Opening Up Evaluation Microdata:
Balancing the Benefits and Risks of Research Transparency

Jennifer Sturdy, Stephanie Burch, Heather Hanson, Jack Molyneaux

Abstract: The Millennium Challenge Corporation (MCC), an independent U.S. Government foreign aid agency, uses a data-based approach to select country partners, and design, monitor, and evaluate programs. With a commitment to transparency built into the MCC model as a means for accountability and learning, the agency has grappled with the ethical, legal, and practical issues related to public release of microdata. In 2013, MCC identified key objectives for public release of its evaluation-related microdata: facilitating verification of evaluation results, maximizing data usability, and minimizing risk to survey respondents. The agency implemented a series of actions, including establishing a microdata dissemination mechanism; data protection principles; an internal Disclosure Review Board; procedures for restricted-access to low-risk data; and allocating staffing and resources to implement the data protection principles. The paper discusses MCC's experience and concludes with a set of lessons learned about finding balance between benefits and risks when releasing evaluation-related microdata.

Keywords: transparency; open data; microdata; de-identification; privacy; data protection; results; evaluation; international development

Acknowledgements: The authors would like to acknowledge the efforts of the additional MCC staff responsible for establishment of the MCC Disclosure Review Board and development of data protection principles: Miguel Adams, Brett Bearce, Katerina Ntep, and Jolyne Sanjak; as well as the external technical advisors, Gerald Gates and Alvan Zarate. In addition, the authors wish to acknowledge the contributions of Olivier Dupriez, Mehmood Asghar, Nicole Styler, and Paul Burgess to support development of MCC's Evaluation Catalog for data dissemination practices, as well as Joseph Green for his substantial contributions as the external data de-identification expert. The authors also thank Sixto Aquino, Scott Fontaine, Nonna Gorilovskaya, Berta Heybey, Tom Kelly, Laura Leussing, Sebastian Martinez, Kari Nelson, Shreena Patel, Bill Savedoff, Sarah Jane Staats, and participants at the 2014 annual meeting of the Berkeley Initiative for Transparency in the Social Sciences (BITSS) for comments on earlier drafts. Finally, none of the actions taken or lessons learned presented in this paper could have been possible without the leadership and commitment of Sheila Herrling.

Disclosure Statement: The authors are employees of the Millennium Challenge Corporation (MCC). All activities reviewed in this paper were funded by MCC. Documentation of these activities is publicly available on MCC's website, www.mcc.gov. The opinions expressed in this paper are those of the authors and do not necessarily reflect the view of MCC, or the United States Government. The authors declare that, other than the compensation they receive as MCC employees, they have no financial interest in the research described in this paper.

1. Transparency in a results-based agency

Demand for evidence-based policy and data driven decisions is growing, and there are important parallel efforts in the social science research community to strengthen the transparency and credibility of research that produces such evidence, particularly in economics, psychology, and political science (Miguel et al 2014). Established in 2004, the Millennium Challenge Corporation (MCC) was envisioned as an institution that would do both—use data for decision-making and be transparent with the data and documentation that underlie decisions.

MCC is an independent U.S. Government foreign aid agency that provides assistance to competitively-selected countries that demonstrate a commitment to good governance, economic freedom, and investing in their citizens. This assistance funds programs designed to achieve MCC's mission: *reducing poverty through economic growth*. Since its inception, MCC has been committed to modernizing the fight against poverty by employing a business-like approach, in which decisions are based on data and analysis, and the results are measured and publicly shared. While laudable, this emphasis on results and transparency is not without risks – including potential reputational harm to the agency, as well as potential loss of confidentiality for those who provide sensitive data.

Reputational harm to an agency may come from publishing data that shows a program failed to achieve the intended impact on outcomes, leading to questions about the capabilities of the agency to deliver on its promises. However, MCC believes that being accountable for and transparent about such results helps MCC to improve its programs and practices (Sturdy, Aquino and Molyneux 2014). In addition, publicly releasing mixed results is useful for other agencies and country partners working to tackle similar issues.¹ Thus, MCC considers the benefits of publishing its program results, including those that are negative or inconclusive, to outweigh any potential reputational harm.

