The False Security of Blind Dates

Chrononymization's Lack of Impact on Data Privacy of Laboratory Data

J.J. Cimino

NIH Clinical Center, US National Institutes of Health, Bethesda, Maryland, USA

Keywords

Patient data privacy, data adjustments, clinical research, clinical informatics, health policy, anonymizatoin, de-identification, dates

Summary

Background: The reuse of clinical data for research purposes requires methods for the protection of personal privacy. One general approach is the removal of personal identifiers from the data. A frequent part of this anonymization process is the removal of times and dates, which we refer to as "chrononymization." While this step can make the association with identified data (such as public information or a small sample of patient information) more difficult, it comes at a cost to the usefulness of the data for research.

Objectives: We sought to determine whether removal of dates from common laboratory test panels offers any advantage in protecting such data from re-identification.

Methods: We obtained a set of results for 5.9 million laboratory panels from the National Institutes of Health's (NIH) Biomedical Translational Research Information System (BTRIS), selected a random set of 20,000 panels from the larger source sets, and then identified all matches between the sets.

Results: We found that while removal of dates could hinder the re-identification of a single test result, such removal had almost no effect when entire panels were used.

Conclusions: Our results suggest that reliance on chrononymization provides a false sense of security for the protection of laboratory test results. As a result of this study, the NIH has chosen to rely on policy solutions, such as strong data use agreements, rather than removal of dates when reusing clinical data for research purposes.

Correspondence to

James J. Cimino, MD
Chief, Laboratory for Informatics Development
NIH Clinical Center, US National Institutes of Health
10 Center Drive, Bethesda, Maryland, 20814, USA
E-mail: ciminoj@mail.nih.gov
Phone: +1–301–443–9696

Appl Clin Inf 2012; 3: 392–403
doi:10.4338/ACI-2012-07-RA-0028
received: July 13, 2012
accepted: October 1, 2012
published: October 24, 2012
Citation: Cimino J.J. The false security of blind dates: Chrononymization's lack of impact on data privacy of laboratory data. Appl Clin Inf 2012; 3: 392–403
http://dx.doi.org/10.4338/ACI-2012-07-RA-0028

© Schattauer 2012

Downloaded from www.aci-journal.org on 2012-10-24 | ID: 77027 | IP: 217.110.19.91
For personal or educational use only. No other uses without permission. All rights reserved.

1. Introduction

When considering personal privacy, the importance of health information protection is on par with that of financial information protection. Yet, society must weigh the potential individual harm that may result from the release of identifiable health data against the potential benefit to the public that may result from research carried out on those data. In particular, electronic health records (EHRs) contain sensitive, personally identifiable data that are of substantial value to biomedical research, ranging from basic understanding of disease processes to the determination of relative effectiveness of diagnostic and therapeutic options [1].

The solution to balancing these two motivations involves, in part, altering EHR data prior to secondary reuse in a way that minimizes the risk of disclosing sensitive information – for example, a person's past and current medical problems, laboratory test results and medications – while preserving the ability of researchers to learn about the human condition in general – for example, the relationships among those problems, test results and medications. An obvious first step in this process is the removal of information from the data that would allow the association of the data with the person from whom they were derived; this is referred to as *de-identification*. The ideal goal is to alter the data sufficiently that they can never be related back to their origins, referred to as *anonymization* [2].

De-identification efforts that seek to produce anonymous records will naturally include the removal of names and facial photographs. Dates and times found throughout the medical record are also a frequent target of de-identification processes either through complete removal – a process we refer to as "chrononymization" – or through some perturbation, usually a shift forward or backward by some undisclosed time span. The widely held view is that patient records from which dates and times have been altered or removed will be safer from deliberate attempts at re-identification than if they were left intact. This paper reports experience to the contrary.

2. Background

2.1 General Approaches to EHR Data De-Identification

El Emam [3] considers – in decreasing order of re-identification risk – *directly identifying information* (which uniquely identifies an individual, such as a Social Security Number, or a combination of name and address), *indirectly identifying relation information* (which probabilistically identifies an individual, such as a combination of demographic and geographic data), *indirectly identifying transactional information* (in which probabilistic re-identification is complicated by the multiple instances of such data for an individual, such as diagnoses and medications, that are also common to many other individuals), and *sensitive information* (such as laboratory test results). In particular, he notes that this last type is "rarely useful for re-identification purposes" [3].

