Landscape Analysis of Privacy Preserving Patient Record Linkage Software (P3RLS)

Final Report Version 2

Prepared By

Synectics for Management Decisions, Inc.

15 January 2020

PI: Jesse Aronson, MS, PE, PMP | jesse.aronson@smdi.com | 703-568-7606

Study performed for Leidos Biomedical Research (LBR) under Agreement 18Q110, issued as a subcontract under contract HHSN261201500003I issued by the National Cancer Institute (NCI), National Institutes of Health (NIH), Department of Health and Human Services (HHS)

This page intentionally left blank

Table of Contents

1 Executive Summary

This version updates the final report to include evaluation of the HealthVerity Census/Marketplace products. HealthVerity opted not to participate in the original survey but was added after subsequent contact with the company

This document presents the findings of the Landscape Analysis of Privacy Protecting Record Linkage Software (P3RLS). This work was performed for Leidos Biomedical Research (LBR) under Agreement 18Q110, issued as a subcontract under contract HHSN261201500003I issued by the National Cancer Institute (NCI), National Institutes of Health (NIH).

The landscape analysis was performed using a structured methodology based on systems engineering best practices. The process started with development and capture of requirements. Development of requirements was a collaborative effort between the study team and the Integrated Project Team (IPT) which include representation from NCI and LBR. From the requirements, evaluation criteria were developed, and the resultant criteria were categorized into those that could be applied during this survey phase and those which would be applicable during a pilot phase involving hands-on testing of candidate software. The criteria to be used during the landscape analysis phase formed the basis for a survey questionnaire to be completed by vendors of candidate software products.

In parallel, the team surveyed the marketplace of record linkage software. While there are many record linkage products on the market addressing application areas in finance, healthcare, and marketing, few included privacy-protecting features. An initial analysis of fifty-two (52) products yielded only eight which appeared to provide the privacy-protecting record linkage features desired by NCI. These eight were selected for further evaluation. For each of the eight vendors, an initial contact by email was followed by an introductory phone call, which was in turn followed by distribution of the survey questionnaire to the vendor. Seven of the eight vendors returned completed surveys.

Based on questionnaire responses, a score was developed for each product. Each question in the questionnaire traces back to a requirement, and was weighted based on whether its associated requirement had been ranked by the IPT as “Must Have” vs. “Should Have” vs. “Could Have.” Each question was also weighted based a requirements category, with PPRL Functionality having the highest weight, followed by Usability and Scalability, with all other categories (Interoperability, etc.) having lower emphasis.

The result is that four products are recommended for further evaluation through a pilot phase. These include software from HealthVerity, Senzing, Crossix, and Datavant, all of which scored highly and closely enough to be considered equally viable as candidates for NCI’s use. The remainder of this draft report provides more details on the study process and results.

***Note**: Throughout this report the terms "privacy-protected record linkage" (PPRL) and "privacy-protected patient record linkage" (P3RL) are used interchangeably, since the products do not distinguish between patient records and other linkable record.*

2 Study Process

The goal of this study was to assess the landscape of currently available privacy preserving patient record linkage software (P3RLS) in the context of NCI needs, and ultimately make a recommendation of one or more software products to be used for pilot testing, along with developing the associated evaluation criteria for that pilot testing.

The landscape analysis used a structured process based on system-engineering best practices. The process started by working with NCI and LBR stakeholders to develop, capture, and prioritize requirements. From the requirements, evaluation criteria were developed, and the resultant criteria were segregated into those that could be applied during this survey phase and those which would be applicable during a pilot phase involving hands-on testing of candidate software. The criteria to be used during the landscape analysis phase formed the basis for a survey questionnaire to be completed by vendors of candidate software products. For each vendor, an initial contact by email or website was followed by a 30-minute introductory phone call, which, if the vendor agreed, was in turn followed by distribution of the survey questionnaire to the vendor. Based on questionnaire responses, a score was developed for each product.

The team conducted extensive research to identify candidate software products to include in the survey. While there is a wide selection of record linkage software available, relatively few of these products offer privacy-protecting features. Fifty-two (52) record linkage products were identified, of which eleven were determined to have privacy-protecting record linkage features. This is a dynamic technical field, and over the course of the survey several of the candidate vendors merged, and one vendor (IBM) did not respond to our contacts, leaving a final field of eight candidate products. The final list of candidate software products, the preliminary scores, and the status of the analysis of that product are shown in Table 1. The complete list of products evaluated is included in this report in Section 14.

Figure 1 summarizes the study process. This report contains additional detail on each step of the process; the text below the figure identifies where in the report information about each step can be found.

Figure 1: Landscape analysis study process

Task 1, *Capturing Requirements*: Determining the needs to meet for P3RLS software, integrating and prioritizing the capabilities expressed by the various stakeholders for the effort. More information on development of requirements can be found in Section 3.

Task 2, *Identify Candidate Software*: Surveying the marketplace to identify candidate record linkage software, narrowing down the set of candidates based on key privacy-protection requirements and initial interviews with vendors. More information on development of requirements can be found in Section 6.

Task 3, *Developing Evaluation Criteria:* Transforming requirements into an unambiguous, measurable form that can be used both in this phase and subsequent pilot testing to evaluate candidate software solutions. More information on development of requirements can be found in Section 5.

Task 4, *Evaluate Software:* Developing a scoring approach for the candidate software products, collecting product information and assessing each candidate software product to determine the extent to which it meets NCI requirements. More information on development of requirements can be found in Section 7.

Task 5, *Prepare Recommendations*: Reviewing the results of the software evaluation to develop recommendations for next steps and documenting the evaluation methodology and results of the systematic review of candidate P3RLS software.

3 Findings

The set of privacy-protecting record linkage products currently on the market is small. The landscape is also highly dynamic; the initial list of PPRL candidate software products shrank from eleven to eight over the course of our analysis as companies merged or went out of business and products went end-of-life. The number of recent research publications in the area reveals that there is vibrant research and development being done in this area, which may lead to new or enhanced products over time. Any product selection in this area should take into consideration these dynamic forces.

This study was chartered to recommend candidates to take forward into deeper evaluation through hands-on experimentation. As such, the scoring performed on the software was not designed to be an absolute ranking. Rather, it was intended to identify those products which sufficiently align with NCI's requirements to merit further experimentation. While no threshold "passing" score for products was established *a priori*, past experience predicted that scores in such evaluations tend to cluster, and indeed the scores of the P3RLS products clustered at three levels: those with total scores in the 30s (best fits, recommended to move forward into pilot evaluation), those in the 20s (products with merit, but which fall short in some way) and those in the teens (not sufficiently aligned with NCI requirements).

The final scores for the products are shown in Figure 2. Four vendors, HealthVerity, Crossix, Datavant, and Senzing, received high scores and are recommended for further analysis. In the second tier, CSIRO Anonlink is an open-source product from Australia. Its lack of commercial support and low ease of use were factors in lower scoring. Privitar is a commercial company based in the United Kingdom. Its product's primary use case involved the UK health system (which has the benefit of a universal patient identifier) and its representative expressed concern about linking more heterogeneous data sets. Privitar's score also suffered somewhat both due to a narrow feature set and the incompleteness of their survey response.

Vendor	Score
Crossix	35.1
HealthVerity	32.7
Datavant	31.5
Senzing	31.5
CSIRO	25.7
Privitar	23.3
Policywise	17.2

Figure 2: P3RLS Product Evaluation Scores

Policywise is a non-profit based in Alberta, Canada. It was clear from our interaction with Policywise that Linkwise was a small effort adjacent to its core mission. It offered a small feature set and had little documentation or ongoing development and support. The eighth product in the final set was GRHANITE from the University of Melbourne, Australia. They participated in an initial interview with the study team but failed to return the full survey and so were not included in the final analysis.

The products have varying feature sets and so a full consideration of how to implement privacy-protected record linkage for NCI should consider an architectural view of how the complete minimal set of requirements will get implemented. For example, pre-linkage data cleaning (Requirement F-13) was not rated by the IPT as a "Must Have" requirement because it can be implemented as a pre-processing step prior to record linkage; however, if a record linkage tool lacking data cleaning features (such as nickname substitution and phonetic name encoding) is selected, that function still somehow needs to be provided.

Survey Results - Candidate Responses & Scores

Legend:	Question Response	Question Response
	Meets	Fully Meets
	Partial	Partially Meets
	Custom	Meets with Customization
	No	Does Not Meet

Crossix	Datavant	Senzing	Anonlink	Privitar	PolicyWise	HealthVerity
35.1	*31.5*	*31.5*	*25.7*	*23.3*	*17.2*	*32.7*

			Metered Score:	*28.5*		*25.1*		*25.1*		*19.4*		*20.2*		*10.9*		*10.9*	
Question #	**Question Category**	**Question**	**Coefficient**	**Crossix Response**	**Crossix Score**	**Datavant Response**	**Datavant Score**	**Senzing Response**	**Senzing Score**	**Anonlink Response**	**Anonlink Score**	**Privitar Response**	**Privitar Scroe**	**PolicyWise Response**	**PolicyWise Score**	**HealthVerity Response**	**HealthVerity Score**
1	PPID Generation and Record Linkage	What can the user specify about the PPID generation process (e.g., which variables go into the PPID, and whether they should be truncated)?	0.50	Meets	1.50	Meets	1.50	Meets	1.50	Meets	1.50	Meets	1.50	Meets	1.50	Partial	1.00
2	PPID Generation and Record Linkage	Can the software generate and link on multiple PPIDs (e.g., hashes composed from different concatenated input variable combinations) either as a single pass or multiple passes? Is there a maximum number?	0.50	Meets	1.50	Meets	1.50	Meets	1.50	Custom	0.50	Meets	1.50	Meets	1.50	Meets	1.50
5	PPID Generation and Record Linkage	How many files can be simultaneously linked (e.g., can the software link more than two files in one pass)?	0.50	Meets	1.50	Meets	1.50	Meets	1.50	Meets	1.50	Meets	1.50	No	0.00	Meets	1.50
6	PPID Generation and Record Linkage	Does the product support deduplication?	0.50	Meets	1.50	Partial	1.00	Meets	1.50	Custom	0.50	No	0.00	Custom	0.50	Meets	1.50
7	PPID Generation and Record Linkage	Can the software support more than pairwise linkages (e.g., find all the records in a file that match)?	0.50	Meets	1.50	Meets	1.50	Meets	1.50	Custom	0.50	Meets	1.50	No	0.00	Meets	1.50
8	PPID Generation and Record Linkage	Does the product support two-party privacy protected linkage (identifying linkages between two files potentially belonging to two owners, where neither party sees the other's unencrypted data)?	0.50	Meets	1.50	Meets	1.50	Meets	1.50	Meets	1.50	Meets	1.50	Meets	1.50	Meets	1.50
9	PPID Generation and Record Linkage	Does the product support three-party protect linkage, where there is a trusted "honest broker" able to resolve possible linkages?	0.50	Meets	1.50	Meets	1.50	Meets	1.50	Meets	1.50	Meets	1.50	Meets	1.50	Meets	1.50
10	PPID Generation and Record Linkage	For three-party linkage, what information is made available to the trusted broker to resolve "possible" linkages (e.g., reports, distance metrics, comparisons of source data - if the broker can see source data)?	0.50	Meets	1.50	Meets	1.50	Meets	1.50	Custom	0.50	No	0.00	No	0.00	Meets	1.50
11	PPID Generation and Record Linkage	Are there any features for authorized reidentification of data?	0.50	Partial	1.00	Meets	1.50	Meets	1.50	Meets	1.50	Meets	1.50	No	0.00	Meets	1.50
12	PPID Generation and Record Linkage	What is tunable about matching criteria/algorithm?	0.50	Meets	1.50	Meets	1.50	Meets	1.50	Meets	1.50	No	0.00	Partial	1.00	Meets	1.50

Question #	Question Category	Question	Coefficient	Crossix Response	Crossix Score	Datavant Response	Datavant Score	Senzing Response	Senzing Score	Anonlink Response	Anonlink Score	Privitar Response	Privitar Scroe	PolicyWise Response	PolicyWise Score	HealthVerity Response	HealthVerity Score
13	PPID Generation and Record Linkage	Does the software have any ability to persist results so that subsequent linkages between data sets (e.g., after updates) can be incremental rather than from scratch?	0.50	Meets	1.50	Meets	1.50	Meets	1.50	Custom	0.50	Meets	1.50	No	0.00	Meets	1.50
14	PPID Generation and Record Linkage	Does the software have the ability to split databases into linked vs. non-linked records, or other splitting and merging capability?	0.50	Meets	1.50	Custom	0.50	Partial	1.00	No	0.00	Custom	0.50	No	0.00	Meets	1.50
15	PPID Generation and Record Linkage	Can the product persist PPIDs so they don't have to be regenerated for future runs?	0.50	Meets	1.50	Meets	1.50	Meets	1.50	Meets	1.50	Custom	0.50	No	0.00	Meets	1.50
19	Operating Environment and Licensing Model	Cloud-based version available? If so, which cloud environment?	0.05	Meets	0.15	Custom	0.05	Partial	0.10	Meets	0.15	Meets	0.15	No	0.00	Meets	0.15
21	Operating Environment and Licensing Model	Does your licensing support "record linkage as a service", either through offering a cloud-based service or by distribution of the software as a utility?	0.05	Meets	0.15	Meets	0.15	No	0.00	Meets	0.15	Custom	0.05	Meets	0.15	Meets	0.15
22	Operating Environment and Licensing Model	Is the system a set of components or a single, integrated software program? Does it require software to be developed to create a complete application?	0.05	Meets	0.15	Meets	0.15	Meets	0.15	Custom	0.05	Meets	0.15	Meets	0.15	Meets	0.15
23	Usability and Security Features	Does the product include a graphical user interface (GUI)?	0.15	Partial	0.30	Custom	0.15	Custom	0.15	No	0.00	Meets	0.45	Meets	0.45	Meets	0.45
24	Usability and Security Features	Does the product include the ability to save configurations to facilitate multiple runs using the same parameters?	0.15	Meets	0.45	Meets	0.45	Meets	0.45	Meets	0.45	Meets	0.45	No	0.00	Meets	0.45
25	Usability and Security Features	Can the software be scripted to perform operations automatically?	0.15	Meets	0.45	Meets	0.45	Meets	0.45	Meets	0.45	Meets	0.45	No	0.00	Meets	0.45
26	Usability and Security Features	Does the software require configuration, or can it be used "out of the box"?	0.15	Meets	0.45	Partial	0.30	Meets	0.45	Partial	0.30	Meets	0.45	Meets	0.45	Meets	0.45
27	Usability and Security Features	Describe the product documentation available (provide link if possible).	0.15	Meets	0.45	Meets	0.45	Meets	0.45	Meets	0.45	Meets	0.45	Custom	0.15	Meets	0.45
30	Usability and Security Features	Is there an active development effort for the product?	0.15	Meets	0.45	Meets	0.45	Meets	0.45	Meets	0.45	Meets	0.45	No	0.00	Meets	0.45
31	Usability and Security Features	Describe the product support available.	0.15	Meets	0.45	Meets	0.45	Meets	0.45	Custom	0.15	Meets	0.45	Partial	0.30	Meets	0.45
33	Usability and Security Features	Does the system contain security features such as requiring login/authentication?	0.15	Meets	0.45	Meets	0.45	Custom	0.15	Meets	0.45	Meets	0.45	No	0.00	Meets	0.45
34	Usability and Security Features	Are there different user roles (e.g., administrator vs. user vs. data manager)?	0.15	Meets	0.45	Meets	0.45	Custom	0.15	Partial	0.30	Meets	0.45	No	0.00	No	0.00
35	Usability and Security Features	If there are different roles, is the user interface segregated and optimized by role (e.g., a researcher would see the features of interest to an end-user, while an administrator would see a more full set of configuration functions)?	0.15	Meets	0.45	Partial	0.30	Custom	0.15	No	0.00	Meets	0.45	No	0.00	No	0.00
36	Usability and Security Features	What execution performance reports are available (e.g., execution time, number of record pair comparisons, etc.).	0.15	Meets	0.45	Custom	0.15	Custom	0.15	Meets	0.45	Partial	0.30	Partial	0.30	Meets	0.45

Question #	Question Category	Question	Coefficient	Crossix Response	Crossix Score	Datavant Response	Datavant Score	Senzing Response	Senzing Score	Anonlink Response	Anonlink Score	Privitar Response	Privitar Scroe	PolicyWise Response	PolicyWise Score	HealthVerity Response	HealthVerity Score
36	Usability and Security Features	What execution performance reports are available (e.g., execution time, number of record pair comparisons, etc.).	0.15	Meets	0.45	Custom	0.15	Custom	0.15	Meets	0.45	Partial	0.30	Partial	0.30	Meets	0.45
37	Usability and Security Features	What linkage performance reports are available (e.g., number of matches, number of possible matches, number of duplicates - if the software does de-duplication, etc.).	0.15	Meets	0.45	Custom	0.15	Custom	0.15	Meets	0.45	Custom	0.15	No	0.00	Meets	0.45
38	Usability and Security Features	Has the system been approved to operate under U.S. government security regulations such as FISMA or FedRAMP?	0.15	Meets	0.45	No	0.00	No	0.00	No	0.00	Partial	0.30	No	0.00	Custom	0.15
39	Usability and Security Features	Can the system run in a mode which does not persist any data (to minimize security risks)?	0.15	Meets	0.45	Meets	0.45	Custom	0.15	Custom	0.15	Custom	0.15	Meets	0.45	Meets	0.45
40	Usability and Security Features	What protections are in place for source data?	0.15	Meets	0.45	Meets	0.45	Custom	0.15	Meets	0.45	No	0.00	Meets	0.45	Meets	0.45
43	External System Integration	Can the software be configured to be flexible about input formats (e.g., mapping input columns to program variables), or must inputs be put into a particular format?	0.05	Meets	0.15	Meets	0.15	No	0.00	Custom	0.05	No	0.00	Partial	0.10	Meets	0.15
45	External System Integration	Can the user customize the outputs?	0.05	Meets	0.15	Meets	0.15	No	0.00	Custom	0.05	No	0.00	Meets	0.15	Meets	0.15
46	Data Cleaning / Pre-Processing Features	Describe any features the tool has to identify data quality issues, and standardize, recode and clean data to improve matching performance (e.g., substitution of nicknames such as "Jim" to "James", address lookup and standardization, geocoding, phonetic matching).	0.05	Meets	0.15	Partial	0.10	Meets	0.15	Meets	0.15	Custom	0.05	Partial	0.10	Meets	0.15
47	Data Cleaning / Pre-Processing Features	Is pre-processing specifiable field by field and file by file (e.g., a different date cleaning for DOB vs. record date, and for DOB in file 1 vs. file 2)?	0.05	Meets	0.15	No	0.00	Meets	0.15	Meets	0.15	Meets	0.15	Custom	0.05	Meets	0.15
48	Data Cleaning / Pre-Processing Features	Is the product extensible to use user-supplied pre-processing modules/services?	0.05	Meets	0.15	Custom	0.05	Meets	0.15	Meets	0.15	Meets	0.15	No	0.00	Partial	0.10
49	Data Cleaning / Pre-Processing Features	Can the software export subsets of the pre-processed data fields (e.g., only certain columns, only certain rows)?	0.05	Meets	0.15	Partial	0.10	No	0.00	Partial	0.10	Partial	0.10	Meets	0.15	Meets	0.15
50	Data Cleaning / Pre-Processing Features	Does the software support an evaluation mode (e.g., to allow researchers to work with their own data sets to clean and tune data before linkage)?	0.05	Meets	0.15	Meets	0.15	Meets	0.15	Meets	0.15	Partial	0.10	No	0.00	Meets	0.15
53	Performance and Scalability	What features does the software have (such as blocking or database indexing) to improve performance? Can the user specify blocking parameters?	0.15	Meets	0.45	Partial	0.30	Meets	0.45	Partial	0.30	Meets	0.45	No	0.00	No	0.00
54	Performance and Scalability	Describe the ability to customize performance improvement features such as blocking?	0.15	Meets	0.45	Custom	0.15	Meets	0.45	No	0.00	No	0.00	No	0.00	No	0.00
55	Performance and Scalability	How can performance be improved by adding computational power (e.g., elastic compute)?	0.15	Meets	0.45	Meets	0.45	Meets	0.45	Meets	0.45	Meets	0.45	No	0.00	Meets	0.45

			Narrative Score:	*6.6*		*6.5*		*6.4*		*6.3*		*3.1*		*6.3*		*6.3*	
Question #	**Question Category**	**Question**	**Coefficient**	**Crossix**		**Datavant**		**Senzing**		**Anonlink**		**Privitar**		**PolicyWise**		**HealthVerity**	
				Response	**Score**	**Response**	**Score**	**Response**	**Score**	**Response**	**Score**	**Response**	**Score**	**Response**	**Score**	**Response**	**Score**
3	PPID Generation and Record Linkage	What is the mechanism for generating PPIDs (e.g., SHA-2 hashing)?	0.50	Meets	1.5	Meets	1.5	Meets	1.5	Meets	1.5	No	0	Meets	1.5	Meets	1.50
4	PPID Generation and Record Linkage	What probabalistic matching capability is available (e.g., Bloom filters on q-grams)?	0.50	Meets	1.5	Meets	1.5	Meets	1.5	Meets	1.5	No	0	Meets	1.5	Meets	1.50
16	Operating Environment and Licensing Model	What Platform/OS(s) does the system run under?	0.05	Meets	0.15	Meets	0.15	Meets	0.15	Meets	0.15	Meets	0.15	Meets	0.15	Meets	0.15
17	Operating Environment and Licensing Model	What other software is required to run your software (e.g., DBMS)?	0.05	Meets	0.15	Meets	0.15	Meets	0.15	Custom	0.05	Meets	0.15	Meets	0.15	Meets	0.15
18	Operating Environment and Licensing Model	Minimum hardware specification.	0.05	Meets	0.15	Meets	0.15	Meets	0.15	Meets	0.15	Custom	0.05	Meets	0.15	Partial	0.10
20	Operating Environment and Licensing Model	Licensing model (per seat, per CPU, open source, etc.).	0.05	Meets	0.15	Meets	0.15	Meets	0.15	Meets	0.15	Meets	0.15	Meets	0.15	Partial	0.10
28	Usability and Security Features	When was the software first released?	0.15	Meets	0.45	Meets	0.45	Meets	0.45	Meets	0.45	Meets	0.45	Meets	0.45	Meets	0.45
29	Usability and Security Features	When was the most recent release of the software?	0.15	Meets	0.45	Meets	0.45	Meets	0.45	Meets	0.45	Meets	0.45	Meets	0.45	Meets	0.45
32	Usability and Security Features	Is the software single-user or multi-user? If multi-user, how does the system manage integrity and security of data and ensure partitioning between users?	0.15	Meets	0.45	Meets	0.45	Meets	0.45	Meets	0.45	Meets	0.45	Meets	0.45	Meets	0.45
41	External System Integration	What file formats can the software use (e.g., delimited and fixed-width text files, MS Excel, XML, JSON)?	0.05	Meets	0.15	Meets	0.15	Meets	0.15	Meets	0.15	Meets	0.15	Meets	0.15	Meets	0.15
42	External System Integration	Does the software integrate directly with data sources for input and/or output (e.g., ODBC/JDBC integration with relational database, web services)? Which ones?	0.05	Meets	0.15	No	0	No	0	Custom	0.05	Meets	0.15	No	0	No	0.00
44	External System Integration	What output formats does the software support?	0.05	Meets	0.15	Meets	0.15	Meets	0.15	Meets	0.15	Meets	0.15	Meets	0.15	Partial	0.10
51	Performance and Scalability	What is the maximum file size/number of records that the software can handle?	0.15	Meets	0.45	Meets	0.45	Meets	0.45	Meets	0.45	No	0	Meets	0.45	Meets	0.45
52	Performance and Scalability	What is the largest use case for the software to date?	0.15	Meets	0.45	Meets	0.45	Meets	0.45	Meets	0.45	Meets	0.45	Meets	0.45	Meets	0.45
56	Use cases, applications and future capabilities	Do you have any use cases, publications/white papers, demos or videos describing applications of your product? (please provide links or describe separately outside of this form as appropriate)?	0.05	Meets	0.15	Meets	0.15	Meets	0.15	No	0	Meets	0.15	Custom	0.05	Meets	0.15
57	Use cases, applications and future capabilities	Do you have any additional features planned or in development that you would be willing to share and feel we should know about?	0.05	Meets	0.15	Meets	0.15	Custom	0.05	Meets	0.15	Meets	0.15	Custom	0.05	Partial	0.10

4 Requirements Development

In order to evaluate candidate P3RLS software, it is important to first identify those functions and attributes that you wish for the software to provide. Direction from the IPT at the kickoff meeting was to focus the landscape survey strictly on P3RLS functionality. The scope therefore excluded broader examination of federated data management, integration, provenance and use.

