Chronic Conditions Warehouse

Your source for national CMS Medicare and Medicaid research data

Chronic Conditions Warehouse Virtual Research Data Center

Data Output Review Process

OCTOBER 2023 | VERSION 2.3

This page intentionally left blank.

Revision Log

Date	Revisions	Version
October 2023	Refined file type text	2.3
March 2023	Refined file type text	2.2
August 2022	Updated output review process	2.1
March 2022	Updated output review process under the 2022 CCW VRDC Cost Pricing Model Plan	2.0
July 2021	Updated help email addresses with @gdit.com to @ccwdata.org	1.94
December 2020	Template update; added clarification regarding outputting zip codes and allowance of minimum, median, maximum, and percentiles; OEDA policy updates; passive to active voice	1.93
March 2020	Added sentence regarding download size; added verbiage and a FAQ regarding output of zip codes	1.92
December 2019	Updated output limits per CMS, format changes	1.91
June 2019	Updated document to reflect CMS feedback	1.9
April 2019	Updated document to reflect CMS feedback	1.8
March 2019	Updated document to reflect CMS enhanced compliance of output reviews	1.7
December 2016	Updated Part D verbiage and additional FAQs	1.6
September 2016	Updated GDIT to HealthAPT	1.5
June 2016	Updated SFTS link	1.4
February 2016	Updated document	1.3
November 2015	Added FAQ section	1.2
October 2015	Updated document	1.1
February 2015	Created initial document	1.0

Table of Contents

List of Figures

1.0 Introduction

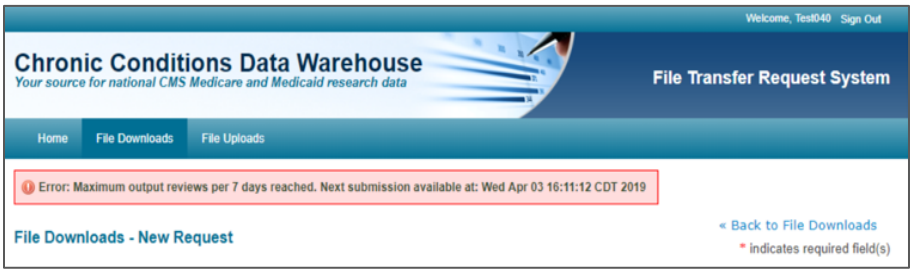
The Chronic Conditions Warehouse (CCW) Virtual Research Data Center (VRDC) output review process exists to help researchers protect Medicare and Medicaid beneficiaries' confidentiality. The purpose of the output review process is to help researchers avoid accidental disclosure or the perceived disclosure of confidential information. The CCW analytical team reviews all output requested for download from the VRDC, and ensures it meets all disclosure checks before allowing the researcher to download it.

While the CCW team conducts a review of all CCW VRDC output, it is ultimately the researcher's responsibility to ensure any output downloaded from the CCW VRDC environment is compliant with the Centers for Medicare & Medicaid Services' (CMS) privacy policies as stated in the Data Use Agreement (DUA). Researchers must review output themselves and should exercise caution when publicly disclosing research findings. In addition, researchers who frequently request to download output that is not compliant with CMS's policies may have their CCW VRDC access suspended or terminated for violation of their DUA.

2.0 Output Review Policies

This section contains specifics of the output review process.

1. **Researcher:** The CCW VRDC File Transfer Request System (FTRS) allows each DUA three output reviews per week (rolling seven-days) shared by all researchers on the DUA. Researchers may submit multiple files in one review. The total size of files cannot exceed 1 gigabyte (GB) per week. The FTRS allows one resubmission of a rejected output review per week, if applicable. The screen notification below is an example of what researchers see should they exceed the permitted output review limit. Individual files within the request need to be 1 GB or less.



2. **Innovator:** The CCW VRDC FTRS allows each DUA six output reviews per week (rolling seven-days) shared by all users on the DUA. Researchers may submit multiple files in one review. The total size of files cannot exceed 2 GB per week. The FTRS allows one resubmission of a rejected output review per week, if applicable. The screen notification above is an example of what researchers see if they exceed the permitted output review limit. Individual files within the request need to be 1 GB or less.

