BioMedware – NAACCR Workshop:

Evaluation of Homomorphic Cryptography for Geospatial Studies with Human Subjects

June 8-9, 2013
Austin, TX

Geoffrey M. Jacquez[1,2], Khaled El Emam[3], Betsy Kohler[4] and Mike Bykowski[5]

1 BioMedware, Ann Arbor, MI
2 Department of Geography, SUNY Buffalo, Buffalo, NY
3 University of Ottawa, Ottawa, ON
4 North American Association of Central Cancer Registries Inc., Springfield, IL
5 Consolidated Solutions and Innovations, Gaithersburg, MD

Source: BioMedware (2014)
http://www.biomedware.com/publications/7920141202naaccrhomomorphiccryptography.pdf

Acknowledgements

This workshop was funded as part of the project "National Library of Medicine (NLM) grant "R21 LM011132 Exploratory evaluation of homomorphic cryptography for confidentiality protection", Geoff Jacquez Principal Investigator, with co-Investigators Khaled El Emam and Betsy Kohler. We thank workshop participants Francis Boscoe, David O'Brien, Glenn Copeland, Rich Pinder, David Stinchcomb, and Xiao Cheng Wu. Charlie Blackburn organized the workshop, with proceedings recorded by Mike Bykowski.

Background

The "Evaluation of Homomorphic Cryptography for Geospatial Studies with Human Subjects Workshop, sponsored by Biomedware and the North American Association of Central Cancer Registries, Inc., (NAACCR), was held on June 8-9, 2013, in Austin, TX. The purpose of the workshop was to:

- Learn of the current status of the National Library of Medicine (NLM) grant "R21 LM011132 Exploratory evaluation of homomorphic cryptography for confidentiality protection" (in progress).

- Identify opportunities, obstacles, and benefits of the application of cryptography in the geospatial analysis of human subjects data.

- Formulate recommendations regarding the use and future research directions of cryptography in human subjects research with registry data and from other data providers (e.g., individual data from National Institutes of Health [NIH] studies, case-control data, and others).

Homomorphic encryption was defined as a procedure or approach that encrypts data in such a way that mathematical operations can be conducted without having to decrypt the data. Results are then decrypted and reported, without revealing confidential information.

Welcome, Logistics, and Introduction

Workshop Chair Dr. Geoffrey Jacquez, President of BioMedware and Professor of Geography at The State University of New York, Buffalo, welcomed participants and explained that this breakthrough in encryption technology—homomorphic encryption—may present new opportunities to accelerate the pace of research and discovery. In collaboration with Workshop Co-Chair Dr. Khaled El Emam, Associate Professor, Faculty of Medicine and the School of Information Technology and Engineering, University of Ottawa, Dr. Jacquez is leading an NLM-funded R21 grant ("Exploratory evaluation of homomorphic cryptography for confidentiality protection") to explore these opportunities.

Both days of the Workshop included sessions dedicated to the project's overview and an introduction to cryptography and multi-party computation. These were followed by group discussions focused on identifying opportunities, obstacles, benefits, and recommendations.

Day 1 participants included:

- Francis Boscoe, New York Cancer Registry
- David O'Brien, Alaska Cancer Registry

Day 2 participants included:

- Glenn Copeland, Michigan Cancer Surveillance Program
- Betsy Kohler, NAACCR
- Rich Pinder, Los Angeles Cancer Surveillance Program
- David Stinchcomb, Westat
- Xiao Cheng Wu, Louisiana Cancer Registry

Dr. Jacquez reviewed the Workshop's logistics and provided additional background information, noting that his hope was that the group would help develop recommendations on future directions on how homomorphic encryption could be used to advance work within the scope of public health. It is anticipated that discussions from this Workshop will help inform and serve as the background for a peer-reviewed publication(s)—very little has been published in the encryption literature with regard to policy implications, and any insights in this area would be particularly useful to the public health and cryptography communities. Should the Workshop discussions and resulting publication(s) identify sufficient policy and technical opportunities for homomorphic encryption to address (in terms of public health and/or research), this R21 grant

may be followed by another full proposal to the NIH (perhaps in collaboration with NAACCR and/or Workshop participants, and likely targeted to the NLM).

Workshop participants were asked to consider the following overarching questions throughout the Workshop:

- Is it acceptable to have the total number of cases in a confidential dataset be public information? This makes it possible to calculate a look up table of likelihoods for a dataset that can then be encrypted and shared among secure parties. This question is relevant because of its potential impact on computation speed.

