Information Security and Computer Fraud, 2014, Vol. 2, No. 3, 33-38
Available online at http://pubs.sciepub.com/iscf/2/3/1
© Science and Education Publishing
DOI:10.12691/iscf-2-3-1

The DISTANCE Model for Collaborative Research: Distributing Analytic Effort Using Scrambled Data Sets

Howard H. Moffet[1,*], E. Margaret Warton[1], Melissa M. Parker[1], Jennifer Y. Liu[1], Courtney R. Lyles[2], Andrew J. Karter[1]

[1]Division of Research, Kaiser Permanente Northern California, 2000 Broadway, Oakland, CA 94612
[2]Center for Vulnerable Populations, University of California San Francisco, 1001 Potrero Ave., San Francisco, CA 94110
*Corresponding author: Howard.H.Moffet@kp.org

Received September 30, 2014; Revised October 17, 2014; Accepted October 20, 2014

Abstract **Background:** Data-sharing is encouraged to fulfill the ethical responsibility to transform research data into public health knowledge, but data sharing carries risks of improper disclosure and potential harm from release of individually identifiable data. **Methods:** The study objective was to develop and implement a novel method for scientific collaboration and data sharing which distributes the analytic burden while protecting patient privacy. A procedure was developed where in an investigator who is external to an analytic coordinating center (ACC) can conduct original research following a protocol governed by a Publications and Presentations (P&P) Committee. The collaborating investigator submits a study proposal and, if approved, develops the analytic specifications using existing data dictionaries and templates. An original data set is prepared according to the specifications and the external investigator is provided with a complete but de-identified and shuffled data set which retains all key data fields but which obfuscates individually identifiable data and patterns; this "scrambled data set" provides a "sandbox" for the external investigator to develop and test analytic code for analyses. The analytic code is then run against the original data at the ACC to generate output which is used by the external investigator in preparing a manuscript for journal submission. **Results:** The method has been successfully used with collaborators to produce many published papers and conference reports. **Conclusion:** By distributing the analytic burden, this method can facilitate collaboration and expand analytic capacity, resulting in more science for less money.

Keywords: *data sharing, privacy rule, information dissemination, collaboration, cohort studies, epidemiology, de-identification*

Cite This Article: Howard H. Moffet, E. Margaret Warton, Melissa M. Parker, Jennifer Y. Liu, Courtney R. Lyles, and Andrew J. Karter, "The DISTANCE Model for Collaborative Research: Distributing Analytic Effort Using Scrambled Data Sets." *Information Security and Computer Fraud*, vol. 2, no. 3 (2014): 33-38. doi: 10.12691/iscf-2-3-1.

1. Background

"*Data should be made as widely and freely available as possible while safeguarding the privacy of participants and protecting confidential and proprietary data.*" (from NIH Statement on Sharing Research Data, February 26, 2003 [1])

Primary data collection and cohort creation are expensive endeavors, and the data generated typically far exceed the analytic capacity and time frame supported by the original grant. Data-sharing is encouraged to fulfill the ethical responsibility to transform research data into public health knowledge; the National Institutes of Health require a data-sharing plan for research applications requesting $500,000 or more of direct costs in any single year. [2] However, data sharing carries risks of improper disclosure and potential harm from release of individually identifiable data.

The Privacy Rule, [3] as part of a federal mandate to safeguard the rights and welfare of human subjects, provides a framework by which health information can be shared (disclosed) for research purposes. Health information which has been "de-identified" may be used and disclosed freely, as it is no longer considered protected health information. [4] There are two approaches to data de-identification: Expert Determination Method or Safe Harbor Method. [4] The Safe Harbor Method requires "the removal of specified individual identifiers as well as absence of actual knowledge by the covered entity that the remaining information could be used alone or in combination with other information to identify the individual." [4] However, while the Privacy Rules permits de-identified data sets to be shared freely, the responsible covered entity may choose to restrict disclosures; further, de-identified data sets are generally regarded as being of limited value because, typically, relevant data have been removed [5].