An agency's willingness to take on this reputational risk often depends on the external incentives in place. In the case of foreign assistance, a greater demand for donor transparency has provided such external incentives. The recent passage of the Foreign Aid Transparency and Accountability Act of 2016 (Poe 2016) now enforces across all U.S. Government agencies what is built into MCC's model—conducting evaluations of programs and sharing underlying data. International efforts, like Publish What You Fund's Aid Transparency Index² and the International Aid Transparency Initiative (IATI)³, were developed in response to commitments made at the High Level Forum on Aid Effectiveness in Accra to generate demand for greater donor transparency. The U.S. Government has also established interagency guidance to ensure comparable reporting of all foreign assistance funding,⁴ which is now published in various formats on the Foreign Assistance Dashboard managed by the U.S. Department of State. Additionally, the Memorandum on Transparency and Open Government,⁵ the Open Government Directive,⁶ and the

¹ For example, with the release of the first five impact evaluations in farmer training (<https://assets.mcc.gov/reports/issuebrief-2012002119501-ag-impact-evals.pdf>) MCC presented findings to Feed the Future, the U.S. Government's global hunger and food security initiative, (<http://www.feedthefuture.gov/>) with the aim of communicating lessons learned both on the projects and the evaluations to inform decisions and practices working in a similar portfolio.

² Available online at - <http://www.publishwhatyoufund.org/index/2014-ati/>.

³ Available online at - <http://www.aidtransparency.net/>.

⁴ Available online at - <https://obamawhitehouse.archives.gov/sites/default/files/omb/bulletins/fy2012/b12-01.pdf>

⁵ Available online at - <https://obamawhitehouse.archives.gov/the-press-office/2015/11/16/memorandum-transparency-and-open-government>

Executive Order on openness⁷ provide external incentives for U.S. Government agencies to remain committed to data transparency. With the passage of the Evidence-Based Policymaking Commission Act of 2016 (Ryan 2016), there are also a growing number of resources available to guide federal agencies on how to produce and disseminate policy-relevant data.⁸

The second risk - centered on potential loss of confidentiality for those who provide sensitive data – and how to balance this risk with the potential benefits associated with transparent dissemination of this data - is the focus of the rest of this paper.

2. Distinguishing microdata

MCC collects and reports a wide range of input, output, and outcome data associated with project selection, program expenditures, monitoring, and evaluation.⁹ While public release of any of this data may be affected in different ways (Hanson and Marschner 2015), this paper focuses on MCC's experience with the public release of a particular form of results data—evaluation-related microdata.

MCC collects microdata through surveys for a variety of purposes—from problem diagnostics, to project selection and design, to program evaluations. Unlike input and output data that is typically aggregated, microdata is unique in that it is collected—and can be reported—at the individual, household, enterprise, and/or community level. It also tends to have two distinguishing characteristics: it is personally identifiable and can be sensitive. Personally Identifiable Information (PII) is any information that can be used, on its own or in conjunction with other information that is linked or linkable to a specific individual, to determine the identity of an individual or otherwise locate or contact the individual. It may include direct identifiers, such as name, date of birth, mailing or home address, email address, telephone number, GPS coordinates, national identification number, physical/biological identifiers (physical appearance, through photo or video data collection, fingerprints, DNA, etc.) or indirect identifiers that may combine to expose unique characteristics, such as the only female-headed household in a village. Additionally, some microdata is sensitive – it is information that may pose a risk to the individual or firm if it is collected or released in a way that is linkable to the individual or firm responding to the survey. This type of data may include income, asset holdings, or other information, such as health status, the public release of which could harm survey respondents. Given its nature, the public release of microdata raises ethical, legal and practical issues.

3. Understanding ethical, legal, and practical issues

To begin to understand the ethical, legal, and practical issues associated with public dissemination of microdata, MCC consulted several agencies, including the U.S. Census Bureau and the Bureau of Labor Statistics, as well as other guidance (Dupriez and Boyko 2010; NISTIR 8053). The following details MCC's experience confronting these issues.

3.1. Ethical

⁶ Available online at - <https://obamawhitehouse.archives.gov/open/documents/open-government-directive>

⁷ Available online at - <http://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government->

⁸ For more information, please refer to white papers produced by the Commission on Evidence-Based Policy making at https://www.obamawhitehouse.gov/omb/management/commission_evidence.

