Healthcare Big Data in the Cloud

cloud security alliance®

The Cloud Security Alliance (CSA) promotes the use of best practices for providing security assurance within Cloud Computing, and provides education on the uses of Cloud Computing to help secure all other forms of computing.

Acknowledgments

Lead Authors:

Dr. James Angle

Key Contributors:

Vincent Campitelli
Patty Ryan

CSA Global Staff:

Alex Kaluza
AnnMarie Ulskey (Design)

Additional Support:

Bowen Close
Frank Guanco

Table of Contents

Abstract 5
Introduction 5
Big Data Characteristics and Analytics 6
Big Data in Healthcare 8
Privacy and Security 9
Conclusion 15
References 16

Abstract

We are living in the information age. Social media, emails, and numerous data sensors generate large and complex data sets every day. As we become a more digitized society, these large data sets present both opportunities and challenges. In healthcare, sharing and analysis of big data can aid in the identification and treatment of diseases as well as predicting epidemics. Additionally, big data analytics can aid in predicting healthcare requirements for specific populations by utilizing modeling and management of healthcare patterns for predictive analysis. The goal is to collect and analyze as much data about a patient as possible, from patient records as well as from wearable devices. Collection and analysis of data can aid in identifying warning signs of serious illness at an early stage, allowing for better treatment options. The challenges with big data are storage capacity, privacy and security, and collaboration.

Introduction

A digital transformation is emerging within the healthcare industry and changing it in ways that were not possible just a few years ago. The use of cloud computing and big data analytics, coupled with the move to consumer-centric healthcare, is reshaping the way healthcare is delivered. Healthcare Delivery Organizations (HDO) have access to large quantities of data that, if collated, analyzed, and properly utilized, can provide tremendous benefit to both the HDO and the patient. This data is coming from sources that were unavailable until recently, including: the Internet of Things, electronic health records (EHR), other clinical data, and social media (Faggella, 2019).

Big data in healthcare refers to various large and complex data, which are difficult to analyze and manage. Big data analytics in healthcare enables analysis of large data sets from large numbers of patients, identifying clusters and correlation between datasets and developing predictive models. The information produced can be shared with other HDOs and research organizations for improving patient outcomes and can also identify health issues and allow for early intervention and treatment. The use of predictive analytics can aid in both patient care and HDO care delivery utilization, which is extremely important in rural areas where healthcare capacity is limited.

The use of cloud computing allows for big data analytics and collaboration, but with the use of cloud computing come the challenges associated with security and regulatory compliance. This paper will look at big data and some use cases for big data in healthcare, the impact of big data on healthcare, regulatory requirements for Protected Health Information (PHI) in the cloud, and securing PHI in the cloud. The paper is organized as follows: 1) Description of the characteristics of big data and big data analytics. 2) Use cases for big data in healthcare. 3) Data privacy and security for big data in the cloud. 4) Discussion and further works.

Big Data Characteristics and Analytics

"Big data" refers to volumes of data so large that they are difficult to process using traditional techniques (Ristevski and Chen, 2018). Data is not referred to as "big data" due to its size alone, but rather the task of deriving usable information from it. Creating large, complex data and gleaning information from the data present a significant challenge. Another challenge with big data in healthcare is that medical data is complex and contains sensitive, private, personal information. The healthcare industry has been moving toward utilization of the vast pool of data created by electronic health records, imaging data, and data from user devices like mobile devices and wearable devices (Alexandru et al., 2016). Although there are various definitions of big data, it is commonly described to include six Vs, which are illustrated in Figure 1.

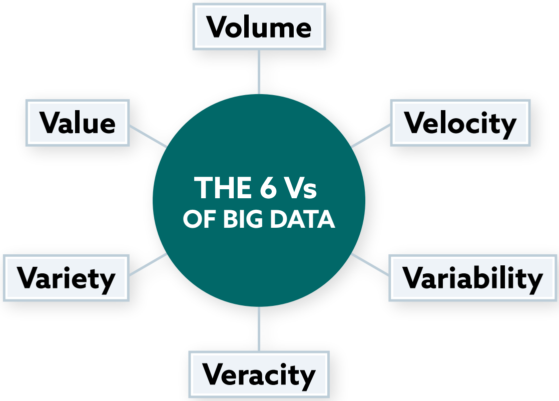

Figure 1. The 6 Vs of big data.

