International Journal of Healthcare Information Systems and Informatics
Volume 15 • Issue 4 • October-December 2020

Blockchain Application to the Cancer Registry Database

Joseph E. Kasten, Pennsylvania State University, York, USA

ABSTRACT

Blockchain, since its 2008 conceptual inception, has largely been contextualized in crypto currencies. Today, blockchain technology has matured to a level that allows the exploration of its application to other and diverse domains, including the management of cancer registries. When collecting and handling data relating to cancer diagnosis and treatment as mandated by law in many municipalities, the process is both time-consuming and requires significant coordination among multiple levels of data collecting jurisdictions. This often leads to inconsistent data vis-à-vis the various levels of data storage. This paper calls for using a blockchain-based mechanism to alert the data users on possible inconsistencies prior to applying the collected data in cancer research. A system framework drawing on the design science research methodology is found to result in increased data quality so as to improve cancer research outcome accuracies.

KEYWORDS

Blockchain Technology, Cancer Registry, Design Science Research (DSR), Distributed Ledger Database

1. INTRODUCTION

A cancer diagnosis can devastate a patient's life physically, emotionally, and if the patient is not properly insured, financially. Once diagnosed with cancer, the patient and his/her family typically begin a planning process of treatment that may lead, for all intent and purpose, to a cure. At the initial planning stage, however, little thought is given to the myriad activities that take place prior to, and concurrent with, the patient's treatment journey. Accordingly, a key aspect of cancer care, which may often be overlooked, is the role of cancer research; for instance, the patient will likely not understand the role that previous research has played in his/her treatment plan. Also, most patients are not understanding that, if they are admitted into a large cancer treatment center, especially one specializes only in cancer care, they will most likely become a participant in one or more studies that are currently underway. Indeed, one of the many forms they will sign to initiate their treatment program will be a release that allows the faculty and staff at the facility to collect and analyze data pertaining to the patient's disease and/or treatment progression.

Broadly, data used in health research come from multiple sources, but one of the most potent sources is the Electronic Medical Record (EMR). As a means to improve the quality, safety, and efficacy of the health care process, EMR development and implementation was mandated by Title XIII of the American Recovery and Reinvestment Act (ARRA) of 2009 (United States Congress, 2009). As a side benefit, EMR also created a much more efficient and effective tool for health researchers to

DOI: 10.4018/IJHISI.2020100105

extract data for their studies. Even so, cancer research requires data that are not often readily available in the EMRs. One challenge is the nature of cancer care itself - it is not atypical for the diagnosis of cancer, for example, to occur in one institution, the labs to be done in another, and the treatment to occur in a third. While all of these facilities will have EMRs, it is conceivable that the different systems will contain data only on the patient's encounter with the specific institution. No single EMR will contain enough data to track the progress of the cancer, its histology and staging, and whether the cancer is a new or recurrent tumor. Thus, while useful for research to inform decisions relating to insurance coverage and costs of treatment, these data are not suitable for the types of knowledge needed to develop effective means of fighting the patient-specific cancer disease.

This paper highlights key aspects of the data collection, organizing, and managing processes within the cancer registry system that sometime lead to inconsistencies in the data held within the databases. Such inconsistencies are a threat to the validity of the cancer research process; accordingly, the paper presents a mechanism, using blockchain technology, to identify these inconsistencies when they occur and to call attention to these inconsistencies for those wishing to use the data for any type of cancer research. As documented in the literature, blockchain is being applied to a growing number of healthcare use cases owing to its ability to protect the integrity of the data. When data are used to make life or death decisions, or to perform research that results in the development of life-saving therapies such as those developed by cancer researchers, the data consumer must be able to trust that the data are accurate and valid. This means having a tool to compare the data being used with its original state to check for tampering or other changes – a requirement that characterizes the blockchain technology. As such, this paper seeks to address the following broader research questions:

- *Can blockchain data management processes be used to help manage the data discrepancies between the various levels of cancer registries?*
- *What improvements can the use of blockchain-based data storage and management brings about in the cancer registry process?*

The rest of this paper will be organized as follows. Section 2 overviews the cancer registry processes, blockchain technology, and the literature that supports the current research. Section 3 details the study methodology and the application of that methodology in cancer registry system design while section 4 presents and highlights important results of the study. Finally, section 5 closes the paper with some concluding remarks while pointing out certain limitations of the present study and offers directions for future research.

2. BACKGROUND

2.1 Cancer Registry Processes

Decades before the call for EMRs is to be realized pervasively, the need for databases created especially for cancer research was recognized in key US jurisdictions. In 1940, the State of New York (NYS), for example, created its statewide cancer registry service, which continues to the present (New York State Department of Health, 2018). After the passage of the Cancer Registries Amendment Act in 1992, States such as NYS began receiving federal funds to upgrade and extend their cancer registry systems.

Currently, each of the fifty (50) US States, the District of Columbia, Puerto Rico, as well as Canadian provinces, and many other countries maintain cancer registries for the purposes of research (National Institutes of Health, NIH website on *Surveillance, Epidemiology, and End Results*). The primary difference between these cancer registries and the typical EMR is that the cancer registry collects data (in some cases over 100 items) from multiple institutions; more specifically, it allows for the stitching together of a more complete picture of a patient's disease, including all factors that

are commonly encountered in cancer development in a patient such as stage of disease diagnosis, histology, laterality, and behavior (NIH website).

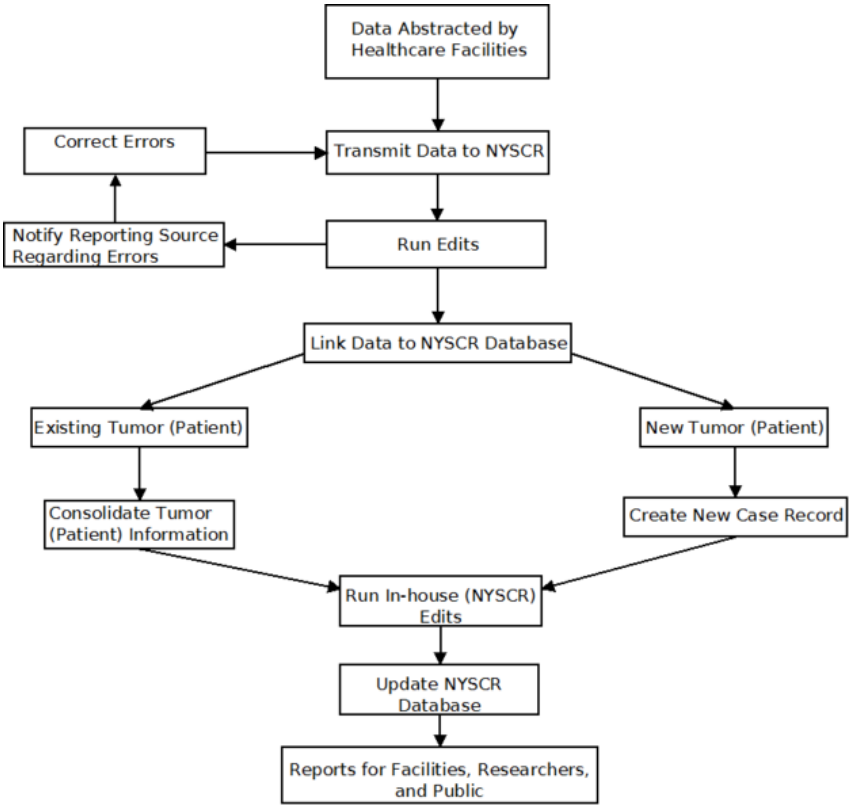
The current process for compiling the tumor registry data is to compel each facility involved in a patient's cancer diagnosis (or treatment) to report these data to the State cancer registry. This includes all institutions from the family doctor to the lab to the largest cancer facility in the State. Each such facility must report their encounters with the patient in a standard format and within a pre-specified time period. The time period varies with the type of patient (encounters with children must be reported more quickly than with adults) and by the type of cancer (NYS Department of Health, 2017). Frequently, data for a specific patient might be reported from many different facilities depending on the treatment progression. Moreover, each patient might have multiple records in the registry as each primary tumor is tracked individually. Further complicating the process is that these data are not added all at once, but rather as the patient proceeds through the diagnosis-treatment process. Hence, a record might remain open for years until the disease resolves; for patients who reach a status of "free of disease," their post-treatment care is also noted in the record for that occurrence.