4.1 Requirements Development Methodology

During the requirements capture process, the project team consulted a comprehensive set of sources and used a collaborative, iterative process to develop draft requirements, which were in turn reviewed by the IPT. Comments from the IPT were incorporated into a finalized requirements baseline.

The following sources were used in developing requirements:

P3RLS Landscape Analysis SOW: The Statement of Work contains information in the Requirements and Background sections as well as indirect, or high-level capabilities. Requirements capture was done from both sections with the assumption that any duplicates would be filtered out at the conclusion of the requirements gathering process.

SME Input: The Synectics team worked with an Integrated Project Team consisting of representatives from NCI Division of Cancer Control and Population Sciences (DCCPS), NCI Center for Bioinformatics and Information Technology (CBIIT), LBR and Synectics to identify potential requirements. Subject-matter expert (SME) input included both interactive sessions focused on NCI DCCPS use cases and a list of potential additional use cases provided by NCI CBIIT. Requirements were gathered through interviews with personnel from the North American Association of Central Cancer Registries (NAACCR) regarding cancer registry integration/de-duplication and from Information Management Systems regarding their current (non-privacy protecting) record linkage work, as well as from subject matter experts within the IPT.

Literature: A review of relevant literature to the patient privacy preserving identifier generation as well as record linkage were the primary functional drivers to the literature review. A combination of directed readings from the NCI Wiki pages as well as other sources related to these two functional drivers were conducted. This included review of a combination of research papers, product documentation, technical texts, and government reference materials related to the subject matter.

Product Review: Documentation of several functionally related products was reviewed to extract any potential requirements. The team was careful to generalize requirements so as not to bias the overall requirements towards any particular product.

4.2 Scope Boundaries

Figure 3 illustrates the Functional Scope of the requirements effort. It is not meant to be prescriptive

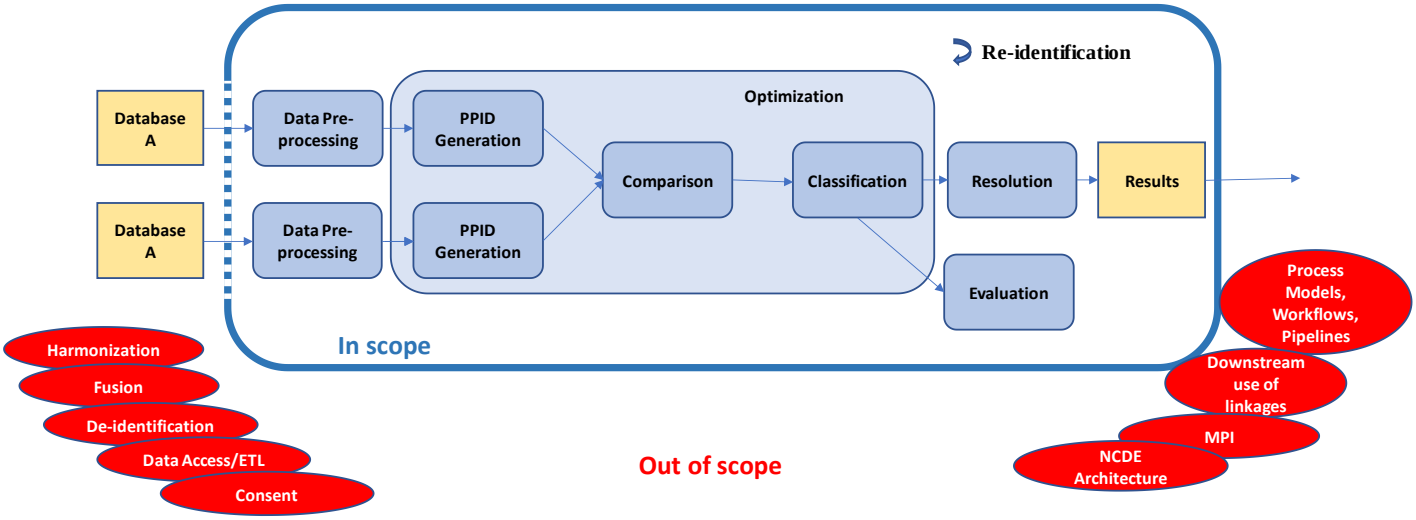

Figure 3: Functional Scope

of a process for P3RLS; rather, the figure, adapted from (Vatsalan, 2012), illustrates the boundaries of scope. The following functional areas were deemed to be in scope for the P3RLS Landscape Analysis:

- **Data pre-processing/cleansing** – Putting data into consistent formats and performing semantic clean-ups (e.g., phonetic spellings of names and substitution of nicknames) to enable more consistent generation of identifiers, improving matching performance
- **Privacy Protecting Identifier (PPID) Generation** – Generation of PPIDs from one or more combinations of input data fields
- **Optimization** – Indexing and other techniques to improve performance and scale
- **Comparison** – Comparison of PPIDs across multiple data sets to identify matches
- **Classification** – Pairwise characterization of data records into "match", "non-match" and "possible match"
- **Resolution** – Resolution of "possible" matches into matches or non-matches
- **Evaluation** – Review of matching results to quantify and assess matching performance
- **Re-identification** – Functions to prevent undesired re-identification from PPIDs and allow desired re-identification

The scope also incorporated non-functional characteristics including performance, scalability, reliability, usability, and other factors related to whether the software is appropriate for enterprise use within the National Cancer Data Ecosystem (NCDE) or similar environments.

The following areas were deemed to be out of scope for the P3RLS Landscape Analysis Project. This is due to a variety of reasons, including avoiding exploration of common database operations as well as avoiding duplication due to overlap with other NCI projects within the overall NCDE:

- **Data harmonization and data fusion** – In a federated environment, this is the responsibility of data stewards or researchers
- **De-identification** – In a federated environment, source data will not be de-identified
- **Data access/ETL** – This is a broader database/data processing concern
- **Consent** – Consent is generally independent of record linkage
- **Process models, workflows and pipelines** – Would be more in the scope of the overall NCDE or would be the province of researchers
- **Downstream use of linkages** – Responsibility or scope of individual researchers
- **Master Patient Index (MPI)** – This is a data integration / harmonization concern
- **NCDE Architecture** – Evaluation of P3RLS functionality is independent larger architecture characteristics

4.3 Additional Assumptions and Constraints

Before and during the requirements identification process, some conditions arose which had a potential impact of the scope of the analysis. Key additional assumptions and constraints included:

- P3RLS software should be able to be offered as a utility or service by NCI – and therefore, deployment models including software distribution as well as Software as a Service should be considered
- Downstream use of identified linkages should not be considered – and therefore, consideration of issues such as statistical disclosure risk (the increased re-identification risk arising from combination of data sets) should not be considered
- While case-level record linkage may be of interest to NCI for certain use cases, the focus of this effort should be patient-level linkage
- While generation of persistent encrypted Unique Identifiers (UIDs) either by patients or researchers may be of interest to NCI, the focus of this effort should be on record linkage rather than on persistent UIDs.

4.4 Requirements Categories

When performing the requirements capture process, requirements were grouped into categories. The categories used were adapted from the ISO/IEC 25010:2011 standard and considering the Sandia COTS Selection Process published in Sandia report SAND2006-0478. Categories support a structured analysis and were used to emphasize the most important categories in the evaluation process (this is described further in Section 7). The categories used for this evaluation include:

Functional Suitability: Describes whether the software performs the specified tasks and user objectives required for privacy protecting record linkage, and with what accuracy. [encompasses PPID generation and record linkage and data cleaning / pre-processing, and use cases/applications and future capabilities in the survey evaluation]

Interoperability/External System Integration: Describes the ability of the software to interface with other systems (such as the ability to ingest data). Also includes deployment considerations such as the ability to deploy to the cloud, and availability on different operating systems.

Performance and Scalability: Describes whether processing times and resources fall within specified constraints and whether the software scales to the required data sizes.

Security: Captures requirements for protection of information such as confidentiality, integrity, non-repudiation, and accountability.

Reliability/Support: Captures maturity both as a piece of the software (e.g., known bugs and execution issues), as well as from the product perspective (e.g., whether it is supported and whether there is adequate documentation available).

Usability: Describes the difficulty or ease for users to become proficient in using the software effectively, quality of the user interface, robustness for error handling, accessibility, installation and maintenance.

4.5 Requirements Prioritization

The set of requirements was intended to cast a wide net, that is, to cover both the essential functions of PPRL software and other desirable but not essential characteristics. This approach was designed to capture the breadth of capabilities available in the Customer off the Shelf and Government off the Shelf (COTS/GOTS) markets, with the expectation that various products considered would offer differing sets of capabilities. To keep focus on the essential requirements, the IPT assigned each requirement a priority from among the following: "Must Have", "Nice to Have", and "Could Have". These priorities were used in scoring the candidate products, as described further in Section 7.

The finalized set of requirements, including categorization and prioritization, can be found in Section 10 (Appendix 1: Requirements).

5 Evaluation Criteria Development

Establishing a set of *evaluation criteria* which can be applied in a uniform, standardized fashion to candidate software provides a foundation for meaningful and systematic comparisons and recommendations. Following finalization of requirements, a set of evaluation criteria were derived for use in the present landscape analysis, with additional criteria identified for use in a potential follow-on hands-on evaluation of candidate software.

5.1 Categories of Evaluation Criteria

Evaluation criteria are of two types: quantitative metrics, and qualitative criteria. Quantitative metrics are those for which numeric values can be measured or computed via analysis, such as the *precision* of a matching algorithm. In contrast, qualitative criteria are those which are evaluated via inspection or judgement, and lend themselves to discrete values, either binary ("Meets" vs. "Doesn't Meet") or multi-level. As an example of the latter category, evaluation of a product's reporting capability may not be quantitatively expressible via a metric but may be able to be

categorized based on its capabilities judged against categories such as: no reporting, a basic fixed set of reports, an extensive fixed set of reports, or customized reporting.

Quantitative metrics fall primarily into three categories: Linkage Quality, Scalability, and Privacy. These are described as:

Linkage Quality: Measures the correctness and completeness of identifying matching records between data sets. Metrics in this area are based on the completeness and correctness of the matches found. In practice, there is a trade-off between privacy and linkage quality, as stronger privacy protection can make it harder to identify matches between records. Linkage quality can be hard to assess with real data where ground truth is not known; however, synthetic data and analysis can be used to determine linkage quality metric values.

Scalability: Measures the ability of the software to handle large data sets, as well as the ability to execute efficiently so as to perform the linkage function within an acceptable runtime duration. Some applications which have been built for research or focused applications may not be able to handle large, real-world numbers of records. In general, the performance of an applications will be based on the complexity of its comparison algorithms as well as any optimizations such as blocking which reduce the number of pairwise comparisons to be performed.

Privacy: Measures of the risk of unintended disclosure of sensitive information. Within this category there are risks of identity disclosure (or reidentification), and attribute disclosure of one or more sensitive attributes of a record.

5.2 Methodology for Developing Evaluation Criteria

The methodology used in developing evaluation criteria used two approaches: (1) derivation of criteria from literature related to P3RLS and (2) spot-checking of published information about commercially available products. There is fairly good consensus in the record linkage literature and similar classification problems regarding Linkage Quality metrics. Likewise, similar scalability metrics are commonly found for a wide range of software applications. Link Quality and Scalability increase in both importance and difficulty with large data sets on the order of millions of records and terabytes of information. At such scales, manual verification of linkage quality and software performance become extremely difficult to perform.

Privacy metrics present a still more difficult challenge. Privacy metrics can be complex to select and assess as they involve assumptions about the attack model (e.g., an honest but curious insider attack vs. a hostile outsider attack), the amount of other information and resources the attacker has available, the type of attack used, and more. There are a wide range of viewpoints on privacy metrics, ranging from opinions that standard measures for Privacy Preserving Record Linkage are "still and immature aspect of the PPRL literature" (Vatsalan, et. al., 2017) to surveys (admittedly, not focused specifically on privacy-protecting record linkage) cataloging large numbers of potentially applicable metrics (see, for example, the 80 metrics cataloged by Wagner & Eckhoff, 2018). Unfortunately, many of these latter metrics are either not directly applicable or can only be evaluated either by theoretical analysis or by experimentally using extensive attempts to mimic attack vectors. A subset of these privacy metrics was selected for inclusion; future evaluators may want to expand this set as this area matures.

The finalized set of evaluation criteria, including traceability back to requirements, can be found in Section 11 (Appendix 2: Evaluation Criteria and Survey Questions).

5.3 Survey Development from Evaluation Criteria

Following review of the Evaluation Criteria by the IPT, the sturdy team formatted the landscape phase criteria into a format amenable to a survey to be sent to vendors. This included simplifying wording to facilitate comprehension, and rewording criteria into question format that could easily be answered. For example:

Requirement P-1: "*Shall operate on record sets up to tens of millions of records*" gave rise to four evaluation criteria as shown in Figure 4.

Evaluation Criteria	Definition
Run Time	How long a given application takes to perform linkage on a given data set using particular computing resources
Memory Space	The amount of memory that an application requires to run properly
Communication Size	The amount of data that passes through a communications network within the IT infrastructure
Reduction Ratio	How much an indexing technique is able to reduce the number of candidate record pairs that are being generated compared to all possible record pairs. A higher reduction ratio value means an indexing technique is more efficient in reducing the number of candidate record pairs that are being generated; however, link quality can be affected if cross-category matches are mixed

Figure 4: Evaluation Criteria generated from Requirement P-1

Not all of these criteria were appropriate for the landscape phase, and the various vendors were unlikely to have tested their code against an identical data set. Therefore, the following survey questions were formulated to elicit relevant information that we could expect the vendors to be able to provide:

Question	Text
51	What is the maximum file size/number of records that the software can handle?
52	What is the largest use case for the software to date?
53	What features does the software have (such as blocking or database indexing) to improve performance? Can the user specify blocking parameters?
54	Describe the ability to customize performance improvement features such as blocking?
55	How can performance be improved by adding computational power (e.g., elastic compute)?

Figure 5: Survey Questions Derived from Requirement P-1

The finalized survey questions provided to the vendors, including traceability back to requirements, can be found in Section 11 (Appendix 2: Evaluation Criteria and Survey Questions).

6 Candidate Software Identification

To assess the universe of possible PPRL solutions, the team first cast a wide net to gather as broad a set of record linkage products as possible, knowing that products on the initial list might satisfy only some of the major requirements. Guided by LBR/NCI's priorities, Synectics gathered product information by leveraging domain knowledge, reviewing relevant publications and

products, engaging in extensive internet searches using multiple search engines, and conducting interviews of experts.

The product search included COTS, GOTS, open source, and research products; however, to be include there had to be something that was clearly offered as a software product rather than a research capability. Both U.S. and non-U.S. products were considered.

In surveying the landscape of record linkage software, a wide selection of capabilities was found, addressing several business problem domains. Few of these products offer privacy-protecting features. Record linkage software clusters in the following problem domains:

Health: Primarily focused on electronic health record (EHR) linkage and health information exchange (HIE) via master patient indices (MPI), for goals such as care coordination and data de-duplication. Also includes some focus on life sciences research.

Financial: Focus on fraud detection, including areas such as Know Your Customer (KYC), which is the process of businesses verifying their clients' identities and assessing risks of illegal intentions, anti-money laundering (AML), focused on preventing criminals from disguising illegally obtained funds as legitimate income, benefits fraud, and other areas where obscure or ambiguous identity could be used for fraudulent purposes. In this domain, the term used is "entity resolution."

Marketing and Intelligence: While these are two distinctly different applications, both are focused on characterizing an individual via a web of personal, product, organizational and belief relationships to which they're connected. These focus mostly on linking unstructured data, with disambiguation of identity (record linkage) being a secondary focus.

The result of this initial step was a master list of 52 products. The team then examined the websites and other published material for each of the products looking for mention of privacy-protecting features. In general, products focused on financial and marketing domains operate only in the clear, and government-focused intelligence products operate in secure enclaves where privacy protection at the record level is not a consideration. In general, privacy-protecting features were found only in products focused on the health domain as an application area.

This first-level investigation of the products significantly narrowed the field of candidates to be evaluated. Initially the list of candidates to be evaluated contained eleven products. During the period of the study of these companies combined through acquisition, one was determined to be out of business, and one product was discovered to have gone end-of-life, resulting in a final list of eight products which were taken forward into the detailed evaluation process described in Section 7.

7 Evaluation Process

7.1 Survey Development and Data Collection

Each candidate software product was evaluated through a structured process, which started with outreach to the vendor and continued with a 30-minute initial teleconference. During that call the team assessed the applicability of the product and asked the vendor's willingness to participate in

the full survey. All the vendors with whom we spoke agreed to participate in the full survey. Of those, all but GRHANITE completed and returned the survey.

7.2 Software Evaluation and Scoring

A structured survey was used to ensure that uniform information was collected from all candidate vendors, and that the vendors were scored in a uniform and unbiased way.

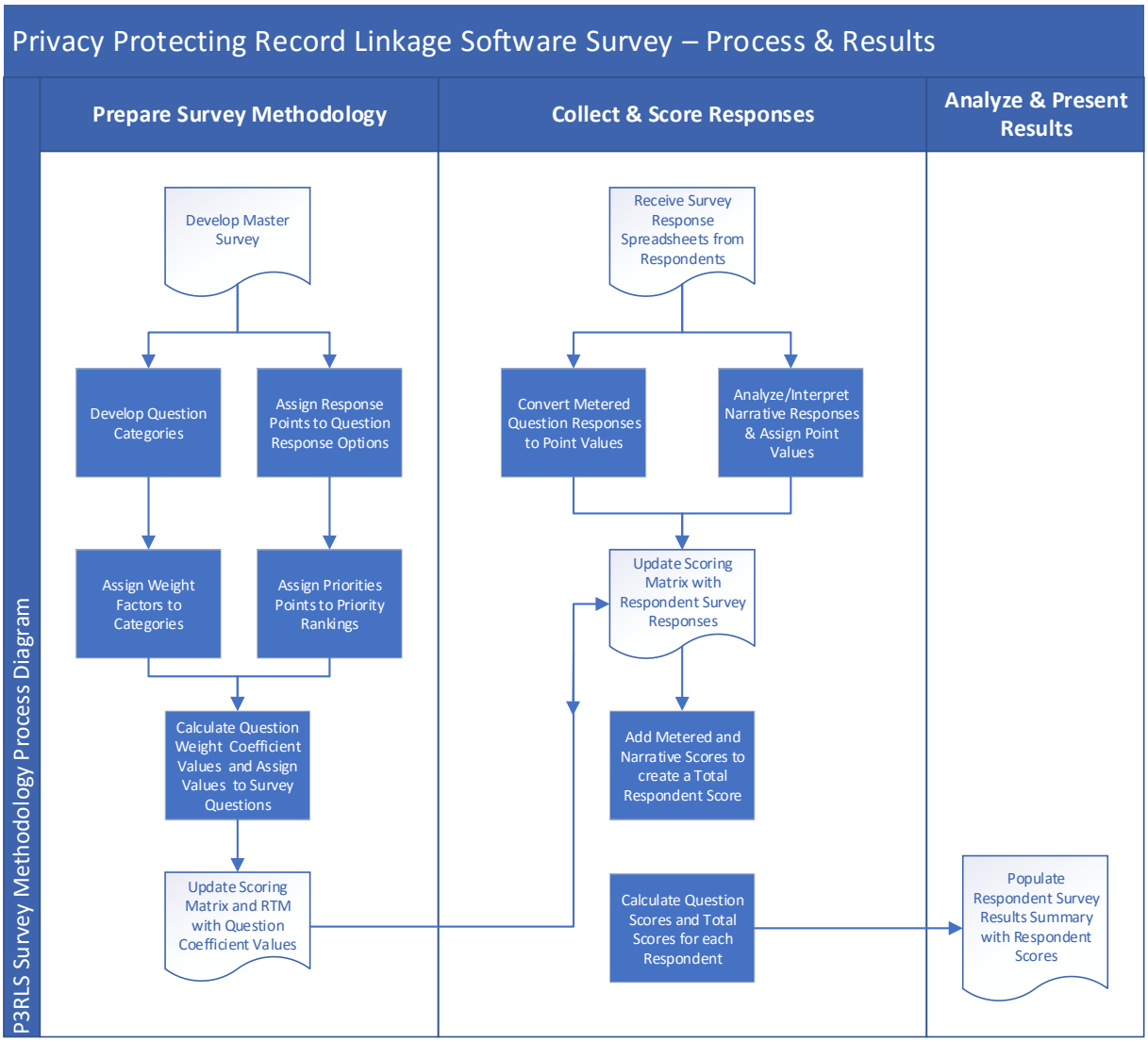

Figure 6: Survey Development Process

The steps in the survey methodology for evaluating and scoring software for the Privacy Preserving Patient Record Linkage Software (P3RLS) is as follows:

Assign Weights to **Question Categories** Based on Customer-provided Input.

There are 57 questions in the survey questionnaire that have each been attributed to one of seven defined categories, where each category refers to a topic area. The categories and the number of questions in each category are shown in Figure 7.

