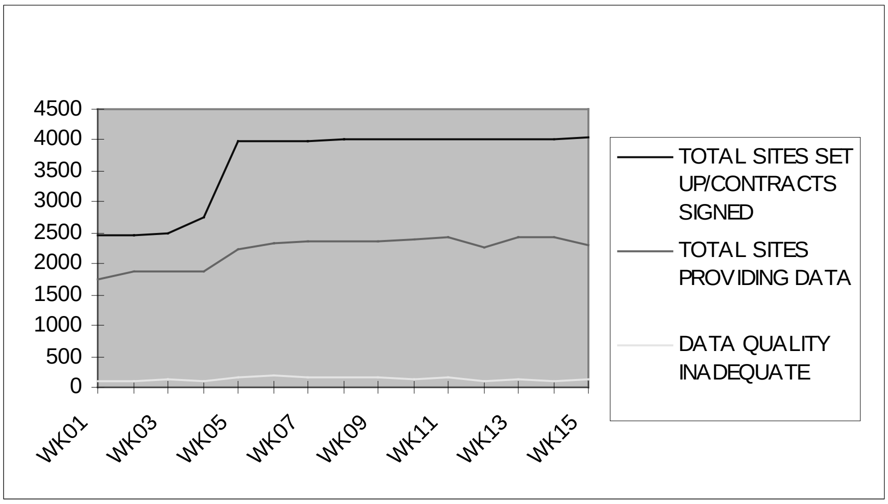

Figure 2 - Pharmacy shop variability

3. Cell-Level Data and History Trends

For the cell-level data only percentage shares will be released (number of prescriptions recorded for a doctor for a single brand of drug expressed as a percentage of her prescribing in the group of similar drugs), no absolute numbers are revealed at all (see the point below and Table 2). See the two following sections about the cell structure and about what data should and should not be reported in this format.

The total volume, in the terms of unprojected data, for the group of similar drugs is revealed only as a quintile (1 for top 20%, 5 for the bottom 20%). The quintile reflects the relative share of the group of drugs in the total volume of doctor's prescriptions for all drugs; these relative shares are then ordered per doctor and the first 20 % of doctors will be ranked at the level 1, next 20 % at the level 2, etc. This will then be stable even during periods of short absence like holiday or sick leave, when the total volume of collected prescriptions decreases, but the relative shares stay more-or-less the same.

Doctor	Rank	Drug 1	Drug 2	Drug 3	Drug 4	Drug 5	Drug 6
Dr 1	1	25%	13%	18%	12%	12%	20%
Dr 2	5	8%	16%	25%	11%	32%	8%
Dr 3	3	12%	21%	23%	25%	10%	9%
Dr 4	1	16%	25%	25%	9%	12%	13%
…..…	……	……	……	……	……	……	……
Dr 21	2	26%	7%	15%	19%	17%	16%
Dr 22	5	19%	13%	17%	23%	21%	7%
Others	1	18%	17%	22%	17%	12%	14%

Table 2 – View at the doctors' prescriptions in a cell.

The *history data will be available only in separate chunks for all doctors on a given period basis*. This data will also be anonymised for all doctors (including the consenting ones). The history view will be "started" at the same time and re-started *on a regular basis* for all doctors and cells, with the periods available for 4-6 months for the start, then after the real data analysis and possible threat assessment we can consider any period extension.

All newly moved doctors will re-appear only at the beginning of the period. Then in month 1 of the period (of say 6 months), only month 1 data is available, in month 2 data for months 1 and 2 is available, … in month 6, data for months 1 to 6 is available, and in month 1 of the next period only data for that month is available. The doctors' positions in the cell and the pseudonyms have to be re-shuffled at the start of each period.

4. Structure of a Cell

The size of a cell is determined primarily by the number of doctors included in that cell. The minimum number of doctors in a cell will be 15, but the average size of a cell will be 20 and the non-consenting doctors have to account for at least 50 % of the group.

The structure and size of a cell also have to reflect other needs and threats possibly due to privacy rivalry, breach of confidence by pharmacists, etc. After undergoing field research and interviewing representatives of the parties involved, we analysed the existing data available from about two thousand pharmacies supplying data for more than 10 weeks on a stabilised level of supply quality. Personal interviews reflected that a doctor would be able to identify prescribing patterns of her colleagues in a group of up to 10 doctors, who would be from two to three neighbouring practices. However, it was also noted by the interviewed that there often exist easier ways to get to their colleagues' through so-called “social engineering”. While we had to devise a scheme with a level of protection better than the standard NHS level of personal data protection (given, e.g. by [2,3]), we had to acknowledge that protecting the privacy against other players in the healthcare market in 100 % of cases is not possible. Analysis of the data using Microsoft Excel or the IMS Health Dataview package was undertaken on the data collection described above. We tried to identify the biggest cell, where identifying a doctor's prescription pattern, excluding rare and sensitive drug, was possible. This lead us to suggest the above techniques for history trends and for cell-level data, as well as to the following suggestions regarding the cell size and structure:

1. Number of GP practices per cell - minimum of four.
2. Number of pharmacies per cell - minimum of four (this, as well as the number of practices could be three in theory, but cells would have to be restructured if a pharmacy or practice participating in the scheme ceased to do so for whatever reason).
3. The non-consenting doctors listed in one cell shall be drawn from at least three different practices.

