BMC Medical Research Methodology

BioMed Central

Debate

Open Access

A proposed architecture and method of operation for improving the protection of privacy and confidentiality in disease registers

Tim Churches*

Address: Centre for Epidemiology and Research, New South Wales Department of Health, Locked Mail Bag 961, North Sydney NSW 2059, Australia

Email: Tim Churches* - tchur@doh.health.nsw.gov.au

* Corresponding author

Published: 6 January 2003

BMC Medical Research Methodology 2003, **3**:1

Received: 21 November 2002
Accepted: 6 January 2003

This article is available from: http://www.biomedcentral.com/1471-2288/3/1

© 2003 Churches; licensee BioMed Central Ltd. This is an Open Access article: verbatim copying and redistribution of this article are permitted in all media for any purpose, provided this notice is preserved along with the article's original URL.

Abstract

Background: Disease registers aim to collect information about all instances of a disease or condition in a defined population of individuals. Traditionally methods of operating disease registers have required that notifications of cases be identified by unique identifiers such as social security number or national identification number, or by ensembles of non-unique identifying data items, such as name, sex and date of birth. However, growing concern over the privacy and confidentiality aspects of disease registers may hinder their future operation. Technical solutions to these legitimate concerns are needed.

Discussion: An alternative method of operation is proposed which involves splitting the personal identifiers from the medical details at the source of notification, and separately encrypting each part using asymmetrical (public key) cryptographic methods. The identifying information is sent to a single Population Register, and the medical details to the relevant disease register. The Population Register uses probabilistic record linkage to assign a unique personal identification (UPI) number to each person notified to it, although not necessarily everyone in the entire population. This UPI is shared only with a single trusted third party whose sole function is to translate between this UPI and separate series of personal identification numbers which are specific to each disease register.

Summary: The system proposed would significantly improve the protection of privacy and confidentiality, while still allowing the efficient linkage of records between disease registers, under the control and supervision of the trusted third party and independent ethics committees. The proposed architecture could accommodate genetic databases and tissue banks as well as a wide range of other health and social data collections. It is important that proposals such as this are subject to widespread scrutiny by information security experts, researchers and interested members of the general public, alike.

Background

Disease registers aim to collect information about all instances of a disease or condition in a defined population of individuals. Usually the population covered by a disease register is defined geographically – for example, all people resident in a particular jurisdiction – and such registers are generally referred to a being "population-based". However, disease registers can also have a more limited scope, such as all people covered by a particular health insurer regardless of where they live, or all people using a

particular health facility, although such registers are often referred to as "research databases".

The core function of most disease registers is to measure the incidence or prevalence of their target disease or condition, although many registers have additional functions, such as providing population-based cases for case-control or cohort studies, and collecting information which can be used to monitor the effectiveness and efficiency of health care delivery [1].

Cancer registries are perhaps the best-known and well-established type of population-based disease register. However, in the last few decades, other types of disease register have started to appear. These include registers of birth defects, diabetes and chronic infectious diseases. This trend seems likely to accelerate, as technical advances in computing and digital communications decrease the cost of establishing and operating disease register databases, as well as broadening the scope of information which can feasibly be collected by them.

These advances also pose a number of concomitant challenges. One particular challenge – that of protecting individual privacy and maintaining confidentiality in an environment in which large volumes of health information can be copied and transmitted *ad infinitum* in just seconds – is attracting increasing attention from health care providers, regulators and consumers alike.

Anderson has observed:

"The likelihood that unauthorised use will be made of information is a function of its value and the number of people who have access to it; and consolidating valuable private information, such as medical records, into large databases increases both of these risk factors simultaneously" [2].

Examples of these concerns can be found in recent debates in Britain over the automatic transfer, either with or without explicit and informed consent, of personal health information to cancer registries, and in Iceland over the establishment of a far more general health research database [3][4][5][6][7][8][9].

This paper does not attempt to address the societal issues underlying such debates. It does however propose an information system architecture and method of operation which would enhance the protection afforded to the large volumes of highly confidential personal health information which disease registers and other centralised health databases necessarily accumulate.

Discussion

Traditional disease registers

Traditionally, disease registers have required that health service providers notify them of each case (instance) of the target disease or condition occurring in a population by sending them the medical or other substantive details of the case, together with identifying information for the person in whom the case has occurred.

Notifications to most disease registers need to be identified in this way to enable the register to assemble a single record for each unique case of the target disease from the multiple notifications which might be received about that case. For example, a patient might receive a clinical diagnosis of a particular type of cancer from their general practitioner (family physician), who will send a notification to the relevant cancer registry. A fine needle biopsy of the tumour may be taken, and this will result in the histopathology laboratory sending another notification to the cancer registry. The patient may then be admitted to hospital for surgery, which results in yet another notification of the same case to the cancer registry. In the absence of a universally shared electronic health record for each patient, such redundancy in the notification process is unavoidable, because each potential notifier has no way of knowing whether anyone else has notified that particular case to the relevant disease register. Redundancy in notification is also desirable because it minimises the likelihood of a case being overlooked by the disease register. However, it also means that the disease register must be able to determine that all these notifications relate to a single case of disease in a single individual. Typically, this is done by examining the identifying information associated with each notification and checking to see if that information matches with any cases already on the database maintained by the register.

Disease registers have successfully used this method of operation for many decades. However recent advances in computing, cryptography and communication networks have made alternative methods of operation feasible. Before considering one such alternative method of operation, some enabling technology and underlying concepts will be reviewed briefly.

Public key cryptography

Public key cryptography uses properties of large prime numbers to encrypt data using a pair of complementary keys (equivalent to passwords) belonging to each party wishing to exchange information in private. These keys are known as the public key and the private key. The public key is published and can used by anyone wishing to encrypt information in such a way that it can only be decrypted (read) by the holder of the matching private key, and by no-one else. In practice, for reasons of compu-

tational efficiency, public key encryption and decryption algorithms are used to pass random "session keys" securely between parties and these session keys are used with conventional encryption algorithms to protect the actual data – however, the effect is the same as if the entire message were encrypted or decrypted using public or private keys. Each party's private key can also be used to digitally "sign" messages to prove to the recipient that the party sending the message is in fact whom they claim to be and that the message has not been altered during the transmission process. Usually a trusted agency known as a certificate authority handles the distribution of public keys and vouches for the authenticity of these keys and the *bona fides* of the parties to whom they belong. Together, this technology is often referred to as "public key infrastructure" (PKI) [10][11].

