10. The Iowa Experiment

On June 6, 2007, a Planning Office in Iowa, along with three community Shelters and the area's HMIS tested the PrivaMix Demonstration System in three experiments. Jointly, we term these "the Iowa Experiment." One experiment concerned the uniqueness of Client source information that did not use Social Security numbers. One experiment used a network of computers to test the ability of the software to de-duplicate. The final experiment examined the identifiability of the de-duplicated results. Below are details of these experiments.

10.1 Materials

Below is a description of the materials used in the Iowa Experiment.

10.1.1. Computers

The Iowa Experiment used five laptops in their original factory configurations. Four of the machines were Toshiba Satellite M115 laptops. Each Toshiba machine had an Intel Celeron M at 1.6GHz processor running the Windows XP operating system, 448MB of RAM memory, and 74GB of hard drive space. Among standard ports, each machine included a PCMCIA port. The original cost was about $500 each.

The Dell laptop had an Intel Centrino Duo (two processors) at 1.83Ghz, running the Windows XP operating system, 2 GB of RAM memory, and 93GB of hard drive space. Among standard ports, this machine included a PCMCIA port. The original cost was about $2000. The Dell laptop was significantly more powerful that the Toshiba machines.

These machines do not reflect the minimum machine requirements, as much as a description of the actual machines used. By providing standard laptops rather than using machines already at Shelters, these experiments were able to focus on performance issues rather than software installation and other secondary problems that can emerge in attempting to load software on unknown machines.

10.1.2. Network

Even though each laptop had all the standard Internet connection options (modem, wireless Internet, and Ethernet) built-in, the Iowa Experiment used five wireless broadband cards (4 Verizon and 1 Sprint), one per machine. The Verizon cards used the PCMCIA slots on the laptops. The Sprint card used the USB port.

A wireless broadband card communicates directly with a wireless mobile phone network to send and receive information over the Internet. This is usually slower than dial-up or cable Internet options.

The PrivaMix Demonstration System does not require the use of wireless broadband access to the Internet. By using these cards in standard laptops, the experiments did not have to assume participants were technically able to provide Internet access to the laptops.

Together, the laptops and network cards provided standardized hardware so that the experiments focus efficiently and narrowly on de-duplication performance.

10.1.3. PrivaMix Demonstration System

Each laptop ran an edition of PrivaMix Demonstration System (version 0.36). One machine designated as the Planning Office machine ran the CoC edition. The other four machines ran the Shelter Edition. Section 9 contains a detailed description of the PrivaMix Demonstration System. Appendix A has a copy of the User's Guide.

10.2 Subjects

Subjects are clients whose data appeared at participating shelters and the HMIS. The actual subjects are not clients of domestic violence ("DV") homeless shelters, but are clients of homeless family shelters (not domestic violence specific). Using non-DV shelters allowed us to compare computed de-identified results with results derived manually using fully identified data.

A downside to using non-DV shelters is that differences in data collected in DV versus non-DV shelters may exist and would not reflect in results. Therefore, the generalizability of these experiments assume there is no difference between DV and non-DV data collection. This assumption seems reasonable given perceived similarities in client populations. (See Section 10.3.6 for a field-level compliance comparison.)

Below is a description of participants.

- Iowa Institute for Community Alliances, participated in its role as the Planning Office or CoC in Des Moines, Iowa.
- HMIS in DesMoines, Iowa participated as a Shelter in its role to de-duplicate across Shelters and the HMIS. These are the same system administrator at Iowa Institute for Community Alliances.
- House of Mercy in Des Moines, Iowa participated as a Shelter.
- New Directions in Des Moines, Iowa participated as a Shelter.
- YWCA in Des Moines, Iowa participated as a Shelter.

For the remainder of this section, the term "Clients" refers to the Clients represented in data, even though they are not actual DV clients. The term "Shelters" refers to House of Mercy, New Direction, YWCA, and sometimes HMIS. Other times, the HMIS is identified separately. The inclusion or exclusion should be obvious by context. The term "Planning Office" refers to the CoC for DesMoines, Iowa.

10.2.1. Data

Data used in the experiments consisted of retrospective Client data (January through June 2006). Shelters previously provided these records to the HMIS for producing an AHAR. Below is further description of data content and handling.