The Health Insurance Portability and Accountability Act of 1996 (HIPAA) [4] includes a privacy rule that, among other issues, distinguishes a *limited data set* (one in which only direct identifiers have been removed) from a *de-identified set* (from which direct and indirect identifiers have been removed). According to HIPAA, the removal of 17 specific types of elements, along with other unique identifying characteristics, meets a "safe harbor" standard for de-identified data sets and permits their use without restriction. The list of elements includes dates that might link the data to an individual either through an indirect relation (such as a birth date) or an indirect transaction (such as a hospital admission date). The Department of Health and Human Services is currently considering additional restrictions on the inclusion of dates [5].

As a result of HIPAA and other legal concerns, healthcare institutions that make their data available for reuse for research strive for full anonymization through a variety of means that go beyond the safe harbor requirements, such as providing only aggregated results with minimum sample sizes and blurring specific numeric results [6]. Methods for obscuring date information include random replacement of numeric values and resetting all dates to be relative to some arbitrary random date [7]. In general, the goal of these efforts is to assure that resulting data sets have some minimum number of individuals, often specified by the symbol κ, so that the known individual cannot be distinguished from others in the cohort, producing "κ-anonymity" [8].

Downloaded from www.aci-journal.org on 2012-10-24 | ID: 77027 | IP: 217.110.19.91
For personal or educational use only. No other uses without permission. All rights reserved.

The objective of these methods is not simply to prevent the accidental disclosure of patient identity – for example, the inadvertent display to a researcher of an acquaintance's medical history. There is real concern that a malicious agent might deliberately attempt re-identification of a patient's record (the target) for personal or corporate gain, such as a check of employability or insurability. The presumption is that such an agent might have some identifiable information about the target, from either their own data or a publicly available source, and could then match those data to a unique individual in a de-identified data set. The link then allows the association of additional data to the target. The ability to involve genetic data, as either the link or the disclosure, is particularly problematic [9], especially when associated with hospital visit data [10]. Although data providers do not generally make explicit their reasons for removing or altering dates in their data, their efforts to do so clearly imply that they consider such information to be helpful to a re-identification attack.

2.2 The Biomedical Translational Research Information System (BTRIS)

One source of de-identified data sets is the Biomedical Translational Research Information System (BTRIS), a trans-institutional resource at the National Institutes of Health (NIH) that provides researchers with access to clinical research data [11]. BTRIS contains data on over 340,000 research subjects seen at the NIH from 1953 to present. It includes data from many different NIH systems, including two EHRs used at the NIH's Clinical Center (the hospital of the NIH at the Bethesda, Maryland campus), from 1976 to 2004 (known as the Medical Information System, or MIS) and 2004 to present (the current Clinical Research Information System, or CRIS). Data in BTRIS include those typically found in EHRs (clinicians' orders, test results, and documents related to patient care) as well as research-specific data (such as case report forms). The data are in fully identified form. Elements that are coded with controlled terminologies are represented with a single local ontology, called the Research Entities Dictionary (RED) that serves as a meta-thesaurus of all the local coding terminologies used over the years across the various data sources. Investigators at NIH can issue their data queries directly using the BTRIS system and obtain immediate results.

One of the principle purposes of BTRIS is the provision of data in de-identified form for NIH investigators who seek to reuse the data to study research questions. While the process for providing de-identified text data (such as clinicians' notes and diagnostic procedure documents) is still in development, certain data are currently provided in de-identified form by simply omitting certain metadata elements. For example, laboratory test results and vital signs measurements are provided without any of the patient identifiers from the "header" portion of the reports and are simply labeled with temporary, sequential patient identifiers.

The availability of BTRIS at the NIH has naturally required careful application of, and adherence to, NIH policies. In some cases, this has resulted in clarification of pre-existing policies to extend to the new capabilities with which investigators find themselves empowered. Although the NIH Office of Human Subjects Research Protections (OHSRP) did not previously require removal of dates from de-identified data sets, the review of policies as they related to BTRIS offered the opportunity to consider this restriction, in light of efforts to do so at other institutions (as noted above). With full recognition that the "chrononymization" of data might adversely affect the usefulness of the data for research, the question was therefore asked: To what degree will removal of dates improve subject privacy in a typical data set, such as the results of common laboratory panels?

3. Methods

3.1 Study Design

In order to resolve this question, the NIH undertook an analysis of the BTRIS repository to determine the relative potential for re-identification of data with date information removed as compared to the same data with date information intact. The study was restricted to common laboratory test results and considered a hypothetical scenario in which a malicious agent has possession of a single identifiable panel of test results and wishes to compare these data with a large, de-identified data set to see if a match can be found. To simulate this situation, BTRIS provided large data sets, from which

Downloaded from www.aci-journal.org on 2012-10-24 | ID: 77027 | IP: 217.110.19.91
For personal or educational use only. No other uses without permission. All rights reserved.

a random subset of results were extracted; the individual results in the subsets were then compared to the larger sets, with and without dates, to see how often the results would match only the original subject's record.