Degrees of identifiability

The definitions of the degrees of identifiability of personal data which will be used in this paper are as follows. These definitions are not intended to be completely general, and relate only to "microdata", in which each record represents an individual or an event associated with an individual, and not to aggregated data. Quantin *et al.* provide a more general discussion of the issue of identifiability [12]. Standard nomenclature and definitions in this area are badly needed.

Directly identifying data items contains sufficient information to allow individuals to be identified or located easily in the absence of additional information. For example, name or residential address are directly identifying data items. Note that the identification does not need to be unambiguous – a particular residential address may be shared by a small number of individuals, but knowledge of it may still compromise the privacy or confidentiality of one of those individuals.

Indirectly identifying data items allow the direct identifiers of individuals to be found using additional, accessible external information. Most telephone numbers are *indirectly identifying* because the name, address and other characteristics of the individuals associated with that number can be obtained by reverse look-up of public telephone lists, or perhaps by simply calling the number. Whether identification numbers assigned to individuals by government institutions, or commercial organisations, are *indirectly identifying* depends on: how widely used that class of identification number is; and how easily accessible the additional information associated with the number is. Credit card numbers are *indirectly identifying* because a record of the card number plus the owner's name (and, often, their signature) is left behind with the vendor every time the card is used. In most cases social security numbers, national healthcare identification numbers or driver's license numbers should also should be regarded as *indirectly identifying* because the names, address and other personal details associated with them are easily accessible to very large numbers of public servants. The existence of legal and administrative sanctions against the misuse of such access does not guarantee that misuse will not, in fact, occur.

Potentially re-identifiable data contains sufficient, usually non-unique, data items to allow identification of individuals to whom the data relate with a reasonable degree of certainty. Additional information, which may require considerable effort to assemble or which may be not be publicly accessible, will often be required to achieve this re-identification. However, it can never be assumed that hostile third parties do not have access to such information – indeed, the range of additional information available to third parties can never be known in advance. For example, by using sources such as electoral rolls it may be possible to narrow down a combination of date of birth and locality of residence to just a handful of individuals. This issue has been investigated in detail by Sweeney and others [13][14].

Anonymous data does not permit re-identification of individuals with anything more than a negligible degree of certainty, no matter how much additional information is available to a third party.

The main design goal for the system proposed in this paper is the separation of *directly* and *indirectly identifying* data items from medical information and other substantive details at the earliest opportunity, and the rigorous maintenance of the separation thenceforth.

Elements of the proposed information system architecture

A *Disease Register* is an organisation which collects relevant information about all incident or prevalent cases of a particular disease or condition which occur in a defined population. The information collected usually comprises demographic attributes and details of the specific diagnosis, disease or condition, but may include information about the treatment, complications and outcomes in each case.

Health Care Providers are organisations or individuals which provide some form of health care service to patients (persons) and which are therefore in a position to capture the information about cases and related health events which might be required by a *Disease Register*. *Health Care Providers* include hospitals, general practitioners (family physicians) and pathology laboratories.

A *Notifiable Health Event* is any event about which a *Disease Register* requires information. Examples of *Notifiable*

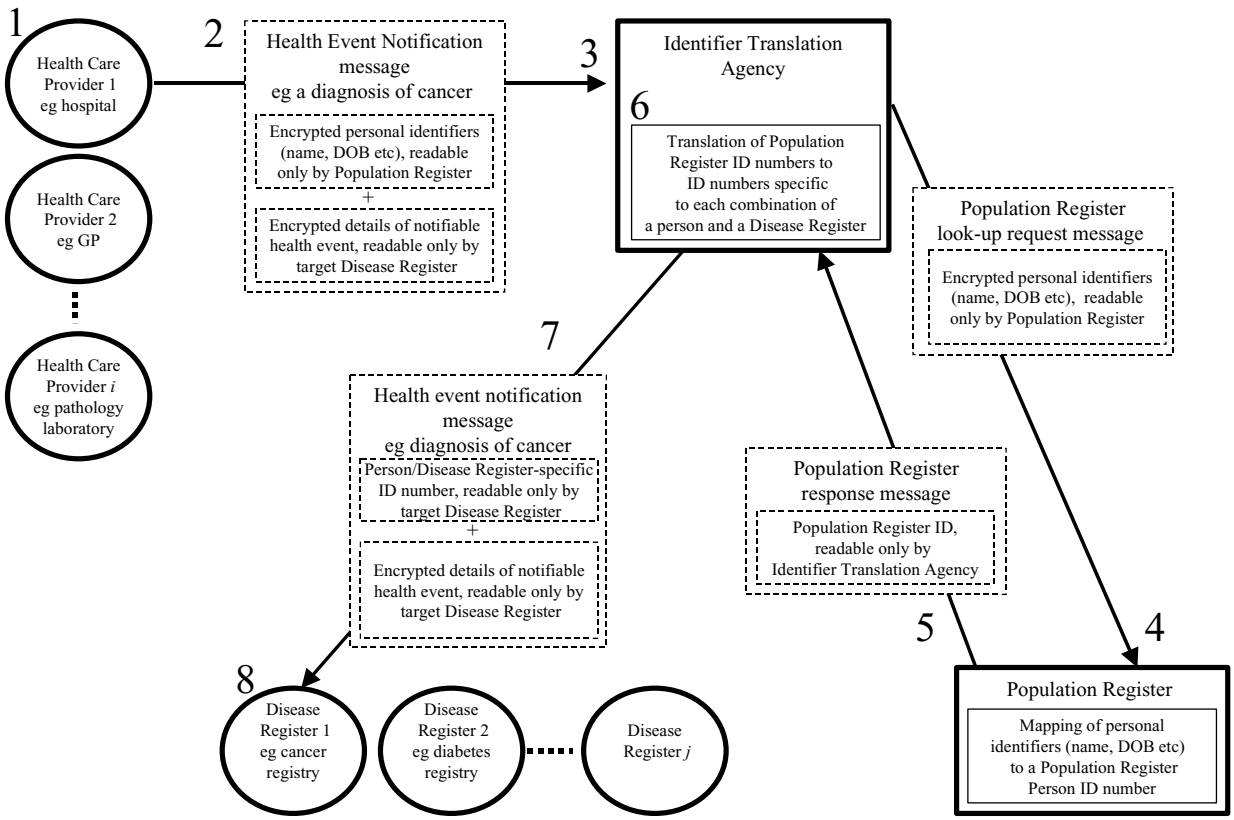

Figure 1
Method of operation. This graph should be examined in conjunction with the commentary provided in the text.