Shelter	Gold Standard Number of Records	Test Database Number of Records	Modified Test Database Number of Records
HMIS	1937	1937	1937
House of Mercy	59	59	59
New Directions	132	132	132
YWCA	------	------	66
Total	2128	2128	2194

Figure 71. Number of Client records in Gold Standard and Test databases by participant. The Gold Standard Database includes manual corrections and inclusion of missing information for those records known to be of the same Clients. The Test Database lacks these modifications, containing the original errors and omissions. The Modified Test Database is the same as the Test Database with 66 records added to generate more common visits across participants. All databases have the same 1570 distinct Clients.

1 The likelihood that the fields contain omissions or errors is small.

2 The likely number of possible distinct combination of values across the fields must be sufficiently large to be unique for each Client.

3 The Client is likely to provide the same values for the fields at each Shelter.

Figure 72. Conditions for selecting fields for Client source information.

Privacy

In order to produce the initial dataset and to analyze some of the experimental results, personnel needed access to identifiable Client information. The only persons who had such access to identifiable data was the existing HMIS and Shelter personnel from whom the data originated.

Gold Standard Database

System administrators[22] at the HMIS took on the laborious task of extracting identifiable Client data from the HMIS originally contributed by House of Mercy and New Directions. System administrators then manually reviewed the data, manually correcting errors and entering omissions, so that records believed to belong to the same person had accurate information in fields that may form the basis of generating UIDs. The fields subject to correction were *first name*, *last name*, *gender*, and *date of birth*. The total number of records was 2128 for 1570 distinct Clients. This comprised our "Gold Standard" database. The data elements are the Universal Data Elements, including name (as *first name* and *last name* fields) and *Social Security number* (see Figure 5). Figure 71 lists the total records by Shelter.

Test Database

The Test Database contains the same records as the Gold Standard Database, except the records are in their originally unchanged form. None of the values reflect the manual cleaning done in the

[22] Eileen Mitchell, HMIS system administrator for the HMIS in Des Moines, Iowa, performed the labor of producing the Gold Standard Database and supervised its use.

Gold Standard Database. Secondly, the Test Database includes an additional 66 records assigned to the YWCA in order to provide additional visits occurring across more Shelters. The total number of records was 2194 for the same 1570 distinct Clients. The data elements are the Universal Data Elements, including name and Social Security number (see Figure 5). Figure 71 lists the total records by Shelter.

Modified Test Database

The Modified Test Database is a copy of the Test Database with a major change to fields and records. Changes to the fields include dropping, modifying, and re-ordering. Specific fields dropped: *last name* and *Social Security number. Fields changed: first name* to be only the first three letters of the first name. The order of fields is: *first 3 letters of the first name*, *date of birth*, *year of birth*, *race*, *gender*, *veteran*, *disability*, *prior residence type*, *prior residence days*, *ZIP*, *entry date*, *exit date*, *provider ID*, *group ID*, and *program ID*. (See Figure 5 for field descriptions.) The Client source information is the two leftmost fields, *first 3 letters of the first name* and *date of birth*. The remaining fields, *year of birth* through *program ID*, comprise the Universal Data Elements. Records added: 66 records assigned to the YWCA in order to provide additional visits occurring across more Shelters. The total number of records was 2194 for the same 1570 distinct Clients. Figure 71 lists the total records by Shelter.

10.3 Experiments: Client source information

A key component in de-duplicating UIDs is the Client source information used to construct the UIDs. Fields having omissions or errors can render UIDs useless. Experiments in this section compared traditional and proposed choices for constructing UIDs.

Figure 72 lists three conditions for fields to satisfy to be good choices for Client source information.

Problem Statement.

> *Given traditional and proposed ways of constructing UIDs (see Sections 10.3.1, Section 10.3.2, Section 10.3.3, and Section 10.3.4), determine which ways best satisfy the three conditions for constructing UIDs listed in Figure 72.*

The next four subsections describe different ways to construct UIDs.

Position	Content
1	First letter of first name
2	First letter of last name
3	Third letter of last name
4	First letter of gender
5	Date of Birth yyyymmdd (or all 0's if present)
12	Soundex of first name
16	Soundex of last name

Figure 73. Servicepoint Client UID encoding. The result is a 20 character code.