Question Category (Feature/Capability)	Total Questions
Usability and Security Features	18
PPID Generation and Record Linkage	15
Operating Environment and Licensing Model	7
Data Cleaning / Pre-Processing Features	5
External System Integration	5
Performance and Scalability	5
Use Cases, Applications and Future Capabilities	2
Grand Total	**57**

Figure 7: Question Category Weightings

Each category is weighted and assigned a level of emphasis per NCI's ranking of relative importance. The weighting values for the categories were developed in collaboration with NCI and FNLCR and were approved by the IPT. The question categories have been assigned weighting values as shown in Figure 8.

Question Category (Feature/Capability)	Category Weighting
PPID Generation and Record Linkage	50%
Performance and Scalability	15%
Usability and Security Features	15%
Data Cleaning / Pre-Processing Features	5%
External System Integration	5%
Operating Environment and Licensing Model	5%
Use Cases, Applications and Future Capabilities	5%
Grand Total	**100%**

Figure 8: Question Category Weighting Values

Assign Requirements Priority Values to Questions based on Customer-provided Input.

Each survey question is derived from a P3RLS requirement. In the course of the project, the IPT assigned each requirement one of the following a priority values: "Must Have", "Should Have", or "Could Have". Weighting each question by Requirement Priority assigns a level of emphasis to each category in keeping with NCI's ranking of the relative priority of the requirements.

The Requirements Priorities are assigned values for calculating a Weight Coefficient to be used in calculating a respondent's question response score. The values assigned to the Requirements Priorities are shown in Figure 9:

Requirement Priority	Points Rank
Must Have	3
Should Have	2
Could Have	1

Figure 9: Requirements Priority Values

Calculate and Assign to Each Question a Weight Coefficient Score

The Weight Coefficient value for each question is calculated by multiplying these two factors. Thus, if a question has a Weight of 50% and the Requirements Priority value is 3 ("Must Have"), the formula and resulting weight coefficient would be .5 x 3 = 1.5.

Identify the Scoring for each question.

Identifying the Types of Questions based on Response (metered or narrative responses). The questionnaire contains two types of questions:

Questions that permit a "metered response." A metered response is one in which users can express a discrete answer to a particular question, for example: "Does your software include a GUI?" The possible responses for metered questions are shown in Figure 10.

Question Response
Fully Meets
Partially Meets
Meets with Customization
Does Not Meet

Figure 10: Possible Metered Question Responses

Questions that require a "narrative response." A narrative response is one in which the answer cannot be answered via a discrete set of responses: What mechanism is used for producing a given result? Answers to narrative responses were captured in an Excel spreadsheet and were analyzed and interpreted by the team to determine whether the question had been answered in enough detail to quantify the question response, and if so, what score should be assigned.

Assigning Response Scores. Responses were given a numeric value. Figure 11 shows the values assigned to each question response/answer:

Question Response	Question Response (full)	Score
Meets	Fully Meets	3
Partial	Partially Meets	2
Custom	Meets with Customization	1
No	Does Not Meet	0
DNAQ	Respondent Did Not Answer the Question	0

Figure 11: Non-metered Question Responses

Calculate **Respondent Question Scores**

The Respondent Question Score were calculated by multiplying each question's weight coefficient value by the response score. Thus, if a particular question has a weight coefficient of 1.5 and the respondent's answer to the question is "3" (fully meets) then the formula for calculating the respondent's score for that question would be 1.5 x 3 = 4.5.

Calculating Respondent Total Score

Weighted question scores for the 57 questions were summed to provide the Respondent's Total Score.

8 Candidate Software Summaries

This section presents summaries for the final candidate software products.

8.1 CSIRO Anonlink

Product Name: Anonlink / CLKHash

Vendor: Commonwealth Scientific and Industrial Research Organization (Australian federal government agency, formed through the merger of predecessor government and private R&D labs). (www.csiro.au),

Summary: Anonlink is set of PPRL libraries. The software is written Python and C has been open sourced on GitHub. The software consists of the following components: CLKHSH, which encodes PII into cryptographic long-term key (CLK) hashes; Anonlink, which performs linkage on CLK hashes; and Entity-Service, a REST API for the system. Anonlink does not have a GUI.

Reviewer Comments: Powerful and flexible tool; however, it offers only the core of PPRL functionality and is not a user-friendly system. Would require development to create a polished, researcher-friendly tool. Does not perform data cleansing. The POC for GRHANITE pointed out that certain country-specific aspects of their tool (e.g., use of Australian postal codes) would have to be re-worked for U.S. use; this may be true for other non-U.S. software such as Anonlink as well.

Major Use Cases: State-to-Federal linkages of health data within Australia. Single digit millions of records per dataset. Goal is to each 10 million X 10 million comparisons without blocking.

Operating Environments: OSX, Windows, Linus, cloud-based using Kubernetes

Most Recent Release: June 2019

First Release: August 2017

Licensing Model: Open Source under Apache 2.0

Support Model: Through contracts to CSIRO.

Documentation: Thorough online documentation including tutorials.

U.S. Security Compliance: None

Additional Links:

- https://data.csiro.au/dap/landingpage?pid=csiro%3A26733
- Video of a technical talk on this software: https://2018.pycon-au.org/talks/44892-privacy-preserving-record-linkage/

- Code links:
    - https://github.com/n1analytics/clkhash/ - client-side library and command line tool for encoding PII into CLKs.
    - https://github.com/n1analytics/anonlink/ - server-side library for fast CLK comparisons and solvers.
    - https://github.com/n1analytics/entity-service/ - HTTP Rest service for record linkage of CLKs. Data61 hosts a demonstration version of the entity service including documentation and tutorials at https://anonlink.data61.xyz/
    - https://github.com/n1analytics/encoding-service/ - HTTP Rest service for encoding PII into CLKs.

8.2 Crossix SafeMine

Product Name: SafeMine

Vendor: Crossix (crossix.com)

Summary: Crossix is a U.S.-based company focused primarily on health marketing data and services for the pharmaceutical industry; however, they also license their PPRL software. Its software includes Tokenizer (linkage based on hashed tokens) as well as SafeMine, which extends this to include matching of identified data in a secure, HIPAA-compliant environment. Also, the system can use third-party data such as address history databases to enhance matching.

Reviewer Comments: Because of its focus on marketing, SafeMine has features focused on building marketing profiles via record linkage (e.g., identifying members of the same household, finding providers likely to be interested in prescribing a particular medication). These extra features may not be of use and may add un-needed complexity. However, the record linkage engine appears robust and well-proven.

Major Use Cases: "Tens of billions of records"

Operating Environments: Java-based, runs on Windows and Linux. Requires MySQL. Deployed to AWS.

Most Recent Release: June 2019

First Release: 2006

Licensing Model: One-time deployment fee per node, plus annual licensing fee. "Flexible"

Support Model: Full commercial support

Documentation: Full commercial documentation

U.S. Security Compliance: HIPAA compliant, HITRUST and NIST certified

Additional Links:

- Product video at https://crossix.com/the-crossix-difference/

8.3 Datavant

Product Name: Datavant

Vendor: Datavant (Datavant.com)

Summary: Datavant offers a set of modules focused on record linkage for health marketing, health data analytics, care management, payers and life sciences. Datavant has acquired Health Data Link, Universal Patient Key, and Prognos's OPAL de-identification, bolstering its capabilities.

Reviewer Comments: Capability appears well matched to NCI requirements. Mostly command line modules – lack of a GUI may be a hurdle to end-users. Software runs on site, but communicates back to Datavant servers for encryption keys, user licensing verification, etc. so may need firewall configuration to run.

Major Use Cases: Processing data sets of over a billion records. For a consumer-focused healthcare website, seven databases are linked, with over 5 billion healthcare transactions with information on nearly 190 million Americans. Has worked with IMSWeb on SEER, ORIEN Cancerlinq.

Operating Environments: Windows 10, Windows Server 2016 and later, Ubuntu Linux. Has been deployed to AWS, Azure, and Google cloud platforms.

Most Recent Release: May 2019

First Release: October 2014

Licensing Model: Annual master license fee, and per partner fee (hub and spoke model). "Flexible"

Support Model: Full commercial support

Documentation: Full commercial documentation

U.S. Security Compliance: SOC 2 Type 2.

Additional Links:

- The following documents were provided:
  - Token Selection Deep Dive
  - User Guide
  - Matching Accuracy of Tokens in De-identified Health Data Set
  - Datavant Overview Deck – highlights of tokenization, linkage, ecosystem
  - Datavant Key Use Case Highlights
  - Summary Technical Requirements
  - Security Compliance Overview

8.4 Privitar Securelink

Product Name: Securelink

Vendor: Privitar (www.privitar.com/securelink)

Summary: Privitar is a UK-based company focused on enterprise applications for privacy protection (they have a U.S. presence). Securelink is a part of a suite of programs all focused on privacy protection. The latest version is integrated with its publisher tool which provides de-identification, governance, and user-friendly interfaces.

Reviewer Comments: Privitar's tool appears to have significant capability; however, its primary use has been in the UK health system which has a unique personal identifier which can be used for linkage. In conversations the company has stressed the need for data sets to have the same identifiers.

Privitar's expressed process model involves four parties: the two (or more) data owners/stewards, an intermediary to do additional encryption, and the trusted party for the comparisons.

Doesn't do cleansing or pre-processing. Can "watermark" data for additional security.

In developmental version, have implemented homomorphic encryption.

Major Use Cases: UK National Health System, 50 TB data of patient claims files. Other customers include Anthem, HSBC and BT.

Operating Environments: Linux (requires Oracle, MySQL or HDFS). Has been deployed to AWS and Azure.

Most Recent Release: July 2019.

First Release: 2018

Licensing Model: "Operations Based"

Support Model: Full commercial

Documentation: Full commercial

U.S. Security Compliance: Approved by UK government for national health system data. No U.S. approvals.

Additional Links:

- None

8.5 Senzing

Product Name: Senzing

Vendor: Senzing (www.senzing.com)

Summary: Senzing is a record linkage / entity resolution software with a long lineage. In 2005 IBM acquired a start-up company with entity resolution technology, which formed the basis for the IBM InfoSphere Identity Insight product[1] and spawned further IBM internal research on privacy-protecting linkage. In 2016 that same team was spun back out of IBM to create Senzing.

The product has many features for in-the-clear resolution (including AI-based similarity recognition between records) that are not of interest for PPRL. Its PPRL seems robust and, as mentioned above, is long-established. Data can be multiply hashed by multiple parties to improve security.

Reviewer Comments: Senzing responded that the product is primarily a set of APIs (available in C, Python or Java), but that there are third party GUI's with which it integrates. The company's web site does describe a GUI, but it may be for a more limited desktop version.

Major Use Cases: 3 billion identity records

Operating Environments: CentOS 7 x86_64, RedHat 7 x86_64, Debian 9 / Ubuntu 16.04 x86_64, Amazon Linux 2016 x86_64 (requires RDBMS: IBM Db2, SQLite, PostgreSQL, MySQL / MariaDB - 5.6.5 / 10.1, or AWS RDS). Cloud based implementation via Docker, Kubernetes, Rancher, Helm, and others.

Most Recent Release: June 2019

First Release: 2012

Licensing Model: "Per record ingested into the database"

Support Model: Bundled with licensing. Support services available. See https://senzing.zendesk.com/hc/en-us/articles/236071408-Support-Services

Documentation: https://sensing.com/developer, http://docs.senzing.com

U.S. Security Compliance: None

Additional Links:

- https://senzing.com/wp-content/uploads/Uniquely-Senzing-WP-042319.pdf
- https://senzing.zendesk.com/hc/en-us/articles/231726307-Principle-based-Entity-Resolution
- Entity Resolution in Slow Motion: (https://www.youtube.com/watch?v=MPHd1eqU_yo)

[1] IBM Identity Insight was initially identified as a candidate for this study, but the product has gone end of life.

- Privacy By Design: (https://jeffjonas.typepad.com/jeff_jonas/2012/06/privacy-by-design-in-the-era-of-big-data.html)
- Senzing Demo: (https://www.youtube.com/watch?v=O7oLUnWet8w)
- Jeff Jonas introducing Senzing: (https://www.linkedin.com/pulse/meet-senzing-g2-say-hello-entity-resolution-20-jeff-jonas/)
- Semantic Reconciliation - Entity Centric Learning: (https://jeffjonas.typepad.com/jeff_jonas/2007/04/to_know_semanti.html)
- Sequence Neutrality: (https://senzing.com/sequence-neutrality/)

8.6 University of Melbourne GRHANITE

Product Name: GRHANITE

Vendor: University of Melbourne, Australia (https://www.grhanite.com/)

Summary: GRHANITE was developed by the University of Melbourne Health and Bioinformatics Centre and provides privacy-protecting record linkage, consent management, and information aggregation and routing.

Reviewer Comments: *GRHANITE participated in an initial screening interview but did not return the full survey. The product was therefore not included in the ranking*. The POC pointed out that a number of aspects of GRHANITE, such as use of postal codes, Australian Medicare IDs and nickname files would have to be modified for U.S. use.

Major Use Cases: 17 million Australian health records

Operating Environments: MS Windows (Windows XP SP3 onwards (Windows 2003 Server, Vista, Windows 7, Windows Server 2008, Windows 8, Windows Server 2012, Windows 10 Pro)

Most Recent Release: Latest release listed on the site is 2015, but per the POC, "We have a significant new phase of development starting in Q3 this year aiming to re-vamp the algorithms based on the latest research."

First Release: Unknown

Licensing Model: Freely available in Australia. Agreement would need to be worked out for U.S. use.

Support Model: Support agreement with vendor.

Documentation: Unknown

U.S. Security Compliance: None

Additional Links: None

8.7 Linkwise Policywise

Product Name: Linkwise

Vendor: Policywise, Alberta, Canada. (https://Policywise.com/2018/03/15/linkwise/)

Summary: Linkwise is a privacy-protecting data linkage software product created by Policywise for Children & Families, a Canadian non-profit focused on ensuring the well-being of families and children. The product was developed after they became unhappy with the record linkage capabilities of a COTS product. Typical use cases have been in the thousands of records, though the software has been tested to "low millions" of records.

Reviewer Comments: Linkwise appears to be a modern piece of software that incorporates current art on record linkage (e.g., the use of Bloom filters). However, Policywise is primarily a policy organization and in speaking with them one gets the idea that Linkwise was a sideline effort which is not a major focus of the organization. It appears development was outsourced; the Policywise POC wasn't clear on certain details and in the survey, response answered a number of questions with "I might have to ask the programmer." Development is "sporadic."

Major Use Cases: None provided.

Operating Environments: MS Windows

Most Recent Release: 2018

First Release: 2018

Licensing Model: Primarily they have been providing linkage as a service via contract; however, they are open to flexible models.

Support Model: Via contract with Policywise. "Policywise staff offer help in using the software."

Documentation: Limited

U.S. Security Compliance: None

Additional Links: None

8.8 HealthVerity Census

Product Name: Census

Vendor: HealthVerity (www.healthverity.com)

Summary: HealthVerity Census is part of a product line aimed at creating a marketplace of linked data sets (managed by the companion Marketplace product) within the HealthVerity environment. The product family includes a robust feature set for linking and managing linked data. The company provides Java-based software that clients can use to generate privacy-protecting hashed identifiers in their own environments (PPID generation can also be accomplished in a secure cloud environment, uploading or calling an API with PII). The hashed

IDs are further encrypted and sent to HealthVerity, which performs linkage in the company's environment, returning linkage results that include the ID's of matched records.

HealthVerity does not match data sources pairwise. Rather, it maintains a persistent index of "HealthVerity IDs" (HVIDs) corresponding to the incoming hashes. When a new data set's hashed identifiers are submitted to HealthVerity the new data is linked against existing known HVIDs using an error-tolerant Bloom filter approach. For each match found, the associated data record is associated with that HVID; if a match is not found, a new HVID is added to the master index. HealthVerity states that their persistent index of HVIDs with linkages to data sets contains 330 million distinct patient IDs, roughly equal in size to the entire population of the United States. This is consistent with their goal of creating a marketplace of data where researchers can come to "shop" for data sets. The HVIDs are based on encrypted versions of hashed IDs, and so do not directly contain PII.

HealthVerity's data market environment has a friendly web-based user interface. Linkage occurs quickly and users can query for data sets via a variety of parameters such as diagnosis, medication, and data type. While the default is to coalesce all linked data into a single data marketplace, individual clients can choose to have their own data space set off from the larger data market. HealthVerity also links associated demographic data about patients from related non-health data sets.

Reviewer Comments: The product appears mature, feature rich, and includes well thought out PPID generation and linkage. In order to make their approach of matching the same IDs against data sets over time work properly, PPID generation has to be fixed, with HealthVerity specifying exactly what goes into hashes of data fields. The HealthVerity Marketplace offers many features applicable to the NCDE, but using HealthVerity locks you into a vendor environment, as all linkages are performed and results managed by HealthVerity.

In surveying products we had come across two vendors, Verato and Occam, which used a similar approach to matching against persistent patient identities (neither of these companies offered privacy-protected matching). Occam creates their baseline patient identities from well-sourced external ground truth from LexisNexis records, while Verato created their own master index from a variety of sources. In contrast, HealthVerity creates a baseline patient identity from the first data point encountered for an individual; it is not clear if there is a cascading linkage impact if that first record contains errors.

Major Use Cases: "HealthVerity de-identification and matching is used by over 100 entities; over 50 billion records processed"

Operating Environments: For on-premise use (PPID generation), HealthVerity Census requires Java Runtime Environment (JRE) v1.8 plus Java Cryptography Extension. HealthVerity also has a secure cloud-based version and API available. Linkage runs in HealthVerity's environment.

Most Recent Release: December 2019

First Release: 2015

Licensing Model: Per configuration + per file processed

Support Model: Documentation, support via HealthVerity deployment engineer

Documentation: Configuration and access documentation

U.S. Security Compliance: HIPAA

Additional Links:

- The following documents were provided:
  - HealthVerity Census Description Whitepaper
  - HealthVerity Overview
  - Matching Accuracy Metrics Whitepaper

8.9 Other Product Outreach

Outreach was made to several additional companies based on the recommendations by the IPT, though these vendors did not appear from their publicly facing web presence to have PPRL products. They included:

IQVIA – Outreach via company "Contact Us" page and directly to Jeffrey Clark, a POC provided by the IPT. No response.

Privacy Analytics – Outreach via email to Dr. Khaled El Emam. No response.

Georgetown University ATRA – Georgetown has developed a high assurance trusted computing environment in which record linkage could be performed in the clear with little risk of PII disclosure. Privacy protection comes from the high security of the computing environment rather than from masking identity via PPIDs. While ATRA has successfully been used by NIAID and CDC for disease surveillance, the approach was not deemed applicable to the provided use cases.

In addition, we initially identified IBM as a potential vendor. We subsequently determined that the primary product of interest had gone end-of-life, and that the IBM product team had been spun out as Senzing. IBM has another record linkage product called Watson Financial Crimes Insight, but it didn't appear to be privacy-protecting; many products in the area of anti-money laundering are not.

9 Future Considerations

This section contains additional observations and considerations for subsequent evaluation of PPRL products.

- **Feature Sets.** The finalist candidate products have differing feature sets. For example, not all products do nickname substitution (e.g., "Alex" to "Alexander"). Care must be taken during the pilot phase to ensure a level playing field, e.g., to make sure that equivalent nickname substitution pre-processing is performed for products which do not perform this function.
- **End-to-End Process.** Similarly, PPRL product(s) ultimately selected may implement only part of the end-to-end process (e.g., data extraction tool, surname phonetic encoding tool,

address standardization tool, etc.). LBR and NCI may want to enumerate the complete set of functions involved in the record linkage process, determine which will be provided as a tool by NCI and which are assumed to be the responsibility of the user, and perform similar landscape analyses to identify how to provide end-to-end functional coverage of the process.

- **Related Capabilities.** The landscape analysis identified products which, while not PPRL tools, provide notable relevant capabilities. For example, both Verato and Occam compare source records against established identity databases such as LexisNexis®. The use of an intermediate "ground truth" for identity or address resolution, or application or other data quality algorithms in advance of PPID generation and record linkage could help performance. Likewise, there exists a selection of address resolution tools that could be applied to input data.
- **Performance.** Performance comparisons should be executed suing the same computational environments. Ideally, both local and cloud-based environments should be tested, and scale of testing should approximate operational data sizes.
- **Testing Protocol Design.** Testing design should consider how to capture qualitative aspects of the software such as ease-of use. Testing design should also make sure to capture unplanned and infrequent, yet important, characteristics such as system crashes. Testing should include a range of data sets (e.g., both registries and clinical trials) and should, if possible include an in-the-clear linkage as a "gold standard" against which to compare PPRL. If there is a desire to test sensitivity to certain conditions (e.g., linkage quality among populations a small number of very prevalent surnames), synthetic data should be considered.
- **Adaptation for U.S. Use.** The final product list includes software from the U.S., Canada, England, and Australia. In considering non-U.S. software, care should be taken to consider required adaptation for U.S. use (e.g., different nickname files) as well as subtle assumptions in algorithm design that could affect linkage performance (e.g., differences in granularity of postal codes in different countries).
- **Case-By-Case vs. Persistent Linkage**. The survey team recognizes that notions of a Master Patient Index (MPI) were out of scope and that each linkage of records should be considered as starting from a clean slate. In practice, linkages may be done on a recurring basis and NCI may want to examine the possibility of persisting the results of linkages to facilitate subsequent linkage, perhaps using emerging secure technology such as blockchain. It is recognized that the computational gains from maintaining an MPI would need to be balanced against the security and privacy considerations of such an index.

