Guidance for Deduplicating Client-Level Data

Deduplication of population-level data for people living with HIV (PLHIV) reduces fragmentation of patient records, data quality gaps, and over-reporting, and can lead to improved clinical outcomes and programmatic decisions for the HIV response.

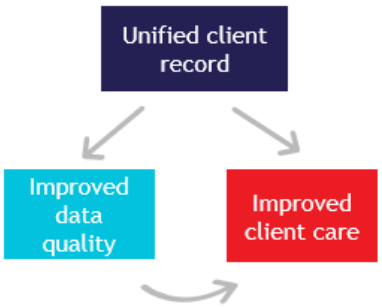
PURPOSE AND AUDIENCE

The purpose of this document is to provide United States Agency for International Development (USAID) Missions with the rationale and options for implementing and routinizing deduplication processes for digitized HIV datasets, with the aim of merging the client record. The primary audience for this document is strategic information (SI) officers at USAID Missions who are responsible for the strategy and implementation of health information systems (HIS) for the President's Emergency Plan for AIDS Relief (PEPFAR).

WHY A DEDUPLICATION PROCESS MATTERS

Deduplication is the process of matching patient records so that one merged client record exists. The deduplication process and resulting merged client record can improve client-centered care and clinical outcomes. For example, multiple client records create a burden for the client to re-share information and reduces each provider's ability to make informed care decisions; further, clients who seek care from multiple sites may be marked as lost to follow-up (LTFU), generating an unnecessary response at each site to support their return to treatment, when they in fact have not left care.

Deduplication is also a critical process for systems interoperability and health information exchange, which enables disparate records systems to be interlinked for automated unification of client records. Here, if duplicate records are correctly identified and unified, viral load results and antiretroviral therapy (ART) dispensing data are recorded into the correct client file. Likewise, if community health information systems are incorporated into the process, information about the services provided to clients at the community level can be exchanged easily with data in electronic medical records (EMRs). This unified patient record improves client treatment monitoring and the efficiency of service delivery.



Finally, to optimize resources to end the epidemic, it is critical for PEPFAR programs to have accurate clinical cascade data to inform programmatic decisions. For example, if a client is registered multiple times within a single service or across multiple services, this may lead to over-reporting,[1] which reduces policymakers' ability to forecast resources required — commodities and human resources for health — and

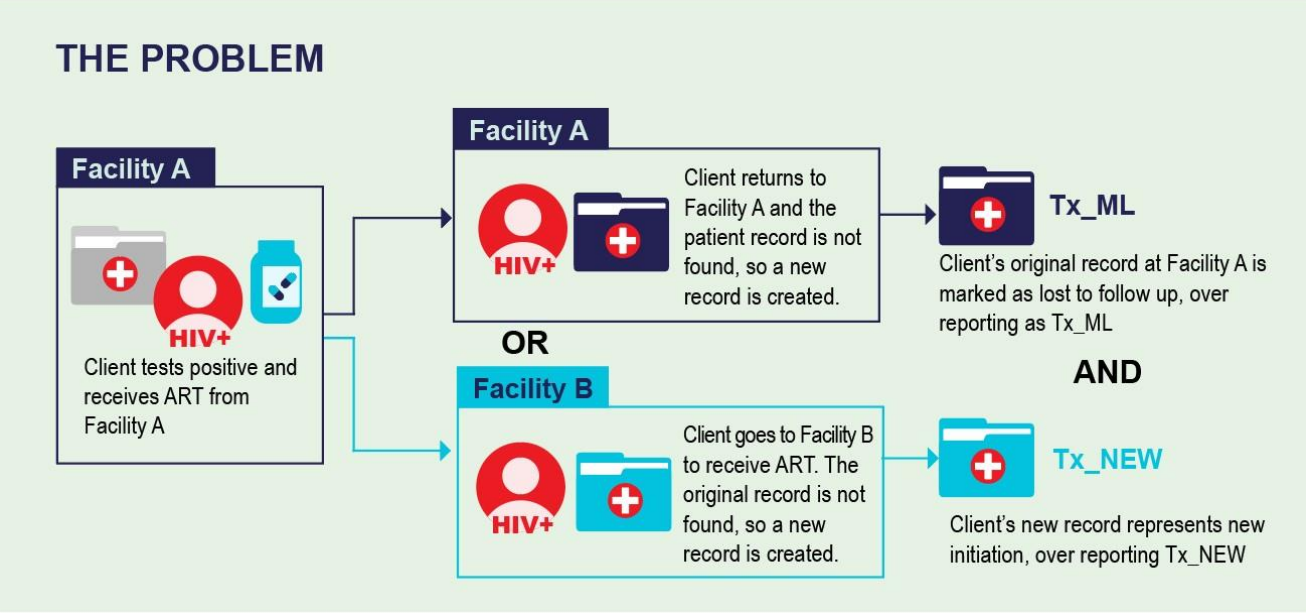
[1] For example, over-reporting of clients lost to follow-up (Tx_ML) or clients new to ART (Tx_New), as shown in the illustration on the next page.