- The novelty in the approach that is emerging from this project is the development and use of higher dimensional homomorphism. This makes it possible for the parties involved in secure multi-party computation to undertake analysis of encrypted data without knowing *any* of the data values (in current secure computation models each of the parties typically has partial knowledge of the data values). This potentially makes it possible for data repositories to: (1) share highly confidential data for external analysis with minimal risk to privacy, and (2) create and share permutations of the data, again with minimal risk to privacy. How might these benefits be of use to your organization?

- Suppose a system could be implemented with the characteristics summarized in the Workshop. Would your organization's present use policies allow it? What might the implications for data and data use policies be?

- What are the dimensions (IT, policy, security, benefits, etc.) along which a homomorphic encryption system along the lines envisioned should be evaluated? Apply these to assess the current proposed system.

- What are the application areas with the greatest potential benefit? These might include accelerated sharing of research data for confirmatory analysis, increased pace of analysis of secondary analysis, and others.

- What are relevant aspects that should be considered when evaluating how such a cryptographic system might be employed to accelerate institutional review board (IRB) approval of studies involving secondary analysis of existing data? What are the implications for human subjects regulations and the Health Insurance Portability and Accountability Act (HIPAA)?

- What is the prospect (if any) for accelerating the pace of applied and translational research?

- What guidelines are available and/or should be formulated for sharing maps and data visualizations of confidential human subjects data?

Project Overview

Dr. Jacquez discussed the impetus for this project, quoting from a 2010 paper by Wartenberg and Thompson:

> "What is most puzzling and distressing is that, in spite of our increasingly sophisticated technology and electronic data systems, researchers' direct online access to federal vital records data has become increasingly limited over time, impeding and sometimes precluding potentially valuable etiologic investigations." [1]

It is widely assumed that removing names and addresses is sufficient to protect personal privacy. However, the problem of de-identification involves personal details that are not obviously identifying. Known as quasi-identifiers, these include the person's address, age, sex, profession, ethnic origin, and income, among others. For geospatial analysis, one significant challenge is protecting confidentiality while using geographic information (e.g. coordinates of place-of-residence) to undertake spatial epidemiology, exposure reconstruction, health care access evaluation, place-specific health disparities analyses, and so on—any analysis that involves geographic locations from health records.

The project's R21 proposal (which is, in general terms, a feasibility assessment) has the following general aims (a detailed description of the project is included as Appendix A to this report):

- Build a prototype secure multi-party computation (SCM) platform for the computation of mathematical operations required in geospatial analyses. Several approaches have been selected to evaluate the performance of this approach, which to the best of Dr. Jacquez's knowledge, has never been used with geospatial data (e.g., spatial weight calculations for residential locations, cluster/hotspot analyses, calculating rates of late-stage diagnosis by rates, calculating relative and absolute disparities in stage at diagnosis).

- Apply the prototype systems to assess racial disparities in stage at diagnosis for prostate and breast cancers.

- Formally evaluate the approach and formulate recommendations using an independent working group convened by NAACCR (which is the purpose of this Workshop).

- Disseminate the recommendations and results of the feasibility analysis through peer-reviewed publications and presentations at scientific meetings.

Potential innovations as a result of this work include the following: (1) software for the statistical analysis of encrypted data in geospatial health studies; (2) protocols for implementing this new approach in a manner designed to be acceptable to IRBs, researchers, and data repository stakeholders; and (3) case studies that demonstrate how to apply these novel

techniques. The "big picture" questions to be addressed through this effort include the following: Does it work? Is it fast enough and scalable? Is it acceptable to stakeholders? What are the policy implications? Can the evaluation of re-identification risk for multiple queries and the addition of outside data be automated?

This work also holds the potential for informing a new field of study, geospatial cryptography, that focuses on the unique problems that can be addressed using cryptographic approaches to geospatial analysis.

Introduction to Cryptography and Multi-Party Computation

Dr. El Emam began by referencing three books that provide information on the de-identification of data:

- Risky Business: Sharing Health Data While Protecting Privacy,[2] a policy/legal review that includes a collection of articles.

- Guide to the De-Identification of Personal Health Information,[3] focused on methodology, with guidance on how to de-identify data and measure risk.

- Anonymyzing Health Data,[4] a book of case studies, expected to be available this fall.