Step	Description
1	Copy the first letter of the string
2	Remove all occurrences of the following unless it is the first letter of the string: a, e, h, i, o, u, w, y
3	From the second letter forward, assign the following number to letters: for b, f, p, v, assign 1 for c, g, j, k, q, s, x, z, assign 2 for d, t, assign 3 for l, assign 4 for m, n, assign 5 for r, assign 6
4	If two or more adjacent numbers repeat, keep only the first.
5	Return the first four characters, padding 0's on the right if needed.

Figure 74. Soundex algorithm. Given a string, the Soundex algorithm provides a 4-character code. Examples: Washington (W252), Robert and Rupert (R163).

Position	Content
1	First letter of first name
2	First letter of last name
3	Third letter of last name
4	Date of Birth yyyymmdd (or all 0's if present)

Figure 75. Servicepoint Client UID encoding variant. This version differs from Figure 73 by not including gender or Soundex. The results is a 11 character code.

Position	Content
1	First three letters of first name
4	Date of Birth yyyymmdd (or all 0's if present)

Figure 76. Privacert proposed Client UID encoding. The result is an 11 character code.

10.3.1. Social Security Number

The Social Security number is perhaps the most common way to reference people in data. This is a 9-digit value, being uniquely assigned to most people in the United States.

10.3.2. Servicepoint Client Unique ID

The most common UID used within HMIS systems, and used by the HMIS in Iowa, is the Servicepoint Client Unique ID.[23] A Servicepoint Client Unique ID is 20 characters, encoded as described in Figure 73.

The Servicepoint encoding uses Soundex[33], which is a phonetic algorithm for encoding names in 4 characters. The Soundex algorithm appears in Figure 74.

10.3.3. Servicepoint Client Unique ID Variant

A simple variant to the Servicepoint UID encoding described above in Section 10.3.2 does not include gender or Soundex. The result is a 11 character encoding, as described in Figure 75.

10.3.4. Proposed Privacert Method

While the PrivaMix Demonstration System works with any Client source information, Privacert proposed one combination of fields for consideration using only the first name and the date of birth. The result is a 11 character encoding, as described in Figure 76.

10.3.5. Experimental design

Using the Test Database, proposed method for constructing UIDs were compared in terms of addressing the three conditions described in Figure 72. Results appear in Section 10.3.6.

[23] Servicepoint is a product of Bowman Systems, servicing more than 30,000 clients in 45 states. They are a national leader in providing HMIS services. For more information, see http://www.bowmansystems.com/products.html.

10.3.6. Results

The following results: (1) compare data compliance of domestic violence shelters to non-DV shelters, (2) report the number of blank values found in the fields of interest to the four methods of constructing UIDs mentioned above; (3) report the number of records effected by data discrepancies in fields relied on by the four methods; and, (4) a summary of how each of the four methods address the three conditions identified as important to the construction of UIDs.

Comparison of DV to non-DV data compliance.

Figure 77 shows previously reported results compiled by Abt[24] from Iowa's Planning Office that compare percentages of missing information in Iowa DV versus non-DV. Data for DV shelters result from on site visits and therefore reflect data maintained by DV shelters internally. Because of the changes to the Universal Data Elements (see Figure 5), first and last name is not required, so DV shelters only collect these fields a half to one-quarter of the time. DV shelters rarely collect Social Security numbers. Values routinely appear for dates of birth. The accuracy of none of the values is known with the ad hoc observation that many dates of birth share the same January 1 value, but with different years.

Number of blank values.

Figure 78 shows the number of blank values found in noted fields in the Test Database, as compiled and previously reported by Abt [34]. Most records lacked a middle initial. Of the fields used by the four methods for constructing UIDs described above, most values were present.

Comparison of UIDs effected by bad or missing data.

Figure 79 shows a comparison of the number of UIDs effected by bad or missing data in the Test Database, as compiled and previously reported by Abt [34]. The comparison is between Servicepoint (Section 10.3.2) and the proposed Privacert (Section 10.3.4) methods. In total, 88 UIDs were negatively impacted using Servicepoint's method compared to only 56 for the proposed Privacert method. Because the proposed Privacert method uses fewer fields that can contain bad or missing data, it performed better. Errors found in the last names or gender fields resulted in bad UIDs for the Servicepoint method while having no adverse effect on the proposed Privacert method.

Summary of UID methods.