Health Events might include the diagnosis of a new case of cancer, an admission to hospital for a particular reason, or births and deaths (in which case the statutory body responsible for registering vital events is regarded as a type of *Health Care Provider*).

The *Population Register* is a trusted agency which is organisationally and physically distinct from all other parties which participate in the system. The function of the *Population Register* is to maintain a database of personal identifying information, such as name, date of birth, sex, country of birth and residential addresses. The database is used to assign a unique *Population Register Identifier (ID)*, typically a number, to each person of whom the *Population Register* is notified, which is not necessarily every person in the wider population. However, unlike other widely used unique health care identifiers, such as the NHS Tracking Number in the UK, the *Population Register ID* has very limited scope and is divulged to only one other party: the *Identifier Translation Agency*.

The *Identifier Translation Agency* is another trusted third party which is also organisationally and physically distinct from all other parties, including the *Population Register*, which participate in the system. Its role is to translate the unique identifier assigned to each person by the *Population Register* into a separate unique identifier which is specific to both each person and to each of the *Disease Registers* which participate in the system. This *person/Disease Register-specific identifier* (again, typically a number) also has very limited exposure and scope: it is shared only with the *Disease Register* to which it relates and with no-one else. The *Identifier Translation Agency* also provides temporary storage and forwarding facilities for encrypted messages.

Method of operation

The following operations correspond to the numbered data flows and procedures shown in Figure 1.

1. A *Health Care Provider* produces or captures information about a *Notifiable Health Event*, such as the diagnosis of a case of cancer.

2. The *Health Care Provider's* information system sends a *Health Event Notification* message to the *Identifier Translation Agency*. This message comprises two parts. The first part contains only the personal identifying details (such as name, address and date of birth) of the person to whom the *Notifiable Health Event* relates. These identifying details are encrypted by the *Health Care Provider's* information system using the public key of the *Population Register*, effectively rendering the information unreadable by any party other than the *Population Register*. The second part of the message contains only the medical or other details of the *Notifiable Health Event* in question, but not the personal identifiers of the person to whom it relates. This second part is also encrypted prior to dispatch, this time using the public key of the target *Disease Register* for this particular *Notifiable Health Event*. This renders the information unreadable by any party other than the target *Disease Register*.

3. Upon receipt of the *Health Event Notification* message, the I*dentifier Translation Agency* "unpacks" the two parts and tags each with the same arbitrary, unique random number (a "nonce") for tracking purposes. The first part of the message, which contains the encrypted personal identifiers, is forwarded to the *Population Register* in the form of a request to retrieve the *Population Register ID* for that person. The purpose of interposing the *Identifier Translation Agency* between the *Health Care Provider* and the *Population Register* is to prevent the *Population Register* from discovering the source of the *Health Event Notification* message and thereby being able to infer information about the *Notifiable Health Event* which triggered it. The *Identifier Translation Agency* may need to randomly delay, re-order or batch the messages it sends to the *Population Register* to achieve this goal. The *Identifier Translation Agency* also temporarily stores the second part of the *Health Event Notification* message, which contains the encrypted medical details of the *Notifiable Health Event* in question.

4. Upon receipt of a look-up request message, the *Population Register* uses its private key to decrypt the personal identifying information which the message contains and attempts to match this information against its database of persons. Probabilistic record linkage or other "fuzzy", error-tolerant matching techniques would be used for this matching or "lookup" operation, possibly assisted by human intervention where required. If a match can be made, then the previously assigned *Population Register* unique identifier (*Population Register ID*) for that person is retrieved, otherwise that set of identifying information is added to the database as a previously unencountered person to whom a new *Population Register ID* is assigned.

5. The *Population Register ID* which has been retrieved or assigned is encrypted using the public key of the *Identifier Translation Agency* and returned to it, together with the nonce, in the form of a response message.

6. The *Identifier Translation Agency* maintains a database which maps each *Population Register ID* to a series of unique alternative ID numbers which are specific to each combination of a person and a *Disease Register*. The *Identifier Translation Agency* uses its private key to decrypt the response message which it has received from the *Population Register* and extracts the *Population Register ID* contained in it, together with the nonce which identifies the message. The *Identifier Translation Agency* then uses the nonce to retrieve the temporarily stored second part of the *Health Event Notification* message which it received previously from the *Health Service Provider*. From this it determines to which *Disease Register* the information should be sent. It then uses the *Population Register ID* to retrieve from its translation table the corresponding *person/Disease Register-specific ID*, or if one does not exist, it assigns one.

7. The *person/Disease Register-specific ID* is encrypted using the public key of the target *Disease Register* and packaged with the retrieved medical details of the *Notifiable Health Event* (which are still encrypted with the private key of the target *Disease Register*) and the nonce. This package is sent as a message to the target *Disease Register*.

8. The target *Disease Register* decrypts both parts of the message using its private key and updates its database with the medical details of the person identified by the *person/Disease Register-specific ID*, without needing to know the identity of the person to whom that *person/Disease Register-specific ID* relates. The nonce is used to guard against replay attacks – the *Disease Register* database is updated with data associated with a particular nonce only once.

Table 1 illustrates this sequence of events expressed in protocol engineering notation.

In practice, each element of the system would acknowledge the receipt of messages and periodically re-send unacknowledged messages in order to guarantee delivery. These "handshaking" messages are not shown in Figure 1 in the interests of clarity. In addition, a cryptographic hash of each message would be encrypted by the message originator using its private key, and this electronic signature would be decrypted by each recipient using the public key

Table 1: Method of operation

Notation	
PT*j*	Patient *j*
HCP*i*	*Health Care Provider i*
ITA	*Identifier Translation Agency*
PR	*Population Register*
DR*i*	*Disease Register i*
NHE_{PID}	Personal identifying deatils for a *Notifiable Health Event*
NHE_{MED}	Medical details for a *Notifiable Health Event*
$\{NHE_{PID}\}_{KPR}$	NHE_{PID} encrypted with the public key of PR
$\{NHE_{MED}\}_{KDRi}$	NHE_{MED} encrypted with the public key of DR*i*
N	A *nonce* (number-used-once)
prlu()	*Population Register* look-up, returns a PRID
PRID	*Population Register* ID number
italu()	*Identifier Translation Agency* look-up, returns a PDRID
drlu()	Returns the name of a *Disease Register*, given a *nonce*
PDRID	*person/Disease Register-specific* ID number
drup()	Updates a *Disease Register* database with the NHE_{MED} for a particular PDRID.
Protocol	
1.	PT*j* $\rightarrow$ HCP*i* : NHE_{PID},NHE_{MED}
2.	HCP $\rightarrow$ ITA : $\{\{NHE_{PID}\}_{KPR},\{NHE_{MED}\}_{KDRi}\}_{KITA}$
3.	ITA $\rightarrow$ PR : $\{\{NHE_{PID}\}_{KPR}, N\}_{KPR}$
4.	PR : PRID = prlu(NHE_{PID})
5.	PR $\rightarrow$ ITA : $\{PRID,N\}_{KITA}$
6.	ITA : PDRID = italu(PRID, drlu(N))
7.	ITA $\rightarrow$ DR*i* : { PDRID,N,$\{NHE_{MED}\}_{KDRi}\}_{KDRi}$
8.	DR*i* : drup(PDRID,N,NHE_{MED})

