Published in final edited form as:
J Geogr Syst. 2017 July ; 19(3): 197–220. doi:10.1007/s10109-017-0252-3.

Geospatial cryptography: enabling researchers to access private, spatially referenced, human subjects data for cancer control and prevention

Geoffrey M. Jacquez[1,2], **Aleksander Essex**[3], **Andrew Curtis**[4], **Betsy Kohler**[5], **Recinda Sherman**[5], **Khaled El Emam**[6], **Chen Shi**[1], **Andy Kaufmann**[2], **Linda Beale**[7], **Thomas Cusick**[8], **Daniel Goldberg**[9,10], and **Pierre Goovaerts**[2]

[1]Department of Geography, State University of New York at Buffalo, Buffalo, NY, USA

[2]BioMedware, Ann Arbor, MI, USA

[3]Department of Electrical and Computer Engineering, Western University, London, ON, Canada

[4]Department of Geography, Kent State University, Kent, OH, USA

[5]North American Association of Central Cancer Registries, Springfield, IL, USA

[6]Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada

[7]Esri, Redlands, CA, USA

[8]Department of Mathematics, University at Buffalo, Buffalo, NY, USA

[9]Department of Geography, Texas A&M University, College Station, TX, USA

[10]Department of Computer Science & Engineering, Texas A&M University, College Station, TX, USA

Abstract

As the volume, accuracy and precision of digital geographic information have increased, concerns regarding individual privacy and confidentiality have come to the forefront. Not only do these challenge a basic tenet underlying the advancement of science by posing substantial obstacles to the sharing of data to validate research results, but they are obstacles to conducting certain research projects in the first place. Geospatial cryptography involves the specification, design, implementation and application of cryptographic techniques to address privacy, confidentiality and security concerns for geographically referenced data. This article defines geospatial cryptography and demonstrates its application in cancer control and surveillance. Four use cases are considered: (1) national-level de-duplication among state or province-based cancer registries; (2) sharing of confidential data across cancer registries to support case aggregation across administrative geographies; (3) secure data linkage; and (4) cancer cluster investigation and surveillance. A secure multi-party system for geospatial cryptography is developed. Solutions under geospatial cryptography are presented and computation time is calculated. As services provided by cancer

Correspondence to: Geoffrey M. Jacquez.

Compliance with ethical standards

Conflict of interest The authors declare no potential conflicts of interest.

registries to the research community, de-duplication, case aggregation across administrative geographies and secure data linkage are often time-consuming and in some instances precluded by confidentiality and security concerns. Geospatial cryptography provides secure solutions that hold significant promise for addressing these concerns and for accelerating the pace of research with human subjects data residing in our nation's cancer registries. Pursuit of the research directions posed herein conceivably would lead to a geospatially encrypted geographic information system (GEGIS) designed specifically to promote the sharing and spatial analysis of confidential data. Geospatial cryptography holds substantial promise for accelerating the pace of research with spatially referenced human subjects data.

Keywords

Geospatial cryptography; Geographic information science; Spatial methods; Human subjects research; Privacy

1 Introduction

Access to spatially referenced health data is vital for numerous reasons, including being able to share and combine information; identify patterns in the search for etiology; and logistical support and planning. Quite understandably, this access continues to cause concerns among data guardians due to the consequences of privacy breaches or the careless release of sensitive information. This concern has increased as the precision and availability of other geographic data make recreating the identity of individuals from even partial information a possibility. Researchers in mathematical cryptography have completed a substantial body of confidentiality-related work to identify possible solutions and to quantify re-identification risk (Aslett et al. 2015; Kim et al. 2013). But to date, no universally accepted best practices guide has been adopted, and mathematically proven cryptographic techniques for solving confidentiality and security concerns when handling individual-level data from our nation's cancer registries have not been implemented, even though the potential benefits are high. It has been suggested that a factor explaining this gap has been the search for a generally applicable mathematical solution. However, it may be more productive to consider niche-specific solutions rather than universal methods or guidelines that are supposed to cover all eventualities. This article therefore focuses on the eventual development of location-specific best practices for application in cancer registries. We begin this effort by considering four types of situations (use cases) in which the precise location of an individual is required to be known to correctly complete a research model in epidemiological research in cancer research.

The recent literature on spatial confidentiality involves geographic masking and related techniques that seek to obscure geographic coordinates and/or analysis results to the minimum extent needed to prevent re-identification, the control of data access or securing the transmission of data (Richardson et al. 2015; Zandbergen 2014). All these approaches have limitations on a continuum ranging from the usefulness to researchers to the effectiveness of reducing confidentiality breaches. An additional consideration is the logistical ease of implementation; a solution may theoretically be perfect, but if it is unlikely for either the data guardian or end user to implement, then it has applied limitations. Here we consider spatial confidentiality problems encountered by cancer registries. We introduce geospatial cryptography, which we define as the 'specification, design, implementation and application of mathematical and cryptographic techniques to address privacy, confidentiality and security concerns that arise in the analysis of spatially referenced data on individuals'. In this article we assume both data guardian and data user have the technical ability to perform the required manipulations. We also assume the data user understands the types of tasks to be performed and is not expecting to perform an exploratory analysis that would involve multiple data queries, the addition of data from external sources and data visualizations involving linked windows and statistical and cartographic brushing. Relaxation of these assumptions is considered in the Sect. 4.

To illustrate that there is a need for such niche confidentiality solutions, we consider the use of geospatial cryptography in cancer control and surveillance by considering four use cases: (1) national-level de-duplication among state or province-based cancer registries; (2) Sharing of confidential data across cancer registries to support case aggregation across administrative geographies; (3) secure data linkage; and (4) cancer cluster investigation and surveillance. These are considered in order of difficulty, concluding with geospatial cryptographic analysis for cancer clustering. This most difficult use case employs homomorphic encryption, with data analysis occurring within a "black box" such that the analyst does not have direct access to the data (Fig. 1, to be discussed in more detail later), yet still allowing for *niche suitable* advanced spatial analysis, such as a space–time cluster analysis of residence geocoded incident cases.

Protection of confidentiality is a paramount concern in human subjects research, yet limits data sharing and access to the very information that is required to accomplish significant, rapid advances in public health (Wartenberg and Thompson 2010). This problem spans the research continuum, from basic discovery to translational research (VanWey et al. 2005). Both epidemiological and laboratory studies may employ data linkage using individual identifiers such as name and address (Rushton et al. 2006; Verykios et al. 2009) to link health data with Medicare and other government databases (National Research Council 2007). Common linking techniques require identifying information to facilitate high match rates, but this is often not permissible for privacy reasons. Protocols for confidentiality protection vary across government agencies and research organizations, with some requiring that researchers visit and undertake their analyses within a secure facility limiting access to data. In other instances, data custodians only release de-identified information such that individual identifiers cannot be reconstructed, while others rely on aggregation to protect confidentiality (Abowd and Lane 2004; Fefferman et al. 2005; Ciriani et al. 2010).

De-identification can result in a loss of precision (due to generalization and aggregation) and statistical power (due to suppression). For geospatial analysis, the problem is to protect confidentiality while using geographic information (e.g., coordinates of place of residence) to undertake cluster analysis, spatial regressions, spatial epidemiology, exposure reconstruction, healthcare access evaluation and place-specific health disparities analysis or any analysis that involves geographic locations from health records. In spite of the necessity of the sharing and analysis of health data, restrictive institutional protocols are increasingly employed to strictly limit data access, with some organizations going so far as to decline to release individual-level data under any circumstance. Yet data sharing is at the very heart of the scientific method; it makes possible validation studies to duplicate results, and the lack of ready access to high-precision federal vital records data makes many etiologic studies difficult if not impossible. Theoretically, the ability of geospatial cryptography to share and analyze confidential data with a minimal risk of releasing private information would be a major advance that would accelerate both basic and applied research. Recent advances in secure multi-party computation and homomorphic cryptography make possible analysis of encrypted data without decryption (Gentry 2009; Smart and Vercauteren 2010), but have yet to be applied in geography. Under these approaches, confidential data would not have to be decrypted to allow analysis and dissemination of the results. However, current limitations to this approach include the widespread usage across all situations and the required technical expertise. The NLM funded project titled "Exploratory Evaluation of Homomorphic Cryptography for Confidentiality Protection" (GMJ Principal Investigator) investigated the feasibility of using homomorphic cryptography in geospatial health analysis of confidential data (http://projectreporter.nih.gov/project_info_description.cfm?projectnumber=1R21LM011132-01A1), some of the results of which are reported in this article.

