11. Identifiability of Iowa's De-duplicated Results

The goal of this work is to accomplish de-duplication with guarantees of privacy protection. The PrivaMix Demonstration System accurately de-duplicates (Section 10), and provably provides privacy protection of the UID, throughout the UID construction and de-duplication processes (see Section 9). However, data linkage problems may still exist (Section 4.2) because they do not involve the UIDs, but the data elements that are shared. While the UIDs have provable protection, the shared data elements may be vulnerable. This section examines the uniqueness and re-identification risks associated with shared data elements.

Problem Statement.

> *Given de-duplicated results, compute the uniqueness of Clients and describe possible re-identification strategies.*

11.1 Statistical description of Iowa's demographic elements

This section reports distributions of Client demographics found in the Iowa data that was de-duplicated by PrivaMix.

11.1.1. Analysis design

De-duplicated Demographics Database.

This section reports distributions of Client demographics found in the Iowa data that was de-duplicated by PrivaMix. This writing terms this data the "De-duplicated Demographics Database." Figure 85 reviews the variations of databases used in the Iowa experiments. The Modified Test Database was the source of PrivaMix de-duplication. The de-duplicated results provided data having the same rows of information as found in the Modified Test Database (Figure 84). The fields are different because Client source fields did not appear in de-duplicated results. Of the fields that do appear, some of them contain demographic values –specifically, *year of birth*, *gender*, *ZIP*, *race*, and *ethnicity*. These are the only fields in the De-duplicated Demographic Database. The rows are the values appearing for the 1614 distinct Clients. In summary, the De-duplicated Demographic Database had 1614 records and 5 fields, where each record represents the demographics of a distinct de-duplicated Client.

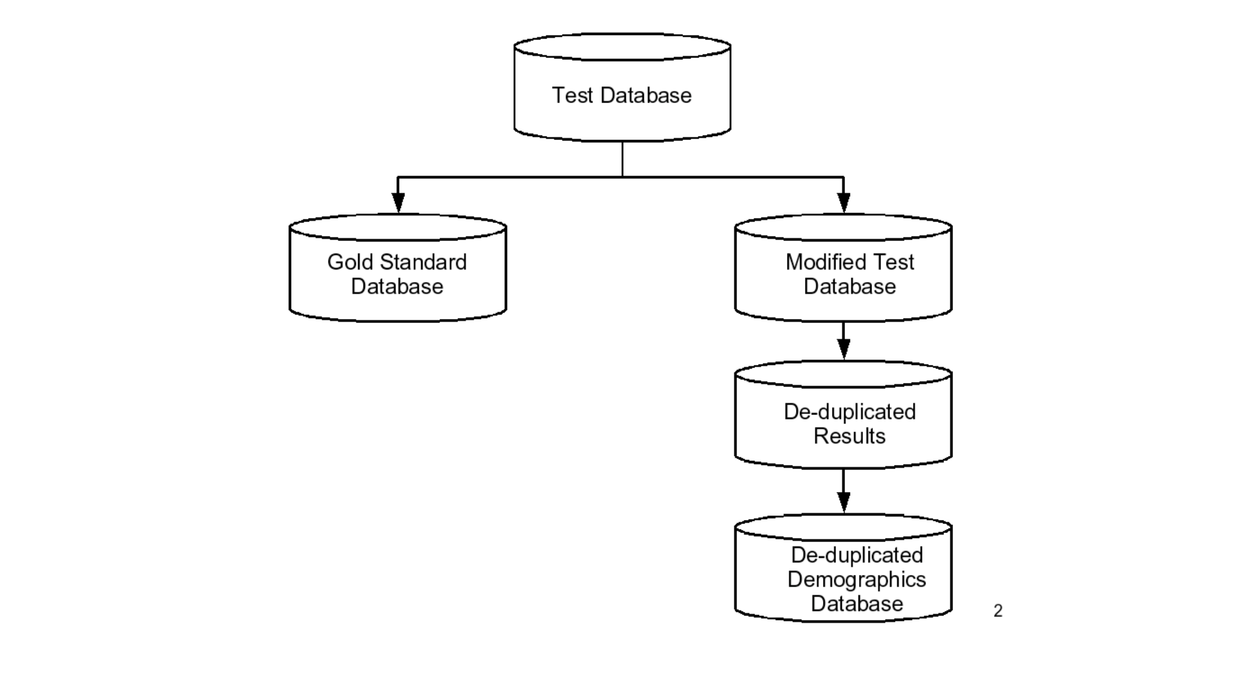


Figure 85. Relationships of databases used in Iowa Experiment. The original Test Database has 2128 records. The Gold Standard Database includes manual corrections to identify 1570 distinct Clients. The Modified Test Database has 66 records added to generate more common visits across participants. After PrivaMix de-duplication, the De-duplicated Results contains a copy of the Modified Test Database with Client source information replaced with sequentially assigned numbers that repeat to identify records belonging to the same Client. The De-duplicated Demographics Database contains only a distinct copy of Client demographic fields in the De-duplicated Results. See Figure 71 for record counts.

