# Deduplication Analysis Guidance

What is Deduplication?

- Overview:
  - PEPFAR is concerned about individual humans and deduplication allows us to better understand data as it relates to individuals. When two (or more) mechanisms reach the same beneficiaries at a site with the same services, this creates Duplication of reporting. Further detail on when Deduplication occurs can be found at the Deduplication Mechanism Overview page.
  - The deduplication process takes place at the Interagency level for the same data element, after all partner data has been entered.
  - Deduplication, often referred to as dedup, should be used any time an aggregation of data is being evaluated (e.g., Reviewing multiple agencies, multiple partners, etc.).

Deduplication Logic Example:
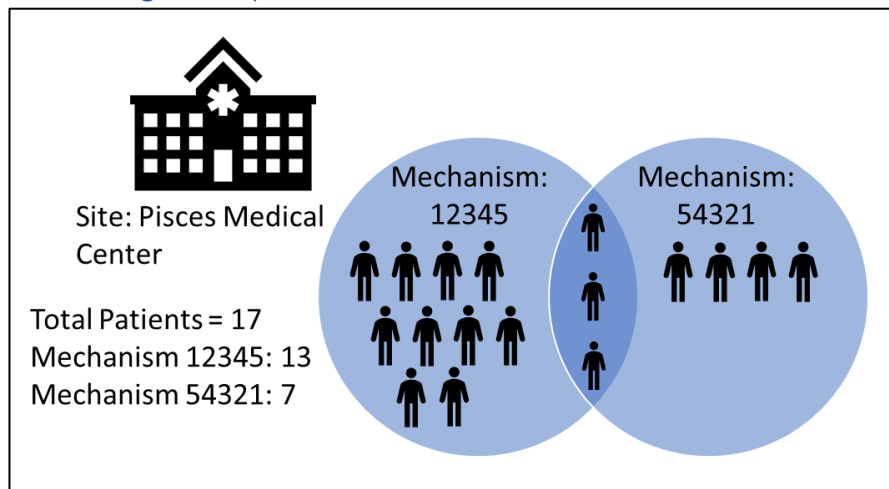
The above image illustrates the Deduplication Logic using the MSD as an example. In this example, two implementing partners provide the same service within the same community for the same group of people. As you can see, there are 17 total beneficiaries that sought care at Pisces Medical Center. Mechanism 12345 saw 13 beneficiaries, and Mechanism 554321 saw 7 beneficiaries; however, 3 of those beneficiaries were seen by both mechanisms. These would be considered duplicates.

The Process of Deduplication

- An in-depth review of the Deduplication Process can be found in the DATIM Deduplication Training Review document.
- Deduplication occurs at the agency level, after all partners report data.
  - Deduplication occurs at the data element level, meaning that only pieces of data with the same disaggregation can be deduplicated. E.g., Indicator/Disaggregate/Age/Sex must all match when applicable.

- Deduplication must also be completed within the Reporting Period for the quarter. Data can not be retroactively deduplicated.
- Types of Deduplication
  - Pure/Normal/Simple Deduplication: All DSD-DSD or TA-TA duplicate (or differing), meaning that deduplication of like-to-like services (e.g. direct service delivery or technical assistance is provided by two mechanisms to the same beneficiaries)
  - DSD-TA Crosswalk/Complex Deduplication: When multiple partners report DSD and TA results at the same site, which double-counts beneficiaries across DSD and TA results; revise TA value within DATIM.
- Deduplication occurs in the **DATIM Dedupe App**, which allows for a variety of Deduplication Resolutions. Examples of the deduplications types and their applications are available in the DATIM Deduplication Training Review Guide.
  - **Standard Duplication Resolutions**: Which logic should be applied to the values?
    - Sum: Add all IM Results together at site level (No duplication of clients by IMs). This is used when the services provided to beneficiaries can only be done once. An example of this would be VMMC_CIRC, where a beneficiary can only be circumcised once.
    - Max: Across all IMs, report the "highest reported value" of each indicator/disaggregation at the site level. (Full or Broad Duplication of Clients by IMs). This is used when sites have multiple partners providing a variety of services, some of which may overlap with other partner's services . An example of this would be if three partners were providing HIV services at a site. Partner A is only providing testing services, while Partner B is providing TB and HIV testing services, meanwhile Partner C is providing PMTCT services, including HIV testing. You would use "max" for all mechanisms, because Partner A is reporting the same individuals as Partner B and C (see Deduplication Training Review Guide for more details).
    - Custom: Across all IMs, report a "custom value" at the site level based on contextual or known programmatic factors. This is used in situations where a custom count is needed. For example, if two partners provide services to specific but potentially overlapping populations, such as when one partner provides ARTs specifically to pregnant women, and another partner is dedicated to providing services to Female Sex Workers (FSWs). Since FSWs may become pregnant, they may receive services from both partners, and a custom count would be entered to account for those who receive services from both partners.
  - **Crosswalk Deduplications for DSD/TA:** When partners reach the same clients, but with different types of service (DSD or TA). A cross walk of the services provided to each individual is done to confirm that they received different types of service. Usually this is resolved by selecting the "max" option in the DATIM Deduplication App.

