IJDC | *Brief Report*

Using Metadata to Promote Transparency in Health Research: Creating the COVID Measures Archive at ICPSR

Megan Chenoweth John Kubale

ICPSR-University of Michigan

Abstract

Data sharing is a key strategy for fostering transparency, reproducibility, and trust in scientific research. Data sharing is endorsed and even required by many funders, such as the National Institutes of Health (NIH) in the United States. However, many NIH-funded projects face obstacles to data sharing, either to protect research participants' privacy, safeguard proprietary data, or remain compliant with data use agreements. Yet even researchers who cannot openly share data still benefit from openness and transparency into one another's work, and from making their own research more transparent where possible. The Social, Behavioral, and Economic COVID Coordinating Center at ICPSR (SBE CCC) has launched a new archive aimed at addressing these challenges within the domain of social, behavioral, and economic (SBE) research into the COVID-19 pandemic. In September 2023, SBE CCC launched the COVID measures archive with the dual goals of a) offering researchers the ability to compare measures across SBE studies of COVID while b) protecting contributors' needs for privacy and confidentiality in health research. The COVID measures archive primarily holds variable-level metadata, which provides visibility into the individual variables and measures employed in studies without necessitating the sharing of confidential or restricted data. This brief report describes the features of the COVID measures archive and illustrates how it can be used to foster transparency and consistency across SBE COVID studies.

Submitted 24 May 2024 ~ *Accepted* 25 June 2024

Correspondence should be addressed to Megan Chenoweth, Email: mmchenow@umich.edu

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

http://dx.doi.org/10.2218/ijdc.v19i1.977
DOI: 10.2218/ijdc.v19i1.977

Introduction

Data sharing is a key strategy for fostering transparency, reproducibility, and trust in scientific research. In January 2023, the United States National Institutes of Health (NIH) released a new policy for scientific data sharing requiring grantees to develop a data sharing plan and to share data in an established data repository.[1] One such repository is ICPSR,[2] a consortium for curating and sharing research data at the University of Michigan Institute for Social Research (ISR). ICPSR houses several dedicated NIH-funded data repositories, including Data Sharing for Demographic Research (DSDR), the National Archive of Computerized Data on Aging (NACDA), and the National Addiction & HIV Data Archive Program (NAHDAP).

Despite the importance of data sharing, and the mandates in place to promote it, health researchers face a number of obstacles when it comes to sharing their data. First is the need to protect the privacy and confidentiality of their research participants. Researchers must avoid disclosure of personally identifiable information and ensure proper sharing protocols for sensitive data. Second, many researchers utilize proprietary data in their research, such as the electronic health records of private healthcare providers or insurers, and the owners of these data may not allow further sharing. Even non-proprietary data, such as survey data held by a data repository, may be subject to data use agreements that limit researchers' ability to freely reshare data—including those generated from secondary use. In these circumstances, data sharing may be limited to comply with those data use agreements.

Yet even researchers facing these obstacles benefit from making their research as open as possible. One such benefit is that openness promotes interoperability across studies. Building knowledge about a topic requires comparing findings across studies. However, comparing studies requires measuring concepts the same way. This is only possible with visibility into measures from other studies. Additionally, being open and transparent about new measures benefits researchers by fostering dissemination and uptake of comparable measures across studies.

Another benefit is that openness fosters trust in institutions. Without transparency, study designers may use incompatible measures that result in conflicting findings. For example, a summary of two NBER working papers (Maas, 2022) highlighted conflicting conclusions about the relationship between school closures and parental employment. One working paper (Garcia & Cowan, 2022) found no significant relationship between school closures and whether parents work at all. Another (Hansen et al., 2022) found a significant relationship between mothers' working status and school closures. Lack of consistent measures does not fully explain, but does contribute to, the differing results between these two papers. These inconsistencies, especially regarding novel research topics such as the COVID-19 pandemic, may reduce trust in institutions when it is needed most.

In this brief report, we will describe an example of a group of researchers facing data sharing needs and challenges, and a tool developed by ICPSR to address those needs and challenges.