3.2 Source Data Sets

The BTRIS database contains two tables of precompiled test results that have been created to handle common data requests. One table corresponds to the hematologic data found in a typical complete blood count and the other corresponds to twenty types of clinical chemistry tests that are often performed together in one or more test panels (Chem20). Each table contains columns that correspond to classes of test concepts in the RED. For example, one column in the Complete Blood Count table contains all whole blood hematocrit results, while one column of the Chem20 table contains all intravascular sodium ion measurements. In both cases, results are drawn from MIS, CRIS, and other NIH systems and include all available results for the NIH subject population. Thus, individual patients may have many records in each table.

For the purposes of this study, only the columns corresponding to the five most common results of the Complete Blood Count (*CBC*: white blood cell count, red blood cell count, hemoglobin, hematocrit and platelet count) and those corresponding to the seven most common results the Chem20 (*Chem7*: sodium, potassium, chloride, bicarbonate, glucose, blood urea nitrogen and creatinine) were considered. Data extraction included dates and times, as well as columns containing the specific test results of interest. Data were tagged with temporary identifiers for each subject. Only rows that contained numeric data in all of the columns of interest were included in these *source sets*. Although the study was performed as part of system development, rather than as biomedical research per se, the approval for the study was nevertheless obtained from the OHSRP.

3.3 Search Set Creation

Ten thousand (10,000) records were randomly extracted from each of the two source sets to create the *search sets*. The randomization process involved the division of each unsorted source set into 10,000 partitions and then selecting one record at random (using a random number generator set to return an integer between one and the number of records in the partition) from each partition. No attempt was made to exclude multiple records for the same subject.

3.4 Re-Identification Process

Each record in each search set was compared to each record in the corresponding source set. Comparisons were made for each permutation of an increasing combination of elements, including single elements (five for CBC and seven for Chem7), pairs of elements (10 pairs for CBC and 20 pairs for Chem7), all the way to the full set of elements for each panel (one set of five elements for CBC and one set of seven elements for Chem20). In all, 156 permutations were examined (2^5–1, or 31, for CBC and 2^7–1, or 127, for Chem7). Comparisons were made using date and time, date only, and without either date or time. Matching results are reported as sums of element sets of identical size (one through five for CBC and one through seven for Chem7).

All matches were counted as true positives if the temporary subject identifiers matched and as false positives otherwise. At least one true positive match was expected for each search record, since it was still present in the original source set. Results are reported in two forms: the *average positive predictive value* (APPV; defined as the total number of true positives divided by the total number of matches) for all the records in the set and as the *match rate* (MR; defined as the percentage of records that matched only to the correct subject) for the entire set. The APPV thus provides a measure of the average correctness of a particular "hit" in a set of matches when matching on a particular record, while the MR provides a measure of the likelihood that a particular record will identify only the correct record.

© Schattauer 2012

Downloaded from www.aci-journal.org on 2012-10-24 | ID: 77027 | IP: 217.110.19.91
For personal or educational use only. No other uses without permission. All rights reserved.

4. Results

4.1 Data Set Sizes

As of November 20th, 2011, the Complete Blood Count and Chem20 tables BTRIS contained 2,854,162 and 3,059,981 rows, respectively. After filtering of rows containing non-numeric or null values in the results columns of interest, the CBC table was reduced to 1,624,750 records (with data on 157,932 unique subjects) and the Chem7 table was reduced to 2,239,603 records (with data on 184,716 unique subjects). The 10,000-record CBC and Chem7 search sets contained data on 8,068 and 8,422 unique subjects, respectively. ▶Table 1 shows the means and standard deviations for each of the CBC and Chem7 values in the source and search sets.

4.2 Matching Search Sets to Source Sets

As expected, the accuracy of the match with a single data element was very poor when no date information was used (CBC: APPV 0.0002, MR 0.0006; Chem7: APPV 0.00001, MR 0). Also as expected, matching improved moderately with the use of date (CBC: APPV 0.5684, MR 0.5378; Chem7: APPV 0.0867, MR 0.0750) and even further with the use of date and time (CBC: APPV 0.9946, MR 0.9946; Chem7: APPV 0.9398, MR 0.9431).

The poor performance of matching without date information was quickly overcome with the addition of multiple data elements to the comparison. ▶Table 2 shows the performance of comparisons without dates, with date only, and with date and time as the number of data elements was increased. Of note, even when no date or time information was used, the APPV and MR increased to 0.9925 and 0.9926, respectively, with four CBC data elements; with all five elements, only one out of 10,000 records matched a single false positive, (APPV and MR 0.9999).* For the Chem7, when no date information was used, the APPV and MR increased to 0.9570 and 0.9674, respectively, with six elements; with all seven elements, only 114 out of 10,000 records matched one or more false positives. ▶Figures 1 and 2 depict these performance characteristics graphically.