The authors here describe a protocol for making de-identified data more productive using a protocol which enables an external investigator to collaborate with an analytic coordinating center (ACC). The ACC de-identifies and then shuffles data to create "scrambled data sets," a process which deletes or obfuscates individually identifiable data and patterns while leaving the population characteristics intact. A scrambled data set is useful to the external investigator as a sandbox to develop and test statistical code which is run against the original data at the analytic coordinating center(ACC), generating output which is used in preparing a manuscript. As the ACC analysts' time is often a limiting factor for productivity in multi-site studies, this method of collaboration based on shared effort and distributed data analysis has been used to leverage resources for greater productivity.

The Diabetes Study of Northern California (DISTANCE) began in 2005 as a survey follow-up study among a racially stratified cohort of 20,000 patients with diabetes (www.distancesurvey.org). [6] The survey data has been linked to extensive data from the Kaiser Permanente Northern California (Kaiser) electronic health record and the Kaiser Diabetes Registry, which was established in 1994 and currently includes over 230,000 patients with diabetes [7]. Today, the DISTANCE collaboration involves over 40 scientists from multiple institutions and is guided by the Publications and Presentations Committee (P&P) which strives to: (i) ensure accurate, uniform, timely, and high quality reporting of research findings; (ii) preserve the scientific integrity of the study; and (iii) safeguard the rights and confidentiality of participants. The P&P oversees the ACC where final research data resides and all analyses are performed.

Because the procedure described here uses de-identified data, it is not necessarily subject to human subjects protection rules; de-identified may be used and disclosed freely, as it is no longer considered protected health information. [4] However, in every instance in which we have applied this procedure, the use of original data has been approved by the Institutional Review Board of the Kaiser Foundation Research Institute. The following is a generalized description of the DISTANCE collaborative research method which distributes the analytic effort using scrambled data sets.

2. Methods

An external investigator with an idea for a study based on ACC data submits a written proposal to the P&P. The ACC provides the investigator with manuscript writing guidelines, data dictionaries and sample statistical specifications. The investigator follows a well delineated protocol and accepts responsibility for some of the analytic effort outside of the ACC (Figure 1). The P&P must give approval for any effort ("manuscript") intended to result in a publication, whether journal article, conference abstract/presentation or public report. The investigator is responsible for adhering to ACC policies and guidelines and for producing the final manuscript for publication. The investigator must possess the necessary skills– or have a qualified analyst– to carry out the analysis required by the proposal.

After P&P approval of the proposal, the investigator works with an assigned ACC analyst to develop detailed analytic specifications. It is helpful for the investigator to review specifications from previous analyses and become familiar with existing data which may include survey responses, clinical and administrative measures and various derived variables. In development of the specifications document, edit mode in word processing is used to track the refinements by investigator and analyst, recording their discussions about questions, comments or changes. In addition, the analytic plan is often presented at collaborator meetings for group feedback.

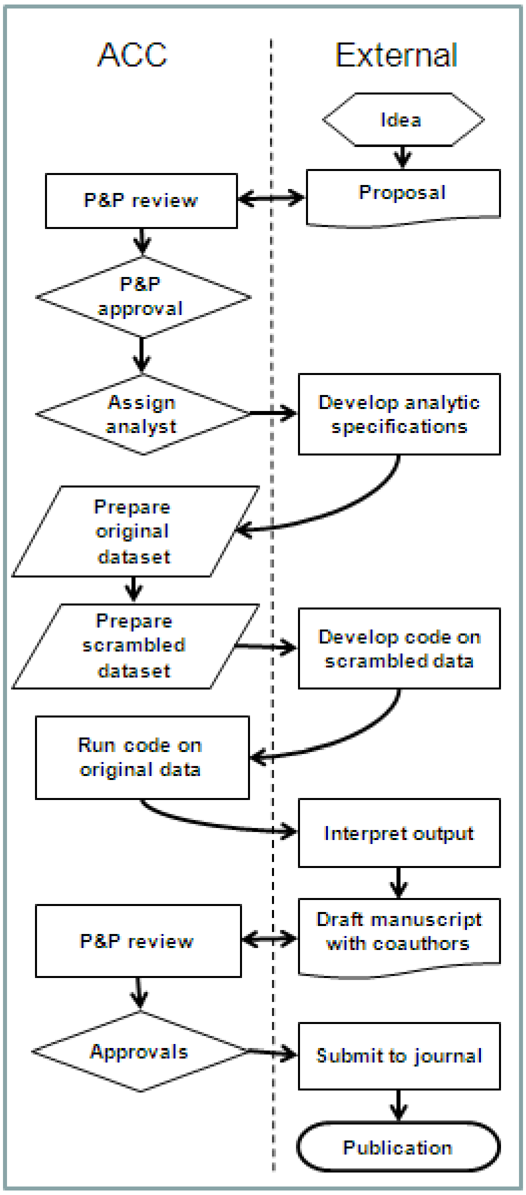


Figure 1.