10 Appendix 1: Requirements

Requirements for Privacy Protecting Record Linkage Software

Requirement ID	Requirement Category	Requirement	Consensus Prioritization	Requirements Comments
F-1	Functional Suitability	PPID Generation: Product generates privacy-protecting identifiers from source data	Must Have	Basic function: generates privacy-protecting identifier. Note techniques used (e.g., secure hashing, secure multi-party computation)
F-2	Functional Suitability	Linkage: Product can compare PPIDs from two (or more) sources and identify matches, non-matches, and possible matches	Must Have	Basic function: data linkage
F-3	Functional Suitability	Supports two-party linkage	Should Have	Two parties work without a trusted third party to identify linkages
F-4	Functional Suitability	Supports multi-party linkage (e.g., with an honest broker)	Should Have	One or more trusted third parties are used to resolve "possible" matches.
F-5	Functional Suitability	Supports linkage of more than two data sets	Should Have	Optimizations for identifying linkages across three or more data sets
F-6	Functional Suitability	Supports human-assisted classification adjudication via features such as masking and distance measures	Could Have	Functions for resolving "possible" matches without fully revealing source PII to the human performing resolution. Useful, but not a priority for this application
F-7	Functional Suitability	Supports manual review of results and resolution of "possible" linkages	Must Have	Allows a workflow which allows linkage results to be manually reviewed and resolved
F-8	Functional Suitability	Provides metrics of classification performance	Must Have	Since the classification software doesn't know ground truth, metrics can reflect only what's observable (e.g., number of linkages found) but not performance (e.g., precision)
F-9	Functional Suitability	Provides flexible PPID comparison techniques	Could Have	Examples include exact and approximate equality, q-gram comparisons, Bloom filters, distance metrics. An individual piece of software may not be required to have more than one - particularly if one technique is found to be a "gold standard"
F-10	Functional Suitability	Provides flexible linkage classification techniques	Could Have	Examples include ability to consider multiple ID comparisons, probabilistic matching, tunable matching criteria
F-11	Functional Suitability	Ability to configure usage of a range of input fields (PII and others) from source schemata to generate identifiers	Must Have	e.g., generating IDs using various combinations of Safe Harbor fields
F-12	Functional Suitability	Product is extensible to incorporate user-supplied data pre-processing (e.g., language-specific phonetic encoding of names)	Could Have	Pre-processing enables greater match accuracy and fewer errors and manual verification. Extensibility allows domain and data-specific functionality to be added by the user

Requirements for Privacy Protecting Record Linkage Software

Requirement ID	Requirement Category	Requirement	Consensus Prioritization	Requirements Comments
F-13	Functional Suitability	Tolerance of minor data inconsistencies via ability to do data cleaning prior to generation of identifiers.	Should Have	Not a "Must Have" but highly desireable. The goal is to fix noisy, incomplete or inconsistent data. This can be simple formatting such as making sure SSNs are in the right format and have valid values, and/or more semantic-based actions such as substituting nicknames ("Joe" for "Joseph") or fixing common spelling errors
F-14	Functional Suitability	Ability to perform data field translation and transformation prior to generation of identifiers	Could Have	Examples include ability to translate between representations used in heterogeneous databases, such as translating "1" to "M" and "2" to "F", phonetic encoding (Soundex), binning k-anonymization, etc.
F-15	Functional Suitability	Allows for splitting and merging of match results files	Should Have	Helps in manual review of linkage results. For example, subset by hi probability vs. low probability matches - this is currently used in registry integration
F-16	Functional Suitability	Supports user-configurable comparison/classification parameters (e.g., to tune recall and precision performance)	Must Have	Manual configuration of parameters supports match optimization
F-17	Functional Suitability	Generates reports	Must Have	Output metrics, performance statistics, etc. to optimize performance for the application, facilitate review of results and quality assurance
F-18	Functional Suitability	Is able to detect duplicate records	Should Have	Useful for identifying duplicate records within or across record sets. For example, a patient linked via PPID based on PII, but has two different cancers in different records. Could be valid primary/secondary of the same patient or two different patients.
F-19	Functional Suitability	Product enables authorized re-identification	Must Have	Where appropriate and authorized, allow PPIDs to be reversed to reveal true identities
F-21	Functional Suitability	Ability to persist, internally or externally, identifiers generated from various PII combinations	Could Have	Storage of linkage results. This effort is not implementing a Master Patient Index so this is not required, but could have utility at some point.
F-24	Functional Suitability	Ability to persist, internally or externally, linkages between data sets	Could Have	Storage of linkage results
I-1	Interoperability	Shall operate on commonly available cloud or on-premise infrastructure	Should Have	Needs to operate in environments of accessible to the community
I-2	Interoperability	Supports multiple data input formats including variable-length, fixed length, ODBC/JDBC, XML, JSON and APIs for input files	Could Have	Broad range of input formats for input and linkage

Requirements for Privacy Protecting Record Linkage Software

Requirement ID	Requirement Category	Requirement	Consensus Prioritization	Requirements Comments
I-3	Interoperability	Product produces results via one or more well-defined output formats (e.g., CSV, XML, JSON) or APIs suitable for persistence, transmission or offline storage	Could Have	Broad range of output results for reporting and future interoperability within NCDE and/or other environments
I-4	Interoperability	Product is flexible in interfacing with databases (relational, NoSQL, etc.)	Could Have	Note: This analysis considers only use of structured data. Free text, imagery, genomic data are out of scope.
P-1	Performance	Shall operate on record sets up to tens of millions of records	Must Have	Note stated scale and performance curves
P-2	Performance	Performance shall be able to be improved via addition of computational resources (e.g., adding memory or processors, or elastic compute)	Could Have	Requirement provided by DCCPS to scale up via adding compute resources
P-3	Performance	Contains explicit optimizations such as indexing and blocking	Should Have	Reduces the number of record comparisons to be done Note types of optimizations implemented
P-4	Performance	Performance optimizations such as indexing and blocking are tunable by researchers	Could Have	For optimization of performance. Such tuning may be more in the realm of information scientists rather than researchers.
P-5	Performance	Product contains optimizations for re-generation of identifiers and/or reclassification based on source database updates	Could Have	Recognizing that source data will change over time, can an incremental update be done rather than fully re-running the ID generation and matching?
P-7	Performance	Solution will run on FedRamp or GovCloud compliant host	Could Have	Such environments may be required
S-1	Security	System does not persist source data	Could Have	Systems should be memoryless to protect source PII
S-2	Security	Product incorporates features to reduce risk of unauthorized re-identification	Must Have	Incorporates techniques such as salted hashes, chaffing, decoupling, etc. to thwart common attacks such as dictionary, frequency, etc. as well as statistical disclosure control
S-3	Security	Source information (PII) is well protected	Must Have	System has adequate data protections
S-4	Security	Product has been used in an application compliant with privacy and information protection regulations (HIPAA, FISMA)	Could Have	May be required to comply with industry and government standard security policies and mandates.
S-5	Security	System has appropriate user access controls and segregation	Must Have	Secure access management is provided, whether the system is deployed on premise, in the cloud or a hybrid of the two
S-6	Security	Ability to use state-of-the-art encryption	Must Have	Examples of secure techniques include SHA-2 at 512 bits, SHA-1, MD-5

Requirements for Privacy Protecting Record Linkage Software

Requirement ID	Requirement Category	Requirement	Consensus Prioritization	Requirements Comments
R-1	Reliability (Support)	Product can be made available as a service or tool for the research community	Must Have	Can be packaged for distribution to and remote use by data stewards, such as was done in NAACCR registry integration project, or can be made available via the cloud
R-2	Reliability (Support)	Technical support is available for the product	Must Have	Technical support will be available for the product whether it is commercial or open source
R-3	Reliability (Support)	Product licensing supports use as a tool for the research community	Must Have	Does the product have a licensing scheme which supports widespread use
R-4	Reliability (Support)	Product includes extensive inline and offline documentation	Should Have	Inline documentation is built into the software and can be accessed during the execution of the product. Offline documentation consists of documents in formats such as docx, html and pdf that are accessed outside the boundaries of the product.
R-5	Reliability (Support)	Has been demonstrated on applications of similar scale and complexity	Should Have	Note whether product has been demonstrated on real and/or synthetic data
R-6	Reliability (Support)	Vendor or Developer Community is stable with long-term support	Should Have	Avoid "one off" and potentially orphaned software
U-1	Usability	Ability to store/recall previous PII hash combinations		Saved parameters for program use
U-2	Usability	Automation: can be scripted to perform operations automatically or driven by configuration files	Should Have	Configuration files will allow for system execution of processes that normally may require human interaction
U-3	Usability	Product navigation is performed with minimal number of clicks	Should Have	
U-4	Usability	Product is "easy to use" via a graphical user interface and/or batch configuration	Should Have	GUIs are standard user interfaces; configuration files are good for batch execution
U-5	Usability	Results are easy to understand for end-users without significant post-processing	Should Have	Avoid complex data structures, machine-unreadable PDF formats, complex unique formats, etc.
U-6	Usability	User Interface segregates by roles	Could Have	Software will be easier to use if it presents UI's tuned to the needs of users (e.g., Data Contributors vs. Scientists performing linkage vs. Users of the results) rather than having only one set of screens.
U-7	Usability	Software allows editing of the data within the program	Could Have	Can you edit the data in the program, rather than having a cycle of edit/re-load

Requirements for Privacy Protecting Record Linkage Software

Requirement ID	Requirement Category	Requirement	Consensus Prioritization	Requirements Comments
U-8	Usability	Software provides a preview of the data (input data and/or linkage results)	Should Have	As with the previous requirements, supports efficient workflows vs. having to edit and re-load data.
U-9	Usability	Software supports an evaluation mode vs. linkage mode	Could Have	Allows researchers to work with their own data sets to tune parameters before performing linkage.

11 Appendix 2: Evaluation Criteria and Survey Questions

11.1 Evaluation Criteria

Evaluation Criteria ID	Evaluation Criteria Category	Evaluation Criteria	Definition
EC-P-1	Performance/Scalability	Run Time	How long a given application takes to perform linkage on a given data set using particular computing resources
EC-P-2	Performance/Scalability	Memory Space	The amount of memory that an application requires to run properly
EC-P-3	Performance/Scalability	Communication Size	The amount of data that passes through a communications network within the IT infrastructure
EC-P-4	Performance/Scalability	Reduction Ratio	How much an indexing technique is able to reduce the number of candidate record pairs that are being generated compared to all possible record pairs. A higher reduction ratio value means an indexing technique is more efficient in reducing the number of candidate record pairs that are being generated; however, link quality can be affected if cross-category matches are mixed
EC-LQ-1	Link Quality	Precision/Positive Predictive Value	The fraction of record pairs classified as matches by a decision model that are true matches. Computed as the number of True Matches TM divided by the sum of true matches and false matches (FM) = TM/(TM+FM)
EC-LQ-2	Link Quality	Recall/Sensitivity	The fraction of true matches that are correctly classified as matches. Computed as true matches divided by True Matches plus False Negatives (FN) = TM/(TM+FN)
EC-LQ-3	Link Quality	F-measure	Combines precision and recall into a single metric = 2 x (Precision x Recall)/(Precision + Recall)
EC-LQ-4	Link Quality	Pair Completeness	Measures the effectiveness of an indexing technique in the record linkage process. Computed as the number of true matches correctly placed by blocking (BM) divided by the true matches plus false non-matches = BM/(TM+FN)
EC-LQ-5	Link Quality	Pair Quality	Measures the efficiency of a blocking technique. Similar to recall. Computed as true matches correctly placed by blocking (BM) divided by the sum of BM and true non-matches = BM/(BM+BN)
EC-LQ-6	Link Quality	Accuracy	The fraction of record pairs correctly classified = (TM + TN)/(TM+FM+TN+FN)
EC-LQ-7	Link Quality	Specificity	The fraction of true non-matches that are correctly classified as non-matches = TN/(TN+FN)
EC-S-1	Security/Privacy	Disclosure Risk	DR is the probability that masked records/QID values can be reidentified by being linked with records or values in a publicly available dataset, or by attacks such as dictionary attack. There are several variants of this metric.

Evaluation Criteria ID	Evaluation Criteria Category	Evaluation Criteria	Definition
EC-S-2	Security/Privacy	Uncertainty: Degree of Unlinkability	Unlinkability measures the adversary's uncertainty about which items are related. This can be a measure, for example, of salt effectiveness.
EC-S-3	Security/Privacy	Time: Time Until Adversary's Success	This time-based metric and assumes that the adversary will eventually succeed - or can be used as a threshold value where the software is acceptable if the time to succeed is greater than a given time value *t*.

11.2 Landscape Analysis Survey Questions

This section contains the survey questions distributed to the vendors in the course of the landscape analysis. It identifies which requirement each question relates to (requirements traceability via the "Source Requirement" column) and the priority of that requirement via the "Requirement Priority" column.

ID	Question	Source Requirement	Priority
	PPID Generation and Record Linkage		
1	What can the user specify about the PPID generation process (e.g., which variables go into the PPID, and whether they should be truncated)?	F-1, F-11	01 - Must Have
2	Can the software generate and link on multiple PPIDs (e.g., hashes composed from different concatenated input variable combinations) either as a single pass or multiple passes? Is there a maximum number?	F-10	03 - Could Have
3	What is the mechanism for generating PPIDs (e.g., SHA-2 hashing)?	S-6	01 - Must Have
4	What probabilistic matching capability is available (e.g., Bloom filters on q-grams)?	F-9, F-10	03 - Could Have
5	How many files can be simultaneously linked (e.g., can the software link more than two files in one pass)?	F-5	02 - Should Have
6	Does the product support deduplication?	F-18	02 - Should Have
7	Can the software support more than pairwise linkages (e.g., find all the records in a file that match)?	F-2	03 - Could Have
8	Does the product support two-party privacy protected linkage (identifying linkages between two files potentially belonging to two owners, where neither party sees the other's unencrypted data)?	F-3	02 - Should Have
9	Does the product support three-party protect linkage, where there is a trusted "honest broker" able to resolve possible linkages?	F-4	02 - Should Have
10	For three-party linkage, what information is made available to the trusted broker to resolve "possible" linkages (e.g., reports, distance metrics, comparisons of source data - if the broker can see source data)?	F-6, F-7	01 - Must Have
11	Are there any features for authorized reidentification of data?	F-19	01 - Must Have
12	What is tunable about matching criteria/algorithm?	F-16	01 - Must Have
13	Does the software have any ability to persist results so that subsequent linkages between data sets (e.g., after updates) can be incremental rather than from scratch?	F-24	03 - Could Have
14	Does the software have the ability to split databases into linked vs. non-linked records, or other splitting and merging capability?	F-15	02 - Should Have
15	Can the product persist PPIDs so they don't have to be regenerated for future runs?	F-21	03 - Could Have
	Operating Environment and Licensing Model		
16	What Platform/OS(s) does the system run under?	I-1	02 - Should Have
17	What other software is required to run your software (e.g., DBMS)?	I-1	04 - N/A
18	Minimum hardware specification.	I-1	04 - N/A
19	Cloud-based version available? If so, which cloud environment?	I-1, P-7	02 - Should Have
20	Licensing model (per seat, per CPU, open source, etc.).	R-3	01 - Must Have
21	Does your licensing support "record linkage as a service", either through offering a cloud-based service or by distribution of the software as a utility?	R-1	03 - Could Have

ID	Question	Source Requirement	Priority
22	Is the system a set of components or a single, integrated software program? Does it require software to be developed to create a complete application?	U-4	04 - N/A
	Usability and Security Features		
23	Does the product include a graphical user interface (GUI)?	U-4	02 - Should Have
24	Does the product include the ability to save configurations to facilitate multiple runs using the same parameters?	U-2	02 - Should Have
25	Can the software be scripted to perform operations automatically?	U-2	02 - Should Have
26	Does the software require configuration, or can it be used "out of the box"?	U-4	02 - Should Have
27	Describe the product documentation available (provide link if possible).	R-4	02 - Should Have
28	When was the software first released?	R-6	04 - N/A
29	When was the most recent release of the software?	NA	02 - Should Have
30	Is there an active development effort for the product?	R-6	02 - Should Have
31	Describe the product support available.	R-2	01 - Must Have
32	Is the software single-user or multi-user? If multi-user, how does the system manage integrity and security of data and ensure partitioning between users?	U-6	03 - Could Have
33	Does the system contain security features such as requiring login/authentication?	S-5	01 - Must Have
34	Are there different user roles (e.g., administrator vs. user vs. data manager)?	U-6	03 - Could Have
35	If there are different roles, is the user interface segregated and optimized by role (e.g., a researcher would see the features of interest to an end-user, while an administrator would see a more full set of configuration functions)?	U-6	03 - Could Have
36	What execution performance reports are available (e.g., execution time, number of record pair comparisons, etc.).	F-8, F-17	01 - Must Have
37	What linkage performance reports are available (e.g., number of matches, number of possible matches, number of duplicates - if the software does de-duplication, etc.).	F-8, F-17	01 - Must Have
38	Has the system been approved to operate under U.S. government security regulations such as FISMA or FedRAMP?	S-4	03 - Could Have
39	Can the system run in a mode which does not persist any data (to minimize security risks)?	S-1	03 - Could Have
40	What protections are in place for source data?	S-3	01 - Must Have
	External System Integration		
41	What file formats can the software use (e.g., delimited and fixed-width text files, MS Excel, XML, JSON)?	I-2	03 - Could Have
42	Does the software integrate directly with data sources for input and/or output (e.g., ODBC/JDBC integration with relational database, web services)? Which ones?	I-4	03 - Could Have
43	Can the software be configured to be flexible about input formats (e.g., mapping input columns to program variables), or must inputs be put into a particular format?	I-2	03 - Could Have
44	What output formats does the software support?	I-3	03 - Could Have
45	Can the user customize the outputs?	I-3	03 - Could Have
	Data Cleaning / Pre-Processing Features		

ID	Question	Source Requirement	Priority
46	Describe any features the tool has to identify data quality issues, and standardize, recode and clean data to improve matching performance (e.g., substitution of nicknames such as "Jim" to "James", address lookup and standardization, geocoding, phonetic matching).	F-13	02 - Should Have
47	Is pre-processing specifiable field by field and file by file (e.g., a different date cleaning for DOB vs. record date, and for DOB in file 1 vs. file 2)?	F-14	03 - Could Have
48	Is the product extensible to use user-supplied pre-processing modules/services?	F-12	03 - Could Have
49	Can the software export subsets of the pre-processed data fields (e.g., only certain columns, only certain rows)?	F-15	02 - Should Have
50	Does the software support an evaluation mode (e.g., to allow researchers to work with their own data sets to clean and tune data before linkage)?	U-9	03 - Could Have
	Performance and Scalability		
51	What is the maximum file size/number of records that the software can handle?	P-1	01 - Must Have
52	What is the largest use case for the software to date?	R-5, P-1	04 - N/A
53	What features does the software have (such as blocking or database indexing) to improve performance? Can the user specify blocking parameters?	P-3	02 - Should Have
54	Describe the ability to customize performance improvement features such as blocking?	P-4	03 - Could Have
55	How can performance be improved by adding computational power (e.g., elastic compute)?	P-2	03 - Could Have
	Use cases, applications and future capabilities		
56	Do you have any use cases, publications/white papers, demos or videos describing applications of your product? (please provide links or describe separately outside of this form as appropriate)?	NA	04 - N/A
57	Do you have any additional features planned or in development that you would be willing to share and feel we should know about?	NA	04 - N/A

11.3 Additional Questions for Use in the Pilot Phase

This section contains questions which were developed during the landscape analysis, but which are appropriate to be investigated during the pilot phase.

ID	Evaluation Question	Initially Asked in Phase	Evaluation Criteria or Requirement Mapping
85	Results are easy to understand for end-users without significant post-processing	Pilot Phase	Req. U-5
86	Run time for a given data set	Pilot Phase	EC-P-1
87	Memory used to link a given data set	Pilot Phase	EC-P-2
88	Communication size (the amount of data transmitted over the network for a given data set)	Pilot Phase	EC-P-3

ID	Evaluation Question	Initially Asked in Phase	Evaluation Criteria or Requirement Mapping
89	Reduction Ratio: How much an indexing technique is able to reduce the number of candidate record pairs that are being generated compared to all possible record pairs. A higher reduction ratio value means an indexing technique is more efficient in reducing the number of candidate record pairs that are being generated; however, link quality can be affected if cross-category matches are mixed	Pilot Phase	EC-P-4
90	Precision/Positive Predictive Value: The fraction of record pairs classified as matches by a decision model that are true matches. Computed as the number of True Matches TM divided by the sum of true matches and false matches (FM) = TM/(TM+FM)	Pilot Phase	EC-LQ-1
91	Recall/Sensitivity: The fraction of true matches that are correctly classified as matches. Computed as true matches divided by True Matches plus False Negatives (FN) = TM/(TM+FN)	Pilot Phase	EC-LQ-2
92	F-measure: Combines precision and recall into a single metric = 2 x (Precision x Recall)/(Precision + Recall)	Pilot Phase	EC-LQ-3
93	Pair Completeness: Measures the effectiveness of an indexing technique in the record linkage process. Computed as the number of true matches correctly placed by blocking (BM) divided by the true matches plus false non-matches = BM/(TM+FN)	Pilot Phase	EC-LQ-4
94	Pair Quality: Measures the efficiency of a blocking technique. Similar to recall. Computed as true matches correctly placed by blocking (BM) divided by the sum of BM and true non-matches = BM/(BM+BN)	Pilot Phase	EC-LQ-5
95	Accuracy: The fraction of record pairs correctly classified = (TM + TN)/(TM+FM+TN+FN)	Pilot Phase	EC-LQ-6
96	Specificity: The fraction of true non-matches that are correctly classified as non-matches = TN/(TN+FN)	Pilot Phase	EC-LQ-7
97	Disclosure Risk: DR is the probability that masked records/QID values can be reidentified by being linked with records or values in a publicly available dataset, or by attacks such as dictionary attack. There are several variants of this metric.	Pilot Phase	EC-S-1
98	Uncertainty: Degree of Unlinkability: Unlinkability measures the adversary's uncertainty about which items are related. This can be a measure, for example, of salt effectiveness.	Pilot Phase	EC-S-2

ID	Evaluation Question	Initially Asked in Phase	Evaluation Criteria or Requirement Mapping
99	Time: Time Until Adversary's Success: This time-based metric and assumes that the adversary will eventually succeed - or can be used as a threshold value where the software is acceptable if the time to succeed is greater than a given time value t.	Pilot Phase	EC-S-3

12 Appendix 3: Vendor Points of Contact

Company	Product	POC	Position	Email	Phone
Crosse	SafeMine	Jeremy Mittler	VP, Industry Solutions	jeremy.mittler@crossix.com	Desk: 212-994-9369 | cell: 201-320-9684
CSIRO/Data61	Anonlink	Brian Thorne	Sr. Software Engineer	Brian.Thorne@data61.csiro.au	(02) 9490 5666
Datavant	Datavant	Jasmin Phua	Head of Solutions, Health Systems & Government	jas@datavant.com	443-794-9427
HealthVerity	Census/ Marketplace	Andrew Kress	CEO	akress@healthverity.com	215 582-2008
Policywise	Linkwise	Jason Lau	Director, Data Operations	jlau@policywise.com	Desk: 780 408-8732 | cell: 780 221-5081
Privitar	SecureLink	Rob O'Brien	Senior Director North America	rob.obrien@privitar.com	973-234-8975
Senzing	Senzing	Brian Macy	Chief of Product Support	support@senzing.com	NA
U. of Melbourne	GRHANITE	Douglas Boyle	Director, HaBIC Research Information Technology Unit	dboyle@unimelb.edu.au	Desk: +61 3 5823 4521 | cell: +61 458 220 820

13 Appendix 4: Candidate Product Survey Responses

This appendix provides the full survey answers returned by the vendors, including comments on each response as provided by the vendor. Comments are exactly as provided by the vendor. To avoid altering any vendor intent or information, the surveys have not been edited.