The data contained in the cancer registry are then made available for research. Each State registry publishes its own statistical analysis, but the real value is the cancer researchers' ability to have a readily accessible and available, standardized collection of cancer-related data, which have been organized based specifically on their usefulness for cancer research. Note that these registries can be used to track issues as widely varying as treatment efficacy to prevailing cancer rates in a specific geographic region. As with any research, the quality of the findings is directly related to the quality of the data used. In this sense, cancer registry data administrators, who are typically employees of the individual State's department of health, are expected to take certain steps to ensure that the data in the registry are of a high quality. For instance, **Figure 1** offers insights into the data submission and revision process for the New York State Cancer Registry (NYSCR).

Some of the measures of data quality include the percentage of "death certificate only" entries in the database, the percentage of diagnoses confirmed by microscopic evidence, and the percentage of cases with a nonspecific diagnosis (NYS Dept. of Health, 2017). Each of these measures might indicate that there are likely pieces of the data trail that have not yet been reported or correctly recorded. In some cases, the patient has sought treatment in another State and so those data would be reported to that State's registry, but in most cases it signifies a data quality issue. To remedy these errors, there is a standard data revision process in place that specifies the method and timing of making changes (NYS Dept. of Health, 2017). Most changes are submitted along with the monthly batch of new records, but some require a different time window. The **Figure 1** NYSCR model describes how the patient-specific cancer-related data are collected, classified, validated, and distributed within and beyond the cancer registry system. This process is typical of other jurisdictions and municipalities that are compliant with the Surveillance, Epidemiology, and End Results (SEER) guidelines. Both the NIH and the National Cancer Institute (NCR) sponsor the SEER program. The program serves to standardize the data format of data collected in cancer registries, thus improving data extraction and analysis. As shown in **Figure 1**, there are two opportunities to improve the quality of the data contained in the registry: (a) once at the facility level; and (b) the other at the State NYSCR level. From a data integrity perspective, the constantly changing data, whether from adding new data or updating existing ones, presents a challenge to ensure that researchers are provided with the most current data. This paper argues for a solution to that challenge via blockchain technology.

Existing cancer registry systems are challenged by weaknesses, which can be resolved with use of blockchain technology. One such limitation is the possibility of researchers assuming the same data to be present in both the local and State databases when, over a period of time, that may not necessarily be the case. This arises from the two-step process of how data are being collected - first, at the institution level and then passing them onto the State level. Data accumulated at the institutional level are included in the local registry; periodically, as a patient's treatment continues, data are added to existing records. Alternatively, as a result of internal quality control procedures, data are revised. At

Figure 1. Flow of data from reporting facilities through the NYSCR (NYS Dept. of Health, 2017)

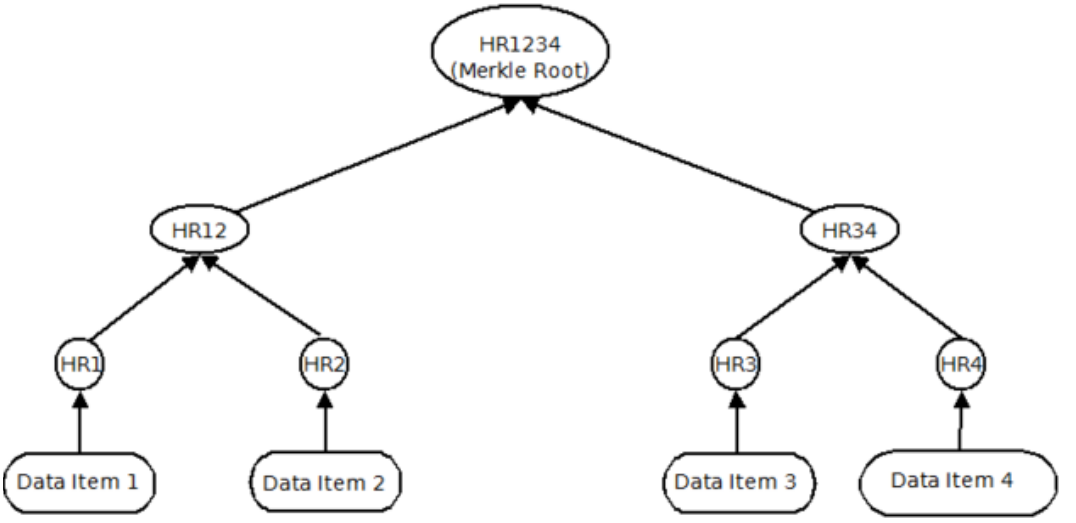


predefined intervals (monthly), these new and revised data are sent to the State level for inclusion in the statewide registry. Prior to being included, the data go through another round of quality check as well as a consolidation process meant to capture data from multiple facilities that pertain to the same patient and tumor. During this time and before the data are finally added to the State level registry, the data exist only at the institutional level. Yet, after being entered into the statewide registry, the data might be changed as the need arises due to internal quality control or consolidation processes and thus the data at the two levels might again not be identical. While the aforementioned occurrence may be expected from time to time, there is no way for any researcher who asks for the data at either level to know about the possible embedded discrepancies between the two databases. With the passage of time, the data that have been changed at the State level may be fed back into the institutional level so as to finally consolidate them, but this process may take months, or longer, to transpire.

2.2 Blockchain

The blockchain approach was originally conceived for organizing and storing financial transactions (such as Bitcoin) on an open network while ensuring that a specific Bitcoin could not be spent more than once by its owner (Nakamoto, 2008). To realize this, blockchain was developed with mechanisms to provide immutability to the data stored on the system, which in this case were Bitcoin transactions. More simply, the system tracked that a particular coin was transferred from one owner to another, and it ensured that the original owner could not spend the same coin again unless it somehow made its way back to the original spender by way of another transaction.