11.1.2. Results

Figure 86 displays the ZIP code distribution of the 1614 distinct Clients in the De-duplicated Demographics Database. Not listed are 492 Clients for which no ZIP was found. ZIP 50309 had the most Clients (224). Average number of Clients per ZIP was 7, with a standard deviation of 25. A total of 101 Clients (or 6%) have a unique 5-digit ZIP code.

Figure 87 displays the distribution of the years of births of the 1614 distinct Clients in the De-duplicated Demographics Database. No year of birth was reported for 12 Clients. The most popular year was 1957 (48 Clients). Average number of Clients per year was 21, with a standard deviation of 13. Years of birth from 1901 through 1937 (6 values) are unique.

Figure 88 displays the distribution of gender, race, ethnicity, and race and ethnicity combined for the 1614 distinct Clients in the De-duplicated Demographics Database.

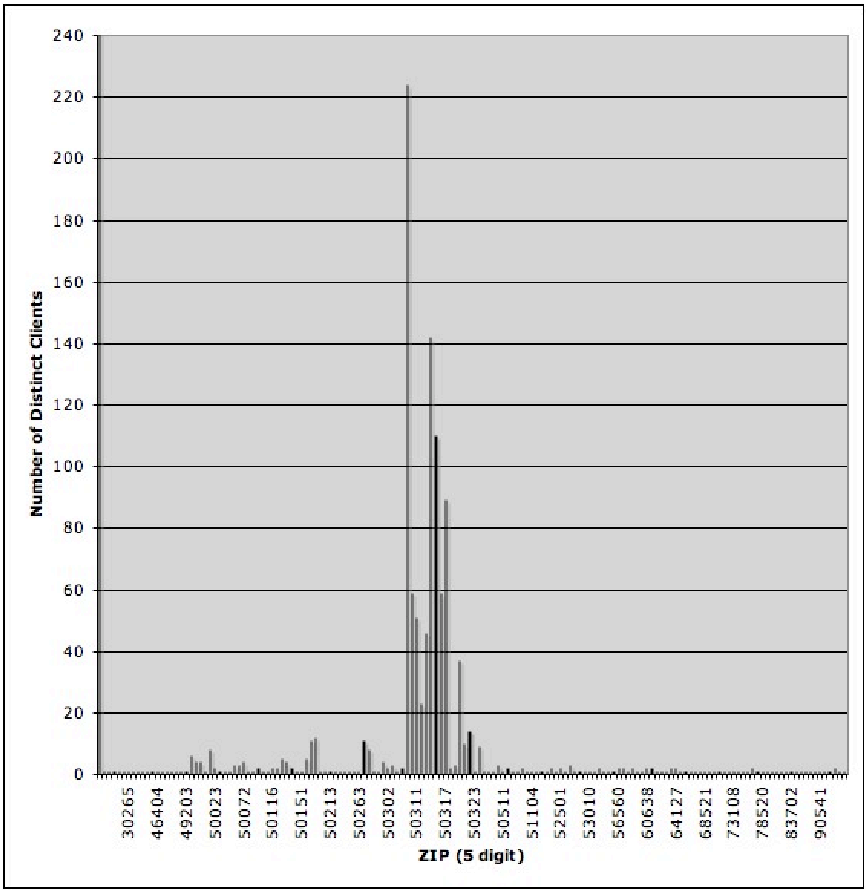


Figure 86. Distribution of 5-digit ZIP codes in Iowa de-duplicated results. Counts based on 1614 de-duplicated clients in the De-duplicated Demographics Database. Not listed are 492 Clients for which no ZIP was found. ZIP 50309 had the most Clients (224). Average number of Clients per ZIP was 7, with a standard deviation of 25.