The Social, Behavioral, and Economic COVID Consortium and Coordinating Center

The Social, Behavioral, and Economic (SBE) COVID Consortium is a group of 15 NIH-funded research groups investigating the COVID-19 pandemic. Each consortium member operates its own independent research team with its own research questions, data, and methods. Consortium members come from a wide range of disciplines, including demography, economics,

[1] Data Management and Sharing Policy: https://sharing.nih.gov/data-management-and-sharing-policy
[2] ICPSR: https://www.icpsr.umich.edu

epidemiology, medicine, statistics, and sociology. As a consortium, their shared goal is to establish consistency and comparability across their research, for example, by establishing common data elements (CDEs) and incorporating shared measures into their studies. ICPSR hosts a coordinating center for this consortium, the Social, Behavioral, and Economic COVID Coordinating Center (SBE CCC). Its purpose is to foster communication and collaboration between the consortium, the social science research community, and the public.[3]

Members of the SBE COVID Consortium utilize a wide variety of data sources in their research. Some, such as the US County COVID Policy database and the National Neighborhood Data Archive (NaNDA), are free and open to the public. However, many are subject to access restrictions due to containing restricted, sensitive, or proprietary data. Examples of restricted data used by consortium research teams include SafeGraph mobility data (propriety), mortality data from the National Vital Statistics System (personally identifiable), electronic health records, and Medicare claims (personal health information). The access restrictions on these data sources represent an obstacle to open collaboration between consortium members.

How SBE CCC's COVID Measures Archive Addresses Data Sharing Challenges

SBE CCC's approach to this challenge has been to launch our COVID measures archive, a repository of variable-level metadata. Consortium members who are unable to share data can use the COVID measures archive to share study- and variable-level metadata from their research studies instead.[4] The COVID measures archive was launched in September 2023. As of February 2025 it contained 18,532 variables from 35 studies contributed by nine research teams (seven consortium members and two outside projects). The COVID measures archive has several key features that facilitate cross-study investigation and comparison at the variable level.