This table should be read in conjunction with the commentary provided in the text.

of the message originator and verified against the actual message. The combination of strong encryption to protect the message "payloads", digital signatures to detect forgery and tampering, and handshaking protocols to guarantee delivery means that any widely accepted store-and-forward delivery mechanism, such as Internet (SMTP) e-mail, could be safely used to convey the messages between parties. The inevitable delays in transmission and processing of messages are unlikely to be a problem because disease registers are rarely required to operate on a real-time or near real-time basis.

Linking data for research purposes

So far, discussion has centred around the partitioning, at source, of notifications into: a) a *directly* and *indirectly identifiable* segment; and b) a *potentially re-identifiable* segment, and the strict maintenance of this separation subsequently. However, the system as proposed above also permits information from different *Disease Registers* to be efficiently linked at the level of individuals without the need for researchers or any of the *Disease Registers* involved to have access to any *directly* or *indirectly identifying* information.

We will first introduce some additional elements.

Privacy and Confidentiality Protection Committees (PCPCs) are independent bodies which authorise record linkage of data held by two or more *Disease Registers* and oversee the use of such data by researchers. The scope of the *PCPCs* oversight would be limited to the use of the data held by the *Identifier Translation Agency* and by the *Population Register*. Ethics committees or institutional review boards of each *Disease Register* would still need to approve the contribution of data to a cross-register linkage study.

Secure Research Facilities are also organisationally and physically distinct from other elements of the system. Their role is to provide a secure environment in which data from different *Disease Registers* can be linked at the person or event level and the resulting compound data sets made available to researchers for analysis. *Secure Research Facilities* never have access to *directly* or *indirectly identified* data, but they do receive *potentially re-identifiable* data on behalf of researchers. Only aggregated information is ever allowed to leave these facilities, after carefully scrutiny by facility staff to ensure that it is *anonymous*.

Researchers first formulate a research proposal which is submitted to a *PCPC* for approval. The *PCPC* then instructs the relevant *Disease Registers* to forward the required data to a nominated *Secure Research Facility*. Each *Disease Register* first maps the *person/Disease Register-specific IDs*, which it uses to identify each case or event, to a new series of arbitrary ID numbers which are specific to each research project and which are used only once, for that project. The *PCPC* also requests that the *Disease Registers* involved forward these mappings to the *Identifier Translation Agency*. Using these project-specific mappings and its database of correspondences between *person/Disease Register IDs*, the *Identifier Translation Agency* returns a mapping to the *Secure Research Facility* which allows it to deterministically (and thus, accurately and efficiently) link the *Disease Register* records it has received for that research project. These linked records are then made available to the researchers for use only within the physical and digital confines of the *Secure Research Facility*.

Some studies may require the direct follow-up of the cases known to a *Disease Register*. In these circumstances, names, addresses and other *directly-identifying* information could be supplied to researchers with the co-operation of the relevant *Disease Register*, the *Identifier Translation Agency* and the *Population Register*. The business rules under which the *Population Register* and the *Identifier Translation Agency* operate might stipulate that approval by two independent *PCPCs* is required before such a release of *directly-identified* information could take place.

In addition, *PCPCs* might be guided by information about individuals' wishes with respect to such direct participation in research. This information about individuals' wishes could be captured as part of the original notification to each *Disease Register*, using a standardised format similar to that being developed by the World Wide Web Consortium as part of the Platform for Privacy Preferences Project (P3P) [15]. This would allow direct contact by researchers to be restricted to those individuals who have indicated their willingness to be approached about follow-up studies or about novel uses of the information which they have contributed to particular *Disease Registers*. Caulfield *et al.* have suggested an authorisation model, as an alternative to "one-time consent", which would be very appropriate for the system proposed here [16]. Whether patients can opt out of having their medical details (but not their identity) automatically forwarded to a *Disease Register* depends on whether the *Disease Register* has been established under legislation which requires mandatory notification – for example, in Australia, notification of cancer to cancer registries is required by law.

Related work

The system as proposed above, which was first articulated in this form in [17], marries the well-established technology of public key encryption with a number of ideas which have been described previously in various guises.

The first idea is the separation of identifying details, such as name and date-of-birth, from all other substantive data items prior to the linkage of records between files. In 1979, Boruch and Cecil described a method for linking survey or administrative data files held by different agencies [18][19]. Agency A sends the identifying details plus arbitrary record-level code numbers, but no other information, to Agency B. Agency B then matches these identifiers to its own file, and after removing the identifying details, sends the file plus the matched Agency A code numbers to the researchers. Agency A also sends its file, similarly stripped of identifying details, to the researchers. The researchers then link the two files using the Agency A code numbers which are present in both files.

Pommerening *et al.* [20] described a technique for improving privacy and security in cancer registries, necessitated by changes in German privacy legislation, which achieves the same effect. Their method involves dividing the cancer registry into two operationally distinct offices. The first office receives notifications and handles all communication with notifying health care providers. The personal identifiers on each notification are encrypted before passing the records to the second office, which links the new data to its database using the encrypted identifiers. Blobel provides further details of this system [21]. Ho subsequently described and obtained a United States patent for a system which stores personal identifiers in a database which is administered separately from another database

used to store medical or other substantive details [22][23]. The first database maps *directly identifying* information such as name and date-of-birth into a code number which is used to identify records belonging to individuals in the second database. Medical or substantive data is encrypted by the originator using a key shared with the second database. This encrypted data is forwarded by the first database to the second database together with the appropriate code number. Quantin *et al.* [24][25] have described a system which uses hashed linkage keys, based on an ensemble of partial identifiers such as name, sex and date-of-birth, for epidemiological and health service evaluation follow-up studies. A United States patent has also been granted to the Ford Motor Company for a system which uses a "privacy data escrow agent" to perform a similar role to that of the *Identifier Translation Agency* proposed in this paper [26].