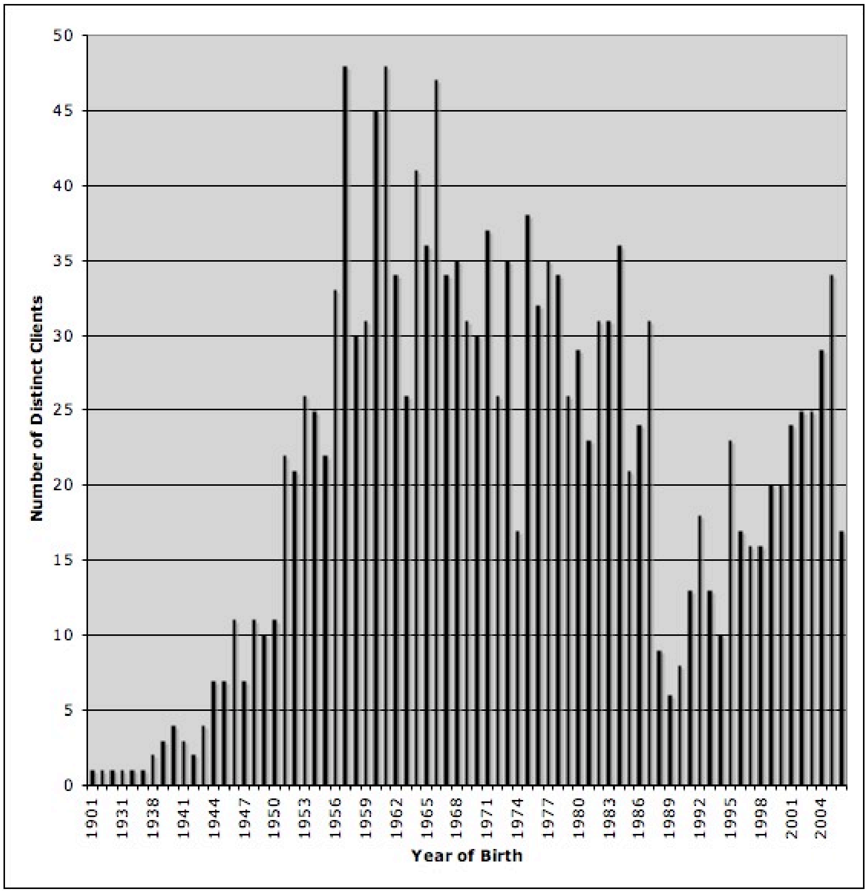
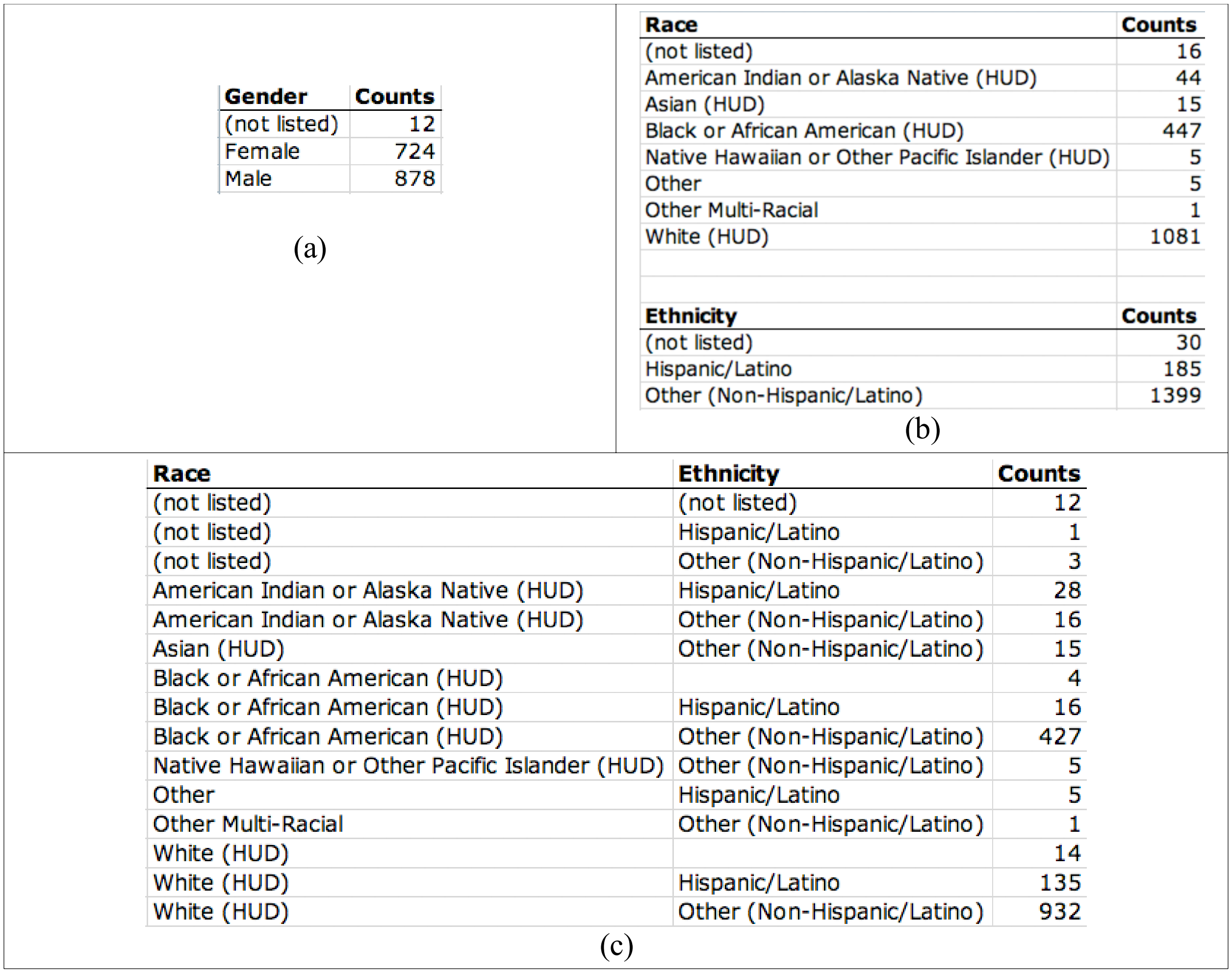


Figure 87. Distribution of Client years of birth in Iowa de-duplicated results. Counts based on 1614 de-duplicated clients in the De-duplicated Demographics Database. No year of birth was reported for 12 Clients. The most popular year was 1957 (48 Clients). Average number of Clients per year was 21, with a standard deviation of 13.