Figure 80 compares results of the four methods of UID construction (Section 10.3.1, Section 10.3.2, Section 10.3.3, and Section 10.3.4) in terms of the three conditions found important to constructing UIDs (Figure 72).

Condition 1: The fewer the number of UIDs adversely effected by omission of errors found in the data, the better the method's performance. Values are copied from the earlier results in Figure 78 and Figure 79 for the SSN (264), Servicepoint (88), and proposed Privacert (56) methods. The 84 UIDs negatively effected by omissions or errors using the Servicepoint variant (Servicepoint2) was

[24] Results that are noted in this writing as being compiled and previously reported by Abt appear in [34]. During the Iowa experiment, some analyses required access to identifiable HMIS data. This was done by Brian Sokol at Abt under the supervision of the system administrators of Iowa's HMIS. They computed their results independent of this author for privacy reasons and for the benefit of having an independent third party perform such analyses. Results of analyses done by this author are those appearing without credit to Abt.

inferred from the 88 effected by Servicepoint. The Servicepoint variant does not use gender, which accounted for 4 errors in Figure 79. Overall, the proposed Privacert method performed best.

Condition 2: The larger the number of distinct combinations of the fields, the greater the number of Clients within a CoC that can use the method.

Assuming all possible digits are possible with a SSN gives 10^9 possible values.

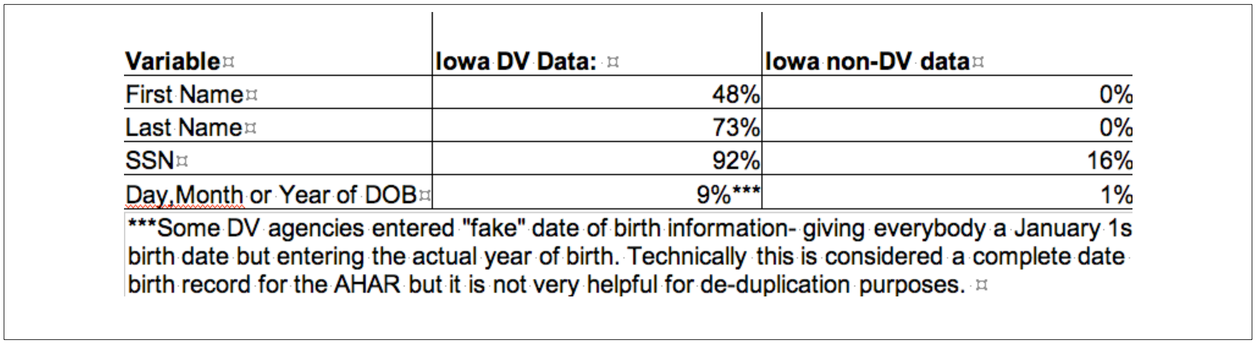
The Servicepoint encoding on has 26 * 2 * 36500 * 156 * 156 = 46 * 10^9 possible values. There are 26 different letters, 2 different genders, 36,500 dates of births (assuming 100 year age range), and 156 possible Soundex values. Using the first letter of the first name and the first letter of the last name is redundant with the Soundex code so those values are not included.

The Servicepoint variant has 26 * 26 * 26 * 36500 = 641 * 10^6 possible values.

Similarly, the proposed Privacert method has 26 * 26 * 26 * 36500 = 641 * 10^6 possible values.

In summary, Social Security numbers and the Servicepoint encoding can accommodate the most number of Clients within a CoC. The Servicepoint variant and the proposed Privacert method are comparable. The numbers computed are the maximum possible. Not all letters are equally likely in names and not all dates of birth over a 100 year range are equally likely to be Clients. Nonetheless, all four methods seem reasonable for the subjects in this study.