Figure 2. Example Merkle Tree structure



In the broader context, blockchain relies on the use of cryptographic hashing, distributed data storage, and the use of hash puzzles to provide control over what data are allowed to join the chain of blocks making up the distributed database, or distributed ledger as some refer to the blockchain methodology (Drescher, 2017). Cryptographic hashing is a process by which data are transformed from their original state into an encrypted "hash reference" string. This hash reference has the following useful properties (Drescher, 2017):

Property 1. It is highly unlikely that a hash reference can be reverse engineered into its original form
Property 2. It is highly unlikely that two different data inputs will result in the same hash reference
Property 3. Two very similar inputs will result in substantially different hash reference values

Cryptographic hash functions can be used to process multiple files simultaneously. This property is very useful in that an entire group of documents or other input data can be hashed together to form a single hash reference value that represents the entire set of documents. When multiple datasets are jointly hashed into a single hash reference value, it is known as a Merkle Tree and the single hash value at the top of the tree is known as its root (Drescher, 2017). **Figure 2** is a representation of a Merkle Tree. In the case of a hash created from a record in the cancer registry system, each data item shown in **Figure 2** would be the data element in one field within a record, with the root of the Merkle Tree representing the hash of the values in the fields of the record. Importantly, any change in the underlying data, no matter how small, will change the root of the tree as well as all intermediate hash reference values. This is a critical ingredient supporting the claim of data immutability in blockchain.

As well, in order for a block of data to be added to the chain, it must go through one last hash function, this time including data from the preceding block in the chain and an unknown random number (nonce). This is known as a hash puzzle - to solve the puzzle, the identity of the nonce must be determined that will result in a hash that meets certain criteria (for example, leads with three zeros). Once the encryption is made (and the block is added), the opportunity for an actor to change data without notice in any block in the chain, or the data that go into making the Merkle Trees represented on the chain, is removed. This is because any change to any data represented by hash values on the chain would result in an incorrect hash value due to the aforementioned **Property 3** of a cryptographic

hash function. Thus, any unauthorized change to the data will immediately be noticed as such an attempt would invalidate all subsequent hash values on the chain (Drescher, 2017).

2.3 Related Literature

Given the specific nature and intent of this study, the literature that serves as a foundation for this work stems from three primary areas of inquiry: the general application of blockchain, research on the creation and use of cancer registries, and applications of blockchain technology in the health care domain. The first subsection reveals that the preponderance of research dealing with blockchain; that is, it deals with how blockchain might be generally useful in a wide variety of applications. Even though cancer registries are the primary source of data for literally hundreds of cancer-centered research, the second subsection demonstrates that past efforts to create and structure these data repositories have been limited. As there are no published examples of a blockchain-based cancer registry, the last subsection discusses other applications of blockchain in the health care industry.

2.3.1 Blockchain Research

There is relatively little in the way of empirical research surrounding blockchain; yet, papers exist to help us identify useful areas of application and means of architecture. As the universe of blockchain applications continues to grow, initiatives in one area will undoubtedly impact on others; thus, it is important to identify and understand the application of these technologies beyond the scope of the current project. Firica (2017) highlights the performance restrictions inherent in a high-volume, blockchain-based application. Others provide guidance for the architecture of blockchain-based systems. Xu et al. (2017) outline a comprehensive list of design considerations for blockchain-based systems and provide insights into how multiple chains may be merged effectively. Xu et al. (2016) provide a roadmap for the off-chain storage of data and detail the performance differences between private (such as the solution proposed in this paper) v. public (such as Bitcoin) chains.

Financial applications of blockchain form the largest area of application, mostly centering on cryptocurrencies. The second largest body of work, and the one most relevant to this paper, is that of supply chain (SC) management. Focusing their design on the challenge of provenance determination, Kim & Laskowski (2018) describe an ontology-based system for SC management (SCM). In fact, provenance determination does not greatly differ from the problems involved when keeping the data of various levels of cancer registry data consistent. In order to overcome the performance challenges inherent in a public blockchain, Gao et al. (2018) promote a hybrid blockchain SC model that puts in place a two-step block creation process. O'Leary (2017) proposes a model of a consortium-based private chain environment in which only those firms involved in the SC have access to the chain. By promoting this approach, the author both increases the performance of the database as well as provides a safer data storage environment by allowing only those organizations within the consortium to have access to the chain's data.

2.3.2 Cancer Registry Research

Three areas of research into cancer registries that are of primary interests to the present study include: (a) the creation and use of cancer registries; (b) the collection of data for inclusion in cancer registries; and (c) the accuracy of the data held in these databases. These three areas of inquiry each have a direct impact on the functionality of the registry. The present paper describes a mechanism that works to improve the accuracy of the data held by the cancer registry such that the data are of the highest quality when used for various types of cancer research.

The current cancer registry landscape has been a multi-decade endeavor. Davis, McCarthy & Berger (1999) point out that research prior to the widespread availability of regional and national cancer registries was primarily based on the data collected in clinical and/or institutional settings. Wöhrer et al. (2009) describe the use of the Austrian Brain Tumour Registry (ABTR) in conjunction with the Austrian National Cancer Registry (ANCR) to provide a number of public and scientific

benefits only possible through the use of a comprehensive and widely available tumor registry program. In another example of the power of tumor registry linkage, Bradley et al. (2007) explore a strategy for linking Medicaid, Medicare, and the Michigan Tumor Registry to examine disparities in cancer diagnoses, care quality, and survival.

The quality of the research performed based on any cancer registry is dependent on the data completeness and accuracy (Hall, Schulze, Groome, Mackillop & Holowaty, 2006). Few researchers however evaluated the quality of the data contained in a cancer registry. Ostrom et al. (2016) found that significant variation exists in the identification of site-specific factors (SSF), year of diagnosis, and histologic type between the databases they examined. They point out that some of the reasons for these discrepancies might be that only half of the SSFs are required by SEER guidelines and that some of the data collected pertained to only specific types of cancer and thus were often overlooked when not pertinent to the cancer at hand.

To date, there is a scarcity of research on cancer registries. Most of the studies that mention these registries do so from the perspective that these registries serve only as a tool of their research, not their primary focus. Those researchers that have centered on the quality of these repositories examine the ways in which data can be gathered more efficiently and effectively, but once collected, these researchers are often no longer interested in how the data had been accumulated into the registries. Nonetheless, the fact that data collection and validation take place over a long period of time and that data are often added to the various levels of registries over different time frames bring a great deal of vulnerability that is yet to be addressed by past research.

2.3.3 Blockchain in Health Care

As this may be the first blockchain application in the cancer registry space, this portion of the literature review focuses largely on the use of blockchain in other health care areas.