Gender	Counts
(not listed)	12
Female	724
Male	878

(a)

Race	Counts
(not listed)	16
American Indian or Alaska Native (HUD)	44
Asian (HUD)	15
Black or African American (HUD)	447
Native Hawaiian or Other Pacific Islander (HUD)	5
Other	5
Other Multi-Racial	1
White (HUD)	1081

Ethnicity	Counts
(not listed)	30
Hispanic/Latino	185
Other (Non-Hispanic/Latino)	1399

(b)

Race	Ethnicity	Counts
(not listed)	(not listed)	12
(not listed)	Hispanic/Latino	1
(not listed)	Other (Non-Hispanic/Latino)	3
American Indian or Alaska Native (HUD)	Hispanic/Latino	28
American Indian or Alaska Native (HUD)	Other (Non-Hispanic/Latino)	16
Asian (HUD)	Other (Non-Hispanic/Latino)	15
Black or African American (HUD)		4
Black or African American (HUD)	Hispanic/Latino	16
Black or African American (HUD)	Other (Non-Hispanic/Latino)	427
Native Hawaiian or Other Pacific Islander (HUD)	Other (Non-Hispanic/Latino)	5
Other	Hispanic/Latino	5
Other Multi-Racial	Other (Non-Hispanic/Latino)	1
White (HUD)		14
White (HUD)	Hispanic/Latino	135
White (HUD)	Other (Non-Hispanic/Latino)	932

(c)

Figure 88. Distribution of gender, race, and ethnicity in Iowa de-duplicated results. Counts based on 1614 clients in the De-duplicated Demographics Database. Race and Ethnicity values are reported separately in (b) and in combination in (c).

11.2 Uniqueness of demographic combinations in Iowa results

Examining demographics individually, as was done in above in Section 11.1, revealed that some values for year of birth (6) and ZIP (101) appeared only once. These values are unique and therefore the demographics of their 107 Clients are unique. However, demographic fields often combine to jointly yield a greater number of unique combinations. This section reports on the uniqueness of combinations of demographic values occurring in the De-duplicated Demographic Database.

11.2.1. Analysis design

When discussing various ways of counting unique combinations of values, the term "binsize" is useful.

A binsize refers to the number of people to whom a record could ambiguously relate. In the De-duplicated Demographic Database, size people have a distinct year of birth. So, each of these

Clients has a bin size of 1 with respect to year of birth. Similarly, two people were born in 1938 and two people were born in 1942. Each of these Clients has a bin size of 2 with respect to year of birth.

Examining how the number of unique combinations changes as the information becomes less specific is useful in understanding how data elements can have fewer unique combinations. Results in this section will look at the following aggregations of fields:

ZIP	ZIP5	5-digit postal code provided in data
	ZIP4	First 4 digits of postal code (larger geography)
	ZIP3	First 3 digits of postal code (largest geography examined)
Year of Birth	Year of birth	4 digit year of birth provided in data
	Age	Computed as a 2-year range computed using year of birth
	5-year age range	5-year range computed using year of birth
	AHAR age ranges	Age ranges used in AHAR, computed using year of birth. Ranges are: under 1, 1 through 5, 6 through 12, 13 through 17, 18 through 30, 31 through 50,, 51 through 61, and 62 and over.

11.2.2. Uniqueness results

Figure 89 shows the percentage and number of unique combinations of values for various aggregations of ZIP and age. All combinations include gender. Unique combinations of {year of birth, gender, ZIP5} occurred in 36% of Clients (or 580 Clients). As fewer rightmost digits appear in the ZIP, the number of unique combinations decreases. Unique combinations of {year of birth, gender, ZIP4} occurred in 21% of Clients (or 346 Clients). Unique combinations of {year of birth, gender, ZIP3} occurred in 16% of Clients (or 255 Clients).

Similarly, using less specific age information reduces the number of unique combinations. Unique combinations of {age, gender, ZIP5} occurred in 26% of Clients (or 423 Clients). Unique combinations of {age, gender, ZIP5} occurred in 18% of Clients (or 294 Clients). And, unique combinations of {AHAR age ranges, gender, ZIP5} occurred in 18% of Clients (or 294 Clients).

The most aggregated combination of {AHAR age ranges, gender, ZIP3} provided the fewest number of unique combinations, 6% of Clients (or 100 Clients). All combinations of aggregations of ZIP, gender, and age examined revealed unique combinations.

Figure 90, Figure 91, Figure 92, and Figure 93 show the cumulative percentage of population as the binsize increases. The leftmost value in each curve is binsize 1. These are the number of unique occurrences described previously in Figure 89. The rightmost point on each curve occurs when the entire population is included.

Figure 94 examines combinations that include race and ethnicity. The percentage and number of unique combinations of values for various aggregations of ZIP and age are copied from Figure 89 for comparison. In each case, additionally including race and ethnicity increased the number of unique combinations. A general observation is that including race and ethnicity almost doubles the number of unique combinations.

Figure 95, Figure 96, Figure 97, and Figure 98 show the cumulative percentage of population as the binsize increases for combinations that include race and ethnicity. Curves are copied from Figure 90, Figure 91, Figure 92 and Figure 93 for comparison to the curves related to combinations that do not include race and ethnicity.

Figure 99, Figure 100, Figure 101, and Figure 102 show the distributions of binsizes for various combinations of demographic values, with and without race and ethnicity.