3. Based on the Cost Pricing Model Plan and CMS approval, researchers may have the option to purchase additional output reviews. The FTRS still limits the download size to 1 GB per file; therefore, individual files within the request need to be 1 GB or less before requesting the output review. Output review submission restrictions are per DUA, not per seat.

4. The CCW output review analyst rejects all output containing fields representing small cell size (N < 11) to ensure beneficiary or patient information privacy. **NOTE:** Zero cell frequencies/counts are acceptable to output.

5. If the cell has a size of 11 or greater (N > = 11), it is generally allowable.

- Researchers must suppress all cells related to beneficiary or patient information with a frequency < 11. Before submitting, researchers should consider recategorizing the variables instead of removing the observations completely. If this is not possible, identify and suppress all cells with frequencies < 11 before resubmitting the output.

6. Researchers must remove personal identifiers, as well as:
   - Remove name, address, Social Security number (SSN), date of birth (DOB), death date, health insurance claim (HIC), admission date, discharge date, beneficiary identifier (Bene ID), Tax Identification Number (TIN), etc.
   - Avoid beneficiary dates of care
   - Use age categories or age ranges, not specific ages, or percentiles
   - Treat contextual variables describing an area with the same caution as geographic indicators
   - Adhere to the Health Insurance Portability and Accountability Act (HIPAA) Safe Harbor provision, prohibiting disclosing geographic information at and below the zip-code level. The CCW output review analyst:
     - Rejects beneficiary-level data with zip code information
     - Rejects output with beneficiary census tract or beneficiary latitude and longitude location variables
     - Approves both provider- and facility-level location zip codes for export when clearly labeled "provider zip codes"
     - Permits beneficiary counts with provider- and facility-level zips to export

     NOTE: Researchers must specify the source (provider or beneficiary) when submitting output containing zip codes. The CCW Help staff contacts researchers if they do not specify the source of the zip codes in the request.

   - Provider identifiers associated with any beneficiary or patient information with small cell sizes (N < 11) are not allowable
   - Unique to Part D data, researchers cannot include in their output for download both the CCW prescriber ID and the unencrypted prescriber ID together. CCW has licensing agreements with several data vendors and releasing the combination of data elements noted above would violate licensing agreements. As a result, researchers should remove one or the other from the output

7. Health information, including counts of beneficiaries identified by diagnosis and cause of death codes, must meet minimum sample size criteria for release.

8. CMS policies prohibit individual values such as extreme observations (e.g., five smallest or five greatest values in a distribution of data). An extreme observation is a sample of size N = 1.
   - The CCW output review policy prohibits extreme values in the output, even though it may not be linkable to a specific beneficiary. When the study sample size is large, minimum, maximum, median, mode, and percentiles *may* be approvable
   - An example of adequate sample size may be greater or equal to 500 observations with no individual strata less than 50 observations. However, the CCW output review analyst approves this type of output on a case-by-case basis
   - Determination of adequate sample sizes for allowing output of this type is at the discretion of the CCW output review analyst. **NOTE:** Extreme observations are never approvable, even when researchers base the observations on a large sample size

9. Researchers are strongly discouraged from submitting zip files as this will increase output review time and the chance of rejection. If the CCW output review analyst rejects one document in the zip file, the analyst must reject the entire zip file. The CCW output review analyst automatically rejects zip folders within folders.

10. A CCW output review analyst must review results, findings, or output before download. Once the CCW analyst has reviewed the output, the researcher receives an email notification that the file is available for download using the Secure File Transfer System (SFTS).

11. The CCW output review analyst reviews output submissions within two business days unless the request requires a more extensive review. Particularly complicated output, multiple files, or output that is a PDF or Word document over 100 pages require additional time.

12. Researchers may submit multiple files in one review, but complex or extensive reviews may require more than the standard two-business day turnaround as noted in the previous policy.