The second idea is the use of unique personal identification numbers with very limited scope. Szolovits *et al.* [27] and Anderson [28] have identified the hazards to personal privacy and confidentiality associated with simple unique health care identification numbers which have a wide scope. Kohane *et al.* subsequently proposed a framework of unique health care identifiers of limited or varying scope, known as the Health Information Identification and De-identification Toolkit (HIDIT) [29], as a means of avoiding some of these hazards.

Together, the *Identifier Translation Agency* and the *Confidentiality and Privacy Protection Committees* have similar roles to those played by the Personal Data Protection Authority in the Icelandic Healthcare Database [30]. However, in the Icelandic system, the Personal Data Protection Authority has access to both *directly* and *indirectly identified* information and the associated medical and genetic details of individuals, and thus represents a single point at which privacy and confidentiality might be compromised.

The proposed system also bears some resemblance to the use of "Federated Network Identities" and pseudonyms proposed by the Liberty Alliance Project, which is an initiative of a broad spectrum of industries intended to promote "e-commerce" without compromising the privacy and security of individually identitifying information [31]. However, the Liberty Alliance proposal allows pseudonyms to be shared amongst consortia of organisations, whereas the pseudonyms (that is, the *person/Disease Register-specific IDs*) in the system proposed here are specific to each *Disease Register* and are not shared with other organisations.

Most recently, Kelman *et al.* [32] have described data handling protocols for epidemiological record linkage studies which incorporate many of the preceding ideas. The administrative procedures which they describe would serve as an excellent foundation for a set of business rules governing the record linkage process as proposed above.

Distinguishing features of the system

The system proposed in this paper differs from those mentioned above in a number of important ways. Firstly, all *directly* and *indirectly identifying* information is effectively separated from the medical or other substantive details at the earliest opportunity – that is, at the source of the notification. Secondly, public key encryption is used throughout the system to maintain the separation between identifying information and medical details at all stages of processing. It is also used to ensure the authenticity, integrity and privacy of all message exchanges between parties to the system. Thirdly, a single *Population Register*, which is responsible for the linkage of all *directly* and *indirectly identifying* information, is effectively shared, via the *Identifier Translation Agency*, by multiple *Disease Registers*. Fourthly, the *Identifier Translation Agency* is used as a proxy to obfuscate the source of information flowing into the *Population Register* and to limit the scope and use of the unique IDs assigned to individuals by the *Population Register*. Fifthly, the *Identifier Translation Agency* does not have access to any privileged information – it acts solely as a conduit for encrypted messages which it cannot itself decrypt, and as a translator between sets of arbitrary ID numbers which have no intrinsic meaning or interpretation. In this way the hazards associated with the use of a unique personal identifying number, which has widespread currency throughout the health system, are avoided.

However, the most important feature of the system proposed here is that the improvement in the protection of privacy and confidentiality stems from its underlying architecture, rather than from the need for perpetual and unfailing observance of additional administrative and procedural safeguards by disease register staff. Anderson has identified the difficulty of effectively instituting "separation of duties" within a single organisation as a weakness in the security of many health information systems [33]. The proposed system structurally enforces those separations.

Disease registers have an excellent track record on security and the maintenance of confidentiality. However, it is important to recognise that as they and other electronic health data collections become more numerous and access to them is extended to more people, there is an increasing likelihood of accidental or deliberate breaches of confidentiality, possibly on a large scale. The protection provided by the system proposed here is based on the administrative, physical and digital separation of data, rather than on the assumption that people will always behave as they should.

Legal, legislative and governance considerations

Careful attention to issues of legal protection and governance would be required in order for a comprehensive system such as this to be accepted by the general community and by health care providers.

The *Identifier Translation Agency* requires the greatest legal protection because it holds the links between health information held by *Disease Registers* and information on individual identities held by the *Population Register*. Ideally, it would be established under legislative arrangements which provide it with complete independence from government departments and ministries, and with immunity from legal processes such as *subpoena* by courts. The *Identifier Translation Agency* and the *Privacy and Confidentiality Protection Committees* should also be legislatively bound to a set of operational rules which make explicit the ethical standards against which proposals to link information held by other elements of the system are evaluated. These rules, and the need for the absolute independence of the *Identifier Translation Agency* and the *Privacy and Confidentiality Protection Committees*, must be well understood and accepted by the community, which is effectively entrusting the ongoing protection of its privacy and confidentiality to them. Community ownership of the principles which underlie the system is also important in protecting it against possible malfeasance by future governments of unknown disposition and ideology.

It would be desirable if the *Population Register* and each *Disease Register* also had similar legislative protection and independent governance, but this is not essential. Indeed, the system could even be implemented within a single organisation, covering only one or a few data collections, provided that the *Identifier Translation Agency* was externally administered and adequate digital and physical independence of the various elements could be established and maintained.

It may be possible for existing organisations to fulfil some of these roles. Apart from some modifications to their information systems, the operation of *Health Care Providers* would not be affected. Existing *Disease Registers* could continue to function, but would no longer need to devote resources to the task of matching the identities in incoming notifications to their databases – this task would be ceded to the *Population Register*, hopefully with some gain in efficiency due to greater automation and economy of scale. In many jurisdictions it is likely that public- or private-sector organisations already exist with the technical and organisational capacity to undertake the roles of the *Identifier Translation Agency* and the *Population Register*. The key question is whether these organisations are sufficiently independent, both physically and administratively, from other entities in the proposed system to ensure adequate separation of roles. This requirement would probably rule out most government agencies, although publicly-owned independent corporations might be suitable. As noted previously, governance of these organisations needs to be via business and operational rules, preferably enshrined in legislation, rather by direct control by elected representatives or the executive arm of government.

Some form of centralised funding, such as a direct budget allocation by a government body, is likely to be necessary to establish the infrastructure required by the proposed system, particularly the *Identifier Translation Agency* and the *Population Register*. Grants-in-aid might also be made available to *Health Care Providers* and existing *Disease Registers* to assist with the modification of their information systems.

Many different mechanisms for ongoing funding are possible: for example, the *Identifier Translation Agency* and the *Population Register* might together charge each *Disease Register* an annual fee for the services which they provide. Similarly, *Secure Research Facilities* and even *Privacy and Confidentiality Protection Committees* might charge researchers a fee on a cost-recovery basis for the use of their services. Clearly some changes would be needed to the way in which research which uses disease registers is funded, but the overall cost to society of undertaking such research should be no greater than at present. Given that the proposed system would facilitate cross-register research, the societal cost-benefit ratio of operating disease registers may actually fall.