5. Discussion

This study demonstrates that a malicious agent, armed only with a single, identified, dated laboratory test result generally would have difficulty picking out the corresponding data from a large data set in which identifiers and dates had been removed. However, most common laboratory tests are performed as parts of panels. If a malicious agent has one result, he probably has a panel of results. While the use of multiple individual, independent results naturally increases APPV and MR (and, therefore, the chance of successful re-identification), this study shows that the whole is greater than the sum of the parts: the *association* between the results provides additional indirect identifying information. For a common panel like a CBC, the match rate is as good as, or better than, that of identifiers used for medical records and Social Security [12–15]. The removal of dates does not guarantee a $\kappa>1$ [16] and therefore, laboratory panels such as CBCs and Chem7s are, in effect, biometric identifiers, making the results themselves subject to HIPAA restrictions.

Other studies have found similar patterns in patient data. For example, Loukides and colleagues showed that diagnosis codes, without date information, can be used to uniquely identify patient records [17], while unpublished findings by McCoy, Malin and Miller showed that sequential series of test results provide unique patient identification (Malin BA: personal communication). The logical implication of these findings is that, for the types of data that occur in patterns unique to a particu-

* In fact, this false positive match was on the same day, a few minutes apart, and may actually be the same laboratory result being reported twice with two different medical records numbers. Technically, this could be considered a true positive for two different subjects.

Downloaded from www.aci-journal.org on 2012-10-24 | ID: 77027 | IP: 217.110.19.91
For personal or educational use only. No other uses without permission. All rights reserved.

lar patient regardless of the time, knowing the time point is not necessary for identifying the patient.

Meanwhile, the chrononymization methods applied by various institutions [7, 18] are not necessarily harmless. Although a particular method may have little adverse effect for a particular purpose, it may have a major effect for another. For example, setting all dates in a record to be relative to some arbitrarily chosen random date will preserve the temporal relationships among the specific events in the record, but may interfere with proper interpretation in relation to external influences, such as seasonal environmental factors like influenza epidemics. Applying different chrononymization techniques to data sets on a case-by-case basis adds administrative burdens, and might be thwarted by a determined malicious agent who requests the same data in different forms. The current study shows that altering dates is ineffective when multiple contemporaneous identified data elements are available for comparison with a data set in which data elements are also contemporaneous. Further de-identification would require removal of the temporal relationships among the data (including individual elements of laboratory test panels), which would drop the match rate close to, but not completely to, zero (as some single values in the test sets were actually unique in the search sets). The resulting data set would, however, be almost worthless for understanding human health and disease.

Various additional methods for altering patient data sets can increase the difficulty of re-identification, as depicted in ▶Figure 3. Each method, however, reduces the reusability of the data in ways that are unrelated to the actual need to know a patient's identity. For example, random perturbations of patient numbers or test results [6] will necessarily reduce the accuracy of conclusions drawn from the data and, in any case, may be filtered out through multiple queries for the same data set.

The data sets used in this study were limited to laboratory results from a single institution – one that specializes in unusual medical conditions. However, the tests chosen for the study are among the most commonly performed panels in any health institution and a large proportion of the source sets, and likely the test sets as well, were from test results performed on normal volunteers. The results of this study can be easily repeated at other institutions. Whether other types of clinical data (such as vital signs or radiologic studies) can be as easily re-identified without date information remains an open question.

This study presumes a hypothetical situation in which a malicious agent has access to part of a person's medical record and seeks to use it as a "key" to access to remainder, which might reside in a publicly available data set. In our data sets, all results for an individual patient share the same temporary identifier – a typical scenario for de-identified data sets intended for use in data mining or epidemiologic studies. Therefore, a malicious agent who knows one panel of results (simulated by our tests sets) can readily re-identify of all other results for the same patient.

It is important to note that in this situation, the de-identified data set must contain the original key; the results do not suggest that performing a new test will allow correlation with previous or future tests. Nevertheless, there are a variety of situations in which the hypothetical attack scenario might arise. For example, a practitioner who is privy to part of a patient's record (perhaps obtained prior to separating from a healthcare institution or transmitted as part of a referral record) might wish to use the partial record to learn additional information about the patient. An insurance company that obtains a clinical record as part of a justification for reimbursement might seek to obtain additional risk information about the insured patient. The likelihood of such scenarios is unknown. Malin and colleagues suggest that laboratory results are not often disclosed with identifiers beyond the healthcare setting [19], although it is common for insurance companies to obtain medical records for reimbursement purposes. Owners of data sets nevertheless appear to believe that the risk of such scenarios is significant enough that they consider the removal or obfuscation of dates when sharing their data.

