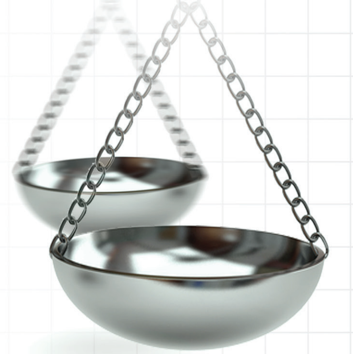
Review

For reprint orders, please contact: reprints@futuremedicine.com



Comparative effectiveness research and big data: balancing potential with legal and ethical considerations

Journal of Comparative Effectiveness Research

Big data holds big potential for comparative effectiveness research. The ability to quickly synthesize and use vast amounts of health data to compare medical interventions across settings of care, patient populations, payers and time will greatly inform efforts to improve quality, reduce costs and deliver more patient-centered care. However, the use of big data raises significant legal and ethical issues that may present barriers or limitations to the full potential of big data. This paper addresses the scope of some of these legal and ethical issues and how they may be managed effectively to fully realize the potential of big data.

Elizabeth Alexandra Gray[1]
& Jane Hyatt Thorpe*[,1]
[1]Milken Institute School of Public Health, George Washington University, 950 New Hampshire Ave, NW 6th Floor, Washington, DC 20052, USA
*Author for correspondence:
Tel.: +1 202/994 4183
jthorpe@gwu.edu

Keywords: big data • comparative effectiveness research • ethics • health information • health information exchange • legal • privacy • research • security

Benefits & current uses of big data

’Big data’ refers to data defined by three ‘Vs’ [1]:

- Volume: massive quantities;
- Velocity: arrives fast and requires rapid processing;
- Variety: includes multiple formats that must be structured and standardized.

These characteristics represent big data’s challenges – traditional database systems do not have the capacity to manage the three Vs, necessitating innovative technical solutions. A fourth ‘V’ is value, generated by analyzing data to uncover meaningful correlations and patterns [2]. Leveraging big data in healthcare can improve decision-making, prevent disease and disease-spread, reduce costs and improve outcomes.

Comparative effectiveness research

Comparative effectiveness research (CER) evaluates and compares health outcomes and the clinical risks and benefits of multiple, commonly offered clinical methods for treating a specific medical condition in a select patient population [3]. CER studies seek to determine the clinical effectiveness of ‘real world’ treatment options and are distinguishable from explanatory clinical trials, which compare an intervention to a placebo under ‘ideal’ treatment circumstances [3]. Comparative effectiveness can be determined using a variety of methods, including [3,4]:

- Methods relying on existing evidence to compare treatment pathways:
  - Systematic review (analyze published studies);
  - Decision analysis (simulation using published evidence).
- Methods generating new evidence of comparative effectiveness:
  - Observational cohort and case–control studies (analyze clinical data);
  - CER-focused randomized control trials (RCT) or ‘practical clinical trials’ (compare intervention to standard-of-care in representative population that is randomly allocated into control and experimental groups).

All these methods except CER-focused RCT can utilize data generated by third

Future Medicine part of fsg

10.2217/CER.14.51 © 2015 Future Medicine Ltd ISSN 2042-6305

parties, which is more cost-effective and less time-consuming than conducting interventional trials [4]. Relying on existing data to conduct CER often means that the data is used for a purpose not contemplated during initial collection. Secondary uses enhance the value of data [5], but are limited in scope by what data was originally collected and the extent of patient consent, if relevant.

Big data solutions enhance secondary uses by aggregating disparate clinical, registry, administrative, claims and patient-generated data into a single set [5]. Integrating multisource data allows researchers to fill in clinical information gaps and study entire populations to determine the effectiveness of drugs or procedures in routine care, identify latent adverse events and compare outcomes across therapies [6]. The size and breadth of integrated databases overcome barriers inherent to small-scale studies, including unrepresentative study groups and insufficient statistical power or precision [7].

The 'Mini-Sentinel' pilot program of the US FDA has implemented big data solutions to advance CER. This is the first step toward building the Sentinel System, a nationwide surveillance system for medical products [8]. The Mini-Sentinel database provides access to routinely collected, standardized healthcare data held by collaborating institutions, representing 153 million patient records [9]. Data is used to assess health outcomes, occurrences of diagnoses and procedures, and the impact of FDA regulatory activities.

Big data algorithms can simplify the study design process, including identification of patient populations and research protocol development, by quickly and efficiently stratifying patient cohorts [5], automatically identifying candidates for studies and identifying disease co-occurrence patterns. The Electronic Medical Records and Genomics (eMERGE) Network, a consortium of medical research institutions funded by the NIH, is an example. eMERGE participants developed algorithms that extract phenotype information from free text in electronic medical records and automatically identify cases and controls [10].

The federal government is a key driver of CER. The Patient-Centered Outcomes Research Institute (PCORI), established by the Patient Protection and Affordable Care Act [3], is an example. PCORI funds selected CER projects based on national priorities; to date, PCORI has provided US$464.4 million to 279 projects [11]. In addition to financing studies, PCORI funds infrastructure development and research on methodology, communication and dissemination [11]. In 2013, PCORI announced its vision for a national data infrastructure to support big data solutions and advance CER, and is building PCORnet, a national network to collect and share standardized clinical data from various settings to facilitate clinical outcomes research [12].

Quality improvement

Service underuse, overuse and misuse plague the US healthcare system, producing poor outcomes and high costs [13]. Healthcare quality improvement modifies administrative and clinical processes to reduce variations in care delivery and administrative process, and improve patient outcomes [14].

Quality improvement seeks to adopt best practices, which can be generated and implemented using big data solutions. These solutions improve business processes, such as tracking hospital assets and administrative transactions, and conducting asset management [15]. The Mayo Clinic successfully uses big data to conduct clinical analysis and assess process connections to improve care. For example, Mayo practitioners sought to improve management of recovery beds using a simulation model to predict bed needs; during the modeling phase, practitioners identified best practice protocols that reduced bed needs by 30% [16].

