Creation of public use files: lessons learned from the comparative effectiveness research public use files data pilot project

Erkan Erdem and Sergio I Prada

IMPAQ International

13. September 2011

Creation of Public Use Files: Lessons Learned from the Comparative Effectiveness Research Public Use Files Data Pilot Project

Erkan Erdem[1] & Sergio Prada[2]
[1]Senior Research Associate, IMPAQ International, LLC, eerdem@impaqint.com
[2]Research Associate at IMPAQ International LLC, contact: sprada@impaqint.com

Abstract

In this paper we describe lessons learned from the creation of Basic Stand Alone (BSA) Public Use Files (PUFs) for the Comparative Effectiveness Research Public Use Files Data Pilot Project (CER-PUF). CER-PUF is aimed at increasing access to the Centers for Medicare and Medicaid Services (CMS) Medicare claims datasets through PUFs that: do not require user fees and data use agreements, have been de-identified to assure the confidentiality of the beneficiaries and providers, and still provide substantial analytic utility to researchers. For this paper we define PUFs as datasets characterized by free and unrestricted access to any user. We derive lessons learned from five major project activities: (i) a review of the statistical and computer science literature on best practices in PUF creation, (ii) interviews with comparative effectiveness researchers to assess their data needs, (iii) case studies of PUF initiatives in the United States, (iv) interviews with stakeholders to identify the most salient issues regarding making microdata publicly available, and (v) the actual process of creating the Medicare claims data BSA PUFs.

Keywords: Public use files, PUFs, re-identification, de-identification, Medicare claims, comparative effectiveness research, confidentiality, data utility

1. Introduction & Background

As administrator of the Medicare and Medicaid programs, CMS accumulates and maintains claims data on all fee-for-service Medicare beneficiaries for different settings categorized into 8 types of claims: Inpatient, Outpatient, Skilled Nursing Facilities, Home Health Agencies, Hospice, Physician/Supplier, Durable Medical Equipment, and Prescription Drug (Part D) Events.[1] The importance of these data for comparative effectiveness research (CER) can hardly be overstated. Currently, however, researchers must not only prepare, submit, and gain approval of applications for data-use agreements (DUAs), but also pay a recovery-of-cost fee to get access to such files. In addition, researchers must be vigilant in observing legal restrictions on the use, maintenance, sharing, and final disposition of the files to which they gain access. For many researchers, the application and approval process, fees, and restrictions represent significant barriers to all kinds of healthcare related research.

As host of these valuable data, CMS understands the importance of providing improved access to them. To further this objective, as part of the *Comparative Effectiveness Research Public Use Data Pilot Project (CER-PUF),* CMS recently initiated an effort to increase access to Medicare claims data through the creation and dissemination of public use files (PUFs) for researchers and data entrepreneurs. In this context we define PUFs as datasets characterized by free and unrestricted access to any user.

[1] https://www.cms.gov/FilesForOrderGenInfo

The CER-PUF project is unique in that, while CMS currently provides aggregated data or tables with Medicare claims information, it has never before released micro-level (claim-level or beneficiary-level) data. Of paramount importance to the project is strict protection of beneficiary and provider confidentiality, in pursuit of which PUFs must comply with existing privacy laws, such as the Health Insurance Portability and Accountability Act of 1996 (HIPAA) and the Patient Safety and Quality Improvement Act of 2005 (PSQIA). The HIPAA Privacy Rule protects the privacy of individually identifiable health information and also defines how protected health information can be disclosed and used. The PSQIA Patient Safety Rule establishes a framework by which hospitals, doctors, and other healthcare providers may voluntarily report information to Patient Safety Organizations (PSOs), on a privileged and confidential basis, for the aggregation and analysis of patient safety events.[2]

The CER-PUF project is aimed at increasing access to CMS data sets through the creation of PUFs that (i) do not require user fees and data use agreements, (ii) have been de-identified and tested thoroughly to assure the confidentiality of the beneficiaries and providers is protected, and (iii) still contain significant analytic utility for end-users. To inform our PUF development process, we undertook a series of information-gathering activities. We also created Basic Stand Alone (BSA) PUFs for each of the 8 claims types listed above.

This paper is organized into lessons learned from the five major CER-PUF activities to date: a statistical and computer sciences literature review on best practices in PUF creation (Section 2), interviews with comparative effectiveness researchers to assess their data needs (Section 3), case studies of PUF initiatives in the United States (Section 4), interviews with stakeholders to identify the most salient issues regarding making microdata publicly available (Section 5), and the actual process of creating the Medicare claims data BSA PUFs (Section 6). Section 7 concludes the paper with an overall assessment of lessons learned. The project also included a review of the laws and regulations governing creation of PUFs based on Medicare claims data, which is not included in our discussion. For details on that part of the project see Thorpe (2011).

2. The Literature Review

The primary question we addressed in the statistical and computer science literature review was the following: Is there a consensus regarding best practices for creating PUFs? We summarize our main findings here; for detail we refer the reader to Prada et al. (2011). Although we found a great deal of information about possible methods, we found no consensus on any of the fundamental questions we sought to answer.