⁹ MCC releases all IATI format data online - <http://data.mcc.gov/>

The heart of the ethical issues around public release of microdata is the responsibility of the agency, its agents, and its country partners to protect the survey respondents from harm that may be caused by such release. For any given microdata file, the researcher's and agency's ethical obligations are first defined through the informed consent process. This is the process by which respondents are informed that their participation is voluntary, and are provided with information about the data collection, such as its purpose and the identities of those who will have access to the information, as well as any risks and benefits involved with their participation. The informed consent process is the first mechanism for giving survey respondents confidence that their information, particularly PII, will be protected. Early in its efforts to publicly release evaluation-related microdata, MCC realized that the level of confidentiality promised to each respondent in the original informed consent was a key factor for determining to what extent the microdata could be publicly released. In many early surveys, respondents were informed that participation in the study would remain entirely confidential. This ethical consideration is part of a social contract widely shared by statistics professionals.¹⁰

3.2. Legal

Protecting the confidentiality of respondents raises a number of legal challenges for MCC. First, many federal agencies that collect PII as part of their programs are obligated by legislation to protect the confidentiality of that information.¹¹ However, MCC's enabling statute does not contain such an obligation. Moreover, the U.S. Government-wide statutes that commonly protect privacy do not apply to MCC's evaluation-related microdata. For example, the Privacy Act of 1974, which protects privacy interests within the United States, does not apply because MCC's program evaluations focus on data collection of foreign citizens overseas. Similarly, the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA)¹², which provides broad legal privacy protections to (and correspondingly strict legal obligations on) federal statistical agencies is unavailable to MCC because it is not a federal statistical agency and does not otherwise meet the requirements for CIPSEA applicability. Without legal protection for the PII collected by MCC through its programs, such data could be subject to disclosure under the provisions of the Freedom of Information Act (FOIA),¹³ which requires all US federal agencies to make their records available to persons who request them. There are exemptions from disclosure under FOIA that allow MCC to provide a measure of protection, thereby making the risk that microdata with PII would have to be released unlikely.¹⁴ However, the risk does exist and it limits MCC's ability to guarantee respondent confidentiality fully.

MCC's use of grants to deliver its assistance is another constraint to MCC providing complete protection of survey respondent confidentiality. Under traditional legal principles for grants, assets acquired with

¹⁰ See, for example, the American Statistical Association Code of Ethics, and the International Statistical Institute Declaration on Professional Ethics.

¹¹ See, for example the authorizing legislation for the National Center for Health Statistics (42 U.S.C. §242k, amended) and the United States Census Bureau (13 U.S.C §9, as amended). It should be noted that the majority of agencies whose authorizing legislation include such provisions are statistical agencies. MCC is a foreign assistance agency and not a statistical agency.

¹² CIPSEA - Pub.L. 107-347, 116 Stat. 2899, 44 U.S.C. § 101.

¹³ The Freedom of Information Act (5 U.S.C. 552 (1996)) was first passed in 1967 and gives the public a statutory right of access to federal agency records with the aim of encouraging government accountability through transparency.

¹⁴ For example, Exemption (b)(6) of FOIA exempts "personnel and medical files and similar files the disclosure of which would constitute a clearly unwarranted invasion of personal privacy" from disclosure under FOIA. However, this provision has been the subject of numerous lawsuits and differing interpretations by the courts.

grant proceeds are the property of the grantee or country partner, including microdata collected and developed as part of project implementation and evaluation. While the terms of its grants allow MCC and its agents full use of the microdata that is funded under the grants, the lack of exclusive ownership rights restricts MCC's ability to control access and dissemination and ultimately, its ability to ensure complete protection of the data.

3.3. Practical

Practical issues facing releasing evaluation-related microdata include both institutional arrangements and establishing the dissemination mechanism. First, the agency needs institutional arrangements – requiring human and financial resources - to develop, implement, and enforce data protection principles and practices in line with what is promised at the informed consent stage. Second, the agency must identify an appropriate microdata dissemination mechanism that facilitates access to data documentation and public-use versions of the data. That mechanism must be developed, tested, and populated, which requires resources and time. Recognizing that it will take time to establish data protection activities is useful for setting goals about what can be released and when.

4. Understanding competing objectives

By 2013, with a clearer understanding of the ethical, legal, and practical issues facing release of the evaluation-related microdata, MCC defined the three, often competing, objectives for releasing the data, which guided the development of its data protection principles and practices. The first—*maintain promise of confidentiality made through informed consent process to minimize risks to survey respondents*—has already been discussed in detail. The additional two objectives are discussed below:

Facilitate verification of evaluation results. For MCC evaluations, independent evaluators (individuals or firms) are contracted to design, implement, and disseminate the results of the evaluations. The credibility of these MCC-financed evaluations does not rest on the evaluators' reputations or MCC's efforts to protect their independence through its internal quality assurance processes alone. Rather, MCC seeks to adhere to norms within the social science research community by ensuring that each evaluation's analytic results are verifiable, meaning MCC's stakeholders—including researchers, implementers, and the general public—have the ability to analyze the same data as the evaluators and produce the same, or at least similar, statistical results.