Legend: From idea to publication. This flowchart illustrates the steps for the DISTANCE collaborative research method which distributes analytic effort outside the ACC using scrambled data

In most cases, but especially in studies aiming to make causal inferences, investigators prepare a directed acyclic graph (DAG), which is a conceptualization of the causal framework underlying the proposed study and which graphically display assumed causal relationships between variables in the analysis, based on subject-matter knowledge. [8,9] A DAG can help clarify the *a priori* assumptions, identify potential confounders and mediators and avoid missing important covariates in the initial steps of building a data set. Thus, in addition to developing the relevant conceptual models, analysis of the DAG facilitates model development, with the aim of specifying the most parsimonious statistical model.

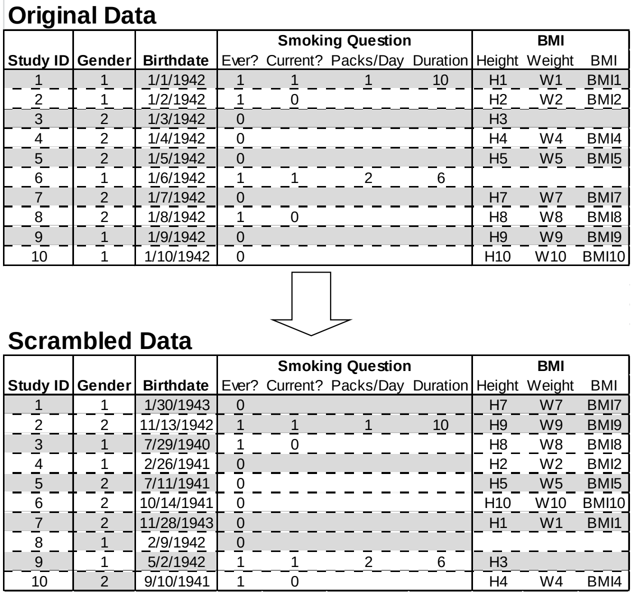
During development of the analytic specifications, the ACC analyst advises the investigator on the available data and its limitations, assists in defining the data cut points or transformations, or suggests analytic strategies and model specifications. In particular, the ACC analyst provides the investigator with univariate statistics for variables in the proposed study data set to facilitate an understanding of variable distributions and rates of data missingness. The investigator and analyst review existing cohort sand derived variables to minimize duplication of effort and to use study resources most economically. In many cases, an existing cohort or data set can be used, but additional or updated clinical or administrative health plan data may also be required for the analysis. In some cases, an existing data set can be used for which a scrambled data set has already been prepared. Collaborating clinicians or other members of the writing group can help identify potential covariates, confounders or mediators of particular clinical measures. Clinical data archiving is often very complex, and ACC analysts have background knowledge that can prove invaluable when designing a study. Issues such as changes in the availability and quality of clinical measures over time and changes in methods of measurement are taken into consideration when creating any variable derived from clinical or administrative data.

Once the specifications are complete (Table), the ACC analyst prepares the “original” data set containing only the data elements necessary for the proposed analysis. The ACC analyst then prepares the scrambled data set (described below) which the investigator will use to develop and test analytic code. The analytic work can be shared using any statistical software that is available to both the investigator and ACC analyst; however, if the external investigator and analyst have different versions of the same software, this can present a challenge which is best identified at the beginning of the process.

