Am J Transl Res 2016;8(3):1560-1580
www.ajtr.org /ISSN:1943-8141/AJTR0020863

Original Article

Privacy and security in the era of digital health: what should translational researchers know and do about it?

Barbara L Filkins[1], Ju Young Kim[2,8], Bruce Roberts[3], Winston Armstrong[4], Mark A Miller[4], Michael L Hultner[5], Anthony P Castillo[6], Jean-Christophe Ducom[7], Eric J Topol[2,7], Steven R Steinhubl[2,7]

[1]Syntax2Semantics LLC, USA; [2]Scripps Translational Science Institute, USA; [3]Cyber Security Institute of San Diego, USA; [4]San Diego Supercomputer Center, USA; [5]Lockheed Martin Health and Life Sciences, USA; [6]USDN Inc., USA; [7]The Scripps Research Institute, USA; [8]Seoul National University Bundang Hospital, Korea

Received November 30, 2015; Accepted February 19, 2016; Epub March 15, 2016; Published March 30, 2016

Abstract: The rapid growth in the availability and incorporation of digital technologies in almost every aspect of our lives creates extraordinary opportunities but brings with it unique challenges. This is especially true for the translational researcher, whose work has been markedly enhanced through the capabilities of big data aggregation and analytics, wireless sensors, online study enrollment, mobile engagement, and much more. At the same time each of these tools brings distinctive security and privacy issues that most translational researchers are inadequately prepared to deal with despite accepting overall responsibility for them. For the researcher, the solution for addressing these challenges is both simple and complex. Cyber-situational awareness is no longer a luxury-it is fundamental in combating both the elite and highly organized adversaries on the Internet as well as taking proactive steps to avoid a careless turn down the wrong digital dark alley. The researcher, now responsible for elements that may/may not be beyond his or her direct control, needs an additional level of cyber literacy to understand the responsibilities imposed on them as data owner. Responsibility lies with knowing what you can do about the things you can control and those you can't. The objective of this paper is to describe the data privacy and security concerns that translational researchers need to be aware of, and discuss the tools and techniques available to them to help minimize that risk.

Keywords: Digital health, cyber security, privacy, confidentiality, translational research

Introduction

Researchers, practitioners and consumers alike are increasingly embracing mobile technology, cloud computing, broadband access, and wearable devices-effectively removing the traditional perimeter defenses around sensitive data. As a result, security measures to protect this information must be initiated at the source and maintained until the information reaches its intended endpoint-whether it be sensors, apps, research databases, websites, electronic health records (EHR), a patient, or a general population. Health care providers and researchers are now working with a digital ecosystem of tools, enabled by the Internet, loosely coupled and easy to deploy, that provides powerful capabilities for care delivery and analysis, but along with this comes formidable challenges in protecting the privacy and security of individuals and their information. The objective of this paper is to describe the data privacy and security concerns that translational researchers need to be aware of, and discuss the tools and techniques available to them to help minimize that risk.

Evolution of digital health

Over the past 40 years, monolithic information technology (IT) systems as well as brick and mortar perimeter defenses of potentially sensitive health data have given way to loosely coupled ecosystems. Although there were multiple earlier efforts, adoption of Veterans Health Information Systems and Technology Architecture (VisTA) in 1980 is often recognized as the start of what is now referred to as digital medicine with the embracing of that first generation EHR. Another milestone occurred in 2000 with the successful implementation of the Computerized Patient Record System (CPRS), a graphical user

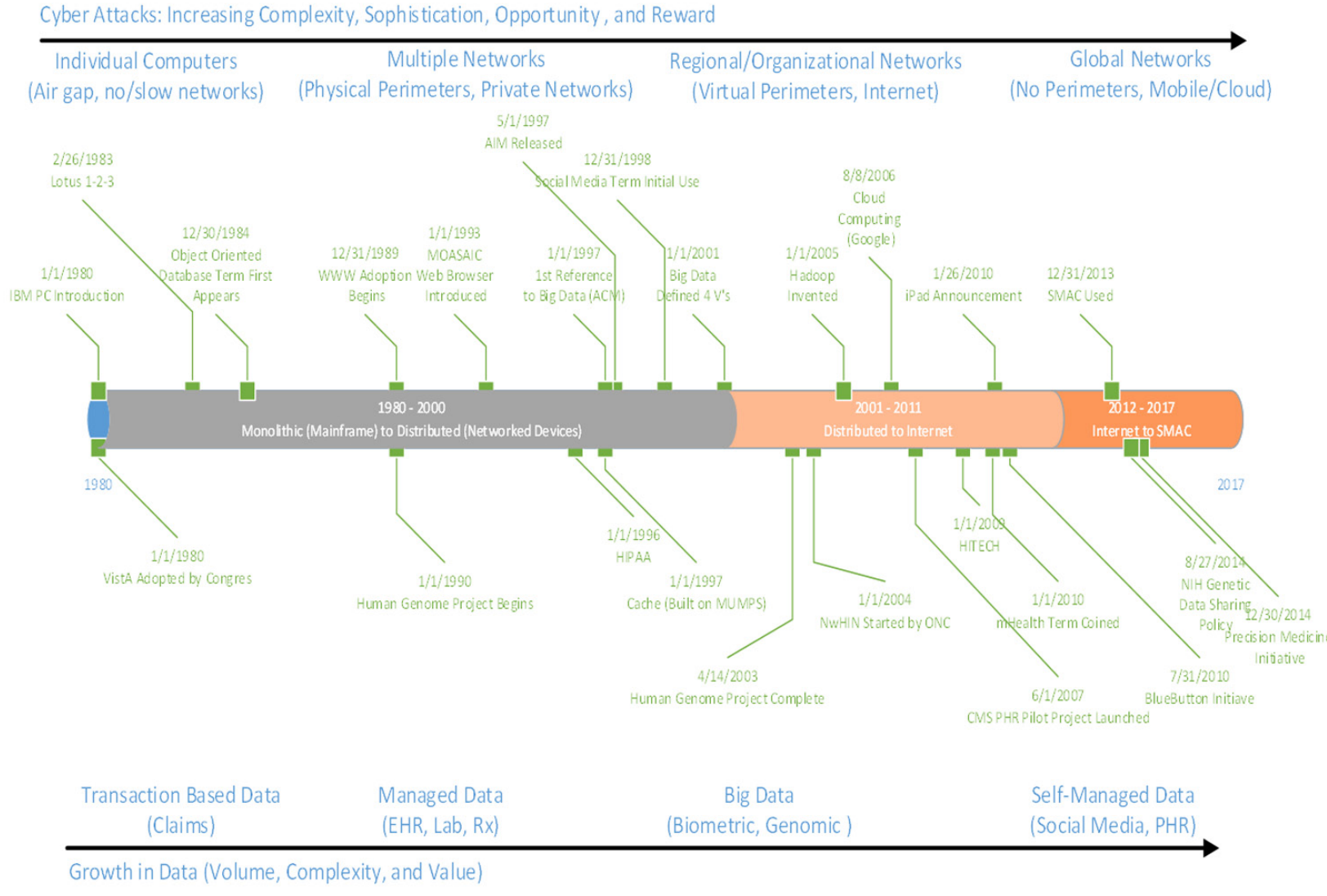

Figure 1. Evolution of clinical technology.

interface to VisTA that allowed providers to review and edit a patient's EHR, marking the true start of medical informatics as a field. The introduction of the first iPhone in 2007, with its potential for ubiquitous mobile computing and connectivity, marked the beginning of an ecosystem allowing for real world tracking and collection of clinical and research quality personal health information through mobile devices.

Today, the convergence and mutual reinforcement of social, mobile, analytics, and cloud (SMAC) reflect a world where consumers are technology-immersed; the Internet of Things (IoT) is extending digital monitoring possibilities as "things" (e.g. cars, homes, work environments, wearables, etc.) become smarter, ubiquitous and autonomous [1]. This trend towards an ecosystem of loosely connected devices means that security safeguards must become data-centric-embedded with the data itself and not necessarily dependent on the infrastructure in which it is found (**Figure 1**).

The quantity of available data specific to an individual has also exploded. Health care data, coupled with an individual's financial profile, social behavior patterns, and, in a growing number of cases, genomic information, is becoming ever more valuable-whether to legitimate commercial entities interested in targeted marketing, individuals seeking to illicitly obtain services at the expense of another, or to criminals profiting from selling this packaged identity or using it to commit fraud worth millions. All this information can potentially be readily accessed globally through the Internet from all types of devices, from traditional desktops and smart phones to wearables.

Technology is moving rapidly, but the risks are moving just as fast. The ability to assure confidentiality, integrity, access and non-repudiation (identity authenticity) of information offers unique opportunities and risks. As the perimeter defenses have dropped away, cyber threats have become more sophisticated, persistent, and impactful. But at the same time, it is important to recognize is that attacks are not necessarily more complex, but it is the shear number of low level, easier to see, targeting users that increase vulnerability. Traditional security measures, like strong complex passwords (when used), are simply becoming insufficient for the modern connected environment.

Why the concern?