13. Researchers should avoid intermediate or draft output (e.g., tables of preliminary descriptive statistics, large numbers, or regression models).

14. Researchers should not use the output review as a method for backing up their code, data, or files.

15. The CCW output review analyst does not approve output results that pose a disclosure risk. If that is the case, the CCW analyst may ask the researcher to collapse categories to increase cell counts. Researchers may also need to suppress other cells or values to maintain confidentiality.

16. The output should not include any method that would facilitate the re-identification of an individual. This is a disclosure violation.

    Before submitting output for review, the researcher should perform a self-review of the output:
    - Remove individual-level data from any output; if there is any doubt, remove
    - Check that summarized SAS datasets meet the CCW VRDC data review requirements and criteria
    - Consider recategorizing variables of all cells related to beneficiary or patient information with a frequency less than 11. Identify and suppress all cells with frequencies less than 11 before submitting the tables for output review
    - Remove extreme values. Remove proc univariate extreme observations (1, 99, and 100 percentiles)

17. List of accepted file types.

    - For fastest output review turnaround time:
      - csv (*.csv), Excel (*.xlsx, *.xls, *.xlsb), SAS data sets (*.sas7bdat), or text formats (*.doc, *.docx, *.rtf, *.pdf, *.lst, *.txt) *.stata, and *.tsv
    - May increase output review time and the potential for questions from the CCW output review analyst:
      - *.dat, *.do, *.dta, *.fts, *.gph, *.htm, *.html, *.jpeg, *.png, *.pptx, *.pqt, *.tab, and *.zip

    - Please submit data files in formats (such as csv, Excel, or SAS). Submit text or figures in text formats

    Researchers who need the SAS code and logs, should submit with a SAS program interface, or copy into a text file format. Researchers should submit Databricks (R, Stata, Python) Notebook files as an Excel file or convert to SAS for large files. Random forest binary model files must convert to SAS or another accepted file type.

18. Helpful file naming conventions aid the review speed and include identifying the output content (e.g., counts of x) and the sampled subjects. For example, Beneficiaries_diabetes_summary_count. Within the output, label columns, fields, and variables clearly and intuitively.

19. Scrub the SAS code and associated logs of all Protected Health Information (PHI) or Personally Identifiable Information (PII).

20. CCW output review analyst rejects all output containing fields representing sample sizes or counts related to beneficiary or patient information if less than 11. The CCW analyst performing the output review looks for fields labeled “N,” “Count,” “Freq,” “Nobs,” and other labels indicating a sample size. To avoid possible rejection and output review efficiency, clearly label all values with a description and label all columns before submission.

    It is the researcher’s responsibility to ensure output does not pose, or appear to pose, a disclosure risk. The researcher risks rejection if submitting output with small cell sizes coupled with vaguely labeled variables (even if those fields do not represent counts of beneficiaries or procedures).