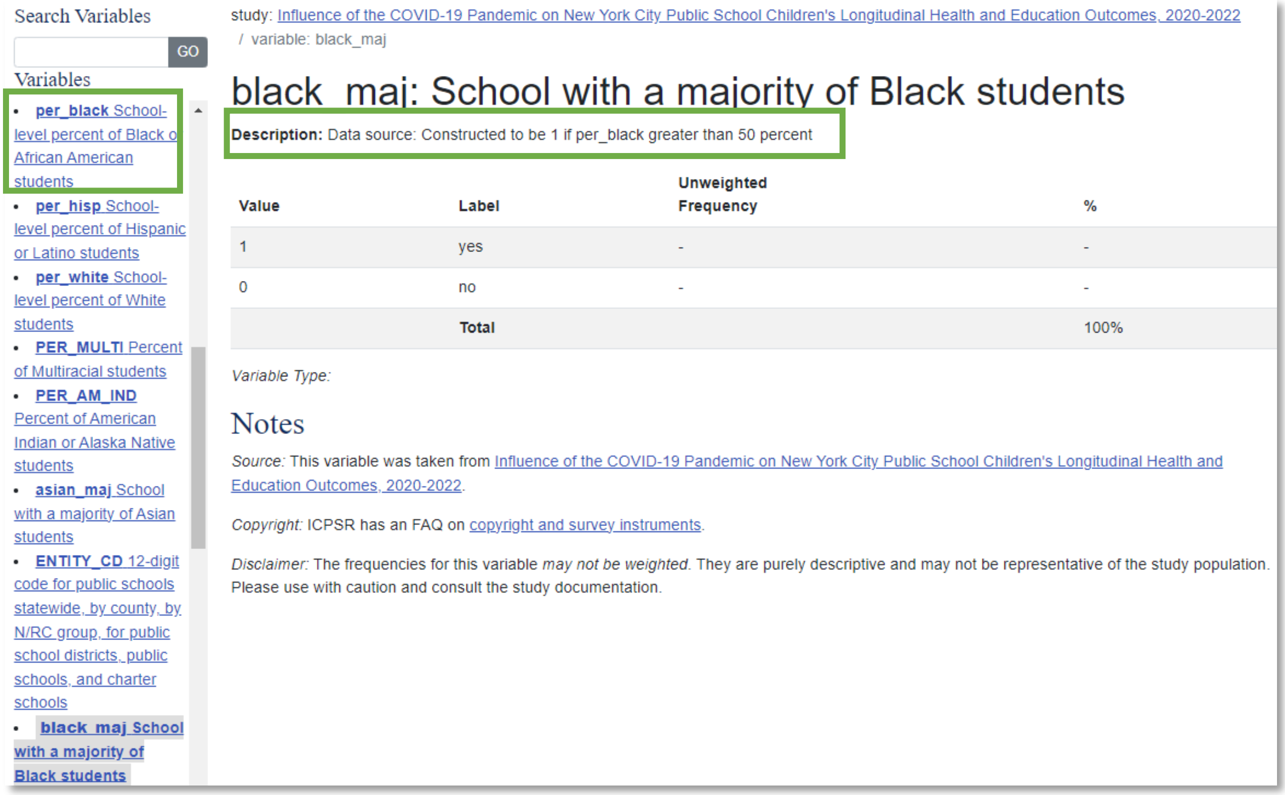
Centrality of Variable-Level Metadata

One key feature of the COVID measures archive is its focus on variable-level metadata. The main purpose of the archive is to foster visibility into the specific measures used in consortium members' studies. As a result, the main point of access is the searchable list of variables from all studies.

The variable-level metadata found in the COVID measures archive consists of a searchable list of variables along with, where relevant, question text and value labels for all possible values. For variables taken directly from other sources, there is a description of the data source as well as information on where and how to obtain the source data. For constructed variables, there is a description of how the variable was constructed. Figure 1 illustrates variable-level metadata for a created variable within the COVID measures archive. The description of how the variable was created and a variable from the same study used in its creation are indicated with green boxes.

[3] Social, Behavioral, & Economic COVID Coordinating Center: https://www.icpsr.umich.edu/sbeccc

[4] COVID Measures Archive: https://www.icpsr.umich.edu/web/sbeccc/search/variables



Search Variables

GO

Variables

- **per_black** School-level percent of Black o African American students
- **per_hisp** School-level percent of Hispanic or Latino students
- **per_white** School-level percent of White students
- **PER_MULTI** Percent of Multiracial students
- **PER_AM_IND** Percent of American Indian or Alaska Native students
- **asian_maj** School with a majority of Asian students
- **ENTITY_CD** 12-digit code for public schools statewide, by county, by N/RC group, for public school districts, public schools, and charter schools
- **black_maj School with a majority of Black students**

study: Influence of the COVID-19 Pandemic on New York City Public School Children's Longitudinal Health and Education Outcomes, 2020-2022 / variable: black_maj

black_maj: School with a majority of Black students

Description: Data source: Constructed to be 1 if per_black greater than 50 percent

Value	Label	Unweighted Frequency	%
1	yes	-	-
0	no	-	-
	Total		100%

Variable Type:

Notes

Source: This variable was taken from Influence of the COVID-19 Pandemic on New York City Public School Children's Longitudinal Health and Education Outcomes, 2020-2022.