Risk and hazard assessment

When assessing the protection of privacy and confidentiality provided by a health information system, it is important to consider not only the risk of security breaches but also the hazards associated with them. Perhaps the worst-case scenario for a disease register, and therefore the maximum hazard, would be the misappropriation of the information held by the register and its publication on the Internet. Such misappropriation might be carried out by external attackers, or by internal staff who misuse their privileged access rights. It might involve direct access to the register database or eavesdropping on and incremental copying of notifications sent to the register. Although such scenarios are unlikely, they are nevertheless possible and must therefore be contemplated.

In the case of a conventional disease register which holds fully identified information, publication could be devastating for individuals whose details were released, and would almost certainly curtail further operation of the register (and perhaps others like it) as a result of public outrage.

For the system proposed here, such an event would still be serious, but not quite so disastrous. If the *Population Register* were compromised, then at worst a list of names, dates of birth, residential addresses and other demographic details of selected members of the population would be discovered. Publication of residential addresses and dates of birth may be distressing for some people. However no information about why the identities of particular individuals appear in the *Population Register* would be released, except for the inference that those people had at some stage been the subject of a notification to a *Disease Register*. As the number of *Disease Registers* which participate in the system increases, the impact of such inference declines. This could also be thwarted by "seeding" the *Population Register* with publicly available lists of identified information which cover a large proportion of the general population, such as electoral rolls or even telephone directories. The publication of the unique identification number assigned by the *Population Register* in conjunction with names and other identifying information would not compromise the entire system because the *Population Register ID number* is used only by the *Identifier Translation Agency*, which, like the *Population Register*, does not hold any medical or health information. Similarly publication of the identification number translation tables held by the *Identifier Translation Agency* would also have only a limited impact since the information has meaning only to *Disease Registers* and the *Population Register*. Because all elements of the system are physically and administratively distinct, operated by different staff and, ideally, using dissimilar computer systems, the probability that one element of the system is compromised, either by external attackers or malicious insiders, should be largely independent of the probability of compromise of any other element of the system.

However, a breach in the security of a *Disease Register* could still have serious consequences. Although each *Disease Register* does not hold any *directly* or *indirectly identifying* data items, the information which is held by it may still be *potentially re-identifiable*. It is therefore important that *Disease Registers* are supplied with only as much medical and other health information as they need to fulfil their core functions. The system proposed here would certainly not obviate the need for careful attention to physical, administrative and electronic security, particularly by *Disease Registers*. A means of dealing with this problem is suggested in a later section of this paper.

Another hazard is deliberate sabotage of one of the elements of the proposed system by an external or internal agent. The proposed system would not change the risk or magnitude of this hazard for *Health Care Providers* and *Disease Registers*. However, the centralised nature of the *Population Register* and the *Identifier Translation Agency* significantly increases the magnitude of the hazard posed by sabotage of these elements. For example, a saboteur might incorrectly merge identities maintained by the *Population Register*, which would result in erroneous merging of cases on *Disease Register* databases. The *Disease Registers* would have no way of knowing whether these changes were justified or not.

It is likely that such sabotage would eventually be detected through audits of the *Population Register* (described below), and the *Disease Register* databases could be corrected by reverting to back-ups and replaying corrected versions of their transaction logs. However such fixes would be costly and research projects undertaken with the incorrect data may be invalidated. One solution would be to operate two independent sets of *Population Registers* and the *Identifier Translation Agencies* in parallel. The risk that both sets of central agencies would be subject to sabotage simultaneously would be very small. *Health Care Providers* would need to send separately encrypted *Health Event Notifications* to each of the duplicate *Identifier Translation Agencies*. *Disease Registers* would only modify their databases if they received consistent information from both "arms" of the system. Business rules to resolve cases in which conflicting information was received would be required. Such redundancy would clearly increase the cost of establishing and operating the proposed system, but might also be justified on disaster recovery and other continuity-of-service grounds.

Extension of the system beyond disease registers

It is possible to extend the system proposed above simply by establishing additional registers. These registers could accommodate almost any type of laboratory or clinical data and could be established at a low marginal cost. This is because: i) the privacy protection arrangements for registers will already be in place and accepted by the community and health care providers; ii) one of the most expensive aspects of register operation, that of matching the identities of incoming notifications to records already on the register database, is handled centrally by the *Population Register*, with consequent economies of scale.

There are many additional types of register which could be established under the umbrella of this system. "Clinical databases" as described by Black and Payne [34] are obvious possibilities, but other candidates include "registers" containing the responses to population-based health status and health risk factor surveys, and even registers of social characteristics, such as unemployment status, income level and educational attainment. Controlled linkage of these "social characteristic registers" to various health status registers would permit the precise and ongoing monitoring of the effects which specific lifestyle and socio-

economic factors have on particular health outcomes, and vice-versa.

The system could also address many of the privacy issues associated with genetic databases. Genetic databases contain realised genotype information about individual subjects. With the advent of micro-arrays and other tools for the characterisation of individual genomes, tissue banks and blood sample collections should perhaps be thought of as databases of latent genotype information [35][36][37]. Tissue banks and blood sample collections would act as physical repositories for samples and as data stores for realised genotype information, but would be required to operate in conjunction with a "tissue register". Institutions would assign a unique, arbitrary ID number to each sample before sending the material to the tissue bank or blood sample collection. Institutions would also send a corresponding "notification" to the *Identifier Translation Agency*, comprising i) the demographic, medical and other details for each tissue or blood sample, encrypted with the public key of the tissue register; ii) the sample ID number, encrypted with the public key of the *Identifier Translation Agency*; and c) the *directly* and *indirectly identifying* personal details, encrypted with the public key of the *Population Register*. The *Identifier Translation Agency* would map the sample ID to a new, unique tissue register ID number and would forward this, together with the still-encrypted medical details, to the tissue register. The tissue register ID would also be mapped by the *Identifier Translation Agency* to the Population Register ID returned to it by the *Population Register*, in the usual fashion. Thus, tissue banks (and blood sample collections) would have access to *anonymous* genetic material and derived genotype information. Tissue registers would hold medical and demographic details for each sample, but no *directly* or *indirectly identifying* information, and would not have access to any genetic material or genotype information. Tissue banks would need to formally apply to a *PCPC* in order to obtain the authorisation required to link their samples to the corresponding medical and demographic details held by the tissue register. Similarly, tissue registers would require a *PCPC* to forward an authorisation to the *Identifier Translation Agency* before the tissue register could link its data to data held by a *Disease Register*.

