Section 3: Privacy and Security Technology for Secondary Use of Health Information 3
Introduction 3
Standard security controls 4
Standard privacy principles and controls 5
De-identification 6
Precision medicine requires pseudonymized health information 6
De-identification is not binary and maximally de-identified data are useless 6
Types and levels of de-identification 7
Use of health information 9
Electronic enforcement of purposes for use 9
Searching and use 10
Consent 11
Types of informational consent 11
An informational consent requirement for the HIN 12
Consent directives and policies 13
Identity management 14
Patient identification and matching, a.k.a. patient resolution 14
Global and federated identity 15
Identity protection 15
Patient provider relationships and the circle of care 16
Technical governance and architecture 16
Technical governance 16
Risk segregation in architecture 17
Specific cloud security risks and controls 18
Emerging technology risks in distributed processing 20

3: Privacy and Security Technology

Privacy preserving data mining (and machine learning) 20

Privacy-preserving query and analysis algorithms 21

Privacy techniques applied to analysis outputs 21

Differential privacy 22

Homomorphic encryption 23

Tokenization 24

Format-preserving encryption for tokens 25

Blockchain 25

Audit logging 33

UI caching, scrolling and paging 33

Audit logging at all points 33

Remote logging 34

Conclusions 34

References 36

Section 3: Privacy and Security Technology for Secondary Use of Health Information

André Carrington and Meng Zhu[1]

We examine privacy and security architecture and technology in a distributed health information network designed to benefit Canadian society through medical innovation, technical innovation and precision medicine. We discuss assumptions, the requirements and concerns which require extra or unique attention in this scenario and the controls needed to address them.

Introduction

We examine the privacy and security architecture and technology in a distributed health information network designed to benefit Canadian society through medical and technical research and innovation as well as precision medicine.

To achieve these aims, the presence or establishment of a distributed health information network (HIN) with proper governance and access to a wide array of patient and administrative data, including real world data and genomic data are assumed. It is also assumed that the HIN would contain a limited amount of centralized data with most patient data kept at the source or custodian and accessed in a distributed manner, since the latter offers privacy advantages.

We provide a brief summary of sources for standard privacy and security controls, however, for greater impact, we focus on discussion of architecture and technology requirements and controls for the HIN which require extra or unique attention as non-standard or quasi-standard.

As a matter of style, we discuss controls as measures which "should" be employed. The governing body of the HIN would need to decide if in fact such controls are mandatory (i.e., "shall" or "must"), optional with or without explanation (i.e., "should" or "may") or not desired based on assessment of the costs, benefits and risks.

[1] Some text written by Meng Zhu on privacy preserving data mining is included.

Standard security controls

At the highest level security is often described in terms of three objectives: confidentiality, integrity and availability (CIA). Another triad which is common, is: prevent, detect and respond—which NIST expanded to five objectives.

Since our focus is on security architecture and technology (not governance outside of technology), the most dominant and authoritative information security reference is the *ISO 27002 Information technology - Security techniques - Code of practice for information security controls.*

ISO 27002 contains 14 topics: information security policy, the organization of information security, human resource security, asset management, access control, cryptography, physical and environmental security, operations security, communications security, system acquisition development and maintenance, supplier relationships, information security incident management, information security for business continuity, and compliance. Within these topics there are 35 control objectives and 114 controls.

The security controls should be built in, not bolted on—as an old security principle.

Additional general references for HIN security controls are:

- Canada Health Infoway Blueprint version 2 Privacy and Security Conceptual Architecture
- NIST SP 800-53 Security and Privacy Controls for Federal Information Systems and Organizations
- NIST SP 500-299 Cloud Computing Security Reference Architecture
- NIST Framework for Improving Critical Infrastructure Cybersecurity
- Cloud Security Alliance Cloud Controls Matrix
- Cloud Security Alliance Reference Architecture
- ASIS Facilities Physical Security Manual

References on security principles that are considered **obsolete** and **defunct**:

- Generally Accepted System Security Principles
- Generally Accepted Information Security Principles
- NIST SP 800-14 Generally Accepted Principles and Practices for Securing Information Technology Systems (withdrawn)

Standard privacy principles and controls

Privacy is concerned with protecting the dignity of people in the collection, use and disclosure of personal information including personal health information (PHI).

In Canada, the most dominant and authoritative privacy principles are the ten principles found in *CAN/CSA-Q 830-96 (now 830-03) Model Code for the Protection of Personal Information*, which was included in the *Protection of Personal Information and Electronic Documents Act (PIPEDA)*.

The ten privacy principles are: accountability, identifying purposes, consent, limiting collection, limiting use disclosure and retention, accuracy, safeguards, openness, individual access and challenging compliance.

Since our focus is on privacy architecture and technology, we are interested in how these principles influence design—and we recommend adherence to the privacy by design principles as well: proactive not reactive, privacy as the default setting, privacy embedded into design, positive-sum not zero-sum (full-functionality), end-to-end security and full life-cycle protection, visibility and transparency of/in standards, respect for user privacy.

Additional general references for privacy principles (not controls) are:

- Privacy laws and regulations
- CMA Principles for the Protection of Patients' Personal Health Information
- ISO/IEC 29100 Privacy Framework
- OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data
- CMA Medical Record: Confidentiality, Access and Disclosure (Update 2000)

More specific than privacy principles are privacy controls, but there is no standard "set" of privacy controls for architecture and technology with standard definitions. Individually, the following controls are common, where applicable: de-identification, opt-in preferences and redaction/masking.

Other controls applicable to the HIN are: consent management, lock-boxes, tokenization (including format preserving encryption) and identity protection.

We discuss these quasi-standard or non-standard privacy controls in later sections.

Finally, additional general references for HIN privacy controls are:

- COACH Guidelines for the Protection of Health Information
- Canada Health Infoway Blueprint version 2 Privacy and Security Conceptual Architecture
- NIST SP 800-53 Security and Privacy Controls for Federal Information Systems and Organizations
- The ISO 29101 Privacy Architecture
- Guidance from privacy commissioners and medical device regulators
- Orders from privacy commissioners

De-identification

Precision medicine requires pseudonymized health information

In oncology, a doctor requires the ability to search for (pseudonymized) individual cases similar to the (pseudonymized) patient at hand in terms of genetic markers and conditions. We refer to treatment of a patient based on knowledge of individual not average responses, as precision or personalized medicine. Precision medicine is critical for oncology but is rarely used in most practice areas of medicine. Similar case search and retrieval is a beneficial use of another patient's pseudonymized, yet individual, health information.

De-identification is not binary and maximally de-identified data are useless

In this section, we explain that de-identification is **not** a binary matter as corroborated by other authors (Emam, Gratton, Polonetsky, & Arbuckle, 2013), i.e. de-identified or not. In general, data are not fully identifiable, nor are data fully de-identified.

If we consider a patient's record in a hospital, with all of the fields intact, including direct identifiers such as the health care number and address, the data are not always fully identifiable because of patient matching and identification problems causing an 8% error rate in the US (Fernandes, Myers, & Viola, 2006; Pew Research, 2018; Sánchez, 2011). This is a known problem in Canada too from the author's experience at Canada Health Infoway. In (Sweeney, Abu, & Winn, 2013) nicknames (e.g., Jim instead of James) accounted for 13% error in patient matching and identification.

Records for the same person in different provinces, at different times, for example, cannot be easily matched or identified to one individual, because health card numbers are specific to each province.