Secure computation is a set of techniques (protocols) developed to allow computations to be performed on large amounts of encrypted data (i.e., to carry out analytics without knowing or exposing the raw data). Example computations include: (1) rates and categorical data analyses associated with public health surveillance, (2) rare adverse drug event detection using regression models for distributed data, and (3) secure matching (i.e., record lookup without revealing the record details, matching databases without revealing matching keys). Homomorphic encryption allows one to carry out the mathematical computations necessary for secure computation without decrypting the data and does not require any hardware additions. The concept of homomorphic encryption has been in existence for many years—only recently, however, has it been operationalized to solve practical problems.

With unrestricted access to databases, users can construct queries in such a way that different types of overlapping information result. Users can then start reconstructing the original datasets. Although this may require running many queries, but it is a plausible scenario for databases that grant have unrestricted access. There have been a number of highly publicized cases in which teams of researchers conducting what they view as a "public service" have been able to de-identify publicly available hospital discharge and other data.

Dr. El Emam then described encryption using a public key that is shared and used to encrypt the data in conjunction with a private key that is not shared and used to decrypt the data. Randomized public key encryption adds a layer of security and complexity such that if the same

value is encrypted multiple times, a different encrypted value results each time because the encrypted value is randomized.

There are three primary concerns when developing these types of protocols:

- Sharing secrets. If one is going to share data among organizations, they should not share "secrets" (e.g., keys, passwords, etc.) so that data are not compromised (which is why public and private keys are used).

- Brute force attacks. If a person's social security number is encrypted, but not through randomized encryption, someone trying to identify an individual could run an algorithm trying all possible social security numbers until a match is found. Randomization and use of a secure, secret key are ways to prevent these types of brute force attacks.

- Frequency attacks. A simple example is using a database of names. If a state's publicly accessible voter registration database is available, and the most common name in the state is "James," then it is possible to take the dataset and examine the frequency of encrypted values representing names and identify "James" as the most common value, and many values for encrypted data can be found. Again, randomization and use of a secure, secret key can protect against frequency attacks.

Additively homomorphic encryption involves performing a mathematical operation (in this case, homomorphic multiplication) on two encrypted values, "a" and "b" ([a] · [b]), which results in the sum of these encrypted values ([a + b]). Those who possess the key can then decrypt this result for the decrypted sum of values "a" and "b." A number of homomorphic encryption systems allow this type of operation. For the most part, Drs. Jacquez and El Emam are using the additive homomorphic cryptosystem of Paillier[5] in this project.

The basic concept behind the typology Drs. Jacquez and El Emam are working on involves a data owner, who receives a public key and utilizes some form of cloud computing infrastructure that runs the computations on encrypted data, which would be available to researchers who make requests for analytical work. Encrypted results are then sent to the researcher, who is given a private key for decryption. In some cases, the person with the private key could also be the data owner. For these types of protocols, three types of assumptions can be made regarding third parties: (1) Trusted third parties can access personal health information, share secrets, and analyze data without encryption. Trusted third parties are very expensive, and their use becomes complicated when they cross jurisdictions. (2) For semi-trusted third parties, the only assumption made is that the semi-trusted third party will adhere to the protocol, and with properly designed protocols, it should be impossible for them to see any confidential information if the inputs and outputs are properly controlled. This is the assumption/model used in health care settings. (3) The assumption for malicious third parties is that they will try to modify the protocol, change the computations, etc. Protocols designed to protect against malicious third parties (e.g., in areas such as national security and banking) are extremely slow and complex.

Dr. El Emam reviewed a number of real-world secure computation protocols. The first involved the surveillance of antibiotic-resistant organism colonization infection in long-term care homes (LTCHs) in Ontario. The objective of this work was to compute colonization rates without knowing the values for any single LTCH, and it provided the first-ever picture of methicillin-resistant *Staphylococcus aureus* in Canadian LCTHs at a time when access to infection rate information among LCTHs across the country was not legally required. Through the protocol, Dr. El Emam and colleagues demonstrated that it secure computation could be used to compute statistics while making it impossible to reverse engineer values for individual LTCHs. While the identities of the LCTHs are not anonymous, the values were protected. This project allowed LCTHs to receive benchmark feedback information while not revealing their individual infection rates to the outside world. It was found that participation in this exercise was much higher than that of a subsequent exercise that did not utilize secure computation (i.e., did not provide assurances that data could be linked back to individual LTCHs).

One similar problem or example relative to NAACCR is the census tract poverty indicator, which is a code indicating the poverty level where a person lives. Currently, the preferred approach is that state registries send in their census tract information, and then NAACCR calculates the data and assigns the census tract code to each person. The problem is that many states (roughly 20) will not send their census tract information to NAACCR, because some registries/states do not view NAACCR as a trusted third party. To increase the number of states involved in this activity, a process similar to the protocol used for the LCTHs in Ontario could be adopted.