As of 2015, hacking has become the leading cause of breaches reported by CMS [2]. Motivation for attack can range from financial gain: intent to commit fraud, profiting from selling packaged identities; unauthorized hospital clerk (insider) idly viewing the health record of a movie star; to just the challenge of defeating a security system. Risks stem from several sources: opportunity, increased motivation, and a lack of understanding by the health care community in the use of technology.

Opportunity

The cybersecurity community has a mantra- "It's not if you will be attacked but when." Others in the hacking community feel even this is too soft and instead assert that every IP address on the internet has already been attacked from the moment it had any connectivity to a public IP address. Research has brought to light the proliferation of attacks that happen on any routable public IP address [3].

Today's game-changing technologies-utilization of social media, mobile devices, the Internet of Things, and cloud-computing-present an increasing number of access points. Security strength varies widely. Available data sources about an individual can easily be stitched together to exploit high value items like financial assets and medical identity. Technology gets more complex and more complex attacks emerge from the simple viruses of yesterday, to multifaceted malware that expose applications, systems and networks on multiple levels for information gain or destructive attacks. Though, not to be forgotten, is that many attacks still successfully use simplistic 1980s-style methods like default passwords to achieve their malicious intent.

Many legacy healthcare and research organizations have not yet fully adapted to this accelerated rate of change, whereas a number of forward-thinking organizations are beginning to embrace the use of lightweight configurable systems that displace or augment legacy IT. Adoption of digital technologies has outpaced the implementation of appropriate safeguards for privacy and security, as well as the ability to anticipate and respond to potential threats. The explosive growth of connected devices that

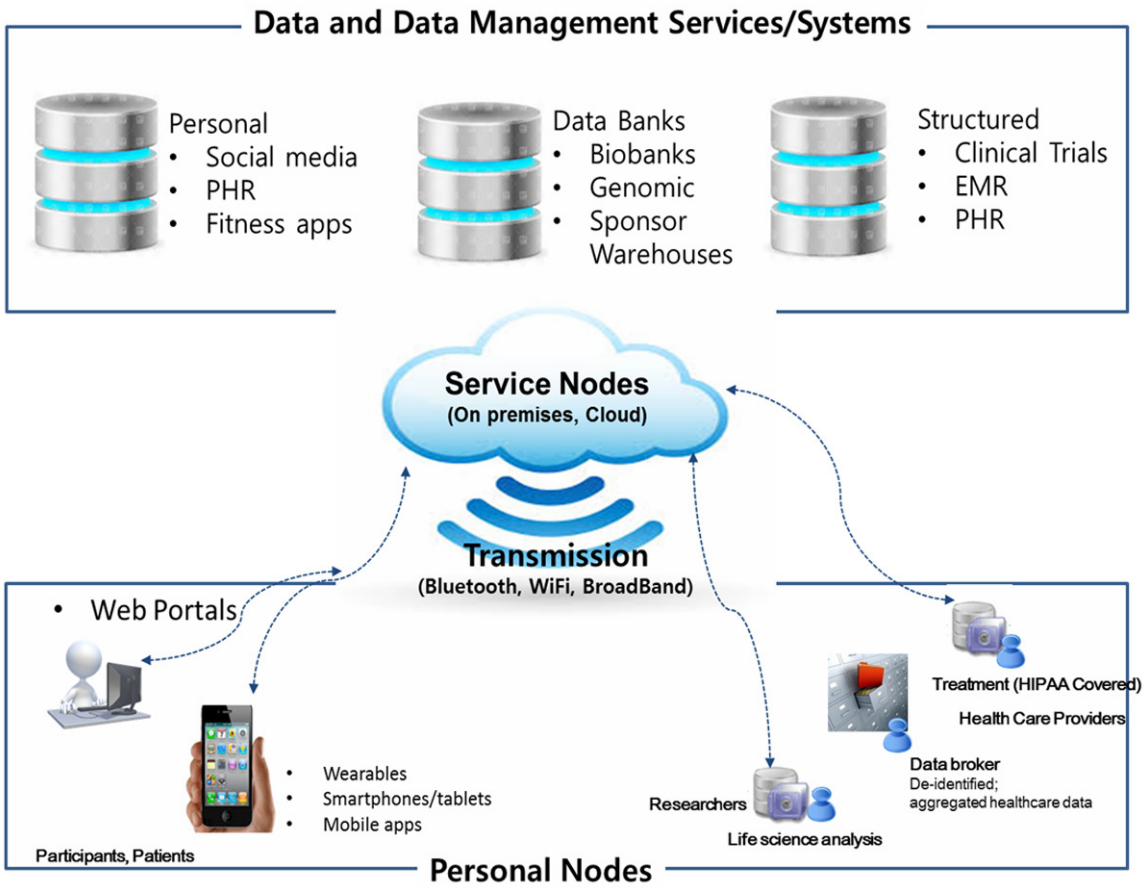

Figure 2. The connected world of translational research in medicine.

contain medical information and their integration into backend systems that contain additional critical data has also opened the door to new compromises. Not unexpectedly, health care in 2014 had the largest increase in the number of potential attack surfaces of any industry [4].

Motivation for attack does not have to involve nefarious intent, cyber warfare, financial gain, or even retaliation against a specific individual. In 2012, Michael Honan, a correspondent for Wired, suffered a major attack against his online identity, compromising almost all and destroying most of his digital assets. Motivation? The hacker wanted Mike's twitter handle [5].

However, the strongest motivator of the mounting attacks on healthcare is the financial value of information. Healthcare has the highest per capita cost for a stolen record ($363) of any industry [6]. According to the 2015 Ponemon report on security of healthcare data [7], the average cost of a data breach for healthcare organization is estimated at more than $2.1 million and criminal attacks are the number one cause of data breaches in health care, up 125 percent compared to five years ago. Medical identity theft is on the rise, with an increase of 21.7% from 2013 to 2014 [8]. The value per set of stolen credentials can vary from $50 to well over a thousand US dollars, depending on how complete the set is. Stolen medical identities can be used for anything from a victim's relative attempting to gain coverage, to massive deception and fraud perpetrated by organized crime.

Fraud is a complex problem that has cost the United States government $6 billion for Medicare alone in the last two years [9]. Prescription drug programs are an especially hot target. Recently, a number of Miami pharmacy owners were charged with paying Medicare beneficiaries for their personal identification numbers, which they used to file fraudulent

Table 1. Inherent risks in the connected world of translational sciences research

Layer	Uses	Risks/Vectors
Personal node (Individual)	Users: Participants and patients Uses: Recruitment, communications, data collection Devices: wearables, smartphones, tablets, apps on mobile devices owned by individual	Risk: Compromise of sensitive data, theft of identification, unauthorized access to study results or patient information Vectors: Loss of device, social engineering to gain control of device, malware installed that results in loss of control by device owner
Personal node (Entities to include healthcare organizations, research institutions)	Users: Researchers and Clinicians Uses: Research and analysis, collaboration point between researcher and clinical provider Devices: Desktops, laptops, tablets, mobile devices owned by organization	Risk: Compromise of sensitive data, theft of identification, unauthorized access to study results or patient information, though potentially lower than devices managed by organization Vectors: Loss of device, social engineering to gain control of device, malware installed that results in loss of control by device owner
Transmission and communication protocols	Uses: Transmit data between endpoints (participants, researches, clinicians, administrators, service endpoints) Protocols: Bluetooth, Wi-Fi, Broadband	Risks: Interception of sensitive data in transit, undetected changes in data due to transmission, denial of service Vectors: Insecure transmission, lack of, or compromised encryption
Services node (including data management services and platforms)	Uses: Web-based applications for email, messaging, file storage Line of business applications to include electronic health records (EHR), personal health record (PHR), web portal, research databases, analytics tools, survey management. App store for mobile applications System/Interfaces: On premises systems in enterprise data center, cloud provider	Risk: Compromise of sensitive data, theft of identification, compromise or theft or intellectual property (such as metadata, research protocols and preliminary results), unauthorized access to study results or patient information, falsification of results, data loss/destruction Vectors: Insider threat (negligent or intentional), lack of proper cloud security, lack of proper IT security, insecure access for reporting study results (i.e., protection against bots), lack of timely audit or awareness

claims for drugs that were never dispensed. These individuals worked with a clinic owner, who forged and altered prescriptions and sold them to the pharmacies. The defendants fraudulently billed Medicare $21.2 million [10].

Understanding

Our personal lives are built around a certain level of physical distrust-we lock our homes, we hide our wallets and purses, we avoid dark alleys. Yet, for the most part, we trust the Internet with less concern for safety, revealing personal information to the Web that we otherwise would not share, reassured by the notion that simple password protection is adequate to protect our sensitive information.

The increased use of applications that rely on cloud computing, when coupled with the rise in mobile and the use of personal devices for work, allows sensitive data to flow outside the traditional enterprise firewalls. In fact, companies wanting to benefit from the cloud's flexibility and the productivity of "bring your own device" or "BYOD" have created new systems and procedures that allow their employees to reach corporate data remotely, giving hackers greater attack surfaces with which to work. Attackers commonly leverage social media to create targeted, convincing user mode attacks like spear phishing to steal employee credentials and use them to access company data. Since employees often have more access to sensitive data than they actually need, companies end up placing their data at risk unnecessarily. This means that hackers can now also use the same pathways that company employees use to access sensitive company data. All they need is employee credentials.