In most simple scenarios, a record is identifiable/matched with the health care number, however there are multiple scenarios that present problems such as, when the health care number is not present in an encounter or record, records in different provinces for people with somewhat common names, multiple people (fraudulently) sharing a health card, etc.

On the other end of the spectrum, complete anonymity or de-identification, is achieved when there is no risk of identification or reidentification—however, that risk cannot be 'perfectly resolved' (Sweeney, 2002) since we do not know what a recipient of data has or will have, to link with data for re-identification.

Nearly complete anonymity is achieved when all the records in a dataset are the same, e.g., the dataset becomes one record—an average of the entire dataset, but an average has minimal use. Useful information contains some degree of uniqueness toward reidentification.

Types and levels of de-identification

The simplest concept in de-identification is the removal of direct identifiers, such as name, health card number, social insurance number, home address and telephone number. This data at the individual level may be used for education in primary care.

Health research often requires data of individuals (Willison, 2009), and so does precision medicine. For periodic updates (e.g., longitudinal tracking) and auditability, a pseudonym is assigned, i.e., a meaningless but unique number (MBUN) to each record and the map between the pseudonym and an original identifier is kept separately and securely in a cross-walk file. Auditability is required in decision support for precision medicine.

The next level of de-identification concept pertains to the removal of quasi-identifiers which are fields or attributes that easily identify a person when combined with other quasi-identiers—such as the triad (or 3-tuple) of zip code, birth data and gender which can be used to re-identify patients by linking. Linking with that triad has been demonstrated with US medical data[2] (Sweeney, 2002), US

[2] 87% of the state of Massachusetts' employee medical records were re-identified.

public genomic data sets[3] (Sweeney et al., 2013) and US hospital-collected genomic data[4] (Malin & Sweeney, 2004). Linking the same triad in Canada to public data sets is also achievable (Emam, 2006).

A better alternative to the previous approach, is to suppress some amount of detail or information in quasi-identifiers, e.g., suppressing the first 3 digits of a postal code, in a dynamic manner according to the rule of k-anonymity. In Ontario, data are considered de-identified according to the Ministry of Health and Long Term Care if any record cannot be associated with less than 5 persons. We say in this case that the cell size is 5, i.e., k=5 in the context of k-anonymity.

While k-anonymity addresses a record being linked to another record, i.e., record linkage, (Fung, Wang, Chen, & Yu, 2010) examine alternative methods and other risks not addressed by k-anonymity. Other risks include a record being linked to an attribute value (attribute linkage), or to another table indicating membership (table linkage). Techniques include record suppression, value suppression (full or partial) and cell suppression. Suppression reduces the risk of inference, in cases such as:

- Rare demographics, age, e.g., 101 (this would typically be fully suppressed)
- Rare conditions, e.g., scissors in lungs
- Information that presents a high risk of unintended linkage, e.g., exact date (or exact date and time) of visit

Alternative algorithms include, for example, t-Closeness (Li, Li, & Venkatasubramanian, 2007) and optimal k-anonymity (Bayardo & Agrawal, 2005).

For continuous variables, adding noise to data or binning it, adds privacy and can allow k-Anonymity to work more effectively—however, the noise and bin size must not be too large or the data becomes invalid from a clinical and statistical point of view.

---

[3] 84% and 97% of contributors to the Personal Genome Project were re-identified by exact name and nicknames, respectively.

[4] For 2 diseases 33-37% of genomes were re-identified, for 3 diseases about 50% were re-identified, and for 3 more diseases, 69%, 75% and 100% of genomes were re-identified.

Changing binary and categorical values as a form of noise/perturbation along with binning as employed by some methods (Ghinita, Tao, & Kalnis, 2008) in health care can be problematic. Clinically invalid scenarios may arise (Dankar & El Emam, 2013) such as inaccurate drug-drug interactions, medications which do not correlated with a patient's condition or demographics, etc. There are scenarios however, where two categories (one or both with a low cell count) may be combined, *based on clinical analysis*, if they are etiologically related. Semantic frameworks may be employed to automatically assist with de-identification of binary and categorical values (Martínez, Sánchez, & Valls, 2013).

In de-identification for health care, one should ensure the clinical validity of data first (as a constraint in most cases)—then simultaneously address data privacy risk, data utility/quality, data/statistical bias and cost (of turn-around time, expert time, tools and training).

The HIN should provide the ability to de-identify data in multiple ways, including at a minimum, the following methods: removal of direct identifiers, pseudonymity, k-anonymity or a similar risk-based suppression method that meets the requirements of Ontario as a general representation of jurisdictional requirements in Canada.

Use of health information

Electronic enforcement of purposes for use

Permitted and consented uses of health information should be enforced electronically and automatically – i.e., not just with governance controls.

The purpose for use, should be encoded using a standard codeset, based on the ISO standard for classification of purposes for use of personal health information (ISO TS 14265), for example. The term, beneficial use, should be added to the codeset.

Enforcing permitted and consented use, requires explicit meta data from the user application and/or user directly, about the purpose of their query or action. Sometimes this may be assumed based on the user's role – i.e., if a user is a clinician-researcher (role) accessing data from a teaching

hospital, it is necessary to ask if the query or action is for individual care or research. Whereas some users, e.g., a drug manufacturer, would only have the role we assumed: beneficial use.

The purpose attribute associated with the user's query/action along with other attributes (such as the minimum data quality) should be compared with the permitted use attributes of each data set (or row of data, if/where applicable) with results only returned for matching purposes and enforced by the HIN component accepting and performing the user's query/action. Access control concepts may be applied in terms of four types of components for administration, information (provides the attributes), decision-making and enforcement.

Technical (and governance) controls should be applied to prevent exporting data from the HIN with workflow and/or audit alerts for any queries which are overly broad, or which exceed quotas for use. To ensure analytics occur within the HIN, a wide variety of analytical capabilities will need to be facilitated.

Searching and use

In theory, merely reading a record, to match it against search criteria, is defined as a use under the law. However, when a patient requests who has accessed their record, what uses should be reported?

Should the report include a doctor browsing through a list of names, health card numbers and addresses to find a match? There is the possibility of a doctor retaining knowledge in this case and/or the doctor using search/browsing as "cover" for their real objective of finding specific information that is not the target of the search. It seems prudent to report this search as a use, even though further access to the patient's record is a sure indicator of use.

Should the report include a computer reading names, health care numbers and addresses on behalf of a person doing a query, to find a match? There is no retention of knowledge by a person in this case, although the data may reside in a temporary workspace which could (with low probability) be compromised. It does not seem necessary to report this as a use, unless a relevant compromise is known or suspected.

Consent

Types of informational consent

In a health care setting, consent for the collection, use and disclosure of information, i.e., **informational consent**, is distinguished from consent for treatment. In a primary care setting where the patient is seeking treatment and will provide information, there are two main types of consent: **implied consent** and **express consent** (also called **explicit consent** or **opt-in**). Other consent concepts and types found in legislation are: **informed consent**, **deemed consent**, **no consent** required and **manifest consent**. That said, a patient may **refuse consent** (**opt-out**) or **withdraw consent**, or create a **consent directive** to proactively withhold consent—except in situations of **deemed consent** and **no consent**.