Clinical decision support

Evidence-based medicine thoughtfully uses evidence to make decisions about patient care by integrating individual clinical expertise with the best possible research findings, including CER [17]. CER generates volumes of clinical knowledge that a practitioner cannot individually process, particularly in the midst of care delivery [17]. The 'inferential gap' between the information available during care and the clinical knowledge required to determine the best treatment must be bridged to facilitate delivery of evidence-based medicine [5]. Big data solutions support evidence-based guideline creation and increase real-time access to knowledge in the practice setting [18]. Libraries of clinical evidence can be collected and programmatically implemented into clinical decision support systems, which can mine and analyze guidelines against real-time electronic health record (EHR) patient data 'at the bedside' to inform decision-making and improve safety [5].

Consumer engagement

Engaging patients improves disease management, facilitates patient–physician communication and promotes wellbeing [6]. CER provides vital information to predict the course of disease and identify best treatments but this is only part of the picture; most of a patient's health is determined by factors other than healthcare delivery, including health behaviors, genetics, socioeconomics and physical environment [19]. Social and behavioral information provides insight on whether a patient will comply with a treatment protocol, levels of engagement in disease management programs and whether a wellness regimen is the right fit for a patient [20]. Big data platforms can collect behavior and sentiment data and combine it with clinical information to enable researchers to learn how to target and retain patients.

Developments in patient-oriented care have expanded patients' involvement in medical decisions and management. These include Blue Button, Personal Health Records (PHRs), mobile health technologies (e.g., FitBit), and social networking sites (e.g., PatientsLikeMe). These tools facilitate patient engagement and are a robust source of data, which can be integrated with clinical records using big data solutions. Patient-generated information 'increase[s] the ability of health researchers to perform translational research, better understand clinical effectiveness of therapeutics and opens doors to increased understanding of environmental and behavioral influences on disease' [6].

Big data can transform healthcare research and delivery. As with all health information, there are legal implications for the collection, analysis and use of big data.

Legal framework

There is no comprehensive framework for health information privacy and security [21]. The patchwork of federal and state laws and regulations often overlap but leave many domains unregulated.

Federal laws & regulations

The Health Information Portability & Accountability Act of 1996 (HIPAA)

The HIPAA Privacy, Security and Breach Notification Rules (the Rules) govern 'protected health information' (PHI), individually identifiable information relating to an individual's care or past, present or future physical or mental health condition or payment for care [22]. Individually identifiable information directly identifies a person or contains information that permits identification (e.g., address). The Rules apply to 'covered entities' (health plans, healthcare clearing houses and most healthcare providers) and their 'business associates' (entities that have access to or use PHI when performing certain functions or services for or on behalf of the covered entity) – collectively referred to herein as 'regulated entities'.

The Rules do not apply to 'de-identified information' [23]. Information is de-identified when eighteen specific identifiers are removed from the record ('Safe Harbor' method; see Box 1) or an expert determines that there is minimal risk that information could be used to identify an individual ('Statistical' method) [24].

The Privacy Rule controls regulated entities' disclosure of PHI. Regulated entities are required to disclose PHI to the individual or his or her designated representative and to the Secretary of the US Department of Health and Human Services (HHS) for enforcement [23]. Regulated entities are permitted (but not required) to disclose PHI without individual authorization in accordance with a permissive disclosure exception. Regulated entities must limit most permissive disclosures to the minimum amount of PHI necessary to achieve the intended purpose for which the information was released.

Disclosures for treatment, payment & healthcare operations

Regulated entities may disclose PHI without authorization to carry out the covered entity's [25]:

- Treatment (provision, coordination or management of healthcare and related services among providers; consultation between providers; or patient referrals);
- Payment (activities associated with obtaining premiums, fulfilling coverage responsibilities, providing benefits and obtaining reimbursement);
- Healthcare operations (six categories of activities) [26].

Regulated entities may also disclose PHI without authorization to:

- Enable another provider's treatment activities;
- Another covered entity or provider to facilitate that entity's payment activities;
- Another covered entity for certain operations, if both entities have (or had) a relationship with the individual and the PHI pertains to that relationship [25].

Disclosures for certain public health activities

Regulated entities may disclose PHI without authorization for specific 'public interest activities', which includes disclosures for:

- Public health surveillance, investigations and intervention;
- Activities related to quality, safety or effectiveness of FDA-regulated products [27].

Disclosures for research

A regulated entity may disclose PHI for research purposes without authorization when:

- Research is on decedents' PHI and the PHI is necessary for research purposes;

Box 1. Safe Harbor method of de-identification.

In order to satisfy the Safe Harbor method of de-identification, all of the following elements must be removed from the record:

- Names
- All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census:
  - The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people.
  - The initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000.
  - All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older
- Telephone numbers
- Fax numbers
- Email addresses
- Social security numbers
- Medical record numbers
- Health plan beneficiary numbers
- Account numbers
- Certificate/license numbers
- Vehicle identifiers and serial numbers, including license plate numbers
- Device identifiers and serial numbers
- URLs
- IP addresses
- Biometric identifiers, including finger and voice prints
- Full-face photographs and any comparable images
- Any other unique identifying number, characteristic, or code

- PHI will be used ‘preparatory to research’ (e.g., prepare a research protocol), is necessary for research purposes and will not be physically removed from the regulated entity; or
- An Institutional Review Board (IRB) or Privacy Board alters or waives the authorization requirement after determining that:
  - The research could not be conducted without the waiver or alteration or without access to and use of PHI;
  - The use or disclosure presents minimal privacy risk to individuals; and
  - The researcher has:
  - A plan to protect identifiers from improper use and disclosure and to destroy identifiers at the earliest opportunity (unless retaining identifiers is required by law or is justified for a health or research purpose); and
  - Assured that PHI will not be reused or further disclosed except as legally required, for research oversight, or for other research for which the Privacy Rule would permit use or disclosure [27].

Generally, regulated entities may not sell PHI, but may levy a ‘reasonable, cost-based fee’ to prepare and transmit PHI for research [23].

Limited data sets

A limited data set (LDS) is PHI with sixteen direct identifiers removed [24]. A regulated entity may disclose a LDS without authorization for research, public health or operations if the parties enter into a data use agreement. The data use agreement must:

- List permitted uses and disclosures;
- Identify who may use or receive the LDS;
- Provide that the recipient will abide by specific requirements to protect privacy and security.