Another secure computation protocol involves computing bivariate statistics from registries conducting data on human papillomavirus (HPV) vaccination, and linking this with other types of registry data to understand who has been vaccinated and to carry out long-term monitoring.[6] In this example, there were two registries that wanted to link records in their datasets. The most effective records to link on are personal identifiers, but the registries were not willing/able to share their datasets due to privacy concerns. A linking and surveillance protocol was developed that allows for computation of bivariate statistics on an ongoing basis while concealing the raw data.

Secure computation also was used to help build a protocol for rare adverse drug event detection.[7] This secure multi-party computation protocol allowed for pooling data from multiple sources using a central analysis center. The protocol computes the model parameters in a distributed way, without any of the raw data belonging to another site being seen. In essence, it simulates the results of the pooling without actually pooling the data and without sharing any of the data, with no loss in accuracy. This is being applied in an international project involving the United States, United Kingdom, and Canada, and shows the potential of this type of approach for overcoming IRB-related concerns that prevent groups from sharing data.

In another project, a secure matching protocol that allows for securely linking data sets without sharing any personal information was developed.[8] This approach is useful for de-duplication of databases and for secure lookup (e.g., to determine whether a dataset meets certain criteria without actually accessing the data set). The mathematical model that performs the encryption is

public. The best fields to link databases on are very sensitive (e.g., health insurance number, social security number, medical record number, etc.). Organizations do not have the authority to exchange these data, but need to de-duplicate databases or perform look-ups. Anonymous linking allows for the linking of records in remote databases without sharing any sensitive or personal information or sharing any secrets. One example with relevance to NAACCR is the case in which a person has residences in two different states and receives cancer care from facilities in both.

Dr. El Emam gave an example of an anonymous linking protocol using the Ontario Brain Institute and a provincial data custodian, walking the group through the following steps: (1) generation and distribution of keys; (2) encryption of OHIP# using a public key; (3) encryption of local OHIP# using the same public key; (4) perform homomorphic equality test on the two encrypted values; (5) decrypt the results of the equality tests using the private key; and (6) results of matches then can be used to de-duplicate, link, or return a lookup outcome. Automating the process using this technique greatly speeds up the process, which for some data sets can take a year or more, to near real time.

In de-identification, granularity is lost to some degree. With secure computation, there is no loss of granularity, the user is working with the original raw data and the only loss in accuracy compared with doing the computation on the original raw data is due to the conversion of integers to real numbers and back. This can be scaled so that the errors are 10-15 decimal digits down (i.e., the practical impacts are very small). Dr. El Emam commented that he has never had any meaningful error introduced as a result of these conversions and has seen no loss in accuracy.

The general problem that Drs. Jacquez and El Emam are trying to solve is referred to as cluster detection (a type of n^2 comparison). In the first iteration of their work, they established a baseline comparing every point against every other point. Moving forward, more efficient nearest neighbor computations are being attempted using the Hilbert Curve and locality sensitive hashing. Secure versions of these schemes are being sought and are expected to improve performance dramatically, particularly with regard to geospatial clustering work. Challenges facing the development of these techniques include determining how much pre-computation the data owner should perform and to what degree the necessary analytics can be anticipated, minimizing the complexity of 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.

Potential Applications/Benefits of Homomorphic Cryptography

Workshop participants identified a number of fundamental and basic applications where this technology can now be applied, as well as some of the benefits associated with its application. These applications are described in the following paragraphs.

A virtual data linkage project being led by Dr. Dennis Deapen, Director of the Los Angeles Cancer Registry may be able to take advantage of homomorphic cryptography approaches. This

effort involves an exhaustive cancer study that links to every registry in the country, and therefore necessitates having to run the protocol by 50 IRBs. If the data can be shared in a protected way, researchers can quickly identify numbers of cases in each state and prioritize before submitting their protocol to an IRB, rather than going through the time-consuming IRB process first and then still facing the prospect of collecting data only to find nothing of interest. Aggregating 10 states' data and identifying a substantial number of cases that are worthy studying could save years' worth of work. This approach could also be used to prioritize areas of research interest. The case can be made to IRBs that the risk of disclosing confidential data is extremely small, and there appears to be willingness for IRBs to accept these types of approaches to allowing the sharing of data. This type of approach has allowed for expedited IRB review in Canada. It was noted that the Centers for Disease Control and Prevention (CDC) funded Westat to develop inventory of IRB approval, starting with virtual linkages, to inventory each state's policy. This resource is available to the public online, and intended for use by researchers conducting national and multistate studies. It includes a glossary of terms, resources, etc., and can be found at www.cancerirbassistance.org.[9]

Workshop participants suggested that homomorphic encryption could be used to expand the number of data elements in NAACCR's CINA Deluxe product, which includes a collection of data items provided by registries. This would make CINA Deluxe more attractive to researchers and likely would increase its use. This could lead to research projects that, for example, analyze rates of rare cancers and compare a given state's rates with data for the entire United States.