When a patient presents themselves at a primary care facility seeking treatment their informational consent is usually **implied**—meaning we infer they have consented even though their consent has not been expressly given in oral or written form. There are two exceptions to note. Firstly, any express wishes (consent directives) indicating that they do not consent to specific uses and disclosures must be respected over and above implied consent. Secondly, an electronic form with a box checked "yes" by default, while in written form, is **not express consent**, it is **implied consent** with the ability to *opt-out*.

Express consent is explicitly given in oral or written form. **Express consent** for information, often arises in the context of private practice and in consent for research. Both **express consent** and **implied consent** must be **informed**, i.e., **informed consent** means the patient has a reasonable understanding of what is being collected, who will be using it and why, and to whom information may be disclosed.

For research, **express consent** is needed by default, unless a research ethics board upon reviewing the research proposal and protocol determines that obtaining consent is unreasonable, impractical or infeasible[5] and decides to waive the need to obtain **express consent** and approves the research.

Manifest consent in Quebec, for quite some time caused confusion and was interpreted and practiced as **express consent** by some or **implied consent** by others. It has since been clarified that the consent must be unambiguous in either form.

No consent is required for the use and disclosure of information by the collecting health information custodian in situations of lawful access, subpoena, public health surveillance and protecting individuals from imminent danger,.

Deemed consent and **no consent** for information and treatment, arise in the situation of involuntary detention of mental health patients. **Deemed consent** exists in B.C. and relates to the involuntary detention of patients for mental health purposes, where the patient is deemed to have consented with no right of refusal, withdrawal nor a substitute decision-maker. Other provinces address the same scenario with **no consent** required for information and treatment.

Finally, a word of caution regarding the concept of **retrospective consent** or **deferred consent.** In the health care literature, it pertains to **treatment** in emergency or intensive care for approved studies and protocols where the decision-maker is incapacitated or under duress (Honarmand et al., 2018; Songstad, Roberts, Manley, Owen, & Davis, 2018). Ethical concerns have been raised with its use (Kim, 2012; Ryan, 1971). For research where incapacity or duress are not at play, but a study is unapproved, retrospective consent for information is not informed by the multiple perspectives of members of the research ethics board. Therefore, it may not be **informed consent**.

An informational consent requirement for the HIN

The HIN presents a new consent requirement: for research and innovation to flourish as an objective of the HIN, the HIN and/or its governing body should be able to collect **express consent** for information **proactively** for multiple retrospective and prospective research studies *and innovation*.

[5] Each province uses one or more of these terms alone or in combination.

This may be done in the form of consent directives (discussed in the next section)—which are typically applied to limit use and disclosure, but may also be used to permit use and disclosure proactively.

The patient or participant will be initially informed of beneficial use and its subtypes, the types of organizations, and the types of information they may use. The consent directive may of course permit research while limiting which information may be used.

Notably, while research is defined as a use in the legislation, beneficial use for innovation as a superset of research would need to be clearly and sufficiently defined, justified and governed by an innovation ethics board and permitted in legislation. Since this document has a technical focus, we do not provide such justification here.

Consent directives and policies

Consent directives, also called **disclosure directives** in some jurisdictions, refers to a patient's express/explicit directions about their consent or non-consent, for the collection, use and disclosure of their personal health information. **Consent directives** are specific to the health care, while the related but less complex concepts of **opt-out** and **opt-in** are used in other industries.

The HIN will in many cases obtain hospital data from a data warehouse instead of the originating system—the hospital/clinical information system (H/CIS)—which puts data into the data warehouse. Hence the HIN will need to apply consent directives itself or assist the organization to apply them at the data warehouse, if the data are to be accessed and used. Also, if the HIN holds any data itself—as it will in the form of results—then it also needs to apply a patient's consent directives, either captured locally by the HIN or captured by the source system.

Consent regarding information may be withheld by the patient (at the outset) or withdrawn afterward except in situations of deemed consent or no consent)—however, withholding informational consent can result in non-treatment as the information enables treatment to be sufficiently informed for safety and it allows billing, follow-up, etc.

The HIN should be capable of receiving and enforcing the consent directives of patients and the consent policies of jurisdictions and organizations. Implied consent and deemed consent, as consent policies, should exist explicitly as policies (or meta data) for clear audit trails and proper enforcement. The HIN should be able to map the policies and consent directives to the HIN's own encoding which must be sufficiently rich and flexible so as to comply with all applicable jurisdictions.

To collect and manage consent, alternatives include patient portals and mobile applications.

Since some jurisdictions such as Ontario allow for consent directives to control access to records in whole or in part. The HIN should have flexible consent directive processing which can handle consent directives that either allow and/or disallow actions if the use/access to data relates to:

- Certain health care providers (e.g., a neighbour or family member)
- Certain domains: drugs, lab information, mental health, etc.
- Certain encounters or episodes (a set of encounters)
- Certain periods of time
- Certain organizations (e.g., an addiction rehabilitation facility)

Challenges with consent directives and policies are that all of them must be:

- Fully tracked and archived with version control for medico-legal reasons[6].
- Not contradictory
- Meaningfully mapped by the system which must enforce it

One can manage explicit consent at the program level with implied informed consent for use of de-identified information via notifications of major new studies, audit trail of queries and time expiration on program consent.

Identity management

Patient identification and matching, a.k.a. patient resolution

Patient matching typically has an 8% error rate [cite VUHID]. Since errors will result in data being incorrectly linked for query or analysis the confidence of matches should be tracked.

---

[6] Tracking of consent directives and policies with start and end dates of applicability within each system, is necessary so that the medical record as a doctor sees it at any given point in time can be known or reconstructed, where consent directives and policies act as filters to what the doctor saw.

If the HIN is used for individual care, as intended with precision/personalized care, business rules should enforce a low tolerance for error (3rd standard deviation in matching) for that use—and the user, if they desire, should be able to choose an even lower tolerance at their discretion. The consequences of error can have grave implications for patient safety.

Similarly, if the use pertains to a regulatory submission, then a business rule should enforce low tolerance as required to meet regulatory requirements for data quality.

For other queries or analyses, the tolerance for matching errors can be higher, although the user/analyst should be able to choose a lower tolerance if they desire.

Global and federated identity

Patients, i.e., subjects of care, can be identified by a voluntary universal health identifier (VUHID) that is either private or open. This concept is standardized in ASTM VUHID standards and a vendor service is available (although potentially US-centric).

For federated identity in the health information network, Canada Health Infoway recommends a model wherein identities of member organizations and centralized identities from the health information network would both be recognized. Example implementations include the US Veterans Affairs Federated Identity Model.

Identity protection

To protect identities in the HIN (for record-level information in personalized/precision medicine), Canada Health Infoway recommends that the health information network not store any organizational identities with personal health information. Instead they recommend only storing data with an enterprise identifier, separate from the organizational identifier—and keeping the mapping file separate from the personal health information to mitigate linking threats in the case of a breach. This control could be applied for any centralized data, or data in staging areas which are accessed in a distributed model.

Patient provider relationships and the circle of care

Patient provider relationships and the circle of care concept only apply to primary use of personal health information—as expected for precision or personalized medicine. Otherwise it does not apply.

A patient's circle of care are the set of health care providers who have a need-to-know the patient's personal health information because they are currently treating the patient. They may be treating the patient regularly as part of their health care team (including alternative health), or once/sporadically in each encounter the patient has with the health care system (e.g., filling a prescription), or during an episode of care (multiple encounters until a condition is resolved or stabilized), or for the duration of a treatment program or study.