- Volume: The size of data produced, typically vast and over one petabyte in volume. In healthcare, EHR alone represent large volumes of data. Additionally, this data can change frequently as new test data is introduced and things like International Classification of Diseases (ICD) codes are updated.
- Velocity: The speed at which a data user can access and analyze the data. Velocity is needed in healthcare, since it allows healthcare providers to exchange and use data in a timely manner.
- Variety: The types of data, including structured, semi-structured, and unstructured data. Healthcare has diverse sources of data including multimedia, social media, and financial transactions.
- Veracity: The quality of data produced. Healthcare data must be relevant, reliable, and error-free because life-and-death decisions depend on accurate information.
- Value: The value derived from analysis of the existing data – the most important aspect of big data. The value of big data in healthcare today is largely limited to research.
- Variability: This regards about consistency of the data over time.

Big data analytics in healthcare involve "the use of data, information technology, statistical analysis, quantitative methods, and mathematical computer-based models to help healthcare providers gain improved insight about these patients and make better, fact-based decisions" (*An Overview of Health Analytics*, as cited in Alexandru et al. 2016). The use of big data in healthcare is aimed at improving patient safety and clinical outcomes while promoting wellness and disease management.

There are four primary functions for big data in healthcare:

1. Descriptive analytics, which examines data to understand healthcare decisions and make new informed ones. Its models can be used to categorize, define, combine, and classify the data to extract useful information.
2. Predictive analytics, which examines old or summarized healthcare data for the purpose of identifying patterns of relationships which it can extrapolate to predict the future. Data mining can be used to identify hidden patterns in health data and thus anticipate medical risks, predict outcomes for patients, and improve health-related services.
3. Prescriptive analytics, which uses information and health and medical knowledge to solve problems which involve many alternatives and thus make descriptive or predictive analytics infeasible.
4. Discovery analytics, which uses knowledge about knowledge to identify unknown facts from data and improve the future. It can help discover new diseases or medical conditions, drugs, and treatments (*An Overview of Health Analytics*, as cited in Alexandru et al., 2016).

Applications of big data analytics can improve patient-based service, detect spreading diseases such as COVID-19, generate new insights into disease mechanisms, and provide better treatment methods. Data mining techniques employed can aid in identifying and revealing disease monitoring and health-based trends.

Big data analytics can provide advantages to HDOs by digitizing, combining, and effectively using big data to realize significant benefits in the delivery of healthcare. The benefits realized can include early disease detection, allowing for treatment options and potentially better outcomes and effectively managing population-specific health issues.

A recent use of big data in healthcare is to collect and analyze with the goal of connecting with patients and providing them a better experience. Big data analytics from patient devices can provide HDOs with biometric measurements, giving the provider a better understanding of the patient's overall health and presenting insight beyond symptoms. This enables HDOs to become patient-centric (Davis and Gourdji, 2019).

Big data analytics is also seen as a cost-savings enabler. The McKinsey Global Institute believes big data could help reduce waste and inefficiency in the following three areas (Manyika et al., 2011):

> "Clinical operations: Within clinical operations are five big data levers [comparative effectiveness research, clinical decision support systems, transparency about medical data, remote patient monitoring, and advanced analytics applied to patient profiles] that mainly affect the way providers, payors, and PMP [pharmaceuticals and medical products] provide clinical care. We estimate that, if fully employed, these five levers could reduce national healthcare expenditure by up to $165 billion a year from a base of $2.5 trillion in 2009."

> “R&D [research and development]: Five big data levers [predictive modeling, statistical tools and algorithms to improve clinical trial design, analyzing clinical trials data, personalized medicine, and analyzing disease patterns] could improve R&D productivity in the PMP subsector. Together, these levers could create more than $100 billion in value, about $25 billion of which could be in the form of lower national healthcare expenditure.”

> “Public health: The use of big data can improve public health surveillance and response. By using a nationwide patient and treatment database, public health officials can ensure the rapid, coordinated detection of infectious diseases and a comprehensive outbreak surveillance and response through an Integrated Disease Surveillance and Response program. This lever offers numerous benefits, including a smaller number of claims and payouts, thanks to a timely public health response that would result in a lower incidence of infection. The United States would also be better prepared—in terms of laboratory capacity, for instance—for emerging diseases and outbreaks. There would also be a greater public awareness of the health risks related to infectious diseases, which, in turn, would lower the chance of infections thanks to accurate and timely public health advisories. Taken together, all these components would help to produce a better quality of life.”

Big Data in Healthcare

Electronic health records (EHR) is one of the most widespread big data use cases in healthcare. EHR keep track of each patient’s health chart and their medical reports, thereby reducing the need for duplicate tests and the associated cost. Information available includes medical history, laboratory and imaging results, and demographics.