Thankfully, many of the risks to security stem from known vulnerabilities, and improving individual awareness that those vulnerabilities exist can minimize the risks to security. The problem is, many individuals tend to embrace myths and misconceptions that enable attack. Working through some of these myths can enhance privacy and security and minimize the risk of attack.

The connected world of translational research

Today's clinical and research environments are evolving towards a reference architecture like that shown in **Figure 2**. Data at varying levels of structure and complexity are collected and held across various platforms-from on-premises systems to data lakes in the cloud-and are accessible from anywhere, at any time through a variety of communication channels (email, SMS) and transmission protocols (Bluetooth, Wi-Fi, broadband).

An individual is commonly found at one end of a communication channel: a consumer looking at

her wearable data on a smartphone; a researcher reviewing remotely-collected patient data on his tablet, or a clinician using his/her desktop to review a patient's health record. The smartphone, the tablet, and the desktop (accounting for both hardware and software on the device) are all considered personal computational nodes, whether managed solely by the individual (such as the case with a personal smartphone), an enterprise (such as the clinician's desktop), or both (as might be the case where the researcher owns the tablet but enters into an agreement with their organization for business use).

At the other end of a communications channel, the 'service' node represents access to specific computational technology, such as file storage, data management platforms, analytics tools, or other web-based applications. These service nodes can be hosted "on premises" (an organization's data center), in the cloud, or as a seamless hybrid of the two. Regardless, traditional security safeguards-the physical boundaries of the data center, perimeter firewalls, user names and passwords-are no longer effective as technology pushes the need for security closer to the actual data.

This connected world offers many advantages in terms of flexibility, elasticity, outreach, and cost but-as will be discussed-the cyber landscape is fraught with potential risks. **Table 1** helps begin the journey by summarizing each layer, its use(s) in translational medicine, and the attack surfaces, defined by the potential risks and attack vectors to be discussed.

The personal node and potential security issues

The personal node interfaces directly with an individual-whether the researcher, a member of the research team, a collaborator, or a participants-and can span mobile devices, fixed assets (desktops) as well as the services and the applications or "apps" on each. As the potential security issues in most desktop environments are relatively well known, we will largely turn our attention in this section to the mobile ecosystem.

Security in mobile devices and applications

Mobile devices that fall outside enterprise management potentially constitute the weakest link in a security infrastructure: not only is the node outside of the system administrators direct control, it is also being managed and operated by a human being within their environment and is subject to the individual's own understandings and possible misconceptions.

Mobile devices are easily lost or stolen. When this happens to a device used for work activities, enterprise data or credentials are put at risk, along with personal information, especially since most of these devices are not adequately protected. Of the 4.5 million smartphones that were lost or stolen in 2013, only 36% were protected with a PIN, only 29% had their data backed up, only 7% protected data with a strong password or some other stronger security protection and only 8% featured software that enabled the owner or an administrator to remotely wipe the contents of the device [11].

Second, an individual may be completely unaware of what they 'authorized' when they install an app on a device, for example: what processes in the device are being accessed, whether private information is being sent to a third party (potentially in violation of any licensing or privacy agreements if one even exists), and whether proper security measures are in force. A 2013 analysis of mobile medical, health, and fitness apps revealed disturbing findings: privacy policies were completely lacking for 40% of paid apps; 40% of the apps collect high risk data (including financial information, full name, health information, geo-location, date of birth and zip code); roughly only 50% of apps encrypted personally identifiable information (PII) being sent over the Internet; 83% of both free mobile health and fitness apps store data locally on the device without encryption [12]. A similar study by the Federal Trade Commission of twelve mobile health and fitness apps revealed that user data was disseminated to 76 third parties; the information included usernames, proper names, email addresses, data on exercise and diet habits, medical symptom searches, zip codes, geolocation and gender [13].

Finally, mobile devices have inherent vulnerabilities-the operating system (e.g., iOS, Android), utilities provided by the carrier, and legitimate third party apps-that can account for data loss and leakage. There were 37,246 health, fitness

Table 2. Guidelines to Protect Personal Mobile Devices

• Be vigilant when granting mobile app permissions, especially those that might access sensitive mobile data, GPS location, or the devices camera and microphone. Consider the permissions that an app requests before installation and evaluate whether the exposure is worth the convenience.
• Enable built-in locking features-PINS, passwords, or the processes for two-factor authentication that Google and Apple are now utilizing that further verifies a user's identity by sending a unique, one-time PIN code to their phone before allowing the individual to fully login with their name and password.
• Protect the device with security software, similar in nature to that used on a desktop or laptop such as personal firewalls, spam filters, anti-virus and anti-spyware tools, but designed for the mobile environment, to keep the device free from malware, spyware, and the threat of infection. An Internet search for "top mobile security software" plus the year reveals the depth of the market. Seek advice in selecting and installing a security product.
• Know how to use device location services and remote wipe-whether provided by the carrier, the organization, or a third party (LoJack). Verify where the service can be reliably executed from most user locations. Apple owners are encouraged to turn on "Find My iPhone," Google Android users should take advantage of the Factory Reset Protection feature.
• Backup important information on the device to a secure location, such as a personal or work computer or on-line service.
• Encrypt data, especially PII, stored on the mobile device. For Android, select a trusted encryption application. Consider encrypting the entire device if it's running Android Version 4.0 or greater with the built-in "encrypt your phone" functionality. iOS devices running iOS 4.0 or higher with a passcode set will automatically encrypt all the data, although this method can be circumvented.
• Be aware how the device may be communicating. Turn off Bluetooth if not needed. Avoid using open Wi-Fi. Make sure that the device isn't set to automatically connect.
• Download mobile apps only from trusted sources and keep credentials to app stores like iTunes or Google PlayStore confidential and secure.

and medical related apps on the iTunes or Android market as of 2013 [14, 15] with significant projected growth rate over the next five years. Estimates are that 90% of Android sensitive medical/healthcare apps have been hacked, 22% of these were FDA approved [16].

Attackers are posed to leverage these vulnerabilities. Although the Verizon 2015 DBIR report downplays the impact of mobile malware [17], the continued growth in the number of health-related mobile apps and their corresponding potential vulnerabilities should not be discounted. A widespread vulnerability in the Android OS, "Android Installer Hijacking," was publically disclosed March 2015 and is estimated to impact almost 50% of all current Android users. The exploit, currently only affecting applications downloaded from third-party app stores, allows an attacker to modify or replace a normally benign Android app with malware, all without the knowledge of the user, allowing the malicious application to gain full access to a device, including usernames, passwords, and sensitive data [18].

As with other advanced computing equipment, user awareness is key to safeguarding the mobile device, both electronically to protect the identity and data it carries and physically to secure the device if lost or stolen. Also, being aware of common signs of infection including abnormal issues with performance, dropped calls and disruptions, abnormal usage patterns such as a devices sending SMS (text) messages to premium-rated numbers or unexplained data plan spikes; unknown apps appearing as installed. **Table 2** provides guidelines that researchers should follow in using mobile devices as well as advice to be provided to study participants in order to protect the personal information collected in a study.

Security for a research application

Privacy, data security, and informed consent are integrally bound together in the research environment, both from the standpoints of protection and compliance. The growing use of mobile devices for recruitment of and communication with study participants, as well as subsequent collection of patient-reported data brings new emphasis on these elements. A researcher's IRB or Ethics Committee establishes the requirements that must be met, but provides little guidance as to actually accomplish them. A software framework, such as Apple's ResearchKit, can aid in building a mobile research app, but still does not address data management, privacy and security controls. The researcher is still responsible for implementing protections for data transmission, storage, and use after collection.

Security is essential to privacy. For any app utilized in a study, the researcher needs to understand what sensitive data will be stored on the mobile device, how and where that sensitive

data will be transmitted from the device, and what procedures or actions reduce the risk of compromise. A researcher should confirm through the assurance of their IT team that:

- Both the app and related data can be completely wiped from a device when the participant leaves the study and assurance can be provided to the participant.

- All secondary agreements (e.g., commercial app, app components like run-time libraries, standard services provided by the carrier) that collect and send data to third parties have been identified and evaluated for risk.

- Credentials are used to control access to the app and its data. At a minimum, this should be a PIN or biometric ID, with two-factor authentication strongly recommended.

- The highest level of file/data protection possible is enabled. For example, files/data stored by the app are automatically encrypted whenever the device is locked.

- All sensitive data on the device collected by the app is deleted as soon as possible (i.e., in accordance with the study's published retention policy.)

- The latest version of SSL/TLS is being used for all (no exceptions) communications between the app and other systems, including user authentication and the transfer of sensitive information. An additional policy may be to encrypt sensitive data before transmission, even using SSL/TLS.