Tracking patient provider relationships can help with auditing the circle of care and meaningfully reporting access to a patient's record if/when they request such access records.

Technical governance and architecture

Technical governance

Technology governance should tie into, and align with, business strategies and governance, which may include, but is not limited to, enterprise risk management and business continuity planning.

Technology should be governed in each of the following life cycles:

- The risk management life cycle
- The solution design life cycle (SDLC)
- Operations and incident management

In the risk management life cycle key technical governance activities include:

- Threat risk assessment (TRA), which identifies and evaluates the security threats, controls and risks for a program, project or infrastructure from a people, process and technology point of view. It may include findings from IT audits and security reviews.
- Privacy impact assessment (PIA), which identifies and evaluates the privacy threats and risks for a program, project or infrastructure from a legal, people, process and technology point of view. It may include findings from audits and security reviews.
- IT audit (or IM/IT audit) which identifies the compliance of a program, project or infrastructure's controls with the organization's IM/IT policies, requirements and standards,

which may include, depending on scope: IM/IT security, privacy, acceptable use, access to information and disaster recovery.
- Privacy and security awareness training which ensures that employees are not exploited by people or technology, are aware of threats and risks, know who to contact, know what they are accountable for, and know what standard procedures to follow.
- Threat intelligence informs TRAs and PIAs by identifying threats relevant for assessment, their likelihood and possible impact.

In the solution design life cycle (SDLC) key technical governance activities include:

- Architecture frameworks (AF): The design of HIN solutions should use an architecture framework like TOGAF, that includes conceptual, logical and detailed design steps so that solutions are not irretrievably locked into vendor solutions which are no longer sufficient or advantageous, or vendors who become insolvent or represent a conflict of interest.
- Security/software quality assurance (SQA), vulnerability assessment and penetration testing.

In operations and incident management key technical governance activities include:

- Security incident and event management (SIEM) to track security events and identify which events qualify as incidents worth investigating.
- Investigation and response
- Disaster recovery

Risk segregation in architecture

It is standard to segregate data and systems based on widely different risk profiles to avoid the possibility of accidental cross-contamination, accidental misuse or unauthorized access and intentional misuse or unauthorized access. Segregation by how identifiable the data are and/or the level of data quality may occur.

If personal health information, i.e., identifiable information, or less pseudonymous information, is used, then the law may apply, risk is higher; requirements are stricter; segregation between identifiable and de-identified data may be required; and services for data may or may not differ.

Specific cloud security risks and controls

Although cloud services have been around for over a decade—e.g., Amazon Web Services offered storage in 2002 and infrastructure as a service in 2006—there are two concerns that arise in cloud computing:

1. Mature cloud technologies still pose challenges for complex privacy and security requirements in four areas: audit logging and configuration management, version control, segregation of services and the geographic location of services.
2. Emerging cloud technologies present general cloud security and privacy concerns.

For mature cloud technologies, the first challenge pertains to audit logging with elastic services, which is challenging because instances (IaaS) and platforms (PaaS) are usually elastically created and destroyed and therefore their associated logs need to be remotely captured and unambiguously named, with the architectural state, transaction states and configuration states, recoverable for any given point. Extra configuration management services for a fee, are required at a minimum.

Another challenge is version control with cloud services. Some cloud service providers do not transparently provide a version history nor even a version number for their own proprietary services, making auditing and regulatory reporting (especially in the context of medical devices) difficult if not precarious. Also, some services are based on a dynamic build of many underlying open source components and in the context of elasticity, two components started at different times may have different configurations. Careful tracking with configuration management may be important.

An example of a third challenge for mature cloud technologies, is that one of the top cloud service providers only provided services and architectural patterns in 2016, to meet the requirement for physical segregation of services from other customers in health care, with dedicated virtual private clouds.

A fourth challenge pertains to Canadian requirements to restrict the storage and transmission of data to within a province (per requirements in B.C. and N.L.) and within Canada (per recommendations from Canada Health Infoway's pre-certification criteria and general privacy

concerns stemming from the U.S.A.'s Patriot Act). Some cloud service providers only provide some services in specific American locations, which violates these requirements.

Two emerging cloud technologies also present general cloud security and privacy concerns: the serverless (or container) model and the function as a service (FaaS).

Compared to mature cloud services, the serverless model and FaaS are not operationally mature nor architecturally/theoretically mature with respect to security, privacy and governance patterns & requirements. The cloud service provider and their clients learn and evolve through events arising from real-world events, tests and attacks.

As an example of growing pains, with elastic IaaS, a privacy and security concern was discovered and resolved in the last couple of years. Virtual IP addresses are assigned from a pool and when one customer releases an IP address, another customer could obtain it, while old inbound transactions from the previous customer persisted. This was not a one-off phenomenon, as a study found this occurrence with a percentage of IP addresses obtained from the pool. Note, this was not in the context of elastic services which should be fronted by a fixed listener, but manually obtained and released IP addresses.

Another concern pertains to functions as a service (FaaS). The concern with FaaS is that it attempts to logically abstract and hide the underlying platforms and infrastructure it uses (uses at that time, not subsequently)—however this is at odds with the need for a clear and traceable audit log, since FaaS still uses PaaS and IaaS underneath. Audit logging should not only document the normal flow of transactions and interactions, but also side channel activity to ensure the integrity of the system. Visibility to side channel activity and attacks may be limited in cloud services. Practically speaking cloud service providers may in some cases do a better job of preventing, detecting, responding to, and recovering from, side channel attacks as compared with the organization using the cloud, however some accountability and visibility have been lost in the process and regulatory bodies may not fully accept this transfer of risk.

Finally, advanced or complex privacy and security requirements may require service which are only available in some regions and/or may not be available with certain other specific services or elastic

capabilities—possibly causing unanticipated or late trade-offs between functionality, elasticity/availability and security.

These concerns are identified as specific potential risks which may be investigated in various phases of procurement and privacy impact assessment.

Emerging technology risks in distributed processing

Federated query, distributed query, cluster/parallel processing and distributed machine learning are emerging technologies with new security and deployment patterns which cause cloud service providers and their clients to learn about some of their shortcomings as the technologies undergo real-world use, testing and attack.

Privacy preserving data mining (and machine learning)

We differentiate privacy-preserving data ***publishing*** from privacy preserving ***data mining*** (and machine learning), where the user or analyst receives data versus the results of queries on data.

The privacy-preserving ***publishing*** approach allows the recipient to perform direct inspection and technically (not legally) any query/analysis despite authorization and governance. Direct inspection is typically important for an analyst to understand the data's distributions, indices (statistics) and types of errors (Dankar & El Emam, 2013) although the first can be provided without direct access. This approach allows the analyst to use their own tools.

De-identification, as discussed in a previous section, is the key method used in privacy-preserving publishing. We also discuss differential privacy in the next section as an alternative that we do not recommend.

The privacy-preserving ***data mining*** (e.g., machine learning) approach does not allow direct inspection of data by the analyst. Instead it only allows queries or analyses which:

- use privacy-preserving query and analysis algorithms, or
- apply privacy techniques to the analysis output, or
- use differential privacy to allow or disallow queries or analyses based on rules, quotas or inspection (automatically or manually).

Each of these approaches, discussed in the subsections below, limit the analyst to installed tools and allowed types of query and analyses.