EHR are used to schedule appointments and to send reminders for appointments, lab work, and prescriptions. In many states, HDO records are connected to the state health information exchange in the cloud, which allows all providers to access a patient’s information. This exchange can be used for patients traveling within their state who may require care from an HDO other than their own. The goal for health information exchanges is to connect them nationwide, so patients can receive care anywhere.

Real-time alerting is an example of big data analytics in healthcare, allowing HDOs to analyze medical data in support of clinical decision making. This type of analytics can prevent, for instance, prescribing medication that may have an adverse reaction with other medications prescribed by a different provider.

With a shift in healthcare moving towards a more consumer-centric environment, this data can be combined with data from patient wearable devices that continuously send data to the cloud. The goal of such data collection is to reduce in-house treatment, thus avoiding costly stays (Lebied, 2018).

Additionally, big data can be used to assess the health of the general population and establish trends and identify patterns. This can help identify diseases, location, and outcomes in a pandemic. The use of big data was instrumental in Taiwan's response to COVID-19. Taiwan was quick to mobilize their response by using big data analytics to respond based on travel history and clinical symptoms. They also were able to classify infectious risk based on flight information and travel history. Patients at low risk received clearance for immigration, and high-risk patients were quarantined at home and tracked through their mobile phone during the incubation period (Wang et al., 2020).

The value of predictive analysis cannot be overstated. Predictive analysis gives us the power to face diseases earlier, allowing early intervention and improved outcomes. One can readily see analytics being utilized, aiding in the decision-making of healthcare personnel and patients. HDOs can earlier identify patients with a high risk of developing chronic conditions, giving the patient the best chance of avoiding long-term health issues (Bresnick, 2018).

Taiwan's response to the COVID-19 pandemic is but one example of this type of predictive analysis. With early recognition, daily briefings, and health messaging, the government of Taiwan was able to reassure the public by delivering timely, accurate, and transparent information regarding the epidemic. Taiwan is an example of how a society can respond quickly to a crisis and protect the interests of its citizens during a pandemic (Wang et al., 2020).

Privacy and Security

Prior to moving any data to the cloud, HDOs should conduct a full risk assessment and control assessment, which will ensure they have a full understanding of the risk associated with cloud computing. It will also identify the controls put in place by the cloud service provider and identify any gaps. Additionally, it makes sure the cloud service provider is compliant with relevant privacy and security regulations.

People often conflate privacy and security; however, privacy and security should be treated as distinct concerns (Bambauer, 2013). This paper looks first at privacy issues and then at security issues as they relate to the data lifecycle. The cloud data lifecycle has six stages:

1. Create: New content is generated.
2. Store: Data is committed to storage.
3. Use: Data is viewed and processed.
4. Share: Data is made available for use by others.
5. Archive: When no longer actively used, data is committed to long-term storage.
6. Destroy: When no longer required, data is permanently destroyed.

Privacy

Privacy is concerned with decisions about legitimate access to, use of, and alteration of information. Privacy establishes a framework for deciding who should legitimately have the capability to access and alter information (Bambauer, 2013). The invasion of patient privacy is considered a growing concern in the domain of healthcare big data analytics due to the emergence of advanced persistent threats and targeted attacks against information systems.

A major goal of the HIPAA Privacy Rule is to assure that PHI is properly protected while allowing for the flow of health information needed to provide and promote high-quality healthcare. Before any disclosure of PHI that is not for treatment, payment, or healthcare operations or otherwise permitted or required by the Privacy Rule, the HDO must obtain the patient's written authorization. All authorizations must be in plain language and contain specific information regarding the information to be disclosed and used, who will receive the information, how the information will be used, the expiration date, and the right to revoke the disclosure. However, the HIPAA Privacy Rule specifies that once the PHI has been de-identified/anonymized, it can be disclosed and used without any restriction. The information is no longer considered protected PHI (Department of Health and Human Services, 2013).

Big data has driven unprecedented innovation, economic value, and improvement in social services. In healthcare, the value associated with big data is often connected to the collection of personal information, the consequences of which individuals may not fully understand. The challenge is to gain benefit from big data while protecting the individual's privacy. Privacy protection must allow for individual choices while providing effective risk mitigation (NIST, 2020).