ZIP3	16% (255)	12% (193)	9% (139)	6% (100)
ZIP4	21% (346)	17% (267)	12% (201)	8% (136)
ZIP5	36% (580)	26% (423)	18% (294)	12% (196)
Gender	**Year of Birth**	**Age**	**5 year ranges**	**AHAR ranges**

Figure 89. Percentage of unique occurrences in combined ZIP, gender, age aggregations in Iowa de-duplicated results. Counts based on 1614 clients in the De-duplicated Demographics Database. Categories of ZIP are 5-digits (ZIP5), the first 4 digits (ZIP4) and the first 3 digits (ZIP3). AHAR age ranges: Under 1, 1 to 5, 6 to 12, 13 to 17, 18 to 30, 31 to 50, 51 to 61, and 62 and above. Age computed as a 2-year range using year of birth. Number of Clients having unique combinations of noted field combination appear in parentheses.

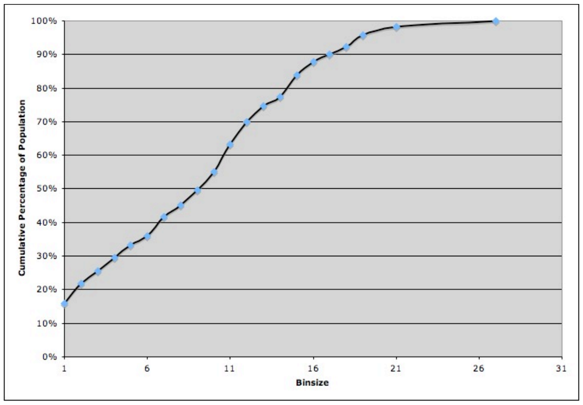
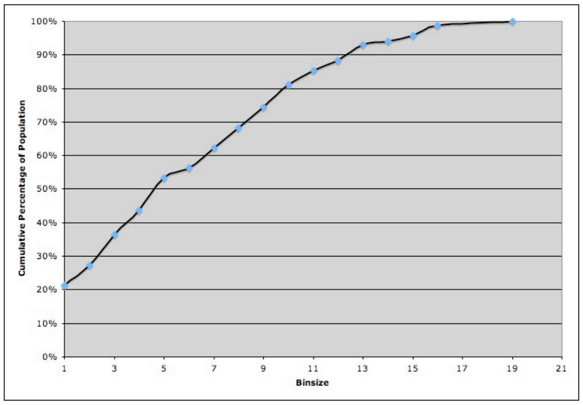
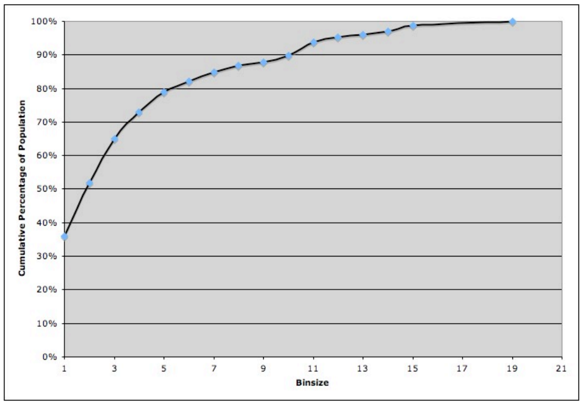
ZIP3	 16% unique (255 Clients)
ZIP4	 21% unique (346 Clients)
ZIP5	 36% unique (580 Clients)
Gender	**Year of Birth**

Figure 90. Binsize distributions for gender, year of birth and ZIP in Iowa de-duplicated results. Population: De-duplicated Demographics Data (1614 total). ZIP: 5-digits (ZIP5), first 4 digits (ZIP4), first 3 digits (ZIP3).

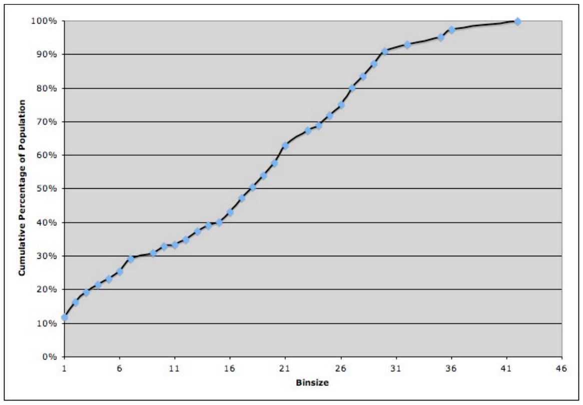
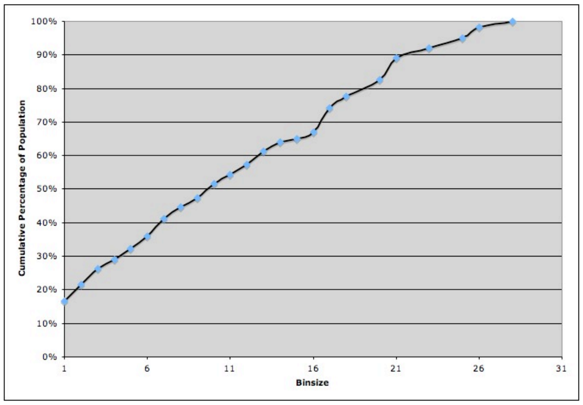
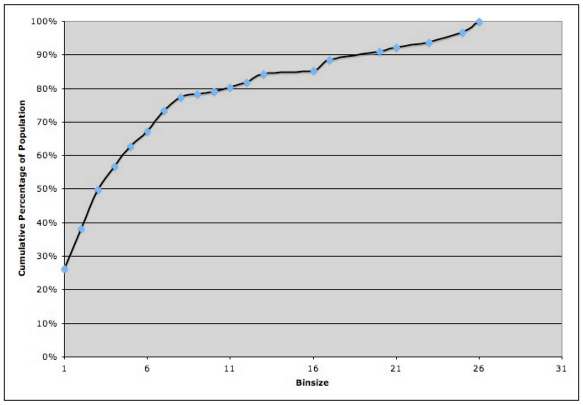
ZIP3	 12% unique (193 Clients)
ZIP4	 17% unique (267 Clients)
ZIP5	 26% unique (423 Clients)
Gender	**Age**