One of the most important facts is the one that *all doctors will be presented as anonymous for the cell view*. For the consenting GPs only, the data will be available (in absolute values, in contrast to the cell level data presentation outlined below) in a separate report.

5. Primary Measures

Drugs dispensed rarely and drugs that are sensitive for whatever reason are excluded from the low geographical level listings provided to the end-users. There are two mechanisms:

- A list of sensitive drugs (such as AZT) will be set up and such drugs will never be included in the data sets provided to end-users, below the top (i.e. national or regional) level.

- A minimum threshold for the number of prescriptions at the national level will be set up and drugs not reaching that threshold will not be included in any non-consenting doctor data.

Information on drugs just recently introduced to the market will not be available at anonymous doctor level during the first year of the project, and possibly not afterwards, depending on the result of the first review. However, it is interesting to look at this issue from the research perspective. To deal with new product introductions, it was suggested that doctor level information would be suppressed on these new products e.g., until

- The product sales have gone for a certain amount of time (e.g., 12 months)
- At least a certain percentage of doctors (e.g., 30%) at the national level have been prescribing the product
- At least a certain percentage (e.g., 20%) of doctors in every cell (or 90-95% of cells) have adopted the product.

One option may be to choose a solution involving the first and one of the other two measures. Another is to consider releasing total sales, new sales, repeated sales, number of prescribers lapsed and never prescribing – all these for larger groups (see Table 2 for cell-level data). What the data users really want to know is when the number of prescriptions increases in total, but only because of new prescribers and with a very limited number of repeated prescriptions. It might be very useful to find out about reasons for a new drug failure and this can be done best with doctors who agree to their identity be known even for new drug prescriptions. This latter option is yet to be considered.

Month	**1**	**2**	**3**	**4**	**5**	**6**
Total	10	15	18	25	33	35
New	10	8	12	19	25	28
Repeated	0	7	6	6	8	7

Table 3 – Possible format for new products' prescription reporting.

We also have to deal with situations, when a doctor practices only for a very limited time. That would imply that her total number of prescriptions is below a certain threshold. This threshold can be calculated as a percentage (e.g., 10%) of the national prescription average from the last data collection. If this happens, then the doctor's data will be included into the group "*others*". This group is normally used for cases like referring data that has not been correctly linked to a doctor, only to a pharmacy.

The group "others" shall always have at data equivalent to at least an average doctor, and this is to be continuously monitored. If this condition is not fulfilled, then the threshold for not listing a doctor will be increased, until enough doctors' data is included in "others". The doctor's data then actually reappears on the list only after the new time series are started (see below).

6. Conclusions

Even if we believe that the above scheme should provide a good level of protection, one should not rely only on such a belief. IMS Health has introduced *strict contractual obligations and controls of the data end-users* not to use the data beyond the specified scope, not to make it available to other parties, etc. However, as we are concerned only with technical measures in this paper, let us review some additional suggestions and possibilities for further research.

It is strongly suggested that an *independent penetration testing* of the devised scheme's strength in protecting non-consenting doctors' identity is undertaken after two months' data for the second time series is available (the data for the first time series will then be also available). This shall target both the doctors' identities and also the possibility of connecting together the consecutive time series of individual doctors, which was identified as the most possible way of attacking the identity protection scheme.

It should be also mentioned that the end-users will be contractually bound to follow IMS Health practice of deleting the individual-level data that is more than two years old. Following this way, cell-level data will be kept only for up to five years, after which only national-level statistics will be kept.

We have identified several additional measures than can be potentially used if "natural" variability of data decreases:

- Regularly keep 10-15 % (minimum of two doctors) missing from the detailed list (vary not only the actual doctors, but also the number of them "amalgamated" in "others") and their data be moved in "others".
- Reordering doctors in the cell everytime the data is provided to the end-users or in very short periods - this would mean no or limited individual time series.

Also, our field exercise has indicated that it would be ethical not to use the consent of the doctors as a one-off thing, but rather approach them regularly, and provide them with an analysis of the data that they would find useful for their work. We suggest, and IMS Health is to implement, that an annual letter should be sent to doctors, with thanks for their participation and allowing them to opt out of the scheme.

Acknowledgements

The author would like to thank IMS Health for providing the original challenge and and to them, as well as to numerous healthcare informatics enthusiasts, for assistance during the project. Particular thanks to Trish Knox, Gail Watts, John Foord, Sati Sian, David Twinberrow and Tom Bradberry. Ross Anderson also provided invaluable criticism and suggestions for improvement.

References

1. *'Security in Clinical Information Systems'*, RJ Anderson, Publication of the British Medical Association, January 1996

2. *'NWN Threats and Vulnerabilities'*, 5 April 1995, IMG Document NWNS/T1.22

3. *'Security Guide for IM&T Specialists'*, 3 April 1995, IMG Document NWNS/T5.11

4. *'Protecting databases from inference attacks'*, TH Hinke, HS Delugach, RP Wolf, Computers and Security v 16 no 8 (1997)

5. *'Neighbourhood Data and Database Security'*, K Yazdanian, F Cuppens, New Security Paradigms Workshops 1992 and 1993, joint proceedings published by IEEE

6. *'Information flow controls: an integrated approach'*, F Cuppens, G Trouessin, Third European Symposium on Research in Computer Security, Brighton, England, proceedings published as Springer-Verlag LNCS v 875

7. *'Cryptography and Data Security'*, D Denning, Addison-Wesley Publishing Co. 1983