Privacy-preserving query and analysis algorithms

Algorithms which preserve privacy include homomorphic encryption, set intersection, secure sum, secure set union, secure size of set intersection and the scalar product (Mendes & Vilela, 2017). All of these algorithms are intended to perform computations over multiple data sources without revealing data between the data sources.

Set intersection (Freedman, Nissim, & Pinkas, 2004) is one of the most useful functions, because before we can perform any analyses based on data from multiple sources, we must first identify the intersection of patients which exist across all of those sources—i.e., we must identify the cohort. However, set intersection requires homomorphic encryption (discussed two subsections below) which has drawbacks in terms of a delayed answer and large energy use and computational expense.

However, prior to finding the set intersection, we may first find the size of the set intersection (Clifton, Kantarcioglu, Vaidya, Lin, & Zhu, 2002) in a secure manner without the expense of homomorphic encryption. This would tell us whether or not we would obtain a cohort of sufficient size for analysis or research—e.g., with sufficient statistical power. If the size is insufficient then there is no need to perform the actual set intersection. This method uses encryption however—so its expense relative to homomorphic encryption would require investigation.

A scalar product is also a very useful operation since it can allows the analyst to check for, or find, similar patients in a precision medicine use case (for oncology or rare diseases). A scalar (or dot) product can be used in a number of similarity functions including kernels such as linear, polynomial, sigmoid and Mercer sigmoid kernels.

Secure sum may be useful for aggregate statistics while secure set union may also serve a purpose for data sets with common data elements but different patients.

Privacy techniques applied to analysis outputs

Techniques applied to analysis outputs include association rule hiding, downgrading classifier effectiveness and query auditing and inference control (Mendes & Vilela, 2017)—and all of these methods incur a trade-off between privacy and utility (ibid).

Association rule hiding can only be applied to data types where the concept of presence and absence applies to both, as in binary or nominal/categorical data types, including text where the keywords may be present or not. It does not apply to real, integer, ordinal or date information (unless these are binned as an uncommon approach).

Downgrading classifier effectiveness can reduce the ability to infer patients who exist within a data set. On a related note, providing only the classification and not the probability of error nor underlying classification score (a continuous value) can help to preserve privacy.

Query auditing and inference control limits the number of queries on the premise that the sum, difference or other set operations, between query results can be used to discover or infer individual records in the data. The downside to this approach is that once the quota of risk (or queries causing that sum of risk) has been expended there is no way for it to be reset, and the quota may not be sufficient for useful investigation. This approach does not foster innovation.

Differential privacy

Differential privacy refers to the difference between a data set and the data set with the addition of a single new record—if the addition does not substantively affect the output from analysis then the privacy of that single record is preserved. If this property applies to any single record within the data set, then privacy is preserved for all of those records. However, our clearly outliers would not meet this property. Also the concept of differential privacy is at often at odds with the utility of data—if each record contains such little information that it has minimal effect on the output, then one of two scenarios are in effect: either we have so many records that any one record (even in more outlying areas) is redundant, or the records we have do not have a lot of information for analysis. For many problems in health care we often do not have an abundance or excess of records, so the latter problem of reduced data utility is likely in order to achieve differential privacy.

Several authors (Dankar & El Emam, 2013; Muralidhar & Sarathy, 2010; Sarathy & Muralidhar, 2011) conclude that differential privacy cannot meet requirements for useful and privacy preserving data ***publishing or mining*** in health care or generally—except for very specific types of queries (e.g., sums) and limited circumstances, such as count data types, small data, non-sparse data, and data with low sensitivity. That said, we review it nevertheless as future developments may overcome limitations.

Differential privacy tries to address two concerns with privacy of queries or analyses performed on data (Dankar & El Emam, 2013):

1. identifiability of individuals from multiple queries/analyses (***mining***), and
2. re-identifiability of an individual in the single ***publication*** of a dataset.

In the first case (***mining***), if we perform a query to obtain a sum or average of elements in subset *A* of the data, and then perform a similar query on subset *A'*=*A*+individual, then we can discover a value for the individual from the difference of results. We generalize the concern of discovery to:

differences of more than one record, multiple queries and other types queries aside from sums and averages.

To prevent discovery of numeric data, differential privacy adds noise (*a priori*) to the original data, sets a global maximum on the number of queries allowed, and tracks the amount of information versus noise in each query. However, the noise can be so large that the output is not useful (Dankar & El Emam, 2013; Sarathy & Muralidhar, 2011) and the noise as privacy protection only applies to numeric variables (Dankar & El Emam, 2013).

The global maximum assumes the possibility of collusion among different users performing queries. For the HIN governance and technical controls may be used to treat the risks of collusion between organizations, but even so, limits on the number of queries and the utility of each query makes this approach impractical for the HIN.

Furthermore, even if some limitations are overcome a fundamental result for one specific type of query demonstrates that no privacy mechanism can overcome the issue of privacy loss if the noise is not sufficiently large compared to the amount (bits) of data one wishes to discover.

The alternative approach, de-identification (of different types and levels) is therefore more attractive.

In the second case (***publishing***), differential privacy seeks to perform de-identification based on theoretical concepts different from k-Anonymity and similar concepts. It seeks to add additive or multiplicative noise to the data to achieve its goal of no single record substantively affecting the output of analysis.

Homomorphic encryption

Homomorphic encryption fulfills the requirement to store encrypted data in an untrusted environment—the cloud perhaps—and then perform operations on that encrypted data (e.g., a query or sum), without decrypting it, to obtain an encrypted result. The encrypted result can then be brought back to a trusted environment and decrypted.

Since homo means the same, and morphe means shape—homomorphic encryption (HE) refers to being able to keep the same shape of data when operating in either encrypted or unencrypted form.

Some HE schemes allow only specific operations on data, whereas fully homomorphic encryption (FHE) schemes allow any operation.

It is a very powerful idea except that operations performed with FHE take about ***1 million times longer*** than they would with unencrypted data. It is expected that the performance of FHE may improve over time since it is a relatively new method for which efficiencies may be discovered.

In the model of the HIN where data are distributed—i.e., left at each source and processing is performed in a distributed manner at each site, and if the processing throughput and storage at each site is sufficient and available, then there is no need for homomorphic encryption of source data. If elasticity of processing or storage in the cloud are needed however, HE or FHE are possibilities, and they could also be used for storage of result data in the cloud.

For any specific use case, the feasibility of transaction throughput/performance of HE or FHE would have to be carefully examined for the intended operations and estimated data size and types.

Tokenization

Tokenization turns a sensitive data element into pseudonyms, i.e., meaningless but unique numbers (MBUNs) which are either mapped (with one-time pads), encrypted or one-way hashed (Stapleton & Poore, 2011), so that the original element is not recoverable without a cross-walk file or an encryption key. It is primarily used in the financial realm for credit card numbers as the literature reflects (Díaz-Santiago, Rodríguez-Henríquez, & Chakraborty, 2016; Stapleton & Poore, 2011), however it also be used in other domains.

We discuss format-preserving encryption in the subsequent section, ***with considerable caution***, as an ***option*** for tokens.

Usually tokens are not used for data elements in general, but rather more sensitive ones such as direct identifiers and perhaps quasi identifiers. However they can be applied for any masking use-case in general—e.g., to enforce a consent directive that masks drugs related to mental health or encounters related to sexual abuse, while maintaining allowing access to other parts of their record.