Figure 91. Binsize distributions for gender, age and ZIP aggregations in Iowa de-duplications. De-duplicated Demographics Database (1614 total). 5-digit (ZIP5), first 4 digits (ZIP4), first 3 digits (ZIP3). Age is 2-year range.

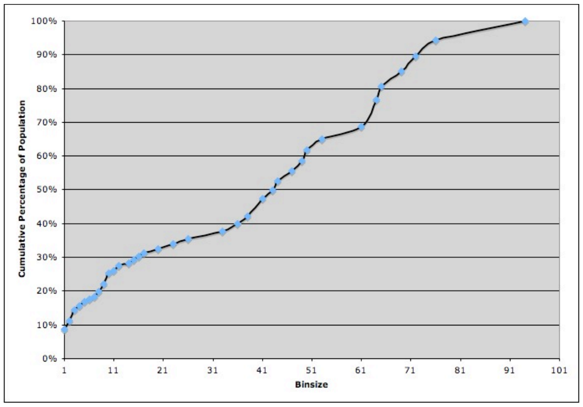
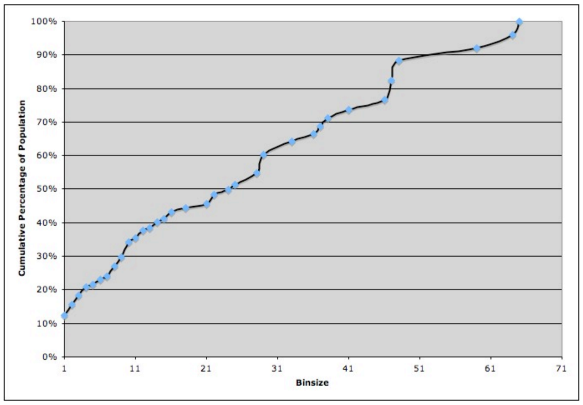
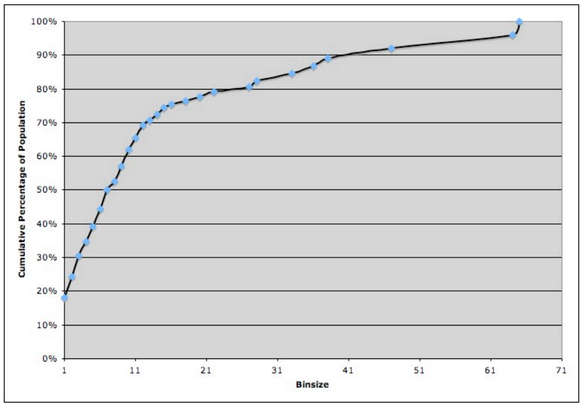
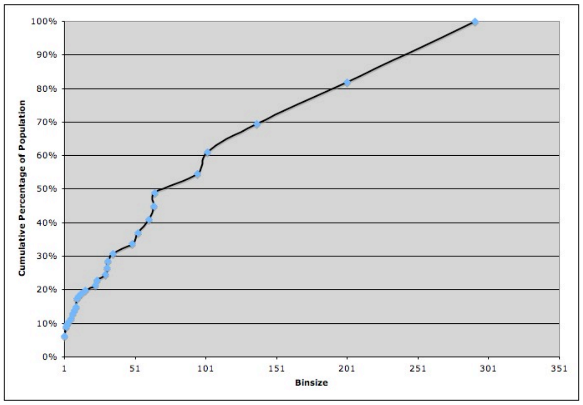
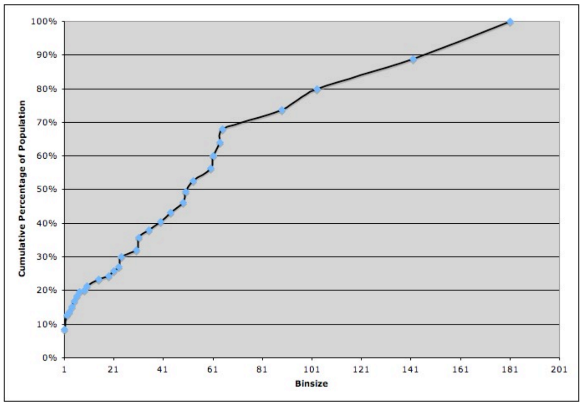
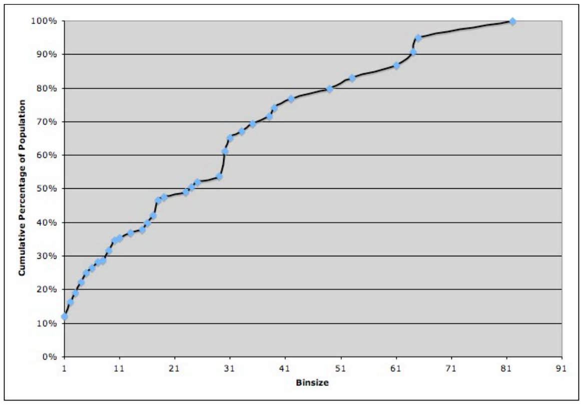
ZIP3	 9% unique (139 Clients)
ZIP4	 12% unique (201 Clients)
ZIP5	 18% unique (294 Clients)
Gender	**5-year Age Ranges**