Based on where the data is collected and where the data is stored, the EU's General Data Protection Regulation (GDPR) may also apply. The main aims of the GDPR is to ensure the personal data of EU "data subjects›" is protected and to increase the rights of EU data subjects over their personal data. Businesses that collect, process, or store the data of EU data subjects must comply with the GDPR regardless of the location of that business.

Looking at data throughout the data lifecycle is necessary for following the GDPR. Ensuring that the GDPR rules are enforced will also help ensure the organization is following U.S. rules.

Create

Personal data (also termed personally identifiable information) is any piece of information that contains an "identifier" that can be used to identify a specific individual or group of individuals. When personal data is collected, the individual whose data is being collected has the right to know what data is being collected, what the data will be used for, and if it will be shared. The collector must obtain consent, which means asking users for permission to process their data. Companies must explain their data collection practices in clear and simple language and users must explicitly agree to them.

Store

When the GDPR refers to the processing of data, it means the handling, use, storage, and destruction of information. Processors and controllers are responsible for ensuring data security at every stage of its lifecycle. When data is stored, individuals retain the right to access their personal data, correct errors, and request the removal of their information.

Use

The GDPR gives individuals certain rights when their data is used. Here are some of the rights outlined by the GDPR:

- You're entitled to know exactly how your data is collected and used.
- You can ask what information has been collected about you.
- If there are mistakes in your data, you can request to have them corrected.
- You can have your data deleted from records.
- You're allowed to refuse data processing (for example, marketing efforts).

All businesses are required to have a privacy policy that explains what they do with users' information.

Privacy policies must:

- Include contact details of the company and its representatives.
- Describe why the company is collecting the data.
- Say how long the information will be kept on file.
- Explain the rights users have.
- Be written in simple language.
- Name the recipients of the personal data (if the company shares data with another organization).

Share

The National Institute of Standards and Technology (NIST) refers to the "data processing ecosystem" to describe how data may be shared between different organizations. The data processing ecosystem encompasses a range of entities and roles that may have complex, multi-directional relationships with each other. A key factor in the management of privacy risk is an entity's role in the data processing ecosystem, which can affect its legal obligations. In the healthcare ecosystem, a formal agreement/ contract between HDOs and cloud service providers is a requirement.

Within this ecosystem, the Data Processing Ecosystem Risk Management function and category refers to the organization establishing and using priorities, constraints, risk tolerance, and assumptions to support risk decisions associated with managing privacy risk and third parties within the data processing ecosystem. In this case, the organization has established and implemented the processes to identify, assess, and manage privacy risks within the data processing ecosystem.

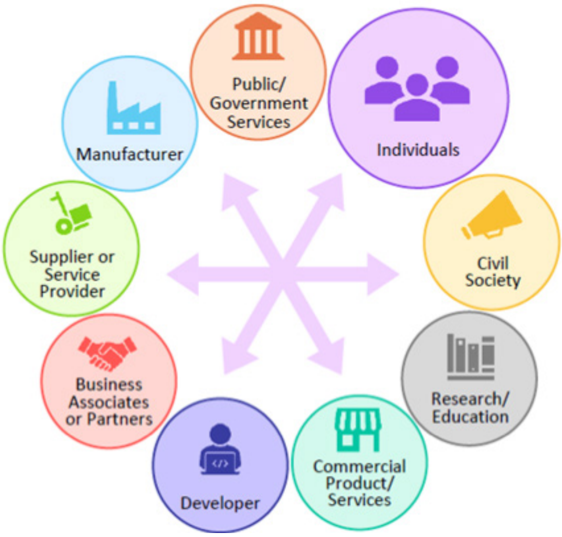

Figure 2. Data processing ecosystem relationships (NIST, 2020).

Archive

In general, except for PHI all personal information should be deleted when it is no longer in use. Under the GDPR, personal data must only be stored for the time taken to achieve the purpose for which the data have been collected. Personal data cannot be stored indefinitely. If the underlying information may be required in the future, the personal information can be disassociated from the data prior to archiving. Disassociating enables the processing of data or events without association to individuals or devices beyond operational requirements (NIST, 2020).

Destroy

Data no longer required must be securely disposed of. All removable media that contains private data should not be disposed of without first ensuring all protected data has been securely removed. All hard copies of such data must be finely shredded before disposal.

Additionally, here are several practices that can be implemented to ensure data remains protected.