Security considerations need to integrate with participant access and use of the app, activities that integrate with privacy and informed consent requirements. A researcher should be involved in the development of policies and procedures for participant access or download of a research app. Their input is important in determining where the participant will acquire the app; whether through a recognized app store (e.g., Apple or Google), another third party site, or directly from the study site. In addition, in development of an explicit privacy policy for every app that collects personal data (Note: This is required for ResearchKit apps posted to Apple's iOS App Store), and to make sure that all related study policies are easily accessible and understandable before the participant downloads the app. Also, it is their responsibility to make available a non-technical explanation, such as a FAQ on the study website, as to which permissions in a participant's mobile device the app requires access, including what the participant can decline and still have the app work effectively. Finally, the researcher must assure that they are able to obtain a clear opt-in from a participant before accessing location data in the mobile device. Precise geolocation information is increasingly considered sensitive information Research from Nanjing University demonstrated how accelerometer-based movements can be easily traced and users identified as a result [19]. Similarly, participants should be notified that geo-tagging may occur if the app takes photos and/or videos as the device may embed metadata that can reveal location coordinates where the photo or video was taken.

A final consideration in the design of a research app is the use of electronic signature. If a study requires written informed consent, the use of electronic, including digital, signatures is permitted. The FDA under 21 CFR Part 11 does not have a preference for electronic or digital signatures, both being valid if regulatory requirements and expectations are satisfied. Researchers should be aware that the two types are not interchangeable.

- An electronic signature is the legally binding equivalent of an individual's handwritten signature; it can be as basic as a typed name, a credential such as a password, or a digitized image of the handwritten signature. Its use is problematic in maintaining integrity and security as nothing binds the signature to the actual record.

- A digital signature is technology that uses cryptographic methods and critical metadata pertaining to an electronic signature to create an electronic "fingerprint" that ensures signer authenticity, provides accountability, secures sensitive data, and guards against tampering.

Research Kit, as a representative framework, does not include digital signature support; the study design must address how electronic signature should be implemented. A researcher should be aware, however, that integrating electronic and digital signatures authenticates

the individual signing the informed consent, ensures the source file, which preserves this record, is secure and verifiable, and potentially could also be used as a method to secure participant data on their mobile device.

Security in email and messaging services

Email and instant messaging protocols were not originally designed with privacy or security in mind. Convenience is the major driver. For the majority of messaging platforms, almost all information is sent in clear text and the validation of the sender and recipient is not mandatory. Given the myriad ways individuals send, receive, store, and use messaging services, trying to fully secure messaging with a technical solution alone is virtually impossible.

Despite these shortcomings, email remains the most ubiquitous method of communication on the Internet today. In 2015, the number of emails sent and received per day total over 205 billion, with an average number of business emails sent and received per user per day totaling over 120 [20]. Health care accounts for part of this traffic. An estimated 50% of family physicians and 67% of other specialists e-mail their colleagues for clinical purposes [21]. About 15% of physicians communicate with their patients using email. And its use, along with instant messaging and social networking, continues to grow as all modalities are inextricably woven into modern lifestyles, both personal and professional.

The key drivers behind secure electronic messaging include: confidentiality (the message is private, cannot be read by other than the intended recipient), integrity (the message hasn't been tampered with in transmission), and authenticity (the message comes from the person who sent it). Security shortcomings can be balanced by safe practices that can be implemented by both individuals and organizations such as:

- Verify the identity of a recipient before sending an electronic message, especially one that may contain sensitive information.

- Encrypt the message payload and any attachments, realizing that parts of the message may not be encrypted. For example, the subject line on most emails are still sent clear text. Better yet, develop preset templates for communication with study participants.

- Send the password to decrypt the message payload/attachments in a separate email.

- Scan all inbound and outbound messages for malware, including deep inspection of attachments. Considering disabling attachments unless absolutely needed.

- Log all email and/or text traffic in accordance with regulations and retain for an appropriate length of time (i.e., six years if HIPAA-regulated).

A researcher can also work with their IT team to implement capabilities specific to a given study, such as establishing study-specific email or instant messaging accounts for participants. Another method is to notify a participant via the email or instant messaging account they enrolled into the study and that a message is waiting for them on the study's secure portal site. The participant would then login the site with a secure key or credentials to obtain information. Consider implementing Sender Policy Framework (SPF), a simple email-validation system designed to detect email spoofing, in the study email used by researchers and staff

Another, potentially greater concern around messaging is around the use of social engineering to compromise an individual. 'Phishing', a word play on 'fishing', uses communication methods, like email and instant messages, to trick individuals into divulging sensitive information directly or directing them to a malicious Web site where malware will be downloaded to their device, resulting in further compromise of other devices, applications, or systems to which now infected device connects.

There are many notable incidents that involve phishing, healthcare organization among them. St. Vincent Medical Group, Inc. said in a statement on its website that approximately 760 patients potentially had their PHI exposed after an employee's username and password was compromised because of an email phishing scam [22]. The Texas-based Seton Healthcare Family, part of Ascension health system, determined that a December 2014 phishing attack exposed sensitive data (including demographics, medical record numbers, Social Security

Table 3. How not to fall for a phishing attack

• Check the email addresses. The email may appear to come from a legitimate organization but the "FROM" address is a personal email such as @gmail.com or @hotmail.com, it is likely an attack.
• Be suspicious of email messages that: come from unfamiliar senders; make unsolicited offers, address topics unrelated to your personal interests or ask to confirm private information over the Internet through a link in the email; aren't personalized; that ask you to call a phone number to update your account information; have obvious grammar or spelling mistakes if from a major business.
• Do not click on links, download files or open attachments in emails from unknown senders. It is best to open attachments only when you are expecting them and know what they contain, even if you know the sender.
• Know how to check the true destination of a link that appears in an email. This shows where you would actually go if you clicked on the link. If this is different than what is shown in the email, chances are this is an indication of an attack.
• Beware of links in emails that ask for personal information, even if the email appears to come from an enterprise you do business with. Phishing web sites often copy the entire look of a legitimate web site, making it appear authentic.

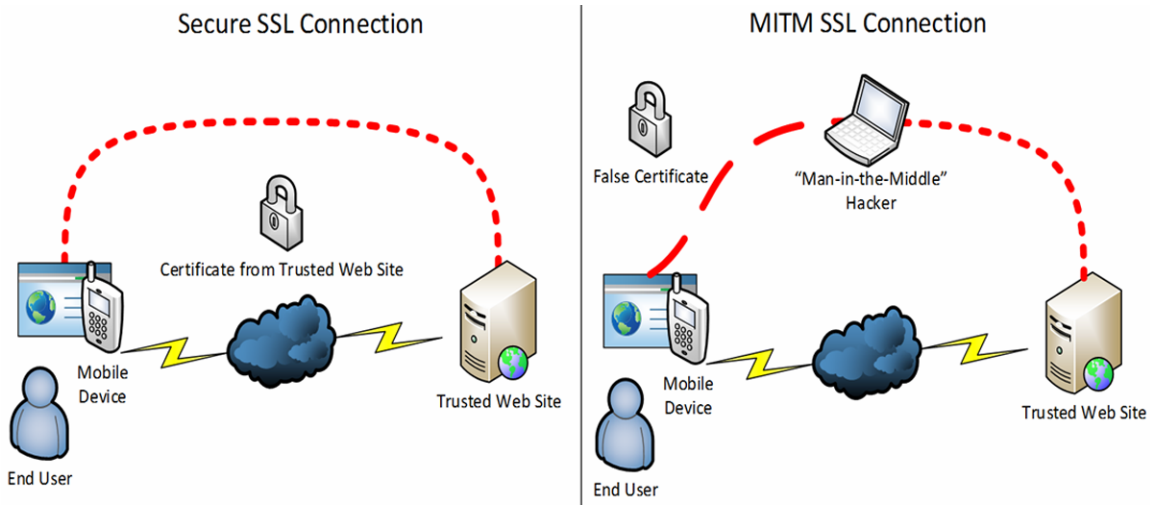

Figure 3. SSL connection, secure and insecure.

numbers, clinical data and insurance information) for 39,000 patients [23].

Malicious emails (phishing) are used in 95% of successful data breaches (Mandiant) and accounts for approximately 80% of malware entering organizations. It has been reported that 23% of users will open a phishing email [17]. Keeping safe and secure as possible depends on understanding and awareness. **Table 3** offers advice to avoid falling for a phishing attack.

Security in internet communications

Two standard protocols-Secure Socket Layer (SSL) and Transport Layer Security (TLS)-are used to secure most Internet communications including email (IMAP, POP3), database, and secure websites connections (HTTPS). In each case, TLS/SSL is combined with the additional protocol to provide secure authentication and session confidentiality. The connection between a client (the Web browser on the smartphone, tablet, or desktop) and a server (the destination website) is being secured by a "behind the scenes" handshake that depends on a digital certificate; a document signed by a trusted authority that the owner of the website is trustworthy.

Compromised certificates can undermine the security of Internet communications based on SSL session. A "man in the middle" (MITM) attack, allows an encrypted session to be easily eavesdropped upon by a third party as shown in **Figure 3**. A fake certificate that impersonates the legitimate certificate from the trusted source in every way except not being signed by the trusted authority is sent to the end user's device. This allows the attacker to send the online traffic to an intermediate site, decrypt it, manipulate it, and then re-encrypt and forward it to the study site, leaving the end user and the researcher unaware that the attacker may have

Table 4. Bluetooth security tactics for researchers

• Disable Bluetooth when not in use. Consider disabling Bluetooth devices in closed environments such as aboard a commercial aircraft.
• Disable unused services, like enable the wireless headset but disable file transfer.
• Place Bluetooth device in non-discoverable mode when not pairing with another device.
• Never accept files or messages over Bluetooth from untrusted devices, use a second factor of verification before accepting a connection.
• Never accept pairing from untrusted or unknown devices. Pairing is permanent until deleted and pairing with an unknown device can provide access to all the services enabled on the mobile device.
• Change PINs semi-frequently by un-pairing the devices, changing the PIN, and re-pairing.
• Use a secondary form of authentication for access to apps, such as a username and password or PIN.
• Be careful about Bluetooth pairings when on the go. Pairing a smartphone to a rental car may leave data behind after the connection is terminated.

captured authentication credentials or violated integrity of transmitted data [24].