2.1. Is there a shared framework for analyzing disclosure risk in the literature?

Disclosure is the communication, either directly or by inference, of information about a member of a dataset that could not be known without viewing the dataset. Disclosure takes place if someone extracts information about any person (or other entity) from the dataset. The literature calls this individual the intruder. Duncan et al. (1993) distinguish three types of disclosure: (i) when a data subject is identified from a released file (identity disclosure); (ii) when sensitive information about a data subject is revealed through the released file (attribute disclosure); (iii) when the released data make it possible to determine the value of some characteristic of an individual more accurately than otherwise would have been possible (inferential disclosure).

[2] U.S. Department of Health & Human Services; http://www.hhs.gov/ocr/privacy/index.html

Hundepool et al. (2010) argues that a general framework for addressing disclosure risk should address the following questions in a cohesive and consistent way:

- Why is confidential protection needed?
- What are the key characteristics and uses of the data?
- What disclosure risks need to be protected against?
- What disclosure methods are the most efficient?
- How much utility is lost?
- What is the risk of re-identification?

We searched the literature for answers to these questions within a single cohesive framework, but found instead two competing research paradigms.

Paradigm 1 holds that it is indeed possible to minimize the risk of disclosure and therefore to release data to the public. This paradigm comes from the statistical literature on disclosure limitation techniques and their achievements, which is extensive (see Prada et al. 2011 and Duncan et al. 2011 for recent reviews). Consistent with Paradigm 1, this literature is devoted to developing methods and software to mask data. Multiple methods are available, from simple (i.e., coarsening) to complex (i.e., synthetic) methods. Paradigm 2, in contrast, holds that privacy and confidentiality cannot be achieved in an environment in which personal information is gathered at an increasing rate by multiple people with multiple interests. This sharply contrasting paradigm comes from the computer science literature. For instance, computer scientists Narayanan and Shmatikov (2010) criticize the types of de-identification techniques developed by Paradigm 1 advocates as based on the assumption that personally identifiable information is a fixed set of attributes such as names and contact information, which "creates the fallacious distinction between 'identifying' and 'non-identifying' attributes." Such a distinction might make sense in the context of one attack, these authors say, but is increasingly meaningless as the amount and variety of publicly available information about individuals grows exponentially. In a similar vein, Cynthia Dwork and Moni Naor argue that the type of privacy defined by Dalenius (1977), on which the statistical literature is based ("access to a statistical database should not enable one to learn anything about an individual that could not be learned without access") cannot be achieved as a general rule. These authors illustrate the intuition for their finding with the following parable:

"Suppose one's exact height were considered a sensitive piece of information, and that revealing the exact height of an individual were a privacy breach. Assume that the database yields the average heights of women of different nationalities. An adversary who has access to the statistical database and the auxiliary information "Terry Gross is two inches shorter than the average Lithuanian woman" learns Terry Gross' height, while anyone learning only the auxiliary information, without access to the average heights, learns relatively little." (Dwork and Naor, 2010, p. 93)

2.2. Are there standards for acceptable risk?

According to Paradigm 1, data should be released if the probability of identifying an individual or entity in the data file is sufficiently small. However, there are as yet no definitive answers for practitioners regarding either a specific definition of universal risk, assumptions about the intruder, or what constitutes "sufficiently small". The definition of what is "sufficiently small" is currently up to the data producer, based on the producer's obligation to the subjects in the dataset.

In the case of healthcare data, HIPAA provides a list of eighteen (18) direct identifiers, collectively known as the Safe-Harbor Method, that must be removed to comply with the HIPAA regulations.[3] This approach, which is intuitive, could lead one to naively think the dataset is then safe from disclosure because no individual is explicitly identifiable. However, as shown by Sweeney (1997, 2000) and Agrawal and Srikant (2000), removal of direct identifiers does not protect all individuals from data disclosure or re-identification. A combination of just a few indirect identifiers (such as birth date, gender, and zip code) can be used to identify a large portion of individuals on any dataset. And these variables can then be matched to another publicly available dataset to identify individuals in the data. This is another area without definitive answers for a practitioner.

2.3. Which are the most frequently applied techniques for limiting disclosure risk?

Limiting disclosure risk can be done in two steps.[4] First, significant protection can be attained simply by using a random sample data as the source file for the PUF, rather than the full (or population) data base. Second, after selection of the source file, additional protection is possible by applying policy rules if any (e.g., HIPAA's Safe-Harbor Rule) and/or disclosure limitation techniques. The latter is known as "treating" the data. We provide a brief summary of the most frequently applied techniques. For a technical description of the limitations of each see Winkler (2007).

2.3.1. Sampling, Global Recoding, and Local Suppression

Sampling from a full database is a powerful method of protecting the confidentiality of data by creating uncertainty about whether the target record exists in the PUF. It is also important to control for that probability, however, by determining the appropriate sample size. Even though this decision is based on the size of the full database, number of variables included in it, and other characteristics of the included variables, a sample size of 1%-5% is widely accepted and used for PUFs in the U.S.[5] For example, the U.S. Social Security Administration provides a 1% sample for the 2004 Benefits and Earnings PUF, 2006 Earnings PUF, 2001 Old-Age, Survivors PUF, and Disability Insurance (OASDI) PUF, and a 5% sample for the 2001 Supplemental Security Income (SSI) PUF. The U.S. Census Bureau has been providing both a 1% and a 5% PUF from the 2000 Census of Population and Housing.