13.1 Anonlink

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
	PPID Generation and Record Linkage					
1	What can the user specify about the PPID generation process (e.g., which variables go into the PPID, and whether they should be truncated)?	x				By creating a linkage schema the user can specify per field behaviour. https://clkhash.readthedocs.io/en/latest/schema.html
2	Can the software generate and link on multiple PPIDs (e.g., hashes composed from different concatenated input variable combinations) either as a single pass or multiple passes? Is there a maximum number?			x		The system is currently designed to encode to a fixed size Cryptographic Linkage Key (CLK). It would be possible to modify the system to handle different CLKs but it would have security implications. The system is well decoupled though, so it is trivial to generate multiple PPIDs with different linkage schemas.
3	What is the mechanism for generating PPIDs (e.g., SHA-2 hashing)?					The system uses two methods: 1) original double hash encoding of the ngrams using the method from Schnell, R., Bachteler, T., & Reiher, J. (2011). A Novel Error-Tolerant Anonymous Linking Code. 2) a hashing mechanism based of the BLAKE2 designed to counter flaws in the original. See https://github.com/data61/clkhash/blob/master/clkhash/bloomfilter.py#L140 and https://github.com/data61/clkhash/issues/33
4	What probabilistic matching capability is available (e.g., Bloom filters on q-grams)?					Bloom filters on q-grams
5	How many files can be simultaneously linked (e.g., can the software link more than two files in one pass)?	x				Example with 5 data providers here https://anonlink-entity-service.readthedocs.io/en/latest/tutorial/multiparty-linkage-in-entity-service.html
6	Does the product support deduplication?			x		Currently working on it. Expect to be in a release by end of 2019.
7	Can the software support more than pairwise linkages (e.g., find all the records in a file that match)?			x		When output type is similarity scores they fully support many to many links. For full support the solver would need modification as it assumes the pairwise constraint.

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
8	Does the product support two-party privacy protected linkage (identifying linkages between two files potentially belonging to two owners, where neither party sees the other's unencrypted data)?					Yes the linkage is identified without either party seeing other data (encrypted or not), but using a third party.
9	Does the product support three-party protect linkage, where there is a trusted "honest broker" able to resolve possible linkages?	x				Yes 2 (or more) data providers trust a third party to act as the honest broker.
10	For three-party linkage, what information is made available to the trusted broker to resolve "possible" linkages (e.g., reports, distance metrics, comparisons of source data - if the broker can see source data)?			x		If the participants agree the distance metrics can be made available to the broker for tuning the linkage schema.
11	Are there any features for authorized reidentification of data?				x	Not possible within the system. However Anonlink is usually deployed within a larger system that usually does include such functionality.
12	What is tunable about matching criteria/algorithm?					The tuning is all contained within the Linkage Schema - tokenization, weighting of each feature etc.
13	Does the software have any ability to persist results so that subsequent linkages between data sets (e.g., after updates) can be incremental rather than from scratch?			x		We have thought about this a lot but decided to focus on performance instead to make precomputation feasible.
14	Does the software have the ability to split databases into linked vs. non-linked records, or other splitting and merging capability?				x	Our software assumes only access to the PII information - not any non linking features. This means another software/script will usually take the linkage results and could easily do the splitting.
15	Can the product persist PPIDs so they don't have to be regenerated for future runs?	x				The clkhash (or encoding service) componts generate CLKs which are stable "longterm" keys. They only need to be regenerated if the Linkage Schema, HMAC Keys, or underlying linking data changes.
	Operating Environment and Licensing Model					
16	What Platform/OS(s) does the system run under?					All. We test on OSX, Windows and Linux.
17	What other software is required to run your software (e.g., DBMS)?					Docker is required to run the anonlink-entity-service component (the third party matching)

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
18	Minimum hardware specification.					The matching component expects a 64bit architecture, solver runs in memory so the size of the problem will be eventually constained by RAM. Minimal resources that pass all our integration tests are here: https://github.com/data61/anonlink-entity-service/blob/develop/deployment/entity-service/minimal-values.yaml
19	Cloud-based version available? If so, which cloud environment?					Yes, recommended deployment is on Kubernetes. We've tested on Azure & Google managed k8s, and run our integration testing on a kops/aws k8s cluster.
20	Licensing model (per seat, per CPU, open source, etc.).					Open source under Apache 2.0
21	Does your licensing support "record linkage as a service", either through offering a cloud-based service or by distribution of the software as a utility?	x				We would consider running/hosting a paid service but don't currently. The software is ready to be run as a cloud hosted service.
22	Is the system a set of components or a single, integrated software program? Does it require software to be developed to create a complete application?			x		The anonlink system doesn't currently include a UI.
	Usability and Security Features					
23	Does the product include a graphical user interface (GUI)?				x	Has been designed as a REST api, we have UX mockups for a frontend but would require funding to develop.
24	Does the product include the ability to save configurations to facilitate multiple runs using the same parameters?	x				We even call them runs.
25	Can the software be scripted to perform operations automatically?	x				Via the rest API
26	Does the software require configuration, or can it be used "out of the box"?		x			A serious deployment to a kubernetes cluster is going to require configuration. However all tests, example ipython notebooks run against a fresh docker-compose deployment without any configuration required.
27	Describe the product documentation available (provide link if possible).					https://anonlink-entity-service.readthedocs.io/en/v1.11.0/# and https://clkhash.readthedocs.io/en/v0.13.0/
28	When was the software first released?					First release of clkhash was made on Aug 2, 2017

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
29	When was the most recent release of the software?					For clkhash June 27 2019 - https://github.com/data61/clkhash/releases/tag/v0.13.0 For anonlink-entity-service June 1 2019. https://github.com/data61/anonlink-entity-service/releases/tag/v1.11.0
30	Is there an active development effort for the product?	x				
31	Describe the product support available.			x		We offer research and development contracts for feature development, support contracts can be negotiated. Otherwise like any open source project support is on a best effort basis.
32	Is the software single-user or multi-user? If multi-user, how does the system manage integrity and security of data and ensure partitioning between users?					Multiuser, access controlled via project scoped tokens. CLK data is not segregated between users. External pentesting didn't reveal any security concerns. Data integrity is not cryptographically guaranteed end to end, however CLKs are transferred over TLS and stored in MinIO or S3 which do guarantee integrity (https://docs.min.io/docs/minio-erasure-code-quickstart-guide.html)
33	Does the system contain security features such as requiring login/authentication?	x				rate limiting, per project access control
34	Are there different user roles (e.g., administrator vs. user vs. data manager)?		x			Within a project there are "data provider" and "analyst" roles. Authentication can also be configured at cluster level.
35	If there are different roles, is the user interface segregated and optimized by role (e.g., a researcher would see the features of interest to an end-user, while an administrator would see a more full set of configuration functions)?					n/a
36	What execution performance reports are available (e.g., execution time, number of record pair comparisons, etc.).	x				Detailed information in the logging, and a rest endpoint publishes per run statistics as well as global statistics.
37	What linkage performance reports are available (e.g., number of matches, number of possible matches, number of duplicates - if the software does de-duplication, etc.).	x				Rest endpoint provides statistics per run (e.g. current progress info), as well as global statistics.
38	Has the system been approved to operate under U.S. government security regulations such as FISMA or FedRAMP?				x	Has not been assessed

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
39	Can the system run in a mode which does not persist any data (to minimize security risks)?			x		The api supports deleting all data from a run or project. This doesn't extend to logs.
40	What protections are in place for source data?	x				Never leaves the security domain where it belongs.
	External System Integration					
41	What file formats can the software use (e.g., delimited and fixed-width text files, MS Excel, XML, JSON)?					csv for PII using clkhash command line tool, json is supported by encoding service. Entity service uses a json rest api.
42	Does the software integrate directly with data sources for input and/or output (e.g., ODBC/JDBC integration with relational database, web services)? Which ones?					No, but it is written in Python so could be modified to do so.
43	Can the software be configured to be flexible about input formats (e.g., mapping input columns to program variables), or must inputs be put into a particular format?			x		Deployment configuration can be via environment variables or config files.
44	What output formats does the software support?					Mapping tables, permutations, groups, similarity scores
45	Can the user customize the outputs?			x		Not directly from the anonlink-entity-service, but it is a json data structure so it is expected that the user will be "using" the output from Rstudio or Jupyter notebook etc.
	Data Cleaning / Pre-Processing Features					
46	Describe any features the tool has to identify data quality issues, and standardize, recode and clean data to improve matching performance (e.g., substitution of nicknames such as "Jim" to "James", address lookup and standardization, geocoding, phonetic matching).					clkhash expects most common cleaning transformations to occur independantly. The schema is strictly defined so various type issues and missing data issues get picked up at encoding time. We have investigated methods for distance aware encoding in a privacy preserving way - https://medium.com/@wilko.henecka/distance-aware-address-encoding-for-privacy-preserving-record-linkage-a6cecdaadc22
47	Is pre-processing specifiable field by field and file by file (e.g., a different date cleaning for DOB vs. record date, and for DOB in file 1 vs. file 2)?	x				
48	Is the product extensible to use user-supplied pre-processing modules/services?	x				

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
49	Can the software export subsets of the pre-processed data fields (e.g., only certain columns, only certain rows)?		x			The clkhash tool processes full files at a time. The encoding service can export subsets of encoded data.
50	Does the software support an evaluation mode (e.g., to allow researchers to work with their own data sets to clean and tune data before linkage)?	x				
	Performance and Scalability					
51	What is the maximum file size/number of records that the software can handle?					Millions of records in each dataset.
52	What is the largest use case for the software to date?					State to Federal linkages within Australia
53	What features does the software have (such as blocking or database indexing) to improve performance? Can the user specify blocking parameters?		x			The similarity scoring workload is carried out in parallel on multiple machines and is written in optimized assembler - we see 50+ million comparisons per second per CPU core. Currently the anonlink-entity-services doesn't expose blocking to the user. Analysts can carry out blocking and use the service to link the blocked data together. We have adding functionality to the anonlink library and may implement support in the future.
54	Describe the ability to customize performance improvement features such as blocking?					
55	How can performance be improved by adding computational power (e.g., elastic compute)?					On a Kubernetes deployment the user may scale the available resources (at runtime) and the service can take advantage of additional compute capability. We have also created proof of concept code for running on Nvidia GPUs but haven't implemented that functionality in the service - indicative performance on a GTX 1080 is ~1.2 Billion comparisons/s including data transfer.

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
	Use cases, applications and future capabilities					
56	Do you have any use cases, publications/white papers, demos or videos describing applications of your product? (please provide links or describe separately outside of this form as appropriate)?					
57	Do you have any additional features planned or in development that you would be willing to share and feel we should know about?					I've mentioned a few directions in the survey e.g. handling deduplication well. We are investing effort investigating and prototyping approaches to cryptographically secure solutions. We have been experimenting with baysian optimization to learn the best Linkage Schema for a particular dataset. By taking into account accuracy against different subpopulations we can optimize on a "fair" linkage.

13.2 Crossix

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
	PPID Generation and Record Linkage					
1	What can the user specify about the PPID generation process (e.g., which variables go into the PPID, and whether they should be truncated)?	X				
2	Can the software generate and link on multiple PPIDs (e.g., hashes composed from different concatenated input variable combinations) either as a single pass or multiple passes? Is there a maximum number?	X				
3	What is the mechanism for generating PPIDs (e.g., SHA-2 hashing)?					SHA-256. Data is further encrypted after hashing as an additional security measure.
4	What probabilistic matching capability is available (e.g., Bloom filters on q-grams)?					SafeMine uses a proprietary algorithm. The algorithm calculates a match score and probability based on configurable data elements.
5	How many files can be simultaneously linked (e.g., can the software link more than two files in one pass)?	X				SafeMine uses a distributed model. This means that files are processed in parallel across multiple sites. The linkage is done separately.
6	Does the product support deduplication?	X				
7	Can the software support more than pairwise linkages (e.g., find all the records in a file that match)?	X				
8	Does the product support two-party privacy protected linkage (identifying linkages between two files potentially belonging to two owners, where neither party sees the other's unencrypted data)?	X				
9	Does the product support three-party protect linkage, where there is a trusted "honest broker" able to resolve possible linkages?	X				

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
10	For three-party linkage, what information is made available to the trusted broker to resolve "possible" linkages (e.g., reports, distance metrics, comparisons of source data - if the broker can see source data)?	X				For situations where a linkage is not definitive, we provide tools to evaluate and the user/trusted broker will be able to decide. The technology shows all rules that agree, rules that don't agree, distance metrics, score (by geography, name, etc.) as well as a threshold of uniqueness.
11	Are there any features for authorized reidentification of data?		X			
12	What is tunable about matching criteria/algorithm?	X				Nearly every parameter.
13	Does the software have any ability to persist results so that subsequent linkages between data sets (e.g., after updates) can be incremental rather than from scratch?	X				
14	Does the software have the ability to split databases into linked vs. non-linked records, or other splitting and merging capability?	X				
15	Can the product persist PPIDs so they don't have to be regenerated for future runs?	X				
	Operating Environment and Licensing Model					
16	What Platform/OS(s) does the system run under?					Full java implementation, so both Linux/Windows works.
17	What other software is required to run your software (e.g., DBMS)?					MySQL, and third party libraries that are integrated with the product.
18	Minimum hardware specification.					Minimum would depend on file size - no set specifications
19	Cloud-based version available? If so, which cloud environment?	X				Already deployed on AWS. Possible to deploy elsewhere.
20	Licensing model (per seat, per CPU, open source, etc.).					Typically, a one-time fee for installation / configuration per node. Annual licensing fee for use of technology per node. We are flexible and can accommodate other models as needed.
21	Does your licensing support "record linkage as a service", either through offering a cloud-based service or by distribution of the software as a utility?	X				We support multiple deployments

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
22	Is the system a set of components or a single, integrated software program? Does it require software to be developed to create a complete application?	X				Several components that together create a complete application. Customization is achieved through configuration.
	Usability and Security Features					
23	Does the product include a graphical user interface (GUI)?		X			Configuration is controlled through standard xml files, multiple GUI editors are available.
24	Does the product include the ability to save configurations to facilitate multiple runs using the same parameters?	X				
25	Can the software be scripted to perform operations automatically?	X				
26	Does the software require configuration, or can it be used "out of the box"?	X				There is a default configuration and may also require additoinal configuration.
27	Describe the product documentation available (provide link if possible).	X				
28	When was the software first released?					First installation was in 2006
29	When was the most recent release of the software?					June 2019
30	Is there an active development effort for the product?	X				Yes
31	Describe the product support available.	X				Crossix technology is used today, at massive scale, across many healthcare companies. We provide all necessary support, including technical, configuration, etc.
32	Is the software single-user or multi-user? If multi-user, how does the system manage integrity and security of data and ensure partitioning between users?					The core engine is used as a service, which matches all data sets uploaded to it. It can be wrapped per case.
33	Does the system contain security features such as requiring login/authentication?	X				We secure the system by protocols such as SSH and SSL. In addition, the system has today a single user which is identifying using username and password. On the data supplier end, we also have an ftp user identified with a username and password on top of SSH. All data is encrypted in rest, and we support encrypting clear data sets immediately upon their upload and before any other processing is applied to it.

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
34	Are there different user roles (e.g., administrator vs. user vs. data manager)?	X				There is an administrator as well as one user who is allowed to upload data and a second user who is allowed to query for data. Last type we have is user which is allowed to check system reports for monitoring purposes.
35	If there are different roles, is the user interface segregated and optimized by role (e.g., a researcher would see the features of interest to an end-user, while an administrator would see a more full set of configuration functions)?	X				
36	What execution performance reports are available (e.g., execution time, number of record pair comparisons, etc.).	X				Multiple metrics, including those listed
37	What linkage performance reports are available (e.g., number of matches, number of possible matches, number of duplicates - if the software does de-duplication, etc.).	X				Multiple metrics, including those listed
38	Has the system been approved to operate under U.S. government security regulations such as FISMA or FedRAMP?	X				SafeMine is HIPAA compliant. We are HITRUST and NIST certified.
39	Can the system run in a mode which does not persist any data (to minimize security risks)?	X				
40	What protections are in place for source data?	X				Encryption at rest. Customer-specific key used.
	External System Integration					
41	What file formats can the software use (e.g., delimited and fixed-width text files, MS Excel, XML, JSON)?	[redacted]				The system reads flat/gzipped delimited/positional files.
42	Does the software integrate directly with data sources for input and/or output (e.g., ODBC/JDBC integration with relational database, web services)? Which ones?					JDBC, MySQL
43	Can the software be configured to be flexible about input formats (e.g., mapping input columns to program variables), or must inputs be put into a particular format?	X				The system responds to queries. Can also generate a map file for generated IDs.
44	What output formats does the software support?	[redacted]				Flat file

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
45	Can the user customize the outputs?	X				
	Data Cleaning / Pre-Processing Features					
46	Describe any features the tool has to identify data quality issues, and standardize, recode and clean data to improve matching performance (e.g., substitution of nicknames such as "Jim" to "James", address lookup and standardization, geocoding, phonetic matching).	X				Soundex. Normalization tools. Statistical analysis, noise detection, frequency tables for popular demographic fields.
47	Is pre-processing specifiable field by field and file by file (e.g., a different date cleaning for DOB vs. record date, and for DOB in file 1 vs. file 2)?	X				
48	Is the product extensible to use user-supplied pre-processing modules/services?	X				
49	Can the software export subsets of the pre-processed data fields (e.g., only certain columns, only certain rows)?	X				
50	Does the software support an evaluation mode (e.g., to allow researchers to work with their own data sets to clean and tune data before linkage)?	X				
	Performance and Scalability					
51	What is the maximum file size/number of records that the software can handle?	[redacted]				System can handle data in tens of billions of transactions
52	What is the largest use case for the software to date?					Tens of billions of records processed
53	What features does the software have (such as blocking or database indexing) to improve performance? Can the user specify blocking parameters?	X				There are multiple performance enhancing features in the system.
54	Describe the ability to customize performance improvement features such as blocking?	X				System can be configured
55	How can performance be improved by adding computational power (e.g., elastic compute)?	X				

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
	Use cases, applications and future capabilities					
56	Do you have any use cases, publications/white papers, demos or videos describing applications of your product? (please provide links or describe separately outside of this form as appropriate)?	[redacted]				Yes; high-level video of our technology can be seen here: https://crossix.com/the-crossix-difference/
57	Do you have any additional features planned or in development that you would be willing to share and feel we should know about?					Please see attached word document

13.3 Datavant

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
	PPID Generation and Record Linkage					
1	What can the user specify about the PPID generation process (e.g., which variables go into the PPID, and whether they should be truncated)?	X				Datavant provides customers with extensive flexibility for the PPID generation process:. Inputs used to generate the token are flexible, supporting a broad variety of use cases based on data availability at the source and the privacy framework in use. Tokens generated through the PPID process can be built from many different combinations of PII elements, and the specific tokens to be created will be specified during the configuration process. Multiple tokens are often created to facilitate matching. Tokens can be built based on fields such as social security number, which allows for deterministic matching; alternately, they can be constructed from fields such as name and date of birth, which in combination can be used to support probabilistic matching. Our Token Selection Deep Dive (in reference document) provides an in-depth look on the possible selection criteria. During the onboarding process Datavant works with data partners on an approved data layout that will specify the accepted input variables and formats and the output format. We do not recommend truncating the generated tokens as it will affect linkage accuracy.
2	Can the software generate and link on multiple PPIDs (e.g., hashes composed from different concatenated input variable combinations) either as a single pass or multiple passes? Is there a maximum number?	X				The software can generate and link on multiple PPIDs. The software outputs the customers desired set of hashed combinations. Currently 26 hash combinations can be generated, and we continue to add new combinations in close partnership with our Expert Determination certifier, ensuring clarity on statistical re-identification risk on tokens intended for de-identified linkages. Please see User Guide pg 27 for the PPIDs available for linkage. Linkability in single or multiple passes is flexible and up to the trusted third party performing the linkage as we understand that depending on the intended use case, different linking techniques should be applied.
3	What is the mechanism for generating PPIDs (e.g., SHA-2 hashing)?	[redacted]				Datavant software combines the PII with a high-entropy master salt and runs the resulting amalgamation through a one-way SHA 256 hash function to create a master token for the patient, which is then further encrypted using AES-128 with PKCS#7 padding using a site-specific key to create the final site-specific token. The hash salt is common to all Datavant software installations to ensure compatibility of the resulting tokens. The encryption key is different for each installation to ensure that the tokens created at each site are unique. The hash salt and all encryption keys are maintained in a secure secrets management system maintained by Datavant.

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
4	What probabilistic matching capability is available (e.g., Bloom filters on q-grams)?					Datavant believes in providing its customers with extensive flexibility on matching capabilities. Today, Datavant can provide a variety of functions for token-based matching, in which a user can select which token combinations constitute a match across records. Longer term, Datavant plans to provide additional matching capabilities to meet client needs. The Datavant token scheme can also be employed in your own environment(s) to allow both deterministic and probabilistic matching, depending on the design you choose. There are many probabilistic matching algorithms one could employ to match tokens, and based on the tokens used, one strategy could be more advantageous than the other. We recommend not picking a single token or token combination for the matching logic, but to instead take advantage of multiple matching options using a "drop through" or “waterfall” technique. In this technique, the most stringent set of tokens are used in the first round to define a match. Any records matched in this round are put aside, and only unmatched records move to the next round, where the next most stringent tokens are used to define a match. The token designs span the spectrum of being optimized for broad matching (with the downside of a higher false positive rate) or for more stringent matching (with the downside of a higher false negative rate). We have been watching the use of Bloom filters with interest but understand it may introduce potential vulnerabilities though challenging to exploit. Bloom filters also add complexity to the linkage process and as the techniques to ensure Bloom filters are sufficiently protective expand, we have concerns that it may impact linkage accuracy and reliability.
5	How many files can be simultaneously linked (e.g., can the software link more than two files in one pass)?	X				There is no limit to the number of files that can be simultaneously linked assuming a reasonable computing capacity is provided. The user may link 2 files or specify a single directory for which all files within the directory, as well as all files within a sub-directory within that directory will be processed. Additionally files can have different sets of columns with different output tokens.
6	Does the product support deduplication?		X			The product supports linkages premised on generated tokens (PPIDs) that are dependent on input variables at the data source. It does not provide de-duplication capabilities in the sense of an Enterprise Master Patient Index or Identity Resolution system where additional variables in a patient record or demographic record may be used as part of the deduplication and filtering process. In the generation of the linked output, an index is created showing the Linked ID and relationship with all linked records from the file(s) linked. In this latter definition, deduplication is support as the end user will be able to ascertain if the linked record was present from the same dataset or from a single data source.

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
7	Can the software support more than pairwise linkages (e.g., find all the records in a file that match)?	X				Depending on the customer's implementation, all records within a file or across a series of files within a directory will be matched and linked.
8	Does the product support two-party privacy protected linkage (identifying linkages between two files potentially belonging to two owners, where neither party sees the other's unencrypted data)?	X				In Datavant's unique privacy protected linkage model, a two-party privacy model either with or without a trusted third party is supported. Each party will tokenize, resulting in Party A tokens and Party B tokens, which are never distributed to each other. Each party then encodes with a transit key where the generated transit tokens can only be used by the receiving party (party performing the linkage) to run linkages on encrypted transit tokens by Party A and Party B. The linkage recipient will never see Party A and Party B site-specific tokens, which are also encrypted tokens. See pg 6 Matching Accuracy of Tokens in De-identified Health Data Sets document for this tokenization and linkage workflow.
9	Does the product support three-party protect linkage, where there is a trusted "honest broker" able to resolve possible linkages?	X				The product supports three-party privacy protected linkages where a trusted honest broker can resolve possible linkages. In Datavant's privacy framework, a data recipient (e.g. data aggregator) peer-to-peer linkage (2-party), and multi-party linkages, where a third party serves as a trusted honest broker for linkage only, linkage and dataset aggregation) are all possible.
10	For three-party linkage, what information is made available to the trusted broker to resolve "possible" linkages (e.g., reports, distance metrics, comparisons of source data - if the broker can see source data)?	X				In a three-party linkage model, the trusted broker will be provided an account on the Datavant Portal which governs the key distribution process based on user credentials, and the ability to process a transit token from authorized data partners for a specific use case/project will be added to their user access controls. The data partners can send only the processed transit tokens to the trusted broker to resolve linkages. No other information needs to be provided in order for linkages to occur. If required in the data specification, additional information can accompany those tokens. For example, if authorized, the trusted broker could receive a file from 2 data partners; 1 containing tokens for linkage, and the other file containing tokens and 3-digit ZIP code. Datavant does not specify or add additional data to partner files.

