AcademyHealth

April 15, 2025

Mehmet Oz, M.D.
Administrator
Centers for Medicare and Medicaid Services
7500 Security Boulevard
Baltimore, Maryland 21244

RE: Research Data Request & Access Policy Changes Feedback Opportunity

Dear Administrator Oz:

As the professional home for health services and systems researchers, AcademyHealth is pleased to offer input to guide the Centers for Medicare and Medicaid Services' (CMS) updated research data request and access policies, processes, and tools.

Health services researchers examine healthcare accessibility, costs, and outcomes, and AcademyHealth members regularly use CMS data in their research. We additionally host the Medicaid Data Learning Network, the State-University Partnership Learning Network, the Medicaid Medical Director Network, and the Medicaid Outcomes Distributed Research Network. We represent many of the top researchers and a large proportion of users of CMS Research Identifiable File (RIF) data.

Our membership, our researchers, and our peer organizations continue to be gravely concerned that the CMS proposal will have a profoundly negative impact on Medicaid and Medicare beneficiaries' access to high-quality and evidence-informed care, threaten the infrastructure of public health and health systems research, create significant barriers to lesser-funded organizations and individual researchers, and stifle the crucial advancements in health care research, especially for junior and future scholars of Medicaid and Medicare research.

We appreciate that CMS is working with stakeholders to design a policy and urge the agency to ensure it balances the critical importance of data security and data access by Medicaid and Medicare researchers, administrators, providers, patients, and health systems. As we noted in our previous letter to your office, these interests need not be mutually exclusive and should not be approached as in conflict.

Below we provide feedback on the updated questions based primarily on responses gathered from our membership and our Medicaid Data Learning Network, but that we believe are consistent with the perception of our broader membership.

1666 K Street NW
Suite 1100
Washington, DC 20006
202 292 6700 PHONE
www.academyhealth.org

CCW VRDC Technical Concerns

We have serious concerns about the CCW VRDC's technical capabilities. It is unlikely that the system would be able to support researchers' current and future CMS projects due to cost and system capacity limitations. It is important to note that without current access to the VRDC, which many researchers do not have due to the current paywall, it is impossible to fully understand the challenges of using this system.

Does your current research utilize any additional data source(s) other than CMS data?

CMS data is often used cross-sectionally, combining multiple datasets. It is critical that researchers continue to be able to combine datasets, including individually identifiable data efficiently to answer important questions to improve health and inform policymakers.

(a) Are the data individually identifiable?

Individually identifiable datasets such as electronic health data, other claims data, birth/death records are being utilized with CMS data. Much of this cross-sectional work with individually identifiable data is federally funded and would not be able to continue using the CCW VRDC. Reasons stated include data limitations in the VRDC, conflicting coding languages, and permitting/privacy constraints.

(c) What is the approximate volume (in GB or TB, as appropriate) of the additional data?

When asked about the approximate volume of additional data, an AcademyHealth member reported their research team that uses approximately five files of additional data across multiple years per year. Each additional file is around 50MB. This equates to about 2GB (50MB * 5 years * 8 years) of storage space being needed. Other researchers report linking CMS data with individual clinical data that require TBs of storage.

(d) What file format(s) are the additional data files, and would you be compressing them prior to uploading to the CCW VRDC?

Reported incorporated file formats include but are not limited to csv, rds, nc, shp, and h5. Some of these must be compressed.

Researchers have requested additional programming languages in the CCW VRDC. Please help us refine the list of potential options by providing a ranked list of languages (up to 5) that you'd like to see available.

We firmly believe that researchers should be able to use the language they prefer and are most familiar with in order to have the agility to answer the toughest questions in our health system in an effective and efficient manner. Given that SAS has been the primary language

used to access CMS data for decades, many CMS researchers are most familiar with that language. AcademyHealth members also reported using programming languages such as R, Python, and SAS. The current fee model is set up to require users to purchase the more expensive "Full VRDC" option to use R in addition to compute resources for processing, which would be cost prohibitive. The VRDC would need to keep up with the programming advancements and ensure the fee structure doesn't discourage innovation by limiting languages.

Are there other additions or improvements to the CCW VRDC beyond to the new analytic tools and programming languages that you'd like CMS to explore?

We have serious concerns about the analytic container sizes compared to the size of the dataset. The intention seems to be that users develop and test code in the analytic container and then submit processing jobs to Databricks. It's unclear if the data are structured in a way that even a sample can be loaded into an analytic container for code development, without exceeding the RAM limitations.

In addition, researchers request shared drives to facilitate collaboration across distinct projects/DUAs; this would allow us to maintain our collaborative approach to research that often involves sharing code and auxiliary datasets across studies.

Are there capabilities —currently available or not—that you would like to have as a la carte purchasing options (e.g., paying for additional compute resources)? Currently researchers can purchase additional Databricks credits, additional storage space, additional output reviews, or access to an analytic container.

Yes, we have several recommendations:

- Larger analytic containers than the current offerings.
- Flexibility to increase the size of an existing analytic container dynamically, instead of at seat fee renewal.
- Flexibility to purchase additional Databricks credits as needed throughout the project year, not just at project renewal or once during the project year.