When determining appropriate sample size, data producers inevitably have to compromise between the utility of the PUF and the risk of re-identification. As the sample size is increased, the precision of the statistical estimates improves. However, this also increases the risk of re-identification by diminishing uncertainty about whether the target is actually in the sample. No agreed-upon method yet exists on the optimal way to make this compromise.

Global recoding is a process of reducing the number of values a single variable can have in a dataset. For example, if an individual's birth date exists in a dataset it can be used as an indirect identifying variable. However, recoding the variable to coarser values, such as birth year, will make it less useful as an indirect identifier. Recoding to even coarser values, such as five-year intervals, will further reduce the identifying power of the information. The appropriate level of

[3] For the list of direct identifiers see: http://www.cdc.gov/mmwr/preview/mmwrhtml/m2e411a1.htm#box2

[4] For a summary of general guidelines see the "Statistical Policy Working Paper 22 (2nd version, 2005)", Report on Statistical Disclosure Limitation Methodology, Office of Management and Budget, 2005.

[5] Using a sample also requires (i) ensuring that the estimates obtained from the sample accurately represent the population and (ii) including sample weights.

recoding also depends on the trade-off chosen between the competing needs for data confidentiality and preservation of data utility.

Local suppression is the process of removing, or suppressing, data from a dataset. This can be done for single variables within a record or for an entire record. The current local suppression rule relies on creating a file without small cells, so that no record is unique to one group of variables. This rule, currently known as *k-anonymity,* has been in use by statisticians for decades (Willenborg and de Waal, 1996). Under this rule, indirect identifying variables are recoded until each combination of recoded variables has at least *k* number of records associated with it. At this point no individual in the dataset can be identified with certainty because no individual has a unique profile.

2.3.2. Perturbation

Perturbation is a process that reduces disclosure risk by altering the values of variables in the dataset. Perturbation can take multiple forms, including selective perturbation, data swapping, substitution, and synthetic treatments. *Selective perturbation* deterministically selects records for treatment to reduce disclosure risk. Also called blank and impute, the method selects values from single records, removes them from the record, then imputes a new value (Skinner, 2009). *Data swapping* transforms a dataset by exchanging values of sensitive information between records (Fienberg and McIntyre, 2005). *Substitution* replaces some or all identifying variables in a record with the same variables from another record; it is different from data swapping in that the data only move in one direction (Singh, 2009). *Synthetic treatment,* a technique that is gaining ground, treats all records in the dataset to create a new, "synthetic" dataset that is representative of the original data file. The indirect identifying variables may be changed by a variety of methods, including perturbation, multiple imputations, and other model-based techniques (Reiter, 2009).

2.4. Is there consensus on how to measure utility loss?

Selecting the variables to be treated depends on the content of the source data, preferences of the data producer, and availability of external data sources with similar information. But every time a variable is treated its utility decreases. There are algorithms whose purpose is to make the process of recoding as efficient as possible by minimizing the amount of information loss while reducing disclosure risk. These algorithms use information loss metrics, which quantify the precision lost in the data from recoding to compare possible recoding and suppression schema and generally navigate through a decision tree comprising all possible recoding/suppression options (see El Emam et al., 2009 for detailed discussion of some of them). Information loss metrics are only useful in making decisions regarding recoding and suppression; they provide no measure of data utility. One way to get such information is by conducting a needs assessment, the topic of the next section.

3. The Needs Assessment

The needs assessment part of the CER-PUF study asked comparative researchers what they needed in PUFs from the Medicare claims database. This section summarizes our findings. For detail see Erdem and Concannon (2011).

3.1. Objective

The purpose of this part of the study was to understand the research needs of comparative effectiveness analysts in PUFs developed from Medicare claims data files. We did this through interviews with researchers selected from academia, government agencies, private companies, and non-profit organizations. The interviews were based on a discussion guide developed specifically for the project. The criterion for interviewee selection was that they had experience with Medicare claims data, were expert in CER, or both. After a screening process, 15 researchers were interviewed.

3.2. Findings

Overall, the researchers praised the availability of PUFs as an appropriate and useful development. The researchers believed that PUFs can be very helpful in health services or health policy research provided they contain a sufficient range of variables. However, the BSA PUFs we created in the first year of the project were viewed as insufficient to satisfy the complex needs of comparative effectiveness research. It was generally agreed that access to the actual claims data would still be necessary.

3.2.1. Current process for gaining data access

According to the current CMS data licensing system, researchers have to submit their research plans and request the necessary data. If approved, they have to sign a DUA with CMS, pay a non-negligible fee, and agree to significant restrictions on maintenance, sharing, and re-using the data. Not surprisingly, our interviews revealed strong agreement that the current process is both expensive and long. This combination is damaging, according to our interviewees, because it inhibits health researchers from independently pursuing research topics that are not funded as part of some large overall contract or grant. There was also agreement that lower costs could increase the quantity of research done on important health topics. Interestingly, even though interviewees complained about the lengthy and restricting approval process, many emphasized the importance of privacy concerns and the necessity of having a strict DUA.