Table. Minimum Requirements for Analytic Specifications

1. Abstract
2. Research question
3. Directed acyclic graph (DAG)
4. Background and research findings to date on the substantive area
5. Study hypotheses or research questions
6. Objectives
7. Inclusion criteria
8. Exclusion criteria
9. Study design with clearly stated observation windows and baseline definitions
10. Explicit variable definitions
11. Explicit model specifications
12. Staged analytic plans

Of the 18 elements (individually identifiable data categories) covered by the Privacy Rule, [3,10] typically only medical record numbers and dates are relevant to the proposed research. Medical record numbers are replaced with anonymous study identification numbers (Figure 2). Dates of birth or medical events (e.g., appointments, procedures, hospitalizations) are perturbed by adding or subtracting a random number of days (e.g., ± 0-365) to each date. Alternatively, especially for longitudinal studies, an index or baseline date (e.g., a diagnosis date, baseline survey date or first medication dispensing date) can be identified and perturbed, and then all other dates can be converted to a number representing days pre- or post-baseline.



Original Data

			Smoking Question				BMI		
Study ID	**Gender**	**Birthdate**	Ever?	Current?	Packs/Day	Duration	Height	Weight	BMI
1	1	1/1/1942	1	1	1	10	H1	W1	BMI1
2	1	1/2/1942	1	0			H2	W2	BMI2
3	2	1/3/1942	0				H3		
4	2	1/4/1942	0				H4	W4	BMI4
5	2	1/5/1942	0				H5	W5	BMI5
6	1	1/6/1942	1	1	2	6			
7	2	1/7/1942	0				H7	W7	BMI7
8	2	1/8/1942	1	0			H8	W8	BMI8
9	1	1/9/1942	0				H9	W9	BMI9
10	1	1/10/1942	0				H10	W10	BMI10

Scrambled Data

			Smoking Question				BMI		
Study ID	**Gender**	**Birthdate**	Ever?	Current?	Packs/Day	Duration	Height	Weight	BMI
1	1	1/30/1943	0				H7	W7	BMI7
2	2	11/13/1942	1	1	1	10	H9	W9	BMI9
3	1	7/29/1940	1	0			H8	W8	BMI8
4	1	2/26/1941	0				H2	W2	BMI2
5	2	7/11/1941	0				H5	W5	BMI5
6	2	10/14/1941	0				H10	W10	BMI10
7	2	11/28/1943	0				H1	W1	BMI1
8	1	2/9/1942	0						
9	1	5/2/1942	1	1	2	6	H3		
10	2	9/10/1941	1	0			H4	W4	BMI4

Figure 2. Example of transformation of original data set into scramble data set

Legend: In this small, mock dataset, original data is transformed into scrambled data. Gender is randomly reordered; each birth date has a random number added or subtracted; a set of smoking questions is randomly reordered; height, weight and calculated body mass index are randomly reordered.

In preparing the scrambled data set, the complete variable structure (and population characteristics) of the data remains intact, but all individually-identifiable data are replaced or randomly modified so that individually-identifiable patterns are disrupted. There is no technical novelty in this approach to de-identification (also known as “data shuffling” [11]), but a description of this simple method is provided here.

The scrambling process is most simply described for a single data set with rows (observations) and columns (variables), although data with more complicated database architecture (e.g., many-to-many structure) can be accommodated. Each cell within a given column is assigned a random number and then sorted (e.g., low to high) by the assigned random numbers. This process is repeated for each individual column. Sets of columns representing variables that form a scale, derived variable or index (e.g., smoking questions or height and weight with calculated BMI) are randomly sorted as a group in order to maintain their internal validity. The scrambling protocol thus disrupts patterns which could identify an individual (e.g., combinations such as an individual’s gender plus smoking status plus weight plus age).

The scrambled data set has a structure identical to the original data set and retains actual values for each variable (except dates) from the original dataset. The scrambled data set is emailed to the external investigator to develop and test analytic code which should run equally well against the original data. In the scrambled dataset, non-marginal statistics and associations are meaningless, but missingness and marginal summary statistics (e.g., mean patient weight) for each variable or derived variable are accurate and valid. This allows the investigator to characterize the study population (corresponding to the manuscript Table 1) directly from the scrambled data.