Researchers have commented on the data architecture within the CCW VRDC. Do you have concerns about the way the CCW VRDC stores data? If yes, please provide clear and detailed feedback about any specific issues as well as what changes, if any, you believe we should consider.

Researchers have invested substantial time and financial investments into their current processing systems that would no longer be efficient in the new system. The computing resources available in the analytic containers seem to be tailored to working in SAS, making it difficult to work in other programming languages that rely on RAM for data management without partitioning the data into smaller pieces, increasing available memory, or both. Additionally, it would be prohibitively expensive to obtain enough parallel

sessions to make parallel processing effective, and that the data aren't currently structured to efficiently access samples.

Data Access Costs Concerns

AcademyHealth is deeply concerned about the additional financial burden that the proposed changes to the CMS research data request and access policies, processes, and tools. In addition to the increased cost associated with the VRDC in general, many research teams would need to implement additional costly changes (e.g. changes to operating systems and/or programming languages) in order to continue to access CMS data. The cost is exceptionally corrosive to the building of talent pipelines and the training of early career researchers.

For example, one AcademyHealth member's research team calculated the differences in costs from the current system to the proposed VRDC model. Under the current system, physical TAF data files at an organization-level is purchased at the cost of approximately $70,000 per year, which is shared across multiple projects via reuse agreements. With six ongoing TAF-related projects, the cost per project over four years is approximately $50,000 ([$70,000 × 4 years] ÷ 6 projects = $47,000 per project). Additionally, each study pays a $2,000 re-use fee under the current model, keeping overall data access costs manageable. By contrast, under the VRDC model, each research project would require individual seat fees, computing resources, and storage costs, dramatically increasing expenses. Based on the typical research team structure—a PI, 1-3 analysts, and part-time data engineering support—it is estimated that a single 4-year project in the VRDC would cost approximately $415,400. This represents an increase of over $350,000 per project compared to the current approach. This does not account for the training costs associated with the new system. With the additional costs, the transition to the CCW VRDC would not be financially viable for this center.

CMS charges fees to researchers to recoup the cost of making data available for research purposes[1]. These fees are based on CMS's costs and allow the agency to continue to make data available to researchers. Historically, CMS has infrequently reevaluated our pricing, leading to large pricing increases when changes are made. Would you rather: a) continue these larger, infrequent cost increases on a multiyear timeline, or b) switch to smaller annual increases?

It would be preferred to have smaller annual increases that could be forecasted so that researchers could appropriately budget for grant purposes.

If CMS were to explore the creation of new CCW VRDC seat options, what would you need included in a lower cost, entry-level seat? Assume CMS would also offer alternative seat types which include all the existing tools and capabilities.

There is a need for lower cost seats. For quality results, multiple scientists must be involved in the coding and review process. To ensure access for entry level scientists such as graduate students, there should be a significantly lower cost seat available.

Would the development of a "viewer" seat in the CCW VRDC be beneficial for you or someone on your team?

We strongly support the development of a "viewer seat". We envision this would be a very low-cost seat (non-transferrable) which could have access to CMS data under multiple data use agreements (DUAs) and primarily be used by those who supervise student work or oversee research teams.

(a) What features would be necessary for you in this type of seat?

It would be necessary for this seat to have the ability to open files and view aggregated tables as well as claim-line level records. The seat should have read privileges but would not require write privileges as the user with this seat would not run code.

(b) Are there any features that would prevent your ability to use this seat type?

Cost could be a limiting factor.

Timeline Concerns

Research is an ongoing and timely process that requires significant programmatic and fiscal planning. Uncertainty in the ability to execute a grant as intended due to policy changes is highly corrosive to the research environment.

What is your average grant funding timeline, from initial proposal development through funding decision and award of funds? Please provide a timeline for each phase, if possible. The following timeline assumes the proposal is not funded upon the first submission (which is common), and a resubmission is required.

Researchers are often working on five-year timelines, and as a result need to be able to plan a 5 year budget. The total funding timeline from initial development to award of funds is 19-36 months. Lengthy timelines for seat changes could impact the overall grant timeline, and the need to address analytical problems in a "revise & resubmit" add additional time into the amount of time a seat must be maintained.

20. When developing your proposal budget do you typically use the Estimate Study Size or the Data Pricing tools found on the CCW website (https://www2.ccwdata.org/web/guest/pricing)?

No, researchers reported referring to previous literature to estimate study sizes.

Conclusion

As it is currently proposed, the shift to the CCW VRDC system would pose significant challenges for health services researchers and the ability of Medicaid and health system policymakers to have access to cutting-edge data to inform policymaking. Not only is the system prohibitively costly, but it would also require researchers to change tried and true methods to conduct effective research to work with the CCW VRDC, which has a strong risk of limiting the types and rigorousness of analysis. AcademyHealth urges CMS to consider the lasting impacts that this change could have on CMS research as a whole, and continue to partner with researchers to ensure the new system meets the needs of those who use it.

Thank you for the opportunity to discuss the perspectives and concerns of the health services research community. For further comment, clarification, or inquiry, please email Josh Caplan at Josh.Caplan@AcademyHealth.org.