3.2.2. Potential of PUFs for Research

Researchers were enthusiastic about the availability of micro level PUFs in general. Our interviewees agreed that Medicare claims PUFs would significantly increase exposure to Medicare claims files--allowing users to create descriptive statistics, analyze basic relationships, formulate hypotheses, answer high-level questions, and perform preliminary analyses for research projects. It was also widely agreed that PUFs could provide an opportunity for researchers to conduct preliminary analyses for pilot studies prior to spending considerable time and money obtaining a DUA. There was consensus, however, that without more variables than in our project's BSA PUFs, such as detailed diagnosis and procedures, PUFs alone would not be sufficient for most CER studies. Since obvious challenges are involved in creating PUFs with high analytic utility for CER while maintaining confidentially, most researchers agreed on the importance of establishing appropriate expectations of what the PUFs involve and how they can best be used. It was also emphasized that even enriched PUFs will not be enough, and that researchers will still need a method by which they can have access to Medicare identifiable files or limited data sets.

Interviewees emphasized that most CER requires multi-year datasets in which beneficiaries can be tracked over time and across different types of care.[6] Hence, stand-alone PUFs created with data from a single year and including only one type of care (e.g., Inpatient) for a specific calendar year would probably not be very helpful. Researchers also agreed that a PUF that is linkable to other PUFs as well as to external data sets would have much higher utility.

[6] Ideally, these should allow for chronological sequencing of treatments and hospitalizations for each beneficiary.

Other information interviewees expressed interest in accessing include:

- Providers, such as an encrypted ID, characteristics of provider and/or healthcare setting (e.g., primary care physician or specialist, size of institution),
- Geography, such as state of residence, zip code, hospital referral region (HRR), or urban/rural designation,
- Race/ethnicity of beneficiaries,
- Supplementary insurance, such as dual eligibility (i.e., Medicaid) or other insurance,
- Health outcomes, such as mortality, morbidity, survival, discharge/transfer, re-admission, and major clinical events.

All interviewees emphasized the importance of a comprehensive codebook and sufficient documentation. There was also some feedback favoring (1) a simple user interface that allows for basic statistical analyses and tables with the PUFs, and/or (2) a responsive technical assistance and support team that would allow users to ask questions.

Finally, several interviewees were concerned that PUFs could inadvertently increase the number of studies with incorrect assumptions and, therefore, misleading results. This led to the suggestion that a framework for training researchers on correct data use be considered. It was widely agreed, however, that the alternative of not expanding public access to Medicare data would be worse than the risks of triggering improper research.

3.2.3. De-identification methods

Given the various methods of de-identification listed above in subsection 2.3, most interviewees favored suppression as the only acceptable approach, on the grounds that suppression would remove outlier observations from the data while leaving all other records unaltered. Some of the researchers feared that the high level of de-identification needed for PUFs based on Medicare claims data would decrease the analytic utility of any PUF significantly.

One suggestion offered for dealing with researcher concerns over working with perturbed data was to offer the service of re-run completed SAS, SPSS, STATA or other programs on unperturbed data,[7] which would allow data users to validate their PUF-based research results. This suggestion was made particularly in the context of addressing peer reviewed journal concerns about publishing results based on perturbed data.

Although all discussion partners viewed lack of a DUA for PUFs as favorable, those with experience creating PUFs suggested that CMS have a registration process, ask for a minimum set of information from the user (e.g., name, address, phone number, organization), and require a commitment to comply with a short list of rules.

One researcher raised the risks of including any geographic identifiers, given the potential for re-identification through the "mosaic effects" of multiple files overlaid on each other, or through data "mash-ups" from multiple files and multiple sources. Such risks are eliminated within the BSA PUFs, which were created from disjoint samples of Medicare beneficiaries.

[7] This could be accomplished by a data enclave or a remote data center.

4. Case Studies of other PUF Initiatives

To elicit lessons learned from other PUF data initiatives, we undertook six case studies of individual-level data initiatives in the US.

4.1. Objective

The objective of the case studies undertaken for the CER-PUF project was to provide CMS with instructive information for use in its CER-PUF initiative. For this purpose, we chose from a number of projects identified as representative of the wide variety of existing PUF initiatives, de-identification methodologies, and data access methods. We ensured inclusion of a range of initiative types, by choosing case studies based on the following project characteristics: domain, sponsoring entity type, project organization, de-identification methods, data access restrictions, and data access methods. For detail on the selected initiatives, choice of data access methods, and initiative-specific features see Prada, S. 2011.

4.2. Findings

Four of the six case studies are PUFs and two are Non-Public Use Files (non-PUF), defined as files characterized by access restrictions that oblige users to reveal their identity and intentions for use. Typically, non-PUFs demand signed DUAs before granting data access.

4.2.1. Disclosure risk standard

Among the four case study institutions that provide PUFs by our definition (i.e., with no restrictions), we found that the Confidentiality and Data Access Committee's (CDAC) *Checklist on Disclosure Potential of Proposed Data Releases* is widely used as the tool to assess disclosure risk of proposed data. We found, in contrast, that the CDAC Checklist is not used by the two non-PUF initiatives. One described its decision-to-release process as a "judgment call" based on its knowledge of both the data and the data users; the other simply relies on the individual risk analysis made at the original data source.