- Clear desk policy: Before any employee leaves, care should be taken to ensure that no materials containing private data are left on the desk and that computers are locked.
- Password security: It is imperative that no passwords are written down. Passwords themselves should be long and complex.
- Practice secure storage: Any material that contains a person's personal private information must be stored in a secure manner. Digital data must be encrypted.
- Ensure that mobile devices are secured: Devices should be adequately secured and be password-protected.
- Ensure secure transmission of data: Private information should not be sent via insecure means.
- Secure disposal of data: Removable media that contain private data should not be disposed of without first ensuring that all protected data has been securely removed from the devices.
- Reporting breaches: If a breach occurs, in most instances an organization has 72 hours to report the breach.

It is essential to train all employees on their responsibilities under various privacy rules and to strictly adhere to policies and procedures to minimize the risk.

Security

While privacy involves decisions about who has access to, use of, and ability to alter information, security implements privacy's choices. Privacy dictates how security's options should be implemented and the conditions under which they are applicable. Security determines who can access, use, and alter data. Security, therefore, is the interface between information and privacy. It facilitates privacy rights, putting them into effect (Bambauer, 2013).

Change is the new norm for the healthcare sector. In fact, digitization of health and patient data is undergoing a dramatic and fundamental shift in clinical, operating, and business models. This shift is spurred by a population demanding that healthcare be more engaged in planning and delivering innovative, personalized healthcare (Abouelmehdi et al., 2018). This demand for evidence-based medicine (as opposed to subjective clinical decisions) requires greater use of big data.

HDOs store, process, and share large volumes of data that are used in big data analytics to support the healthcare industry. Data is a critical asset that requires HDOs to implement security solutions in order to secure the data and comply with healthcare mandates. Big data for healthcare can contain information requiring protection based on multiple regulations. In addition to the Health Insurance Portability and Accountability Act (HIPAA), the HDO must also comply with all regulations for Personally Identifiable Information (PII).

The GDPR checklist is useful to know some of the terminology and the basic structure of the law. The following is some of the items from the checklist:

> Under the GDPR, you must follow the principles of data protection by design and by default, including implementing appropriate technical and organizational measures to protect data. In other words, data protection is something you now have to consider whenever you do anything with other people's personal data. You also need to make sure any processing of personal data adheres to the data protection principles. Technical measures include encryption and organizational measures include actions such as limiting the amount of personal data you collect or deleting data you no longer need.
>
> Most of the productivity tools used by businesses are now available with built-in end-to-end encryption, including email, messaging, notes, and cloud storage. The GDPR requires organizations to use encryption or pseudonymization.
>
> Even if your technical security is strong, operational security can still be a weak link. A security policy can ensure that your team members are knowledgeable about data security. It should include guidance about email security, password, two-factor authentication, device encryption, and VPNs.
>
> A data protection impact assessment is a way to help you understand how a product or service could jeopardize your customers' data as well as how to minimize risks. The GDPR requires organizations to carry out this kind of analysis whenever they plan to use personal data in such a way that it's likely to result in a high risk to customers' rights and freedoms.
>
> Another part of data protection by design and by default is making sure someone in your organization is accountable for compliance. This person should be empowered to evaluate data protection policies and the implementation of those policies.
>
> You should have signed data protection agreements between your organization and any third parties that process personal data on your behalf. This includes any third-party services that handle personal data of your subjects, including analytical software, email, and cloud services (GDPR, 2020).

As with privacy, looking at data throughout the data lifecycle is done so that all requirements are followed.

Create

First, it is important that the data is required and that there is a business need for it. Second, you must have consent to collect PHI or PII. Depending on where the data is created, there are regulatory requirements; for example, the GDPR requires that security be built-in from when the data is created. Likewise, HIPAA requires security for all PHI, from inception to destruction. The result of these regulations is that the data must be created in a secure environment.

In order to meet these requirements, the cloud customer must understand the legal and contractual requirements to ensure proper handling of their data. The customer should also know where the data will be stored and where it will be processed to comply with all jurisdictional requirements. Ensure, through a risk analysis, that the cloud provider can adequately protect the data.

Store

There are numerous areas to consider when storing data in the cloud. The data owner must determine where the data comes from as well as where it is stored. The service provider must protect the data in the cloud, including access control and encryption. Access control should be implemented within the management plane, application level, and internet sharing controls. Data can be protected through encryption or tokenization.

In addition to these controls, the cloud service provider should have a secure architecture in place that utilizes standard security best practices, including a robust monitoring, auditing, and alerting capability. A data loss prevention system can help identify who is using the data and from where.