Authorizations

Any disclosure not identified as required or permissive requires individual written authorization [28]. Authorization is also required for most disclosures of psychotherapy notes, sale of PHI, and most marketing uses and disclosures. An authorization to use or disclose PHI for research may be combined with any written permission for the same or another research study (e.g., informed consent).

The Security Rule requires that regulated entities protect electronic PHI (e-PHI) by maintaining reasonable and appropriate safeguards [29]:

- Administrative (e.g., procedures for accessing e-PHI);
- Physical (e.g., specified use of workstations that access e-PHI);
- Technical (e.g., audit controls for systems containing or using e-PHI); and
- Organizational (e.g., business associate agreements contain applicable specifications).

Entities may use any security measures to reasonably and appropriately implement standards and specifications.

The Breach Notification Rule governs 'unsecured PHI' (not rendered unusable, unreadable or indecipherable to unauthorized individuals) [30]. A breach occurs when a prohibited use or disclosure 'compromises the security or privacy of the PHI'. When a breach occurs, business associates must notify the covered entity, and covered entities must notify affected individuals and HHS.

The Common Rule protects most human subjects involved in federally-funded research [31]. A human subject is an individual about whom a researcher obtains data through intervention, interaction or identifiable private information [32]. 'Private information' is information provided for a specific purpose that the individual reasonably expects will not be made public (e.g., a medical record). The Common Rule only applies to individually identifiable private information (i.e., subject identity may be readily ascertained by the researcher or associated with the information). The following types of research are exempt from the Common Rule:

- Research using survey or interview procedures or observation of public behavior, if:
- The results are recorded in a way that does not permit identification;
- Disclosing subjects' responses could not expose them to liability or damage.
- Research involving data, records or bio-specimens that exist when the study commences, if the results are recorded in a way that does not permit subject identification or if the sources are publicly available [31].

An IRB will only approve research where subjects (or their legally authorized representatives) have given informed consent to participate [33]. Researchers must provide potential subjects with specific information about the research and give them adequate time to consider whether to voluntarily participate. An IRB may waive or alter all or part of the informed consent requirements if:

- The research involves no more than minimal risk to subjects;
- A waiver will not adversely affect the subjects' rights/welfare;
- The research could not be carried out without waiver or alteration; and
- Subjects will be provided with pertinent information whenever appropriate [34].

Informed consent must be documented via a signed consent form, though an IRB may waive this requirement if: the form would be the only record linking the subject and the research, and the principal risk to the subject would be harm resulting from a breach of confidentiality; or the research presents minimal risk of harm to subjects and does not involve procedures normally requiring written consent outside the research context [34]. In July 2011, HHS issued an advance notice of proposed rulemaking, seeking comment on improving the Common Rule's effectiveness, and proposed to:

- Establish data security protections to maintain confidentiality of identifiable information, applicable to all (even 'exempt') research – including adopting HIPAA's definitions of individually identifiable information and de-identification;
- Require written consent for research using biospecimens, including those stripped of identifiers;
- Broaden exemptions for research using existing data to permit secondary uses of identifiable information;
- Eliminate continuing review requirement for research limited to obtaining follow-up clinical information and analyzing research data [35].

The Genetic Information Nondisclosure Act of 2008 (GINA)

GINA prohibits health plans and issuers from using genetic information to make eligibility, coverage, underwriting or premium-setting decisions [36]. Generally, health plans and issuers may not request or require that beneficiaries undergo genetic testing or provide genetic information, but may request voluntary provision of genetic information for research.

GINA prohibits employers from using genetic information to discriminate against employees or applicants and from acquiring employee or applicant genetic information, subject to exceptions [36]. Genetic information acquired by an employer may be disclosed to an occupational or health researcher and to a public health organization in limited circumstances.

The Privacy Act of 1974 & the United States Freedom of Information Act (FOIA)

The Privacy Act protects identifiable information about individuals held or collected by the federal government [37]. A federal agency may release information to individuals or their designees with written consent or pursuant to a disclosure exemption, including disclosures for statistical research, agency-specific routine uses and as required by FOIA.

FOIA provides that any person may access information contained in federal agency records unless information is exempted from disclosure [38]. FOIA Exemption 6 prohibits disclosures of information about individuals in ‘personnel, medical, and similar files’ that ‘would constitute a clearly unwarranted invasion of personal privacy’. Extensive legal scrutiny of this exemption has sought to balance public interest in information with individual privacy rights. In 2013, this balance shifted in favor of public interest via a court order overturning a 1979 decision, declaring that provider-identifiable Medicare claims data was no longer protected by Exemption 6 [39].

The Centers for Medicare & Medicaid Services (CMS) subsequently announced its intention to release individual physicians’ Medicare billing data in response to FOIA requests [40], and in April 2014 released Part B claims data for 2012, which may be used by anyone for any purpose [41]. CMS recently issued a proposed rule to expand access to Medicare Part D data, which would make unencrypted prescriber, pharmacy and plan identifiers in prescription drug event records available to external researchers, subject to limitations [42].

42 CFR Part 2 (Part 2)

Part 2 limits disclosure and use of identifying information that could or does reveal an individual received substance abuse treatment and applies to federally-assisted programs providing substance abuse diagnosis, treatment or referral – this includes programs that participate in Medicare, have a US Drug Enforcement Administration (DEA) number or are federally tax-exempt [43].

Written patient consent is required for disclosure, with limited exceptions, including to qualified researchers [43]. A single consent form may authorize disclosure to multiple parties or for multiple purposes, and must include a statement prohibiting the recipient from further disclosing the information without consent or unless permitted by Part 2.

State framework

States define their own privacy framework. [44]. State laws often apply to the same entities, activities or types of health information as federal laws (e.g., genetic information), but may be contrary to federal requirements; generally, entities must comply with whichever law is most protective [45]. State laws also protect certain sensitive information and vulnerable populations not subject to the federal framework. These laws vary in scope and may govern HIV/AIDS and sexually transmitted disease information [46], mental health information [47], or minors and those legally declared incompetent.