Figure 92. Binsize distributions for gender, 5-year age ranges and ZIP aggregations in Iowa de-duplications. De-duplicated Demographics Database (1614 total). ZIP: 5-digits (ZIP5), first 4 digits (ZIP4), first 3 digits (ZIP3).

<table>
<tr><td><b>ZIP3</b></td><td>


6% unique (100 Clients)</td></tr>
<tr><td><b>ZIP4</b></td><td>


8% unique (136 Clients)</td></tr>
<tr><td><b>ZIP5</b></td><td>


12% unique (196 Clients)</td></tr>
<tr><td><b>Gender</b></td><td><b>AHAR age ranges</b></td></tr>
</table>

Figure 93. Binsize distributions for combined gender, AHAR age ranges and ZIP aggregations in Iowa de-duplicated results. Counts based on 1614 clients in the De-duplicated Demographics Database. Categories of ZIP are 5-digits (ZIP5), the first 4 digits (ZIP4) and the first 3 digits (ZIP3). AHAR age ranges: Under 1, 1 to 5, 6 to 12, 13 to 17, 18 to 30, 31 to 50, 51 to 61, and 62 and above.

ZIP3	**Gender**	16% (255)	12% (193)	9% (139)	6% (100)
	Gender, Race, Ethnicity	29% (464)	21% (346)	15% (245)	11% (178)
ZIP4	**Gender**	21% (346)	17% (267)	12% (201)	8% (136)
	Gender, Race, Ethnicity	35% (571)	28% (455)	20% (326)	14% (223)
ZIP5	**Gender**	36% (580)	26% (423)	18% (294)	12% (196)
	Gender, Race, Ethnicity	55% (882)	44% (704)	30% (488)	20% (317)
		Year of Birth	**Age**	**5 year ranges**	**AHAR ranges**

Figure 94. Percentage of unique occurrences in combined demographic aggregations in Iowa de-duplicated results. Counts based on 1614 clients in the De-duplicated Demographics Database. Categories of ZIP are 5-digits (ZIP5), the first 4 digits (ZIP4) and the first 3 digits (ZIP3). AHAR age ranges: Under 1, 1 to 5, 6 to 12, 13 to 17, 18 to 30, 31 to 50, 51 to 61, and 62 and above. Age computed as a 2-year range using year of birth. Number of Clients having unique combinations of noted field combination appear in parentheses.

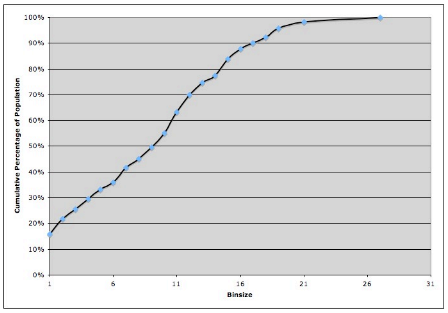
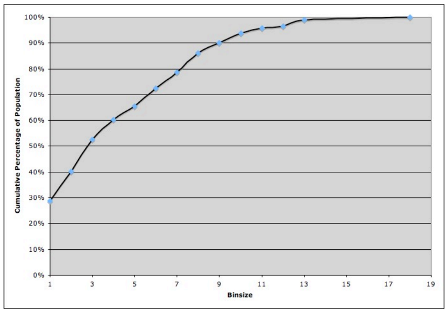
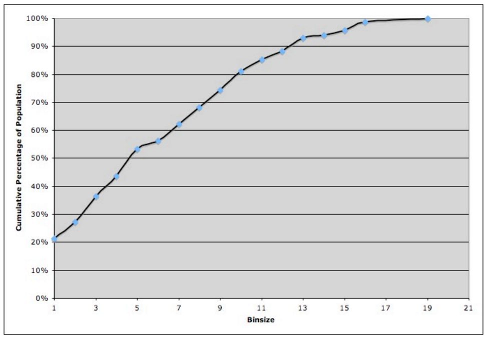
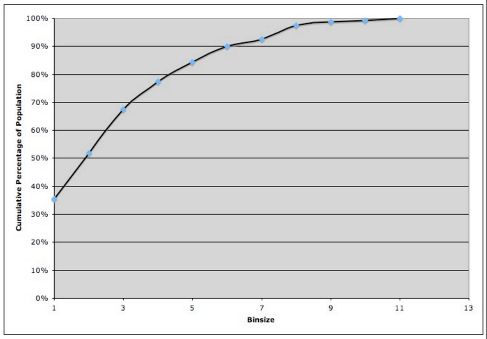
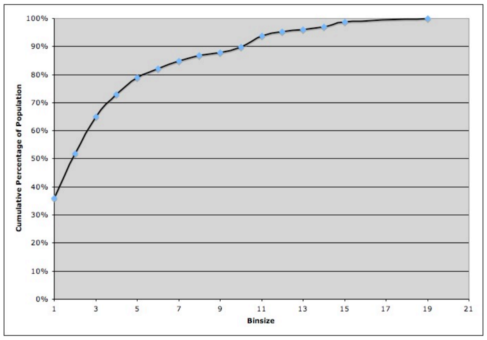
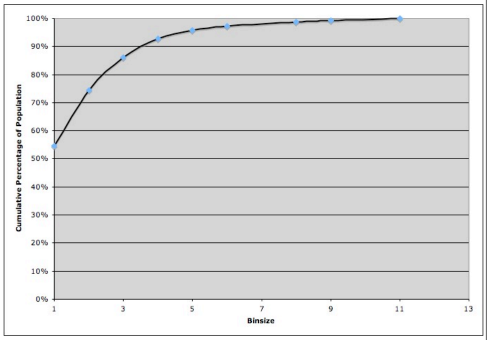
<table>
<tr><td><b>ZIP3</b></td><td>
16% unique (255 Clients)</td><td>
29% (464 Clients)</td></tr>
<tr><td><b>ZIP4</b></td><td>
21% unique (346 Clients)</td><td>
35% (571 Clients)</td></tr>
<tr><td><b>ZIP5</b></td><td>
36% unique (580 Clients)</td><td>
55% unique (882 Clients)</td></tr>
<tr><td></td><td><b>Gender</b></td><td><b>Gender, Race, Ethnicity</b></td></tr>
<tr><td></td><td colspan="2"><b>Year of Birth</b></td></tr>
</table>