Data management and protection is critically important to the successful treatment of a patient's disorder and the use of blockchain to safely and accurately handle these data is quickly becoming the focus of much research. This subsection provides an overview of the types of healthcare systems that are being developed using blockchain. Numerous articles have been written extolling the virtues of blockchain technology in the healthcare field (Yue, Huiji, Jin, Li & Jiang 2016; Engelhardt, 2017; Roman-Belmonte, De la Corte-Rodriguez & Rodriguez-Merchan, 2018; Angraal, Krumholz & Schulz, 2017). Alonzo, Arambarri, López-Coronado & de la Torre Diez (2019) provide a comprehensive review of the blockchain literature in health care while Jayaraman, Saleh & King (2019) discuss the potential use of blockchain in the healthcare SC.

A more specific application of blockchain, and perhaps the first likely to be pursued within a commercial environment, are those efforts toward the storage and management of EMRs. Blockchain is an obvious choice for such an application due to its ability to secure data and maintain data immutability even though there may be certain apprehensiveness on the part of some researchers as to its ability to revolutionize the storing of medical records (Pirtle & Ehrenfeld, 2018). Esposito, De Santis, Tortora, Chang & Choo (2018) champion the security benefits of cloud-based blockchain medical record storage. Azaria, Ekblaw, Vieira & Lippman (2016) as well as Fan, Wang, Ren, Li & Wang (2018) have now advocated more specific designs involving the internal workings of a blockchain-based EMR system. As well, Dubovitskaya, Xu, Ryu, Schumacher & Wang (2017) suggest that the security and trustability of blockchain EMRs would make them a likely candidate for sharing data among multiple databases (such as sharing data between the local cancer registry and the State or regional registry. More recently, Chen, Ding, Xu, Zheng & Yang (2019) as well as Kauer, Alam, Jameel, Mourya & Chang (2018) propose a cloud-based system that may be more suited to storing and organizing heterogeneous medical data, which are characteristically prevalent in health care. Finally, Tian, He & Ding (2019) propose an EMR system that promotes the sharing of data using blockchain architecture while Viola (2018) describes how blockchain can be used to store pointers

to data held securely off-chain, thereby saving storage and processing costs, a central feature of the system proposed in the current work.

Altogether, researchers largely see blockchain as a means to improve the security of medical data stored by a healthcare facility. Wang & Song (2018), for example, describe a system that uses both attribute-based and identity-based encryption to encrypt medical data. Griggs et al. (2018) blend the Internet of Things (IoT) and blockchain to develop a tool that uses smart contracts to facilitate the analysis and management of medical sensors and store the data collected in a blockchain environment. Zhou, Wang & Sun (2018) use blockchain to securely store medical insurance data to make these both easily available to insurance companies and to protect the data from those entities (hackers) that should not have access to it. Indeed, blockchain technology is not impervious to attack. The work of Firdaus et al. (2018) offers an approach using practical swarm optimization to detect root exploit malware.

In summary, this brief literature review has demonstrated two important points. First, the cancer registry system fulfills an important role in medical research and its data handling and protection processes suffer from having a number of participants involved, each with their own internal processes and challenges. This often leads to the possibility of inconsistent data and thus incorrect research results. The other revelation is that while not widely implemented, there is great enthusiasm for using blockchain technology in health care, especially because of its ability to maintain the immutability of the data contained within the system. Using these two pointers as inspiration, the rest of this paper details how blockchain may be used to support the storage, organization, and safeguarding of data being submitted into local and regional cancer registries.

3. METHODOLOGY

In addressing the aforementioned research gaps, this paper advocates a Design Science Research (DSR) approach (Hevner, March & Park, 2004). The use of DSR is now well accepted by the IT community and appears in an increasing number of peer-reviewed journal articles (Indulska & Recker, 2010). The DSR approach is chosen for two reasons. First, the lack of research specifying the use of blockchain in the cancer registry space, or any closely related application, provides no solid foundation upon which to analyze the specific applicability of the technology. Second, the DSR provides a framework within which a system can be designed that offers a platform for a particular technology and provides a base to further explore the pairing of a specific technology with a specific area of application. **Table 1** presents the DSR guidelines and how they are unfolded in this paper.

3.1 System Design

As a starting step in designing the system, the full requirements of the organization must be understood and explored. For this reason, representatives of one State's cancer registry system were invited to participate in the study as a source of information about the processes and pitfalls of the existing cancer registry mechanism. Information was also drawn from the documentation published by the applicable governmental agencies and accrediting organizations. Owing to the SEER certification process, examining a single State's processes generalize well to other States; indeed, this generality extends to almost all other US municipalities where cancer registry processes are observed, including those in countries outside of the US such as Canada. Following a number of interactive meetings and a series of email exchanges, the actual process of collecting and evaluating cancer data was uncovered. The results of these information-gathering meetings can be conveniently displayed within the framework of a traditional SWOT (Strengths, Weaknesses, Opportunities, and Threats) analysis as presented in **Table 2**.

For generating alerts on data inconsistencies found in the institutional v. State level registries, the design uses the concept of a blockchain to keep track of the current version of the data on the institutional database. This current version is kept on a blockchain-based distributed storage network such that all parties with access to the data can compare the version of the data at the institutional level

Table 1. Design-science research guidelines (Hevner, March, & Park 2004)

1. Design as an Artifact	This study provides a set of models that describe the system's basic architecture as well as its functionality.
2. Problem Relevance	The relevance of the problem is demonstrated in the Introduction as well as the literature review.
3. Design Evaluation	The design is evaluated within the organization it is developed for in terms of its ability to overcome the problems set forth in the Introduction.
4. Research Contributions	This research contributes in its novel application of an emerging technology (blockchain).
5. Research Rigor	The rigor of this study rests in its grounding in current, well documented, organizational processes and the level of scrutiny applied by members of the target organization who carry out those processes.
6. Design as a Search Process	The design that results from this study is the product of a search for an appropriate technology to improve a process that contributes to a field of research with significant societal impact (i.e. cancer research).
7. Communication of Research	The results of this research will be communicated through its publication in an appropriate research outlet.

Table 2. SWOT Analysis Results

Strengths	As described in both the SOP and by the personnel involved, the system makes significant efforts to identify all incidents of cancer within its jurisdiction and include these data in the registry.
Weaknesses	Data contained at the local and state levels are thought to be identical, but because of the organizational processes in place might not be. Researchers using these data are not alerted to the potential differences.
Opportunities	The primary contributor to the data mismatch that occurs between the institutional and state level registries is organizational rather than technological and therefore beyond a technological solution. However, the use of an appropriate technology can enable the researcher or any data steward to understand which records in a particular registry have been reconciled with the other registry.
Threats	The threat that we are concerned about for this project is the use of data that is perceived to be ready for research, when in fact it is still not completely reconciled. For this project, this can happen in two ways.

with that at the State level. If the comparison shows agreement, the data can be used in its current state; otherwise, the researcher can try to ascertain the nature of the differences.