While etiologic investigations and exposures reconstructions in small geographic areas are realizing substantial insights using local information (Bell et al. 2006; Meliker et al. 2009), the danger of reengineering actual or approximate addresses from analytical output is a significant concern (Curtis et al. 2006a; Gutmann et al. 2008; Wieland et al. 2008; Boulos et al. 2009). Conducting space–time analyses using encrypted geographic coordinate information could make possible individual-level spatial and spatiotemporal epidemiological analyses without compromising individual privacy. *Geospatial cryptography* may fundamentally transform how we access, share and analyze confidential geospatial health data, within a predetermined framework of likely need, expected utility and technical ability. As such, this paper is organized around the four use cases introduced earlier that have been structured to illustrate the benefits and limitations of geospatial cryptography in the analysis of data from our nation's cancer registries.

1.1 Contributions of this paper

There are three substantive contributions of this paper. First, secure computational platforms have yet to be defined for geospatial cryptography, a task accomplished in this paper. Second, this paper is the first to define four real-world use cases that involve the analysis of confidential data from cancer registries where the present state of practice substantially reduces the pace of cancer research. These are identified and evaluated in this paper, and geospatial cryptographic solutions for each use case are proposed. Finally, this research demonstrates how geospatial cryptography has the potential to accelerate the pace of cancer research through secure data sharing.

2 Materials and methods

2.1 A methodology for geospatial cryptography

Before considering each use case, we first describe an approach to geospatial cryptography involving partial homomorphic encryption and secure multi-party computation.

For decades, a major goal of cryptography has been the development of encryption methods for which one can add and multiply ciphertexts, where ciphertexts are the encrypted data. This would enable a fully homomorphic encryption scheme, since once addition and multiplication are possible, one can compute any function. Because of its inherent data security (the data keeper does not have to share the key; hence, security is very high), and because re-identification risk can be mathematically quantified (hence, data security can be directly assessed), it appears homomorphic encryption could accelerate human subjects research whenever confidential data must be shared for analysis to proceed. Conceivably, one could conduct the geospatial analysis in the encrypted space using original coordinates (not “fuzzed” or aggregated) and return the results (e.g., odds ratios for exposure; incidence/mortality rate; and cluster results), without ever revealing individual locations. As noted in the discussion, the development of a geospatially encrypted geographic information system (GEGIS) would support mapping and visualization.

Formally, homomorphic cryptography is defined as a procedure that encrypts data in such a fashion that mathematical operations can be conducted without having to decrypt the data (Fontaine and Galand 2007). Results are then decrypted and reported, without ever revealing confidential information. In the additive homomorphic cryptosystem of Paillier (Paillier 1999), the sum of two messages is equal to the decryption of the product of their corresponding ciphertexts:

$$D(E(m_1, e) \otimes E(m_2, e), d) = m_1 + m_2 \quad (1)$$

Here m_1 and m_2 are the two plaintext messages, E is the encryption function, D is the decryption function, e is the public encryption key and d is the private decryption key. The product of a ciphertext may also be calculated in the Paillier cryptosystem:

$$D(E(m_1, e)^q, d) = m_1 \times q \quad (2)$$

Here q is just an arbitrary exponent. In addition, Paillier encryption is probabilistic in that its encryption algorithm uses pseudorandom number generators. Hence, encrypting the same message several times will yield different ciphertexts, making it very difficult for even an informed adversary to compare encrypted messages in order to ascertain the original value that was encrypted.

Recently, a fully homomorphic encryption scheme was developed that allows for computing arbitrary functions over encrypted data without the decryption key (Gentry 2009). While the Gentry system is an active research area in computer science (Gentry and Halevi 2011;

Smart and Vercauteren 2010; Santos et al. 2015), our experiments found it not to be practical at this time as it requires large ciphertexts and considerable computation time. However, a hub-based system for disease surveillance based on Paillier encryption has been successfully implemented (El Emam et al. 2011). For that specific problem, an efficient and scalable protocol was developed which also provided strong privacy guarantees. This gives some confidence that a Paillier-based system provides a viable framework for geospatial cryptography in cancer control and surveillance, and is the system used in our research.

2.2 Secure multi-party computation

The computer science community has developed a number of secure multi-party computation (SMC) protocols for basic mathematical operations, such as multiplication and the dot product. These have been integrated into protocols that perform more complex data mining analysis, such as association rule mining. However, there have been very few applications of SMC to solving healthcare problems, such as public health surveillance (El Emam et al. 2011) and running basic queries on genomic sequence databases (Kantarcioglu et al. 2008). Geospatial analysis imposes a further challenge: When locations of place of residence are displayed, individual privacy can be breached by reverse geocoding (Rushton et al. 2006), as occurred for mortality maps after Hurricane Katrina (Curtis et al. 2006b). To date and to our knowledge, SMC platforms have yet to be defined for geospatial cryptography, a unique contribution of this paper. Here we propose a prototype secure multi-party computation platform for secure data analysis in general and the analysis of geospatial cancer data in particular.

The configuration of the prototype SMC platform is defined in terms of the parties involved and their communications. For confidential health data, as maintained by cancer registries, three types of parties may be involved: the researcher, data custodians (DC) which are the cancer registries themselves, and semi-trusted third parties (TTPs). TTPs are working in between the two other parties, by receiving the private shares of data from the DC and performing the distributed computations to extract and send the query results requested by the researcher. The TTPs never receive keys to decrypt the data. The data are encrypted while under the provenance of the data custodians, and they are never decrypted once they leave the DC and are shared with the TTP. Hence, the TTP can receive whatever private data the DC wishes to share for analysis (e.g., either complete data or a subset of information for each individual). Notice that the TTP could be defined to reside within the domain of the cancer registry, or be implemented on the cloud or at an external data center or server. Where the TTPs would be located thus is flexible, and in a given implementation would be placed to comply with the registry’s regulatory environment and bureaucratic decision process.

The desired characteristics of the protocol are threefold. First, it ensures that the researcher cannot get access to, view, or inadvertently reveal any personal health information. Second, the DC will give service to the researcher through two or more TTPs, and the TTPs have no access to the original data. Finally, the DC may or may not know the queries requested by the researcher, depending on the specific requirements of the data steward to ensure appropriate use of data.. Note that collusion among all of the TTPs is required to compromise the original data. However, even if all except one TTP collude, they would not be able to reach the original data owned by the DC. This protocol thus suits the call, for example, for an NIH-wide GIS capability (Richardson et al. 2013) that will support intra-and extra-mural researchers by providing geospatial analysis of confidential health data without compromising individual privacy. It also provides a computational platform suitable for the four use cases considered here.

2.2.1 Setup phase—There are two main steps in the protocol: setup and operation. Figure 1 illustrates the setup phase of the protocol. There are four steps in the setup phase that ultimately result in the establishment of security keys to the involved TTPs and metadata and a data map being sent to the researcher. The following notation is used: a denotes the raw data stored by DC; E_i is the encryption of the raw data using the encryption key initiated by TTP$_i$; and μ is the mean value of the instances of the selected attribute. To establish secure data distribution, each data item a is privately distributed among N trusted third parties, TTP$_1$,…, TTP$_N$, as follows:

1. Each TTP generates a set of public and private keys, and broadcasts its public key.
2. DC randomly selects numbers a_1,…, $a_{N\text{-}1}$, and calculates a_N such that $a=\sum_{i=1}^{N} a_i$.
3. DC encrypts each private share using the corresponding public key received from the TTPs and sends E_i (a_i) to TTP$_i$, for i = 1,…, N.
4. Each TTP decrypts the received value from the DC and privately stores its own dataset.

The DC then sends the metadata and data map to the researcher, who uses this to make queries and conduct secure analyses.

2.2.2 Operations phase—In the operations phase, queries sent by the researcher are performed on the secure data by the TTP and the final result is then sent back to the researcher. Figure 2 illustrates the operation phase of the protocol, which involves 3 steps:

1. The researcher creates her query and sends it to the TTPs interfaces.
2. As called for by the query, the TTPs perform the analyses using the Paillier cryptosystem and secure building blocks (such as secure multi-party multiplication and dot product).
3. Once the computations are complete, the researcher is sent the final result wrapped with her private random value (key), and extracts the analysis result by unwrapping the received value.

Here we present a simple example of computing the mean to illustrate how our platform could work. The example uses two TTPs, TTP$_1$ and TTP$_2$, but the approach applies without loss of generality when there are more than two TTPs. We let n be the number of records in the data set.

1. The researcher generates a random number, r_{RC}, and sends its encryption, $E_1(r_{\mathrm{RC}})$, to TTP$_2$.
2. The researcher sends the name of the attribute whose mean is to be calculated, A, to both TTPs.
3. TTP$_2$ computes $\mu_2 = \frac{\sum_{i=1}^{n} a_{i,2}}{n}$, encrypts μ_2 and sends $E_1(\mu_2) \otimes E_1(r_{\mathrm{RC}})$ to TTP$_1$.
4. TTP$_1$ decrypts $E_1(\mu_2) \otimes E_1(r_{\mathrm{RC}}) = E_1(\mu_2 + r_{\mathrm{RC}})$, adds $\mu_1 = \frac{\sum_{i=1}^{n} a_{i,1}}{n}$ and sends the result to the researcher.
5. The researcher subtracts the random number r_{RC} from the received value to reach the mean value of the selected attribute.