to pinpoint where new services should be offered. Deduplicated, better-quality data can catalyze facility performance improvements, as achievements and challenges can be systematically tracked by stakeholders, with interventions informed; likewise, deduplicated data of community health information systems, orphans and vulnerable children (OVC) and Determined, Resilient, Empowered, AIDS-Free, and Safe (DREAMS) systems, and others, can greatly improve the impact of these interventions.

CAUSES OF DUPLICATION

Duplication of client records can occur for a number of reasons within the field setting, including the following:

At the **client level,** clients may be migratory and may receive care at multiple geographically disparate locations; clients may provide aliases because they want anonymity; and clients may be receiving more than one service at a facility that has a different registration process for each type of service. The figure below depicts this challenge: a client may register for services at multiple sites, or multiple times at one site, and in each instance, different providers have no insight into services received or information shared.



Systemic issues at the facility level may enable duplication of records through the existence of multiple disconnected records systems, poor registration practices, poor system design, a lack of regulatory measures around the standardized capture of demographic data or standards for interoperability, or other related reasons.

At the above-site level, multiple circumstances could permit duplication of records, including poor data system architecture and poor health data governance, such as fragmentation of reporting streams. Data system initiatives have grown with the global HIV response and are designed to improve client services; these include laboratory and pharmacy systems and systems that track community support interventions — OVC, DREAMS, among others. However, one significant downside of the proliferation of these systems,