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. NAACCR has a Fitness for Use Workgroup that is trying to develop benchmarks for whether data should or should not be used for various purposes based on level of completeness—the potential uses of homomorphic cryptography this may help inform their work.

It was noted that some state registries (e.g., Missouri, Kansas, Minnesota, etc.) have more severe restrictions on sharing data compared with others. In fact, some states can only participate in research in a very limited capacity, if at all, because their data are so heavily protected. Some state registries cannot submit data to NAACCR in the way NAACCR would prefer due to state regulations (e.g., legislation in some states allows them to submit county-level data to the CDC, but not to NAACCR). Workshop participants suggested sharing this approach with these more restrictive registries in an attempt to ascertain whether its use would help relieve some of these issues. It was further suggested that NAACCR leadership could be tapped to help identify candidate states.

Similarly, some organizations share data with CDC, NAACCR, and other entities, but do so at a greater level of detail with groups that can provide certain legal protections that NAACCR cannot (e.g., CDC). If NAACCR were to create this type of environment, might persuade these groups to share more detailed data and could allow for more freely exchanging information.

For the most part, derivations on collected data items are done by registries. Each of the programs run to derive a value costs time, and more of these are accumulating. If these derivations could be carried out on an encrypted server, it would save significant time. In fact, the prospect of saving significant amounts of time, which is tied to many of these applications, was cited by Workshop participants as a significant argument for "buy-in" on the part of NAACCR, registries, and other stakeholders.

NAACCR has been running the same 16 duplicate protocols for the last 14 years—a program that utilizes encryption to take on this task would save significant time and could be used for multi-state de-duplication, which is extremely challenging (and impossible for NAACCR, which does not have the personally identifiable data). Some of this de-duplication is done using the National Death Index, but this is a labor- and time-intensive process. It was suggested that a pilot of willing registries could be initiated to determine if a homomorphic cryptography approach to these issues would show promise for use by NAACCR.

Similarly, expanding cancer cluster analyses and de-duplication beyond state lines (and potentially, the U.S.-Canada border) is a potential application for this homomorphic encryption. Workshop participants envisioned a system through which data could be uploaded to an encrypted system at NAACCR that would provide the results for all states in an area of concern without identifying the data from each state. Currently, there is no practical way to accomplish this type of geospatial data sharing. Not communicating across state lines leads to duplication and creates false hot spots. Solving the de-duplication issue would represent a tremendous advance, and a protocol that could be used in this capacity is up and running in Ontario. Dr. El Emam and colleagues are at the point of addressing minor usability issues, and the technical work is complete.

Workshop participants agreed that issues relating to data privacy represent substantial obstacles to basic geographic information system operations. Dr. Daniel Goldberg created a geocoder which is being hosted at Texas A&M University (and is free to use for up to approximately 2,500 records).[10] NAACCR members can submit the addresses of their cases to this geocoder (a number of private organizations use it as well). The address information is known to Texas A&M and/or NAACCR, and then returned to the user and deleted. Although there are vulnerabilities in this approach, most NAACCR registries have grown comfortable with using this geocoder. With online geocoding systems, whoever is providing the data is essentially providing personal information (e.g., address) to the geocoding system. Using a secure string comparison scheme, it may be possible to carry out secure geocoding such that whoever is providing the data encrypts it, submits it, and then can do the geocoding without actually knowing the exact addresses. More information is needed to determine whether releasing address information to geocoding systems is enough of a privacy concern to warrant work in this space. Workshop participants indicated that the NAACCR community would be very enthusiastic about the use of homomorphic encryption to safeguard data in conjunction with this geocoder. Dr. El Emam suggested that this could serve as an interesting use case representing a fundamental step forward in geospatial analysis and could easily be done with their protocol.