One might ask why cannot data custodians manage the distribution of encrypted data?

In fact, the TTP could be the data custodians themselves, and it will depend on the specific implementation and institutional considerations. A second consideration is why do analytical results need to be encrypted, as results do not necessarily violate data confidentiality? Whether the results need to be encrypted will depend on the re-identification risks of the results. Multi-party computation provides a wide range of flexibility and precision in terms of what can be inferred about individual subjects only given access to the result. Consider some examples relevant to the use cases. In some situations, it may be appropriate to deliver specific values, such as the identifiers of records matching a certain criterion (e.g., secure data linkage). In other instances, it may be more appropriate to only produce a count (e.g., case aggregation). Or in cases of high re-identification risk, a computation could be specified to only return a single bit: yes or no, there were matches (e.g., de-duplication via record matching). Such Boolean responses would also be used in a cluster surveillance system; yes or no, there was global clustering; yes or no, there was local clustering; yes or no, there was focused clustering; yes or no, there was a statistically significant departure from the baseline in incidence or mortality (e.g., cluster detection and surveillance). Additional information acceptable to the data custodians might also be included, such as measures of comparative risk associated with the cluster, number of cases in the cluster and so on. The key is to provide information to support informed decision-making that is also acceptable to the data custodians and does not increase re-identification risk. This section has provided a description of how geospatial cryptography can be implemented as a secure system, with specific examples.

3 Results

We now consider the four use cases that illustrate how applied geospatial cryptography can be used in cancer control and surveillance at our nation’s cancer registries.

3.1 Use case one: national-level de-duplication among state or province-based cancer registries

Duplication of health records arises when multiple records of the same health event are recorded by several different data stakeholders. One example is cancer, where duplicate records regarding the same health event (for example, tumor diagnosis) arise when a person visits health providers in more than one administrative jurisdiction. Every state, province and territory in the USA and Canada has a population-based cancer registry which collects information on every newly diagnosed cancer cases in their respective geographic jurisdiction. Registry data are certified annually by the North American Association of Central Cancer Registries (NAACCR) based on meeting objective data quality standards. The standards include measures for completeness of coverage, accuracy (passing edits) and timeliness of reporting. The certification process requires de-duplication of data within a state or province, but currently does not address the possibility of duplication between states and provinces, a deficiency addressed by this use case. Some states also include additional, non-certified registries, creating additional duplication challenges. For example, a collection of papers appeared in 2015 in the journal Statistics and Public Policy that conducted cluster analyses on data described in Amin et al. (2014) as coming from the Florida Association of Pediatric Tumor Programs, which is separate from the Florida Cancer Data System. Inclusion of data from data sources other than certified registries may pose additional concerns and is an issue for future work.

The registration of a specific cancer depends on the place of residence at the time of diagnosis. Since an individual may have multiple different cancers over the course of a lifetime, sometimes these cancers are recorded as occurring in different jurisdictions. It is also possible that a resident of one jurisdiction would be diagnosed or treated outside of their home state or province. This arises, for example, for people who reside in different places at different times of the year (e.g., “snowbirds”), and the same cancer is “diagnosed” in each jurisdiction for each place of residence (e.g., both Florida and New York). Data on out-of-state residents are routinely shared with their state of residence through data exchange agreements. However, when both jurisdictions have a valid address for diagnosed cases, the cases can become duplicates, counted in each jurisdiction, thus overestimating the number of cancer cases.

Further, confidentiality concerns and concomitant regulations restrict cancer registries from sharing the necessary personal identifying information that is required to de-duplicate cases across cancer registries. As a result, an unknown proportion of cancer incidence is overestimated due to an inability to undertake de-duplication across registries, a problem that is readily solved through geospatial cryptography.

From a technical perspective, the non-cryptographic approach to de-duplication requires plaintext (i.e., non-encrypted) access to the other party’s dataset. Duplicates are identified by directly comparing every element in the first list with every element in the second list and recording any matches. The obvious limitation of this approach is that the de-duplication cannot proceed unless at least one party agrees to share its data with the other. In practice and as noted earlier, there are privacy laws and policies in place that expressly prevent this, or, in cases where such sharing can be accomplished, it may be further restricted by legal, regulatory or procedural bottlenecks governing the access process itself. Even if a data release is possible in principle, the specific data to be released may require a risk assessment, and in some cases, a partial de-identification.

Secure computation, by comparison, does not require one party to release their data to the other. Rather the party sends their data in an *encrypted* form, under a key only they know. De-duplication is performed only on the encrypted data. The results can be provided in several forms, depending on the regulatory and bureaucratic environment. First, as a response indicating whether or not duplicates were found (yes or no), and hence, further scrutiny is warranted. Second, as a list of the duplicates, encrypted and returned to each of the registries that contain the duplicate. Third, as a decrypted list of the duplicates, sent to each registry. Notice that access to personally identifiable information in another registries data is **not** a precondition to conducting the de-duplication.

What are the benefits of secure de-duplication using geospatial cryptography? Under current practice, de-duplication requires that state registries share data records for purposes of identifying duplicate records. Some state registries have more severe restrictions on sharing data compared with others. In fact, some state registries can only participate in research in a very limited capacity, if at all, because their data are so heavily protected. Further, some registries limit the variables submitted to NAACCR (e.g. some states submit tract-level data to their funding agencies but not to NAACCR). A program that utilizes encryption would save significant time and could be used for multistate de-duplication, which is extremely challenging. National-level de-duplication must be done through interstate cooperation, because the NAACCR Cancer in North American (CiNA) dataset does not contain the personal identifiers required for de-duplication. A substantial opportunity thus exists to enhance data sharing and to improve data accuracy among multiple data stakeholders through geospatial cryptography. Currently, NAACCR is undergoing a pilot study to determine the feasibility and accuracy of applying a hashing encryption approach to inter-state deduplication. NAACCR is currently evaluating using hashing algorithms for national-level de-duplication. It is unknown if this method is accurate or feasible in terms of financial, FTE, and IT burden on cancer registries. Parallel evaluation of homomorphic cryptography and comparison of accuracy and feasibility to hashing is a clear next step to determine not only the promise of encryption for de-duplication but also to delineate the limitations of each approach to guide future directions.

Nationally, some organizations share data with NCI, CDC, NAACCR and other entities, but do so at a greater level of detail with groups that can provide certain legal protections. If such a de-duplication service were created, its enhanced security might allow these organizations to share more detailed data by meeting legal mandated security requirements, thereby increasing data sharing and accelerating the pace of human subjects research.

3.2 Use case two: sharing of confidential data across cancer registries to support case aggregation

Currently, confidentiality concerns impede the study of rare cancers. To study rare cancers, researchers may wish to aggregate all cases of a specific cancer across geographic boundaries to pool data from a sufficient number of cases for statistical analyses. Indeed, such an activity is required simply to undertake the basic statistical power analysis needed in research proposals. Simple questions such as “will there be enough cases in the study to detect an effect of a given size?” at present must be answered speculatively, rather than by an accurate assessment of how many cases will be available for inclusion in the study. However, because many states and provinces will not release information on cases if the aggregated cell size is less than 6, even for de-identified data, the researchers are prevented from assembling a cohort. Individual approvals from each state/province IRB would be required to assemble such a cohort, with multiple applications, reviews and approvals, delaying the study, incurring costs, and reducing efficiency. Further, after considerable effort and expense it sometimes occurs that a sufficient number of cases were not achieved.

We envision a procedure through which data could be uploaded to an encrypted system that would provide the results for all states in an area of concern without identifying the data from each state. Currently, only de-identified cohorts can be aggregated this way using national datasets, such as CiNA. Presently, there are two main strategies in place: attempt to de-identify data before sharing it with researchers, or simply do not release it at all. Not communicating across state lines leads to data duplication and leads directly to elevated risk estimates and potentially to the creation of false hot spots. Furthermore, data with personal identifiers or below-county geographic identifiers remains largely sequestered across jurisdictional lines. For example, a nationwide study may involve up to 50 separate IRB requests to 50 registries. In some cases, data might be made available to researchers following de-identification. Fundamentally, however, de-identification is often accomplished at the expense of data precision and quality, and case aggregation remains an issue. A secure computation approach could, in principle, overcome data fragmentation across jurisdictional lines by enforcing separation along technological (i.e., cryptographic) lines. There are numerous technical questions for future work to explore, such as which entities control the decryption key, and which entities participate in the secure computation protocol. All this, however, points toward a vision for the future in which data can one day be aggregated in a common registry with the goal of providing higher-quality, faster results while simultaneously delivering on legal and policy obligations to protect patient privacy.