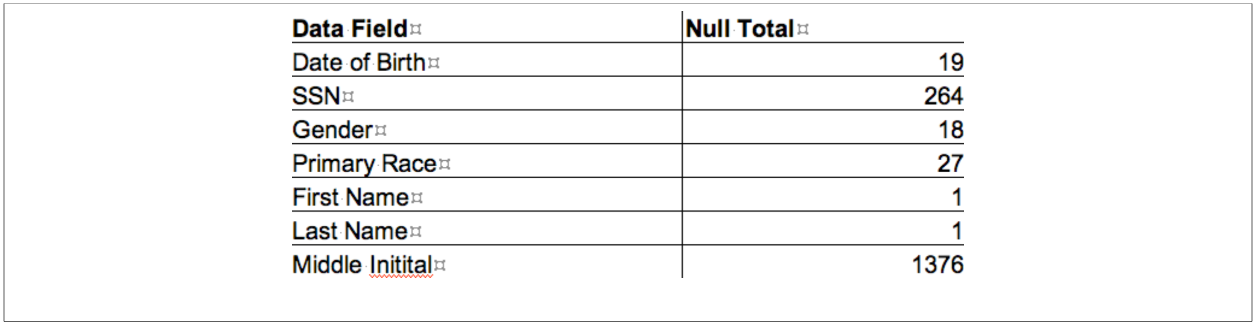
Condition 3: Consistency of values cannot be measured without de-duplication which was not done in this set of experiments (see Section 10.4).



Variable¤	Iowa DV Data: ¤	Iowa non-DV data¤
First Name¤	48%	0%
Last Name¤	73%	0%
SSN¤	92%	16%
Day,Month or Year of DOB¤	9%***	1%

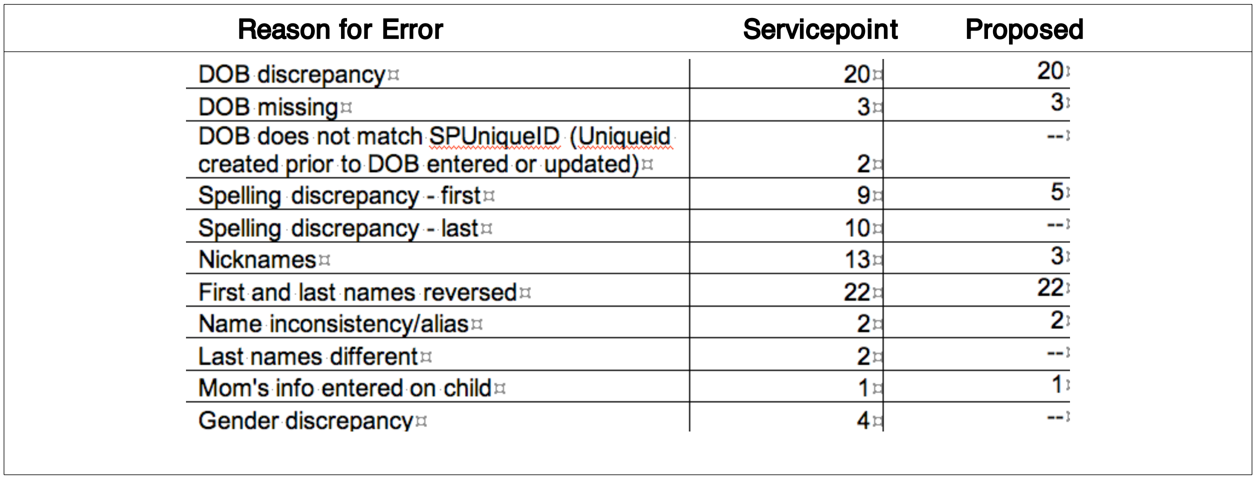
***Some DV agencies entered "fake" date of birth information- giving everybody a January 1s birth date but entering the actual year of birth. Technically this is considered a complete date birth record for the AHAR but it is not very helpful for de-duplication purposes. ¤

Figure 77. Percent of missing data for DV and non-DV shelters. Fields are first name, last name, Social Security numbers (SSN), and dates of birth (DOB) of Clients. Courtesy Abt Associates [34].



Data Field	Null Total
Date of Birth	19
SSN	264
Gender	18
Primary Race	27
First Name	1
Last Name	1
Middle Initital	1376

Figure 78. Number of missing values found in Test Database. A missing or "null" value has no value appearing in the database. Counts based on 2128 records. Courtesy Abt Associates [34].



Reason for Error	Servicepoint	Proposed
DOB discrepancy	20	20
DOB missing	3	3
DOB does not match SPUniqueID (Uniqueid created prior to DOB entered or updated)	2	--
Spelling discrepancy - first	9	5
Spelling discrepancy - last	10	--
Nicknames	13	3
First and last names reversed	22	22
Name inconsistency/alias	2	2
Last names different	2	--
Mom's info entered on child	1	1
Gender discrepancy	4	--

Figure 79. Comparison of UIDs effected by bad or missing data. Compares Servicepoint's UID construction (Section 10.3.2) to the proposed Privacert method (Section 10.3.4) using records in the Test Database. Counts based on 2128 records. Courtesy Abt Associates [34].