In addition to patient privacy laws and regulations, some states have passed legislation requiring physician consent to release certain information, including performance measurement data and prescribing practices. However, in 2011, the US Supreme Court declared one such state law unconstitutional. In Sorrell versus IMS Health, the Court held that pharmaceutical and data-mining companies’ free speech rights were violated by a Vermont law that prohibited the sale of records containing a physician’s prescribing practices or the use of such records for marketing without physician consent [48].

Legal & ethical challenges

Confusion and fear of penalties surrounding the scope and applicability of the legal framework is a barrier to robust use of big data. Researchers must understand when laws apply and that laws apply to the same data differently depending on who holds the information, who wants to use it and for what purpose. There are also ethical implications relevant to certain activities.

Privacy & Security Rule breaches

The use of identifiable information involves privacy risks; the Security Rule mitigates these by requiring that regulated entities implement security mechanisms and policies, but affords flexibility to meet these standards [29]. When acquiring data from a regulated entity, researchers must tailor their protocols to that entity’s practices, which can be cost and time prohibitive if protocols mandate extensive training or use of specific software. This is especially burdensome for CER studies, which may have limited budgets and less time flexibility than other forms of clinical research [49].

Breach risk can be reduced by applying data minimization practices, which limit information collected to that necessary to accomplish a specified purpose and call for data destruction once that purpose is achieved [2]. The Privacy Rule’s minimum necessary requirement is a form of data minimization, but does not limit the scope of PHI that may be disclosed [50] or mandate data destruction. Although adhering to the

Privacy Rule achieves legal compliance, privacy ethics may require application of data minimization. This would significantly reduce, if not eliminate, the value of big data.

Big data is about the power of discovery, and its business model incentivizes collection of more data for longer time periods to enable unanticipated secondary uses of data [51]. Using existing data saves time and money, and permits access to a cross-section of information [52], but strict application of data minimization prohibits secondary use of clinical, claims and administrative data for research. These existing data sources are necessary for observational studies, a primary method of CER. Destroying information collected for research upon study completion means the data is unavailable for future studies, which significantly inhibits CER that relies on comparing previous studies and existing data.

Concerns with de-identification

The federal framework for privacy and security does not apply to de-identified information, because it poses minimal risk to privacy. Many IRBs waive informed consent requirements if data is de-identified [53]. However, ‘identifiable information’ is inconsistently defined across the framework. The Privacy Rule is the only federal regulation that specifies a process for de-identification or delineates specific identifiers that must be removed before information is considered ‘de-identified’. Adhering to these standards is sufficient to comply with all other laws and regulations [54], though the Common Rule and other laws are likely satisfied by a lesser degree of anonymization. Where the laws offer flexibility, researchers must balance patient privacy rights against practical considerations in selecting which protocols to implement. However, researchers should become accustomed to using these methods - the HIPAA standards for de-identification may soon be incorporated into the Common Rule, as suggested by the 2011 advance notice of proposed rule-making [35].

Questionable utility

De-identified data is not useful for all research [5]. For example, genetic information is PHI, precluding the use of de-identified information for genetic research [55]. Causal relationships are often determined by linking records from multiple settings or time periods to a single patient, which cannot be accomplished using anonymized records [56]. A LDS may be an effective alternative, because it retains data elements to link patient records. However, an LDS is still PHI subject to HIPAA, and aggregating disparate LDSs is impossible unless each set has retained the exact same data elements [56].

There are ethical concerns with using de-identified information for research as well. De-identified information does not enable researchers to follow-up with patients, avoid duplicative work and communicate findings to patients [14,57]. The good resulting from protecting patient privacy may be outweighed by the value these functions offer patients.

Eroding concept of de-identified data

General public perception is that technological improvements and vast quantities of available public data have rendered de-identification impossible [56]. One study showed that those who declined to disclose information for anonymous research did so because they did not believe it was possible for information to be de-identified [57]. Although following HIPAA’s de-identification protocols is legally sufficient, ethical concerns exist if patients do not trust the system.

A patient may withhold information from his provider to avoid having it used for anonymous research where he believes that de-identification is impossible [50]. Good healthcare depends on accurate and reliable information; failure to disclose information to a provider can negatively impact care delivery and increase safety risks. Researchers must consider options for accessing patient information that build stakeholder trust, such as obtaining patient consent.

Lack of consistent framework for patient consent/authorization

Every federal law permits use and disclosure of patient data for research with patient authorization – it is the key common denominator. Researchers can do anything with patient data by obtaining effective consent, but numerous legal and ethical issues related to consent exist.

There is no common consent architecture. The same information may be governed by some laws but not others; for example, the Common Rule does not apply to studies of existing clinical records, but HIPAA requires an authorization to disclose such records absent IRB waiver. Entities may have different obligations depending on the type of information; for example, Part 2 prohibits the recipient of substance abuse records from re-disclosing that information without additional consent, but HIPAA does not limit re-disclosure of PHI by an unregulated entity [50]. The same researcher may have different obligations depending on the intended use of the information; for example, researchers may access PHI without authorization to identify patients for recruitment but HIPAA requires an authorization to contact those patients about participation [58].

It may be impracticable to obtain patient consent, especially for retrospective records-based studies such as observational CER studies [53]. Finding patients is time-consuming and costly [58]. Seeking consent may produce biased results; individuals who do not consent to disclosure may possess different clinical, demographic and socioeconomic characteristics to consenters [5,53,59].

Also, the process by which one obtains consent may raise ethical concerns. Most consent forms are technical and phrased to obtain consent. Patients may authorize disclosure of their information without fully understanding what they are agreeing to [50] or may voluntarily provide information in one context without realizing that information may be used for another purpose without their knowledge [52]. This is common with respect to social media sites, which mine user-generated information to characterize health behaviors and sell that data to researchers. This information can reveal valuable insights about patient characteristics that impact treatment effectiveness, which may provide significant value for CER. Because patient-generated data exists outside the traditional healthcare domain, it is not subject to the federal framework for health information privacy and consent requirements are inapplicable [50]. However, this does not obviate ethical considerations and researchers must consider whether using this information is appropriate.

Challenges presented by data segmentation

Data segmentation walls off data that is considered undesirable to share [56]. Because some sensitive information may be subject to heightened protection, data segmentation simplifies the consent process, as users need only comply with laws applicable to the remaining information. Further, patients may be more willing to consent to disclosures that will not include information about which they feel most vulnerable. Despite its benefits, data segmentation can be technically and logistically complicated, and expensive.