Once the code runs without error on the scrambled dataset, it is emailed to the ACC analyst to run against the original dataset. The ACC analyst corrects minor coding errors as needed and sends to the investigator the output (stripped of individually identifiable data, if any),notes on any changes made to the code and, if needed, the log files. If more complex errors occur, particularly in the code for model specification, the analyst alerts the investigator and asks for revised code. The process is repeated until analyses are complete. By having the collaborating investigator develop the analytic code, a substantial burden is removed from the ACC analyst, whose time may be a limiting factor, and thus ACC productivity is increased.

During this iterative process, the ACC analyst frequently runs code without closely checking the external investigator's methodology or the output. These un-monitored runs are time-saving and acceptable during the development of the model or method, given that the models often change. However, once the process approaches its final iteration, the investigator will ask the ACC analyst to review and approve the code and final output before the investigator prepares the draft manuscript. Once a draft manuscript is completed, the ACC analyst performs a final review, checking the appropriateness of analysis and the consistency between output and manuscript. As with all manuscripts, the investigator actively involves coauthors throughout the process. Targeted calls at critical junctures (e.g., to discuss specifications or focus of the manuscript or discuss a reviewer's comments) are very useful.

When the investigator and the ACC analyst are satisfied and all co-authors have given their final approvals, the journal-ready manuscript is submitted to P&P for final review and approval. After institutional approvals are obtained, the manuscript is ready for submission to the target journal.

The increased ACC productivity motivated the development of a database to track the progress of each manuscript from proposal to publication and to monitor the workload of each ACC analyst. Each manuscript is linked to its supporting grants and, upon publication, the database record is completed with its PubMed hyperlink, PubMed ID, PubMed Central ID and a 100-word summary. The database is also used in generating progress reports.

3. Results and Discussion

To date, this method has been used in several published papers [12-20] and conference abstracts [21,22,23], as well as manuscripts in progress.

The DISTANCE collaborative research method uses scrambled data sets and a protocol which distributes some of the analytic effort outside of the ACC but presents no risk to patient privacy. Scrambled data sets provide a "sandbox" for investigators external to the primary data collection site to develop and test code for statistical models; because it is based on real data, it allows the external investigator to preview summary statistics (e.g., for the manuscript's Table 1) or to independently assess univariate data patterns (e.g., to identify appropriate cut points to categorize variables). In general, the external investigator is responsible for developing the proposal, specifications, analytic code, interpretation of analytic output and preparation of a manuscript for journal submission; the ACC analyst is responsible for preparing the original and scrambled data sets, running the code against the original data set and reviewing the final code and manuscript. The P&P provides guidance and oversight to ensure scientific integrity and quality.

3.1. Limitations

The specific method of shuffling data is not novel and there are likely other ways to accomplish the same end. [24] For example, instead of scrambling the data, one could insert random values or dates; however, basic characteristics of the data, such as means, would be lost with no saving of effort. While it is possible that de-identified data could be re-identified [25], the scrambling procedure eliminates the patterns which might permit re-identification and loss of privacy. Occasionally, the back-and-forth in the development of the specifications and the analytic code creates delays, but the external investigator usually drives the process: typically, the code is submitted to the ACC analyst, run on the original data and promptly returned. This protocol works well even when the model specification becomes very complex.

This method is most compatible with hypothesis-driven research; it is less compatible with "data mining" since associations observed in the scrambled data sets are not meaningful. It expands collaborative opportunities to external investigators, especially junior faculty and fellows who may have sufficient funding (e.g., a K-award) to cover their time and who want to hone their analytic skills but lack access to quality data. This approach has been used successfully with many outside investigators, including a former doctoral student (Dr. Lyles) who successfully used scrambled data sets to produce her dissertation and four peer-reviewed journal articles. [13,14,15] The protocol has been replicated and used by the P&P of a DISTANCE sub-study, ("Diabetes and Aging in a Multi-ethnic Population," R01-DK081796) and we are developing the methodology to create and manage scrambled datasets for a longitudinal study with differential follow-up across subjects.

While the initial creation of a scrambled data set requires effort, it is small compared to the effort saved by distributing some of the analytic effort to the external investigator. Additionally, scrambled data sets can be reused or amended: the scrambled DISTANCE data set, based on subject responses to the 2005-2006 DISTANCE Survey [6], can be used for subsequent research questions or amended with additional or updated clinical or administrative scrambled data.