Tokens may be used as pseudonyms in search criteria with record-level data, as in personalized/precision medicine—e.g., a patient record would be identified by a meaningless but unique number such as 4822317, which is not a health care number nor medical record number. Otherwise, tokens do not play a role in analytics for detection, recommendation or prediction—machine learning cannot use them (unless they are filtering criteria as a pre-processing step).

Format-preserving encryption for tokens

With considerable caution, format-preserving encryption (FPE) (Dworkin, 2016) is an option which can be used with tokens. FPE allows a token to retain the same data type and length for schematic (not semantic) compatibility with an existing database schema or application with data format checks.

CAUTION: Although FPE was first proposed in 2002 (Black & Rogaway, 2002), it was only standardized in 2016 (Dworkin, 2016) with two modes: FF1 and FF3. FF3 was broken[7] (within a year of standardization) while FF2 was broken between the draft and final publication. This is an unusual occurrence for a NIST standard—however FPE for the middle/masked digits of credit cards and for social insurance/security numbers is a tall order since the space (domain) of possible numbers is relatively small from a cryptographic perspective. Hence, tokenization without FPE is preferred, although FF1 may be used if necessary and warranted by risk analysis.

Blockchain

Blockchain is a diary or ledger, that uses security mechanisms (cryptography) to ensure the integrity of ledger entries, and that uses computers (nodes) operated by many different parties to store and

[7] NIST's notification of FF3 being broken is found at https://csrc.nist.gov/News/2017/Recent-Cryptanalysis-of-FF3

check the authenticity of entries. Blockchain is widely known as the technology which enables Bitcoin.

As originally designed, blockchain is a publicly visible ledger, or **public blockchain** with the following key features:

1. it is a shared distributed data base, with integrity protections
2. it provides escrow and automated transfer of value (money, goods, access)
3. it is accessible to the public
4. it can run code per contracts—code which can be downloaded & inspected
5. distributed participating computers can join or leave dynamically
6. it has low throughput and should only be used for small pieces of data

McKinsey reports that blockchain is an "immature technology" in a "nascent" market without clear paths to success, but provides two main categories for use "beyond the hype": storage of static information, and acting as a registry of tradeable information. We later review six use cases they propose within those categories.

Clarity on use is important since, two decades ago, many of blockchain's underlying technologies[8] and requirements[9] also caused a hype cycle for public-key infrastructure (PKI). Hence we note that new technologies are adopted or used when they:

- address a clear and present pain point or need (e.g., trusting websites); or
- allow you to do something entirely new (e.g., mobile music streaming), or
- are dictated by a central authority for a necessary task (e.g., submitting data to a "payer"), ***and***
- are not stymied by competing interests, legal hurdles, lack of understanding, lack of standardization, or user/developer resistance to behaviour change

Private or permissioned blockchains have also been designed, which are only visible to authorized parties—however this limits or defeats the ability to address some of the decentralized needs which are uniquely met by blockchain.

---

[8] digital signature, identity & attribute certificates and e-cash [Warren; Brands].
[9] non-repudiation, decentralized web-of-trust [Warren; Brands]

Key features of a **private/permissioned blockchain** are largely the same as public blockchains, with notable changes underlined:

1. it is a shared distributed data base, with integrity protections, but does not protect off-chain content or management of identities, permissions & keys
2. it provides escrow and automated transfer of value (money, goods, access)
3. it is accessible only to authorized users
4. it can run code per contracts—code which can be downloaded & inspected
5. authorized distributed participating computers can join or leave, but joining is not automatic
6. it has low throughput and should only be used for small pieces of data
7. requires identity/permission management, mostly likely centralized

In more technical terms: blockchain is a protocol and architecture that uses cumulative hashing, nonces, proofs of work, digital signature via PKI and Byzantine fault tolerance in ways which reduce the efficiency of attempts at fraud. Public blockchains use compensation for work to incentivize a population of legitimate miners.

The distributed nature of blockchain is intended to resist fraud or forgery on the assumption that it would be difficult to cause collusion among the quorum[10] of nodes needed to agree upon the validity of entries in the blocks of the chain. Blockchain assumes that the vast majority (if not all) of the nodes and miners are legitimate (not fraudulent or malicious) and that monetary incentives given to miners helps ensure the legitimacy of miners.

From our analysis, blockchain meets various fundamental requirements and objectives (Table 1) as well as advanced requirements and objectives (Table 2). Notably, our analysis in Table 2 can be mapped to McKinsey's six categories: static registry, identity, smart contracts, dynamic registry, payments infrastructure, and other combinations of the aforementioned.

[10] The quorum may be a majority (more than half) or super-majority (two-thirds).

Transparency of transaction history and current state with the blockchain (but not necessarily transactions outside of it in the larger context) and transparency of entries/data in the blockchain. In the broader transaction context, read transactions may not be easily audited.

Pseudonymity is an additional requirement for crypto-currencies, on top of blockchain.

Table 1. Fundamental requirements and objectives met by blockchain, including security in terms of confidentiality, integrity and availability (CIA)

Requirements met	Objectives met		
	Trust	Security (C,I,A)	Func-tionality
1. Non-repudiation, integrity and access (transparency) of blockchain content, history and distributed application code	✓	I	
2. Decentralized/redundant storage, control and processing	✓	I, A	
3. Distributed application code and automatic triggers, escrow, transfer	✓	I	✓

Table 2. Advanced blockchain requirements and objectives met by blockchain, including security in terms of confidentiality, integrity and availability (CIA)

Advanced requirements met	Objectives met		
	Trust	Security	Func-

		(C,I,A)	tionality
1. Directories	✓	I, A	✓
2. Ledgers	✓	I, A	✓
3. Crypto-currency	✓	I, A	✓
4. Pay per use	✓	I, A	✓
5. Other smart contracts (e.g., signoff workflow, time-limited access tokens)	✓	I, A	✓

Simple but reasonable **privacy** for the **identities** in transactions with Bitcoin is achieved with pseudonyms (wallet identifiers)—however, there are cases where e-mail addresses have been used instead and persons involved with the transaction have been re-identified.

There are numerous faulty assumptions about blockchain (Table 3). For example, using encryption (and cryptography) in, or on top of blockchain, for **confidentiality** among multiple stakeholders (not just the data subject) subverts blockchain's benefits of decentralized control, enforcement and non-repudiation, contrary to faulty assumption #4 (Table 3). Managing keys is necessarily off the blockchain and usually requires a central authority, i.e., a certification authority, except in the rare web-of-trust model.

Similarly, using indices within blockchain pointing to content off the blockchain also subverts blockchain's decentralized benefits. You can ensure that the pointers are correct, but you cannot ensure that the off-chain content at the pointer is the same, is still there, nor if it is valid upon first submission.

If blockchain is used to store data, it cannot check whether or not the stored data is meaningful or valid. Blockchain cannot detect fraudulent audit trail content sent to it for storage (Table 3, faulty

assumption #6). It can only detect fraudulent attempts to change or misrepresent an entry in the blockchain, after it has been stored.