This paper calls for adopting the Xu et al. (2016) framework as a guide to the creation of a blockchain-enabled system to address the key design challenging issues. This framework poses a set of design trade-offs or questions that must be decided in order to create a system that best utilizes blockchain technology to solve a design problem (**Table 3**). These steps represent the flexibility inherent within blockchain, which allows the application to be customized for use in an environment

Table 3. Design decision tradeoffs (Xu et al, 2016)

Blockchain Design Decision 1	Mechanisms of improving transaction processing rate.
Blockchain Design Decision 2	Mechanisms of selecting the next block added to the blockchain.
Application Design Decision 1	Scope: On-chain vs. off-chain
Application Design Decision 2	Public chain vs. private chain
Application Design Decision 3	Single chain vs. multiple chains
Application Design Decision 4	External validation vs. internal validation
Application Design Decision 5	Permissionless vs. permissioned blockchain

such as the cancer registry. There is no requirement that these decisions be addressed in the order listed, but they identify the aspects of a blockchain design that should be considered in order to maximize its potential to solve organizational problems.

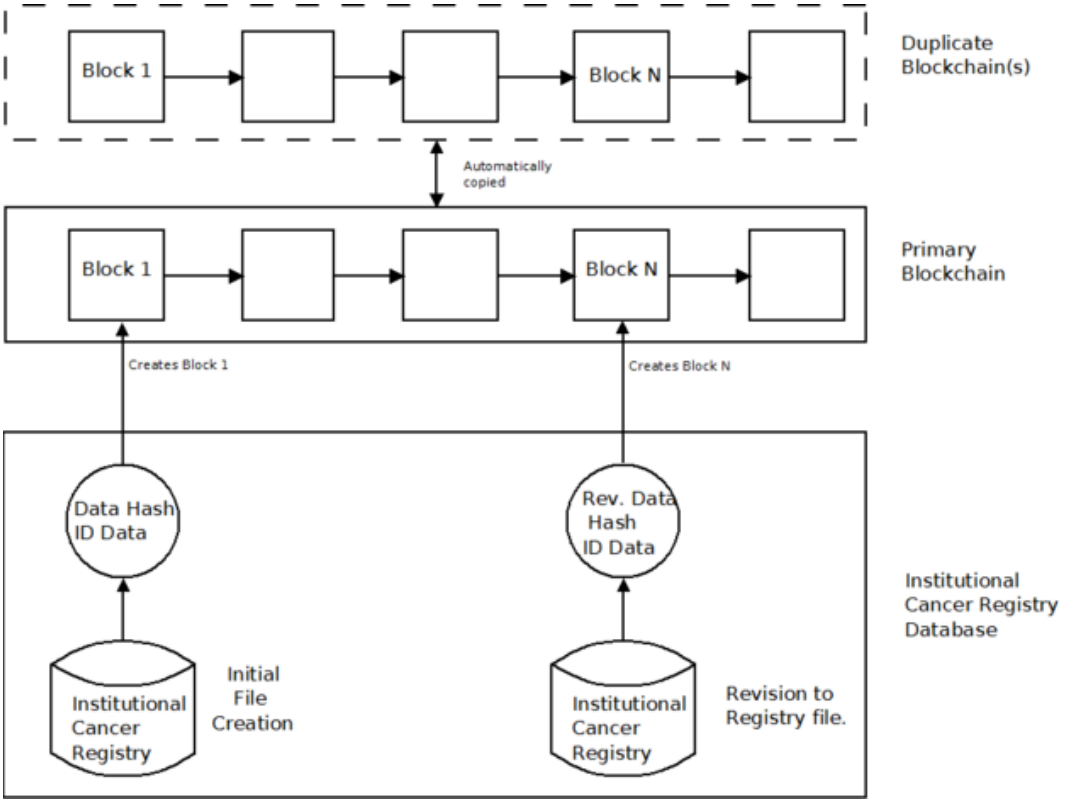
As data are added to each institutional registry, they are hashed to create a Merkle Tree and the root of the tree is added to the blockchain. Each time the data are changed or added to, they are re-hashed and a new block is added to the chain representing the revised data. While not all data items need to be hashed and added to the chain, certainly those that are important to research and patient/tumor identification would be added. Concurrently, as data are deposited in the State level registry, they are also hashed using the same algorithm to create a Merkle Tree comprising the same data points as those used at the institutional level. Since the data at the State level may have been derived from multiple sites (as determined by the patient's treatment path), a tree will be created for each contributing institution containing only those data collected by that specific institution. When data are extracted from either the State or institutional registry, the resulting hash value is then compared with the corresponding hash value at the matching database; if the value matches, the data are consistent across registries. Otherwise, an unmatched hash value signals that the data at the two registries do not match and the researcher must exercise caution.

The first decision for the system designer to undertake is to decide which data to store on the chain and which to store off-chain (*Application Design Decision 1*). There are many good reasons to store the data off-chain rather than on the chain. For instance, by storing only the root value of the Merkle Tree that is comprised of the registry data, the data will be protected from unauthorized or unnecessary viewing and the speed by which blocks are added to the chain will be greatly improved. Thus, the data contained in the registry is never stored on the blockchain, but only the hash value of the Merkle Tree as well as an identifier for that particular file. The identifier will actually consist of a group of fields because of the complex nature of cancer and cancer registries.

The initial configuration of the file identifier comprises:

- Accession Number: A nine-digit number including year of first contact and the order in which the file was added to the registry
- Medical Record Number: The medical record could span multiple visits and/or illnesses but it forms a mechanism for collecting multiple reports
- Patient Control Number: Assigned at admission, sometimes linked to medical record. Does not designate a specific illness or treatment
- Permanent Facility Number: Identifies specific physical location of treatment
- Social Security Number: Identifies patients in the United States uniquely
- Primary Site Code: Locates the specific site where a primary tumor (and secondary tumor, if applicable) is located; uses ICD-O-3.1 codes (International Classification of Diseases for Oncology version 3.1) as the current standard (NYS Dept of Health, 2017)

Figure 3. Overview of cancer registry blockchain design



Each of the above data item is considered a mandatory field in a cancer registry, though each might not always be available (e.g. Social Security Number) or they might change over time as diagnoses proceed from preliminary to final. Moreover, it is hoped that having as many identifiers as possible will allow for positive identification of the appropriate file. With this important design issue dealt with, the basic system design is presented in **Figure 3**.

Figure 3 shows the basic workings of the proposed blockchain-based cancer registry system along the perspective of the institutions collecting and storing the data. The left of **Figure 3** shows that the initial file creation takes place upon the collection and input of pertinent data by the cancer registrar at the facility. Those fields crucial to cancer research, as determined by various cancer research organizations, are combined into a Merkle Tree, the root of which is stored on the blockchain along with the identification data noted above (Block 1). When more data are collected or a revision of some sort occurs, the registry data are updated and the revised Merkle Root is used to create a new block on the chain (Block n). Simultaneously, when data are added at the State level, the cumulative data are hashed and the hash reference is stored with the data. When data are requested for research from the State level, the root value of those data can be compared to the root value of the corresponding data on the blockchain to determine the level of agreement. For data requested at the local level, the corresponding hash value can be compared with those for the same data at the State repository.