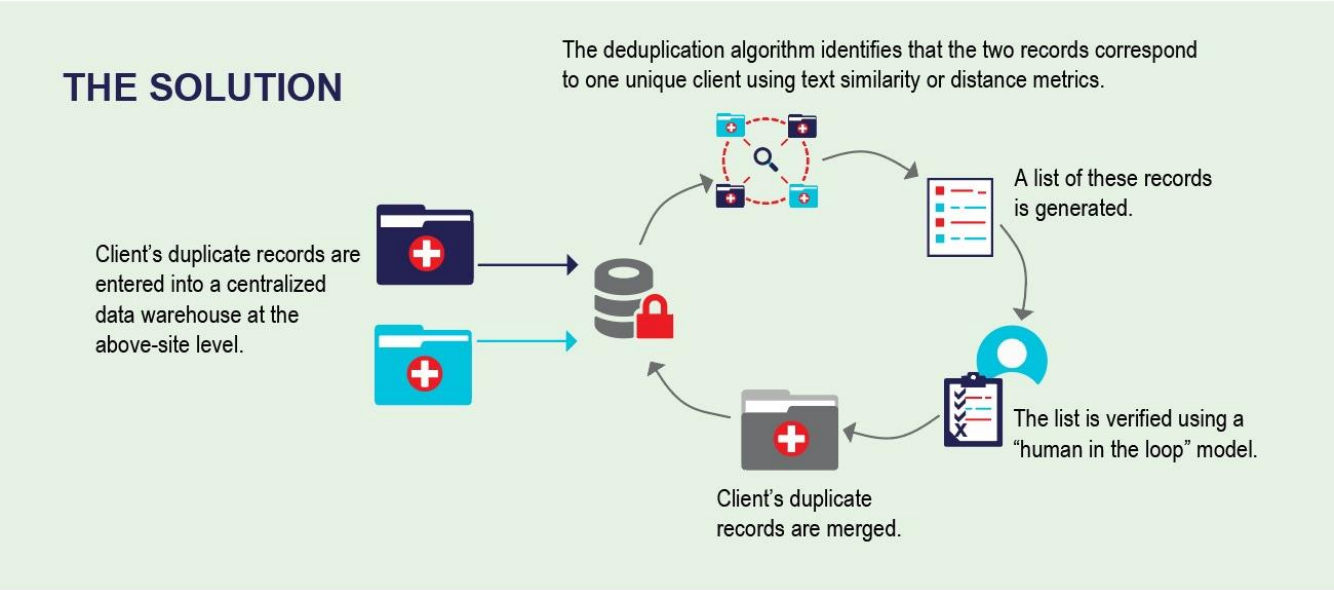
especially in the absence of clear policy and robust infrastructure,[2] is that client records may become duplicated across services and geographic locations, across all of these EMR systems.

THE CASE FOR DEDUPLICATION AS A METHOD FOR ACHIEVING A MERGED CLIENT RECORD

The prevailing approach of country programs toward elimination of duplication has been the pursuit of the unique client identifier, or unique ID (UID). Nevertheless, the promise of most country-led unique health identification initiatives, decades in the planning, is unmet. UID projects face multiple barriers to success. They require whole-of-government support, policy, and consensus; they are expensive to implement; the use of personally identifiable information (PII) involves privacy, confidentiality, and security considerations; and the level of intervention also brings considerations, such as provision of UID at birth or at a later time or event, or using disease-specific UIDs versus UIDs for all healthcare or publicly offered services.

This list of barriers is not exhaustive. Even where UID initiatives are underwritten by policy and have received wide public sector and stakeholder support, they alone are not a panacea for duplicate records. There could be exceptions to the reliability and universality of UID use, or inconsistencies in the processes relating to UID use, such as how and where UIDs are disseminated or how the UIDs are coded within different electronic records systems.

Deduplication of client records should, therefore, be a routine quality management practice for any digital data ecosystem. A deduplication undertaking involves leveraging existing investments in digitization of client-level data, pooling the data, and applying deduplication procedures (as depicted below) to identify a significant proportion of duplicate data; then merging these data towards unifying the client record, and eliminating these data from the reportable data. Applying deduplication procedures on pooled data can help countries achieve the promise of UID initiatives — merged HIV client records — with less cost and effort, and less risk.



[2] Examples of clear policy and robust infrastructure include enterprise architecture, eHealth strategy, interoperability framework, and/or subsystem standards that delineate the regulatory consensus on how best to mitigate duplicative efforts.