None of the six initiatives uses a theoretically established risk framework. Nor did we find any formal definition of a "safe threshold" (beyond the HIPAA "Safe Harbor" method) to judge whether a candidate file will be considered ready for release. By "safe threshold" we mean a specified percentage of records at risk of re-identification within the file above which a file would not be considered ready for public release. In all six interviewed institutions, such decisions are made on a case by case basis.

4.2.2. Disclosure Review Board or similar panel

All four of the case study initiatives that release PUFs have followed the recommendations in Statistical Policy Working Paper 22 and centralized their review of disclosure-limited data products by establishing a DRB or similar panel. A common practice is for a completed CDAC Checklist memo to be submitted to the organization's DRB or similar panel for review. Interestingly, DRBs are nonexistent for the non-PUF initiatives in the case study. None of the documents reviewed by DRBs or any of the DRB decisions made are available to researchers; and no information on data disclosure avoidance steps taken is released by non-PUF institutions. Interestingly, in the one case where the PUF includes information from two different agencies, the DRBs of both are required to approve release of the file.

4.2.3. Geographic information

Detailed geographic indicators are generally stripped from the PUFs included in our case studies. This practice follows CDAC's Checklist suggestions, as geography is a key factor in enabling identification. Geographic indicators are not stripped from the two non-PUF case study initiatives, however. In one of the two cases (the Surveillance, Epidemiology and End Results (SEER) Program of the National Cancer Institute) state and county identifiers are available, and there are no minimum population requirements. In the other case (the Healthcare Cost and Utilization Project of the Agency for Healthcare Research and Quality), some states release zip codes at the 3- and some even at the 5-digit level.

4.2.4. Weights

PUFs whose data are collected through surveys only release final weights. They do not release the components that make up the final weight because these may be indicative of geographic areas. They usually do not release the actual PSU and strata identifiers either, because that can be risky as well. Instead, they provide pseudo-strata and pseudo-PSU variables containing less information than the replicates. Replicate and/or bootstrap weights are provided for variance estimation. In the case of the Census, each PUF file contains a weighting factor for each population and housing unit record to enable full population estimates. These weights are not adjusted after disclosure limitation methods are applied, however.

4.2.5. Professional expertise required

We found that the level of expertise these institutions use to create both PUFs and non-PUFs is high, interdisciplinary, and considered a scarce resource. Several senior statisticians and mathematicians are involved in the typical PUF creation process. Additional experts are also involved, including (i) DRB members and other senior staff who review the files before release, (ii) data analysis experts such as economists and epidemiologists (depending on the data collected), and (iii) expert statisticians at the firms contracted to create the actual PUFs. One of the case studies revealed that there is a scarcity of professionals with background in de-identification methods, an unexpected finding that seems rooted in lack of interest in the topic at U.S. graduate schools.

4.2.6. Identifying risk

The main criterion used by our case study institutions to identify potentially identifiable records in PUF initiatives, shared by both PUF and non-PUF initiatives, is whether a record is unique with respect to a combination of key variables (known as *k-anonymity)*. Typically, demographic indicators such as gender, age, race, education, marital status, number of children, geographical location, are used to define combinations. Which variables and the exact nature of the combination(s) is confidential information.

4.2.7. Disclosure limitation methods and software

We found no preference among the many methods available for limiting data. The agencies included in our case studies use all the well-known technique (coarsening, suppression, top and bottom coding, rounding, random rounding, and data swapping). The decision on which method to use is typically case-specific and even variable-specific. It also depends on internal deliberation. Each initiative has its own algorithm for disclosure avoidance. And all refrain from using disclosure avoidance software because of the possibility of reverse engineering. As expected, PUF initiatives use more sophisticated masking techniques than non-PUF initiatives.

4.2.8. Risk of match to other datasets

We also found that agencies take into consideration other files available to the public (e.g., online) when evaluating the risk of disclosure. This is true for both external datasets and previously released data (e.g. reports, tables) from the same source. These comparative activities are conducted both in-house and/or by outside contractors (particularly data security firms). The degree and level of sophistication of these activities vary by initiative, with Census, NCHS, and NCES exemplifying initiatives that are highly concerned and highly cautious, and non-PUF initiatives such as those in our case studies much less so.

4.2.9. Re-identification certification

None of the cases studied has a re-identification certification procedure in place. Re-identification refers to the possibility of an intruder being able to identify someone in a PUF and learn information that he/she would not be able to learn otherwise. Re-identification certification tests the vulnerability of PUFs to external sources and need to be conducted by a third party.

4.2.10. Data utility

We found little information on what is done regarding data utility (e.g., an assessment of information loss) after applying statistical disclosure limitation (SDL) methods. While the PUF initiatives in our case studies reported that they do conduct such analyses, these are not available to researchers. The data utility analyses we were told about concentrate on comparisons of means and distribution tests before and after disclosure treatment, to determine their effect on pre-treatment statistical characteristics of the data. In-house and consultant statisticians also do multivariate tests to study effects on relationships among multiple variables. Despite the limited nature of these tests, all PUF-initiative case studies highlighted the importance of data utility analyses, and, in particular, the coordination of such analyses between statisticians and program directors (topic experts), to avoid unnecessary distortions in the data to be released. These concerns were less pronounced for the non-PUF initiatives.