Use

As pointed out in the storage section, where the data is processed has an impact on what regulatory requirements there are. In order to access the data, organizations should use federation and multifactor authentication whenever possible. Identity and access management (IAM) is an important part of securing data in use. IAM is the process of managing access of individuals to digital resources. The process determines who has access and what authorization they have regarding the data. Additionally, only the required data should be collected and used, and only for the specified collection reason. Application programming interface (API) requires digital signatures to ensure security.

Share

As a reminder, not all data should be shared. When data sharing is required, the organization responsible for the data must ensure its security. As with data use, IAM is critical for data security. At a minimum, a data loss prevention (DLP) solution should be employed to monitor the use of data. DLP is used to ensure that sensitive data is not lost or misused. DLP software classifies regulated and confidential data and identifies violations of policies defined by organizations or within a predefined policy pack, typically driven by regulatory compliance such as HIPAA, PCI DSS, or the GDPR. In addition, the organization should consider an information rights management (IRM) technology. IRM is a security technology that protects documents containing sensitive information from unauthorized access.

Sharing data requires the data to be transmitted from the cloud to all the users of the data. Data should be encrypted while in transit. When encrypting data in transit, ensure that a secure protocol is used. Using Transport Layer Security (TLS) 1.2 is essential.

Archive

Data that is important but that does not need to be accessed or modified frequently is stored in a data archive. Archiving your data provides many benefits, especially in terms of efficiency. Archived data should be encrypted and access to the data should be controlled. Personal data or healthcare data can only be kept if it is still required for the intended purpose of collection.

Destroy

When no longer required, data should be permanently destroyed. Since data in the cloud is in a shared environment, typical data destruction methods such as magnetic destruction cannot be used, leaving encryption as the only real option. Encryption is applied to data and encryption keys are destroyed, guaranteeing your data is inaccessible.

Conclusion

Big data in healthcare is not going away. With healthcare big data in the cloud, there is the opportunity to manage healthcare research in order to derive benefits in care delivery and predictive analytics. However, as with any other cloud implementation, there are several challenges that impede its potential. A primary concern is privacy and security issues. Maintaining the sanctity and integrity of healthcare data is of paramount importance.

While there are potential opportunities for big data in healthcare, big data security and privacy are considered huge obstacles. In this paper, we have presented privacy and security issues in each phase of the data lifecycle and discussed methods to mitigate privacy and security compliance concerns.

References

Abouelmehdi, Karim et al., 2018. *Big Healthcare Data: Preserving Security and Privacy*, Journal of Big Data, Vol. 5.

Alexandru, Adriana et al., 2016. *Healthcare, Big Data and Cloud Computing*, WSEAS Transactions on Computer Research, Vol. 4.

Bambauer, Derek E., 2013. *Privacy Versus Security*, The Journal of Criminal Law and Criminology, Vol. 103, No. 3.

Bresnick, Jennifer, 2018. *10 High-Value Use Cases for Predictive Analytics in Healthcare*, Health IT Analytics, retrieved from https://healthitanalytics.com/news/10-high-value-use-cases-for-predictive-analytics-in-healthcare.

Davis, Scott M., and Gourdji, Jeff, 2019. *Making the Healthcare Shift: The Transformation to Consumer-Centricity*, Morgan James Publishing, New York.

Department of Health and Human Services, 2013. *Summary of the HIPAA Privacy Rule*, retrieved from https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html.

Faggella, Daniel, 2019. *Where Healthcare's Big Data Actually Comes From*, Emerj, retrieved from https://emerj.com/ai-sector-overviews/where-healthcares-big-data-actually-comes-from/.

GDRP.EU, 2020. *GDPR Checklist for Data Controllers*, retrieved from https://gdpr.eu/checklist/.

Lebied, Mona, 2018. *12 Examples of Big Data Analytics In Healthcare That Can Save People*, Retrieved from https://www.datapine.com/blog/big-data-examples-in-healthcare/.

National Institute of Standards and Technology (NIST), 2020. *NIST Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk Management*, retrieved from https://www.nist.gov/system/files/documents/2020/01/16/NIST%20Privacy%20Framework_V1.0.pdf.

Manyika, James et al., 2011. *Big Data: The Next Frontier for Innovation, Competition, and Productivity, McKinsey Global Institute*, retrieved from https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation.

Ristevski, Blagoj, and Chen, Ming, 2018. *Big Data Analytics in Medicine and Healthcare*, Journal of Integrative Bioinformatics, Vol. 15, No. 3.

Wang, Jason C. et al., 2020. *Response to COVID-19 in Taiwan: Big Data Analytics, New Technology, and Proactive Testing*, JAMA, Vol. 323, No. 14.