One of the primary challenges relates to categorizing data elements – for example, to segment HIV status, the system must strip direct references to HIV and HIV treatments as well as any other data elements that, when viewed together, may indicate that the patient is HIV-positive [60]. Building algorithms to effectively infer such correlations is a complex process that is not yet well-developed. It also may be possible to combine other information (e.g., consumer behavior data) with segmented records to make predictions about sensitive information, which raises serious ethical concerns.

Looking ahead: balancing considerations

Researchers need a comprehensive understanding of varying legal requirements in order to effectively harness the potential of big data. Researchers must also consider ethical issues raised by the use of identifiable information, even where it occurs in accordance with the legal framework. There are many ways to balance these concerns to facilitate the use of big data.

Develop consistent framework for patient consent

Patient authorization and consent is the master key that unlocks all patient health information; the pace of research and discovery increases dramatically when individuals voluntarily disclose information [61]. Generally, patients are willing to participate in research if they are asked [56]. Seeking consent facilitates patient engagement, increases transparency and may improve data quality [62].

The HIPAA rules permit authorizations to be combined with any other legal permission related to the same or other research studies, and all laws permit the inclusion of consent elements that are not contrary to those required. Given this flexibility, a common consent form would alleviate some of the practical challenges related to patient consent. Table 1 highlights consent elements shared across the legal framework, which need only appear once in a common consent form. A dynamic form generated using big data solutions includes only those consent elements applicable to the information in a patient's record or the intended use of the data. For example, where the patient has never received substance abuse treatment, or where researchers have no need for such information, the consent form would not include Part 2 consent requirements. Disclosures based on dynamic consent forms may only include those parts of a record for which a patient has given consent, which requires implementation of data segmentation procedures.

HIPAA also permits authorization for unspecified future research, if it describes the future research such that an individual could reasonably expect that his PHI could be disclosed for such a purpose [28]. Individuals prefer to be asked permission for their records to be used, even in a general way, but do not need to know about the specific study or its timing [57]. Covered entities and researchers should collaborate to seek consent for future research where possible to simplify the consent process, avoid overwhelming patients with multiple successive consent forms and ensure that patients understand how their information may be used.

Researchers should also consider whether to partner with providers to establish consumer-interactive consent management systems, which allow patients to actively manage their research preferences [56]. Patients can opt in or out of consideration for certain types of research or for inclusion in research databases through remote access to their electronic medical record or via a PHR offered by their provider.

Table 1. Federal requirements: authorization to disclose protected information

	HIPAA	Common Rule[†]	GINA	Part 2	Privacy Act[‡]
Element					
1. Specific description of information	X	X	X	X	X
2. Identify person(s) or entity authorized to make the requested disclosure	X			X	
3. Identify person(s) or entity authorized to receive the requested information	X	X	X	X	X
4. Describe the intended use(s) of the requested information.	X	X	X	X	
5. The expiration date or event	X	X		X	X
6. Date signed	X			X	
7. Signature of the individual or his personal representative	X	X		X	
Include the following information					
The individual's right to withdraw authorization (if any) and any applicable exceptions to that right.	X	X		X	
Whether any benefits may be conditioned on releasing the information and applicable consequences of refusal to consent. This includes stating that refusal will involve no penalty or loss of benefits where relevant (e.g., GINA).	X	X	X[§]		
The potential for recipient re-disclosure of the information, if any. This includes stating that information may not be re-disclosed without further authorization, where applicable (e.g., Part 2).	X			X	
The authorization must be written in plain language	X	X			
Provide the individual with a copy of the form	X	X			

[†] These requirements apply to informed consent forms and would be relevant to research protocols that seek access to identifiable information from patient records.
[‡] These requirements apply where an individual authorizes an agency release records – requirements are agency-specific; the ones listed here are per HHS.
[§] This requirement applies only to disclosures by health plans and issuers.
GINA: Genetic Information Nondisclosure Act of 2008; HSS: Department of Health and Human Services; HIPAA: Health Information Portability and Accountability Act of 1996.

The consent process can be built into a clinical decision support system. The DISCERN project at Duke University is an example – when a provider enters information in a patient's record, the system queries the EHR for additional patient information, checks it against relevant research protocols and automatically alerts researchers if the patient is a potential candidate for a study [63]. This system could automatically generate a consent form during the patient visit, giving the provider an opportunity to discuss it with the patient. This is beneficial, as patients are more likely to consent to disclosure of their information when asked by their treating provider than any other entity [57].

Re-conceptualize de-identification

Most patients do not object to their information being appropriately used by healthcare professionals if safeguards are in place [59]. De-identified information can be one of those safeguards, but claims that data can be re-identified with ease have cast doubt on the public's belief that identifiable information can truly be de-identified [53]. However, these beliefs may be unfounded. A 2011 review of articles detailing re-identification of health data discovered that only one attack occurred on a data set de-identified in accordance with HIPAA [53]. Further, only 0.013% of that information was re-identified [53] – within the statistical limits of a 'minimal risk' to privacy. To date, there appears to be no published evidence of a re-identification attack on a data set de-identified in full compliance with HIPAA. This underscores the critical importance of de-identification in accordance with HIPAA standards, which may be the only way to protect against re-identification.

Big data platforms improve the utility of de-identified data for research. Recent studies have demonstrated that it is possible to conduct longitudinal studies using de-identified data [64]. Other options for enhancing the value of de-identified data include the creation of federated databases for research. In such a system, researchers conduct statistical analyses of distributed databases and receive summary information without direct identifiers

[65]. PHI is held only at the source, where it is cleaned and analyzed in a common way [66]. Summarized data is sent to a centralized data repository, which releases data relevant to a specific research question.

Consider purpose-specific data minimization standards

A common big data refrain is 'collect it now, decide what to do with it later' [67]. This allows the data to 'speak', telling users what questions it can answer. This method is contrary to data minimization principles, and may result in unnecessary costs and time. To manage this, researchers must understand what data is available and the questions it can answer [2]. Non-traditional information sources, including patient-generated data, may contain useful signals for CER, particularly in combination with traditional sources such as EHRs and claims data. Developing a better understanding of data enables researchers to conceive of possible uses prior to collection, identify what data is needed and build study protocols around available data.