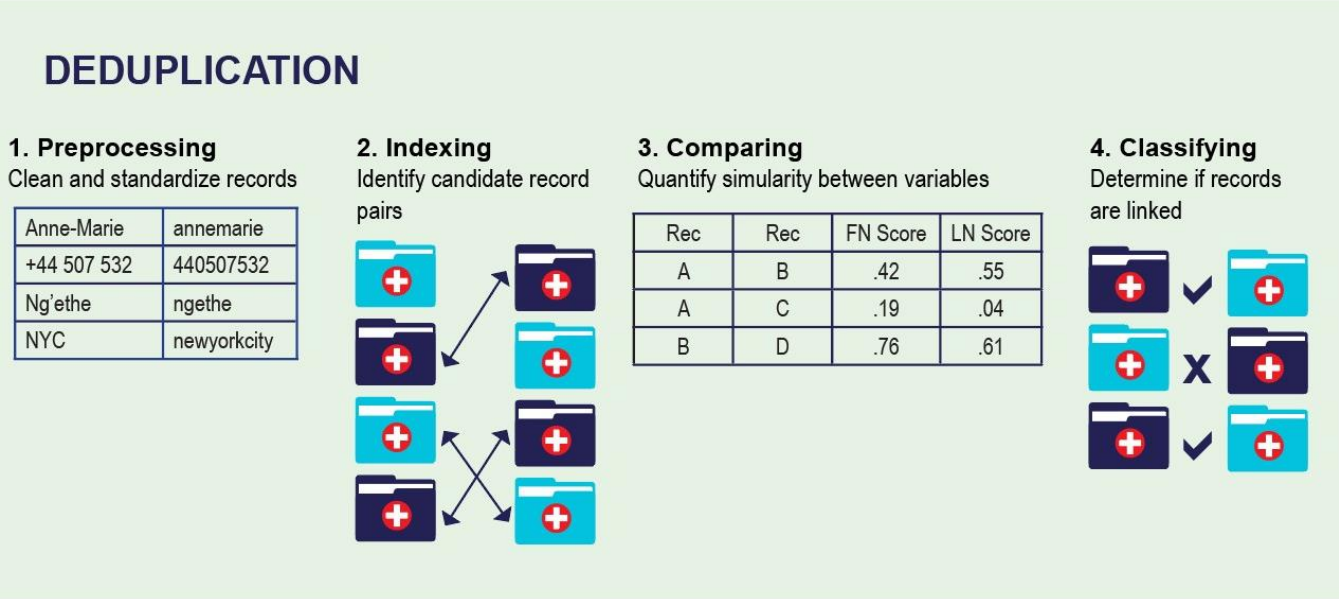
THE DEDUPLICATION PROCESS

Deduplication can be achieved by **deterministic** or **probabilistic** linking methods, whereby a range of stable fields within the datasets are compared for similarity. Essentially, the deduplication process is about *identifying likely links* in separate datasets, verifying whether the likely link is a *true link*, and then *using these true links to connect data from the different datasets* relating to the same entity. Deterministic and probabilistic deduplication are defined further below.

The core steps of the deduplication process are outlined below: preprocessing the client records, indexing, and comparing, before applying a classification algorithm.

Preprocessing

Preprocessing data — which involves such processes as cleaning and standardizing — entails removing stylistic differences in how data are recorded. With text fields, such as client names or places, preprocessing



may include removing non-letters that may appear (e.g., dashes, brackets, or parentheses), removing accents, converting to lowercase, and removing white spaces. For example, in one location, a client may be recorded as “Billy-Jean,” while in another as “billy jean,” which should be identified as the same name. Following preprocessing, both will appear as “billyjean,” facilitating subsequent matching. Similarly, for numeric fields such as phone numbers, preprocessing may include removing non-numbers, such as dropping the “+” before a country code. Other types of fields that may be recorded differently by different actors and should be standardized include street addresses and dates.

Another core step in preprocessing is phonetic encoding of text fields. In this step, text fields are represented by how they are pronounced to mitigate differences from different spellings of the same or similar sound. For example, while “Anne” and “Ann” are spelled differently, their phonetic encodings are the same. A common phonetic-encoding algorithm is the soundex algorithm, though others such as NYSIIS, metaphone, and match rating are also often used.

Indexing

Indexing records is the process of identifying candidates for subsequent matching algorithms. If all records from one facility are to be matched against all records from another facility, that is a full indexing. However, this may materially slow the matching algorithm. If each facility has 1,000 records, full indexing creates

1,000,000 candidate pairs to consider.[3] In practice, the number of records may be much larger, and the number of candidate pairs would scale quadratically.