The DISTANCE collaborative research method has advantages over other common methods of data sharing and collaboration. Unlike public data archives or limited data sets, there is no risk to confidentiality. Unlike typical de-identified data sets, there is no loss of data quality and the original data set can be easily supplemented with additional data or updated with new follow-up data. The scrambled method is simple and effective and, unlike encryption, cannot be undone with any key and has no risk of re-identification without access to the scrambling records which were applied to the original data. Unlike typical analytic coordinating centers, the analysts' time is much less of a limiting factor, so there is no significant bottleneck or loss of productivity. Unlike data enclaves, there is no need to maintain office space or computers designated as secure data access points. There is no need to track the custody of data or its disposition and no risk of improper disclosure, though data agreements are advisable. This method avoids other legal, technical and cultural barriers to data sharing that often complicate multi-site studies, such as the administrative workload associated with executing data use agreements or the other paperwork required when patient data is involved. Scrambled data sets could also be used in studies in which access to original data depends on a lengthy approval process; an investigator could develop analytic code on a scrambled dataset while awaiting receipt of original data. Studies which use a common data model could also use this protocol.

4. Conclusions

The DISTANCE collaborative research method distributes the analytic effort using scrambled datasets and has been used successfully in collaborations with external investigators. The process creates minimal burden on the ACC and mitigates analytic bottlenecks while also eliminating the risk of improper disclosure of confidential patient data, mitigating some of the privacy concerns endemic to collaborative, data sharing endeavors. Finally, it has greatly expanded analytic capacity, resulting in more science for less money.

List of Abbreviations

Abbreviations: Diabetes Study of Northern California (DISTANCE); Publications and Presentations Committee (P&P); analytic coordinating center (ACC); Directed acyclic graph (DAG)

Competing Interests

The authors declare that they have no competing interests.

Acknowledgements

Funding: This work was supported by National Institute of Diabetes and Digestive and Kidney Diseases (grants R01-DK081796, R01-DK080744, R01-DK090272, R01-DK080726,R01-DK086178,P30-DK092924).

Authors' Contributions

HHM drafted the manuscript. All authors contributed to the development and refinement of the procedure. EMW, MMP, JYL and CRL implemented the procedure. All authors contributed to, read and approved the final manuscript. Collaborating investigators used the procedure and contributed valuable feedback in its refinement.

References

[1] NIH Statement on Sharing Research Data. http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm accessed March 21, 2013.

[2] FINAL NIH STATEMENT ON SHARING RESEARCH DATA, NOTICE: NOT-OD-03-032. 2003. http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html accessed March 21, 2013.

[3] U.S. Department of Health and Human Services: Standards for Privacy of Individually Identifiable Health Information. 45 C.F.R. Parts 160 and 164.

[4] Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html Accessed Mar 21, 2013.

[5] Miller JD: Sharing clinical research data in the United States under the Health Insurance Portability and Accountability Act and the Privacy Rule. *Trials* 2010, 11:112.

[6] Moffet HH, Adler N, Schillinger D, Ahmed AT, Laraia B, Selby JV, Neugebauer R, Liu JY, Parker MM, Warton M *et al*: Cohort Profile: The Diabetes Study of Northern California (DISTANCE)--objectives and design of a survey follow-up study of social health disparities in a managed care population. *Int J Epidemiol* 2009, 38 (1): 38-47.

[7] Karter AJ, Schillinger D, Adams AS, Moffet HH, Liu J, Adler NE, Kanaya AM: Elevated Rates of Diabetes in Pacific Islanders and Asian Subgroups: The Diabetes Study of Northern California (DISTANCE). *Diabetes Care* 2012, 36 (3): 574-579.

[8] Greenland S, Pearl J, Robins JM: Causal diagrams for epidemiologic research. *Epidemiology* 1999, 10 (1): 37-48.

[9] Hernan MA, Hernandez-Diaz S, Werler MM, Mitchell AA: Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *American journal of epidemiology* 2002, 155 (2): 176-184.

[10] Flegal KM, Ezzati TM, Harris MI, Haynes SG, Juarez RZ, Knowler WC, Perez-Stable EJ, Stern MP: Prevalence of diabetes in Mexican Americans, Cubans, and Puerto Ricans from the Hispanic Health and Nutrition Examination Survey, 1982-1984. *Diabetes Care* 1991, 14 (7): 628-638.