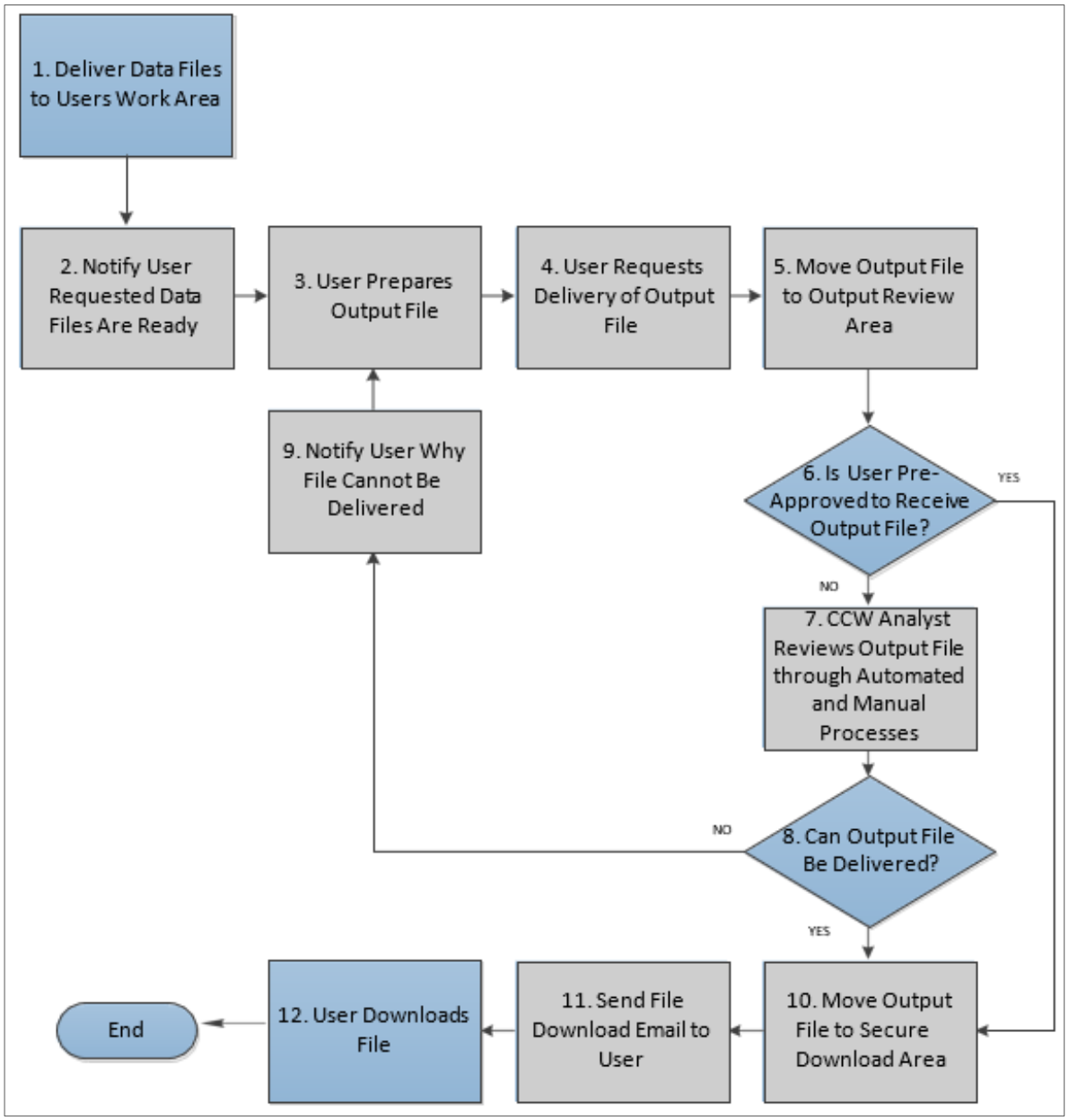
NOTE: Output containing zero cell frequencies/counts are acceptable to output.

3.0 Data Output Review Process Flow

Figure 1 shows the CCW VRDC Process Flow. Assuming the files are ready for output review, begin by accessing the CCW VRDC virtual desktop and CCW FTRS at https://www2.ccwdata.org/tools/ to log in and submit the request, as in step four.

Beginning at step seven, a CCW analyst reviews the output results before the researcher may download the files. Researchers receive notification if there is unacceptable output. If so, revise any rejected files and repeat the process. If the CCW analyst approves the output results, the researcher receives an email advising the files are ready for download using the CCW SFTS at https://sfts.ccwdata.org/.

Figure 1. CCW VRDC process flow — data output review process



4.0 Output Review with SAS

The CCW output review analyst may use SAS code to detect potential PHI/PII/small cell size violations for large SAS-importable files impractical for them to review manually. A CCW analyst manually reviews everything flagged by this code to confirm a violation. The SAS code scans for the following:

- Identifier field names, such as SSN, HIC, DOB, etc., to detect their possible existence
- Identifier field formats such as nine digits for SSN and dates for DOB or date of service
- Cells with beneficiary or patient information count less than 11. **NOTE:** Output containing zero cell frequencies/counts are acceptable to output

For more information, review the Frequently Asked Questions section.