With the basic concept of the blockchain determined, other design decisions can be addressed. *Application Design Decision 3* requires a determination of the number of chains in the system. The original blockchain definition called for an unlimited number of publicly held chains (Nakamota, 2008), which is not necessary for the present purpose. However, having multiple chains to be held by multiple stakeholders would provide a measure of security by creating a controlled amount of

redundancy. Thus, multiple chains could be maintained by major institutions in the State as well as the State itself, each contributing a small amount of computing resources to the cause of higher quality cancer data (*Application Design Decision 2*). A single chain held at a member institution or at the State level would work as well, but it would not provide the safety of having data stored at each institution as a hedge against system or communications failure. The final decision on this aspect of the system would have to be made with reference to the current funding atmosphere. Most funding for the cancer registry at the institution level has been provided "somewhat grudgingly" by the institution with no outside subsidies.

The method of selecting the next block added to the chain (*Blockchain Design Decision 2*) is simplified by taking the data in the order that they are created. The method for improving the speed at which the blocks are added (*Blockchain Design Decision 1*) centers on the difficulty of the hash puzzle. As a private chain, the need to protect the integrity of the data is reduced, but there still needs to be a method to ensure that changed data can be detected. Thus, a simplified hash puzzle (often known as "proof of work") can be employed.

The last two Application Design Decisions (*4 & 5*) ask the designer to deal with the amount of access granted to the "owners" of the blockchain and whether the transactions that take place on the blockchain require an external entity to validate them. To the first point, the design considered here is a permissioned blockchain (Drescher, 2017) in that only those entities (in this case, member institutions and the State) have permission to add or interrogate blocks on the blockchain. Only these institutions have the legal access to the data. Indeed, there is no need for others to have any access to the distributed data. This approach is extended to the need for external validation. The transactions involved on this blockchain are only those that add data to the chain. Unlike bitcoin in which financial transactions are recorded and represent the transfer of wealth (bitcoins), no value is created on this chain and therefore there is no need to make sure we are "keeping score" properly. In the current system, there is no external oversight of data collected and stored except for the accrediting bodies involved and their oversight comes only when the facilities are audited.

4. RESULTS

A full description of the proposed system was provided to the participants from the cancer registries at both the State and institutional levels including the mechanism by which the differences in the data between the various repositories will be highlighted to the researcher. As the inspiration for this system was not to improve transactional performance, but to provide a mechanism to alert potential data users of embedded data discrepancies due to the organizational processes taking place, it appears that the most appropriate means of evaluating the system is to utilize the descriptive method suggested by Hevner, March & Park (2004). In the end, the possibility of installing a prototype system was dismissed due to regulations governing the use of patient data. Moreover, the study participants were more concerned over the organizational aspects of the system than emphasizing the user perspective. The participants were asked to evaluate the system based on its performance (i.e. effectiveness at providing alerts), organizational processes, and cost/value.

The participants were presented with the processes of the system so that they fully understood its impacts on the various user groups (cancer registry personnel, researchers, and administrators). As the hash reference comparison process is performed in the background during the data extraction process, there is no additional burden on either the data entry or data quality personnel. Moreover, as these data are being provided for research purposes and not for use in a clinical setting, the slight potential delay to accommodate the comparison process is not perceived as a performance issue. The background nature of the hash comparison process also removes any doubts about complicating the data access process for the researcher, except that the researcher will have to decide what to do with those records flagged as being inconsistent across the local and statewide databases. In the opinion of the participants, the researcher will have one of three options: to use the data anyway, to discard

the data, or to investigate the reason for the flag. Each of the aforementioned options leads down a different path:

- To use the data anyway would indicate that the data items flagged as inconsistent are not pertinent to the research at hand.
- To discard the data is probably the easiest course of action to take except in cases where the population is so small that the loss of even a single observation would negatively impact the study's validity.
- This leads to the third option, which would likely be a laborious and time-consuming process, especially considering that most research data are stripped of most Personally Identifiable Information (PII) before being provided to the researcher. It was unclear to the participants whether that option would ever prove to be viable given the funding issues surrounding most cancer research as well as the time pressure that often accompanies this type of work.

There was wide consensus that there would be an initial hesitancy on the part of both the institution and the State to undertake a program of this type, though the source of that hesitation varied among the participants. The participants from both the hospital and the State had an immediate negative reaction to the security aspects of the system. These fears were alleviated once they understood fully the type of data being stored on the chain and that all clinical data would be stored in the current manner. There was also concern over a loss of control of one's data. These concerns were relieved when shown that this system serves solely to compare the hash values of data stored on both levels of registry and has no mechanism to revise or remove, or even read, data. All data transactions including initial file creation, revision, or consolidation is done through the current processes only.

The last area to discuss is the cost, and again the participants each concluded that any additional cost would not be easily tolerated. Still, the system was designed with cost minimization at its heart. First, the size of the data to be stored on the blockchain database is minimal and imposes only a very small storage increase. Likewise, there was concern about the computational burden to be placed on the institution's computing resources during the solution of the hash puzzles required to add data to the chain. Again, because of the relaxed hash puzzle burden imposed on the blockchain, the increased computational burden, and therefore the stress on the system and increased energy cost, is negligible. Moreover, this system could easily be hosted on a cloud provider because it contains limited PII and most medical cloud providers are now HIPAA (Health Insurance Portability and Accountability Act) compliant. In many cases, cancer registry storage is already cloud-based.

A study participant from the State cancer registry identified a potential for substantial cost savings. During the usual quality control processes, data at both levels are checked for consistency and accuracy. However, due to organizational limitations, only a sample of the total data held is actually analyzed. The representative from the State pointed out that with this system in place, there would be an assurance that a substantial portion of the data at both levels was already consistent, thus greatly reducing the population from which to sample quality assurance data and thereby reducing the costs associated with quality control activities. At the same time, she pointed out, the efficacy of the quality program would increase as a higher percentage of inconsistent cases could be evaluated within the organizational and budgetary constraints currently in place.

5. DISCUSSION

The initial skepticism of the study participants is completely understandable. They are part of an industry that depends on data to make very important decisions and any change in the processes by which data are collected or managed is usually met with caution. However, as a fuller understanding of the system and its underlying concepts emerged, their resistance subsided. While issues of cost

and institutional control still exist, the study participants were eventually convinced of the system's efficacy in providing a check on data inconsistency.

Any marginal costs incurred in the implementation of this system would be shared across all participating institutions, including the State. This is a common practice in many industries as, to promote the individual firm's success, firms on a single SC find it advantageous to contribute to the overall benefit of the SC (Lambe, Wittman & Spekman, 2001). The Social Exchange Theory (SET) suggests that individual buyers and vendors will cooperate with each other and try to resolve SC problems as long as the outcomes are beneficial to both parties. There is a long history of firms having fewer but closer vendors (Seydel, 2005) and of vendors being asked to provide feedback on the buyer's operations in order to improve their functionality. Buyers also recognize that vendors often have knowledge and expertise that will allow them to improve their operations (Wong, Tjosvold & Zhang, 2005). In the case of the cancer registries, these same concepts apply except that instead of a profit motive as is the case with many other industries, it is the mission of cancer care organizations to contribute to the continued cancer research success that is paramount. Such intent has been codified in the mission statement of the SEER program, which states that their goal is to "improve the quality and completeness of cancer information" (NIH). At its core, the suggested system design approach represents a quality improvement tool, a tool that can increase the level of trust between the researcher and the registry data. Building trust between entities in an information intensive environment such as this is one of the major strengths of a blockchain-based architecture (Turk & Kline, 2017).