Method	Omissions or Errors in Fields	Number of distinct combinations	Consistency of values
SSN	264	10^9	84.60%
Servicepoint	88	$46 * 10^9$	96.40%
Servicepoint2	84	$641 * 10^6$	97.50%
Proposed	56	$641 * 10^6$	97.90%

Figure 80. Comparative summary of four UID methods in real-world data. Results are using only Social Security numbers (SSN), the Servicepoint method (Section 10.3.2), the Servicepoint variant (Section 10.3.3), and the proposed Privacert method (Section 10.3.4) in the Test Database (Section 10.2). The fewer the number of UIDs adversely effected by omission of errors found in the data, the better the method's performance. The larger the number of distinct combinations of the fields, the greater the number of Clients within a CoC that can use the method. Consistency of values is 1-error percentage in Figure 83 (Section 10.4.2). Counts based on 2128 records.

10.4 Experiments: de-duplication

A primary motivation for this work is the utility of de-duplicating UIDs in order to match Client visit information across Shelters. Experiments in this section measured the performance of the PrivaMix Demonstration at de-duplicating real-world data.

10.4.1. Experimental design

Records in the Modified Test Database were divided into smaller databases, one for each of the participants that originally provided the information. The result was four smaller databases, one for each of the three Shelters, and one for the HMIS. Figure 71 shows the distributions of records.

There was a total of 2194 records, with more than half (1937) originating from the HMIS alone. Records originating in the HMIS in this experiment reflect "non-DV" services provided to DV and non-DV clients, indistinguishably. Client information held in the Shelter databases represent "DV" Clients in this experiment. The goal is to de-duplicate visits across DV and non-DV services.

Problem Statement.

> *Given three Shelters, an HMIS, and a Planning Office participating in a PrivaMix network, use the PrivaMix Demonstration System to de-duplicate visits.*

Each of the smaller databases was loaded onto a laptop as a comma-delimited file. Figure 68 lists the fields that comprised the comma-delimited file. The first two fields denote the Client source information. These are *FirstName* and *DateOfBirth*. The remaining 13 fields of the Universal Data Elements (Figure 5) stored values describing the service received by the Client.[25]

The HMIS used the faster Dell machine. The remaining Shelters and the Planning Office used the Toshiba machines. The Dell also used the Sprint wireless modem card, whereas the other Shelters and the Planning Office used the Verizon cards.

The files were saved on each computer with a filename matching the default setting in the PrivaMix Demonstration System. The number of leftmost fields designated to use as Client source information for generating UIDs (2) also matched the default setting in the PrivaMix Demonstration System. The goal was not to assess the flexibility of the software or user's ability to use the laptop per se. Operation was made to be as simple as possible. Upon powering on the machine, the broadband wireless card automatically connected to the Internet and the PrivaMix software loaded. The user need only power on the machine and click the De-duplicate button at the designated time. See Quick Start in Appendix A (page 5 of 16).

The three Shelter machines and the HMIS machine contained the Shelter Edition of the PrivaMix Demonstration System (Section 9). The Planning Office machine contained the CoC Edition of the PrivaMix Demonstration System (Section 9).

Personnel from the HMIS physically visited each Shelter, one at a time. The machine containing that Shelter's information was left with the Shelter. A five minute discussion reviewed the security

[25] Personnel from the HMIS actually loaded the data onto the laptops and maintained control of the laptops until providing the machines to the respective Shelters.

of the machine, the agreed upon time at which de-duplication would occur, the process of powering on the machine, and the need to click the "De-duplicate" button to start.

The agreed upon date and time to start the process was June 5, 2007 at 3pm. At that time, each participant would power on their respective machine at their physical location and then start the de-duplication process.

Once the process begins, there are four distinct phases.

In Phase I, all participants, including the Planning Office, start their machine and click the De-duplicate button. The software will register the machine by sharing IP addresses among only those computers previously known to be participants in the PrivaMix network. See Section 9.1.1 for details.