Activities related to data linkages was identified as an area that could benefit from the use of homomorphic encryption. In Alaska for example, many addresses are P.O. boxes and not the

physical location where individuals live, especially in remote areas. Furthermore, some states are reluctant to give physical location information to registries.

The National Cancer Institute (NCI) is planning to establish a cancer repository focused on genetics through a distributed data model. This "cancer knowledge cloud" is aimed at carrying out next-generation computational capabilities to support biomedical data infrastructure and the NCI is and is looking to the research community to help determine what shape it could take.[11] Participants agreed that homomorphic cryptography could play a role in this entire enterprise.

Although not used widely by the cancer registry community now, homomorphic encryption could eventually lead to the possibility of incorporating secure genotypic information into the record and making this information available to researchers. Another potential future use of this approach could be flexibly aggregating geocoded spatial data to another scale through a secure data aggregator.

Dr. Jacquez sorted some of the more promising potential applications in aspatial and spatial groups. Aspatial applications include de-duplicating, aggregation across boundaries, secure geocoding (although it was noted that at this point, it is unclear whether there is a benefit in this area and whether it is feasible), and secure data linkages.

Additional input from Workshop participants included the addition of suppressed data elements that researchers would like to use (e.g., birth year) but have been unable to do so under certain circumstances. It was noted that national data that has this type of information in an aggregated form are lacking for some fields. Detailed facility/provider information also was identified as a suppressed data element that would be useful (facility identifiers are not allowed to be submitted to NAACCR).

In terms of potential spatial applications, work related to weights, use of a surveillance/scan statistic to identify clusters, and new ways of accessing and using Census data and tax information (e.g., identifying the denominator) were identified. In the future, it may be that these approaches lead to the development of a "geospatial toolbox" for geospatial analysts. They also may play a role in hypothesis generation while allowing for freer access to data within a controlled environment (i.e., less stringent protocols and more efficient, expedited work without unnecessarily compromising privacy). In general, it was noted that homomorphic encryption techniques could enrich surveillance data in many research areas.

Potential Stakeholders and Dimensions to be Considered

Workshop participants engaged in an exercise to identify stakeholders (i.e., those who will need to be convinced of the utility of this approach). Identified stakeholders include IRBs, researchers, funding agencies, data repositories/registries, individuals, facility/provider data, industry/drug companies, and CDC's National Center for Health Statistics (NCHS endorsement would move this advance this process and could be a potential funding agency in the future for this work). Participants also identified a number of dimensions to be considered for assessing the usability, benefits, and implementation. These include technical, policy, existing regulations, legal, capacity/infrastructure, and governance.

Next Steps

Workshop participants commented that for an approach such as homomorphic encryption to be adopted by NAACCR, the most likely mechanism would be testing it with some volunteer state registries over a few years. For some of the potential applications discussed (e.g., cluster analyses), two states sharing a common border would be necessary. NAACCR leadership has been open and enthusiastic about moving the field forward and considering novel approaches for doing so.

Candidate Projects/Activities

Workshop participants discussed candidate projects/activities that could serve as demonstrations to address significant problems that can be done with existing tools but without much investment to address important problems. It was agreed that the outcome(s) of any of these activities would be of great interest to NAACCR. A number of scenarios were discussed, including:

- Pilot projects (which do not require significant effort or funding and can be put in place quickly to demonstrate value). Any pilot would have to be small if no additional resources are brought in, and should be designed to produce immediately useful results that could be used to strengthen an R01/U01 application.

- A U01/R21/Cooperative Agreement-type project on geocoding or de-duplication/aggregation that would require some research component and would likely be lead by an NIH Institute/Center (or possibly NAACCR) that would participate in a steering project to some extent, with a particular outcome in mind.
- An R01 (which requires roughly 1.5 years of research) that would advance the work of the current R21 grant.

- A de-duplication demonstration funded by the CDC (which would have an interest in trying to make this operational because of the impact on national statistics).

It was suggested that the outcome of a pilot project could be a feasibility assessment for a "national tumor index" involving a few states and a focus on de-duplication/aggregation. For the pilot, two (or more) states could be selected along with one or two statistics that would be computed from the pool of de-duplicated data to demonstrate feasibility and that the data are protected. With regard to de-duplication, challenges arise in determining ownership of the data once it is identified (and aggregation cannot be done until this is performed).

One Workshop participant suggested de-duplication and ascertaining frequency by cancer site. It is likely that duplicates are skewed in certain cancers across state lines differently, and demonstrating this could generate significant interest among funding agencies. It was also

suggested that a potential de-duplication project focus on retirement states that have high populations of "snowbirds" (e.g., Florida, New York, Texas, Arizona, etc.).