To accelerate the deduplication process, indexing typically includes steps to eliminate candidate pairs of records that are clearly not duplicates. Such activities include blocking and neighborhood indexing. **Blocking** requires two records to agree on a selected field to be considered as a candidate pair. For example, if blocking on gender, then any pair of records with different genders will be dropped from the list of candidate pairs for subsequent matching. Alternatively, blocking can be applied to multiple variables, so that only records that match on any of a set of fields are identified as candidate pairs (e.g., records must match on either first name or last name). Identifying which fields are good candidates for blocking requires an understanding of the norms of the context in which deduplication is being applied, such as knowing which fields would be provided in a consistent manner and which fields might vary in how they are reported or captured.

Because blocking requires an exact match to qualify as a candidate pair, it may be too restrictive and may eliminate actual matches with minor spelling differences. Sorted **neighborhood indexing** is an approach that seeks to cast a wider net for identifying candidates. In this approach, one sorts a variable in a list or series and considers record pairs as candidate pairs if they are near each other. Users would determine how near is near enough by selecting a so-called "window" size. For example, if applying this approach to first names, one might sort the names in alphabetical order and identify candidate pairs that are close to each other. If considering two records with the first names "Duane" and "Dwayne," strict blocking will eliminate this as a candidate pair, whereas sorted neighborhood matching may identify this pair as a candidate based on the window size. Once again, careful consideration of the local context and examination of the actual records to be matched should inform the best approach to indexing.

Comparing

With candidate record pairs identified, the next step is comparing variables among candidate pairs and quantifying how similar or different the variables are. Techniques in this stage include text similarity measures, numeric similarity scores, and distance metrics. Numerous types of variables can be compared, and some techniques are reviewed below. The result is a feature vector for each candidate pair containing the outputs of the comparing techniques for each variable.

The simplest technique is an exact comparison, which returns 1 if two fields match exactly and 0 if they do not, and can be applied to any type of variable — text, numeric, date, categorical, or other. As a more nuanced approach for text fields, a variety of text similarity approaches can be applied to quantify the similarity of two text fields. Such approaches include Jaro-Winkler and Levenshtein distance, among others. Broadly, these algorithms consider what share of letters match, whether there are common letters that appear in different orders, and the overall number of letters. The output may be a number between 0 and 1, with numbers close to 1 representing highly similar fields and numbers close to 0 representing dissimilar fields. Similar approaches may be applied to fields such as zip codes, street addresses, and phone numbers.

Similarly, there are a variety of approaches for comparing numeric fields, which include dates (e.g., ART start dates) and latitude-longitudes, among others. One such approach used by Elasticsearch calculates the distance between the variables (such as kilometers, if considering global positioning system [GPS]

[3] Richard Ngethe, R., & Friedman, J. (2020). It's a puzzle, it's an algorithm, it's deduplication. Available at https://sciencespeaksblog.org/2020/12/11/its-a-puzzle-its-an-algorithm-its-deduplication

coordinates) and produces a score that diminishes as distance increases. The output again creates a feature vector with values that range between 0 and 1.

Classification

With feature vectors created for each candidate pair, the next step is to classify whether candidate pairs are likely or unlikely matches. Here, a variety of approaches are available that range in sophistication, with different considerations for each. At the most basic, if one lacks the infrastructure or capacity to create the comparison vectors, one could apply a **simple deterministic matching approach** that identifies candidate pairs as matches if they match exactly on a user-selected field or fields. If a reliable unique ID exists, this approach may work well; if not, one might look to match on other fields, such as name and date of birth. The need for an exact match means that simple deterministic methods will identify fewer links and neglect to identify a potentially large number of links that lack exact matches (low sensitivity, or recall). However, the need for an exact match means that suggested links are very likely correct (high positive predictive value, or PPV) and that few incorrect links will be suggested.