In the case of HTTP, the padlock icon visible when connecting to a secure website server reassures the user that the connection between their device and the website is trusted, encrypted and secure. If the user's browser issues a warning, however, this can mean there is an error with the web site's certificate, such as the name to which the certificate is registered does not match the site name or the certificate has expired. Any uncertainty as to the validity of the certificate or security of that site should be a red flag to the user not to submit sensitive information and to confirm the validity of certificate and web site involved.

The researcher needs to ensure that certificates used on any study Web sites or other user-facing interfaces are valid and that they remain so. SSL/TLS certificates have become an attractive target for the underground economy that seeks to monetize data stolen from compromised hosts. In fact, the resale of stolen but valid digital certificates may be the next global black market as they can undermine trust in a variety of ways, from access to business websites to passing off malware as legitimate executables and scripts through code signing [25].

Transmission protocols and potential security issues

Transmission mediums and protocols-Bluetooth, WiFi, and broadband (cellular) services can be thought of as the binding glue connecting personal with service nodes, whether the latter is located in the cloud or the enterprise data center. Insecurities can compromise sensitive data. "Proximity-based hacking" is a new form of attack that compromises the NFC (Near Field Communications) chip now being embedded in most smartphones available in 2015. This chip is used for contactless payment (in lieu of a credit card) or data collection (from another NFC-enabled device). The technology will not be discussed in this paper.

Security in Bluetooth: Released in 1998, Bluetooth is a short-range (1 to 100 meters), low-power wireless communication technology, commonly integrated into modern devices for interface with wireless printers, headsets, and automobiles as well as transfer information between two local devices. Updates to this protocol have been designed to consume less energy, spurring innovation in the mHealth market with the ability to link apps with various sensors, both embedded in and external to most mobile devices including wearables to capture rich datasets-video recordings, accelerometer data, and physical characteristics-that attract researchers and clinicians alike.

Bluetooth is vulnerable to unauthorized direct data access, and eavesdropping on conversations or video. (See **Table 4** for related security tactics). It has also been used to spread malware, but the short range over which Bluetooth operates has tended to hinder the effectiveness of this method [26]. However, the short range of Bluetooth can be extended to several kilometers by attaching a high-gain antenna to a standard Bluetooth radio, so attacks can be carried out at a much greater distance. In 2007, a researcher demonstrated how to eavesdrop on conversations in his neighborhood Starbucks, underscoring how easy this protocol is to compromise [27].

Security in WiFi networks: Public WiFi hotspots-cafes, restaurants, hotels, libraries, public

Table 5. Tactics for secure use of Wi-Fi hotspots

• Make sure all mobile devices have antivirus, anti-malware and a personal firewall all installed and updated.
Install a privacy screen to avoid "shoulder surfing" where an attacker might "look over your shoulder" to gather info or passwords as you type
• Make sure that the connection and session is encrypted. At a minimum, ensure that https is present in a web address before accessing a secure site (i.e., webmail, social media, or any site that requires a login). Make sure that this connection stays encrypted for the entire online session. Some websites encrypt the log-in and then return the user to an unsecured, vulnerable session.
• Do not use unsecure, unencrypted methods to transfer data, like FTP to upload data.
• Don't let the mobile device automatically connect to an open WiFi source. Many devices can be set to notify you when an open network is available and have you determine whether to connect. Don't connect to an "open" network, even the one at Starbucks, if you are working on a project and have no other means of encrypting the connection (like VPN).
• Use a virtual private network (VPN) to protect your data. A VPN adds a layer of encryption and security that is valuable when using any unknown or open connection.
• Limit exposure when using a hotspot by using a personal or a business-issued mobile hot spot configured securely. Many smartphones and tablets have hotspot capability built-in.

places-are all potential open invitations for electronic *eavesdropping.* The danger of open WiFi connectivity was exemplified by a recent experiment where Finn Steglich, a German company, set up a rogue hotspot on the streets of London. Within just 30 minutes, 250 devices had connected to this rogue hotspot, demonstrating the following common concerns around public WiFi [28]:

• Splash pages for WiFi networks that offer Terms and Conditions, a password or other login method, do not make a network safe, especially as people don't read the fine print of the T&Cs and the login method is intended just to gain access to the network, not to really authenticate or protect the user. Here, people accepted a Term and Conditions page that required they give up their first born child or favorite pet in order to be able to use the hotspot!

• Many connections were made automatically without the owner of the devices even knowing.

• Connecting to open "hotspots" makes a user's device visible to other devices on the network. Emails, passwords, and unencrypted instant messages can be easily viewed, unsecured logins to popular websites hijacked. Basically, if a device is visible, it is hackable as the 32 MB of personal data collected during this experiment demonstrated.

In general, most open WiFi hotspots should be considered insecure, even the one at the local Starbucks or aboard a commercial airliner. **Table 5** provides guidelines around the secure use of WiFi hotspots.

Service nodes and potential security issues

Researchers are not normally concerned with the technical management of the IT infrastructure but they are responsible for data management and protection, regardless of where data services are hosted-on premises, in their organization's data center, or in the cloud. As a result, the researcher should better understand how the technical aspects of these systems and applications could impact privacy and security of their data.

Information- or data-centric security is an approach to information security paradigm that emphasizes the security of the information or data itself rather than the security of networks, applications [29]. Meta data needs consideration, as the connections between various data sources can be as or more sensitive than the data upon which such information is based. In helping define a data-centric security focus, the researcher should focus on four key elements, coordinating with their IT team as needed:

• Understand the characteristics of the data, both that collected for the research and that created by the research, to anticipate any unexpected variations in integrity or quality that might flag a possible privacy or security concern.

• Know where that data resides or might reside, whether on the mobile device of a participant, residing in the cloud, or being extracted from a covered entity's EHR, together with the related regulatory requirements around compliance or privacy might be for each source.

• Assess potential risks to the information based on the first two steps and determine

appropriate controls. An essential part of this effort is to identify each authorized user, establish what permissions they have, and document as part of the overall study design so that user authorization can be referred to each time a request for access to data or a corresponding service is made. For example, a patient may have access to their personal information, but not necessarily to the medical data being collected. A researcher will have access to all project data but not necessarily to individually identifiable personal information on a participant. Administrators will probably have very different access requirements.

● Select and apply the most appropriate security practices and controls, both administrative (policies and procedures) and technical (automation) that manage access to the data and are integrated with normal workflows around that data. Methods to protect the data and information, including encryption, masking, and tokenization, need to be evaluated and a determination of where and when to apply them must be made. While no method is perfect, a well-thought out implementation can limit exposure to both the researcher and their institute if a security breach occurs. The researcher should work with their IT team to explore emerging techniques in data science, machine learning, and behavioral analysis to detect malicious behavior that might adversely affect the data being held in a loosely coupled environment.

Security in authentication: pitfalls of passwords

Usernames and passwords are still the most widely used method of secure authentication because they are inexpensive and convenient to implement and use. But for over twenty years, passwords have been a security Achilles heel, due to poor password selection, management, or protection. Back in 1997, the Computer Emergency Response Team (CERT) estimated that about 80 percent of reported security incidents were related to poorly chosen passwords [30]. In 2012, hackers from Eastern Europe exploited a weak password of a system administrator to gain complete access to the Utah Dept. of Technology Service's (DTS) server, breaching 780,000 Medicaid patient health records. And today, what are the most common passwords? Not surprisingly, '123456' and 'password' [31].

Multi-Factor Authentication (MFA) uses more than one authentication factor to logon or process a transaction: "something you know" (account details or passwords), "something you have" (tokens or mobile phones), and "something you are" (biometrics-fingerprints, voice recognition). Various service providers, including Apple and Google, have implemented two-factor authentication, a simpler version of MFA requiring "something you know" (the user password) and "something you have" (a one-time code via text message that is needed to gain access to their account).

Authentication standards are moving to protocols that require no passwords. In one scenario, a participant's mobile app connects to the study's cloud-based patient portal without having a username/password stored on the device, eliminating a potential attack surface that can expose participant's data to attackers. While a detailed discussion is beyond the scope of this paper, researchers should be aware of the following standards: Open Authorization (OAuth), OpenID (single sign-on (SSO)) across various Internet applications; and two created by the Fast Identity Online (FIDO) Alliance. The Universal Authentication Framework (UAF) focuses on authentication without passwords. The Universal Second Factor (U2F) protocol adds a second authentication factor by taking advantage of current technologies available on devices such as fingerprint sensors, cameras (face biometrics), and microphones (voice biometrics).