Participants discussed NAACCR's 16 criteria for de-duplicating a registry (e.g., last name, first name, middle initial, social security number, etc.) and the potential for incorporating them and/or other variables that work across all cancers and are recorded easily into a project that would use the de-duplication protocol across each of these criteria. It was suggested that age, sex, cancer stage, and cancer site be included (some participants suggested using only age, stage, and site) to demonstrate bias in the data and identify an area where basic enrollment would yield a sufficient number of events to support a research project. Dr. El Emam pointed out that a homomorphic encryption protocol already exists that could be used for de-duplication. It would require only minor tweaking and could be adapted for this use quickly if a few registries are willing to participate in a project (participants indicated that they felt a number of registries would be eager to participate).

It was noted that for a pilot project, the geospatial work is not ready to be scaled up, and that the focus should be on de-duplication and aggregation (one participant commented that in documenting the effect of the de-duplication, the ability to aggregate is being demonstrated).

One approach could be to encrypt the data and submit it to NAACCR, which would de-duplicate and compute the statistics. NAACCR then would send the results to the investigators, who would decrypt the data. Part of a pilot project could also involve documenting any issues/challenges with getting IRB approval. It was suggested that a pilot could be simplified by not including NAACCR, and instead slightly modifying an existing IRB protocol with two registries sending encrypted data to each other as a proof of concept.

Workshop participants also discussed a potential pilot project addressing geospatial problems related to distance to facility analyses (e.g., at late-stage diagnosis, how far was a patient from the screening facility?). This could evolve into a full research project, with NAACCR as a potential partner. Showing the value of the secure computation approach over using synthetic data (e.g., the same data quality with a faster return of results) may be an interesting topic for a future proposal as well.

The NCI, in collaboration with NAACCR (and possibly other groups) is the most likely funding agency that could provide funding for these types of pilot projects.

Potential Publications

There was general agreement that these discussions identified sufficient potential benefits and applications to warrant drafting a peer-reviewed publication on the policy implications related to the use of cryptography for improving public health. A broad policy paper was envisioned that outlines the types of obstacles and issues that can be addressed in conjunction with a discussion on the enormous potential held by data de-duplication and aggregation using this approach. Additional areas for consideration were identified, such as cross-disease studies and studies using claims data. Policy issues could be based on overall experiences and/or presented from

NAACCR's perspective, relating to cancer registry data. Participants suggested the *Journal of Cancer Registry Management* and the *Journal of the American Public Health Association* as candidate publications.

Workshop participants also suggested the possibility of drafting a separate paper focused on issues related to the implementation, potential uses, and limitations of secure computation. A simple way to document and communicate with lay people to inform them that this is a reliable way to protect privacy is needed. A peer-reviewed article describing the protocol could serve this purpose. It was noted that a technical publication based on the R21 grant is in preparation.

Potential Challenges

Workshop participants discussed some of the potential challenges facing the implementation of homomorphic encryption in the cancer registry environment. The quality of data varies widely across registries from state to state. One participant noted that it may even be possible to visualize these variations geographically. The ability to mitigate or qualify differences in data quality is necessary. There are also significant differences across registries in terms of capabilities, budgets, and confidentiality/privacy regulations that must be accounted for.

Educating IRBs will be necessary to increase the use of homomorphic encryption in the research enterprise.

NAACCR was identified as a logical semi-trusted partner for pilot activities, and it is likely that NAACCR would be a motivated partner, but it likely does not have the internal IT infrastructure necessary for these activities. NAACCR could potentially contract out this service.

Closing

In closing, Workshop participants emphasized that homomorphic cryptography approaches hold great potential for improving work relative to NAACCR's mission. It can open doors leading to new work that is not currently being done due to concerns about keeping data secure. There are likely national cancer studies that are not being conducted because of these concerns. Fields other than cancer, many of which have less sensitive issues related to protecting confidential data, likely will have a strong interest in this area as well. Homomorphic cryptography carries with it broad appeal and great promise, and it was suggested that it be demonstrated in the cancer world first. Following these comments, Dr. Jacquez thanked Workshop participants, noting that the discussions have been extremely useful and informative, and adjourned the Workshop.