Once deduplications are resolved in the DATIM Deduplication app, the results are calculated and reported as negative numbers (example: -352) in a separate "Dedup" mechanism row in the MSD, Panorama and other spaces that use the DATIM data. This allows for accurate calculation of the true number of individuals reached.

How to use Deduplication when Analyzing Data

Knowing when to use and when not to use dedup data is important to ensure accurate analysis. Below we will describe when you should use dedup in your analysis, and when you should exclude it from your analysis.

- How do Deduplications Appear in Different PEPFAR Analytic spaces? How can you Utilize Deduplication in your Analysis?
    - o Panorama

        In Panorama, you can utilize the *Dedup Mechanism* selector (under *IM + Partner*) to review SNU or Partner Level achievement. These mechanisms are automatically selected, to allow for OU or SNU level achievement to be reviewed with accurate information.

        - When the *Dedup* mechanism is selected, you are seeing the deduplicated results. When selected, results can be viewed at the OU, PSNU, Community, or Site level.
        - When this is not selected, results should only be viewed at the Agency or IM level.
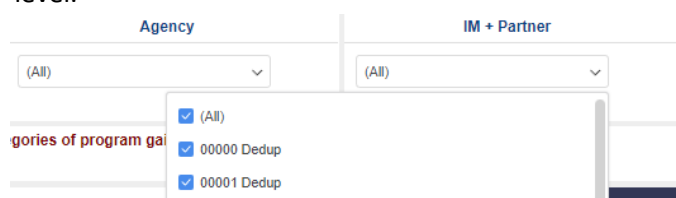
    - o MER Structured Data Sets

        The highlighted box below indicates the deduplication that occurs between Mechanism A and Mechanism B. The deduplication information is designated under the 'Mechanism' column, though this is not associated with any specific Implementing Partners.[1]

        You will notice that the deduplication value is negative. As was discussed above, this is to allow summing of the columns, where the negative duplication values will be removed from the total if a mechanism or partner is not selected.

Mechanism	Facility	Indicator	Indicator Type	disaggregate	Age	Sex	Targets	Approval Level Description
Mechanism A	Facility A	TX_CURR	DSD	Age/Sex/HIV Status	40-44	Male	98	Submitted by Inter-Agency
Mechanism B	Facility A	TX_CURR	DSD	Age/Sex/HIV Status	40-44	Male	98	Submitted by Inter-Agency
Dedup	Facility A	TX_CURR	DSD	Age/Sex/HIV Status	40-44	Male	-98	Exempt from workflow

---

[1] datim.zendesk.com/hc/en-us/articles/203850119-Deduplication-Mechanism-Overview

MSD – Mechanism Pivot Table

When viewing data at an OU, PSNU, Community, or Site level, be sure to include the Dedup mechanism in your pivot table to ensure accurate totals are being reflected, otherwise the total will look falsely inflated.

Mechanism Level Comparison

Row Labels	Column Labels
0	
Dedup	-676
Mechanism 12345	
Partner A	1,857
Mechanism 23456	
Partner B	1,356
Grand Total	**2,537**

MSD – PSNU Pivot Table

When viewing total data at a PSNU level, be sure to include the Dedup Mechanism as a filter or within your pivot table, to ensure that accurate totals are reflected.