4.2.11. Data access

Access to data is granted via online query systems, PUFs, licenses, and onsite at Research Data Centers (see section 7.3 for further detail). We found no access method to be preferred over another. Our case study institutions have adopted such methods in response to user demands. The degree to which access to detailed information is allowed (of the type that is not available in PUFs) depends on both the pertinence of the research question and the degree by which an individual (or individuals) can be held accountable for the use of the data. None of our case study institutions had plans to stop releasing data to the public.

4.2.12. Confidentiality requirements to users

We found that the degree to which PUF initiatives warn users-to-be on confidentiality issues (such as to explicitly avoid actions aimed at re-identifying individuals) varies greatly, from short statements on the webpage where the data are located to agreeing to online DUAs before download. The two non-PUF initiatives required users to sign and submit DUAs for agency revision and approval before granting data access.

4.2.13. Documentation

All initiatives investigated were similar in stressing documentation as a key success factor. Descriptions of data fields are provided in data dictionaries. Descriptions cover the content of each field, method of presentation, and disclosure avoidance steps taken to provide confidentiality. However, PUF initiatives are cautious of not revealing unnecessary information, making the language used rather generic. Our review of the documentation available online for all the case study initiatives suggested that what has been done to protect the data is not discussed at length nor easy to find in documentation. It is our understanding that the SDL techniques should not be revealed unless the data producer is certain that such information does not increase the re-identification risk by allowing re-engineering of the actual data.

4.2.14. Communication channels and user feedback

There are three main channels of two-way communication between each of the case study initiatives and its users: email, phone numbers, and personal communication at national conferences. Regarding diffusion of data releases, the main channels are via listserv and notifications on their respective websites. As for user feedback, we found that data are not generally modified in response to user demands, as these typically involve requests for more detailed information. However, when errors are discovered, either by staff or by users, corrections are made and the PUFs are re-released. Social media channels such as Facebook and Twitter were not in use at the time of our case studies. However, our review of websites suggests that these agencies are moving rapidly in that direction. The Centers for Disease Control and Prevention (CDC), the US Census Bureau and the Social Security Administration, for example, can all be followed on Facebook and Twitter.

4.2.15. Strong success record

Despite the variety of approaches to guarantee privacy and confidentiality (methods and access) among our case studies, we found a strong overall record of success. Perhaps the only problem cited, by non-PUF institutions, is that occasionally researchers publish papers with a small-count cell in a table (e.g., less than 10 individuals, less than 3 institutions). In those cases researchers are asked to immediately take the necessary steps to remove or retrieve such information from where it has been published. There has been no reported breach of confidentiality to date in any of the cases studied.

5. Stakeholder Interviews

In this section, we provide a summary of the stakeholder interviews conducted as part of the CER-PUF project.

5.1. Objective

The issues regarding de-identification of data for PUF creation are complex, with the interest of many parties involved. In the stakeholder interviews we met with three types of experts: experts on de-identification, health information privacy experts or advocates, and governmental representatives from organizations that provide PUFs. The goal was to identify the most salient issues regarding making microdata publicly available.

5.2. Findings

The principal concern identified by our stakeholder experts was that no legal enforcement mechanisms exist to hold individuals accountable for attempting to re-identify data that has been de-identified, since once data have been de-identified they fall outside the boundaries of relevant Federal legislation, such as Health Insurance Portability and Accountability Act (HIPAA) and the Privacy Act. Stakeholders also expressed the belief that de-identification methods traditionally used by many government agencies to create public use datasets – such as the HIPAA safe harbor method – may no longer provide adequate protection against skilled data intruders, due to advances in re-identification methods. They emphasized that agencies need to keep informed of the constantly changing landscape of de- and re-identification methods, to ensure they apply effective treatments and/or access restrictions to safeguard the data they release.

The stakeholders interviewed also explained that agencies intending to release PUFs must contend with a constantly shifting environment, in which an ever-increasing amount of publicly and privately held data about individuals is becoming ever more easily accessible. Improvements in data mining technology and the science of data re-identification magnify the concern that external data sources can be linked to PUFs and used to re-identify the individuals in them.

These experts noted that there are no uniformly accepted data de-identification methods, and that the choices of particular de-identification techniques used by government agencies are primarily based on the specific details of the data being treated. The consensus among our experts was that there are no generally agreed-upon standards for an acceptable level of disclosure risk, or even for methods to objectively quantify such risk. Ultimately, decisions about levels of risk need to be made by each agency, based on what it considers sound policy.

Last, according to stakeholders the trade-off between ease of data access and analytic utility noted earlier is inherent. The data access method chosen by an agency for a particular dataset is dependent on the level of detail about individuals in the dataset. Typically, the more useful a dataset is to researchers, the more detail it contains that can be used to identify the individuals within it, and, thus, the more restrictive must be the access methods provided for it. For this reason, the stakeholders recommended a tiered system of access, based on the level of detail in a dataset, rather than focusing on the PUF option only. A tiered access system, in which more detailed data are provided to data recipients conditional on additional restrictions and obligations, can help balance this trade-off.