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
11	Are there any features for authorized reidentification of data?	X				The Datavant software can be configured to produce crosswalk tables in order to facilitate re-identification by covered entities in approved contexts. However, requests for these configurations are closely reviewed and require approval by Datavant's internal privacy review. Re-identification is otherwise prevented technically and contractually. Datavant maintains a secure central repository for tracking and managing all data source token configurations. As a result, Datavant has an inventory of the entities that are able to generate a crosswalk so as to better support data sources in their decision-making around data sharing partners who may be able to re-identify data. Datavant encourages all users of the software to retain the services of an expert who can assist with and validate that the software's configuration adequately de-identifies the data in the context of its intended use.
12	What is tunable about matching criteria/algorithm?	X				Datavant tokens allow the customer to implement a broad array of matching designs, both deterministic and probabilistic matching, depending on the design. Many of Datavant's customers choose to institute their own series of matching criteria and algorithms based on specific use cases. For example, some customers prefer to have matches optimized for broad matching (with the downside of a higher false positive rate) or for more stringent matching (with the downside of a higher false negative rate). The primary factor impacting tunability are the series of possible tokens generated from the data source.
13	Does the software have any ability to persist results so that subsequent linkages between data sets (e.g., after updates) can be incremental rather than from scratch?	X				Customers can choose to persist results such that subsequent linkages can be incremental. Since the software is available as a set of components, customers are able to integrate it within their existing data pipelines and operational data requirements. The software itself does not manage the data persistence that a Master Data Management tool would provide.
14	Does the software have the ability to split databases into linked vs. non-linked records, or other splitting and merging capability?			X		The software does not directly interface with a database or database systems. During the record linkage process, a Master Index ID is generated for all records. Linked records will have the same Master Index ID, and therefore database views and filters could easily be used to sort linked vs. non-linked records by the Master Index ID and the presence of multiple instances of that ID. The same functionality could be used to also identify duplicates within a dataset.
15	Can the product persist PPIDs so they don't have to be regenerated for future runs?	X				PPIDs can be persisted locally if customers data governance practices permit. Since the software is available for use locally, customers can choose to persist or destory PPIDs according to their local requirements. The software outputs the PPIDs to a flat file with the configurations of the customer's choosing.

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
	Operating Environment and Licensing Model					
16	What Platform/OS(s) does the system run under?					Windows 10 and Windows Server 2016 or later; MacOS 10.12 and later; Linux Ubuntu
17	What other software is required to run your software (e.g., DBMS)?					No other software is required to run Datavant software. Our software is self-contained. There are 2 operational requirements: (1) Ability to retrieve encryption keys over HTTPS port 443, (2) Download of a security authentication file that we provide to the end-user to ensure appropriate authorizations and configurations.
18	Minimum hardware specification.					We recommend using 3GHz Quad-Core Processor, 8GB RAM. No database management system is needed, and servers are optional for small datasets. The software runs on premise behind your firewall, and is able to run on ordinary workstation hardware, or even a laptop. If additional performance is desired, server-class hardware may be used. Typically, use is CPU-bound, so a server with higher clock speed or multiple cores will improve performance. The hardware machine or server should have storage for approximately 2x the volume of your data to allow for output to be written. Clients typically calibrate their CPU needs based on volume of data required to be processed in a fixed time. Datavant has supported multiple clients with initial and ongoing system needs calibration.
19	Cloud-based version available? If so, which cloud environment?			X		Datavant can provide a cloud-based version in your cloud environment of choice with some customization; we are able to support deployments in customer private cloud, Amazon AWS. Microsoft Axure, Google Cloud Platform, and Snowflake. We have extensive experience supporting a broad array of customers on their cloud environments.
20	Licensing model (per seat, per CPU, open source, etc.).					Datavant has a flexible licensing model and does not meter for usage. Our license model is a hub-spoke model where we charge an annual fee for the master license which is typically signed with the data aggregator/recipient (hub) and a single-use link fee or yearly unlimited link fee for each data partner (spoke). In situations where hubs prefer that Datavant run and operate the overlaps and matching, there may be an additional fee.
21	Does your licensing support "record linkage as a service", either through offering a cloud-based service or by distribution of the software as a utility?	X				Since Datavant does not charge on a per link basis and simply distributes software as the utility, data aggregators and hubs are free to provide record linkage as a service. We do have customers that request that we run linkage as a service for them, and those are negotiated on a case-by-case basis.

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
22	Is the system a set of components or a single, integrated software program? Does it require software to be developed to create a complete application?	X				The system is a set of components that can be implemented as needed within customers' data pipelines and environments. Depending on customer's deployment preferences, software may need to be developed to complete the application — for example, some of our enterprise customers prefer to run their own matching scheme on the generated tokens, while others prefer to use Datavant's software. By providing customers with a flexible set of components, we find that implementations are able to accommodate varying data partner system environments and workflows.
	Usability and Security Features					
23	Does the product include a graphical user interface (GUI)?			X		The product consists of a set of command line tools. The user onboarding and management web portal is a web-based graphical user interface. Datavant also has a web-based Discovery platform that enables self-service linkage (overlaps) with high-level aggregate summaries.
24	Does the product include the ability to save configurations to facilitate multiple runs using the same parameters?	X				During the onboarding process, Datavant works with customers to generate all the relevant configurations. During runtime, the customer can run different configurations each with different parameters. For example, if a trusted third party is used solely for a linkage service, the configuration would process and output only tokens needed for linkage, whereas if the data recipient were a central registry that was aggregating the substantive clinical record, that configuration would process both the tokens and output the relevant data that the site would transport to the central registry.
25	Can the software be scripted to perform operations automatically?	X				Since the software consists of a set of standalone components, end users integrate the different components into their current data pipelines as needed. For example, if a customer has an existing ETL (extract-transform-load) process, they can call the tokenization component (DeID) when needed to generate the output tokens.
26	Does the software require configuration, or can it be used "out of the box"?		X			The software is usable "out of the box". The only configurations that need to occur are during the onboarding process where the Datavant team will generate the authorized input and output configurations so the customer has a consistent set of configurations that can be used. Although it is possible for a customer to perform this function out of the box, we have chosen to generate these configurations to ensure sufficient oversight of the types of tokens generated and to ensure that the generated output meets HIPAA expert determination certification.
27	Describe the product documentation available (provide link if possible).	X				Please see Reference Links document for Datavant User Guide.

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
28	When was the software first released?					October 13, 2014
29	When was the most recent release of the software?					V3.1 (May 3, 2019)
30	Is there an active development effort for the product?	X				Datavant has an extensive product roadmap, and works iteratively with our customers on product development. We use an established product roadmap and software development life cycle (SDLC) process including automated and manual testing Minor updates are tested internally before rolling out across customers. Major updates go through beta testing with select customers prior to a broader rollout.
31	Describe the product support available.	X				Datavant provides responsive support. All customers who contact us via our support@datavant.com address receive a response within 3 hours of contact. Additionally, we provide a User Guide that all our customers have found helpful. Most customers are up and running in under 30 minutes with token generation. As part of the onboarding and implementation process, all customers have accounts on the Datavant Portal where guided support, the latest User Guide, and any frequently asked questions are maintained.
32	Is the software single-user or multi-user? If multi-user, how does the system manage integrity and security of data and ensure partitioning between users?					The software is intended to be single-user. Each user has their own account and installation related to the organization / entity / data partner they're responsible for tokenizing. User access controls are enforced through the Datavant Portal. Since Datavant does not process institutions' data, all data processing and partitioning occurs locally on-premise within the institution's data processing environment. Additionally, an authentication file related to the user's institutions and projects is required at runtime, providing a further layer of security to operate the software.
33	Does the system contain security features such as requiring login/authentication?	X				In order to use the software an institution and its authorized users must be registered on the Datavant Portal. Only authorized users with the appropriate roles are able to download the various software components. An authentication file that can only be obtained from the portal is required at runtime in order for the software to run. Therefore, in a situation where a user at an institution has been deactivated but may still have access to the software and data within the institution's local environment, the software will not run.
34	Are there different user roles (e.g., administrator vs. user vs. data manager)?	X				The different user roles are provisioned by Datavant for customers and the end users during the onboarding process. We do not currently provide institutions with administrator access to add their own users. Given the sensitive nature of the data processed, we work closely with our customers to onboard end users and provide the appropriate permissions to download software, generate tokens for different data networks and use cases.

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
35	If there are different roles, is the user interface segregated and optimized by role (e.g., a researcher would see the features of interest to an end-user, while an administrator would see a more full set of configuration functions)?		X			Since the user roles primarily control the ability for the end user to generate tokens and process site tokens for linkage through a command line interface, the role separations are not visible on a user interface. On Datavant's Discovery platform, which is a different product, user interfaces are segregated and optimized by role.
36	What execution performance reports are available (e.g., execution time, number of record pair comparisons, etc.).			X		The software generates a series of local logs including error logs so users have visibility into potential errors, summaries and data missingness. Some examples are included in pg 22 of the Datavant User Guide. Execution performance reports tend to vary based on the customers choice of deployment; in situations where the customer uses their own matching schemes, the customer will need to generate their own execution performance reports. We work with customers to generate performance reports that are appropriate for their execution environments.
37	What linkage performance reports are available (e.g., number of matches, number of possible matches, number of duplicates - if the software does de-duplication, etc.).			X		Linkage performance reports vary based on the software implementations; for example, customers that apply their own matching schemes generate their own performance reports. When the Datavant Discovery platform is used for Overlaps, summary statistics such as Total Unique Individuals in the source dataset, Total Overlapping Individuals between datasets, percentages of overlaps, and frequency distribution graphs are provided.
38	Has the system been approved to operate under U.S. government security regulations such as FISMA or FedRAMP?				X	Datavant's software does not currently meet a FISMA or FedRAMP designation. We are SOC 2 Type 2 certified. Through our acquisition of Health Data Link, we are also certified for use within the Veteran's Administration systems.
39	Can the system run in a mode which does not persist any data (to minimize security risks)?	X				All tokenization and linkages are run locally by the appropriate parties on premise within their own data processing environments. The end user can choose to persist or destroy data as required by their compliance and operational requirements.
40	What protections are in place for source data?	X				All source data is processed locally by the source institution or its designee. Datavant does not have access or ability to view or process source data. During the tokenization process, the HTTPS transaction over port 443 is solely used to obtain encryption keys and confirm user authorizations, thereafter the port is closed before any source data is processed locally. . The security of Datavant's software has been independently reviewed and certified by Rhino Security, who conducted both a code review and a penetration test of the software, network call, and secrets system. Datavant does not have access to client PHI. Nevertheless, all Datavant employees are trained in HIPAA guidelines and Datavant has instituted all policies and processes necessary to safeguard PHI should it be exposed during any interaction with our staff.

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
	External System Integration					
41	What file formats can the software use (e.g., delimited and fixed-width text files, MS Excel, XML, JSON)?					Input files encoded in UTF-8 delimited by pipe, comma, tab, semicolon etc. are accepted formats for the software.
42	Does the software integrate directly with data sources for input and/or output (e.g., ODBC/JDBC integration with relational database, web services)? Which ones?					The software does not currently directly integrate with data sources for input and output through Electronic Data Interchange (EDI) or database interfaces (ODBC, JDBC). We are working on a software development kit (SDK) that will provide end-users and software vendors with ways to integrate using the SDK.
43	Can the software be configured to be flexible about input formats (e.g., mapping input columns to program variables), or must inputs be put into a particular format?	X				The software can be configured to be flexible about input formats. During the onboarding process, we work with the customer to define requested data layouts, and generate the configuration files needed to support the customers data input needs. See pg 15 of the Datavant User Guide for the extensive configuration capabilities that Datavant supports.
44	What output formats does the software support?					Similar to the input file format, the software supports outputs to a flat file with variables and delimiters as requested by the customer.
45	Can the user customize the outputs?	X				Datavant works with the user to ensure to customize the outputs which are specified through the configurations.
	Data Cleaning / Pre-Processing Features					
46	Describe any features the tool has to identify data quality issues, and standardize, recode and clean data to improve matching performance (e.g., substitution of nicknames such as "Jim" to "James", address lookup and standardization, geocoding, phonetic matching).		X			The tool has a set of pre-cleaning and standardizing and normalization processes that are applied. A validation routine is run and then the cleaning routines are run subsequently. Pg 21 of the Datavant User Guide provide some examples of pre-processing that the tool performs. We would be glad to share further details on our data cleaning and pre-processing features under confidentiality. The tool does not currently handle nickname substitutions; does not manipulate nicknames or extend them to typical full names when creating tokens. We are working on a feature that would "clean" the name field so that if it is a common nickname, we would convert it to a full name and create tokens for both the full name vs input name in the record. So Alex would get "cleaned" to Alexander. We are evaluating how this affects matching overlaps and performance currently.
47	Is pre-processing specifiable field by field and file by file (e.g., a different date cleaning for DOB vs. record date, and for DOB in file 1 vs. file 2)?				X	PII attributes that are used for tokenization are processed consistently, ie. We do not pre-process token inputs on a field by field basis. Additional variables that accompany the tokens (PPIDs) are pass through variables for which certain rules and formats can be applied as long as they are specified during the confirguration process but we would not apply different pre-processing features to a DOB vs record date in different files.

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
48	Is the product extensible to use user-supplied pre-processing modules/services?			X		Since the product consists of a set of command line components, the product can be integrated into end-user data pipelines as neeed. The product does not use user-supplied pre-processing modules as a software plugin.
49	Can the software export subsets of the pre-processed data fields (e.g., only certain columns, only certain rows)?		X			Columns can be exported based on the configurations created. The software does not export specific rows. It processes all rows in the input file and outputs all rows.
50	Does the software support an evaluation mode (e.g., to allow researchers to work with their own data sets to clean and tune data before linkage)?	X				The software generates an error log file that provides researchers with insight on problems with the data. See pg 22 of the Datavat User Guide for error types that can be used by researchers to inform their data processing.
	Performance and Scalability					
51	What is the maximum file size/number of records that the software can handle?	[redacted]				There is no maximum file size related to the software.Typically, use is CPU-bound, so a server with higher clock speed or multiple cores will improve performance. The hardware machine or server should have storage for approximately 2x the volume of your data to allow for output to be written. Clients typically calibrate their CPU needs based on volume of data required to be processed in a fixed time. Datavant has supported multiple clients with initial and ongoing system needs calibration.
52	What is the largest use case for the software to date?					Datavant's de-identification software has been used to de-identify and link a dataset of 1 billion records at one time, and for longitudinal records over time, it has been used to de-identify and link mutiple datasets that contain four to five billion records each.
53	What features does the software have (such as blocking or database indexing) to improve performance? Can the user specify blocking parameters?		X			Datavant's software is multi-threaded providing flexibility for end users to scale processing given numbe threads available as needed. The software does not use blocking techniques or parameters at the moment. We are exploring such approaches but have not found those performance techniques to be immediately necessarily. For reference, in a recent performance test, Datavant software has been shown to process 62,500 records per second (or a million records in 0.27 minutes) to transform two tokens (hashes) into transit tokens to accommodate record transfer. The performance benchmarks were based on an Intel(R) Xeon(R) CPU E5-2686 v4 @ 2.30GHz machine with 32 GB RAM and eight (8) threads. Given the availability of elastic compute cloud environments, most of our customers scale their machine configurations and threads as needed.
54	Describe the ability to customize performance improvement features such as blocking?			X		Blocking techniques can be used to improve performance but will require customization to implement and will depend on the matching scheme used.

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
55	How can performance be improved by adding computational power (e.g., elastic compute)?	X				Datavant's software is multi-threaded and can be deployed in an environment that takes advantage of machine configurations such as an elastic compute environment or a multi-threaded environment across a cluster.
	Use cases, applications and future capabilities					
56	Do you have any use cases, publications/white papers, demos or videos describing applications of your product? (please provide links or describe separately outside of this form as appropriate)?	[redacted]				Links to use cases and applications are provided in the Reference document.
57	Do you have any additional features planned or in development that you would be willing to share and feel we should know about?					**Discovery Platform**: Datavant is working on a Discovery Platform for which data partners are able to collaborate with each other directly in a completely de-identified manner to understand potential overlaps and characterization of the dataset to decide whether there is a need to engage with each other. For example, the New York, New Jersey, and Connecticut departments of public health could choose to allow self-service linkages between their cancer registries. No actual data is exchanged, but basic HIPAA expert determination certified data profiles are shared with aggregate-level counts. We believe in empowering our customers and end-users to facilitate safe, private, real-time data collaborations in a privacy-protecting manner. See Datavant Overlaps on Discovery. We have launched our Discovery platform in beta mode and are working with customers and HIPAA expert determination certifiers on continuing to add additional features. **Advances in Unstructured Data Processing**: Datavant is working on additional privacy-preserving and de-identification methods related to unstructure data fields. We understand the nature of healthcare data and current available formats. Through our experience, we realize that highly identifiable information tends to be contained in unstructured data fields such as in Chief Complaint summaries in an EMR or in Lab Notes. We have a working version of a HIPAA expert determination certified de-identifcation process that has been certified for lab notes, and have a roadmap on making such features available on our future roadmap. **Data De-identification Libraries**: Datavant's solution provide de-identification rules and libraries that help customers apply frequently used data de-identification requirements such as excluding/suppressing rare disease ICD codes, processing ZIP codes to 3-digit ZIP codes that include suppression of small population ZIPs. Datavant continues to work with our customers to continue buliding out a library of de-identification modules that can be applied based on the customer's use cases and with our expert determination certifiers to

		[redacted]	provide the appropriate privacy frameworks for data de-identification, data linkage, and data re-identification risks. **Match Configurations and Libraries:** In the past year, Datavant has performed over 50 overlap studies that matched different token sets from various clients to determine how many individuals are common across the aggregated set. The sensitivity and specificity of our matching algorithm is dependent on the quality of the data (how "clean" it is in terms of duplication) and the tokens that are available to use in the exercise. We are working with key partners, including the insights gained from the Health Data Link implementations, which are the industry's current only peer-reviewed published gold standard references on linkage performance for privacy-preserving record linkages within the healthcare sector. These tokenization and match configurations will be available in the coming months with baseline performance standards for linkage precision and recall.

13.4 PolicyWise

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
	PPID Generation and Record Linkage					
1	What can the user specify about the PPID generation process (e.g., which variables go into the PPID, and whether they should be truncated)?	x				User can choose which identifiers are used for linkage. LinkWise doesn't need truncation. Date formats can be specfied in advance.
2	Can the software generate and link on multiple PPIDs (e.g., hashes composed from different concatenated input variable combinations) either as a single pass or multiple passes? Is there a maximum number?	x				Multiple can be used, like names, birthdates, etc.
3	What is the mechanism for generating PPIDs (e.g., SHA-2 hashing)?					Bloom filter
4	What probabilistic matching capability is available (e.g., Bloom filters on q-grams)?					Bloom filter
5	How many files can be simultaneously linked (e.g., can the software link more than two files in one pass)?				x	No. The software supports creating only one single linkable file at a time. Linkage is done post-hoc. However, the advantage is that it allows many deidentified linkable files to be created which are then linked post hoc.
6	Does the product support deduplication?			x		Since we sometimes want to know entities who appear multiple times in a dataset (e.g. for numbers of service usages over time) we retain duplicates. If we only want to identify the overlap in entities between files we can remove duplicates prior to linking .
7	Can the software support more than pairwise linkages (e.g., find all the records in a file that match)?					Ooh sorry this I might have to ask the programmer.
8	Does the product support two-party privacy protected linkage (identifying linkages between two files potentially belonging to two owners, where neither party sees the other's unencrypted data)?	x				The software is run independently by each data provider who never see other providers' data. Linkage is run post-hoc by PolicyWise.
9	Does the product support three-party protect linkage, where there is a trusted "honest broker" able to resolve possible linkages?	x				Not sure I understand this one completely. If you mean having special data covenentors like for ICES in Ontario, then PPRL was designed to get around needing such a special previliedged individual. It certainly wouldn't preclude such a process.