Maximize usability of data. MCC also aims to maximize the usability of evaluation-related microdata for analysis above and beyond the original purposes of the study, while respecting the terms of the informed consent. For example, the evaluation-related microdata could be used for analysis on other outcomes, or to examine differential impacts by geography, socio-economic, or gender characteristics. Such analysis may fall outside the original scope of the evaluation, yet have important development policy dimensions or business value. MCC has already seen examples of researchers using data from evaluations for other analysis, such as in the Mozambique rural water impact evaluation (see Ho, Russel, and Davis 2013).

So how are these three objectives competing? If an agency is purely committed to verification of results and usability of data, then it would commit to public release of all microdata, or at least as much as possible. However, when seeking to also protect the confidentiality of respondents, it may not

be possible to publicly release all microdata. Release of certain data may increase the risk of revealing participation in the study, or the risk of releasing sensitive information, either of which could pose harm to the respondent. Keeping this balance in mind while also understanding the legal, ethical, and practical challenges of public release of microdata, helped guide MCC as it developed data protection principles and resources.

5. Taking Action: Data Protection Principles and Resources

Based on an understanding of the complex issues and competing objectives facing the agency in the public release of its evaluation-related microdata, MCC moved forward with a series of interrelated actions to develop and institutionalize principles and resources for publicly releasing the evaluation-related microdata. The following describes MCC's principles for collection, storage, and dissemination of evaluation-related microdata, as well as resources to establish a data dissemination mechanism and a Disclosure Review Board, enable restricted access to low risk identifiable data, and dedicate staff and other resources necessary to implement these actions.

5.1. Data Protection Principles

In 2013, MCC refined its evolving set of data protection principles. These principles apply to collection, storage, and dissemination, and are informed not just by MCC's experience but by useful discussions in the literature (Glennerster and Powers 2016; Alderman, Das, and Rao 2016).

1. Identify what PII and confidentiality assurances are necessary. Upon reviewing its collection practices, MCC found that more PII was being collected, either by data collection firms, evaluators, other consultants, or MCC and country partner staff, than was necessary for achieving analysis goals as outlined in Evaluation Design Reports. In addition, confidentiality was unnecessarily promised in cases where the data being collected was already publicly accessible. Therefore, the first principle in data collection is to identify what PII and/or confidentiality promises are necessary given the type of data to be collected for specific analysis.

2. Submit the research protocol to an IRB. A research protocol should include not only the study design and methodology, and survey materials such as questionnaires, but also a data protection plan. Since 2013, MCC has required that all independent evaluators submit the research protocol package to an independent Institutional Review Board (IRB),¹⁵ for formal review and approval, in addition to satisfying any country-specific requirements. The objective of this requirement is to ensure that the evaluator, as well as any data collection firm under their direct or indirect management, establishes appropriate procedures for protecting survey respondents' confidentiality and minimizing risk of harm to the respondents as a result of their participation in the data collection. Though an important principle, Alderman, Das, and Rao (2016) provides an extensive discussion on the challenges that traditional, clinically-focused IRB members can face reviewing social and behavioral research in a developing country context, noting the potential limitations of US-based and/or local IRBs to provide

¹⁵ The IRB must be registered with the Office for Human Research Protections (OHRP) within the U.S. Department of Health and Human Services (HHS) and acceptable to MCC. While many countries have established country-level IRBs, not all do, and the quality of these IRBs vary across countries. For this reason, while the evaluator must follow any country-specific requirements, if these are insufficient (i.e., do not meet the qualification of an IRB registered with OHRP), the evaluator must also submit the research protocol package to an international IRB.

sufficient guidance for individual research projects given their country context, research methodology, and other key elements of data collection and research design.

3. *Establish informed consent expectations.* As part of the data protection plan, evaluators and other data handlers are required to have an approved informed consent process that defines the promises of confidentiality. Although every survey is unique, MCC has developed an informed consent template¹⁶ with standardized language for the evaluators and data handlers to use as a guide for laying this foundation at the beginning of data collection. Although MCC expects a clearly defined data protection plan, including the final informed consent, to be reviewed and cleared by the IRB, it also recognizes that every survey has its own unique challenges.