Ultimately, the protection of personal health information cannot depend on data-alteration methods alone. After careful consideration of the costs, benefits, and reliability of different de-identification approaches, based in part on the data in this study, the US National Institutes of Health (NIH) has determined that the best approach to protecting the privacy of human subjects involved in intramural research must include a strong policy to deter re-identification. This policy includes restricting information access to investigators who have completed training on privacy policy and security, who have a formal relationship with the NIH that includes the potential of disciplinary actions, and who have a legitimate need for the information. These investigators sign terms of use that include, in part, the assertion that they will not make a conscious attempt to re-identify these data.

Downloaded from www.aci-journal.org on 2012-10-24 | ID: 77027 | IP: 217.110.19.91
For personal or educational use only. No other uses without permission. All rights reserved.

Action to the contrary would be considered scientific misconduct, which is dealt with harshly. As a result, the use of limited data sets (direct identifiers removed, but times, dates, and other information retained) is permitted by appropriate personnel who have documented a legitimate purpose.

The restriction on access to clinical research data is regrettable, since this presents obstacles not only to malicious efforts, but to well-intentioned, ethical ones as well. If there was assurance that no external, identifiable information was available on the research subjects, perhaps such restrictions would not be required. However, privacy is a contextual phenomenon. Any patient attribute might, in a particular circumstance, be considered to be sensitive information. In an age when anyone can locate the owner of a personal media player based on the musical recordings it contains [20], or can re-identify an "anonymous" rape victim in a newspaper article, using only a few simple searches in Google [21] protection of sensitive data must extend beyond reliance on technical solutions to include the behavior of information systems users as well.

6. Conclusion

This study found that the removal of date information from panels of laboratory test results provides only a false sense of security about improvement in the privacy of clinical research data. Rather than risk reducing the usefulness of the data it collects, the NIH has chosen to allow limited data sets to contain dates but requires users to acknowledge that attempts at re-identification constitute scientific misconduct.

Clinical Relevance

This study examined the uniqueness of laboratory test results commonly seen in clinical settings. Removal of dates from such data is often held to be helpful for reducing the ability to re-identify such data but the current study showed that such removal has almost no effect the re-identifiability of such data.

Conflict of Interest

The author declares that he has no conflicts of interest in the research.

Human Subjects Protections

This study was approved by the NIH Office of Human Subjects Research Protection.

Acknowledgements

The author thanks Adam Wilcox for stimulating discussions on this topic, Vojtech Huser, Xia Jing and Andria Cimino for helpful input on a draft of the manuscript and Bradley Malin and other anonymous reviewers for suggestions on improving an earlier version of this paper. This study was conducted with intramural support from the NIH Clinical Center and the National Library of Medicine.

Federal Work for Hire

This paper is the result of a study that was conducted by the US Federal Government and is not subject to US copyright.

Downloaded from www.aci-journal.org on 2012-10-24 | ID: 77027 | IP: 217.110.19.91
For personal or educational use only. No other uses without permission. All rights reserved.