After all machines complete Phase I, the machines automatically begin Phase II. Each of the Shelter and HMIS machines load the comma-delimited file containing Client information specific to that Shelter or HMIS. The machine then randomly selects a private value (see Section 9.3). The machine then computes UIDs and forwards results to the Planning Office machine. See Figure 68 for examples.

Once the Planning Office machine receives the Client information from the other machines, it initiates mixing, which constitutes Phase III. The Planning Office contacts each Shelter and HMIS, one a time, to mix UIDs and mixes from the other Shelters and HMIS. See Section 9.7 for details.

Once all Shelters and HMIS have mixed all UIDs, Phase IV, the last phase begins. The Planning Office machine de-duplicates UIDs by matching records based on complete mixes. It then re-numbers UIDs and GroupID values sequentially. Finally, comma-delimited results are then stored to the hard drive of the Planning Office machine. See Figure 69 for examples. See Section 9.8 for processing details.

10.4.2. Results

Results: (1) show the time taken for each phase of de-duplication; and, (2) compare de-duplication results.

Time spent.

From the start of the de-duplication process at the designated time until the delivery of the de-duplicated results on the Planning Office machine took 71 minutes. During this time, Shelters forwarded 2253 Client records, thereby mixing 2253 UIDs over four Shelters and HMIS machines. Figure 81 shows the amount of time spent in each phase, as compiled and previously reported by Abt [34].

Despite the Toshiba computers being identical and having to mix the same number of UIDs, the first Shelter took twice as long (20 minutes) to complete mixing in comparison to the other two Shelters using Toshiba machines (11 minutes). The reason for the discrepancy is not clear. It may delay in the Internet connection or start-up overhead. The Dell laptop, used by the HMIS, is much faster. It took only 3 minutes!

During operation, one of the Shelters (House of Mercy) accidentally shut down their machine and had to restart. The program had not anticipated a restart in Phase 1. The result was the double inclusion of their 59 records. It was as if each of their Clients visited them twice.

Phase	Description	Time Completed
Phase 1	All participants run PrivaMix software.	03:00:00 PM
	Network registration.	03:01:00 PM
	Restart by House of Mercy after accidental shutdown.	03:08:00 PM
Phase II	Compute UIDs and forward data	03:25:00 PM
Phase III	Mix Shelter 1 (House of Mercy)	03:45:00 PM
	Mix Shelter 2 (HMIS)	03:48:00 PM
	Mix Shelter 3 (New Directions)	03:59:00 PM
	Mix Shelter 4 (YWCA)	04:10:00 PM
Phase IV	Produce de-duplicated result (CoC)	04:11:00 PM

Figure 81. PrivaMix Demonstration: time spent per phase. Courtesy Abt Associates [34].

Manual de-duplication of Gold Standard Database.

Figure 82 reports manually produced de-duplication results on the Gold Standard Database for three different UID methods, as compiled and previously reported by Abt [34]. Manual de-duplication was done by constructing UIDs with a noted method and then matching results to get a distinct count. Servicepoint and the proposed Privacert method both provided an accurate de-duplicated count of 1570 Clients. Matching Social Security numbers (SSNs) only found 1330 of the Clients because 240 records had no SSN.

Manual de-duplication of Test Database.

Figure 83 reports manually produced de-duplication results on the Test Database for four different UID methods, as compiled and previously reported by Abt [34]. Manual de-duplication was done by constructing UIDs with a noted method and then matching results to get a distinct count. Some Social Security numbers (SSN) were missing, leading to an undercount by that method. All other methods had an over count. The proposed Privacert method performed best, though comparable to the Servicepoint variant. Because Privacert did not use the last name field, an error found there did not effect its performance as it did with the Servicepoint variant.

Consistency of values.

The third condition for selecting fields for Client source information (Figure 72) involves computing the likelihood a Client will provide the same values at each Shelter visited, based on the fields used by the noted UID method. This writing terms this “the consistency measure.” The error percentage in Figure 83 provides a basis for a consistency measure as the inverse of the error percentage, computed as [1.0 – (error percentage)]. This measures the accuracy of de-duplication across records from different Shelters. In manual de-duplication of records in the Test Database, Social Security numbers had the worst consistency (84.6%) because the value was sometimes missing. The proposed Privacert method had the best consistency (97.9%). The Servicepoint variant was comparable (97.5%). The Servicepoint method did a little worse (96.4%). These values appear in Figure 80, as part of a comparative summary of UID methods in terms of selecting Client source information.