4. *Identify secure storage, transfer, and access protocols.* Depending on the context, for storage this can include using lockable cabinets for paper files, encrypting data files, or employing password protection on data systems. While MCC has made great strides to establish a central storage system for all evaluation-related microdata, establishing a consistent and standard practice for limiting access to the identifiable data remains a challenge for all the various stakeholders: data collection firms, implementers, MCC, country partners, and evaluators. Depending on the context, for transfer and access this can include using secure electronic file transfer systems and requiring relevant stakeholders to sign non-disclosure agreements.

5. *Establish Documentation and De-Identification Guidelines and Tools.* The process for dissemination of public-use evaluation-related microdata files is now defined in MCC's Evaluation Microdata Guidelines.¹⁷ These guidelines detail the necessary documentation to facilitate understanding and use of the microdata, as well as the forms of data perturbation that are expected (such as removal, top and bottom coding, categorization) in order to produce de-identified, public-use data files as summarized and documented in the De-Identification Worksheet.¹⁸ As applicable, MCC follows the standards discussed in NIST (2016): “*de-identification is a term used within the US Government to describe the removal of personal information from data that are collected, used, archived, and shared. De-identification is not a single technique, but a collection of approaches, algorithms, and tools that can be applied to different kinds of data with differing levels of effectiveness. In general, the potential risk to privacy posed by a dataset's release decreases as more aggressive de-identification techniques are employed, but data quality decreases as well.*”

6. *Identify and establish dissemination mechanism.* As the Evaluation Microdata Guidelines were developed, practical questions of how to disseminate the public-use evaluation-related microdata and its corresponding documentation emerged, including:

- How can users easily search the MCC website to determine what microdata is available?
- How will all the corresponding documentation be posted in order to facilitate understanding of the study context and use of the data by new data users?

¹⁶ Available online at - https://www.mcc.gov/documents/reports/Short_Informed_Consent_Statement.docx.

¹⁷ Available online at - <https://www.mcc.gov/resources/doc/guidance-evaluation-microdata-guidelines>

¹⁸ Available online at - <https://assets.mcc.gov/documents/drj-data-package-worksheet.docx>

As MCC grappled with these questions, it learned that the World Bank's Development Data Group and International Household Survey Network, an informal network of international agencies, had already made contributions in this area with the development of an open-source software data catalog.¹⁹ MCC worked with the World Bank to develop a template for the MCC Evaluation Catalog²⁰ (launched in 2013) that packages the materials (microdata, evaluation design report(s), questionnaire(s), analysis report(s), microdata file(s), etc.) by evaluation, rather than by survey. This allowed MCC to establish a searchable catalog of its evaluations and the corresponding microdata, providing users with a one-stop shop for all documentation needed to understand and use the evaluation-related microdata.

7. Institutionalize an internal Disclosure Review Board. Noting a gap in the existing institutional structures – where IRBs may guide how evaluators manage data collection and storage, but have less guidance and oversight on data dissemination – MCC established an internal body responsible for overseeing microdata dissemination. The MCC Disclosure Review Board (DRB) was established in May 2013 with the objective of reviewing all evaluation-related microdata packages intended for public release on the MCC Evaluation Catalog.²¹ The DRB includes representatives from all MCC departments,²² as well as non-voting external advisors.

Every evaluation-related microdata package ready for public dissemination is submitted to the DRB with the following components: completed de-identification worksheet, corresponding public-use data file, metadata describing the study, questionnaire(s), and informed consent(s). The package is then presented to the DRB for discussion to determine the extent to which the proposed de-identification efforts meet the three objectives of maximizing verification, maximizing usability, and minimizing risk.

8. Consider how to manage restricted-access microdata. Once the DRB was fully functional, MCC identified a new hurdle: how to manage data whose release is required for meeting the objectives of maximizing verification and/or usability, but which cannot be publicly released due to anticipated risks related to re-identification of survey respondents. In other words, when it comes to public release of microdata, the agency must err on the side of reducing verification and/or usability in order to minimize risk to the respondents.

To tackle this issue, MCC decided, when possible, to ensure a restricted-access data file is produced. To do so, the team responsible for de-identification must determine all de-identification efforts that can still be maintained (particularly removal of any direct PII) without affecting verification and/or usability. This produces a data file that typically contains indirect identifiers. Guided by what is stated in the informed consent, researchers who agree to certain terms are allowed to have restricted-access to this data file for a clearly defined time period. Chief among the terms for restricted-access is a

¹⁹ Available online at - <http://www.ihsn.org/software/nada>.