Table 3. Faulty assumptions people may have about blockchain

Faulty assumption	Objectives at risk		
	Trust, Privacy	Security (C,I,A)	Use, Function
1. Technology is the biggest challenge in multi-stakeholder problems and solutions.	✓		✓
2. Blockchain throughput is sufficient, e.g., minutes per write transaction.			✓
3. Blockchain can store any type/size of data			✓
4. Blockchain can protect the integrity of off-blockchain content, which in includes identities, keys and permissions in a private or permissioned blockchain	✓	I, A	
5. The blockchain protocol assumes legitimate nodes and miners (Orcutt, 2018)	✓	I, A	✓
6. The content stored in blockchain entries are the truth, the whole truth and nothing but the truth	✓	I	✓
7. Access to distributed application code for contracts equals transparency	✓	I	✓
8. The history of blockchain content will never be altered	✓	I, A	

3: Privacy and Security Technology

Another faulty assumption (Table 3, #7) is assuming that access to code provides transparency and trust. The DAO blockchain incident exposed that faulty assumption as well as #8.

There are also other privacy and security risks with blockchain (Table 4) which nascent work and proposals for blockchain in health care (Zyskind & Pentland, 2015) do not address. Some risks may be specific to Canada versus the US but that is largely not the case.

Table 4 identifies **some** risks which clearly illustrate several important stumbling blocks with blockchain and privacy (without being redundant of table 3). Table 3 and 4 combined are **not** a comprehensive **nor** systematic set of risks, and the risks are not evaluated. Additional risks would likely be identified in a project's privacy impact assessment (PIA) and threat risk assessment (TRA) which would have a more specific scope and context.

Table 4. Illustration of some privacy and security risks to compliment Table 3, for consideration in privacy impact assessment (PIA) and threat risk assessment (TRA)

Risk	PIA	TRA (C,I,A)
1. Denial of service (DoS) attacks target legitimate nodes to cause a quorum of malicious nodes (Orcutt, 2018)	✓	I, A
2. Decentralized storage and processing may or may not meet medical and/or legal requirements for: • within province geographic limit on personal health information for B.C. and N.L. residents (transit to/from nodes may violate the law despite encryption-at-rest) • custodianship, • a medical record, • hospital certification and accreditation of software and processes (if versions are not controlled, validated, frozen), • auditability of implementations, • medical device validation re decision support, or • provenance and use as evidence in regulatory submissions	✓	I
3. Advanced blockchain requirements and applications sometimes use/propose security and cryptographic protocols (Orcutt, 2018)—	✓	C, I, A

however Orcutt does not warn the reader that these protocols are not validated by key standards bodies and the cryptographic community at large.		
4. Once encrypted records and keys are put in a public or even a private/permissioned blockchain, they exist there potentially forever and can be copied for offline sustained attack. Exposure depends on read access rights. Threats come from quantum computing, weak keys, cryptographic implementation flaws and the accumulation of cryptotext for cryptanalysis.	✓	C, I
5. The industry is immature in its experience with the life cycle of decentralized processing and storage for blockchain applications outside of cryptocurrency—e.g., due to nodes being naturally decommissioned, error, or malicious attack. This risk is likely a black swan risk (low likelihood, high impact) regarding loss of data. It is difficult to decide how to treat black swan risks[11].	✓	A

In summary, blockchain has a number of potential applications for specific needs, however it suffers from a few key stumbling blocks in its application to personal health information. It should be avoided if other technologies would suffice, because it is an inefficient, energy wasting and complex compared to alternatives.

Any health care initiative considering blockchain should answer the following four questions at the beginning:

[11] The Long-Term Capital Management bailout due to mis-estimation of risk with asset backed securities as well and other near-catastrophic market failures are examples of black swan risks.

1. What is a competitive response time, which makes your application enjoyable and what is the maximum tolerable response time? Blockchain is known to take minutes to perform a write transaction (e.g., if it is used to acquire/distribute access tokens), whereas users prefer a subsecond response time and will tolerate perhaps a 10 second response time.
•
2. Have you done a privacy impact assessment (PIA) to ensure that the proposed system can comply with medical and privacy laws, regulations and directives, and meet your organization's risk tolerance for reputation management and enterprise/program viability? Is the PIA informed by a security TRA (next)?
•
3. Have you done a security threat risk assessment (TRA) to ensure that the proposed system can meet your organization's risk tolerance for reputation management and enterprise/program viability?

4. For a private blockchain, e.g., a consortium of members, how many members will be willing and able to run one or two blockchain nodes? Is the number of nodes and members sufficient to realize the decentralization benefits that blockchain offers?

Audit logging

UI caching, scrolling and paging

Front-end processing and caching should not hide or obfuscate which data were read, e.g., the granularity of a patient list or search result should not exceed 1 page of visible text since an audit trail for scrolling is not realistic. Applying paging is a better solution for audit requirements and the audit trail should be sent to the back-end for storage as a more trustworthy and reliable component.

Audit logging at all points

Apply audit logging at all points to ensure that unauthorized external/internal side channel access is observed and tracked. This is a quasi-standard practice because audit logging is often recommended and implemented – but not for all components, at all levels and for all accounts (e.g., sometimes administrative activity in a database is not logged). The requirement sounds simple enough, but the amount of space required in a infrastructure grows very quickly, so it is not a trivial task to log elastic services, rotate logs and archive audit trails, possibly with compression, with integrity, recoverability and easy replay/use in the context of a security incident and event management (SIEM) system.

Logging at all points includes applying file system audit logging at the operating system (kernel) level to monitor and audit all activity including administrative activity.

While databases keep transaction logs, applying audit logging for system and administrator-level activity is a separate and less common step, but an important step.

Remote logging

Apply remote logging to a secure location to ensure their integrity, with journaling if available (so that any past state can be recovered and no data are overwritten). Or take hashes of the log with digital signatures so that any tampering becomes evident.

Conclusions

In this paper we focused on controls which are not commonly implemented because they are non-standard or quasi-standard controls—e.g., controls which pertain to advanced and optional requirements, emerging technologies and/or multi-stakeholder distributed enterprise architectures. We provide conclusions in various categories as follows.

In de-identification (privacy), we conclude that:

- Precision medicine requires pseudonymized health information.
- De-identification is not a binary nor standard process/outcome, although it sounds like it—instead, there are different approaches, decisions and levels.
- Clinical validity and privacy should be set or assessed as constraints or objectives—where the impact of de-identification on accuracy and other measures can be reported without revealing data or results.

In consent, use and disclosure (privacy), we conclude that:

- Whereas consent and disclosure directives are applied to limit access in primary care, we recommend the electronic enforcment of all types of consent directives for permitting and limiting secondary use, the codification of purposes for use, and electronic requests for data to fulfill a patient's express and informed wishes that their data be used for defined purposes for multiple projects meeting those purposes and standard/previously defined governance criteria and controls.
- Processing records as input to a search, list or query is a use or disclosure, even if that record is not returned in the output, nor read by a person.
- The system should enforce consent/disclosure directives in sequence to include or exclude of a patient's whole or partial record based on time, topic (e.g., physical abuse), organization (e.g., rehab), information type (drugs, genomics, lab results) or users.

3: Privacy and Security Technology

In other privacy principles/topics, we conclude that:

- Globally unique IDs for patients should be privacy-protective
- Patient access/use/disclosure reports support privacy, however the HIN can only report on the logs it keeps—it cannot provide system-wide reports.

In semantic data integrity (security), for data quality, we conclude that:

- Users should be able to set and assess the error tolerance and confidence in records consolidated across multiple sources via patient matching.
- The system should use globally unique IDs for patients. This assists the consolidation of distributed records and it allows patient matching to be static, for consistent results/quality in query and analysis, instead of dynamic.
- The software configuration, i.e., the versions of all components including their dependencies, should be the synchronized across all distributed nodes at regular intervals, to mitigate the risk that query/analysis tasks are performed differently across nodes. Elastic back-end services with inexact dependencies in dynamic builds, e.g., Node.js, exacerbate this risk.