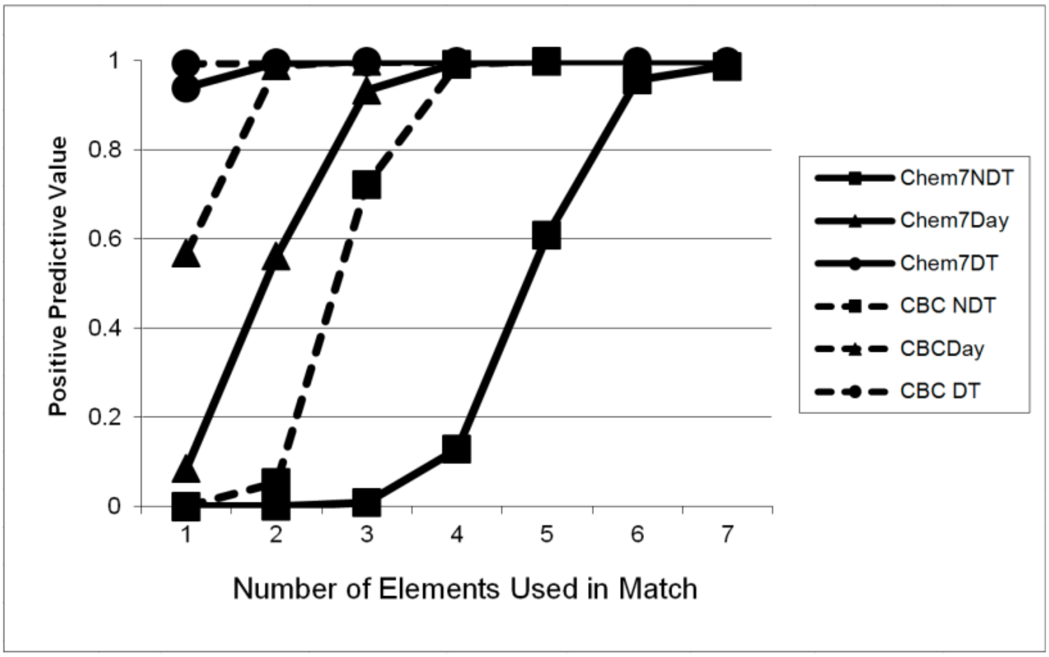

Fig. 1 Average Positive Predictive Value (PPV) for CBC and Chem7, without date or time information (NDT), with date-only (Day) and with date and time (DT). The graphs show that with no date information (circle), data are relatively unique (low PPV) when a small number of elements are available for matching, but rapidly become unique (high PPV) when more of the panel is available. While addition of date and date and time increase the PPV with few elements (moving the graphs to the left), there is essentially no effect when most or all of the panel is available, since PPV is high for all three cases. Note that CBC matches one to five elements, while Chem7 matches one to seven elements.

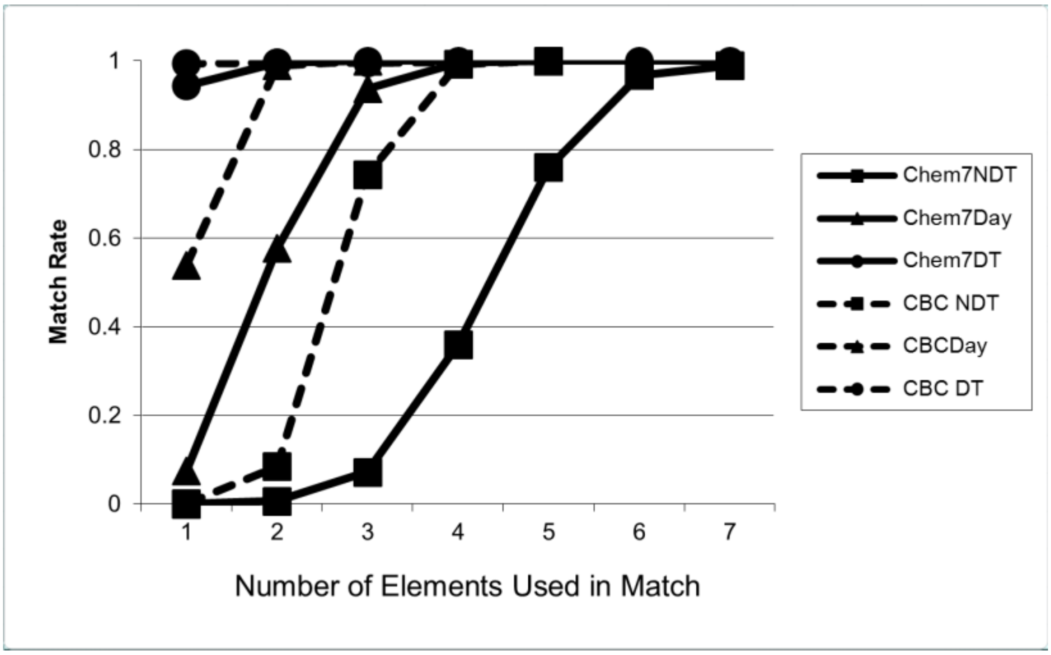

Fig. 2 Match Rate (MR) for CBC and Chem7, without date or time information (NDT), with date-only (Day) and with date and time (DT). The graphs show that with no date information (circle), data are relatively unique (low MR) when a small number of elements are available for matching, but rapidly become unique (high MR) when more of the panel is available. While addition of date and date and time increase the MR with few elements (moving the graphs to the left), there is essentially no effect when most or all of the panel is available, since MR is high for all three cases. Note that CBC matches one to five elements, while Chem7 matches one to seven elements.

Downloaded from www.aci-journal.org on 2012-10-24 | ID: 77027 | IP: 217.110.19.91
For personal or educational use only. No other uses without permission. All rights reserved.