Security around the data management in cloud

Cloud services are especially attractive for data-oriented projects, given its essential characteristics: on-demand self-service that does not require human interaction at the cloud provider, ubiquitous network access, rapid elasticity in scaling resources up and down, and measured service.

As recently highlighted in an NIH notice, the researcher is also responsible for the security issues in data management system [32]. Cloud computing represents significant unknowns

such as lack of direct control over hardware and software, lack of visibility into audit/system activities, physical locations of data, and impact of different jurisdictions where the data may be held.

The researcher needs to be aware and be part of the evaluation of any cloud service provider and needs to demand transparency in certain aspects. While these suggestions don't require a detailed understanding of the technology, they do require some technical literacy to ensure the proper questions are being asked of the cloud provider that balance privacy, security, and legal requirements with functional needs [33]:

- Privileged user access. This includes both cloud provider staff access to information owned by the researcher as well as the methods available to authenticate, manage and track anyone who might have access or might gain access to the sensitive information and applications.

- Regulatory compliance. Know what scrutiny the cloud service provider provides over its operation, what certifications are necessary, and what the provider's destruction or electronic shredding policies are so the company can have evidence that its data is no longer resident on the provider's systems and, therefore, not subject to attack or e-discovery. Look to whether the cloud provider is FedRAMP accredited or ask what assurance level they have achieved in the Cloud Security Association (CSA) Security, Trust & Assurance Registry (STAR). Both of these require a documented independent assessment of the cloud provider by a third party. Use and expectations of the cloud provider must be adjusted accordingly.

- Data location. Ask the provider to commit to storing and processing data in specific jurisdictions, and whether they will make a contractual commitment to obey local privacy requirements on behalf of their customers.

- Data segregation. Data in the cloud is typically not segregated in a multitenant environment. Known whether your data will be stored on dedicated hardware and, if not, what protective measures the cloud provider takes to ensure that your data will not be compromised in that shared environment. For example, the cloud provider should provide evidence that your data (or the environment housing the data) is encrypted and that approved encryption schemes have been implemented and tested by experienced specialists.

- Recovery. A cloud provider should tell you what will happen to your data and service in case of a disaster or outage. Ask about a full recovery and how long it would take.

- Investigative support, such as breach investigation and forensics. Get terms for visibility and incident response report up front and in writing. Will the provider routinely provide the correct level of logs if requested by a customer?

- Liability and Indemnification. Will the cloud provider stand behind their security and privacy assertions and defend the researcher should a breach occur? Make sure that, if dealing with ePHI subject to HIPAA rules, that the cloud provider will sign a Business Associate Agreement (BAA) compliant with HIPAA rules.

- Termination or long-term viability. If the provider goes out of business or gets acquired, make sure that there is a way to get the data back in a format that is usable.

Security in online participant recruitment

Beyond data protection is the need to maintain the integrity and validity of the collected data. Patient recruitment is increasingly being done on-line, using crowd sourcing or social media to attract and engage individuals globally for participation. Verification of individual identity is either done after the fact or not at all, leaving the door open to falsification of identity.

The risk is that convincing 'false' individuals, including multiple electronic identities can be created to access and/or subvert a study. This practice is called "sock puppetry" (a reference when a toy puppet is created by inserting a hand in a sock to bring it to life). Hundreds of thousands of on-line identities can be created through the use of computer scripting, Web automation, and social networks [5]. For example, in 2011, the US military has contracted with a California-based company, Ntrepid Corporation, to create and manage false identities online. The purpose was to spread pro-US pro-

paganda overseas by making it appear that the sentiments are coming from actual living humans and not digital sock puppets [34]. This is a critical reason why each user must be identified and authorized with specific permissions.

An attack on a popular survey site gives another example. Accessing the results from a survey poll, one of the authors (BF) noted an unusually high number of responses for a particular day. Examination of the demographic information left by respondents showed a variety of names, emails, and addresses-all different. However, examining the metadata captured about the individual responses showed some striking similarities. Digging a little deeper, the author found that all responses had emanated from two or three single Internet addresses that were associated with a commercial data center in California over a relatively short period of time.

This relatively crude attempt to bias survey results points out three critical issues in maintaining reliable on-line data collection and management:

● Establish the tools and techniques needed to maintain control over the integrity of data being collected and managed through the cloud provider. This includes restricting access to applications that can commit final data updates to only trusted users with or capturing and storing associated sources for review before final integration into the study "golden" data store. If an incident occurs, correlation of events between the cloud provider and the source of the compromise may be essential.

● Know the characteristics and behavior of the data collection. Is the activity for a given day unusually high? Do there appear to be surprising trends in the data that were not expected or appear strange?

● Design a process that establishes some confidence in authenticating participants, especially if individuals are being recruited through crowd sourcing or social media sites. Select additional demographics that can be used for authentication, take time to perform additional research through a common search engine. Make sure any informed consent clearly tells the participant what metadata may be collected to validate their identity.

Impact of data sharing and genomic data on privacy and security

Open data sharing avoids the duplication of research effort and facilitates the work of researchers who are able to build on and advance the work of others. Properly de-identified health data is an invaluable tool for scientific and health research advances. In most cases the National Institutes of Health [35] requires researchers to make data available to other investigators via an NIH-designated database or an approved alternative.

Concerns for the individual-patient or consumer-remains at the heart of the data sharing issue, especially as the personal data continuum continues to evolve, with increasing fidelity in the data about a person that can be tied to their identity. Anonymization of individual identifiable data figures prominently in both policy development around data sharing and in research into effective ways to prevent re-identification, yet retain the usability of datasets for use in research. Recently the International Cancer Genome Consortium (ICGC) announced the data protection policies for open and controlled access data elements especially re-identification issues [36, 37].

Regardless of the methods, there is always a possibility of re-identification. Identifiable markers can be used to determine the presence of an individual in a dataset, even without explicit personal information or when the genomic data has been aggregated. According to the systemic analysis of re-identification attacks [38, 39], success rate was approximately 26%, though this occurred on a small database with considerable heterogeneity among the studies. Scientific achievements as well as health policy decision-making comes at a cost, with some potential risk for re-identification, so balancing between the conflicting metrics of information quality and privacy protection needs to be considered [40]. Clinical trials frequently require collaborations across multiple healthcare institutions, or networks of diverse research organizations with private industries. These research collaborations often involve the release of de-identified patient level information between institutions, potentially increasing the probability of accidental disclosure of protected health information [11, 40].

Table 6. Privacy risk assessment for data re-identification [46]

Principle and Description	Examples
Replicability: Prioritize health information features according to the probability they consistently occur in relation to an individual. Consider the sensitivity of the information when making the determination of high, medium, or low	*Low:* Results of a given treatment *High:* Personal demographics that are relatively stable (date of birth)
Availability: Determine which external resources contain patients' identifiers and the replicable features in the health information, as well as who is permitted access to the resource and the level of confidence placed in the data integrity of the source. Consider study resources and other channels beyond the study	*Low:* The results of laboratory reports are not often disclosed with identity beyond dental environments *High:* Patient identity and demographics are often in public resources, such as vital records-birth, death, and marriage registries
Distinguish: Determine probability to which the subject's data can be distinguished in the health information source for correlation with additional information from the source	*Low:* What combinations of information have a low probability for identification such as date of birth, gender, and 3-digit zip *High:* What combinations of information have a low probability for identification such as date of birth, gender, and 5-digit zip
Assess Risk: The greater the replicability, availability, and distinguishability of the health information, the greater the overall risk for re-identification	*Low:* Assessment values may be very distinguishing, but they may not be independently replicable and are rarely disclosed in multiple resources to which many people have access *High:* Demographics are highly distinguishing, highly replicable, and are available in public resources
Establish safeguards: Select the safeguards and approach to anonymize the data in the research dataset given the associated risk	*Low Risk:* No safeguard *Medium Risk:* Encryption, tokenization *High Risk:* Application of differential privacy, try before release of information

Anonymizing large datasets is extremely difficult, especially as detailed information is available from unregulated sources that can be correlated with clinical data. The following are some studies that point the way as to how available information can be used as a basis for compromise:

- Healthcare planning data: This level of information is collected by most states. In 2013, news information was used to put names to patient-level health data related to hospitalizations, available publically from the State of Washington, for 43% of the cases examined [41]. This correlation is one reason the recent announcement by CMS of releasing the Medicare Provider Utilization and Payment Data is so chilling.

- Demographics in genomic datasets: Publicly available profiles in the Personal Genome Project (PGP) at Harvard (http://www.personalgenomes.org/) were linked to names and contact information in 84 to 87% of the cases through correlating PGP demographics with public records. The vulnerability is created by the demographics captured by the project itself, leaving the door open to how participants might protect themselves by providing accurate, but less specific information that is more difficult to match with the dataset [42].

- Recreational genealogy databases: Methods have also been reported that successfully link records in a dataset (even those without personal identifiers) to surnames based on genomic information in the dataset and querying recreational genetic genealogy databases. A key feature of this technique is that it entirely relies on free, publicly accessible Internet resources [43].