Reconsider data segmentation process

Until data segmentation procedures are better-developed and reliable, data users must consider other ways of protecting sensitive information. Individuals will disclose their information if they believe that it will be used in a 'just way that does not negatively impact them in the future' [61]. This requires healthcare organizations and researchers collaborate to develop acceptable information use standards.

Education and security

Patient response to data sharing is impacted by factors including their involvement, how the request for use of information is frame and how well they understand the technology [61]. Patients will relinquish some privacy if the incentives are adequate [61]; it is incumbent upon researchers to explain why research is important, how research may be relevant to patients and why medical records are essential to conducting research, especially CER [57].

Information security is also critical to build patient trust and minimize risk. Organizations should implement reasonable security measures, which may include encrypting data or adding verification controls to confirm that a user has permission to access certain information [56]. Education on the benefits of research and the safeguards that are in place increases patient willingness to disclose [68,69].

Conclusion

Big data solutions have the potential to transform comparative effectiveness research, by facilitating the collection and aggregation of volumes of multi-source data to enable comparisons across care settings, patient populations and treatment combinations. Big data use relies on access to health information, but must contend with varying and often misunderstood legal requirements. Ethical considerations related to patient privacy complicate the big data universe and require that researchers navigate the tension between privacy and promoting discovery. Understanding complicated legal requirements while balancing ethical obligations is a daunting task. Researchers can achieve this balance by engaging patients, streamlining research processes and reframing the way they think about information in a big data world.

Future perspective

The potential of big data to support comparative effectiveness research is both great and dynamic. Improvements in technology such as EHRs and PHRs, which enable greater data collection and accessibility, will permit more real-time randomized observational studies, and the integration of personal genomics data into EHRs will allow patient-specific comparisons with others on the same treatment regimens. Furthermore, integrating patient-generated information into CER opens up the possibility for patients to share their feedback with the public and other affected users, allow for cross-comparisons of responses, and feed this data into PHRs, all of which open up previously unavailable avenues for research. Finally, analytics capabilities that are more predictive can integrate clinical data with contextual, real-world information to improve patient-risk stratification and preventive care. As the technology is evolving, there is also a major shift in public perceptions of privacy (online information sharing etc.) that may fundamentally change the way society views confidentiality and the benefit of disclosure for the public good. It is unclear whether the current legal framework is sufficient to transcend this evolution. While there are benefits to aligning federal and state laws, expanding access to data (e.g., release of identifiable Medicare claims data) and enhanced efforts to obtain patient consent will reduce or eliminate many barriers. As such, researchers and policymakers must carefully consider possible unintended consequences that may come with greater regulation.

Acknowledgements

The authors wish to thank J Lerner and the ECRI Institute for addressing big data at their November 2013 annual conference and inviting J Thorpe to speak.

Financial & competing interests disclosure

E Gray and J Thorpe were funded by an external award from the Robert Wood Johnson Foundation (Washington, DC) for

Executive summary

Benefits & current uses of big data

- Managing the three Vs of big data (volume, velocity and variety) is the first step in harnessing the possibilities of big data.
- Data is only valuable if one can make sense of it; extracting value from big data is the most important consideration, which requires: analytics to uncover patterns and correlations, and the ability to mine massive stores of data to quickly return relevant information.
- Effective use of big data is subject to legal considerations governing privacy and security, which are complex and varied; misconceptions and lack of understanding about the legal framework applicable to health information operate as a barrier to robust use of big data in comparative research, clinical decision support, quality improvement and consumer engagement.

Legal framework

- There is no overarching, comprehensive legal framework applicable to health information; federal and state laws and regulations governing health information often overlap and in some cases may contradict each other.
- The HIPAA Rules govern only disclosure of individually identifiable health information held by certain entities. HIPAA does not apply to de-identified data. Researchers may use health information without an individual's authorization in several situations; changes to HIPAA permit compound authorizations and authorizations for unspecified future research uses.
- The Common Rule does not apply to research using existing data that does not identify subjects or de-identified data; informed consent may be waived where risk to the subject is limited to a breach of confidentiality. Recent proposed changes to the Common Rule would align it more closely with HIPAA's de-identification requirements and ease burdens on comparative effectiveness researchers.
- A recent change in the law made individual physician's Medicare payment data available for research use, and a recent proposed rule may make Medicare Part D prescription drug data more readily available to researchers as well.

Legal & ethical challenges

- Privacy and security
  - Researchers will face differing security protocols when dealing with different entities.
  - Data minimization is antithetical to big data best practices of collecting as much data as possible and holding on to it indefinitely.
- De-identified data
  - No consistent definition of de-identification across laws.
  - De-identified data is not useful for many purposes, including determining causal relationships and conducting genetic research; it cannot be used for various aspects of patient engagement and provider performance measurement.
  - The public does not trust that data can be truly anonymous; using de-identified data may be ethically questionable if it negatively impacts patient-provider relationships.
- Lack of consistent framework for patient consent
  - There is no common consent architecture – different consent elements and processes may apply depending on the situation.
  - Seeking consent may be impracticable and can bias results, and patients may not fully understand how their data will be used and by whom.
- Challenges presented by data segmentation
  - Developing systems capable of completely segmenting records is technically challenging and expensive.
  - Segmented data can be combined with other data to infer sensitive information.