3.3 Use case three: secure data linkage

With the advent of personalized medicine and cancer treatment, there is growing interest in multiple primary cancer research. It has become increasingly important to understand the genetic components that contribute to the occurrence of cancers in individuals. In order to track these cancers, we must be able to identify multiple primaries within jurisdictions as well as across jurisdictions. An individual may develop a cancer while a resident of one state/province and a second or third in another. Without the ability to link these cancers to the individual, we are underestimating the number of multiple primary cancers, are overestimating first primary cancers and are unable to contribute effectively to research in this arena.

Much of population-based cancer research relies on matching cancer registry data with other data sets to supplement existing data, to determine a cancer outcome following exposure, or various other research applications. These research efforts are often hindered by a registry's inability to provide data for a given linkage project due to concerns surrounding confidentiality. Additionally, researchers can coordinate some types of studies, such as cohort linkages, through the NAACCR Virtual Pooled Registry Cancer Linkage System

(VPR-CLS), but such projects require multiple IRB approvals and often state-by-state (or province) permissions and approvals, an inefficient and laborious methodology.

Under geospatial cryptography, the data would be linked in a secure, privacy-preserving manner, and researchers could quickly identify the numbers of cases in each state that meet the study requirements, and prioritize before submitting their protocol for IRB approvals. This would augment the current VPR-CLS approach and help ease the burden of IRB and other approval requirements prior to data collection only to find nothing of interest. For example, aggregating 10 states' data and identifying a substantial number of cases that are worth studying could save substantial effort that would otherwise be expended on non-expedited IRB approval processes and could directly support quantification of the statistical power of the study during the planning phase.

The case can be made to IRBs that the risk of disclosing confidential data is extremely small under encryption, and, based on feedback from members of IRBs present at the NAACCR-BioMedware geospatial cryptography workshop, there appears to be willingness for IRBs to accepting cryptographic approaches to data sharing with expedited IRB review.

Similar to the de-duplication model, the current method for conducting data linkages ultimately requires one party to share its data with the other. Conversely in the secure geospatial cryptography model, one party sends its data in an encrypted form to the other party who performs the secure linkage. Depending on the complexity of the linkage method, several rounds of encrypted messages may need to be exchanged between the parties. Exact linkage in a secure multi-party setting (El Emam et al. 2012) has been implemented in the context of public health surveillance. A greater technical challenge in this setting is *approximate* linkage, in which matching fields should be detected even in the presence of small variations (e.g., typos and misspellings), and probabilistic linkage in which a threshold probabilistic model is applied across fields to detect a match. As a downside to this approach, considerable overhead is incurred in this setting because of the potentially exponentially large number of possible outcomes an encrypted computation must explore as a consequence of not leaking information about the encrypted data. Highly efficient secure protocols for records linkage remain an open but promising research area.

3.4 Use case four: cancer clustering and surveillance

Small-area analysis, below county scale, is rarely conducted at a national level. Due to issues with confidentiality, many central cancer registries are unable to supply geographic location, such as latitude/longitude or even census tract to a national research database, such as NAACCRs CiNA dataset. This precludes the ability to conduct national-level cluster detection and, due to edge effects and missing geographic data outside the study area, limits the interpretability of state-based cluster detection analysis conducted along state or national borders. Notably, no geography-based information below province is currently releasable to researchers outside of Canada using North American datasets.

The use of area-based social measures (ABSM) at the national level is also hampered due to issues with confidentiality. Currently, three tract-level ABSM are calculated at the time of national call for data (Kreiger poverty codes, urban/rural status and urban/rural commuting codes). The codes are intended to maintain confidentiality of location but are not always applicable for research purposes. Often the variables of interest extend beyond these three ABSM, for instance education or housing information or locally used ABSM. And the current ABMS, particularly the poverty code, are not applicable for all regions or for stratified race/ethnicity analysis due to issues with residual confounding, which leads to biased results.

Masking, aggregating, or spatially blurring the location data is not recommended practice (Jacquez 2004) due to lack of precision often leading to Type II error, for cluster detection, and computational issues, for ABSM analysis. In addition, research combining individual cancer patient data with ABSM requires multilevel analysis (Diez-Roux 1998; Subramanian 2010), which requires tract identification.

So far in our discussion of secure computation, we have considered a scenario in which two data holders securely interact to answer questions of mutual importance. In this model, we assume that both parties have plaintext access to their respective datasets, and thus, the goal is to complete the computation without requiring plaintext access to the *other* dataset. Let us now consider a slightly different scenario in which there are potentially multiple data holders who provide their data in an encrypted form to a centralized registry. Here the goal would be for the registry, or perhaps a third party (e.g., researcher) to complete the computation without plaintext access to the respective datasets. An example scenario is one in which a researcher is interested in performing geospatial clustering of cancer incidences for the purposes of hot spot detection. Under current practice, the researcher would seek access to geographic data. Depending on the data holders' confidentiality requirements, the registry may undertake to de-identify the data before releasing it to the researcher. Ultimately, however, de-identification represents a zero-sum trade-off between data quality and re-identification risk. Once again, geospatial cryptography could potentially be used to side-step this trade-off. And again, instead of sending de-identified data to the researcher, the registry sends the full-quality data in encrypted form, and the researcher conducts the secure computation (e.g., geospatial clustering) with the interaction of the registry. A semi-trusted third party (TTP) can be designated as the key holder enforcing a separation between the entity that possesses the encrypted data (i.e., the registry) and the entity that can decrypt it (i.e., the TTP). Finally, owing to the flexibility of the cryptography, arbitrary decryption access structures are possible, such as distributed decryption (e.g., two out of two parties are required to perform decryption), or even threshold decryption (e.g., any two out of three).

But how might spatial statistics, necessary for cluster analysis and surveillance, be undertaken in geospatial cryptography? Spatial weights are required in geospatial health analysis, and underpin most spatial models. Given M points (e.g., coordinates of places of residence of cancer cases) in the geographic plane define a matrix of spatial weights, **W**, of the form:

$$S=f(\mathbf{W},\mathbf{D}) \quad (3)$$

Here S is a spatial statistic one wishes to calculate and **D** is a matrix of measures calculated from the attributes, which may include case–control identifiers and exposure metrics. Many inferential spatial statistics conform to this general form (Marshall 1991; Lawson 2006). Examples include the Bernoulli spatial scan statistic (Kulldorff 1997), Cuzick and Edwards' test (Cuzick and Edwards 1990), Mantel's test (Mantel 1967), the Knox test (Knox 1964) and the Vesta and Janus statistics (Jacquez et al. 2007), to name a few. Geospatial modeling approaches such as spatial regression (Anselin and Bera 1998; Waller and Gotway 2004), geostatistics (Goovaerts 1997), geographically weighted regression (Fotheringham et al. 2002) and others may also be written in this form. The weights themselves may be nearest neighbor, adjacency or based on geographic distances (e.g., Euclidean distance).

To assess whether geospatial cryptographic techniques can be used in geospatial analysis we first calculated Euclidean and nearest neighbor relationships among a set of points in the plane, and then assessed clustering among these points by associating a binary attribute with each point (to represent a case–control identifier) and applied the Bernoulli spatial scan statistic (Kulldorff 1997) to assess spatial clustering. While by no means exhaustive or even necessarily representative of the rich complement of algorithms used in spatial analysis, we believe this to be a useful first exploration of geospatial cryptography for use in cancer clustering and surveillance.

3.4.1 Spatial weight calculations for residential locations—Places of residence are frequently used in human subjects research to record place of residence at time of diagnosis or death. These also are used in GIS operations to query data layers. Assume data of the form (x_{it}, y_{it}, $\tilde{a}_{it}$), here x_{it}, y_{it} are the geographic coordinates (e.g., longitude and latitude) of human subject i at time t, and $\tilde{a}_{it}$ is a vector of individual-level attributes (e.g., case–control status, BMI, smoking status and age). We wish to evaluate the Euclidean distance between human subjects i and j at time t, and the nearest neighbor of person i at time t. Without loss of generality suppose the input to the protocol is a pair of coordinate-wise encrypted points (here we drop the time subscripting for simplicity): $E(x_1)$, $E(y_1)$ and $E(x_2)$, $E(y_2)$. The output of the protocol is an encrypted number $E(d)$ representing the geographic Euclidean distance d between the two points:

$$d=\sqrt{(x_1-x_2)^2+(y_1-y_2)^2}=\sqrt{x_1^2+x_2^2-2x_1x_2-2y_1y_2+y_1^2+y_2^2}$$

Although protocols exist for securely computing Euclidean distances in a two-party setting (Mouffron 2008), i.e., in a setting in which each party knows its own coordinates, our protocol computes the result entirely on encrypted data, i.e., in a setting in which the TTPs receive the data from the DC's in encrypted form, and neither party learns any information about the points (including even result d). The intuition of the protocol is as follows: two parties, TTP$_1$ and TTP$_2$ interact to compute the individual terms of the Euclidean distance as defined above, and then homomorphically sum them to produce the encrypted result. One approach to the protocol would be to share the private decryption key k_d between the TTPs. This, however, requires a complex key agreement subprotocol. The approach we take instead is to separate privilege of the TTPs as follows: TTP$_1$ receives the encrypted data from the DCs and TTP$_2$ knows the private decryption k_d. We use the notation $[\![x]\!]$ to denote the encryption of x, i.e., $E(x)$. Finally, we assume the TTPs are honest-but-curious (i.e., that they will follow the protocol below, but will try to extract information about the points through passive observation).