PrivaMix de-duplication of Modified Test Database.

Figure 84 compares de-duplication results from the PrivaMix Demonstration System on the Modified Test Database with the earlier manual results on the Test Database, as compiled and previously reported by Abt [34]. The PrivaMix Demonstration System performed exactly the same as the manual results predicted (Figure 83). The System did not introduce any errors and made the same decisions on all records as constructing encodings in plain text (Figure 76). Shelters constructing UIDs from the plain text did not generate any mismatches. Mixing UIDs and then matching on the complete mixes introduced no omissions or errors. The PrivaMix Demonstration System performed exactly as if plain text encoding was used even though the Client information was provably never shared with the Planning Office or the other Shelters.

UID Method	Unduplicated Count (A)	False Negatives (B)	False Positives (C)	Error Percentage
SSN	1330	0	240	11.3
Servicepoint	1570	0	0	0
Proposed Privacert	1570	0	0	0

Figure 82. Manual de-duplication results on Gold Standard Database. In the 2128 records, the number of distinct Clients is 1570. Some Social Security numbers (SSN) were missing, leading to an undercount by that method. A false positive results when two distinct Clients are counted as one. A false negative results when a record belonging to a known Client is missed. The error percentage is (B) + (C)/2128 * 100. Courtesy Abt Associates [34].

UID Method	Unduplicated Count (A)	False Negatives (B)	False Positives (C)	Error Percentage
SSN	1360	59	269	15.4
Servicepoint	1646	76	0	3.6
Servicepoint 2	1619	51	2	2.5
Proposed Privacert	1614	44	0	2.1

Figure 83. Manual de-duplication results on Test Database. In the 2128 records, the number of distinct Clients is 1570. A false positive results when two distinct Clients are counted as one. A false negative results when a record belonging to a known Client is missed. The error percentage is (B) + (C)/2128 * 100. Courtesy Abt Associates [34].

UID Method (Database)	Unduplicated Count (A)	False Negatives (B)	False Positives (C)	Error Percentage
Proposed Privacert (Test Database)	1614	44	0	2.1
PrivaMix Demo (Modified Test Database)	1614	44	0	2.1

Figure 84. PrivaMix de-duplication results. The predicted results (top row) match the actual results (bottom row) exactly. A false positive results when two distinct Clients are counted as one. A false negative results when a record belonging to a known Client is missed. The error percentage is (B) + (C)/2128 * 100. Courtesy Abt Associates [34].

10.5 Summary

In a real-time experiment with three shelters, an HMIS and a Planning Office, a "PrivaMix Demonstration System" computed an accurate unduplicated accounting using real-world data from homeless programs in Des Moines, Iowa ("the Iowa Experiment"). Here is a summary of experimental results.

The experiment used laptops with wireless broadband network, with the software loaded and pre-configured for operation. Standardizing the machines allowed the experiments to focus efficiently and narrowly on performance.

Subjects were clients whose data appeared a participating shelters and the HMIS in a previous six-month time period. The actual subjects are not clients of domestic violence ("DV") homeless shelters, but are clients of homeless family shelters (not domestic violence specific). Using non-DV shelters allowed us to compare computed de-identified results with results derived manually using fully identified data. Of course, the generalizability of these experiments assume there is no difference between DV and non-DV data collection.

A key component in de-duplicating UIDs is the Client source information used to construct the UIDs. Fields having omissions or errors can render UIDs useless. While the PrivaMix Demonstration System works with any Client source information, Privacert proposed to use the first three letters of the first name and the date of birth. Experiments compared Privacert's proposed method with using Social Security numbers, and two methods currently in use by Servicepoint. Privacert's method encountered fewer fields having omissions or errors than the other methods, and used fields in which clients provided more consistent values than the fields used by the other methods. In performing an unduplicated accounting, the Privacert method had the lowest number of errors.

After constructing UIDs, shelters, the HMIS, and Planning Office conducted a real-time duplication using the laptops located at their facilities. The PrivaMix Demonstration System performed exactly as if plain text was used even though sensitive Client source information was provably never shared with the Planning Office or the other Shelters. No errors were introduced.