²⁰ Available online at - <http://data.mcc.gov/evaluations/index.php/catalog>.

²¹ The DRB Charter is available online at <https://data.mcc.gov/evaluations/index.php/catalog/155/download/895>

²² The DRB is chaired by the Department of Policy and Evaluation (DPE) Vice President, and comprises two technical staff from DPE, and one staff from each of the Office of General Counsel, the Department of Compact Operations, the Department of Administration and Finance, and the Department of Congressional and Public Affairs.

certification and agreement by the researcher's institution not to share the data with any other person or entity, to only use the data for statistical purposes and to make no attempt to re-identify any survey respondent. Further, access to this restricted data is provided only after the DRB has reviewed and approved the full data package for both the public-use and the restricted-access microdata files, and the researcher has submitted to MCC a Restricted-Access Certification Form²³ that sets forth the full terms of their agreement.

5.2. Resources

Over the course of developing its principles for, MCC identified a range of additional required resources.

MCC and country partner staff. As a result of adopting the new principles and practices, MCC staff across the agency needed to increase their level of effort around these activities. The additional tasks in the Monitoring and Evaluation (M&E) unit included completion of training on protection of human subjects,²⁴ sharing and discussing principles with the country partners, MCC's evaluators, and other stakeholders, improving documentation practices internally in order to facilitate public release of materials, pilot testing the Evaluation Catalog, participating in preparation and delivery of data packages to the DRB, and developing work plans with the evaluators on new tasks. In addition, MCC staff at various levels, from technical to senior management, were identified to serve on the DRB, requiring a bi-monthly commitment to review materials and participate in meetings.

Independent evaluators. For independent evaluators, the introduction of new tasks required changes to contract requirements and budgets. The new tasks included budgeting for and procuring IRBs as necessary, building the IRB preparation and review process into the data collection timelines – which may require 1-3 months – executing the data documentation and de-identification requirements for public-release and/or restricted-access microdata, as well as preparing and presenting the de-identification process to the DRB and executing any required follow up.

Technical consultants. There were two additional technical support activities required. First, given this was a new area for MCC, the agency contracted two external advisors²⁵ in data de-identification and dissemination to advise the DRB from 2013-2015, and provide guidance and technical support to the agency on reviews of the evaluation-related microdata packages. While the ultimate goal was for the DRB to rely on the technical capabilities of staff, the guidance of these advisors was fundamental in the creation of the principles and resources and strengthening MCC's internal capacity. Second, at the time that the DRB was established, there was a backlog of identifiable data that MCC stored from closed evaluation contracts. As the contracts had expired, the evaluators were no longer contractually obligated to undertake the microdata documentation and de-identification, so MCC engaged a de-

²³ The Restricted-Access Certification Form template is available online at – https://www.mcc.gov/documents/reports/MCC_Restricted_Use_Data_Form.docx.

²⁴ See National Institute for Health free, online course - <https://phrp.nihtraining.com/users/login.php>.

²⁵ Each expert had a minimum of 20 years of experience working on data privacy issues at the United States Census Bureau and the National Center for Health Statistics, respectively.

identification consultant²⁶ to review the stock of data and execute the de-identification process, including presentation to the DRB.

6. Lessons learned

There are several lessons learned that may be useful for other funding institutions, development agencies, and country partners facing similar issues.

Learn from and collaborate with others. There are many U.S. agencies, particularly statistical agencies, that have tackled some, if not all, of the challenges discussed above. Identifying who they are and reaching out to hear their successes and failures was a means by which MCC identified key advisors, such as the external technical advisors for the DRB. Similarly, there are many academic resources, such as the Inter-university Consortium for Political and Social Research (ICPSR) resource library, to inform how to approach data stewardship issues.²⁷

Establishing clear principles improves practice and reduces institutional risk. Though MCC is not a signatory to the Common Rule - the U.S. Government system of protection for human research subjects in federally funded or regulated research²⁸ - the Common Rule served as the guide for the principles and resources MCC established, including requiring IRB review. With these principles in place, MCC's M&E data protection practices have improved. For example, local stakeholders (such as the project implementers) have made requests to access evaluation-related microdata files that include respondent PII in order to conduct their own analysis. In many cases this poses an immediate risk to the respondents as local stakeholders may attempt to re-identify respondents. The decision around whether or not these stakeholders may have access to the identifiable data, including PII, is now made by the IRB if requested during data collection, or by the MCC DRB if requested during data preparation and dissemination, in each case, based on the content of the informed consent. This places the decision to allow access to this type of data in an appropriately technical and neutral setting, rather than the sometimes political environment in which MCC's project implementation teams, their country partners, and evaluators operate.