3.5 Secure pairwise euclidean distance (SPED) protocol

Public input Public encryption key k_e (i.e., public modulus n in the case of Paillier).

$TTP_1^{'}s$ *Private input* Encrypted coordinates $[\![x_1]\!]$, $[\![y_1]\!]$ and $[\![x_2]\!]$, $[\![y_2]\!]$.

$TTP_1^{'}$ *Private input* private decryption key k_d:

Output Encrypted Euclidean distance $[\![d]\!]$.

3.5.1 Phase 1: computing squared terms—For each of the encrypted points $[\![x_i]\!]$, $[\![y_i]\!]$, securely compute their respective squares, $[[x_i^2]], [[y_i^2]]$:

1. For each point $[\![x_i]\!]$, $[\![y_i]\!]$, TTP$_1$ selects random blind factors r_x, $r_y \in_R \mathbb{Z}_n$ and computes:

$$[\![x_i]\!]^{r_x}=[[r_x x_i]]$$

and

$$[\![y_i]\!]^{r_y}=[[r_y y_i]]$$

and sends the result to TTP$_2$.

2. For each blinded point $[\![r_x x_i]\!]$ and $[[r_y y_i]]$, TTP$_2$ performs the following:

   a. Decrypt to recover blinded plaintext values $r_x x_i$ and $r_y y_i$,

   b. Compute the square of blinded plaintext values: $(r_x x_i)^2$ and $(r_y y_i)^2$

   c. Encrypt the squared blinded plaintext values $[[(r_x x_i)^2]]$ and $[[r_y y_i)\ ^2]]$and send to TTP$_1$

3. For each blinded squared point, TTP$_1$ computes r_x^{-2} mod n and r_y^{-2} mod n. TTP$_1$ then homomorphically strips off the blind factor by computing:

$$[[(r_x x_i)^2]]^{r_x^{-2}}=[[r_x^{-2} r_x^2 x_i^2 \bmod n]]=[[x_i^2]]$$

and

$$[[(r_y y_i)^2]]^{r_y^{-2}}=[[r_y^{-2} r_y^2 y_i^2 \bmod n]]=[[y_i^2]]$$

3.5.2 Phase 2: computing pairwise terms—Securely compute the pairwise terms $[\![-2x_1x_2]\!]$ and $[\![-2y_1y_2]\!]$:

1. Using the blinded terms $r_{x_1}x_1$ and $r_{x_2}x_2$ received in Step 2 of the previous phase, TTP_2 performs the following:

   a. Compute $r_{x_1}x_1 \cdot r_{x_2}x_2 \bmod n$

   b. Encrypt and send the result $\llbracket r_{x_1}\, r_{x_2}\, x_1\, x_2 \rrbracket$ to TTP_1

2. TTP_1 computes $(r_{x_1}\, r_{x_2})^{-1} \bmod n$ then homomorphically strips off the blind factors by computing:

$$\llbracket r_{x_1} r_{x_2} x_1 x_2 \rrbracket^{(r_{x_1} r_{x_2})^{-1}} = [[r_{x_1}^{-1} r_{x_2}^{-1} r_{x_1} r_{x_2} x_1 x_2 \bmod n]] = \llbracket x_1 x_2 \rrbracket$$

Finally, TTP_1 computes:

$$\llbracket x_1 x_2 \rrbracket^{-2} = \llbracket -2x_1 x_2 \rrbracket.$$

The TTPs repeat these steps similarly for computing $\llbracket -2y_1y_2 \rrbracket$.

3.5.3 Phase 3: homomorphically computing the sum of terms—

Using the encrypted terms computed in the previous steps, TTP_1 computes the encrypted Euclidean distance as follows:

$$[[x_1^2]] \cdot [[x_2^2]] \cdot \llbracket -2x_1 x_2 \rrbracket \cdot \llbracket -2y_1 y_2 \rrbracket \cdot [[y_1^2]] \cdot [[y_2^2]]$$

$$[[x_1^2 + x_2^2 - 2x_1 x_2 - 2y_1 y_2 + y_1^2 + y_2^2]] = \llbracket d \rrbracket$$

TTP_1 can now send the Euclidean distance $\llbracket d \rrbracket$ to TTP_2 to be decrypted, or alternatively it can be given as input to another secure protocol.

3.5.4 Security argument (sketch)—

We provide a brief argument that the SPED protocol is secure, i.e., that neither TTP learns any information about points x_1, y_1 and x_2, y_2. Security from the perspective of TTP_1 is straightforward: Through the course of the protocol, TTP_1 only ever receives encrypted values, for which it does not know the corresponding decryption key, and therefore rests on the semantic security of the Paillier encryption system. Security from the perspective of TTP_2 is similarly straightforward: through the course of the protocol, TTP_2 only ever receives encryptions of blinded values. TTP_2 can decrypt the blinded values, but does not know their blind factors. If TTP_1 chose the blind factors independently and uniformly at random, then the blinded values are perfectly hiding.

3.6 Performance evaluation

We implemented the Paillier cryptosystem in C++ using the MIRACL[1] software library using current recommended minimum key length $|n| = 2048$ bits (Barker and Roginsky 2015). We utilized MIRACL's built-in Karatsuba–Comba–Montgomery (KCM) multiplication assembly optimization for AMD64 architecture, and conducted tests on an

Intel Core i7-3770 3.4 GHz desktop running Debian. The timings are for a single thread and averaged across 1000 trials. Encryption took approximately 16 ms per trial, as did scalar homomorphic multiplication. Using Chinese Remaindering, decryption took approximately 10 ms per trial. The timings of other operations, such as homomorphic addition, computing single multiplications or inverses and random number generation, were all negligible in comparison. As a simplifying assumption, we discount the network communication time between parties. Table 1 shows the timings for the SPED protocol at each step to compute the encrypted Euclidean distance between two points. In total it takes approximately 316 ms per point pair.

To get a sense of scale, Fig. 3 shows the time to compute all pairwise distances with the SPED protocol as a function of the total number of records. Since multiple invocations of SPED can be parallelized, we also show how the execution times can be reduced through the use of multiple computing threads/cores.

While substantial research opportunities exist for accelerating computation (since we did not optimize the algorithm), these experiments demonstrated spatial cryptography makes possible secure computations of spatial weights and spatial statistics for individual-level health data such that the identity of individuals cannot be reconstructed or deduced.

4 Discussion

What opportunities might exist for adoption of geospatial cryptography in health research and analysis? To address this question a small working group “Evaluation of Homomorphic Cryptography for Geospatial Studies with Human Subjects” was convened at the 2013 meetings of the North American Association of Central Cancer Registries (NAACCR). Over the course of two days, six participants (Francis Boscoe, New York Cancer Registry; David O’Brien, Alaska Cancer Registry; Glenn Copeland, Michigan Cancer Surveillance Program; Rich Pinder, Los Angeles Cancer Surveillance Program; David Stinchcomb, Westat; and Xiao Cheng Wu, Louisiana Cancer Registry) met with organizers Geoffrey Jacquez (BioMedware and SUNY Buffalo), Khaled El Emam (University of Ottawa) and Betsy Kohler (NAACCR) to formulate recommendations regarding the use of geospatial cryptography in human subjects research.

4.1 Secure geocoding

Potential applications in addition to the prior use cases were identified and include secure geocoding. Workshop participants agreed that issues relating to data privacy represent substantial obstacles to basic geographic information system operations and the sharing of data. Dr. Daniel Goldberg created a geocoder which is available to the cancer surveillance community and is hosted at Texas A&M University. NAACCR members can submit the addresses of their cases to this geocoder to convert addresses into latitude–longitude coordinates suitable for data aggregation (e.g., calculation of county-specific incidence rates) and analysis. The address information is transmitted to the Texas A&M data servers, geocoded, and then returned to the registry that supplied the data. The records are then

[1] https://github.com/CertiVox/MIRACL.

deleted from the Texas A&M servers. Security vulnerabilities in this approach exist when the data are transmitted, as well as when they reside on the Texas A&M servers.

With online geocoding systems, whoever is providing the data are providing personal information (e.g., addresses) to the geocoding system. Using an encrypted secure string comparison scheme, it is possible to carry out secure geocoding such that whoever is providing the data encrypts it, submits it, and the geocoding is then accomplished on the encrypted addresses. Security vulnerabilities during data transmission, while resident on servers, and during the geocoding process itself are greatly minimized as the data are never decrypted.