In a step above simple deterministic methods, **more nuanced deterministic methods** apply rules to the comparison vectors to make matching determinations. For example, if the feature vector contains four values that reflect scores for a first name comparison, a last name comparison, a date comparison, and a geographic comparison, then one might classify as a match any record pair for which each value exceeds a user-defined number. Alternatively, one could simply add the scores together and classify as a match any candidate pair with a sum score above a user-defined threshold. Practically, the approach would need to be tailored and refined to the local context through testing and evaluation of performance.

More sophisticated yet, **probabilistic matching** considers the likelihood that candidate pairs will have similar comparison vectors given that they are a match or not a match. One core element in probabilistic approaches is that they take into consideration the likelihood of data quality errors, such as errors in data entry or coding. The idea is to consider, for each variable, the probability that the two fields will be similar, given that records are indeed a match. For example, as names may be entered with a relatively high frequency of misspellings, this probability may be low. Hypothetically, there may be a higher probability that gender will match among candidate pairs that reflect the same client. The second core element is the probability that values will match or be highly similar on any random pair of records. Extending the previous example, the probability that a random pair of records will match on gender is roughly 50 percent, whereas the probability of matching on month of birth is 8.25 percent (1/12).

Relatedly, the Sequoia Project crafted a framework to help identify which data fields are most useful for probabilistic matching, which are in line with these concepts. The project identified the following dimensions:

- **Completeness:** The rate at which the field is recorded on patient records.
- **Validity:** Is the field recorded correctly (e.g., when precise dates of birth are not known, estimated dates may be provided)?
- **Distinctiveness:** The extent to which the field will vary across patients (e.g., as stated above, gender is not distinctive, whereas a master patient index [MPI] would be distinctive).
- **Comparability:** If the trait is structured (e.g., numeric fields), it may be easier to compare than unstructured fields, such as free text (e.g., street addresses).
- **Stability:** The extent to which the value for a patient remains constant over time (e.g., date of birth is stable, but street addresses and phone numbers may change).

In practice, the Sequoia Project, as part of a consortium with a not-for-profit health system, evaluated fields along these dimensions. In the study, it found that first name, last name, gender, and date of birth ranked highly and were good candidate fields. Other fields, such as street address and phone number, were more distinctive but were harder to compare and less stable. Country programs are encouraged to be proactive in determining useful indicators for deduplication, based on these criteria and on other local considerations.

Probabilistic approaches use these concepts to determine the likelihood of seeing the values in a comparison vector given the assumption that records are in fact linked or not linked. Probabilistic approaches produce a likelihood score for each record pair, in which, again, users would set a threshold to classify links. Because probabilistic approaches do not rely on exact matches, they are more likely to identify a greater number of true links (higher sensitivity), while at the same time they may be more likely to identify records as linked that should not in fact be linked (lower PPV).

Overall**, probabilistic approaches are preferred to deterministic approaches** as they are more flexible, more easily achieve higher sensitivity, and can be refined through continuous testing and evaluation to produce better combinations of sensitivity and PPV.

Verifying True Duplicates and Merging Duplicate Records

Verification and merging are the last steps of the deduplication process. **Verification** requires a person to review whether potential links are true links. Over time, through machine-learning approaches, this step can become more and more automated. Deduplication teams must consider the human resource implications of this verification step, which — depending on data quality, the size of datasets, and the sensitivity of the matching process — may be significant.

Merging is also, at least initially, a human-driven step. It can occur at the above-site level, where deduplication is being conducted to mitigate reporting discrepancies, or at the facility level, where client records are merged for comprehensive consolidation of patient history for optimum clinical service provision. Similar to the verification process, the human resource requirements for merging should be well considered and depend on the scale of the duplicate records, the level of intervention, and several other local, contextual factors.