Continuous learning and adaptation are key. MCC's first attempt to release de-identified data was in 2010. At that time, the M&E unit established internal guidelines for de-identification and managed the release of four microdata files on the MCC website. Once the agency understood the diverse issues and challenges with de-identification and public release of this type of data, it removed the four microdata files from its website to send them through the more formal review process established in 2013. The formal review process raised issues that had not been previously considered resulting in more rigorous de-identification of the original four public-use microdata files. MCC learned quickly that establishing principles and resources was only the first step – frequently returning to the principles and resources to update based on lessons learned was also critical.

²⁶ This expert had 10 years of experience working on statistical analysis of microdata and was trained in de-identification methods.

²⁷ <http://www.icpsr.umich.edu/icpsrweb/content/about/data-stewardship.html>.

²⁸ Each of the federal agencies participating in this system adapts the basic ethical principles outlined in the 1979 Belmont Report to requirements of the agency's own governing regulations. A summary of the rule, and relevant references are available here - <http://www.hhs.gov/ohrp/regulations-and-policy/statutes/index.html>

Getting the right balance between competing objectives is hard. The challenges with striking the right balance between the three objectives became abundantly clear. For example, de-identification efforts required to minimize risk to a respondent often require *necessary* data perturbation that limits verification of evaluator results and/or usability. Specific MCC examples include:

- Upon review of one data package, it was determined that given the small number of treatment villages, which were announced through a public lottery, inclusion of the ‘treatment’ variable for the program would allow for re-identification of villages in the data. Given the small population size within villages, inclusion of this variable was thought to significantly increase the risk of re-identification of households sampled in these villages. While this variable was required for the impact evaluation analysis, and therefore any potential future verification of results, including this variable also introduced too high a risk to the respondents in these villages. With this in mind, the DRB decided that the treatment variable would not be included in a public use file, and instead considered for restricted-access.
- Upon review of another data package, it was determined that inclusion of disaggregated data on local languages significantly increased the risk of re-identification of households given the small population associated with these local languages. Because this evaluation focused on a local language reading program, the ability to disaggregate data by local language was an important analytical requirement. With this in mind, the DRB determined that the local languages would be collapsed in the public-use data files, combining rare and potentially identifying languages into another category, but the full set of languages would be made available on a restricted-access basis.

In addition, placing the responsibility for minimizing risk to respondents on the de-identification analyst creates an incentive for that analyst to become overly risk averse, sometimes resulting in *unnecessary* data perturbation that severely limited verification and/or usability. For example:

- For the review of one data package, the de-identification analyst proposed removal of variables related to child labor and unemployment. However, the DRB discussed how unlikely it was that a household and individual could be re-identified based on any of these factors and also weighed the cost of removing these variables to the overall usability of the data, and determined these variables should remain in the public-use microdata file.

An important overall lesson is that it is the responsibility of the evaluators – under the review of IRBs - to design and implement a credible and ethical research plan, including appropriate data collection, management, and analysis. And it is the responsibility of MCC – through its DRB - to work with the individuals responsible for microdata de-identification to manage a balanced approach for public and restricted-access release of microdata, understanding the associated ethical, legal, and practical issues on a case-by-case basis.

7. Challenges Ahead

With these valuable lessons in hand, MCC looks forward to continuing to improve its transparency efforts for microdata and to tackling the challenges ahead:

- *Transparency Documentation.* The majority of analysis presented in evaluation reports uses microdata that is never made public because it uses confidential identifiable data. Evaluators and

MCC must consider how evaluator reports present results using the de-identified, public-use data—in the report or in annexes—to facilitate improved understanding of any limitations for verification of results. In addition, while the DRB reviews the de-identification efforts in detail, this information is not publicly available to reduce opportunities for an intruder to ‘re-identify’ respondents. MCC is now considering what documentation is needed to better document the data de-identification efforts for any new users of the data.