In the example below, the dedup mechanism shows that PSNU A had 9,615 beneficiaries (sum of all mechs), of which 688 were duplicates, therefore there was a grand total of 8,927 (sum of all mechs – dedup), after removing duplicates. Similarly, PSNU B had 19,054 beneficiaries across all 5 mechs, of which 550 were duplicates, meaning that there was a grand total of 18,504 beneficiaries.

Notice that PSNU C has no value in the 'dedup' column. That means that there were no known duplicate beneficiaries across the 3 reporting mechanisms in that PSNU.

	Dedup	Mech 12345	Mech 23456	Mech 34567	Mech 45678	Mech 56789	Grand Total
PSNU A	-688	2,235	2,128	1,208	3,696	348	8,927
PSNU B	-550	5,595	4,230	961	7,986	282	18,504
PSNU C		2,226	1,590		3,192		7,008
Grand Total	**-1,238**	**10,791**	**8,208**	**2,169**	**15,802**	**630**	**36,362**

MSD – Agency/IM Pivot Table

Conducting agency specific analysis is perhaps the most common reason for excluding dedups. This is because agencies usually only have one mech at a site, and therefore would not have duplicate beneficiaries. Since dedup rows are not assigned to a specific partner, they would not appear in the data when the data is filtered for a specific agency.

Ensure that the Dedupe Option under Agency/Mechanism is deselected.

| TX_CURR			
	2020	**2021**	**2022**
HHS/CDC	353,334	414,788	482,846
USAID	8,301	12,563	15,780
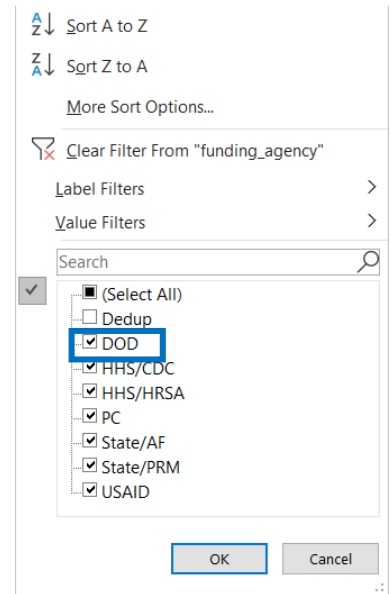
In the above example, agency totals can be viewed without the overall deduplication, to see agency contribution to MER results. As shown below, this data can be further refined to see Mechanism Level data without deduplication, to help identify Partner contribution to the program.

TX_CURR	Partner A	Partner B	Partner C	Partner D	Partner E
HHS/CDC		963,743			287,225
USAID	7,482		9,360	19,802	
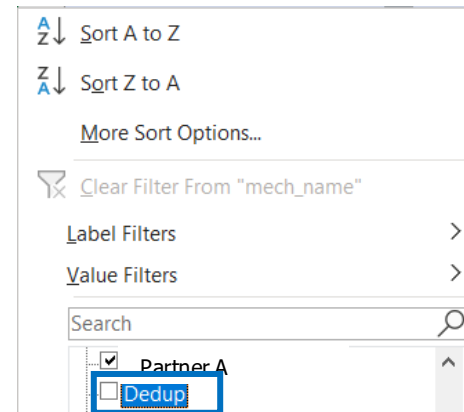
DATIM

When viewing Pivot Tables in DATIM, the deduplication is automatically included. To see how totals would change without deduplication (i.e., at the Agency or IM level), the De-Duplication Dimension can be selected.
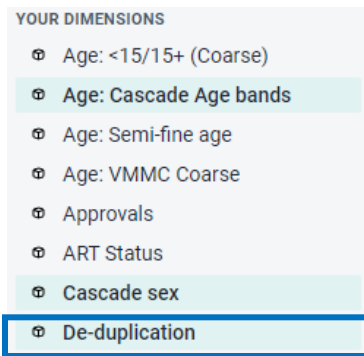
All mechanisms without deduplication	Deduplication adjustments	Total
19,026,489	-128,001	18,898,488