Traditional methods, such as controlled access databases that involve segmentation, encryption, and other de-identification methods, along with data use agreements, may fall short in the long run, given the complexity of and inherent risks of unregulated data sources that can be involved. Innovative and common sense approaches to information and data governance are needed that result in the establishment of clear and, most importantly, actionable policies for data sharing. Research into using differential privacy, a cryptographic process that maximizes the accuracy of queries from statistical databases while minimizing the chances of identifying its records, can be useful. These techniques allow a data owner to publish a pilot de-identified data set so potential data users can test various algorithms, including those not known to the data owner, before requesting full access to the dataset from the data owner [44].

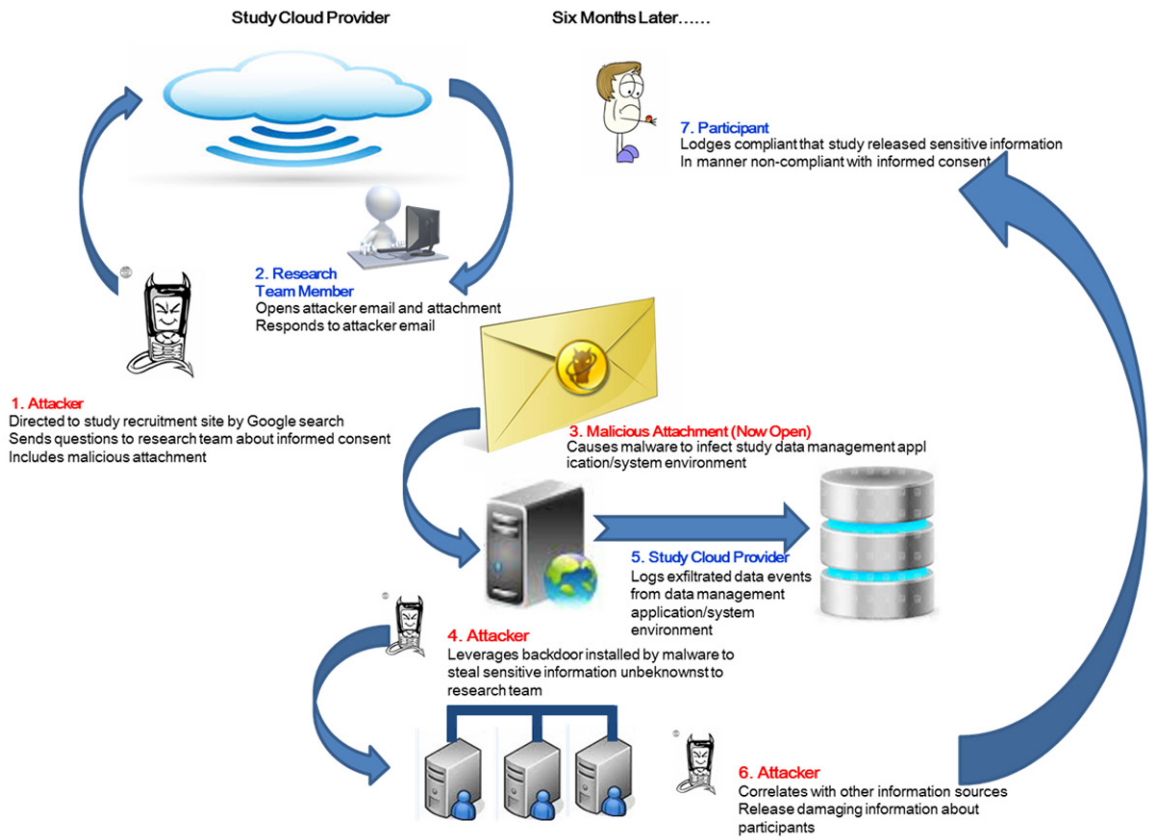

Figure 4. The security risk landscape for today's connected environment.

Germane to effective security is coupling the user who requests access to the data with the data or information that is actually accessed or acted upon. Each user must be robustly identified, their role and permissions defined in order to grant and monitor access to any application that may touch any study-related data store and any utility that permits a user to view, copy, or modify data from that store.

User roles, permissions, and related actions also needs to be captured as part of the data's provenance, the tracing and recording the origin and changes to data as it moves between data stores. A researcher may need to tie medical diagnosis, treatments and outcomes to the associated genome structure to evaluate possible relationships. A physician may wish to understand what an individual's genomic structure indicates about potential threats when attempting to make a diagnosis or prescribe a treatment. This can be achieved whether the data is available in one or more data stores provided the applications required to make the association are available to the user based on their role and permissions.

The development of a privacy risk assessment plan should help the researcher educate participants about the benefits and risks of genetic studies, mapping this to informed consent (to articulate how PII will be shared through unrestricted or controlled access repositories), describing what is being done to protect the data collected, and what a participant's expectations should be as relates to possible exposure. **Table 6** provides some basic guidance for how a risk assessment plan could be organized.

A security approach for the researcher

In the 1990's, information security was based on a layered defense or "defense in depth" that protected the sensitive information and data through strictly enforced logical and physical layers of security, the cyber version of walls and moats. Design of these safeguards was based

Table 7. A Security risk reduction strategy from the researchers' point of view

The Scenario	The Problem	What Can be Done
Attacker discovers the recruitment site and notes that identity validation is limited. A participant can send an email directly to a research team member asking questions about the study and that attachments are allowed.	No verification of participant identification Process that allows delivery of malware to identified research team member	Establish and validate participant identification procedures, even before completion of informed consent Do not provide direct email contact with research team members Do not allow attachments
Research team member, believing she is communicating with a possible participant, trusts the attachment that had been sent and opens it. The actions taken by the malware are stealthy and are not apparent to the team member. She does make a mental note to remind herself to update the anti-virus signatures and make sure her device is patched.	This compromise is successful because of social engineering, one of the main reasons why phishing can be a successful attack vector. The research team member trusts that the person she is communicating with is being honest about participating. The malware could possibly be detected by the appropriate software, but regardless her devices are already compromised because of configuration problems.	Establish a firm policy of not opening attachments from anyone not personally know to the recipient Scan all email attachments for malicious code. Make sure all mobile devices, just like all desktops and laptops, are patched and have the latest anti-virus and malware signatures installed.
A study participant reports the breach of sensitive information. He maintains that his information was released to a 3rd party in violation of informed consent. Upon review of the logs provided by the cloud service provider, it becomes apparent that sensitive data has been leaked from the study environment. A note was also made by the forensic team that most of the data at rest was not encrypted. The infiltrated data has been de-identified but has enough resolution and granularity to create some pretty damaging collateral about the participant. The logs also show that other participants may be adversely affected as well.	The researcher has worked with the cloud provider to set up procedures to monitor data leaving the data management application/system environment (egress). Here the problem is that the logs were not being reviewed in a timely manner. Data encryption at rest might not have helped contain the problem but the fact that the review exposed the lack of encryption will raise flags. The researcher will have to prove that the correlation of the released data with other sources is not something over which the study had control.	Establish a system activity review policy that calls for periodic review of all system events and logs. Stick to the schedule. Make sure the cloud provider provides access to logs with sufficient details. Encrypt all data at rest. Make sure that the informed consent allows for the situation where data may be correlated with other sources and inferences drawn that are outside the scope of the study. Lastly, conduct a privacy risk assessment on the study data to provide guidelines on what might be the risk in releasing de-identified or aggregate data.

on a formal risk assessment, usually assigned to the security or IT staff. The approach assumed the primary location of the sensitive information was a dedicated server, physically isolated and locked away in a data center. The entire process, often overly complicated by regulation such as HIPAA or HITECH, is long, involved, and essentially not user-friendly.

Securing sensitive data today requires an approach that is aware, agile, and adaptable as protective boundaries around information are more tenuous with mobile cloud computing and given that the criticality of the information (i.e., genomic information) is directly tied to the individual [45]. Security must be related to both the nature of a threat and the actual data that can be protected. The interplay across the security layers shown in **Figure 4** do not require a deep knowledge of the technology, but do demand an understanding of the possible threats and probable attack surfaces.

The scenario presented in **Figure 4** starts with an assumption that a study will be using an unsecured, publically available web site to attract possible participants. A participant will complete the informed consent, download the app, and begin to participate in the study. Validation of identity will occur upon the first required visit to a known healthcare professional for evaluation. Here an attacker is shown taking advantage of this 'loose' validation to compromise the study at an early stage.

Table 7 presents the assumptions for each step, what the resulting risks are, and what actions can be taken to mitigate them. A researcher needs to approach the problem of data protection from the standpoint of risk, understanding potential threats, impacts and outcomes, as well as accept the strong possibility of the unknown occurring. This is especially true in light of some the concerns mentioned so far: limitations in de-identification and uncertainties in the location and access to data, loosely coupled ecosystems for data capture and analysis, lack of visibility into the technical infrastructure, especially with mobile and cloud computing, and the ever-expanding number of cyber threats. The Data Protection Plan

Table 8. Data protection plan advice [47]

Data Plan Requirement	Safeguard Category
Identify who has access to the data	Identification, Authentication, and Access Methods: The actors in the study/use case need to be identified and roles established. Methods for identification and subsequent authentication should be defined. Access should be monitored, either individually or through the researcher's organization. Non-repudiation: Did the communication come from the designated person?
Identify who is maintaining confidentiality of the data	Data Governance: What data is being collected, what is the expected behavior (such as how many responses per day), and what are the data sharing policies and procedures across all data sources that will be correlated in the study?
Describe measures for protecting physical and software security of the data	Data Confidentiality and Integrity: How is the data stored and how is it encrypted? What are the de-identification rules and methods and what is the chance of re-identification? Application Confidentiality and Integrity: Same set of questions, including evaluation of the mobile apps that may be deployed.
Ensure authentication and authorization are required for those who have access to medical data by providing firewalls, data encryption, and password protection	Data use and data sharing agreements, implementation of policy around data. Have an action plan around data re-identification that includes both known and unknown (ancillary) methods. Protect the metadata that establishes relationships.
Contingency plan for dealing with any breach of confidentiality	Availability and service levels: Establish contractual terms with the cloud provider as embodies in service level agreement-how long does it take to response to a service request? How long to resolve?

must balance what is required with achievable safeguards under the researcher's control, an example is outlined in **Table 8**.