[11] Muralidhar K SR: Data Shuffling—A New Masking Approach for Numerical Data. *Management Science* 2006, 52 (5): 658-670.

[12] Laiteerapong N, Karter AJ, Liu JY, Moffet HH, Sudore R, Schillinger D, John PM, Huang ES: Correlates of quality of life in older adults with diabetes: the diabetes & aging study. *Diabetes Care* 2011, 34 (8): 1749-1753.

[13] Lyles CR, Karter AJ, Young BA, Spigner C, Grembowski D, Schillinger D, Adler N: Patient-reported racial/ethnic healthcare provider discrimination and medication intensification in the Diabetes Study of Northern California (DISTANCE). *J Gen Intern Med* 2011, 26 (10): 1138-1144.

[14] Lyles CR, Karter AJ, Young BA, Spigner C, Grembowski D, Schillinger D, Adler N: Provider factors and patient-reported healthcare discrimination in the Diabetes Study of California (DISTANCE). *Patient Educ Couns* 2011, 85 (3):e216-224.

[15] Lyles CR, Karter AJ, Young BA, Spigner C, Grembowski D, Schillinger D, Adler NE: Correlates of patient-reported racial/ethnic health care discrimination in the Diabetes Study of Northern California (DISTANCE). *J Health Care Poor Underserved* 2011, 22 (1): 211-225.

[16] Stoddard PJ, Laraia BA, Warton EM, Moffet HH, Adler NE, Schillinger D, Karter AJ: Neighborhood Deprivation and Change in BMI Among Adults With Type 2 Diabetes: The Diabetes Study of Northern California (DISTANCE). *Diabetes Care* 2012, 36 (5): 1200-1208.

[17] Sudore RL, Karter AJ, Huang ES, Moffet HH, Laiteerapong N, Schenker Y, Adams A, Whitmer RA, Liu JY, Miao Y *et al*: Symptom Burden of Adults with Type 2 Diabetes Across the Disease Course: Diabetes & Aging Study. *Journal of General Internal Medicine* 2012, 27 (12): 1674-1681.

[18] Moskowitz D, Lyles CR, Karter AJ, Adler N, Moffet HH, Schillinger D: Patient reported interpersonal processes of care and

perceived social position: The Diabetes Study of Northern California (DISTANCE). *Patient Educ Couns* 2013, 90 (3): 392-398.

[19] Lee SJ, Karter AJ, Thai JN, Van Den Eeden SK, Huang ES: Glycemic Control and Urinary Incontinence in Women with Diabetes Mellitus. *J Womens Health (Larchmt)* 2013, 22 (12): 1049-1055.

[20] Jones-Smith JC, Karter AJ, Warton EM, Kelly M, Kersten E, Moffet HH, Adler N, Schillinger D, Laraia BA: Obesity and the food environment: income and ethnicity differences among people with diabetes: the Diabetes Study of Northern California (DISTANCE). *Diabetes Care* 2013, 36 (9): 2697-2705.

[21] Rees CA KA, Young BA, Spigner C, Grembowski D, Schillinger D, Adler N Correlates of Self-Reported Discrimination in the Diabetes Study of Northern California (DISTANCE). In *31st Annual Meeting & Scientific Sessions of the Society of Behavioral Medicine*. Seattle, WA.

[22] Moskowitz D RC, Adler N, Karter AJ, Moffet HH, Schillinger D. : Effect of the social hierarchy on patient-physician communication: results from the DISTANCE study. In *UCSF Disparities Symposium*. San Francisco, CA.

[23] Lee SJ K, Van den Eeden SK, Cenzer IS, Liu JY, Moffet HH, Huang ES: Glycemic Control and Incontinence in Older Women. In *Amer Diabetes Assn meeting*. San Diego, CA.

[24] Hrynaszkiewicz I, Norton ML, Vickers AJ, Altman DG: Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. *Trials* 2010, 11: 9.

[25] Loukides G, Denny JC, Malin B: The disclosure of diagnosis codes can breach research participants' privacy. *J Am Med Inform Assoc* 2010, 17 (3): 322-327.