Workshop participants suggested the NAACCR community would be enthusiastic about the use of geospatial cryptography to safeguard data in conjunction with such a geocoding protocol, thereby enabling the sharing of data and accelerating the pace of human subjects research.

4.2 Other applications

Registries collect hundreds of data items, many of which have not been evaluated for fitness for use or usability (there is significant interest in analyzing treatment data that registries collect). Identifying where there are missing/unusable data from registries while masking the identity of the individual registries (to encourage participation from as many registries as possible) would be extremely useful. Other potential applications of spatial cryptography include place-specific health disparities analyses, distance to facility analyses (e.g., at late-stage diagnosis, how far was a patient from the screening facility?), and new ways of accessing and using Census data and tax information (e.g., identifying the denominator), all within under secure geospatial encryption.

4.3 Challenges

In spite of the potential applications noted above, challenges facing the development of geospatial cryptography techniques include determining how much precomputation the data owner should perform and to what degree the necessary analytics can be anticipated, minimizing the complexity of cipher key management; developing a broad enough library of secure computation routines/functions that an analyst can use; avoiding information leaks from multiple queries and model results; and improving performance to handle large data sets. Additional considerations include analyst training and implementing strong security controls and audits.

Substantial research opportunities exist in this emerging area of geospatial cryptography. First, spatial data structures and algorithms such as r trees, quad trees, and kd trees (Samet 1990) have yet to be developed in geospatial cryptographic systems. How do we store, sort and rapidly access geospatial data in the encrypted space? Such questions must be addressed for a variety of different data types, including points, space–time paths, polygons and networks. This issue must be solved for large-scale GIS operations to become practical for encrypted data.

Second, protocols for undertaking spatial statistics and models under geospatial encryption remain to be developed. This is a substantial undertaking at the interface of statistics, mathematics and computer science, and involves the development of optimized protocols for matrix algebra as a starting point. That such statistical protocols appear feasible was demonstrated, at least for the spatial scan statistic, in this research. Full feasibility demonstration will involve performance evaluation over a host of complex data arrangements, and test deployment of a geospatial cryptographic system at a cancer registry. But an enormous amount remains to be accomplished for a host of spatial pattern analysis and modeling approaches. Ultimately, a spatial statistical toolbox under geospatial cryptography would be a technological development of considerable importance.

Third, the impacts of geospatial cryptography on the individual and society remain to be assessed. What are the individual and societal implications of geospatial cryptography for confidentiality, data sharing and privacy? Clearly a substantive benefit is enabling collaboration through secure data sharing, but what are the societal implications? This question has yet to be addressed.

Fourth, important mathematical problems exist and likely more will be revealed in geospatial cryptography. The mathematical specification of the data sharing problem has yet to be fully addressed. The examples above can be formulated in such as way that the data providers are able to approve sharing of results with others before such sharing is undertaken. For statistical and GIS analyses, data may be aggregated and reported as a statistical summary—for example, the number of cases that meet certain query criteria within a given area. By itself such a single query may be secure. But when other data from outside sources are added, and other queries of the secure system are undertaken, re-identification risk can increase (Curtis et al. 2006b, 2011). This problem has yet to be properly specified mathematically and solved. Its solution ultimately will provide the closed-form calculation of privacy risks under repeated queries and the addition of data from outside sources. This is of considerable practical importance since it would allow registries a formal mathematical approach for assessing re-identification risks.

Finally, the potential impacts on visualization for geospatially encrypted systems (we might call this a GEGIS—geospatially encrypted geographic information system) at this juncture are largely unknown. The examples posed earlier have carefully avoided visualization, as this involves repeated queries, the addition of multiple external data sources, and the solution of computational issues noted above (e.g., geospatially encrypted data structures, and the ability to undertake spatial statistics under geospatial encryption). Hence, visualization poses an intriguing future research direction. The authors believe a key issue will be specifying and solving the problem of quantification of re-identification risk under multiple queries and when external data sources are included. One could then imagine a system that tracks queries as well as external data sources brought to a GEGIS session. As these operations take place to support visualization, the GEGIS would continually quantify re-identification risk using a closed-form solution. When this risk exceeded a threshold set by the data owners, the system might inform the user and stop performing requested functions. While this would be annoying to the researcher, it would support the maximum amount of visualization and analysis that would be acceptable to the provider of the confidential/private data.

5 Conclusion

An important limitation is that we have focused our discourse on place of residence as the relevant location for analyses. However, for occupational exposures the place of employment may be of greater importance for cancer etiology. But while residence location, including residential histories, is often included in registry data, place of employment may not be. This is an important omission that merits further discussion.

Confidentiality protection is a requirement for studies involving human subjects, but can slow the pace of basic and applied research that is fundamental to improving our nation's health. Spatial cryptography supports the analysis of confidential geospatial data in the encrypted space (e.g., make it homomorphic)—meaning analyses can be conducted on encrypted data with potentially little if any risk of revealing confidential information. This has enormous potential for accelerating basic and translational research. The protocols proposed in this paper, while focused on geospatial analyses of public health data, will broadly apply to knowledge domains that analyze confidential geospatial information.

Acknowledgments

Financial support This study was supported by the National Library of Medicine Grant R21 LM011132-01A1 (PI G. M. Jacquez).

References

Abowd, JM., Lane, J. New approaches to confidentiality protection: synthetic data, remote access and research data centers. In: Domingo-Ferrer, J., Torra, V., editors. Privacy in statistical databases, proceedings. Vol. 3050. Annals of the New York Academy of SciencesSpringer-Verlag; Berlin: 2004. p. 282-289.

Amin R, Hendryx M, Shull M, Bohnert A. A Cluster analysis of pediatric cancer incidence rates in Florida: 2000–2010. Stat Public Policy. 2014; 1(1):69–77.

Anselin, L., Bera, A. Spatial dependence in linear regression models with an introduction to spatial econometrics. In: Giles, D., Ullah, A., editors. Handbook of economic statistics. Marcel Dekker; New York: 1998. p. 237-289.

Aslett LJ, Esperança PM, Holmes CC. A review of homomorphic encryption and software tools for encrypted statistical machine learning. 2015 arXiv preprint arXiv: 150806574.

Barker E, Roginsky A. Transitions: recommendation for transitioning the use of cryptographic algorithms and key lengths. NIST Special Publication. 2011; 800:131A.

Bell BS, Hoskins R, Pickle L, Wartenberg D. Current practices in spatial analysis of cancer data: mapping health statistics to inform policymakers and the public. Int J Health Geogr. 2006; 5(1):49. [PubMed: 17092353]

Boulos MNK, Curtis AJ, AbdelMalik P. Musings on privacy issues in health research involving disaggregate geographic data about individuals. Int J Health Geogr. 2009; 8:8. [PubMed: 19208225]

Ciriani V, Di Vimercati SD, Foresti S, Jajodia S, Paraboschi S, Samarati P. Combining fragmentation and encryption to protect privacy in data storage. ACM Trans Inf Syst Secur. 2010; 13(3):33.

Curtis A, Mills JW, Leitner M. Keeping an eye on privacy issues with geospatial data. Nature. 2006a; 441(7090):150.

Curtis A, Mills JW, Leitner M. Spatial confidentiality and GIS: re-engineering mortality locations from published maps about Hurricane Katrina. Int J Health Geogr. 2006b; 5:44. [PubMed: 17032448]

Curtis A, Mills JW, Augustin L, Cockburn M. Confidentiality risks in fine scale aggregations of health data. Comput Environ Urban Syst. 2011; 35:57–64.

Cuzick J, Edwards R. Spatial clustering for inhomogeneous populations. J R Stat Soc Ser B Methodol. 1990; 52(1):73–104.

Diez-Roux AV. Bringing context back into epidemiology: variables and fallacies in multilevel analysis. Am J Public Health. 1998; 88(2):216–222. [PubMed: 9491010]

El Emam K, Hu J, Mercer J, Peyton L, Kantarcioglu M, Malin B, Buckeridge D, Samet S, Earle C. A secure protocol for protecting the identity of providers when disclosing data for disease surveillance. J Am Med Inf Assoc. 2011; 18(3):212–217.

El Emam K, Samet S, Hu J, Peyton L, Earle C, Jayaraman G, Wong T, Kantarcioglu M, Dankar F, Essex A. A protocol for the secure linking of registries for HPV surveillance. PLoS One. 2012; 7(7):e39915. [PubMed: 22768321]

Fefferman NH, O'Neil EA, Naumova EN. Confidentiality and confidence: is data aggregation a means to achieve both? J Public Health Policy. 2005; 26(4):430–449. [PubMed: 16392743]

Fontaine C, Galand F. A survey of homomorphic encryption for nonspecialists. EURASIP J Inf Secur. 2007; 1:013801.

Fotheringham, AS., Brunsdon, C., Charlton, M. Geographically weighted regression: the analysis of spatially varying relationships. Wiley; West Sussex: 2002.