Tools for Deduplication

Useful tools for deduplication differ based on several factors, including the technology platform of the health information systems in use and the types of indicators being processed. Two commonly used open-source deduplication tools are Python Record Linkage Toolkit[4] and Registry Plus Link Plus.[5]

FOLLOW-ON ACTIVITIES

Once data has been deduplicated and merged at the above-site level, there is great potential to use these validated data for improved client care and facility-level management of the HIV response, for improved community-level health interventions, and for strengthened reporting. An immediate impact of implementing deduplication at the facility level would be improved data quality and clinical service delivery; the case of the client, above, is an excellent example of how duplication of a client's record results in incomplete data history at any one point of care, potentially hindering care, and results in misclassification of LTFU and over-reporting of new, current, and cumulative on ART. Likewise, an immediate impact of implementing deduplication at the community level could be improved contract tracing, improved accuracy of LTFU and/or

4 https://recordlinkage.readthedocs.io/en/latest/about.html

5 https://www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm

identification of undocumented mortality. As such, deduplication is complementary to data quality initiatives and can lead to improvements in continuous quality improvement.

Therefore, an important follow-on activity of deduplication is for country programs to create the architecture and processes to update facility and community-level information systems. This could entail implementing simple bi-directional reporting systems to routinely provide line lists of duplicated data back to data managers for data cleaning of duplicates and merging of patient records within the primary record, via email or by data managers downloading data from a secure website. More sophisticated architecture could include country programs developing master patient indexes (MPIs) — in which unique, deduplicated, and merged records are curated centrally — and to build out the infrastructure and processes, including health information exchange, to enhance information system functionality to search and compare patient records against the MPI for true duplicates to be merged.

FOR MORE INFORMATION

To learn more about deduplication and how it can be applied in your program, contact Julianna Kohler (jkohler@usaid.gov) or Jacob Buehler (jbuehler@usaid.gov), USAID/OHA/SIEI, Health Informatics Team.

To discuss how Data.FI may be able to support the Mission with deduplication, please contact Emily Harris, Data.FI AOR (emharris@usaid.gov) and Jenifer Chapman, Project Director (jenifer.chapman@thepalladiumgroup.com).

FURTHER READING

Probabilistic record linkage, Adrian Sayers, Yoav Ben-Shlomo, Ashley W Blom and Fiona Steele, International Journal of Epidemiology, 2016, Advance Access Publication Date: 20 December 2015.

Design and implementation of a privacy preserving electronic health record linkage tool in Chicago, Abel N Kho et al, Am Med Inform Assoc Journal, 2015.

A framework of identity resolution: evaluating identity attributes and matching algorithms, Jiexun Li1 and Alan G. Wang, Security Informatics Journal, 2015.

A Probabilistic Matching Approach to Link De-identified Data from a Trauma Registry and a Traumatic Brain Injury Model System Center, M. Kesinger, RG. Kumar, AC. Ritter, JL. Sperry, AK. Wagner, American Journal of Physical Medicine and Rehabilitation, 2017.

A Framework for Cross-Organizational Patient Identity Matching, The Sequoia Project, 2018.

SB-20-01

Data for Implementation (Data.FI) is a five-year cooperative agreement funded by the U.S. President's Emergency Plan for AIDS Relief through the U.S. Agency for International Development under Agreement No. 7200AA19CA0004, beginning April 15, 2019. It is implemented by Palladium, in partnership with JSI Research & Training Institute (JSI), Johns Hopkins University (JHU) Department of Epidemiology, Right to Care (RTC), Cooper/Smith, IMC Worldwide, Jembi Health Systems, and Macro-Eyes, and supported by expert local resource partners.

This publication was produced for review by the U.S. President's Emergency Plan for AIDS Relief through the United States Agency for International Development. It was prepared by Data.FI. The information provided is not official U.S. Government information and does not necessarily reflect the views or positions of the U.S. President's Emergency Plan for AIDS Relief, U.S. Agency for International Development, or the United States Government.

JANUARY 2021

FOR MORE INFORMATION

Contact Data.FI:

Emily Harris, Data.FI AOR
emharris@usaid.gov

Jenifer Chapman, Data.FI Project Director
datafiproject@thepalladiumgroup.com

https://datafi.thepalladiumgroup.com/