Looking ahead: balancing considerations

- Develop consistent framework for patient consent
  - Patients are likely to give consent when asked, particularly by their provider.
  - A common consent form streamlines the consent process and can be created using big data to tailor the form to patients or the research.
  - The consent process can be built into clinical decision support systems and patients can manage their consents using an interactive online process.
- Re-conceptualize de-identification
  - Research shows that de-identification in accordance with a HIPAA method is sufficient.
  - De-identified data can be used for longitudinal studies using big data solutions.
  - Researchers can engage with federated databases that collect data and share it according to specified research protocols to access more and better de-identified data.
- Consider purpose-specific data minimization
  - Researchers must first understand the data that is available and what questions it can answer, develop ideas for how to use the data and only then collect the data.
- Reconsider data segmentation process
  - Data segmentation process is not well-developed; until it improves, researchers and healthcare entities should develop strategies to protect sensitive information without compromising valid uses of data.
- Education & security
  - Patients will disclose information when they understand how the information will be used and protected; researchers must educate patients on the value of research and ensure that security protocols are in place to protect data.

work to develop and maintain an online resource of federal and state laws related to health information (www.healthinfolaw.org). J Thorpe was partially funded under a subcontract with ResDAC to provide guidance related to the Centers for Medicare & Medicaid Services' data release policies. J Thorpe also serves as a senior advisor in the US Department of Health and Human Services Office of the National Coordinator for Health Information Technology. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

References

Papers of special note have been highlighted as:
• of interest; •• of considerable interest

1 Howles T. Data, data quality, and ethical use. *Software Quality Professional* 16(2), 4–12 (2014).

2 Murphy M, Barton J. From a sea of data to actionable insights: big data and what it means for lawyers. *Intellectual Property & Technology Law Journal* 26(3), 8–17,11 (2014).

•• **Article discussing various legal considerations for big data beyond the healthcare domain and into property, control and intellectual property issues.**

3 Concato J, Peduzzi P, Huang GD, O'Leary TJ, Kupersmith J. Comparative effectiveness research: what kind of studies do we need? *J. Investig. Med.* 58(6), 764–769 (2010).

4 Giuliano KK, Ferguson M, Silfen E. Medical technology innovation and the importance of comparative effectiveness research. *J. Med. Marketing* 12(1), 55–66 (2012).

• **Article discussing business dimensions of comparative effectiveness framework and how to leverage technology to improve results.**

5 Jensen P, Jensen L, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat. Rev. Genet.* 13(6), 395–405 (2012).

•• **Review of methods and opportunities to extract clinical and genetic information from health information technology applications to improve outcomes.**

6 Pearson J, Brownstein C, Brownstein J. Potential for electronic health records and online social networking to redefine medical research. *Clin. Chem.* 57(2), 196–204 (2011).

7 Hoffman S, Podgurski A. The use and misuse of biomedical data: is bigger really better? *Am. J. Law. Med.* 39(4), 497–538 (2013).

8 US Food and Drug Administration. FDA's Sentinel Initiative – Background (2010).
www.fda.gov

9 US Food and Drug Administration. Welcome to Mini-Sentinel
http://minisentinel.org

10 National Human Genome Research Institute, National Institutes of Health. Electronic Medical Records and Genomics (eMERGE) Network (2014).
www.genome.gov

11 Patient-Centered Outcomes Research Institute (PCORI). Statement of PCORI Executive Director Joe Selby, MD, MPH, Following Center for American Progress Event on Comparative Effectiveness Research (2014).
www.pcori.org

12 Patient-Centered Outcomes Research Institute (PCORI). PCORnet: The National Patient-Centered Clinical Research Network (2013).
www.pcori.org

13 Chassin MR. Improving the quality of health care: what's taking so long? *Health Aff. (Millwood)* 32(10) 1761–1765 (2013).

14 Roszell S. Research in quality improvement: a safety checklist creates an uproar. *J. Nurs. Law* 13(1), 19–24 (2009).

15 Weitner M, Garvin T. Predictive analytics in safety and operational risk management. *IBM Corporation Somers, NY* (2012).
www-01.ibm.com

16 Marmor YN, Rohleder TR, Cook DJ, Huschka TR, Thompson JE. Recovery bed planning in cardiovascular surgery: a simulation case study. *Health Care Manag. Sci.* 16(4), 314–327 (2013).

17 Eddy D. Evidence-based medicine: a unified approach. *Health Aff. (Millwood)* 24(1), 9–17 (2005).

18 Stewart W, Shah N, Selna M, Paulus R, Walker J. Bridging the inferential gap: the electronic health record and clinical evidence. *Health Aff. (Milwood)* 26(2), w181–w191 (2007).

• **Article defining the complexity of clinical knowledge and the opportunity for health information technology to pair with comparative effectiveness research to manage that complexity.**

19 Robert Wood Johnson Foundation Commission to Build a Healthier America. Beyond health care: New directions for a healthier America. (2009).
www.rwjf.org

20 Fox B. Using big data for big impact. *Health Manag. Technol.* 32(11), 16 (2011).

21 Edwards JE, Halawi LA. Analysis of variation in HIT privacy & security laws. *Cambridge Business Review* 18(1), 240–248 (2011).

22 HIPAA General Administrative Requirements. Applicability. 45 CFR § 160.103 (2013).

23 HIPAA Privacy Rule. Uses and disclosures of protected health information. 45 CFR § 164.502 (2013).

24 HIPAA Privacy Rule. Other requirements relating to uses and disclosures of protected health information. 45 CFR § 164.514 (2013).

25 HIPAA Privacy Rule. Uses and disclosures to carry out treatment, payment, or health care operations. 45 CFR § 164.506 (2013).

26 HIPAA Privacy Rule Definitions 45 CFR § 164.501 (2013).

• **Provides a list of the broad categories of activities.**

27 HIPAA Privacy Rule. Uses and disclosures for which an authorization or opportunity to agree or object is not required. 45 CFR § 164.512 (2013).

28 HIPAA Privacy Rule. Uses and disclosures for which an authorization is required. 45 CFR § 164.508 (2013).

29 HIPAA Security Rule. Security Standards for the Protection of Electronic Protected Health Information. 45 CFR Part 164, Subpart C (2013).

30 HIPAA Breach Notification Rule. Notification in the Case of Breach of Unsecured Protected Health Information. 45 CFR Part 164, Subpart D (2013).

31 The Common Rule. To what does this policy apply? 45 CFR § 46.101 (2013).

32 The Common Rule. Definitions. 45 § 46.102 (2013).

33 The Common Rule. General requirements for informed consent. 45 CFR § 46.116 (2013).

34 The Common Rule. Documentation of informed consent. 45 CFR § 46.117 (2013).

35 US Department of Health and Human Services. Advanced Notice of Proposed Rulemaking: Advance Notice of Proposed Rulemaking: Human Subjects Research Protections: Enhancing Protections for Research Subjects and Reducing Burden, Delay, and Ambiguity for Investigators. 76 Fed. Reg. 44512, (2011).