5.0 Output Review Checks

The output review process, as stated above in the Output Review Policies section, includes the following data checks:

- **Personal identifiers**
  - Removal of all personal identifiers, including name, SSN, HIC, BENE ID, DOB, TIN, and address
  - Ensure there are no beneficiary dates of care
  - Ensure no single row of data represents a single beneficiary
  - Confirm using age categories, not age range and percentiles
  - Treat contextual variables describing an area with the same caution as geographic indicators

- **Provider identifiers**
  - Generally, researchers may output provider identifiers or provider summary statistics. An exception is when provider identifiers or provider summary statistics are associated with any beneficiary or patient information with small cell sizes (N < 11)
  - Another exception is a special requirement unique to Part D data. In this case, researchers cannot take out the CCW prescriber ID and the unencrypted prescriber ID together. CCW has licensing agreements with several data vendors and releasing the combination of data elements noted above would violate licensing agreements. As a result, the CCW output review analyst removes one or the other in the output
  - A third exception is provider TINs. Researchers cannot output TINs
  - HIPAA Safe Harbor provision prohibits disclosing geographic information at and below the zip-code level. The CCW output review analyst rejects beneficiary-level data with zip-code information
    - CCW review team can approve both provider- and facility-level location zip codes for export when clearly labeled “provider zip codes”
    - The CMS policy allows beneficiary counts with provider- and facility-level zips to export
    - Researcher must specify the source (provider or beneficiary) when submitting output containing zip codes. The CCW Help Desk staff contacts researchers if they do not specify the source of the zip codes in the request
    - The CCW review team does not attempt statistical perturbation as a disclosure limitation technique. The identifiable output needs adjusting
  - If the CCW analyst’s review reports differences, there must be large enough samples for each group included in the difference (e.g., cannot report a difference when only one group has a large enough sample to report count)

- **Health Information**
  - Counts of beneficiaries identified by diagnosis and cause of death codes must meet minimum sample size criteria for release

6.0 Frequently Asked Questions

Q: Now that the output reviews are at the DUA level, can one person use all the weekly output?

A: Yes. Researchers should collaborate with others on the DUA.

Q: When undergoing the CCW output review, why might the CCW output review analyst reject my output request?

A: The CCW VRDC output review process exists to help researchers protect Medicare and Medicaid beneficiaries' confidentiality. The purpose of the review process is to help researchers avoid disclosure or the perceived disclosure of confidential information. The CCW analytical team reviews all output and ensures it meets all disclosure checks before transferring it to the researcher for download. Researchers cannot export individual-level data.

While the CCW team conducts a review of all CCW VRDC output, it is ultimately the researcher's responsibility to ensure any output removed from the CCW VRDC environment is compliant with the CMS privacy policies as stated in the DUA and as stated within this document. Researchers must review the output and exercise caution when publicly disclosing research findings.

The CCW VRDC output review process minimizes the risk of identifying a person, patient, or beneficiary. The three most common causes for rejection are:

1. **Personal identifiers:** Name, address, SSN, DOB, death date, HIC, admission date, discharge date, beneficiary identifier (bene_id), TIN, etc.

2. **Small cell size (N < 11):** Output review looks for counts (N, frequency, count, quantity, etc.) of persons/patients/beneficiaries/claims or other cells that may be a people count or procedures (number of visits, stays, transfers, etc.).

    - If the cell has a size of 11 or greater (N >= 11), the CCW team generally allows it
    - **NOTE:** Zero cell frequencies/counts are acceptable to output

3. **Individual values:** Rules prohibit individual values such as extreme observations (e.g., five smallest or five greatest values in a distribution of data). An extreme observation is a sample of size N = 1. Accordingly, no single row of data may represent a single beneficiary.