Gentry, C. Annual ACM symposium on theory of computing. Assoc Computing Machinery; New York: 2009. Fully homomorphic encryption using ideal lattices. Stoc'09: Proceedings of the 2009 ACM symposium on theory of computing; p. 169-178.

Gentry, C., Halevi, S. Implementing Gentry's fully-homomorphic encryption scheme. In: Paterson, KG., editor. Advances in cryptology—Eurocrypt 2011. Vol. 6632. Lecture notes in computer scienceSpringer-Verlag; Berlin: 2011. p. 129-148.

Goovaerts, P. Geostatics for natural resources evaluation. Oxford University Press; New York: 1997.

Gutmann MP, Witkowski K, Colyer C, O'Rourke JM, McNally J. Providing spatial data for secondary analysis: issues and current practices relating to confidentiality. Popul Res Policy Rev. 2008; 27(6):639–665. [PubMed: 19122860]

Jacquez GM. Current practices in the spatial analysis of cancer: flies in the ointment. Int J Health Geogr. 2004; 3(1):22. [PubMed: 15479473]

Jacquez GM, Meliker J, Kaufmann A. In search of induction and latency periods: space–time interaction accounting for residential mobility, risk factors and covariates. Int J Health Geogr. 2007; 6:11. [PubMed: 17362510]

Kantarcioglu M, Jiang W, Liu Y, Malin B. A cryptographic approach to securely share and query genomic sequences. IEEE T Inf Technol Biomed. 2008; 12(5):606–617.

Kim J, Mu Y, Obaidat MS. Advanced computer mathematics based cryptography and security technologies. Int J Comput Math. 2013; 90(12):2512–2514.

Knox EG. The detection of space–time interactions. Appl Stat. 1964; 13(1):25–30.

Kulldorff M. A spatial scan statistic. Commun Stat Theory Methods. 1997; 26(6):1481–1496.

Lawson, AB. Statistical methods in spatial epidemiology. 2nd. Wiley; New York: 2006.

Mantel N. The detection of disease clustering and a generalized regression approach. Cancer Res. 1967; 27(2 Part 1):209–220. [PubMed: 6018555]

Marshall RJ. A review of methods for the statistical analysis of spatial patterns of disease. J R Stat Soc Ser A Stat Soc. 1991; 154:421–441.

Meliker JR, Goovaerts P, Jacquez GM, AvRuskin GA, Copeland G. Breast and prostate cancer survival in Michigan. Cancer. 2009; 115(10):2212–2221. [PubMed: 19365825]

Mouffron, M. Transitive q-ary functions over finite fields or finite sets: counts, properties and applications. In: von zur Gathen, J.Imaña, JL., Koç, ÇK., editors. Arithmetic of finite fields: 2nd international workshop; WAIFI 2008 Siena; Italy. July 6–9: 2008 proceedings; Berlin: Springer; 2008. p. 19-35.

National Research Council. Putting people on the map: protecting confidentiality with linked social-spatial data. The National Academies Press; Washington, DC: 2007.

Paillier, P. Public-key cryptosystems based on composite degree residuosity classes. In: Stern, J., editor. Advances in cryptology—Eurocrypt'99. Vol. 1592. Lecture notes in computer scienceSpringer- Verlag; Berlin: 1999. p. 223-238.

Richardson DB, Volkow ND, Kwan M-P, Kaplan RM, Goodchild MF, Croyle RT. Spatial turn in health research. Science. 2013; 339(6126):1390–1392. [PubMed: 23520099]

Richardson DB, Kwan M-P, Alter G, McKendry JE. Replication of scientific research: addressing geoprivacy, confidentiality, and data sharing challenges in geospatial research. Ann GIS. 2015; 21(2):101–110.

Rushton G, Armstrong MP, Gittler J, Greene BR, Pavlik CE, West MM, Zimmerman DL. Geocoding in cancer research—a review. Am J Prev Med. 2006; 30(2):S16–S24. [PubMed: 16458786]

Samet, H. The design and analysis of spatial data structures. Vol. 85. Addison-Wesley; Reading: 1990.

Santos, LCD., Bilar, GR., Dac, F., Pereira, FD. Implementation of the fully homomorphic encryption scheme over integers with shorter keys; 2015 7th International conference on new technologies, mobility and security (NTMS), 27–29 July 2015; 2015. p. 1-5.

Smart, NP., Vercauteren, F. Fully homomorphic encryption with relatively small key and ciphertext sizes. In: Nguyen, PQ., Pointcheval, D., editors. Public key cryptography—Pkc 2010, proceedings. Vol. 6056. Lecture notes in computer scienceSpringer-Verlag; Berlin: 2010. p. 420-443.

Subramanian, SV. Multilevel modeling. In: Fischer, MM., Getis, A., editors. Handbook of applied spatial analysis: software tools, methods and applications. Springer; Berlin: 2010.

VanWey LK, Rindfuss RR, Gutmann MP, Entwisle B, Balk DL. Confidentiality and spatially explicit data: concerns and challenges. Proc Natl Acad Sci USA. 2005; 102(43):15337–15342. [PubMed: 16230608]

Verykios VS, Karakasidis A, Mitrogiannis VK. Privacy preserving record linkage approaches. Int J Data Min Model Manag. 2009; 1(2):206–221.

Waller, L., Gotway, C. Applied spatial statistics for public health data. John Wiley and Sons; New Jersey: 2004.

Wartenberg D, Thompson WD. Privacy versus public health: the impact of current confidentiality rules. Am J Public Health. 2010; 100(3):407–412. [PubMed: 20075316]

Wieland SC, Cassa CA, Mandl KD, Berger B. Revealing the spatial distribution of a disease while preserving privacy. Proc Natl Acad Sci USA. 2008; 105(46):17608–17613. [PubMed: 19015533]

Zandbergen PA. Ensuring confidentiality of geocoded health data: assessing geographic masking strategies for individual-level data. Adv Med. 2014

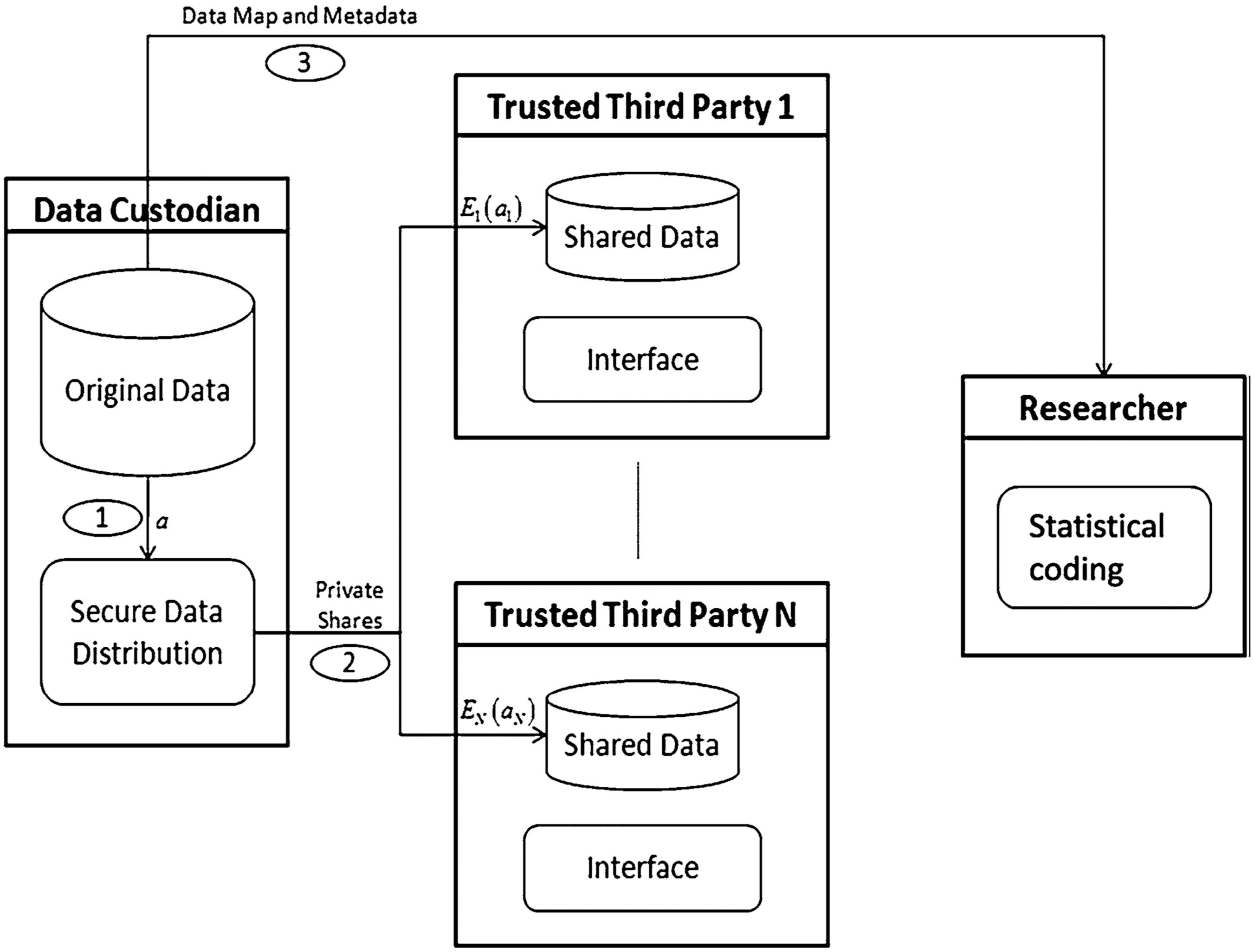

Fig. 1.
Setup phase of the prototype SMC platform. “Statistical coding” may include SAS, R, Python, ArcGIS and SpaceStat