Appendix A: Exploratory Evaluation of Homomorphic Cryptography for Confidentiality Protection: Project Description

Protecting confidentiality of patients and study participants is mission-critical across the health care continuum, yet poses obstacles for the sharing and analysis of health data. Confidentiality protection can retard research from the individual to the population level, but solutions to this important problem are often unsatisfactory and even absent in many applied settings. Recent advances in cryptography for the first time support the analysis of confidential 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 research since confidential data would not have to be decrypted to allow analysis and dissemination of the results, but this potential has yet to be evaluated within the context of human subjects research. This project will perform an exploratory evaluation of homomorphic cryptography for the geospatial analysis of confidential health data and will accomplish four aims:

Aim 1: Build a prototype secure multi-party computation (SCM) platform for the computation of mathematical operations required in geospatial analyses. The platform will implement geospatial analysis protocols on encrypted data such that the identity of individuals cannot be reconstructed or deduced. We will evaluate security of the approach using external expert testing designed to reconstruct identity of individuals. This exploratory project will evaluate the computational performance of the following geospatial operations on encrypted data: (1) Spatial weight calculations for residential locations; (2) cluster/hotspot analysis; (3) calculation of rates of late stage diagnosis by race; and (4) calculation of relative and absolute disparities in stage at diagnosis. These have been selected to be representative of the geospatial computations frequently undertaken in geohealth analyses.

Aim 2: Apply the prototype systems to assess racial disparities in stage at diagnosis for prostate and breast cancers. This will evaluate practical feasibility using previously analyzed data, and will determine whether the results with the not-encrypted data are reproducible.

Aim 3: Formally evaluate the approach and formulate recommendations using an independent working group convened by the North American Association of Central Cancer Registries to include stake-holders including health researchers, IRB Chairs and committee members, experts in confidentiality protection, Directors of disease registries and cryptographers.

Aim 4: Disseminate the recommendations and results of the feasibility analysis though peer-reviewed publications and presentations at scientific meeting.

This highly innovative and high-impact project potentially will accelerate human health research, leading to earlier advances in treatment and improvements in our nation's health. The National Institutes of Health is investing 100's of millions in interoperable electronic health records that are expected to revolutionize health care and disease control and surveillance. Most of the data records for these systems include personal identifiers – Names, addresses, and

related health information, whose confidentiality must be protected under HIPPA and other regulations. However, confidentiality protection is proving to be a major impediment to public health research. This project will evaluate the feasibility of homomorphic encryption technology that make possible, for the first time ever, analysis without having to decrypt the data. This advance is expected to significantly accelerate research that involves accessing and/or linking confidential data.

Appendix B: References/Additional Resources

1. Wartenberg DE, Thompson WD. Privacy Versus Public Health: The Impact of Current Confidentiality Rules. *Am J Public Health*. 2010;100(3): 407-411.

2. El Emam K, ed. Risky Business: Sharing Health Data While Protecting Privacy. Bloomington, IN: Trafford Publishing; 2013.

3. El Emam K. Guide to the De-Identification of Personal Health Information. Boca Raton, FL: CRC Press; 2013.

4. El Emam K, Arbuckle L. Anonymyzing Health Data. Sebastopol, CA: O'Reilly Media, Inc. [in press].

5. Paillier, P. Public-Key Cryptosystems Based on Composite Degree Residuosity Classes. EUROCRYPT'99, Prague; 1999.

6. El Emam K, Samet S, Hu J, Peyton L, Earle C, Jayaraman GC, 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. Epub 2012 Jul 2.

7. El Emam K, Samet S, Arbuckle L, Tamblyn R, Earle C, Kantarcioglu M. A Secure Distributed Logistic Regression Protocol for the Detection of Rare Adverse Drug Events. *J Am Med Inform Assoc*. 2013 May 1;20(3):453-61. Epub 2012 Aug 7.

8. 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 Inform Assoc*. 2011 May 1;18(3):212-7.

9. www.cancerirbassistance.org. (The CDC funded Westat to develop inventory of IRB approval, starting with virtual linkages, to inventory each state's policy. This resource is available to the public online, and intended for use by researchers conducting national and multistate studies. It includes a glossary of terms, resources, etc.)

10. http://geoservices.tamu.edu/Services/Geocode/. (A widely used geocoder developed by Dr. Daniel Goldberg that is hosted at Texas A&M University.)

11. http://cbiit.nci.nih.gov/ncip/nci-cancer-genomics-cloud-pilots. (The National Cancer Institute (NCI) is planning to establish a cancer repository focused on genetics through a distributed data model. This "cancer knowledge cloud" is aimed at carrying out next-generation computational capabilities to support biomedical data infrastructure and the NCI is and is looking to the research community to help determine what shape it could take.)