    - CMS policy prohibits extreme values in the output, even though it may not be linkable to a specific beneficiary. When the study sample size is large, minimum, maximum, median, mode, and percentiles may be approvable
    - An example of adequate sample size may be greater or equal to 500 observations with no individual strata less than 50 observations. However, the CCW analyst approves this type of output on a case-by-case basis
    - Determination of adequate sample sizes for allowing output of this type is at the discretion of the CCW analytic review team. **NOTE:** Extreme observations are never approvable, no matter how large the researcher's sample size

Q: When undergoing CCW output review, will the CCW output review analyst approve output where specific cells are null?

A: Yes. If researchers suppress cells with small cell sizes somehow, AND the other cells cannot back-calculate values based on the other values on the table, then the CCW analyst allows it.

- E.g., if the researcher suppresses a cell, but there is a column total or row total that allows one to back-calculate what the small cell size is, then the CCW analyst rejects that output.
  - For example, if there are five cells in a row (20, 16, 24, *, 15) with a row total of 80, back-calculation: 80 - 20 -16 - 24 - 15 = 5; thus * = 5, and the CCW analytic review team rejects this output
- However, if there are two suppressed cells and back-calculation cannot produce exact values for the suppressed cells, this is allowable
  - For example, if there are five cells (20, *, 15, *, 15) with a row total of 60, back-calculation: 60 - 20 - 15 - 15 = 10; thus * + * = 10, but it is not known if the cells are 1 and 9, or 2 and 8, etc., this is allowable
- CCW analyst also rejects output where small cell sizes may be back calculated from rates, ratios, or prevalence

Q: Will the CCW output review analyst allow minimum, median, and maximum values when undergoing CCW output review?

A: The CCW analyst reviews minimum, median, and maximum values on a case-by-case basis and may or may approve the output depending on the values. These values may represent metrics for an individual beneficiary, particularly with small sizes and rare outcomes. While the values derive from CCW files and not directly available in the files, the values are at a beneficiary level. The minimum and maximum values may indicate values for a specific beneficiary.

Q: Will the CCW output review analyst reject all field values less than 11 (< 11) when undergoing CCW output review?

A: No. The CCW analyst rejects only fields representing sample size, counts, claims, frequencies of the beneficiary, or patient information data if N < 11.

- The CCW analyst performing CCW output review looks for fields labeled "N," "Count," "Freq," "Nobs," and other labels indicating a sample size
- To avoid possible rejection and output review efficiency, clearly label all values with a description and label all columns before submitting the output review. **NOTE:** Zero cell frequencies/counts are acceptable to output

Q: When a file undergoes CCW output review, will the CCW output review analyst reject a request if the cell size is less than 11 (as in minimum, median, maximum) for fields that are NOT identifiers and where the results cannot identify a beneficiary?

A: This depends on the sample size. The CCW output review policies do not allow any "individual values," which may include minimum, maximum, median, quartiles, quantiles, extreme observations, etc. However, when the study sample size is large, minimum, maximum, median, mode, and percentiles may be approvable. This is because it is unlikely a common value belongs to any single beneficiary in a large study. The CCW analyst approves this type of output on a case-by-case basis.

Determination of adequate sample sizes for allowing output of this type is at the discretion of the CCW analyst reviewing the output. Note that extreme observations are never approvable, even when researchers base the observations on a large sample size. Researchers may need to aggregate or combine data to remove small cell sizes.

- For example, if a researcher could identify beneficiaries when cost and utilization metrics show a low dollar amount or day count, indicating a small number of beneficiaries in the area (especially if there is only one). The CCW analyst rejects the request and asks researchers to suppress this data
- To avoid possible rejection and help the CCW analyst, clearly label all values with a description and label all columns before submitting the output review. If researchers are unsure whether the sample size is large

enough to allow output minimum, maximum, median, mode, and percentiles, reach out to ccwhelp@ccwdata.org

- Also, researchers who frequently request to download output that is not compliant with CMS's policies may have their CCW VRDC access suspended or terminated for their DUA violation

Q: May I output patient-level zip codes?