Technical implementation

There do not appear to be any major technical impediments to the implementation of the system proposed in this paper. PKI software is widely available now that the United States patent on the most popular algorithm for public key cryptography has expired and export restrictions on cryptographic software have been relaxed by most governments. Standards and frameworks for the communication of structured health information, such as HL7 [38] or CorbaMED [39], are now widely accepted. The functionality of the *Population Register* is available in a number of off-the-shelve software products which conform to the CorbaMED Person Identification Service (PIDS) specification [40].

There are a number of reasons why it would be desirable to implement the system proposed in this paper using free, open source software components. [41] Firstly, much of the data processing infrastructure required by the system needs to be shared by many participants. It therefore makes sense to defray the cost of developing, customising and maintaining these components for many different computing platforms by making the program code freely available and modifiable under an open source license. Secondly, there needs to be a high degree of community confidence in the security and technical excellence of the system components. A fundamental principle of cryptography, enunciated 120 years ago by Auguste Kerckhoffs, is that the only parts of a cryptographic system which should be kept secret are the keys – it should be assumed that the implementation and algorithmic details for any system will eventually fall into the hands of the enemy, so it is best to make them available for open scrutiny from the outset [42]. Open source licensing would permit broad-based and ongoing auditing of the software components used by the system.

At the time of writing, at least one scalable, open source, probabilistic record linkage engine suitable for use in the *Population Register* was known to be under development [43]. Mature public key infrastructure and store-and-forward communication components are also freely available, but other parts of the system would need to be specially written. There is no reason why this work could not be shared by groups located in many countries.

Weaknesses and possible solutions

There are a number of potential weaknesses in the system proposed in this paper. These include its apparent complexity and the need for all participating parties to adopt and adhere to information standards and protocols. Despite the apparent complexity, it should be possible to promote the system to legislators and the general public in quite simple terms: names, addresses and other identifying information are split off from medical details at source and are separately transmitted and stored at all stages thereafter. Incorporation of the required data processing standards and protocols should not present difficulties for new information systems or existing systems undergoing major revision, but may be problematic for "legacy" systems.

Another weakness of the system as proposed is the residual hazard posed by the *potentially re-identifiable* information held by each *Disease Register*. One method of

reducing this hazard would be to "vertically partition" those *Disease Registers* which hold particularly sensitive data into a series of "*sub-Registers*", each of which would receive and accumulate only a small subset of the data items required by a *Disease Register* as a whole for each case of disease (or for each health event). From a security standpoint, it would be quite acceptable for *sub-Registers* belonging to different *Disease Registers* to be co-located and co-administered. Such sharing of resources by unrelated *sub-Registers* would substantially reduce the marginal cost of vertically partitioning *Disease Registers* in this way. Subsets of *sub-Register* data would generally need to be re-assembled into complete *Disease Register* data sets (which are *potentially re-identifiable*) inside *Secure Research Facilities* in order to carry out complex statistical analyses. However, selection of those research subsets, as well as routine reporting and other descriptive epidemiology, could be carried out without re-assembling the vertically partitioned sub-Register data, through the use of "fusion queries" as described by Yernini *et al.* [44]. This would appear to be a fruitful area for further research and development.

Perhaps the most significant weakness of the system is the high degree of trust which *Disease Registers* must place in the *Population Register* to do its job correctly. If the *Population Register* fails to determine that two identities represent the same person, or vice-versa, then *Disease Registers* may incorrectly interpret two notifications of the same case of disease as representing two cases, or two notifications as incorrectly representing the same case. This is a problem which already afflicts most disease registers to some degree – the critical question is how much worse would a centralised identity matching agency, which does not have access to any medical details, be at this task? Kelman *et al.* [32] have suggested that the inevitable reduction in the effectiveness of identity matching may be an unavoidable cost which researchers and *Disease Registers* have to pay in order to permit their work to continue in an environment in which privacy is given greater importance than in the past. At the very least, the magnitude of this currently unquantified trade-off in matching effectiveness needs to be measured through empirical studies. Even if the loss of matching effectiveness is small, it would be possible to allow *Disease Registers* to link their databases with that of the *Population Register* (via the *Identifier Translation Agency* mappings) from time to time in order to carry out audits of the effectiveness of the identity matching provided by the *Population Register*, and to suggest merges and splitting of identities based on the extra medical information held by the *Disease Registers*. Such audits would need to be undertaken at a *Secure Research Facility*, under strict and independent supervision, to ensure that the identity and disease information were brought together only transiently and that no copies of the linked information were retained by anyone.

Summary

The system proposed in this paper offers the prospect of a federation of disease registers and genetic and social databases which would simultaneously provide better protection of personal privacy and maintenance of confidentiality, while enabling cheaper and more efficient linking of data for research purposes.

To many readers, it might seem unlikely that the level of community consensus necessary for the commissioning of such a system could ever be achieved. This may be true, but we must be careful not to overlook the potential for building intrinsically more secure population-based health information systems provided by modern, networked computing environments. It is therefore important that alternatives to traditional methods for collecting, storing and using biomedical and social research data are proposed and consequently exposed to scrutiny by information security experts, researchers and interested members of the general community, alike. This will take time. In the interim, it may be possible to implement the proposed system in the more circumscribed environments of consortia of health care institutions which are engaged in collaborative research.

Author Contributions

The author was responsible for all aspects of this paper.

Competing interests

None declared.

Acknowledgements

The author thanks the reviewers for their prompt, detailed and helpful comments, which motivated substantial improvements to the paper.