In retrospect, it is also useful to review this relationship via the lens of Deutsch's (1973, 1980) theory of cooperation and competition. Here, Deutsch points out that organizations, and people, interact in different ways depending on the nature of their relationships. In a cooperative setting such as that between the various hospital registries and the State, there is a much higher likelihood that a parallel blockchain arrangement as described herein will succeed. In this type of arrangement, the partnering organizations will act in concert and contribute or share resources that will eventually result in a stronger, and therefore more successful, product. In contrast, for a competitive setting such as systems that are to be developed within a commercial setting, these concepts will be applicable directly to the present situation by substituting "cancer research" for "product."

6. CONCLUSION

This paper highlights the use of blockchain technology to evaluate the level of consistency between the data contained in the cancer registry of an institution and that of the State's registry. Owing to due processes required to maintain as well as increase the quality of the data contained in both registries, there might be a significant period when the data contained in one registry may be different than those contained in the other. As there is no way for a researcher, or any other users, to know if such inconsistencies exist for a particular batch of data, there is a chance that incorrect or obsolete data might be used in cancer research. The solution proposed here is to adopt a mechanism so as to verify whether the data extracted from one source are the same as the data at the other.

The contributions of this work are twofold. First, it demonstrates how blockchain technology might be applied to an environment beyond financial transactions. The immutability of data housed on a blockchain proves to be an ideal mechanism to ensure the validity of data given that a control source is not always available. In the present case, this system cannot vouch for the accuracy of the data contained in the record (there are other processes that do that) but it can alert the user to the fact that two databases do not agree when, in fact, they are believed to contain the very same data or datasets. Understanding the limitations of data is the first step to improving any data collection systems.

The second significant contribution is to apply theories developed for a commercial setting (e.g., SET and Deutsch's theory of cooperation and competition) to a situation that may not be driven by profit. The need to provide accurate data for cancer research links the participants together in the cancer registry program (which can be thought of as a data SC) and that can substitute for the profit

motive that these theories were developed to address. Further research is necessary to determine the extent to which such parallel thinking is applicable.

Finally, the breadth of the present study is limited to a single institution's registry and a single State's registry, and while the organizational procedures, and thus the technical processes, would be similar across the different registries within the same country, different geographical and cultural environments might have the effect of modulating the acceptance of these registry systems in other locations. Hence, future studies are still needed to assess the practicality of employing the proposed type of system in changing contexts.

REFERENCES

Alonzo, S. G., Arambarri, J., López-Coronado, M., & de la Torre Diez, I. (2019). Proposing new blockchain challenges in eHealth. *Journal of Medical Systems*, *43*(3), 64. doi:10.1007/s10916-019-1195-7 PMID:30729329

Angral, S., Krumholz, H. M., & Schulz, W. L. (2017). Blockchain technology applications in healthcare. *Circulation: Cardiovascular Quality and Outcomes*, *10*(9), 1. doi:10.1161/CIRCOUTCOMES.117.003800

Azaria, A., Ekblaw, A., Vieira, T., & Lippman, A. (2016). MedRec: Using blockchain for medical data access and permission management. *Proceedings of the 2nd International Conference on Open and Big Data,* 25-30. doi:10.1109/OBD.2016.11

Bradley, C. J., Given, C. W., Luo, Z., Roberts, C., Copeland, G., & Virnig, B. A. (2007). Medicaid, Medicare, and the Michigan Tumor Registry: A Linkage Strategy. *Medical Decision Making*, *352*(July), 352–363. doi:10.1177/0272989X07302129 PMID:17641138

Chen, Y., Ding, S., Xu, Z., Zheng, H., & Yang, S. (2019). Blockchain-based medical records secure storage and medical services framework. *Journal of Medical Systems*, *43*(1), 5. doi:10.1007/s10916-018-1121-4 PMID:30467604

Davis, F. G., McCarthy, B. J., & Berger, M. S. (1999, July). Centralized databases available for describing primary brain tumor incidence, survival, and treatment: Central Brain Tumor Registry of the United States; Surveillance, Epidemiology, and End Results; and National Cancer Data Base. *Neuro-Oncology*.

Deutsch, M. (1973). *The Resolution of Conflict*. Yale University Press. doi:10.1177/000276427301700206

Deutsch, M. (1980). Fifty years of conflict. In L. Festinger (Ed.), *Retrospections on Social Psychology*. Yale University Press.

Drescher, D. (2017). *Blockchain Basics*. Apress. doi:10.1007/978-1-4842-2604-9

Dubovitskaya, A., Xu, Z., Ryu, S., Schumacher, S., & Wang, F. (2017). Secure and trustable electronic medical records sharing using blockchain. *AMIA Annual Symposium Proceedings,* 650-659.

Englehardt, M. A. (2017). Hitching healthcare to the chain: An introduction to blockchain technology in the healthcare sector. *Technology Innovation Management Review*, *7*(10), 22–34. doi:10.22215/timreview/1111

Esposito, C., De Santis, A., Tortora, G., Chang, H., & Choo, K. R. (2018). Blockchain: A panacea for healthcare cloud-based data security and privacy? *IEEE Cloud Computing*, *5*(Jan), 31–37. doi:10.1109/MCC.2018.011791712

Fan, K., Wang, S., Ren, Y., Li, H., & Yang, Y. (2018). MedBlock: Efficient and secure medical data sharing via blockchain. *Journal of Medical Systems*, *42*(8), 136. doi:10.1007/s10916-018-0993-7 PMID:29931655

Firdaus, A., Anuar, N. B., Ab Razak, M. F., Hashem, I. A. T., Buchok, S., & Sangaiah, A. K. (2018). Root exploit detection and features optimization: Mobile device and blockchain based medical data management. *Journal of Medical Systems*, *42*(6), 112. doi:10.1007/s10916-018-0966-x PMID:29728780

Firica, O. (2017). Blockchain technology: Promises and realities of the year 2017. *Quality - Access to Success*, *18*(S3), 51–58.