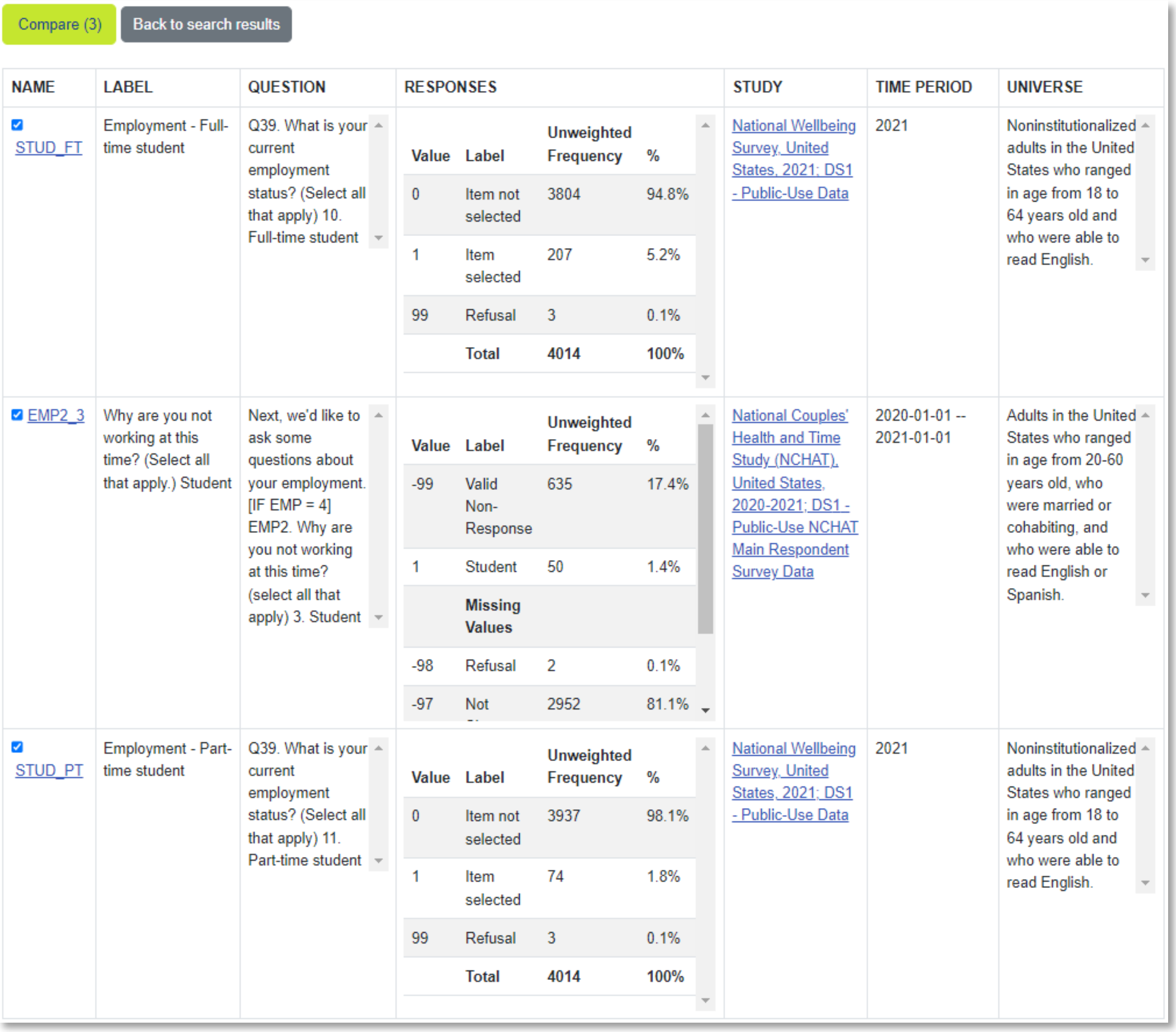
Copyright: ICPSR has an FAQ on copyright and survey instruments.

Disclaimer: The frequencies for this variable *may not be weighted*. They are purely descriptive and may not be representative of the study population. Please use with caution and consult the study documentation.

Figure 1 A screenshot of a variable, black_maj, in the COVID measures archive. The description of how the variable was created as well as another variable from the same study that is referenced in the description are indicated by green boxes.

Comparability Across Studies

A second important feature of the COVID measures archive is the ability to compare variables and measures across studies. The 'Compare Variables' function allows users interested in measures on a particular topic to search, select, and view relevant variables side by side. The comparison view includes variable names and labels, question text for survey questions, and values and labels for categorical variables. It also lists the study that each variable is taken from, as well as that study's time period and the universe, or population, to which it applies. This enables users to assess whether variables are identical, comparable, need harmonization, or are incompatible or inconsistent with one another. Figure 2 shows the results of a three-variable comparison.