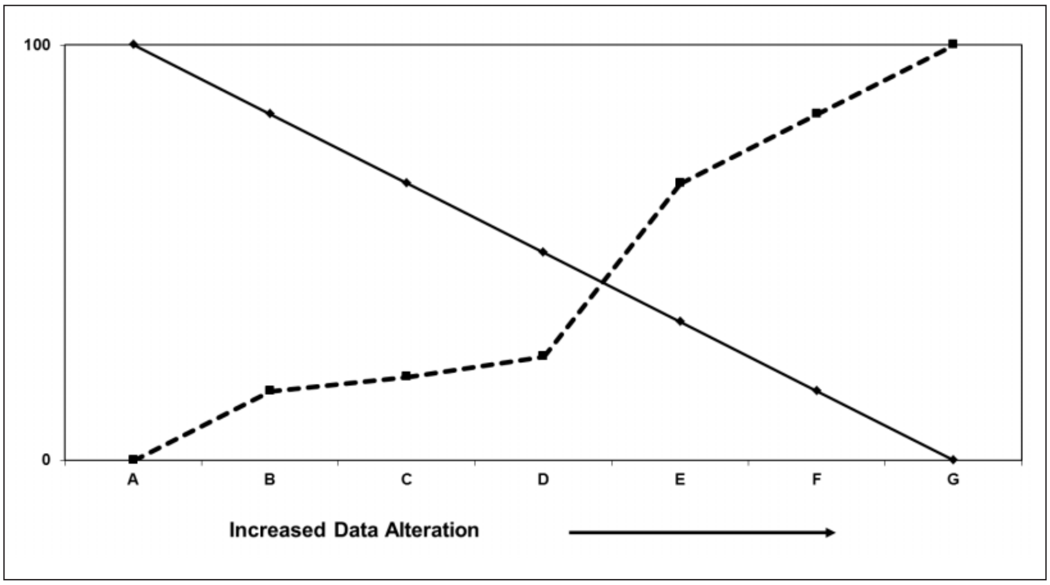

Fig. 3 The relationship between the difficulty of identifying subjects in a data set (broken line), and the usefulness of the data (solid line). The vertical axis represents the percentage of difficulty or usefulness from 0% to 100%; the scale is arbitrary. The definitions of the points on the horizontal scale are after El Emam [3] and are defined in the text. **A**: fully identified; **B**: removal of directly identifying information; **C**: Obscuring indirectly identifying transactional information; **D**: removal of indirectly identifying transactional information; **E**: Obscuring sensitive information; **F**: Decoupling sensitive information; **G**: Removal of sensitive information. The current study shows that there is minimal improvement in anonymization achieved by removal of date information in steps C and D.

© Schattauer 2012

Downloaded from www.aci-journal.org on 2012-10-24 | ID: 77027 | IP: 217.110.19.91
For personal or educational use only. No other uses without permission. All rights reserved.

Test	Source Sets		Search Sets	
	Mean	SD	Mean	SD
Sodium	138.58	3.49	138.53	3.45
Potassium	4.14	0.50	4.15	0.50
Chloride	104.25	4.44	104.45	4.34
Bicarbonate	26.17	3.49	26.12	3.52
Glucose	112.80	44.24	111.95	42.00
Urea (BUN)	17.21	13.77	17.29	14.57
Creatinine	1.09	0.84	1.09	0.81
White Blood Cells	7.406	10.564	6.818	5.037
Red Blood Cells	4.003	0.882	4.005	0.872
Hemoglobin	11.907	2.356	11.908	2.340
Hematocrit	35.332	7.030	35.336	6.978
Platelets	215.460	141.957	214.555	128.231

Table 1 Averages and standard deviations (SD) for each of the individual tests in the source and search sets.

Downloaded from www.aci-journal.org on 2012-10-24 | ID: 77027 | IP: 217.110.19.91
For personal or educational use only. No other uses without permission. All rights reserved.

Table 2 Performance of matches of Complete Blood Count (CBC) and the 7-test Chemistry Panel (Chem7) based on number of data elements used. APPV = average positive predictive value of all matches; MR = average number of records matched only to the correct subject. Each set of rows shows the sum of the matching results for all permutations of panel elements of a particular size. For example, the first set of rows show the pooled matching results for each of the five CBC elements or each of the seven Chem7 elements, while the last set of rows shows the matching results for the single set of all seven Chem7 elements. Note that the CBC panels contain only five elements and therefore do not show results for sets of size six or seven.