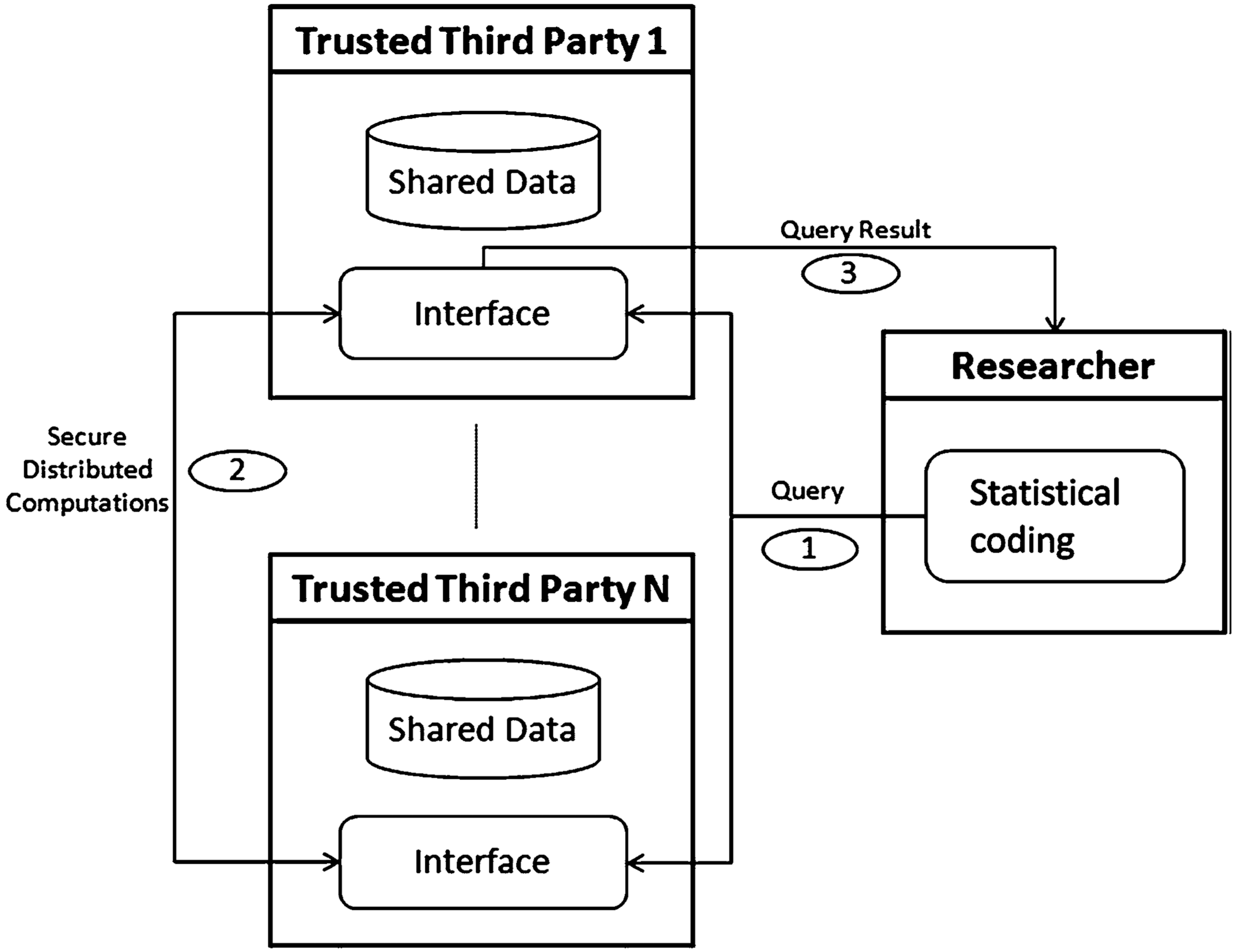

Fig. 2.
Operations phase of the prototype SMC platform

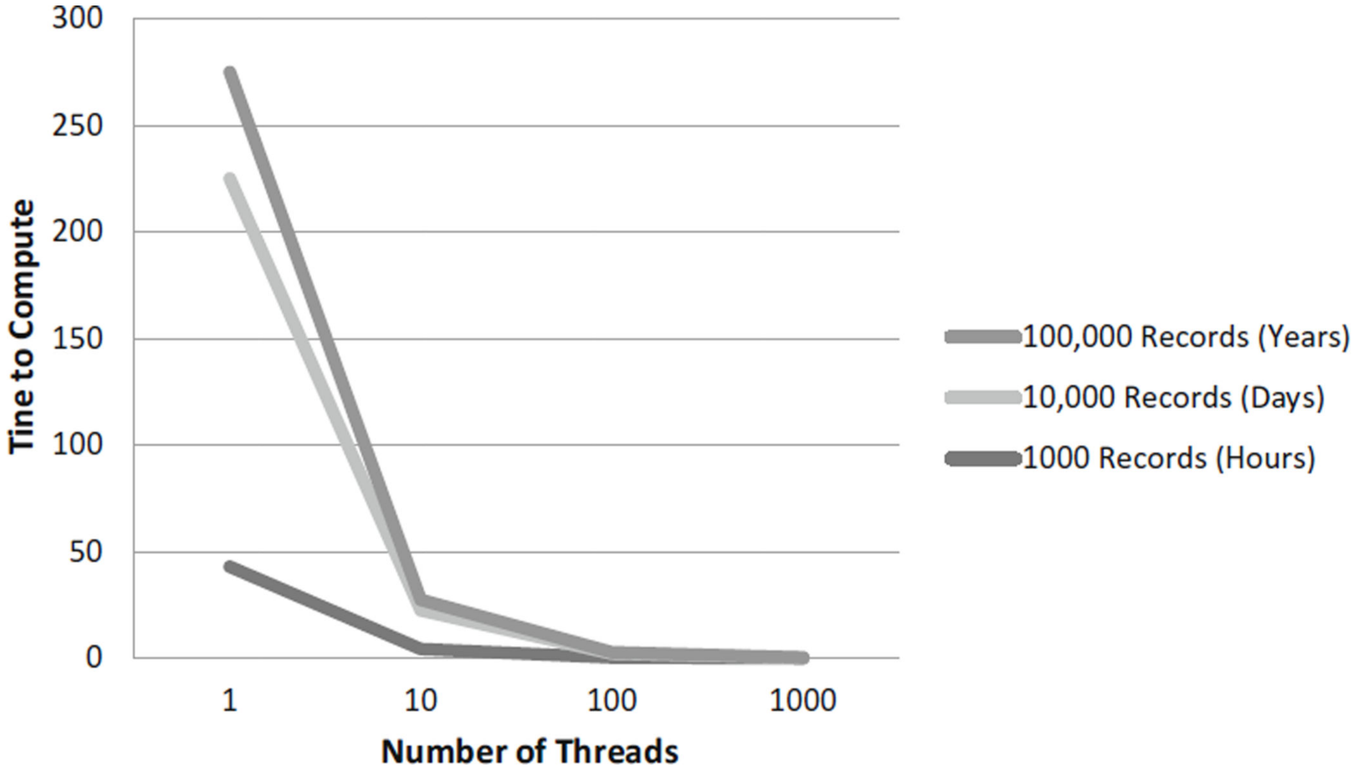

Fig. 3.

Computing all pairwise distances with SPED

Table 1

Timings of operations in the SPED protocol per point pair

	Encryptions	Decryptions	Scalar homomorphic multiplications
Computing square terms (Phase 1)	4 × 16 ms	4 × 10 ms	8 × 16 ms
Computing pairwise terms (Phase 2)	2 × 16 ms	n/a	2 × 16 ms
Computing sum of terms (Phase 3)	n/a	n/a	n/a
Total	6 × 16 ms = 96 ms	6 × 10 ms = 60 ms	10 × 16 ms = 160 ms
Combined total	316 ms		