- *Identifiable Data Repository and Stewardship.* MCC is actively exploring how it may work with evaluators to identify appropriate mechanisms for storing and managing identifiable data (that includes any PII and sensitive data that may be required for further statistical analyses) that is deemed too risky for the current restricted-access mechanism. This is an area where learning from other agencies, particularly federal statistical agencies, may help inform a way forward. For example, the Office of Management and Budget recommends that non-statistical agencies consider partnering with one of the federal statistical agencies to conduct the required data collection.²⁹
- *Qualitative Data.* MCC recognizes that qualitative data poses unique risks and challenges that have largely not been tackled by the process described here. While many of the same principles apply broadly, MCC is now exploring³⁰ how its tools and process for management and release of microdata can apply to the qualitative research it funds to complement many evaluations.

This paper documents MCC’s experience with working toward open data goals for one particular type of data – evaluation-related microdata – while also balancing a key risk to making this type of data open – loss of confidentiality for the human subjects who provide this microdata. As presented here, striking this balance presents many tradeoffs and there is no ‘one size fits all’ solution given what may be considered sensitive in one context for one survey may not be in another context and another survey. Although statistical agencies have been tackling these issues for years, with the dramatic increase in demands for data-driven policy and policy relevant research, non-statistical agencies like MCC have had to tackle these issues in unique contexts – without the same standards around legal protection of the data or clear, specific, and consistent definitions of data ownership. By presenting the MCC process here, we hope that not only similar agencies can learn from our experience, but improve upon it so that we are all working toward improved global practices in the management and dissemination of microdata.

²⁹ Implementation Guidance for Title V of the E-Government Act, Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA), October 2006

(https://www.obamawhitehouse.gov/sites/default/files/omb/fedreg/2007/061507_cipsea_guidance.pdf).

³⁰ This is a broader discussion in social science research. MCC is following specific conversations, like [the Qualitative Transparency Deliberations](#) out of the [Data Access-Research Transparency](#) efforts in political science.

References

Alderman, Harold, Jishnu Das, and Vijayendra Rao. 2016. "Conducting Ethical Economic Research: Complications from the Field." In *The Oxford Handbook of Professional Economic Ethics* edited by George DeMartino and Deirdre McCloskey. Oxford University Press, April.

Dupriez, Olivier and Ernie Boyko. 2010. Dissemination of Microdata Files. Formulating Policies and Procedures. International Household Survey Network, IHSN Working Paper No 005.

Glennerster, Rachel, and Shawn Powers. 2016. "Assessing Risk and Benefit: Ethical Considerations for Running Randomized Evaluations, Especially in Developing Countries." In *The Oxford Handbook of Professional Economic Ethics* edited by George DeMartino and Deirdre McCloskey. Oxford University Press, April.

Hanson, Heather and Catherin Marschner. January 2015. "Transparency" Millennium Challenge Corporation Principles into Practice. <https://assets.mcc.gov/reports/paper-2015001163301-principles-transparency.pdf>

Ho, Jeff C., Kory C. Russel, and Jennifer Davis. 2013. "The challenge of global water access monitoring: evaluating straight-line distance versus self-reported travel time among rural households in Mozambique". *Journal of Water and Health*. IWS Publishing. DOI:10.2166/wh.2013.042

Miguel, E., C. Camerer, K. Casey, J. Cohen, K. M. Esterling, A. Gerber, R. Glennerster, et al. 2014. "Promoting Transparency in Social Science Research." *Science* 343 (6166): 30–31. doi:10.1126/science.1245317.

NISTIR 8053, De-Identification of Personal Data, Simson Garfinkel, September 2015, National Institute of Standards and Technology, Gaithersburg, MD. <http://dx.doi.org/10.6028/NIST.IR.8053>.

NIST Special Publication 800-188 (2nd DRAFT) De-Identifying Government Datasets, Simson Garfinkel, December 2016, National Institute of Standards and Technology, Gaithersburg, MD. http://csrc.nist.gov/publications/drafts/800-188/sp800_188_draft2.pdf

Poe, Ted. (10/20/2015). "H.R.3766 - Foreign Aid Transparency and Accountability Act of 2016." Legislation. July 15. <https://www.congress.gov/bill/114th-congress/house-bill/3766>.

Ryan, Paul. 2016. "H.R.1831 - 114th Congress (2015-2016): Evidence-Based Policymaking Commission Act of 2016." Legislation. March 30. <https://www.congress.gov/bill/114th-congress/house-bill/1831>.

Sturdy, Jennifer, Sixto Aquino, and Jack Molyneaux. 2014. "Learning from Evaluation at the Millennium Challenge Corporation". *Journal of Development Effectiveness*. Taylor & Francis. DOI:10.1080/19439342.2014.975424