Elements	Date Match	CBC APPV	CBC MR	Chem7 APPV	Chem7 MR
1	None	0.0002	0.0006	0.00001	0
	Date Only	0.5684	0.5378	0.8667	0.0750
	Date and Time	0.9946	0.9946	0.9398	0.9431
2	None	0.0548	0.0842	0.0004	0.0067
	Date Only	0.9889	0.9890	0.5623	0.5770
	Date and Time	0.9999	0.9999	0.9963	0.9963
3	None	0.7217	0.7422	0.0089	0.0718
	Date Only	0.9999	0.9890	0.9347	0.9366
	Date and Time	0.9999	0.9999	0.9998	0.9998
4	None	0.9925	0.9926	0.1289	0.3598
	Date Only	0.9999	0.9999	0.9951	0.9951
	Date and Time	1.0000	1.0000	0.9998	0.9999
5	None	0.9999	0.9999	0.6098	0.7595
	Date Only	0.9999	0.9999	0.9997	0.9997
	Date and Time	1.0000	1.0000	0.9998	0.9999
6	None	-	-	0.9570	0.9674
	Date Only	-	-	0.9997	0.9997
	Date and Time	-	-	0.9998	0.9999
7	None	-	-	0.9886	0.9883
	Date Only	-	-	0.9997	0.9997
	Date and Time	-	-	0.9999	0.9999

Downloaded from www.aci-journal.org on 2012-10-24 | ID: 77027 | IP: 217.110.19.91
For personal or educational use only. No other uses without permission. All rights reserved.

References

1. Prokosch H, Ganslandt T. Perspectives for medical informatics: Reusing the electronic medical record for clinical research. Methods of Information in Medicine 2009; 48: 38–44.
2. Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. Journal of the American Medical Informatics Association 2007; 14(5): 550–563.
3. El Emam K. Methods for the de-identification of electronic health records for genomic research. Genome Medicine 2011; 3(4): 25.
4. http://www.hhs.gov/ocr/privacy/hipaa/administrative/
5. http://www.hhs.gov/ohrp/humansubjects/anprm2011page.html
6. Murphy SM, Chueh HC. A security architecture for query tools used to access large biomedical databases. Proceedings of the Annual Symposium of the American Medical Informatics Association 2002: 552–556.
7. Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. Journal of the American Medical Informatics Association 2007; 14(5): 550–563.
8. Sweeney L. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 2002; 10(5): 557–570.
9. Malin BA. An Evaluation of the Current State of Genomic Data Privacy Protection Technology and a Roadmap for the Future. Journal of the American Medical Informatics Association 2005; 12: 28–34.
10. Malin B, Sweeney L. How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. Journal of Biomedical Informatics 2004; 37(3): 179–192.
11. Cimino JJ, Ayres EJ. The clinical research data repository of the US National Institutes of Health. Studies in Health Technology and Informatics 2010; 160(Pt 2): 1299–1303.
12. Simon MS, Mueller BA, Deapen D, Copeland G. A comparison of record linkage yield for health research using different variable sets. Breast Cancer Research and Treatment 2005; 89(2): 107–110.
13. Brice JH, Friend KD, Delbridge TR. Accuracy of EMS-recorded patient demographic data. Prehosp Emerg Care 2008; 12(2): 187–191.
14. Beauchamp A, Tonkin AM, Kelsall H, Sundararajan V, English DR, Sundaresan L, Wolfe R, Turrell G, Giles GG, Peeters A. Validation of de-identified record linkage to ascertain hospital admissions in a cohort study. BMC Medical Research Methodology 2011; 11: 42.
15. Migowski A, Chaves RB, Coeli CM, Ribeiro AL, Tura BR, Kuschnir MC, Azevedo VM, Floriano DB, Magalhães CA, Pinheiro MC, Xavier RM. Accuracy of probabilistic record linkage in the assessment of high-complexity cardiology procedures. Revista de Saúde Pública 2011; 45(2): 269–275.
16. Malin BA. k-Unlinkability: A privacy protection model for distributed data. Data & Knowledge Engineering 2008; 64: 294–311.
17. Loukides G, Denny JC, Malin B. The disclosure of diagnosis codes can breach research participants' privacy. J Am Med Inform Assoc 2010; 17(3): 322–327.
18. Anderson N, Abend A, Mandel A, Geraghty E, Gabriel D, Wynden R, Kamerick M, Anderson K, Rainwater J, Tarczy-Hornoch P. Implementation of a deidentified federated data network for population-based cohort discovery. J Am Med Inform Assoc 2011, Aug 26. [Epub ahead of print]
19. Malin B, Loukides G, Benitez K, Clayton EW. Identifiability in biobanks: models, measures, and mitigation strategies. Hum Genet 2011; 130(3): 383–392.
20. Schnieder H. Via songs, lost iPod reshuffled to owner. The Washington Post, December 21, 2011; C1,3.
21. Brisbane AS. The Public Editor: Name Withheld, but Not His Identity. New York Times, December 17, 2011; Sunday Review:12.

Downloaded from www.aci-journal.org on 2012-10-24 | ID: 77027 | IP: 217.110.19.91
For personal or educational use only. No other uses without permission. All rights reserved.