36 The Genetic Information Nondisclosure Act (GINA). Pub. L. No. 110–233, § 2000ff, 122 Stat. 881 (2008).

37 The Privacy Act. Pub. L. No. 93–579, § 552a, 88 Stat. 1896 (1974).

38 The Freedom of Information Act (FOIA). Pub. L. No. 89–487, 552, 80 Stat. 250 (1967).

39 Fla. Med. Ass'n v. US Dep't of Health, Educ., & Welfare. 947 F.Supp.2d 1325 (M.D. Fla. 2013).

40 US Department of Health and Human Services. Notice: modified policy on disclosure of amounts paid to individual physicians under the Medicare program. 79 Fed. Reg. 3205 (2014).

41 Centers for Medicare & Medicaid Services. Medicare provider utilization and payment data: physician and other supplier. (2014).
www.cms.gov

42 US Department of Health and Human Services. Final rule: Medicare program; contract year 2015 policy and technical changes to the Medicare Advantage and the Medicare prescription drug benefit programs. 79 Fed. Reg. 29843 (2014).

43 Confidentiality of Alcohol and Drug Abuse Patient Records (Part 2). 42 CFR Part 2, (2013).

44 Legal Barriers Project. Health Information & the Law. Deparment of Health Policy, Milken Institute School of Public Health at the George Washington University, Robert Wood Johnson Foundation (2012).
www.healthinfolaw.org

- **provides summaries of health information federal and state laws and regulations, 50-state comparative maps, and analytical briefs.**

45 Pritts J, Lewis S, Jacobson R, Lucia K, Kayne K. Privacy and security solutions for interoperable health information exchange: report on state law requirements for patient permission to disclose health information. US Department of Health and Human Services. (2009).
www.healthit.gov

46 Centers for Disease Control and Prevention. State HIV Laws. (2013).
www.cdc.gov

47 Pritts JL. The importance and value of protecting the privacy of health information: the roles of the HIPAA Privacy Rule and the Common Rule in health research. National Academy of Sciences. (2008).
www.iom.edu

48 Sorrell v. IMS Health, Inc. 131 US 2653 (2011).

49 Harrington SE. Incentivizing comparative effectiveness research. Ewing Marion Kauffman Foundation, Kansas City, MO (2011).
www.kauffman.org

50 Mcgraw D, Dempsey JX, Harris L, Goldman J. Privacy as an enabler, not an impediment: building trust into health information exchange. *Health Aff. (Millwood)* 28(2), 416–427 (2009).

- **Article advocating framework that builds security and privacy protections into health information technology - some recommendations have been implemented since article publication, inviting comparisons of current framework against promised outcomes of proposal.**

51 Tene O, Polonetsky J. Big data for all: privacy and user control in the age of analytics. *Northwestern J. of Technology and Intellectual Property* 11(5), 34 (2013).

52 Vayena E, Mastroiann A, Kahn J. Ethical issues in health research with novel online sources. *Am. J. Public Heath* 102(12), 2225–2230 (2012).

53 El-Emam K, Jonker E, Arbuckle L, Malin B. A systematic review of re-identification attacks on health data. *PLoS* 6(12), e28071 (2011).

54 Secretary's Advisory Committee on Human Research Protections (SACHRP). SACHRP chair letter to HHS Secretary on HIPAA: appendix F. (2004).
www.hhs.gov

55 Cate FH. Protecting privacy in health research: The limits of individual choice. *Cal. Law Rev.* 98(6), 1765–1803 (2010).

56 Peddicord D, Waldo AB, Boutin M, Grande T, Gutierrez L, Jr. A proposal to protect privacy of health information while accelerating comparative effectiveness research. *Health Aff. (Millwood)* 29(11), 2082–2090 (2010).

•• **Analysis of current privacy laws and practices that limit researcher data access; proposes new framework to allow access to information-based research under a research safe harbor.**

57 Kass NE, Natowicz MR, Sara Chandros H, Faden RR, Plantinga L, Gostin LO *et al.* The use of medical records in research: what do patients want? *J. Law Med. Ethics* 31(3), 429–433 (2003).

58 Marti A. Addressing the challenges of HIPAA compliance in research, part II. *J. Health Care Compliance* 7(2), 59–64 (2005).

59 Crook MA. The risks of absolute medical confidentiality. *Sci. Eng. Ethics.* 19(1), 107–122 (2013).

60 Chan EM, Lam PE, Mitchell JC. Understanding the challenges with medical data segmentation for privacy. Stanford University (2013).
www-cs-students.stanford.edu

61 Angst CM. Protect My Privacy or Support the Common-Good? Ethical Questions About Electronic Health Information Exchanges. *Journal of Business Ethics* 90(S2), 169–178 (2010).

62 Hodge JG, Gostin LO, Jacobson PD. Legal issues concerning electronic health information: privacy, quality, and liability. *JAMA* 282(15), 1466–1471 (1999).

63 Duke Center for Health Informatics. Secondary Data Use (n.d.).
www.dchi.duke.edu

64 El-Emam K, Arbuckle L, Koru G, Eze B, Gaudette L, Neri E *et al.* De-identification methods for open health data: the case of the Heritage Health Prize claims dataset. *J. Med. Internet Res.* 14(1), e33 (2012).

65 Hoffman S, Podgurski A. Balancing privacy, autonomy, and scientific needs in electronic health records research. *Southern Methodist University Law Review* 65(85), 69 (2012).

66 Diamond CC, Mostashari F, Shirky C. Collecting and sharing data for population health: A new paradigm. *Health Aff. (Millwood)* 28(2), 454–466 (2009).

67 Croll A, Voytek B. Big Data is Our Generation's Civil Rights Issue, and We Don't Know It. In: *Big Data Now* (*2012 Edition*). O'Reilly Media, Inc., Sebastopol, CA (2012).

68 Rodwin MA. Patient Data: Property, Privacy & the Public Interest. *Am. J. Law Med.* 36(4), 586–618 (2010).

69 Miller FG. Research on medical records without informed consent. *J. Law Med. Ethics* 36(3), 560 (2008).