Gao, Z., Xu, L., Chen, L., Zhao, X., Lu, Y., & Shi, W. (2018). CoC: A unified distributed ledger based supply chain management system. *Journal of Computer Science and Technology*, *33*(2), 237–248. doi:10.1007/s11390-018-1816-5

Griggs, K., Ossipova, O., Kuhlios, P., Baccarini, A. S., Howson, E. A., & Hayajneh, T. (2018). Healthcare blockchain system using smart contracts for secure automated remote patient monitoring. *Journal of Medical Systems*, *42*(7), 130. doi:10.1007/s10916-018-0982-x PMID:29876661

Hall, S., Schulze, K., Groome, P., Mackillop, W., & Holowaty, E. (2006). Using cancer registry data for survival studies: The example of the Ontario Cancer Registry. *Journal of Clinical Epidemiology*, *59*(1), 67–76. doi:10.1016/j.jclinepi.2005.05.001 PMID:16360563

Hevner, A. R., March, S. T., Park, J., & Ram, . (2004). Design science in information science research. *Management Information Systems Quarterly*, *28*(1), 75–105. doi:10.2307/25148625

Indulska, M., & Recker, J. (2010). Design science in IS research: A literature analysis. In S. Gregor (Ed.), *Information Science Foundations: The Role of Design Science* (p. 285). ANU E Press. doi:10.22459/ISF.12.2010.13

Jayaraman, R., Saleh, K., & King, N. (2019). Improving opportunities in healthcare supply chain processes via Internet of Things and blockchain technology. *International Journal of Healthcare Information Systems and Informatics*, *14*(2), 49–65. doi:10.4018/IJHISI.2019040104

Kauer, H., Alam, M. A., Jameel, R., Mourya, A. K., & Chang, V. (2018). A proposed solution and future direction for blockchain-based heterogeneous Medicare data in cloud environment. *Journal of Medical Systems*, *42*(8), 156. doi:10.1007/s10916-018-1007-5 PMID:29987560

Kim, H. M., & Laskowski, M. (2018). Toward an ontology-driven blockchain design for supply-chain provenance. *Intelligent Systems in Accounting, Finance & Management*, *25*(1), 18–27. doi:10.1002/isaf.1424

Lambe, C. J., Wittman, C. M., & Spekman, R. E. (2001). Article. *Journal of Oncology Practice / American Society of Clinical Oncology*, *7*(2), 111–116. doi:10.1200/JOP.2010.000097

Nakamoto, S. (2008). *Bitcoin: A peer-to-peer electronic cash system*. www.bitcoin.org/bitcoin.pdf

National Institutes of Health. (n.d.). *Surveillance, Epidemiology, and End Results.* https://seer.cancer.gov/

New York State Department of Health. (2017). *The New York State Cancer Registry: Facility Reporting Manual*. Author.

New York State Department of Health. (2018). *About the New York State Cancer Registry.* https://www.health.ny.gov/statistics/cancer/registry/about.htm

O'Leary, D. E. (2017). Configuring blockchain architectures for transaction information in blockchain consortiums: The case of accounting and supply chains. *Intelligent Systems in Accounting, Finance & Management*, *24*(4), 138–147. doi:10.1002/isaf.1417

Ostrom, Q. T., Gittleman, H., Kruchko, C., Louis, D. N., Brat, D. J., Gilbert, M. R., Petkov, V. I., & Barnholtz-Sloan, J. S. (2016). Completeness of required site-specific factors for brain and CNS tumors in the Surveillance, Epidemiology and End Results (SEER) 19 database (2004-2012, varying). *Journal of Neuro-Oncology*, *130*(1), 31–42. doi:10.1007/s11060-016-2217-7 PMID:27418206

Pirtle, C., & Ehrenfeld, J. (2018). Blockchain for healthcare: The next generation of medical records? *Journal of Medical Systems*, *42*(8), 172. doi:10.1007/s10916-018-1025-3 PMID:30097733

Roman-Belmonte, J. M., De la Corte-Rodriguez, H., & Rodriguez-Merchan, E. C. (2018). How blockchain technology can change medicine. *Postgraduate Medicine*, *130*(4), 420–427. doi:10.1080/00325481.2018.1472996 PMID:29727247

Seydel, J. (2005). Supporting the paradigm shift in vendor selection: Multicriteria methods for sole-sourcing. *Managerial Finance*, *31*(3), 49–66. doi:10.1108/03074350510769569

Tian, H., He, J., & Ding, Y. (2019). Medical data management on blockchain with privacy. *Journal of Medical Systems*, *43*(2), 26. doi:10.1007/s10916-018-1144-x PMID:30603816

Turk, Ž., & Klinc, R. (2017). Potentials of blockchain technology for construction management. *Proceedings of the Creative Construction Conference (CCC 2017)*. doi:10.1016/j.proeng.2017.08.052

United States Congress. (2009). *H.R.1 - American Recovery and Reinvestment Act of 2009.* https://www.congress.gov/bill/111th-congress/house-bill/1

United States Congress. (2009). *American Recovery and Reinvestment Act of 2009*. https://www.congress.gov/bill/111th-congress/house-bill/1/text

Viola, A. (2018). Blockchain's role in health IT. *Journal of American Health Information Management Association*, *89*(9), 34–35, 54.

Wöhrer, A., Waldhör, T., Heinzl, H., Hacki, M., Feichtinger, J., Gruber-Mösenbacher, U., & Hainfeliner, J. A. et al. (2009). The Austrian Brain Tumor Registry: A cooperative way to establish a population-based brain tumor registry. *Journal of Neuro-Oncology*, *95*(3), 401–411. doi:10.1007/s11060-009-9938-9 PMID:19562257

Wang, H., & Song, Y. (2018). Secure cloud-based HER system using attribute-vased cryptosystem and blockchain. *Journal of Medical Systems*, *42*(8), 152. doi:10.1007/s10916-018-0994-6 PMID:29974270

Wong, A., Tjosvald, D., & Zhang, P. (2005). Developing relationships in strategic alliances: Commitment to quality and cooperative independence. *Industrial Marketing Management*, *34*(7), 722–731. doi:10.1016/j.indmarman.2004.12.007

Xu, X., Pautasso, C., Zhu, L., Gramoli, V., Ponomarev, A., Tran, A. B., & Chen, S. (2016). The blockchain as a software connector. *Proceedings of the 2016 13th Working IEEE/IFIP Conference on Software Architecture.* doi:10.1109/WICSA.2016.21

Xu, X., Webeer, I., Staples, M., Zhu, L., Bosch, J., Bass, L., & Rimba, P. et al. (2017). A taxonomy of blockchain-based systems for architecture design. *Proceedings of the 2017 IEEE International Conference on Software Architecture.* doi:10.1109/ICSA.2017.33

Yue, X., Huiju, W., Jin, D., Li, M., & Jiang, W. (2016). Healthcare data gateways: Found healthcare intelligence on blockchain with novel privacy risk control. *Journal of Medical Systems*, *40*(10), 218. doi:10.1007/s10916-016-0574-6 PMID:27565509

Zhou, L., Wang, L., & Sun, Y. (2018). MIStore: A blockchain-based medical insurance storage system. *Journal of Medical Systems*, *42*(8), 149. doi:10.1007/s10916-018-0996-4 PMID:29968202

Joseph Kasten is an Assistant Professor of Information Science and Technology at the Pennsylvania State University in York, PA. He earned a PhD in Information Science at Long Island University, an MBA at Dowling College, and a BS in Engineering at Florida Tech. Before joining academia, Joe was a senior engineer with Northrop-Grumman where he worked on various projects such as the X-29, Space Shuttle, and the Boeing 777. His research interests center on the implementation of data analytics within the organization as well as the application of blockchain technology to emerging organizational requirements. Professor Kasten's recent research has appeared in the American Journal of Business and Industrial Management and International Journal of Strategic Information Technology and Applications, as well as a number of book chapters.