<table>
<tr><th>NAME</th><th>LABEL</th><th>QUESTION</th><th>RESPONSES</th><th>STUDY</th><th>TIME PERIOD</th><th>UNIVERSE</th></tr>
<tr><td>☑ STUD_FT</td><td>Employment - Full-time student</td><td>Q39. What is your current employment status? (Select all that apply) 10. Full-time student</td><td><table><tr><th>Value</th><th>Label</th><th>Unweighted Frequency</th><th>%</th></tr><tr><td>0</td><td>Item not selected</td><td>3804</td><td>94.8%</td></tr><tr><td>1</td><td>Item selected</td><td>207</td><td>5.2%</td></tr><tr><td>99</td><td>Refusal</td><td>3</td><td>0.1%</td></tr><tr><td></td><td><b>Total</b></td><td><b>4014</b></td><td><b>100%</b></td></tr></table></td><td>National Wellbeing Survey, United States, 2021: DS1 - Public-Use Data</td><td>2021</td><td>Noninstitutionalized adults in the United States who ranged in age from 18 to 64 years old and who were able to read English.</td></tr>
<tr><td>☑ EMP2_3</td><td>Why are you not working at this time? (Select all that apply.) Student</td><td>Next, we'd like to ask some questions about your employment. [IF EMP = 4] EMP2. Why are you not working at this time? (select all that apply) 3. Student</td><td><table><tr><th>Value</th><th>Label</th><th>Unweighted Frequency</th><th>%</th></tr><tr><td>-99</td><td>Valid Non-Response</td><td>635</td><td>17.4%</td></tr><tr><td>1</td><td>Student</td><td>50</td><td>1.4%</td></tr><tr><td></td><td><b>Missing Values</b></td><td></td><td></td></tr><tr><td>-98</td><td>Refusal</td><td>2</td><td>0.1%</td></tr><tr><td>-97</td><td>Not</td><td>2952</td><td>81.1%</td></tr></table></td><td>National Couples' Health and Time Study (NCHAT), United States, 2020-2021: DS1 - Public-Use NCHAT Main Respondent Survey Data</td><td>2020-01-01 -- 2021-01-01</td><td>Adults in the United States who ranged in age from 20-60 years old, who were married or cohabiting, and who were able to read English or Spanish.</td></tr>
<tr><td>☑ STUD_PT</td><td>Employment - Part-time student</td><td>Q39. What is your current employment status? (Select all that apply) 11. Part-time student</td><td><table><tr><th>Value</th><th>Label</th><th>Unweighted Frequency</th><th>%</th></tr><tr><td>0</td><td>Item not selected</td><td>3937</td><td>98.1%</td></tr><tr><td>1</td><td>Item selected</td><td>74</td><td>1.8%</td></tr><tr><td>99</td><td>Refusal</td><td>3</td><td>0.1%</td></tr><tr><td></td><td><b>Total</b></td><td><b>4014</b></td><td><b>100%</b></td></tr></table></td><td>National Wellbeing Survey, United States, 2021: DS1 - Public-Use Data</td><td>2021</td><td>Noninstitutionalized adults in the United States who ranged in age from 18 to 64 years old and who were able to read English.</td></tr>
</table>

Figure 2 View of three variables using the 'Compare Variables' tool in the COVID measures archive. Users can see that the three variables being compared are comparable but not identical.

Interoperability with ICPSR Tools

A final important aspect of the COVID measures archive is that it harnesses the technical capabilities and features that ICPSR employs for making data findable, accessible, interoperable, and reusable (FAIR). One key ICPSR feature leveraged by the COVID measures archive is the social science variables database. The SSVD combines all variables within studies held and curated by ICPSR and makes them searchable within one large cross-study database. This means that COVID measures can be located not only when searching the COVID measures archive directly, but also when searching and exploring across ICPSR. This further improves the findability and interoperability of metadata from restricted data sources beyond the COVID measures archive itself and into ICPSR's broader data holdings.

ICPSR tools also enhance the findability of restricted metadata in other ways. Not just variables, but also studies can be found in a general search of ICPSR's data holdings. Study and variable descriptions are encoded using the Data Documentation Initiative (DDI) standard. Digital object identifiers (DOIs) uniquely identify each study. These features extend the visibility of restricted COVID data outside the SBE COVID Consortium, benefit other health researchers outside the consortium, and enhance the FAIRness of social, behavioral, and economic research on the COVID pandemic.

Conclusion

The COVID measures archive offers a way of balancing the challenges of data sharing – such as the requirement to protect participant privacy or safeguard proprietary data – with the benefits of openness and transparency. It enables sharing of detailed information about variables and measures without requiring the sharing of restricted or proprietary data. This helps our consortium members design interoperable studies, foster trust in research, and disseminate novel measures produced in their own research studies. These benefits make the COVID measures archive a useful tool for both the SBE COVID Consortium and the health research community more broadly.

Acknowledgements

This brief report was supported by grant U24AG076462 from the National Institute on Aging and the Office of Behavioral and Social Science Research at the National Institutes of Health.

References

Garcia, K. S., & Cowan, B. (2022). *The Impact of U.S. School Closures on Labor Market Outcomes during the COVID-19 Pandemic* (w29641; p. w29641). National Bureau of Economic Research. https://doi.org/10.3386/w29641

Hansen, B., Sabia, J., & Schaller, J. (2022). *Schools, Job Flexibility, and Married Women's Labor Supply* (w29660; p. w29660). National Bureau of Economic Research. https://doi.org/10.3386/w29660

Maas, S. (2022, April). Pandemic school closures and parents' labor supply. *NBER Digest*, 4.