The road ahead

The array of threats and the technologies that they target affect all of us, not just researchers and practitioners. The growth in advanced threats, reaching down to even the individual, should dispel the myth that "it won't happen to me". In the coming decade, attackers will be driven by adoption of the applications and systems we most utilize.

For the researcher, the solution is both simple and complex. Cyber-situational awareness is no longer a luxury-it is fundamental in combating both the elite and highly organized adversaries on the Internet as well as taking proactive steps to avoid a careless turn down the wrong digital dark alley. The researcher, now responsible for elements that may/may not be beyond his or her direct control, needs an additional level of cyber literacy to understand the responsibilities imposed on them as data owner. Responsibility lies with knowing what you can do about the things you can control and those you can't. For if cyber risk is viewed from an inaccurate standpoint, there is a danger of coming up with controls and solutions for the unsophisticated hacks and not the sophisticated ones that have existed forever. Consequently, solutions based on a flawed understanding of the security landscape will give too much weight to certain assumptions which will lead in the wrong direction, while leaving the serious threats to exist unchecked and unidentified.

No one can depend on the traditional cyber walls and moats in the new paradigm of loosely connected computing and data devices-what is needed is more aggressive self-assessment with the thought that "offense can inform defense". Just as the move towards patient-generated data is transforming care, the growth in personally-generated identity is transforming health-related information security. Proactive self-assessment and self-security is needed to allow identification and remediation at the individual level. The researcher needs to know the data, the source, and the risks both the granular (individual) and collective (aggregate) levels to identify the risks and the possible threats. Then, the researcher can truly make decisions about relevant privacy and security controls based on this specific assessment rather than on general observations about the cybersecurity landscape.

Acknowledgements

This work was supported by the National Institutes of Health (NIH)/National Center for Advancing Translational Sciences grant UL1-TR001114.

Disclosure of conflict of interest

None.

Address correspondence to: Dr. Steven R Steinhubl, Scripps Translational Science Institute, 3344 North Torrey Pines Court, Suite 300, La Jolla, CA 92037, USA. Tel: 858-554-5757; Fax: 858-546-9284; E-mail: steinhub@scripps.edu

References

[1] Rouse M. SMAC (social, mobile, analytics and cloud) definition. TechTarget 2014.
[2] Snell E. Hacking Still Leading Cause of 2015 Health Data Breaches. http://healthitsecurity.com/2015.
[3] Kaâniche M, Deswarte Y, Alata E, Dacier M and Nicomette V. Empirical analysis and statistical modeling of attack processes based on honeypots. IEEE/IFIP International Conference on Dependable Systems and Networks (DSN-2006), Workshop on Empirical Evaluation of Dependability and Security (WEEDS) 2006; 119-124.
[4] Healthcare is a Growing Target for Cybercrime, and It's Only Going to Get Worse. United States Cybersecurity Magazine 2014; 1: 56.
[5] Goodman M. Future Crimes. New York: Doubleday, 2015.
[6] 2015 Cost of Data Breach Study: Global Analysis. 2015.
[7] Fifth Annual Benchmark Study on Privacy & Security of Healthcare Data. 2015.
[8] Fifth Annual Study on Medical Identity Theft. 2015.
[9] Humer C and Finkle J. Your medical record is worth more to hackers than your credit card. 2014.
[10] http://www.justice.gov 2015.
[11] Smart phone thefts rose to 3.1 million in 2013. http://www.consurmerreports.org 2014.
[12] Ackerman L. Mobile Health and Fitness Applications and Information Privacy, Report to California Consumer Protection Foundation. 2013.
[13] Kyle K. FTC: Fitness Apps Can Help You Shred Calories-and Privacy. http://adage.com 2014.
[14] Dehling T, Gao F, Schneider S and Sunyaev A. Exploring the Far Side of Mobile Health: Information Security and Privacy of Mobile Health Apps on iOS and Android. JMIR mHealth and uHealth 2015; 3: e8.
[15] Essany M. Mobile Health Care Apps Growing Fast in Number. 2013.
[16] State of Mobile App Security. 2014.
[17] 2015 Verizon Data Breach Investigations Report. 2015.
[18] Xu Z. Android Installer Hijacking Vulnerability Could Expose Android Users to Malware. http://researchcenter.paloaltonetworks.com 2015.
[19] Hua J, Shen Z and Zhong S. We Can Track You If You Take the Metro: Tracking Metro Riders Using Accelerometers on Smartphones. arXiv preprint arXiv:1505.05958 2015.
[20] Email Statistics Report, 2015-2019. 2015.
[21] Ladouceur R. Family physicians and electronic communication. Canadian family physician Medecin de famille canadien. 2014; 60: 310.
[22] Substitute Notice-Email Phishing Incident St. Vincent Medical Group, Inc. 2014.
[23] McCann E. Phishing scam breach compromises data of 39K. Health IT News 2015.
[24] Mutton P. Fake SSL certificates deployed across the internet. NetCraft 2014.
[25] Filkins B. New Critical Security Controls Guidelines for SSL/TLS Management. 2015.
[26] Dunning JP. Taming the Blue Beast: A Survey of Bluetooth-Based Threats. IEEE Privacy and Security 2010; 20-27.
[27] Wright J. Eavesdropping on Bluetooth Headsets. 2007.
[28] DaSilva J. WiFi experiment shows just how unsafe WiFi "hotspots" can be. Spot on Networks 2015.
[29] Rouse M. Information-Centric Security. Tech Target 2012.
[30] Kessler G. PASSWORDS-STRENGTHS AND WEAKNESSES. http://www.garykessler.net 1996.
[31] http://www.splashdata.com 2015.
[32] Notice for Use of Cloud Computing Services for Storage and Analysis of Controlled-Access Data Subject to the NIH Genomic Data Sharing (GDS) Policy. National Institute of Health 2015.
[33] Brodkin J. Gartner: Seven cloud-computing security risks. Network World 2008.
[34] Fielding N and Cobain I. Revealed: US spy operation that manipulates social media The Guardian 2011.
[35] NIH policy supports broader sharing of genomic data, strengthen informed-consent rules: American journal of medical genetics. 2015; 167A: viii-ix.
[36] Milius D, Dove ES, Chalmers D, Dyke SO, Kato K, Nicolás P, Ouellette BF, Ozenberger B, Rodriguez LL and Zeps N. The International Cancer Genome Consortium's evolving data-protection policies. Nature biotechnology 2014; 32: 519-523.
[37] NIH policy supports broader sharing of genomic data, strengthens informed-consent rules: research participants must give consent for secondary sharing, even if data are de-identified. Am J Med Genet A 2015; 167a: viii-ix.
[38] Emam KE, Jonker E, Arbuckle L and Malin B. A systematic review of re-identification attacks on health data. PLoS One 2011; 6: e28071.
[39] El Emam K, Jonker E, Arbuckle L and Malin B. A systematic review of re-identification attacks on health data. PLoS One 2011; 6: e28071.

[40] Matsui S. Genomic biomarkers for personalized medicine: development and validation in clinical studies. Comput Math Methods Med 2013; 2013: 865980.
[41] Sweeney L. Matching Known Patients to Health Records in Washington State Data. DataPrivacy.org 2013.
[42] Sweeney L, Abu A and Winn J. Identifying Participants in the Personal Genome Project by Name. dataprivacylab.org/ 2013.
[43] Gymrek M, McGuire AL, Golan D, Halperin E and Erlich Y. Identifying Personal Genomes by Surname Inference. Science 2013; 339: 321-4.
[44] Zhao Y, Wang X, Jiang X, Ohno-Machado L and Tang H. Choosing blindly but wisely: differentially private solicitation of DNA datasets for disease marker discovery. J Am Med Inform Assoc 2015; 22: 100-8.
[45] Daniel GW and Romine M. The significance of President Obama's Precision Medicine Initiative. Brookings 2915.
[46] Malin B. A De-identification Strategy Used for Sharing One Data Provider's Oncology Trials Data through the Project Data Sphere® Repository. 2013.
[47] Ohman C, Kuchinke W, Canham S, Lauritsen J, Schade-Brittinger C, Wittenberg M, McPherson G, McCourt J, Gueyffier F, Lorimer A and Torres F; ECRIN Working Group on Data Centres. Standard requirements for GCP-compliant data management in clincial trials. Trials 2011; 12: 85.