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
10	For three-party linkage, what information is made available to the trusted broker to resolve "possible" linkages (e.g., reports, distance metrics, comparisons of source data - if the broker can see source data)?					Ooh sorry this I might have to ask the programmer.
11	Are there any features for authorized reidentification of data?				x	No
12	What is tunable about matching criteria/algorithm?		x			User can choose as many or as few fields they wish to use for identification.
13	Does the software have any ability to persist results so that subsequent linkages between data sets (e.g., after updates) can be incremental rather than from scratch?				x	No. The senvelope builder offers better results with more data ingested. Adding only a small portion of the total data available would return less accurate or unreliable results.
14	Does the software have the ability to split databases into linked vs. non-linked records, or other splitting and merging capability?				x	No. The software supports creating only one single linkable file at a time. However, the advantage is that it allows many deidentified linkable files to be created which are then linked post hoc.
15	Can the product persist PPIDs so they don't have to be regenerated for future runs?					Ooh sorry this I might have to ask the programmer. I believe so though as this was an issue we had to specifically address in the IBM software.
	Operating Environment and Licensing Model					
16	What Platform/OS(s) does the system run under?					Windows. What other platforms does C# run under?
17	What other software is required to run your software (e.g., DBMS)?					Nothing
18	Minimum hardware specification.					Very little, Entirely dependent on dataset.
19	Cloud-based version available? If so, which cloud environment?				x	
20	Licensing model (per seat, per CPU, open source, etc.).					Our default business model is to conduct linkage as a service. However, we are flexible and can entertain other models.
21	Does your licensing support "record linkage as a service", either through offering a cloud-based service or by distribution of the software as a utility?	x				

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
22	Is the system a set of components or a single, integrated software program? Does it require software to be developed to create a complete application?					The software has 2 components: A. an envelope builder which allows users to create deidentified hashed linkable files. B. A resolver, which links and identifies common entities based on the results from a. PolicyWise uses this as a service to link files provided in a.
	Usability and Security Features					
23	Does the product include a graphical user interface (GUI)?	x				
24	Does the product include the ability to save configurations to facilitate multiple runs using the same parameters?				x	
25	Can the software be scripted to perform operations automatically?				x	
26	Does the software require configuration, or can it be used "out of the box"?					Out of the box
27	Describe the product documentation available (provide link if possible).			x		It's very limited right now. We are working on new validation documentation. https://policywise.com/2018/03/15/linkwise/
28	When was the software first released?					2018
29	When was the most recent release of the software?					2018
30	Is there an active development effort for the product?					It's sporadic depending on client needs and resourcing.
31	Describe the product support available.					PolicyWise staff offer help in using the software.
32	Is the software single-user or multi-user? If multi-user, how does the system manage integrity and security of data and ensure partitioning between users?					Having an independent envelope builder supports the multi-user model. Not even the user of the linkage resolver (ie. PolicyWise) ever sees the identifiable data. Since linkage is actually performed post-hoc identifiable microdata is opaque to all users submitting their data.
33	Does the system contain security features such as requiring login/authentication?				x	No, it's just local.
34	Are there different user roles (e.g., administrator vs. user vs. data manager)?				x	No

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
35	If there are different roles, is the user interface segregated and optimized by role (e.g., a researcher would see the features of interest to an end-user, while an administrator would see a more full set of configuration functions)?				x	
36	What execution performance reports are available (e.g., execution time, number of record pair comparisons, etc.).		x			The linkage resolving component shows the time elapsed to resolve linkages.
37	What linkage performance reports are available (e.g., number of matches, number of possible matches, number of duplicates - if the software does de-duplication, etc.).				x	Again, these would be measured post-hoc. Output from the linkage resolver is joined using SQL or other language. These results can be obtained by interpreting the number of records linked across files.
38	Has the system been approved to operate under U.S. government security regulations such as FISMA or FedRAMP?				x	
39	Can the system run in a mode which does not persist any data (to minimize security risks)?	x				The envelope builder is only run onsite by contributing data providers. Data providers only send hashed results to PolicyWise for resolving.
40	What protections are in place for source data?	x				Source data always remains on data providers' machines. If they desire data providers have the option of sending other microdata. However, each record is only identified by the hashed identifier generated with the envelope builder.
	External System Integration					
41	What file formats can the software use (e.g., delimited and fixed-width text files, MS Excel, XML, JSON)?	[redacted]				Delimited text files: comma, tab, space delimited.
42	Does the software integrate directly with data sources for input and/or output (e.g., ODBC/JDBC integration with relational database, web services)? Which ones?					No
43	Can the software be configured to be flexible about input formats (e.g., mapping input columns to program variables), or must inputs be put into a particular format?					Users can specify the format of a particular variable e.g. birth date as YYYY/MM/DD, DD/MM/YYYY, etc.
44	What output formats does the software support?	[redacted]				.csv

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
45	Can the user customize the outputs?					Yes. The user is able to choose which microdata fields are to be exported with the hashed identifier fields.
	Data Cleaning / Pre-Processing Features					
46	Describe any features the tool has to identify data quality issues, and standardize, recode and clean data to improve matching performance (e.g., substitution of nicknames such as "Jim" to "James", address lookup and standardization, geocoding, phonetic matching).		x			The software uses SoundEx principles to identify similar names, other common names are resolved by the envelope builder. However, these are never reported by LinkWise as statistics.
47	Is pre-processing specifiable field by field and file by file (e.g., a different date cleaning for DOB vs. record date, and for DOB in file 1 vs. file 2)?			x		The envelope builder can accept a number of different date formats for date of birth. Formats do not need to be standardized across the input files.
48	Is the product extensible to use user-supplied pre-processing modules/services?				x	No
49	Can the software export subsets of the pre-processed data fields (e.g., only certain columns, only certain rows)?	x				
50	Does the software support an evaluation mode (e.g., to allow researchers to work with their own data sets to clean and tune data before linkage)?				x	No
	Performance and Scalability					
51	What is the maximum file size/number of records that the software can handle?					We have tried low millions thus far.
52	What is the largest use case for the software to date?					Use cases have been thousands. Testing has been low millions.
53	What features does the software have (such as blocking or database indexing) to improve performance? Can the user specify blocking parameters?				x	
54	Describe the ability to customize performance improvement features such as blocking?				x	

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
55	How can performance be improved by adding computational power (e.g., elastic compute)?				x	
	Use cases, applications and future capabilities					
56	Do you have any use cases, publications/white papers, demos or videos describing applications of your product? (please provide links or describe separately outside of this form as appropriate)?					Under development
57	Do you have any additional features planned or in development that you would be willing to share and feel we should know about?					We would like to add more advanced machine learning techniques but it depends on resourcing in the future

13.5 Privitar

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
	PPID Generation and Record Linkage					
1	What can the user specify about the PPID generation process (e.g., which variables go into the PPID, and whether they should be truncated)?	✓				
2	Can the software generate and link on multiple PPIDs (e.g., hashes composed from different concatenated input variable combinations) either as a single pass or multiple passes? Is there a maximum number?	✓				
3	What is the mechanism for generating PPIDs (e.g., SHA-2 hashing)?					
4	What probabilistic matching capability is available (e.g., Bloom filters on q-grams)?					

5	How many files can be simultaneously linked (e.g., can the software link more than two files in one pass)?	✓				Multiple
6	Does the product support deduplication?				✓	
7	Can the software support more than pairwise linkages (e.g., find all the records in a file that match)?	✓				
8	Does the product support two-party privacy protected linkage (identifying linkages between two files potentially belonging to two owners, where neither party sees the other's unencrypted data)?	✓				
9	Does the product support three-party protect linkage, where there is a trusted "honest broker" able to resolve possible linkages?	✓				
10	For three-party linkage, what information is made available to the trusted broker to resolve "possible" linkages (e.g., reports, distance metrics, comparisons of source data - if the broker can see source data)?					

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
11	Are there any features for authorized reidentification of data?	✓				
12	What is tunable about matching criteria/algorithm?					Need more information
13	Does the software have any ability to persist results so that subsequent linkages between data sets (e.g., after updates) can be incremental rather than from scratch?	✓				
14	Does the software have the ability to split databases into linked vs. non-linked records, or other splitting and merging capability?			✓		
15	Can the product persist PPIDs so they don't have to be regenerated for future runs?			✓		
	Operating Environment and Licensing Model					
16	What Platform/OS(s) does the system run under?					Linux
17	What other software is required to run your software (e.g., DBMS)?					Oracle, MySQL, or HDFS
18	Minimum hardware specification.					Depends upon scale of data
19	Cloud-based version available? If so, which cloud environment?					AWS and Azure
20	Licensing model (per seat, per CPU, open source, etc.).					Operations based
21	Does your licensing support "record linkage as a service", either through offering a cloud-based service or by distribution of the software as a utility?			✓		
22	Is the system a set of components or a single, integrated software program? Does it require software to be developed to create a complete application?	✓				Should not require
	Usability and Security Features					
23	Does the product include a graphical user interface (GUI)?	✓				

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
24	Does the product include the ability to save configurations to facilitate multiple runs using the same parameters?	✓				
25	Can the software be scripted to perform operations automatically?	✓				Supports REST APIs
26	Does the software require configuration, or can it be used "out of the box"?	✓				
27	Describe the product documentation available (provide link if possible).	✓				
28	When was the software first released?					Early access in 2018
29	When was the most recent release of the software?					Next release is early July 2019
30	Is there an active development effort for the product?					Yes
31	Describe the product support available.					During US buisiness hours
32	Is the software single-user or multi-user? If multi-user, how does the system manage integrity and security of data and ensure partitioning between users?					multi-user with role-based access control
33	Does the system contain security features such as requiring login/authentication?	✓				
34	Are there different user roles (e.g., administrator vs. user vs. data manager)?	✓				
35	If there are different roles, is the user interface segregated and optimized by role (e.g., a researcher would see the features of interest to an end-user, while an administrator would see a more full set of configuration functions)?	✓				
36	What execution performance reports are available (e.g., execution time, number of record pair comparisons, etc.).		✓			
37	What linkage performance reports are available (e.g., number of matches, number of possible matches, number of duplicates - if the software does de-duplication, etc.).			✓		
38	Has the system been approved to operate under U.S. government security regulations such as FISMA or FedRAMP?		✓			Approved by UK government for use in national health care (NHS)

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
39	Can the system run in a mode which does not persist any data (to minimize security risks)?			✓		
40	What protections are in place for source data?					Need more information
	External System Integration					
41	What file formats can the software use (e.g., delimited and fixed-width text files, MS Excel, XML, JSON)?					CSV
42	Does the software integrate directly with data sources for input and/or output (e.g., ODBC/JDBC integration with relational database, web services)? Which ones?					Can be integrated into streaming data flow or batch processing
43	Can the software be configured to be flexible about input formats (e.g., mapping input columns to program variables), or must inputs be put into a particular format?					
44	What output formats does the software support?					Multiple supported formats
45	Can the user customize the outputs?					Need more information
	Data Cleaning / Pre-Processing Features					
46	Describe any features the tool has to identify data quality issues, and standardize, recode and clean data to improve matching performance (e.g., substitution of nicknames such as "Jim" to "James", address lookup and standardization, geocoding, phonetic matching).			✓		Support for lookup/substitution tables
47	Is pre-processing specifiable field by field and file by file (e.g., a different date cleaning for DOB vs. record date, and for DOB in file 1 vs. file 2)?	✓				
48	Is the product extensible to use user-supplied pre-processing modules/services?	✓				Via REST APIs
49	Can the software export subsets of the pre-processed data fields (e.g., only certain columns, only certain rows)?		✓			Columns can be redacted

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
50	Does the software support an evaluation mode (e.g., to allow researchers to work with their own data sets to clean and tune data before linkage)?		✓			Evaluation licences can be negotiated
	Performance and Scalability					
51	What is the maximum file size/number of records that the software can handle?					N/A
52	What is the largest use case for the software to date?					50 TB of patient claims, call center,
53	What features does the software have (such as blocking or database indexing) to improve performance? Can the user specify blocking parameters?	✓				Indexing supported by underlying data store
54	Describe the ability to customize performance improvement features such as blocking?					Need more informaiton
55	How can performance be improved by adding computational power (e.g., elastic compute)?					Yes
	Use cases, applications and future capabilities					
56	Do you have any use cases, publications/white papers, demos or videos describing applications of your product? (please provide links or describe separately outside of this form as appropriate)?					https://www.privitar.com/securelink
57	Do you have any additional features planned or in development that you would be willing to share and feel we should know about?					Full integration of Privitar SecureLink with Privitar Publisher in the July 3.0 release

13.6 Senzing

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
	PPID Generation and Record Linkage					
1	What can the user specify about the PPID generation process (e.g., which variables go into the PPID, and whether they should be truncated)?	X				Senzing allows one-way hashes "Identity Attributes" and then performs entity resolution on those attributes. Any type of "Identity" data can be utilized within the Entity Resolution process
2	Can the software generate and link on multiple PPIDs (e.g., hashes composed from different concatenated input variable combinations) either as a single pass or multiple passes? Is there a maximum number?	X				Yes and no maximum number
3	What is the mechanism for generating PPIDs (e.g., SHA-2 hashing)?					By default, Senzing ER uses an HMAC-SHA2-256 one-way hashing algorithm with a 1024-bit secret key. While this construct was developed for use in IPSec, we use it here for entity resolution with no modifications.
4	What probabilistic matching capability is available (e.g., Bloom filters on q-grams)?					Senzing uses a next generation Principle based Entity Resolution technology (see https://senzing.zendesk.com/hc/en-us/articles/231726307-Principle-based-Entity-Resolution). It combines statistical, deterministic, and active machine learning capabilities to provide a highly automated answer. It is not a traditional probabilistic matching engine.
5	How many files can be simultaneously linked (e.g., can the software link more than two files in one pass)?	X				Senzing loads data transactionally, this allows the system to be continuously loaded without having to be refreshed.
6	Does the product support deduplication?	X				
7	Can the software support more than pairwise linkages (e.g., find all the records in a file that match)?	X				

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
8	Does the product support two-party privacy protected linkage (identifying linkages between two files potentially belonging to two owners, where neither party sees the other's unencrypted data)?	X				Information is locally hashed at the source with a private salt key prior to being sent to the central system. (https://senzing.zendesk.com/hc/en-us/articles/360000970834-Selective-Field-Hashing)
9	Does the product support three-party protect linkage, where there is a trusted "honest broker" able to resolve possible linkages?	X				Information is locally hashed at the source with a private salt key prior to being sent to the central system. Once arriving at the central system the information is hashed again with a different salt. The providers salt key is not know to the third party and the third party salt key is not known to the providers. (https://senzing.zendesk.com/hc/en-us/articles/360000970834-Selective-Field-Hashing)
10	For three-party linkage, what information is made available to the trusted broker to resolve "possible" linkages (e.g., reports, distance metrics, comparisons of source data - if the broker can see source data)?	X				Due to our "Selective Hashing" Senzing can support at any level based on business requirements. All fields can be hashed or only those that are required to be hashed. The hashing process is one way and can not be undone. This means that once it is hashed then no one can see the data in the clear.
11	Are there any features for authorized reidentification of data?					No
12	What is tunable about matching criteria/algorithm?	X				Yes, Senzing comaprisons, principles and other areas are tunable via configuration and plugins, though the vast majority of users use the out of the box configuration.
13	Does the software have any ability to persist results so that subsequent linkages between data sets (e.g., after updates) can be incremental rather than from scratch?	X				
14	Does the software have the ability to split databases into linked vs. non-linked records, or other splitting and merging capability?		X			Unclear about this question, however Senzing can export the linked records or non-linked records as required.
15	Can the product persist PPIDs so they don't have to be regenerated for future runs?	X				Because the data is loaded in an incremental process, and the results are persisted, new data can be added without having to regenerate the intial results.

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
	Operating Environment and Licensing Model					
16	What Platform/OS(s) does the system run under?					CentOS 7 x86_64, RedHat 7 x86_64, Debian 9 / Ubuntu 16.04 x86_64, Amazon Linux 2016 x86_64
17	What other software is required to run your software (e.g., DBMS)?					RDBMS - IBM Db2, SQLite, PostgreSQL, MySQL / MariaDB - 5.6.5 / 10.1, AWS RDS
18	Minimum hardware specification.					16 GB RAM, 4 Modern CPU Cores, 100 GB Solid State Drive (SSD) or NVMe storage
19	Cloud-based version available? If so, which cloud environment?		X			Currently via our open source GitHub Senzing supports multiple options for cloud deployments (capabilities such as Docker, Kubernetes, Rancher, Helm, and others)
20	Licensing model (per seat, per CPU, open source, etc.).					Per Record ingested into the database
21	Does your licensing support "record linkage as a service", either through offering a cloud-based service or by distribution of the software as a utility?				X	
22	Is the system a set of components or a single, integrated software program? Does it require software to be developed to create a complete application?	X				Senzing is a set of API libraries that are available in C, Java, or Python. These libraries can be integrated into your application and wrapped with your business needs for entity resolution.
	Usability and Security Features					
23	Does the product include a graphical user interface (GUI)?			X		Senzing is a set of API libraries, however the Senzing GitHub Community has many graphical open source GUI components that can be leveraged or integrated into your solution.
24	Does the product include the ability to save configurations to facilitate multiple runs using the same parameters?	X				
25	Can the software be scripted to perform operations automatically?	X				Yes, because Senzing is a set of libraries they can be utilized as the end user needs. Additionally, we provide a complete set of Python tooling that can be utilized via any scripted process.
26	Does the software require configuration, or can it be used "out of the box"?	X				It can be utilized out of the box, and or fine tuned to the users needs if necessary.
27	Describe the product documentation available (provide link if possible).	X				Yes, https://senzing.com/developer/ or http://docs.senzing.com
28	When was the software first released?					2012

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
29	When was the most recent release of the software?					Jun-19
30	Is there an active development effort for the product?	X				Yes
31	Describe the product support available.	X				Bundled with licensing, Senzing provides Support Services (https://senzing.zendesk.com/hc/en-us/articles/236071408-Support-Services)
32	Is the software single-user or multi-user? If multi-user, how does the system manage integrity and security of data and ensure partitioning between users?					Because Senzing is a set of libraries, it can be deployed as you see fit, most use it as a multi user product. Security of data is dependent on how the database is configured. Senzing does not support "Partitioning" of data at the user level. This can be obtained by managing differnt "Database Schemas" for individual users.
33	Does the system contain security features such as requiring login/authentication?			X		Login and Authentication would be part of the application that is developed around the Senzing libraries.
34	Are there different user roles (e.g., administrator vs. user vs. data manager)?			X		User roles would be part of the application that is developed around the Senzing libraries.
35	If there are different roles, is the user interface segregated and optimized by role (e.g., a researcher would see the features of interest to an end-user, while an administrator would see a more full set of configuration functions)?			X		Please refer to the answer for #34
36	What execution performance reports are available (e.g., execution time, number of record pair comparisons, etc.).			X		The data to support such reports is available. Senzing does not provide load reports itself as that would be part of the solution that integrates the Senzing API libraries.
37	What linkage performance reports are available (e.g., number of matches, number of possible matches, number of duplicates - if the software does de-duplication, etc.).			X		Senzing tracks metrics within this area and can be queried through the API via multiple ways. Get Relationship Statistics. A few examples are: “Get Relationship Details”, “Get Entity Size Breakdown”, “Get Data Source Counts”, “Get Mapping Statistics” and others.
38	Has the system been approved to operate under U.S. government security regulations such as FISMA or FedRAMP?				X	
39	Can the system run in a mode which does not persist any data (to minimize security risks)?			X		There are some customized ways to utilize Senzing in a "Dynamic" mode vs "Persistent"

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
40	What protections are in place for source data?			X		Because Senzing is an API library, your business security requirements can be written into your application while integrating Senzing API and the RDMS chosen. In addition, Selective Hashing (https://senzing.zendesk.com/hc/en-us/articles/360000970834-Selective-Field-Hashing) and Secure Keystores (https://senzing.zendesk.com/hc/en-us/articles/360010578894--Advanced-Selective-Hashing-and-SoftHSM-How-To) can be utilized to support data base is one-way hashed prior to matching.
	External System Integration					
41	What file formats can the software use (e.g., delimited and fixed-width text files, MS Excel, XML, JSON)?					JSON or CSV
42	Does the software integrate directly with data sources for input and/or output (e.g., ODBC/JDBC integration with relational database, web services)? Which ones?					No
43	Can the software be configured to be flexible about input formats (e.g., mapping input columns to program variables), or must inputs be put into a particular format?				X	Inputs must be put into the expected JSON format.
44	What output formats does the software support?					JSON or CSV
45	Can the user customize the outputs?				X	No

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
	Data Cleaning / Pre-Processing Features					
46	Describe any features the tool has to identify data quality issues, and standardize, recode and clean data to improve matching performance (e.g., substitution of nicknames such as "Jim" to "James", address lookup and standardization, geocoding, phonetic matching).	X				Senzing is not a data cleansing tool. Data cleansing operations should be applied to your data before the data is submitted to Senzing. What we do is standardize the data received (see Uniquely Senzing https://senzing.com/uniquely-senzing/, in particular the section on Minimal Data Preparation.) to make sure it is formatted the same. This is easy enough with dates, ID numbers, phone numbers, etc. It's bit harder with names and addresses. To assist with Name we utilize a name matching software call IBM GNM. Addresses have historically been one of the most burdensome fields to deal with. Nearly all analytical engines require addresses to be parsed and standardized which is difficult and can require expensive and time consuming software options. New to Senzing, we now prefer a single ADDR_FULL address for scoring and have been seeing excellent results. This eliminates the burden in processing addresses and is key to fast time to value. Additionally, the engine will detect and stop using values that are overused or don't follow the "Behaviors" that the attribute type as been assigned to. Please refer to the Principle Based Entity Resolution article for additional information (https://senzing.zendesk.com/hc/en-us/articles/231726307-Principle-based-Entity-Resolution-FAQ-)
47	Is pre-processing specifiable field by field and file by file (e.g., a different date cleaning for DOB vs. record date, and for DOB in file 1 vs. file 2)?	X				Senzing does not provide ETL. In general, we take the data fields in the form you provide it (e.g. date in YYYY-MM-DD, 4th of July 1950, chinese gov't format, etc). See Uniquely Senzing (https://senzing.com/uniquely-senzing/), in particular the section on Minimal Data Preparation. However special processing may be required to derive, compare, or standardize data. Senzing allows you, as a programmer, to develop plugin modules to perform that special processing. Senzing can dynamically load plugin modules and invoke their routines, making the Senzing architecture indefinitely extensible. You can develop plugin modules to customize data standardization, expressed feature creation, and relationship and feature scoring. (https://senzing.zendesk.com/hc/en-us/articles/231970628-Developing-Your-Own-G2-Plug-ins)
48	Is the product extensible to use user-supplied pre-processing modules/services?	X				Please refer to answer for #47
49	Can the software export subsets of the pre-processed data fields (e.g., only certain columns, only certain rows)?				X	

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
50	Does the software support an evaluation mode (e.g., to allow researchers to work with their own data sets to clean and tune data before linkage)?	X				
	Performance and Scalability					
51	What is the maximum file size/number of records that the software can handle?					Senzing has been utilized against billions of records and as of yet we have not found our ceiling.
52	What is the largest use case for the software to date?					3 billion identity records
53	What features does the software have (such as blocking or database indexing) to improve performance? Can the user specify blocking parameters?	X				Senzing leverages blocking ("Expressed Features") and Database indexes for scale. The user can configure the system with different blocking parameters and even write plugins to generate new blocking concepts or simply provide the blocking values with the record.
54	Describe the ability to customize performance improvement features such as blocking?	X				Per #53, blocking parameters and concepts can be specified by the user. Senzing also implements active learning on specific data values and types to identify improperly behaving data and automatically adjust to handle it. The customer can provide input into the active learning process to help define its behavior if needed.
55	How can performance be improved by adding computational power (e.g., elastic compute)?	X				The Senzing API is a share nothing but the RDBMS configuration so the API compute nodes are horizontally scalable. The RDBMS itself can be easily spread over 3 DB nodes to near linear scaling.

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
	Use cases, applications and future capabilities					
56	Do you have any use cases, publications/white papers, demos or videos describing applications of your product? (please provide links or describe separately outside of this form as appropriate)?	[redacted]				Entity Resolution in Slow Motion: (https://www.youtube.com/watch?v=MPHd1eqU_yo) Privacy By Design: (https://jeffjonas.typepad.com/jeff_jonas/2012/06/privacy-by-design-in-the-era-of-big-data.html) Senzing Demo: (https://www.youtube.com/watch?v=O7oLUnWet8w) Jeff Jonas introducing Senzing: (https://www.linkedin.com/pulse/meet-senzing-g2-say-hello-entity-resolution-20-jeff-jonas/) Semantic Reconciliation - Entity Centric Learning: (https://jeffjonas.typepad.com/jeff_jonas/2007/04/to_know_semanti.html) Sequence Neutrality: (https://senzing.com/sequence-neutrality/)
57	Do you have any additional features planned or in development that you would be willing to share and feel we should know about?					We would be happy to follow up with a roadmap discussion.