Overall, stakeholders agreed that creating de-identified individual-level Medicare claims public use data files, while providing significant analytic utility, is an extremely challenging endeavor. They expressed the belief that data can never be completely de-identified with certainty, and that it is prudent, therefore, to apply additional measures of protection, such as DUAs, to help mitigate disclosure risk and dissuade would-be intruders from attempting to re-identify data.

6. Creating the BSA PUFs

In the first phase of the CER-PUF project, 8 BSA PUFs have been produced; one per type of claim (e.g., inpatient, carrier). Because priority was given to protecting the privacy of Medicare beneficiaries, the amount of information released was limited. Each BSA PUF contains 7 to 10 analytic variables, the selection of which was based on recommendations received from researchers (see section 4). Even with this limited number, some of these PUFs contain millions of records. (The carrier BSA PUF has close to 68 million line items, for example, and the Part D Events BSA PUF more than 50 million events).

Among the set of SDL methods available in the literature, we chose non-perturbative methods, such as rounding and coarsening, and local suppression, following the preferences of the researchers. Examples of variables that were coarsened include: age into 5-year intervals, 5-digit ICD-9 codes into 3-digit ICD-9 codes, and number of visits and days into categories. Rounding was applied to monetary values, such as Medicare payment amounts, with different rounding rules depending on the range of values. We found the best substitute for a risk measure in the CMS' DUA for use of CMS data in the creation of any document, which stipulates that "… no cell (e.g., admittances, discharges, patients, services) 10 or less may be displayed. Also, no use of percentages or other mathematical formulas may be used if they result in the display of a cell 10 or less." Hence, every cell (i.e., unique combination of all variables) in the BSA PUFs contains at least 11 beneficiaries in the population with the same claim information; smaller cells are suppressed (after rounding and coarsening steps). With a 5% random sample and the "rule of 11", the maximum risk for any given cell is 0.0045 (i.e., 0.05*(1/11)), which corresponds to the probability of an intruder claiming that he/she found his/her target for a randomly drawn record from any of the BSA PUFs. As 11 is the smallest cell size in the file, the risk of the whole PUF may be significantly smaller than 0.0045, depending on the distribution of records across cells.

To avoid utility loss (though without a metric) we searched iteratively for the optimal number of variables and coarsening/rounding decisions to keep the suppression rate (i.e., number of records suppressed divided by the total number of records in initial 5% sample) under 10%. Also, we compared the frequency distributions of variable values before and after the suppression to ensure that the PUFs provided information consistent with the actual claims files.

7. Conclusions Regarding Data Access Options

Many government agencies and institutions host data collected through surveys or administrative records (in the case of Medicare, for example, actual claims). Researchers need access to these data to analyze important policy issues. Given the need to protect the privacy of the individuals in the data sets, however, access to the files can only be made possible by either restricting access through stringent DUAs, removing information that might lead to the identification of individuals and issuing PUFs, or giving researchers access, not through their own computers but through secure portals of some type. We review the advantages and disadvantage of each briefly below.

7.1. The DUA approach

The great benefit of abiding by restrictive DUA terms is researcher access to either the actual or minimally altered data sets. In most such cases, researchers can even have personal copies of the data on their computers. But the financial costs of accessing these datasets can be high, restricting de facto access by researchers without the requisite funds. A possibly more important detriment is that the datasets can be shared, which can lead inadvertently to a breach of confidentiality, or stolen, when the breach is deliberate. A breach of unsecured health information can harm an individual in multiple ways, but it can also be disastrous for the agency involved. In the U.S. healthcare system, any breach has to be reported to the Department of Health and Human Services and to the affected individuals because of the Breach Notification Rule of HIPAA.[8] By then, of course, the damage will have been done.

7.2. The PUF approach

The benefit of PUFs is that they are typically available for free download on the data host's website. This gives researchers full and unrestricted access to the PUF and the ability to perform analyses at will without incurring access costs or having to submit a research plan. However, as detailed in this paper, PUFs have the non-negligible disadvantage of diminished data utility.

The CER-PUF project provided the opportunity to fully assess the costs and benefits of disseminating CMS claims datasets as PUFs. The lessons we learned can be summarized in the follow advice to PUF developers:

- Review underlying laws and regulations regarding the dataset and the institution that hosts it;
- Review PUF plans with stakeholders to assess viability;
- Identify SDL techniques that are acceptable to not only the de-identification and privacy experts but also the data end-users;

[8] http://www.hhs.gov/ocr/privacy/hipaa/administrative/breachnotificationrule/index.html

- Understand the needs of end-users in terms of variables and how they are presented in the PUF;
- Use samples (rather than the full database) to provide added protection;
- Define a risk assessment method and a risk measure, by determining the most likely intruder scenario;
- Define a utility loss metric to quantify the effect of de-identification on the information on the data file;
- Do not make the SDL technique public if such information would reduce the safety of the PUF;
- Prepare detailed documentation, codebook, data dictionary, and FAQs;
- Invest in a range of dissemination methods to increase visibility and ease of use, such as preparing a dashboard, producing short briefs, designing challenges, and making conference presentations.