References

1. Stroup NE, Zack MM and Wharton M **Sources of Routinely Collected Data for Surveillance.** *Principles and Practice of Public Health Surveillance (Edited by: SE Teutsch, RE Churchill) Oxford, Oxford University Press* 1994, 51-56
2. Anderson RJ **Information technology in medical practice: safety and privacy lessons from the United Kingdom.** *Med J Aust* 1999, **170:**181-184
3. Kelly G **Patient data, confidentiality, and electronics.** *BMJ* 1998, **316:**718-719
4. Vandenbroucke JP **Maintaining privacy and the health of the public.** *BMJ* 1998, **316:**1331-1332
5. Al-Shahi R and Warlow C **Using patient-identifiable data for observational research and audit.** *BMJ* 2000, **321:**1031-1032
6. Helliwell T, Hinde S and Warren V **Cancer registries fear collapse [letter].** *BMJ* 2001, **322:**730
7. Kmietowicz Z **Registries will have to apply for right to collect patients' data without consent.** *BMJ* 2001, **322:**1199
8. Verity C and Nicoll A **Consent, confidentiality, and the threat to public health surveillance.** *BMJ* 2002, **324:**1210-1213
9. Snaedal J **The ethics of health sector databases.** *eHealth International* 2002, **1:**6
10. Schneier B **Applied Cryptography** *New York, John Wiley and Sons Inc.* 1996, 31-32
11. Etheridge Y **PKI (public key infrastructure) – how and why it works.** *Health Manag Technol* 2001, **22:**20-21

12. Quantin C, Allaert F-A and Dussere L **Anonymous statistical methods versus cryptographic methods in epidemiology.** *Int J Med Inf* 2000, **60:**177-183
13. Sweeney L **Computational Disclosure Control for Medical Microdata: The Datafly System.** *In: Record Linkage techniques – 1997: Proceedings of an International Workshop and Exposition, Arlington, VA, March 20–21, 1997 (Edited by: Alvey W, Jamerson B) Washington DC, Federal Committee on Statistical Methodology, Office of Management and Budget* 1997, 442-453
14. Mayer TS **Privacy and Confidentiality Research and the U.S. Census Bureau: Recommendations Based on a Review of the Literature.** *Washington DC, U.S. Census Bureau, Statistical Research Division* 2002,
15. World Wide Web Consortium **Platform for Privacy Preferences (P3P) Project**
16. Caulfield T, Upshur REG and Daar A **DNA databanks and consent: A suggested policy option involving an authorization model.** *BMC Medical Ethics* 2003, **4:**1
17. Churches T **A method for improving the protection of personal privacy and confidentiality in disease registers [abstract].** *Australasian Epidemiologist* 2001, **8:**31
18. Boruch R and Cecil J **Assuring the Confidentiality of Social Research Data.** *Philadelphia, University of Philadelphia Press* 1979,
19. United States General Accounting Office **Record Linkage and Privacy: Issues in Creating New federal Research and Statistical Information (GAO-01-126SP)** *Washington DC* 2001, 79-82
20. Pommerening K, Miller M, Schidtmann I and Michaelis J **Pseudonyms for cancer registries.** *Methods Inf Med* 1996, **35:**112-121
21. Blobel B **Clinical Record Systems in Oncology: Experiences and Developments on Cancer Registers in Eastern Germany.** *In: Personal Medical Information. Security, Engineering, and Ethics (Edited by: Anderson R) Berlin, Springer* 1997, 39-56
22. United States Patent Office **US Patent 6,148,342: Secure database management system for confidential records using separately encrypted identifier and access request.** *Washington DC* 2000,
23. Ho AP **A Secret Splitting Method for the Protection of Confidentiality in Computer Records.** *In: Research Advances in Database and Information Systems Security (Edited by: Atluri V, Hale J) Boston, Kluwer Academic Publishers* 2000,
24. Quantin C, Kerkri E, Allaert F-A, Bouzelat H and Dusserre L **Security aspects of medical file regrouping for the epidemiological follow-up.** *In: Medinfo 98 (Edited by: McCray AT, Cesnik B, Scherrer J-R) Amsterdam, IOS Press* 1999, 1135-1137
25. Borst F, Allaert F-A and Quantin C **The Swiss Solution for Anonymously Chaining Patient Files.** *In: MedInfo 2001 (Edited by: Rogers R, Haux R, Patel V) Amsterdam, IOS Press* 2001, 1239-1241
26. United States Patent Office **US Patent 6,449,621: Privacy data escrow system and method.** *Washington DC* 2002,
27. Szolovits P and Kohane I **Against simple universal health-care identifiers.** *J Am Med Inform Assoc* 1994, **1:**316-319
28. Anderson RJ **Remarks on the Caldicott Report.** *Cambridge, Cambridge University* 1998,
29. Kohane I, Dong H and Szolovits P **Health Information Identification and de-Indentification Toolkit.** *In: Proceedings, Annual Fall Symposium of the American Medical Informatics Association (Edited by: Chute C) Florida, Hanley and Belfus Inc* 1998, 356-360
30. deCODE genetics **What is the Icelandic Health Database?** *Reykjavik* 2002,
31. Liberty Alliance Project **Liberty Architecture Overview Version 1.0.** *Piscataway NJ* 2002,
32. Kelman CW, Bass AJ and Holman CD **Research use of linked health data-a best practice protocol.** *Aust N Z J Public Health* 2002, **26:**251-255
33. Anderson RJ **The DeCODE Proposal for an Icelandic Health DataBase.** *Cambridge, Cambridge University* 1998, 7
34. Black N and Payne M **Infopoints: Improving the use of clinical databases.** *BMJ* 2002, **324:**1194
35. Gulcher J and Stefansson K **The Icelandic Healthcare Database and Informed Consent.** *N Engl J Med* 2000, **342:**1827-1830
36. Rose H **The Commodification of Bioinformation: The Icelandic Health Sector Database.** *London, The Wellcome Trust* 2001,
37. Kaye J and Martin P **Safeguards for research using large scale DNA collections.** *BMJ* 2000, **321:**1146-1149
38. Health Level Seven Inc. **An Application Protocol for Electronic Data Exchange In Healthcare Environments, Draft Version 2.3.** *Ann Arbor* 1996,
39. Object Management Group **CORBAmed: OMG's Healthcare domain task force.** *Needham, MA* 1999,
40. Object Management Group **Person Identification Service specification (document formal/99-03-05).** *Needham, MA* 1999,
41. Carnall D **Medical software's free future.** *BMJ* 2000, **321:**976
42. Kerckhoffs A **La cryptographie militaire.** *Journal des sciences militaires* 1883, **IX(5–83):**161-191
43. Christen P and Churches T **Joint Computer Science Technical Report TR-CS-02-05: Febrl – Freely extensible biomedical record linkage.** *Canberra: Australian National University* 2002,
44. Yernini R, Papakonstantinou Y, Abiteboul S and Garcia-Molina H **Fusion Queries over Internet Databases.** *In: Proceedings of the 6th International Conference on Extending Database Technology (EDBT), Valencia, Spain (Edited by: Schek H-J, Saltor F, Ramos I, Alonso G) Heidelberg, Springer-Verlag* 1998, 57-71

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2288/3/1/prepub

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral