Developing a Plan to Assess the Quality of Administrative Data on Non-Degree Credentials for Skilled Technical Workers: Models and Options

Including a preliminary literature review and outline of critical decisions to be made

Potential Value of Administrative Data and Need for Quality Assessment

The most common sources of indicators on the attainment of non-degree credentials for American workers – regardless of their field – is national survey data collected by various units of the federal government. From 2010 through 2018, the federal government made tremendous strides in collecting data on non-degree attainment through the Interagency Working Group on Expanded Measures of Enrollment and Attainment (GEMEnA), which facilitated the placement of indicators of non-degree attainment on a range of federal surveys, including the Census Bureau's Survey of Income and Program Participation and Current Population Survey, NCSES' National Survey of College Graduates, and an original survey of adult educational attainment administered by the Department of Education, the Adult Training and Education Survey (ATES).

These surveys, however, are costly to administer and have inherent limitations in their sampling frame and the level of detail they can collect on each respondent/credential. Given these drawbacks and the apparent promise of new and emerging administrative data sources, we are planning to conduct ***an assessment of the quality of administrative data sources with the intention to build a repository of administrative data on non-degree attainment*.** This literature review is intended to clarify and inform the decisions that will need to be made as we plan our quality assessment.

Definitions of administrative data vary, though usually administrative data is framed in contrast to survey data: unlike a survey that samples a limited portion of a larger population on which one can draw inferences, ***administrative data are intended to paint a comprehensive picture of an entire population*** – whatever that population may be. Typically, administrative data is collected or generated in the course of performing administrative tasks, such as licensing vehicles, treating patients, conferring degrees, or collecting taxes. Unlike surveys, administrative datasets are not designed with researchers' considerations or priorities in mind, which tends to be reflected in how variables are designed and collected.

Administrative data has recently been defined in contrast to "big data" resulting from digital data collection by software apps and electronic devices, which can create massive datasets (potentially billions of records) when every click or movement is tracked (Graeff and Baur 2020), though other researchers have defined administrative data as a subset of big data (Connelly et. al. 2020).

Most research on the quality of administrative datasets to date has occurred in the context of medicine and public health, an environment relatively rich in records (e.g., medical charts,

billing records) that are largely digitized, but often existing in "walled" systems such as individual hospitals or state healthcare agencies. Other relevant research, which includes efforts to develop rubrics for the evaluation of the quality of administrative datasets, has been conducted by national statistical agencies.

Researchers affiliated with the ***statistical and/or census agencies of several countries*** – including the United States, Sweden, the Netherlands, New Zealand, and Australia – have published ***frameworks by which they evaluate the quality of administrative datasets***. In some cases, statistical agencies published the results of their evaluations of national data systems. These evaluations range in their level of completeness and in the extent to which they attempt to quantify the quality of a given dataset, rather than offering more qualitative assessments. While some assessments have been undertaken by researchers looking to evaluate the quality of datasets for academic or theory-driven research, others have focused on identifying ways to enhance the quality of datasets for administrative purposes.

As we analyze prior literature and develop any instruments or rubrics for our assessment, we will attempt to account for some of the differences in the nature of the healthcare and other national datasets that have been the subject of most data quality assessments to date and the data sources we anticipate encountering in our research on non-degree credentials.

Dimensions of Quality

To summarize literature reviews on this topic, we note that ***several dimensions of quality seem to come up again and again*** in quality assessment rubrics for administrative data:[1]

- Timeliness: There is no agreement in the literature on how recent a dataset should be before it becomes "stale," and requirements for recency likely depend on one's research question as well as the availability of alternate data sources.
- Accuracy: Data should have been recorded accurately and consistently regardless of individual differences between different offices and individuals who might have recorded data. Coding of variables should be consistent and free of errors.
- Relevance: Data should be relevant to an administrative dataset's stated purpose. For example, an administrative dataset intended to measure educational performance could include grades, test scores, or credentials completed. Records of an educational institution might also include data on extracurricular activities, dining and parking purchases made by students, or disciplinary actions. Such records not related to academic performance might not be relevant if we are attempting to measure educational attainment, though they can be removed or ignored by researchers.

[1] We will avoid devoting too much time and space to issues that other researchers have already written on at length. See, for example, Chen et al. 2010; Daas et al. 2010; Iwig et al. 2013; Laitila, Wallgren, and Wallgren 2011; Smith et al. 2017; Hand 2018.

- Compatibility/Interoperability: In general, records in higher quality datasets can be linked to external data sources because they contain unique identifiers and codes that are common to other datasets. For example, administrative data on worker pay may be higher quality if workers are accurately coded by occupation or industry using Occupational Information Network (O*Net) or North American Industry Classification System (NAICS) codes, enabling one to do analyses with occupation-level covariates. Unique identifiers, such as social security numbers, that allow for the matching of individual workers, are particularly valuable.
- Completeness: In general, datasets should have as little missing data as possible. This means as few "holes" in terms of missing data for specific variables, and the dataset should contain records for as much of the population it purports to cover as possible. Some researchers may find that imputed data where missing values exist is adequate, but, in general, researchers consider imputed data to be less desirable than data accurately recorded from a primary source. Similarly, researchers want to avoid datasets with duplicate records, especially if it is difficult for the end user to find and remove duplicates (Rothbard 2015).
- Granularity: The level of detail in measures of particular data elements in an administrative dataset. For example, a dataset that codes individuals' occupations into a 300-category coding scheme may be more granular than one that codes individuals into 30 distinct occupations, regardless of whether either coding scheme is interoperable with other datasets. Other common areas in which datasets differ in granularity include geographic detail (e.g., the unit of geography used to identify a location, ranging from a country or state down to a street address or even detailed latitude and longitude), and units of time (years vs. months or dates).
- Integrity: The process of collecting administrative data should be free of undue political influence and should not be subject to unauthorized modification. Established scientific methods should be used in the collection of data and creation of variables. This requires that datasets be secure and that confidentiality be maintained, if confidentiality is promised to the individuals/organizations from which data is collected. Credibility is another sub-component of integrity; datasets may be more credible if past datasets published by the same organization were of high quality.[2]

In addition to the above, accessibility is also a factor that some may consider to be a dimension of quality of relevance for our project. The logic in considering accessibility is simple: in order for an administrative dataset to realize its value for research, researchers must be able to access it. Given that academic researchers often have limited funds to pay data access fees and may face challenges in meeting the security standards set by the owners of administrative

[2] Note that integrity is emphasized as a distinct concept by the U.S. Federal Committee on Statistical Methodology, in contrast with other national frameworks that fold elements of integrity into other dimensions of quality; see the 2020 report "A Framework for Data Quality."

datasets, from the researcher's perspective some datasets that score highly on other dimensions of quality may be unacceptable (or even impossible) for use in research.

Even if one is able to overcome organizational obstacles to arranging data access, one may face difficulty when it comes to obtaining support once access is obtained or be forced to work from inadequate metadata (Rothbard 2015). Given that "messiness" has been described as a common characteristic of administrative datasets commonly used by social scientists (Connelly et. al. 2016), the availability of support after access is acquired is another important consideration for researchers working with datasets that may not have an active community of data users.

Two Approaches to Quality Assessment

The project team ascertains that there are two major approaches to assessing data quality. The first approach relies on ***direct determination of data quality***. The second assesses ***data producer methodologies and quality control practices***.

Outcomes Approach: In the first approach, some researchers examine the ***extent to which a data source captures reliable data*** on a particular set of characteristics of interest. This approach has often been favored by researchers affiliated with U.S. federal agencies. It also has been used frequently in the context of international public health, for example in a study of records generated by HIV clinics in Malawi (Makombe et. al. 2008). In such public health studies, a common approach has been for researchers to "audit" the records generated by an agency creating an administrative dataset and attempting to gauge to what extent records were created in accordance with statistical best practices. Of course, the "audit" approach depends on the efforts of researchers to ensure that the audit itself is done to acceptable quality standards. In the case of public health audit studies, researchers often have resources to verify records that clinics administering records may not have had at their disposal (for example, medical testing resources).

The direct approach to quality assessment can also rely on statistical analyses, especially with respect to imputed data. Checks can be performed to look for outliers and anomalies that may point to data collection errors or biases (Lenk et. al. 2014; Seeskin, Ugarte, and Datta 2019). Indeed, Seeskin, Ugarte, and Datta (2018, 2019) have even published a program for the "R" statistical package designed to assist data users in the identification of quality issues in administrative datasets.

In theory, such checks could be used to quantitatively score the quality of administrative datasets against each other, though any such scoring would be limited to those "variables" that can be measured quantitatively and may miss more subjective measures (such as whether the questions posed by the organization that collected the data were accurately understood by individuals providing data).

Process Approach: In the second approach, favored by some foreign governments and independent scholars, evaluators determine the ***strengths and weaknesses of how a dataset is constructed***. The process of collecting and publishing data, rather than the final dataset, is the object of analysis. For example, researchers may take note of how much missing data exists in a dataset and the strategies used for dealing with missing data (e.g., omission, imputation). Evaluation may ascertain the validity and quality of how data is classified in a dataset, including whether a dataset uses commonly accepted occupation coding schemes and whether a dataset uses definitions of non-degree credentials that are consistent with those used in the wider research community.

A process-oriented approach requires some level of cooperation from individuals and organizations familiar with how a dataset was constructed, and/or the ability to analyze documentation explaining how a dataset was created. There is some risk that one may be misled by such informants (intentionally or unintentionally) or by one's analysis of archival documents if the creation of a dataset was insufficiently documented or if the creators of a dataset want to project an unjustified perception of quality (e.g., if they stand to profit from the sale of a data license).

The process approach does not, however, necessarily require one to have access to the raw data itself. Therefore, it may be suitable in situations where researchers find that they must assess the quality of a dataset without access to microdata. The UK Statistics Authority (2019) also notes that assessments of the process of creating datasets can be used to determine the risk of problems with data quality that would warrant the performance of direct quality assessments.

Determination of Relevance

Defining the relevance of the measures and variables in a dataset requires one to be able to clearly articulate what one wants to get out of the administrative dataset.

In our use case – non-degree credentials in the skilled technical workforce – we would need to have a ***clear idea of what we mean by non-degree credentials and the skilled technical workforce*** that we could benchmark a dataset against. If we do not have a clear definition of, for example, the minimum number of contact hours for a program of study to be considered a non-degree credential or a clear set of criteria for deciding if a credential falls within the domain of the skilled technical workforce, we may struggle to decide how to judge a given dataset on the metric of relevance.

We also need to decide whether we are interested in all non-degree credentials, or whether certain subsets of credentials are of interest. Credentials may be held by members of the skilled technical workforce that are not necessarily relevant to technical work (for instance, because one changed careers). If we decide that we would want to filter out such credentials from our analysis, we would need to be able to identify them. Similarly, we may have access to quality datasets on certain types of credentials, such as certifications or licenses, and lower quality

datasets on other types of credentials (such as "bootcamp" certificate programs in information technology). The decision to focus on certain credentials and certain occupations will influence the quality of data available for analysis, which may lead to issues when we begin to analyze data.

Determination of Assessment Rigor

Different research purposes may lend themselves to different levels of rigor in the conduct of a data quality assessment (Romano 1993; Laitila 2014). For example, research involving life-or-death medical outcomes may be held to higher standards of evidence than research that affects less consequential policy decisions. Researchers may also disagree on whether there is a necessary minimum level of quality for a given purpose. If a dataset does not meet standards of quality data but no alternative sources of data are identified, researchers may find it worthwhile to proceed in spite of their concerns.

A data quality assessment need not assign a score to each dataset under consideration. Assessments may be qualitative in nature and can be valuable to researchers regardless of whether datasets are categorized in any meaningful way with respect to their overall level of quality. The UK Statistics Authority (2019) notes in its administrative data quality framework that it may be reasonable to hold datasets concerning phenomena of lower levels of public interest (without providing guidance for determining such interest) to lower standards of quality, or to accept less through assessments of such datasets. Ultimately, researchers will make qualitative judgements about which trade-offs are acceptable when choosing a dataset, or will choose to draw upon multiple complementary data sources to reach nuanced answers to their research questions.

We can also potentially apply principles of total survey error to the measurement of administrative data, even if administrative data is inherently different from survey data in important respects. Many of the quality issues that affect administrative data are comparable to those that affect surveys, such as the potential for error in how data is imputed (Groen 2012). A "total error" framework has been proposed for considering sources of error that could occur throughout the entire lifecycle of creating and analyzing non-survey datasets (Amaya et. al. 2020) and considering all aspects of a dataset holistically. Such a framework would lead us to evaluate datasets according to multiple potential sources of error and attempt to draw conclusions on the basis of a holistic consideration of all potential error sources.

Considerations

We have identified the following considerations as particularly relevant to our effort to evaluate sources of administrative data on non-degree credentials:

Interoperability: The need to unify unique and disparate data sources is driving our effort to design and create a data repository on non-degree credentials. Given that there is not one single widely accepted list of non-degree credentials, our repository will rely on linkages

between datasets. Evaluating the suitability of datasets for interoperability (e.g., the presence of common identifiers for individual credentials) may be especially important for our purposes.

This is especially important if we are to maximize the value of the many sources of survey data that can potentially yield complementary insights when linkages are made at a given level of geographic detail or at the level of individual firms, credentials, occupations or industries. Links can be made at the individual level using a common identifier – for example, a social security number – or by inferring that an individual appears in multiple datasets with less precise data fields like names (which could be prone to misidentification if one gives different names or nicknames to different data collectors or at different points in time), addresses (which may be inaccurate if individuals move between residences), and dates of birth.

In general, the more personal detail that a dataset contains about an individual, the more likely that a linkage can be made and the lower risk of error in the linkage process (Berka *et al.* 2010). Privacy concerns may prevent (indeed, in most situations, should prevent) personally identifiable information from being included in datasets that can be directly downloaded by researchers without going through a licensing process; whether more onerous access procedures should be treated as an indicator of lower dataset quality from the perspective of researchers is a subjective decision that should be considered carefully when designing quality assessments.

Timeliness and suitability for longitudinal research: Unlike the healthcare records studied in most administrative data quality assessments, administrative records on educational and occupational attainment are often point-in-time, not clearly dated, or only updated sporadically. Anecdotally, we know that non-degree credentials, while not a new phenomenon, have proliferated in recent years – but by in large we lack consistent data (survey or administrative) that would enable longitudinal analysis of rates of prevalence prior to the addition of certifications to the Current Population Survey in 2015.

The scarcity of longitudinal records in our space may lead researchers to place higher value than would be accepted in other fields on administrative records that are not point-in-time and/or that contain inconsistencies in how data is recorded over time. However, the suitability of records for longitudinal research should be prioritized by researchers looking to identify trends in the value and attainment of non-degree credentials – trends that are likely to be of great interest to researchers and policymakers seeking to evaluate the effect of past policy choices.

Inclusion of relevant covariates, especially concerning income: As noted by scholars cited above, datasets may be of higher quality when they contain not just more data, but richer data – which may include more (and more detailed) variables. All things equal, more covariates tend to be better from the researcher's perspective – especially when it comes to measuring a phenomenon with multiple dimensions.

In the context of credentials, for example, knowing that one has a credential is important but so is the field of study, the institution that awarded the credential, the cost of tuition, the length of the course of study, and so on. All datasets containing indicators of non-degree attainment could potentially have value for researchers, but given the research community's interest in identifying credentials of labor market value there is potentially greater value attached to administrative datasets that, either in and of themselves or through linkages, enable us to identify the relationship between the attainment of individual credentials and income.

Detailed income measurements that go beyond annual total income – for example, hourly pay rates and the value of non-wage compensation – would be especially useful for some research questions our researchers may ask and have been flagged as especially valuable covariates in previous research on administrative data quality (Allard et. al. 2018).

Practicability: Though not extensively covered in the literature on data quality assessments, research teams attempting to assess the quality of a dataset must work within the constraints of their resources – which include time, money, expertise and access.

For the purposes of our study, we anticipate that all four types of resource constraints will affect us on at least some level. Money and access are intertwined insofar as some data sources are privately held and must be licensed for a fee, and costs may be associated with meeting the security requirements for accessing a dataset (for example, the fee charged to researchers by some restricted access data facilities). If a fee is prohibitive for researchers seeking to use a dataset to create generalizable knowledge about the credentials held by the skilled technical workforce, it is most likely also prohibitive for researchers seeking access for the purpose of conducting an initial data quality assessment – which would potentially force resource-constrained researchers to rely on research on how the dataset was constructed to assess its quality.

For the purposes of our project, we plan to evaluate as many datasets as is justified by researcher and stakeholder interest – though interest in an overwhelming number of datasets (particularly if we choose to analyze each state's longitudinal data system separately) may force us to reduce the quality of our assessments to complete our project on time and on budget. In this case, it may be warranted to exclude datasets that are unlikely to be accessible to most academic researchers.[3]

Conclusions

While there are common themes in the emerging frameworks for data quality assessments with administrative datasets, ***there are still many approaches to choose from*** – forcing researchers

[3] Some precedent for such a decision exists in the Australian Data Quality Framework (Australian Bureau of Statistics 2008), which defines the availability of microdata to qualified researchers as one of seven official dimensions of quality against which administrative datasets should be evaluated.

to think carefully and strategically when making decisions about what to prioritize in a quality assessment.

There is no commonly agreed upon rubric or instrument for assessing the quality of administrative data, nor is there agreement on whether or how one should quantify the value of a dataset. Nonetheless, we have described some of the considerations that may be especially relevant to conducting an assessment of the quality of administrative datasets for research on non-degree credential attainment in the skilled technical workforce with the goal of helping stakeholders provide useful input at this early stage in the development of our repository.

It is important to be realistic about what is possible with respect to assessing data quality. In the context of applying a total error framework to non-survey data, Amaya et. al. (2020: 116) note:

> "The information required to conduct the ideal investigation is rarely, if ever, available. In some cases, this is because the truth may be un-knowable, the information is proprietary, or the effort required would be cost prohibitive. Instead, we must use what is available and attack the problem from several different angles, relying on the resulting big picture as opposed to the individual tests."

Thus, ***decisions about the methods to use in assessing administrative data quality will ultimately be influenced by the availability of necessary information***. Some dimensions of quality can be assessed to a reasonable extent through the analysis of metadata alone, whereas others will require access to and cooperation from the "owners" of the data and/or the data itself.

The process-oriented and outcome-oriented approaches to determining quality both have distinct advantages and drawbacks. However, they are not mutually exclusive. As we conduct our assessment of the quality of datasets covering the non-degree credential attainment of the skilled technical workforce, ***we can use both approaches*** as needed to holistically review the qualities of each dataset.

From these qualitative and quantitative data points we can develop ***recommendations as to the relative quality of each dataset for the purposes of our research needs***, keeping in mind that our needs may be different from those of other researchers. As there is no single definition of quality for administrative datasets, each researcher can make their own judgement about what to prioritize in their own assessment. Nonetheless, we hope that our assessment will provide a useful model for other researchers to borrow from as they design their assessments and serve as a useful resource for all researchers interested in non-degree attainment in the skilled technical workforce.

References

Allard, Scott W., Emily Wiegand, Colleen Schlecht, A. Rupa Datta, Robert M. Goerge, and Elizabeth Weigensburg. 2018. "State Agencies' Use of Administrative Data for Improved Practice: Needs, Challenges, and Opportunities." *Public Administration Review* 78(2): 240-250.

Amaya, Ashley, Paul P. Biemer, and David Kinyon. 2020. "Total Error Framework in a Big Data World: Adapting the TSE Framework to Big Data." *Journal of Survey Statistics and Methodology* 8: 89-119.

Australian Bureau of Statistics. 2008. *ABS Data Quality Framework.* Canberra, Australia: Australian Bureau of Statistics. https://www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/1520.0Main%20Features1May%202009?opendocument&tabname=Summary&prodno=1520.0&issue=May%202009&num=&view=

Berka, Christopher, Stefan Humer, Manuela Lenk, Mathias Moser, Henrik Rechta, and Eliane Schwerer. 2010. "A Quality Framework for Statistics Based on Administrative Data Sources Using the Example of the Austrian Census 2011." *Austrian Journal of Statistics* 39(4): 29-308.

Chen, Hong, David Hailey, Ning Wang and Ping Yu. 2014. "A Review of Data Quality Assessment Methods for Public Health Information Systems." *International Journal of Environmental Research and Public Health* 11: 5170-5207.

Connelly, Roxanne, Christopher J. Playford, Vernon Gayle, and Chris Dibben. 2016. "The role of administrative data in the big data revolution in social science research." *Social Science Research* 59: (1-12) https://doi.org/10.1016/j.ssresearch.2016.04.015.

Daas, Piet, Saskia J.L. Ossen, and Martin Tennekes. 2010. *Determination of Administrative Data Quality: Recent Results and New Developments*. The Hague: Statistics Netherlands.

Federal Committee on Statistical Methodology [U.S.]. 2020. *A Framework for Data Quality*. Report FCSM-20-04. Washington, DC: Federal Committee on Statistical Methodology. https://nces.ed.gov/FCSM/pdf/FCSM.20.04_A_Framework_for_Data_Quality.pdf

Graeff, Peter, and Nina Baur. 2020. "Digital Data, Administrative Data, and Survey Compared: Updating the Classical Toolbox for Assessing Data Quality of Big Data, Exemplified by the Generation of Corruption Data." *Historical Social Research / Historische Sozialforschung* 45 (3): 244–69. https://www.jstor.org/stable/26918412.

Groen, Jeffrey. 2012. "Sources of Error in Survey and Administrative Data: The Importance of Reporting Procedures." *Journal of Official Statistics* 28(2): 173-198.

Hand, David J. 2018. "Statistical challenges of administrative and transaction data." *Journal of the Royal Statistical Society* 181(3): 555-605.

Iwig, William, Michael Berning, Paul Marck, and Mark Prell. 2013. “Data Quality Assessment Tool for Administrative Data.” Washington, DC: Federal Committee on Statistical Methodology. https://nces.ed.gov/FCSM/pdf/DataQualityAssessmentTool.pdf

Latilla, Thomas, Anders Wallgren, and Britt Wallgren. 2011. *Quality Assessment of Administrative Data.* Stockholm: Statistics Sweden. https://ec.europa.eu/eurostat/cros/system/files/S13P3.pdf

Lenk, Manuela, Franz Astleithner, Eva-Maria Asamer, Predrag Cetkovi, Henrik Rechta, Stefan Humer, Eliane Schwerer, and Mathias Moser. 2014. *Quality assessment of administrative data: Assessment of methods.* Vienna: Statistics Austria. http://www.statistik.at/web_de/static/documentation_of_methods_077211.pdf

Makombe, Simon D. Mindy Hochgesang, Andreas Jahn, Hannock Tweya, Bethany Hedt, Stuart Chuka, Joseph Kwong-Leung Yu, John Aberle-Grasse, Olesi Pasulani, Christopher Bailey, Kelita Kamoto, Erik J Schouten and Anthony D Harries. 2008. "Assessing the quality of data aggregated by antiretroviral treatment clinics in Malawi." *Bulletin of the World Health Organization* 86: 310-314.

Romano, Patrick. 1993. “Can Administrative Data Be Used to Compare the Quality of Health Care?” *Medical Care Review* 50(4): 451-477.

Rothbard, Aileen. 2015. *Quality Issues in the Use of Administrative Data.* Philadelphia: School of Social Policy and Practice, University of Pennsylvania. https://www.aisp.upenn.edu/wp-content/uploads/2015/06/Data-Quality-Paper_Final.pdf

Seeskin, Zachary H., Gabriel Ugarte, and A. Rupa Datta. 2019. “Constructing a toolkit to evaluate quality of state and local administrative data.” *International Journal of Population Data Science* 4(1): 1-11.

Smith, Mark, Lisa M Lix, Mahmoud Azimaee, Jennifer E Enns, Justine Orr, Say Hong, and Leslie L Roos. 2017. Assessing the Quality of Administrative Data for Research: A Framework from the Manitoba Centre for Health Policy.” *Journal of the American Medical Informatics Association* 25(3): 224-229. https://doi.org/10.1093/jamia/ocx078

UK Statistics Authority. 2019. *Administrative Data Quality Assessment Toolkit.* London: UK Statistics Authority. https://osr.statisticsauthority.gov.uk/wp-content/uploads/2019/02/qualityassurancetoolkit_updated_Feb19_2.pdf