7.3. Access through secure data portals

If access to the actual database is deemed absolutely essential, Remote Data Centers (RDCs) or Data Enclaves are two very secure options. These options require significant investments in information technology infrastructure and maintenance, and may also need a mechanism output review. They are more secure than providing access via DUAs and data access fees, simply because data files are stored in safe locations (not on researchers' laptops or mobile data storage devices). In the RDC option, researchers must be physically at the RDC location in order to access the data. For example, the National Center for Health Statistics (NCHS) has two RDC locations (Hyattsville, MD and Atlanta, GA), where researchers can work under supervision during regular business hours. Users have to satisfy a list of requirements, such as submitting a proposal and seeking approval, following strict guidelines while on site, and providing their computer software codes (SAS, Sudaan, etc.) to the RDC staff. They also have the option to access NCHS data from locations that belong to the U.S. Census Bureau following additional requirements and security procedures. An RDC option may not be convenient to researchers, however, given the limited number of locations and the need to travel to where the database is.

The Data Enclave option allows researchers to access the data remotely from any computer with a secure internet connection at any time. Users can submit their queries or procedures using statistical software and ask the automated system to send the results, the log file, or even an output file via email. The disadvantage in comparison to the RDC option is that research activity may be limited by the options in the available software programs or what the researcher can observe in terms of output. This is because of the automated nature of the remote access framework and lack of an in-person review mechanism and a controlled environment, as are present in an RDC.

7.4. A Mixed Dissemination Strategy

A dissemination strategy that involves some or all of the options above may be optimal. It is clear the PUFs can not only increase the availability of data without requiring DUAs or fees but also be sufficient for many studies. But it is also clear that certain research questions cannot be adequately addressed without all the details of the actual data, including individual identities or identifiers. Such questions can only be answered by making the actual files available, either by delivering the file to researchers or by placing them into RDCs or Data Enclaves. How all these options should be weighted in a mixed strategy depends importantly on the costs associated with each, which are not covered in this paper.

References

Agrawal, R. and Srikant, R. (2000). "Privacy-preserving data mining." Proc ACM SIGMOD International Conference on Management of Data, pp. 439-450.

Dalenius, T. (1977). Toward a methodology for statistical disclosure control. *Statistik Tidskrift 15*: 429-444.

Duncan et al. (1993). *Private Lives and Public Policies*. National Academy Press.

Duncan, G. T., Elliot, M. and Salazar-González, J. (2011). *Statistical Confidentiality: Principles and Practice*. New York: Springer.

Dwork, C. and Naor, M. (2010). "On the Difficulties of Disclosure Prevention in Statistical Database or the Case for Differential Privacy." *Journal of Privacy and Confidentiality 2(*1): 93-107

Emam et al., (2009) "A Globally Optimal k-Anonymity Method for the De-Identification of Health Data." *Journal of the American Medical Informatics Association,* 16(5): 670–682.

Erdem, E. and Thomas W. Concannon. "What Do Researchers Say about Medicare Claims Public Use Files?" Prepared for the Centers for Medicare & Medicaid Services, U.S. Department of Health & Human Services, 2011 (Submitted).

Fienberg, S. and McIntyre, J. (2005). "Data Swapping: Variations on a Theme by Dalenius and Reiss." *Journal of Official Statistics, 21(*2): 309-323.

Hundepool, A. et al (2010). *Handbook on Statistical Disclosure Control*. ESSNet SDC (Available at http://neon.vb.cbs.nl/casc/..%5Ccasc%5Chandbook.htm)

Narayanan, A., and Shmatikov, V. (2010). "Myths and Fallacies of 'Personally Identifiable Information'", *Communications of the ACM 53(*6): 24-26.

Prada, S. (2011) "Creating Public Use Files: What makes a successful initiative?" (Submitted).

Prada, S. et al (2011) "Avoiding Disclosure of Individually Identifiable Health Information in Public Use Files: A Literature Review" (*SAGE Open* forthcoming).

Reiter, J. (2009). "Multiple Imputation for Disclosure Limitation: Future Research Challenges." *Journal of Privacy and Confidentiality, 1(*2): 223-233.

Singh, A. (2009). "Maintaining Analytic Utility while Protecting Confidentiality of Survey and Nonsurvey Data." *Journal of Privacy and Confidentiality,* 1(2): 155-182.

Skinner, C. (2009). "Statistical Disclosure Control for Survey Data." in (D. Pfeffermann and C.R. Rao, eds., Handbook of Statistics Vol. 29A, Amsterdam: Elsevier, 381-396.

Sweeney, L. (2000). "Uniqueness of simple demographics in the U.S. population." LIDAP-WP4. Laboratory for International Data Privacy, Carnegie Mellon University.

Sweeny, L. (1997). "Weaving Technology and policy together to maintain confidentiality." *Journal of Law, Medicine and Ethics, 25(*2-3): 98-110

Thorpe, J. (2010) "Medicare Public Use Files for Comparative Effectiveness Research – Analysis of Relevant Laws and Regulations." Prepared for the Centers for Medicare & Medicaid Services, U.S. Department of Health & Human Services.

Willenborg, L. and de Waal, T. (1996). "Statistical Disclosure Control in Practice." Lecture Notes in Statistics Vol. 111, Springer-Verlag, New York.

Winkler, W. E. (2007). "Examples of Easy-to-implement, Widely Used Methods of Masking for which Analytic Properties are not Justified." Research Report Series #2007-21, Statistical Research Division, U.S. Census Bureau.