In security architecture, we conclude that:

- Federated identity management leverages the security of local identity proofing and user authorization as repeatable and standardized processes, i.e., secure processes. Whereas, alternative approaches have disadvantages.
- The open standard for authorization does not assist user or subject identity management needs in the HIN (without further investigation).
- Cloud computing abstraction and elasticity offer attractive functionality but present audit, logging and configuration management challenges and new risks from less mature design patterns and implementations
- Cloud computing containers (or serverless architecture) present new risks from less mature design patterns and implementations
- Standard architecture methods—with conceptual, logical and detailed design—provide the forethought and flexibility needed to avoid getting locked into specific vendors and products which may become insolvent (an availability risk) or unsupported (an integrity risk).
- In centralized storage, high-risk data should be segregated from low-risk data.
- Federated query, distributed query, machine learning clustering/distribution & notebooks are new design patterns – new risks
- The feasibility and utility of differential privacy and homomorphic encryption are not readily apparent and would require careful investigation prior to use. Format preserving encryption for tokens, would also require careful consideration prior to use.
- Blockchain presents both opportunities and challenges that would require careful consideration. Otherwise, any of the other controls discussed in this paper can provide benefits over the status quo.

References

Aldeen, Y. A. A. S., Salleh, M., & Razzaque, M. A. (2015). A comprehensive review on privacy preserving data mining. *SpringerPlus*, *4*(1), 1–36. https://doi.org/10.1186/s40064-015-1481-x

Black, J., & Rogaway, P. (2002). Ciphers with Arbitrary Finite Domains. *CT-RSA 2002 - The Cryptographers' Track at the RSA Conference 2002*, 114–130. https://doi.org/10.1007/3-540-45760-7_9

Dworkin, M. (2016). *NIST Special Publication 800-38G, Recommendation for Block Cipher Modes of Operation: Methods for Format-Preserving Encryption*. Retrieved from http://dx.doi.org/10.6028/NIST.SP.800-38G

Ghinita, G., Tao, Y., & Kalnis, P. (2008). On the anonymization of sparse high-dimensional data, *00*, 715–724.

Sarathy, R., & Muralidhar, K. (2011). Evaluating laplace noise addition to satisfy differential privacy for numeric data. *Transactions on Data Privacy*, *4*(1), 1–17.

Stapleton, J., & Poore, R. S. (2011). Tokenization and other methods of security for cardholder data. *Information Security Journal*, *20*(2), 91–99. https://doi.org/10.1080/19393555.2011.560923

Díaz-Santiago, S., Rodríguez-Henríquez, L. M., & Chakraborty, D. (2016). A cryptographic study of tokenization systems. *International Journal of Information Security*, *15*(4), 413–432. https://doi.org/10.1007/s10207-015-0313-x

Muralidhar, K., & Sarathy, R. (2010). Does differential privacy protect terry gross' privacy? *Lecture Notes in Computer Science*, *6344 LNCS*, 200–209. https://doi.org/10.1007/978-3-642-15838-4_18

Ryan, K. J. et. al. (1971). The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research. *Biochemistry*, *10*(4), 570–576. https://doi.org/10.1021/bi00780a005

Songstad, N. T., Roberts, C. T., Manley, B. J., Owen, L. S., & Davis, P. G. (2018). Retrospective Consent in a Neonatal Randomized Controlled Trial. *Pediatrics*, *141*(1), e20172092. https://doi.org/10.1542/peds.2017-2092

Kim, W. O. (2012). Institutional review board (IRB) and ethical issues in clinical research. *Korean J Anesthesiol*, *62*(1), 3–12. https://doi.org/10.4097/kjae.2012.62.1.3

Emam, K. El. (2006). Overview of Factors Affecting the Risk of Re-identification in Canada. *Information and Privacy Division of Health*, 29. Retrieved from http://www.ehealthinformation.ca/wp-content/uploads/2014/07/2006-Overview-of-Factors.pdf

Sweeney, L., Abu, A., & Winn, J. (2013). Identifying Participants in the Personal Genome Project by Name. *Ssrn*, 1–4. https://doi.org/10.2139/ssrn.2257732

Honarmand, K., Belley-Cote, E. P., Ulic, D., Khalifa, A., Gibson, A., McClure, G., … Cook, D. J. (2018). The Deferred Consent Model in a Prospective Observational Study Evaluating Myocardial Injury in the Intensive Care Unit. *Journal of Intensive Care Medicine*, *33*(8), 475–480. https://doi.org/10.1177/0885066616680772

Willison, D. J. (2009). *Use of Data from the Electronic Health Record for Health Research – current governance challenges and potential approaches*.

Zyskind, G., & Pentland, A. S. (2015). Decentralizing Privacy : Using Blockchain to Protect Personal Data. https://doi.org/10.1109/SPW.2015.27

Orcutt, M. (2018). How secure is blockchain really? *MIT Technology Review*. Retrieved from https://www.technologyreview.com/s/610836/how-secure-is-blockchain-really/

Sweeney, L. (2002). k-Anonymity: A Model for Protecting Privacy, *10*(5), 1–14. https://doi.org/10.1142/S0218488502001648

Malin, B., & Sweeney, L. (2004). 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*, *37*(3), 179–192. https://doi.org/10.1016/j.jbi.2004.04.005

Li, N., Li, T., & Venkatasubramanian, S. (2007). t-Closeness: Privacy Beyond k-Anonimity and l-Diversity. *IEEE 23rd International Conference on Data Engineering*. https://doi.org/10.1080/15423166.2004.565623507173

Fung, B. C. M., Wang, K., Chen, R., & Yu, P. S. (2010). Privacy-Preserving Data Publishing: A Survey of Recent Developments. *Proceedings - International Conference on Data Engineering*, *42*(4), 305–308. https://doi.org/10.1109/ICDEW.2010.5452722

Bayardo, R. J., & Agrawal, R. (2005). Data privacy through optimal k-anonymization. *Proceedings - International Conference on Data Engineering*, 217–228. https://doi.org/10.1109/ICDE.2005.42

Kindermans, P.-J., Schütt, K. T., Alber, M., Müller, K.-R., Erhan, D., Kim, B., & Dähne, S. (2017). Learning how to explain neural networks: PatternNet and PatternAttribution, 1–12. https://doi.org/10.1021/jm900403j

Dankar, F. K., & El Emam, K. (2012). The application of differential privacy to health data. *Proceedings of the 2012 Joint EDBT/ICDT Workshops on - EDBT-ICDT '12*, 158. https://doi.org/10.1145/2320765.2320816

Emam, K. El, Gratton, E., Polonetsky, J., & Arbuckle, L. (2013). The Seven States of Data: When is Pseudonymous Data Not Personal Information? *Journal of Science & Technology*, *24*(1), 1–14.

Sánchez, X. M. M. (2011). Patient identification errors. *Enfermería Clínica*, *21*(5), 295–296. https://doi.org/10.1016/j.enfcli.2011.07.006

Fernandes, L., Myers, S., & Viola, A. (2006). *A Global View of Patient Matching and Patient Identification*.

Pew Research. (2018). Patients Want Better Record-Matching Across Electronic Health Systems, 1–6.