13.7 HealthVerity

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
	PPID Generation and Record Linkage					
1	What can the user specify about the PPID generation process (e.g., which variables go into the PPID, and whether they should be truncated)?		X			While the user can specify different variable inputs, the matching is controlled centrally.
2	Can the software generate and link on multiple PPIDs (e.g., hashes composed from different concatenated input variable combinations) either as a single pass or multiple passes? Is there a maximum number?	X				HealthVerity's Census software is designed to use all of the input variables that are available, without restricting the system to any "required" variables. Our probabilistic matching engine treats any missing variables as hidden variables, essentially marginalizing across all possible values based on their typical expression rate. This lets the system use all of the available information for a potential link, without biasing towards a particular link based on more or less available information.
3	What is the mechanism for generating PPIDs (e.g., SHA-2 hashing)?					Each variable is transformed either with a salted SHA-256 hash or else with Bloom Filters that are subsequently hashed. The name (given name and surname) and the address (street and city) are each combined into their own record-level Bloom filters (RBF), while all other fields are SHA hashed or Bloom-filtered separately.
4	What probabilistic matching capability is available (e.g., Bloom filters on q-grams)?					Names are transformed with trigram Bloom filters, while other fields (address, phone numbers, emails, etc.) are transformed with bigram Bloom filters. Bigram Bloom filters in names were considered to be a re-identification risk through frequency analysis due to the well-known characteristics of bigram distributions in name. Probabilistic matching takes into account empiric distributions of bit differences between potentially linked Bloom filters from common forms of typos and misspellings. Subset analysis on the Bloom filter matching also supports probabilities for subsequences (such as accidental truncation, initials, and nicknames). Probabilistic matching is further supported with Bayesian probabilities to address frequencies of certain values (e.g. John Smith) as well as conditional frequencies (e.g. patient moving from NYC to either Philadelphia or Dayton, Ohio). Continual machine learning supports these probabilities.

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
5	How many files can be simultaneously linked (e.g., can the software link more than two files in one pass)?	X				All files are matched against a central referential database, which is continuously augmented with new information observed from the matching files. There is no inherent limit to the number of files that can be matched against the central repository in parallel. Any new patients which are not found in the central repository are matched against each other in a final resolution pass.
6	Does the product support deduplication?	X				
7	Can the software support more than pairwise linkages (e.g., find all the records in a file that match)?	X				After matching, each patient receives a unique HVID which is persistent across all records
8	Does the product support two-party privacy protected linkage (identifying linkages between two files potentially belonging to two owners, where neither party sees the other's unencrypted data)?	X				All matching is performed on de-identified tokens. Tokenization may be performed behind the owners firewall.
9	Does the product support three-party protect linkage, where there is a trusted "honest broker" able to resolve possible linkages?	X				All files are matched against a central referential database, which is continuously augmented with new information observed from the matching files. There is no inherent limit to the number of files that can be matched against the central repository in parallel. Any new patients which are not found in the central repository are matched against each other in a final resolution pass.
10	For three-party linkage, what information is made available to the trusted broker to resolve "possible" linkages (e.g., reports, distance metrics, comparisons of source data - if the broker can see source data)?	X				A list of the top N candidates, including relative probabilities of match, as well as which fields were present and which fields matched. For Bloom filters, counts of bit differences are available.
11	Are there any features for authorized reidentification of data?	X				While HealthVerity deals exclusively with de-identified data, Data Owners can include a record ID to support reidentification via the persistent HVID linked to the source record ID. Additionally, while HealthVerity's Consent product can help manage patient consent to enable reidentification, the actual reidentification would be done by the data owner.

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
12	What is tunable about matching criteria/algorithm?	X				The default Census matching criteria aims to maintain a 10:1 risk ratio, preferring missed links 10x more than incorrect links (false negative:false positive). This ratio is configurable. Also, the system operates on assumed error rates for mismatched variables and typos, which are driven by empirical observation and continual machine learning. A minor customization could allow these assumed error rates to be tuned.
13	Does the software have any ability to persist results so that subsequent linkages between data sets (e.g., after updates) can be incremental rather than from scratch?	X				The central matching repository remembers the tokens used by previous matching runs, and trivially applies the same identity number (HVID) to all subsequent matching requests. This HVID persists temporally and longitundinally.
14	Does the software have the ability to split databases into linked vs. non-linked records, or other splitting and merging capability?	X				Can also be visualized using the HealthVerity platform architecture
15	Can the product persist PPIDs so they don't have to be regenerated for future runs?	X				This is the default behavior
	Operating Environment and Licensing Model					
16	What Platform/OS(s) does the system run under?					Uses Java for Data Owner-side installs
17	What other software is required to run your software (e.g., DBMS)?					For on-premise use, HealthVerity Census requires Java Runtime Environment (JRE) v1.8 plus Java Cryptography Extension. HealthVerity also has a secure cloud-based version and API available.
18	Minimum hardware specification.					
19	Cloud-based version available? If so, which cloud environment?	X				AWS
20	Licensing model (per seat, per CPU, open source, etc.).					Per configuration + per file processed
21	Does your licensing support "record linkage as a service", either through offering a cloud-based service or by distribution of the software as a utility?	X				Both cloud-based and utility-based linkage as a service is available. HealthVerity also supports linkage as an API-based service.
22	Is the system a set of components or a single, integrated software program? Does it require	X				The de-identified tokenization and the matching engine are distinct programs. No software development is required.

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
	software to be developed to create a complete application?					
	Usability and Security Features					
23	Does the product include a graphical user interface (GUI)?				X	N/A - Data Owners do not interact with the matching directly.
24	Does the product include the ability to save configurations to facilitate multiple runs using the same parameters?	X				
25	Can the software be scripted to perform operations automatically?	X				
26	Does the software require configuration, or can it be used "out of the box"?	X				Typically, HealthVerity assists by configuring the tokenization software to conform to the user's file layout and formatting. However, the user can leverage the API-based service or conform to an existing layout for out-of-the-box performance.
27	Describe the product documentation available (provide link if possible).	X				HIPAA certification, configuration and access documentation
28	When was the software first released?	[redacted]				2015
29	When was the most recent release of the software?					Latest version 5.1.11, December 9, 2019
30	Is there an active development effort for the product?	X				Yes, we are continuing to expand capabilities as well as improve the performance and accuracy of Census.
31	Describe the product support available.	X				Documentation, support via HealthVerity depolyment engineer
32	Is the software single-user or multi-user? If multi-user, how does the system manage integrity and security of data and ensure partitioning between users?	[redacted]				N/A based on HealthVerity architecture
33	Does the system contain security features such as requiring login/authentication?	X				For transmission of files for matching
34	Are there different user roles (e.g., administrator vs. user vs. data manager)?				X	No

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
35	If there are different roles, is the user interface segregated and optimized by role (e.g., a researcher would see the features of interest to an end-user, while an administrator would see a more full set of configuration functions)?				X	N/A based on HealthVerity architecture
36	What execution performance reports are available (e.g., execution time, number of record pair comparisons, etc.).	X				We measure rows/sec (which vary by system) for de-identificaiton and matching
37	What linkage performance reports are available (e.g., number of matches, number of possible matches, number of duplicates - if the software does de-duplication, etc.).	X				We measure number of HealthVerity ID matches and duplicates (based on availability of person identifier on source data) as well as a host of information on variations and statistics of types of matches and non-matches.
38	Has the system been approved to operate under U.S. government security regulations such as FISMA or FedRAMP?			X		
39	Can the system run in a mode which does not persist any data (to minimize security risks)?	X				No PII is persisted by the system
40	What protections are in place for source data?	X				No unhashed, unencrypted source data is made available outside of Data Owner
	External System Integration					
41	What file formats can the software use (e.g., delimited and fixed-width text files, MS Excel, XML, JSON)?	[redacted]				De-limited or fixed width text files, JSON
42	Does the software integrate directly with data sources for input and/or output (e.g., ODBC/JDBC integration with relational database, web services)? Which ones?					No
43	Can the software be configured to be flexible about input formats (e.g., mapping input columns to program variables), or must inputs be put into a particular format?	X				Yes

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
44	What output formats does the software support?					Proprietary
45	Can the user customize the outputs?	X				Yes, based on input field mappings
	Data Cleaning / Pre-Processing Features					
46	Describe any features the tool has to identify data quality issues, and standardize, recode and clean data to improve matching performance (e.g., substitution of nicknames such as "Jim" to "James", address lookup and standardization, geocoding, phonetic matching).	X				The system automatically detects dozens of common data quality issues (e.g. first/last name swaps, salutations and name suffixes, date formatting, extraneous markups, etc) and automatically corrects many of these to ensure a standardized input to the tokenization. Additional diagnosis information is available to assist in debugging low match rates, including characterization of outliers in range and frequency, internal consistency (e.g. zip code and state), and aggregate statistics that can be mapped to common troubleshooting suggestions. Wherever possible, we prefer to allow the matching engine resolve standardization issues (such as nicknames) through advanced analysis of the Bloom filters. This allows the central referential database to maintain the observed variations for an individual and leverage those aspects to improve accuracy. It is our experience that excessive client-side standardization leads to more problems than it solves.
47	Is pre-processing specifiable field by field and file by file (e.g., a different date cleaning for DOB vs. record date, and for DOB in file 1 vs. file 2)?	X				Each variable type is configured to its own set of rules in the initial configuration process. If these rules need to change between files, the software would need to specify a new configuration.
48	Is the product extensible to use user-supplied pre-processing modules/services?		X			The product relies on text or JSON inputs, so as long as output is available in that format, yes.
49	Can the software export subsets of the pre-processed data fields (e.g., only certain columns, only certain rows)?	X				Yes, based on field mappings
50	Does the software support an evaluation mode (e.g., to allow researchers to work with their own data sets to clean and tune data before linkage)?	X				Linkage performed centrally so no need for researchers to tune- test mode is available
	Performance and Scalability					
51	What is the maximum file size/number of records that the software can handle?					Unlimited

ID	Question	Fully Meets	Partially Meets	Meets w/ Customization	Does Not Meet	Description/Comments
52	What is the largest use case for the software to date?					Currently processes over 20 billion records per year- typical run time for de-id exceeds [1mm] records per minute
53	What features does the software have (such as blocking or database indexing) to improve performance? Can the user specify blocking parameters?				X	Unneccessary to use blocking with the HealthVerity system
54	Describe the ability to customize performance improvement features such as blocking?				X	Unneccessary to use blocking with the HealthVerity system
55	How can performance be improved by adding computational power (e.g., elastic compute)?	X				Currently supports multi-processor matching
	Use cases, applications and future capabilities					
56	Do you have any use cases, publications/white papers, demos or videos describing applications of your product? (please provide links or describe separately outside of this form as appropriate)?					HealthVerity de-identification and matching is used by over 100 entities; over 50 billion records processed
57	Do you have any additional features planned or in development that you would be willing to share and feel we should know about?					The models and probabilities are continuously being improved for greater accuracy and flexibility while the backend infrastructure's goals focus around more real time and service level offerings; current enhancements include incorporating email and phone into matching if available.

14 Appendix 5: Full list of Products Examined

Key: Color Code

Y	Product description mentions PPRL capability
R	Related capability
N	Not PPRL software

Software List

PPRL?	Company/University	Product	Comment
Y	Datavant (acquired Health Data Link and Universal Patient Key)	Health Data Link/Datavant	Candidate
Y	Senzing	Senzing	Candidate
Y	GRHANITE	GRHANITE Entity Resolution	Candidate (didn't return survey)
Y	CSIRO (Australia)	Anonlink	Candidate
Y	Policywise	Linkwise	Candidate
Y	Crossix	SafeMINE	Candidate
Y	Privitar	Securelink	Candidate
Y	IBM	IBM Watson Financial Crimes Insight	Candidate (no response)
Y	HealthVerity	HealthVerity Marketplace, Census	Candidate
R	IQVIA	IQVIA Privacy Analytics	Related capability
R	Prognos AI (formerly Medivo)	NA	Related capability
R	Acxiom	Acxiom	Related capability
R	DataLadder	DataMatch	Related capability
Y	IBM	IBM Anonymous Resolution	Product end of life
Y	University of Chicago	DCIFIRHD (Distributed Common Identity for the Integration of Regional Health Data)	Not a product
Y	Universal Patient Key (UPK)	UPK Core	Acquired by Datavant
N	Verato	Iniversal MPI, Auto-steward	
N	Occam	EMPI	
N	SAIL	SAIL Databank	
N	Signet Accel	Avec	Out of business
N	Infoglide	Identity Resolution Engine	

PPRL?	Company/University	Product	Comment
N	CAPriCORN	CAPriCORN (Chicago Area Patient-Centered Outcomes Resrearch Network)	
N	FDA	FDA Sentinel	
N	IBM	IBM Entity Analytic Solutions package (IBM EAS)	
N	IBM	IBM Bigmatch	
N	IBM	IBM Quality Stage	
N	Privacy Analytics	Privacy Analytics Eclipse, PARAT	
N	Electronic Health Information Laboratory (EHIL)	EHIL	
N	CDC	Link Plus	
N	Link King	Link King	
N	Choicemaker	ChoiceMaker 2	
N	FEBRL	ANU Data Mining Group	
N	LinkageWiz	LinkageWiz	
N	NORC (at U. of Chicago)	G-Link	
N	LinkSolv	LinkSolv	
N	DataVance	DataVance	
N	MatchPro	MatchPro	
N	LinkPlus	LinkPlus	
N	Novetta	Novetta Entity Analytics	
N	Digital Reasoning	Digital Reasoning	
N	Feedzai	Feedzai	
N	Basis Technology	Rosete Entity Resolver	
N	NA	Latanya Sweeney's lab	
N	SAS	Dataflux	
N	RadiantOne	RadiantOne ICS	
N	Vynca	Patient Matching	
N	PICSURE	NA	
N	Information Softworks	EMPI	
N	Informatica	Allsight	
N	GDIT	NetOwl	
N	Imprivata	Imprivata	
N	Georgetown University	ATRA	High assurance computing, not PPRL

15 Appendix 6: Glossary and Acronyms

Term	Definition
Candidate Software Survey Results Table	A scoring matrix table showing, for each respondent, the question score, weight coefficient, and calculated Response Score for all questions in the Survey.
Feature/Capability	Group of related questions in a Question Category
Master Survey Questionnaire	A list of all questions included in the Final Version of the Survey. There is also a list of questions that were excluded from the final version.
Metered Response	A metered response to a survey question is one in which users can express a definitive answer to a particular question (i.e. Yes, No, etc.)
Narrative Response	A narrative response is one in which the answer cannot be quantitively measured like that of a metered response.
P3RLS	Privacy Protecting Patient Record Linkage Software. Software that performs PPRL functions, specifically for health and life sciences research domains.
PII	Personally Identifiable Information. Any data that could potentially identify a specific individual
PPID	Privacy Protecting Identifier. An identifier that is generated from PII, but from which the patient cannot be identified, for example, and encrypted, hashed identifier.
PPRL	Privacy Protecting Record Linkage. The process of linking disparate records together based on shared PII, but doing so in a privacy-protecting fashion
Priority	Priority is assigned a value for calculating a Weight Coefficient to be used in calculating a respondent's question response score.
Priority Point Ranking Value	Priority points are assigned to priority levels (i.e. Must Have = 3). Highest priority questions are scored with the highest point total, and successively less points for lower priorities.
Question Categories	Categories are assigned to types of questions based on the general function or capabilities under which the questions can be grouped.
Question Category Weight	The customer assigns a weight, expressed as a relative percentage of the value assigned to each question category. The sum of all category weights will equal 100%.
Question Response	The respondents answer (or response) to a survey question
Requirements Traceability Matrix (RTM)	A list of all feature/capability/performance requirements that constitute the scope of the survey questions.
Respondent Question Scores	The Respondent Question Score is calculated by multiplying each questions weight coefficient value by the respondent's response score.
Respondent Survey Results Scores	The sum total of a Respondent's question scores (by adding the scores for all 57 questions.)
Response Score	Responses to survey questions are assigned a numeric value on which to compute the response score. For example: a response of "Fully Meets" (the requirement/question) is given the highest score of 3. Lesser responses are given lower response score values.
Survey Results Summary Table	A table showing the Respondent Survey Results Scores for all Respondents who participated in the survey.

16 Bibliography

Baker, D. et al. *Technology Primer: Overview of Technological Solutions to Support Privacy-Preserving Record Linkage*. IRDiRC Technical paper. Version 4.0. 7 December 2017. Retrieved from http://www.irdirc.org/wp-content/uploads/2018/03/PPRL-Technical-Primer-V4-2.pdf.

Baker, D. et al. Privacy-Preserving Linkage of Genomic and Clinical Data Sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 31 July 2018. doi: 10.1109/TCBB.2018.2855125

Boyd, J., et al. Technical Challenges of providing record linkage services for research. *BMC Med Inform Decis Mak*. 31 Mar 2014:14;23. doi: 10.1186/1472-6947-14-23. Retrieved from ttp://www.biomedcentral.com/1472-6947/14/23.

Dimitropoulos, L. Privacy and Security Solutions for Interoperable Health Information Exchange. *Perspectives on Patient Matching: Approaches, Findings, and Challenges*. 30 June 2009. Retrieved from https://www.healthit.gov/sites/default/files/patient-matching-white-paper-final-2.pdf.

Durham, E., Xue, Y., Kantarcioglu, M., and Malin, B. Private medical record linkage with approximate matching. *AMIA Annu Symp Proc. 2010*:182-6.

Dusetzina, S.B., Tyree, S., Meyer, A.M., Meyer, A., Green L., and Carpenter, W.R. *Linking Data for Health Services Research: A Framework and Instructional Guide*. AHRQ Publication No. 14-EHC033-EF. Rockville, MD: U.S. Dept. of Health and Human Services, Agency for Healthcare Research and Quality; September 2014.

El-Emam, K., ed. *Risky Business: Sharing Health Data While Protecting Privacy*. Trafford Publishing; 2013.

El-Emam, K. *Guide to the De-Identification of Personal Health Information*. Auerbach Publications; 2013.

Emery, J. and Boyle, D. Data linkage. *Australian Family Physician*, Vol. 46, No. 8, Aug 2017: 615-619.

Gliklich, R.E., Dreyer, N.A, and Leavy, M.B, eds. Registries for Evaluating Patient Outcomes: A User's Guide. Third Edition. *AHRQ Publication No. 13(14)-EHC111*. Washington, D.C.: U.S. Dept. of Health and Human Services, Agency for Healthcare Research and Quality; 2014

Hendler, J. Data Integration for Heterogenous Datasets. *Big Data*. 2014;2(4):205–215. doi:10.1089/big.2014.0068

Hosek, S. and Straus, S. *Patient Privacy, Consent, and Identity Management in Health Information Exchange: Issues for the Military Health System*. Santa Monica, CA: RAND Corporation. 2013. Retrieved from https://www.rand.org/pubs/research_reports/RR112.html

International Organization for Standardization. Systems and software engineering -- Systems and software Quality Requirements and Evaluation (SQuaRE) -- System and software quality models. March 2011. Standard ISO/IEC 25010:2011.

Kho, A.N., Cashy, J.P., Jackson, K.L., Pah, A.R., Goel, S., Boehnke, J., Humphries, J.E., Kominers, S.D., Hota, B.N., Sims, S.A., et al. Design and implementation of a privacy preserving electronic health record linkage tool in Chicago. *Journal of the American Medical Informatics Association*. 2015;22(5):1072–1080. doi:10.1093/jamia/ocv038

Kum, H-C., Krishnamurthy, A., Machanavajjhala, A., Reiter, M.K., and Ahalt, S. Privacy preserving interactive record linkage (PPIRL). *Journal of the American Medical Informatics Association*. 2014;21(2):212–220. doi:10.1136/amiajnl-2013-002165

Lin H., Lai A., et. al. COTS Software Selection Process. Sandia National Laboratories Report SAND2006-0478. May 2006. Retrieved from https://prod-ng.sandia.gov/techlib-noauth/access-control.cgi/2006/060478.pdf.

Morris, G., Farnum, G., Afzal, S., et al. *Patient Identification and Matching Final Report. Prepared for the Office of the National Coordinator for Health Information Technology*. 7 February 2014. Retrieved from https://www.healthit.gov/sites/default/files/patient_identification_matching_final_report.pdf.

Pal, D., Chen, T., Zhong, S., and Khethavath, P. Designing an Algorithm to Preserve Privacy for Medical Record Linkage with Error-Prone Data. *JMIR Medical Informatics*. 2014;2(1):e2. doi:10.2196/medinform.3090

Pew Charitable Trust. *Enhanced Patient Matching Is Critical to Achieving Full Promise of Digital Health Records*. 2 October 2018. Retrieved from https://www.pewtrusts.org/en/research-and-analysis/reports/2018/10/02/enhanced-patient-matching-critical-to-achieving-full-promise-of-digital-health-records.

Philps, M., et al. *Privacy-Preserving Record Linkage: Ethico-Legal Considerations*. IRDiRC Technical Report. March 2018. Retrieved from http://www.irdirc.org/wp-content/uploads/2018/03/Rare-Genetic-Diseases-Workshop_final-version_public.pdf.

Schmidlin, K., Clough-Gorr, K.M., and Spoerri, A. Privacy Preserving Probabilistic Record Linkage (P3RL): a novel method for linking existing health-related data and maintaining participant confidentiality. *BMC Medical Research Methodology*. 2015;15(1).

Schnell, R., Bachteler, T., and Reiher, J. Privacy-preserving record linkage using Bloom filters. BMC Medical Informatics and Decision Making. 2009;9(41). Doi:10.1186/1472-6947-9-41.

Sequoia Project. *A Framework for Cross-Organizational Patient Identity Management*. 2018. Retrieved from https://sequoiaproject.org/resources/patient-matching/.

Sweeney, L. *Privacy-Enhanced Linking*. ACM SIGKDD Explorations. Dec 2005;7(2).

Swire, P. Research Report: Application of IBM Anonymous Resolution to the Health Care Sector. Feb 2006. Retrieved from http://peterswire.net/archive/anon.resolution.whitepaper.pdf.

Vatsalan, D., Christen, P., and Verykios, V.S. A taxonomy of privacy-preserving record linkage techniques. *Information Systems*. 2013;38(6):946–969. doi:10.1016/j.is.2012.11.005

Vatsalan, D., Christen, P., O'Keefe, C., and Verykios. V.S. An Evaluation Framework for Privacy-Preserving Record Linkage. J. Privacy and Conf. 2014;6(1):35-75.

Vatsalan, D., Christen, P., Sehili, Z., and Rahm E. *Privacy-Preserving Record Linkage for Big Data: Current approaches and Research Challenges*. 2017. In: Zomaya A, Sakr S (eds). Handbook of Big Data Technologies. Springer, Cham. doi:10.1007/978-3-319-49340-4_25.

Vatsalan, D., Sehili Z., Christen, P. and Rahm, E. *Privacy-Preserving Record Linkage for Big Data: Current Approaches and Research Challenges*. (2017). Doi:10.1007/978-3-319-49340-4_25.

Wagner, I., and Eckhoff, D. *Technical Privacy Metrics: A Systematic Survey*. ACM Computing Surveys. June 2018, Vol. 51:(3), Article 57. doi:10.1145/3168389.

This page intentionally left blank.