A: HIPAA Safe Harbor provision prohibits disclosing geographic information at and below the zip-code level. The CCW analysts reject beneficiary-level data with zip code information:

- The CCW analysts can approve both provider- and facility-level location zip codes for export when clearly labeled "provider zip codes"
- The CCW team allows beneficiary counts with provider- and facility-level zips to export
- Researchers must specify the source (provider or beneficiary) when submitting output containing zip codes. If researchers do not specify the zip codes' source in the request, CCW Help Desk staff contacts the researcher. The CCW analyst rejects any output containing beneficiary-level data with zip code information
- Researchers cannot output beneficiary census tract or beneficiary latitude and longitude location variables

Q: Will the CCW analyst reject all my files if I submit multiple files in one submission and only one file poses a disclosure risk?

A: No. The rejected file(s) are only those that pose a disclosure risk. The CCW analyst approves the other files.

Q: Are output reviews for low cell counts performed by an automated algorithm, by an individual, or both?

A: A CCW output review analyst performs all reviews for low cell counts. Some reviews may employ SAS code or Excel functions/macros (on a case-by-case basis), but they manually review as well. That is why explicit and intuitive labeling of fields/columns should be a best practice, and this practice facilitates the expeditious approval of the review. The CCW analyst rejects the output if they cannot determine what a field or variable represents, and it contains a small cell size. With this rejection, the CCW analyst requests a small cell redaction or additional information and resubmission. It is both the CCW analyst AND the researcher's responsibility to ensure the output does not pose an unacceptable identification risk.

Q: Will I always have my output approved (or rejected) within two business days?

A: The normal response time for output review is within two business days (excludes weekends and federal holidays). Exceptions to this may be when the CCW analyst is incurring high request volumes or when complicated files need review and/or the output is a .doc or .pdf file type that is more than 100 pages. A good rule of thumb when submitting .doc or .pdf file is that if it is something the researcher would not print, it is probably something the researcher should not submit as a .doc or .pdf. Consider exporting from SAS/STATA to .csv or .xlsx file type for a more efficient output review.

Q: What if I need multiple files reviewed during the seven-day week?

A: If the researcher needs multiple files or anticipates needing multiple files within the rolling seven-day week, the researcher should wait until they are ready to submit all the files for review before submitting any. Researchers may need to allow the files to accumulate to submit all files as a single review. System validation is in place to limit output for researchers and innovators per week. Thoughtfully consider what files the researcher needs outside the CCW VRDC environment and which ones the researcher does not. Using the single review to submit draft

documents or non-quality-controlled documents is not advisable. Using the review to output all SAS code, SAS logs, or everything the researcher has worked on in the CCW VRDC environment is not advisable.

Q: What if I submit a file accidentally and would like it recalled, so it does not count against my allowed reviews?

A: If researchers submit a file inadvertently, email ccwhelp@ccwdata.org immediately for help with the file. **NOTE:** The CCW Help Desk staff cannot cancel the output request. If researchers contact the CCW Help Desk early, depending on the timing, the CCW team may be able to reject the submitted review; however, it counts as a rejected review.

Q: What if a CCW analyst rejects my single submission because of an analyst or technical error?

A: CCW Help Desk staff helps researchers resubmit an erroneous file rejection. Email ccwhelp@ccwdata.org, including the Review ID and the DUA number, as well as the circumstances of the rejection.

Q: What if I'm not sure my file contains prohibited information and I want to know before submitting a request?

A: First, read this document and FAQs thoroughly. After that, if the researcher is still unsure, email ccwhelp@ccwdata.org before the researcher submits the file. Describe the file contents (DO NOT INCLUDE any PHI and/or PII in the description or any emails). CCW Help Desk staff forwards the information to the CCW analytic review team, who emails the researcher with help.

7.0 Where to Get Assistance

The CCW Help Desk staff provides assistance between 8:00 AM to 5:00 PM ET, Monday through Friday (excluding most federal holidays). Contact the CCW Help Desk at ccwhelp@ccwdata.org or 1-866-766-1915.