Figure 95. Comparison of cumulative binsize distributions for combined demographics, year of birth and ZIP aggregations in Iowa de-duplicated results. Counts based on 1614 clients in the De-duplicated Demographics Database. Categories of ZIP are 5-digits (ZIP5), the first 4 digits (ZIP4) and the first 3 digits (ZIP3).

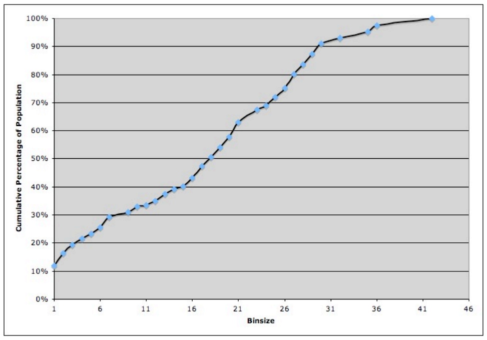
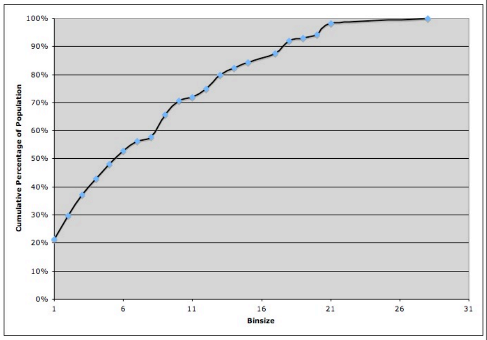
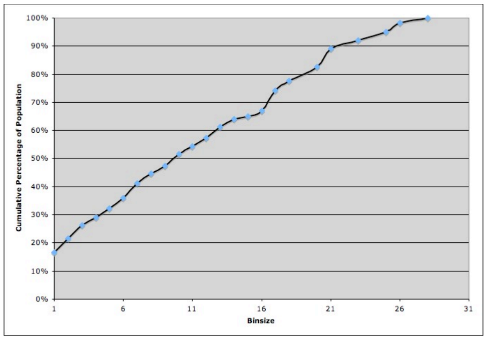
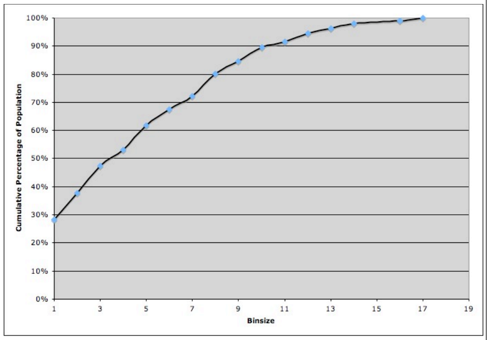
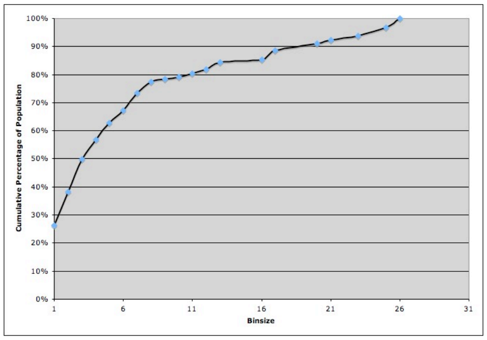
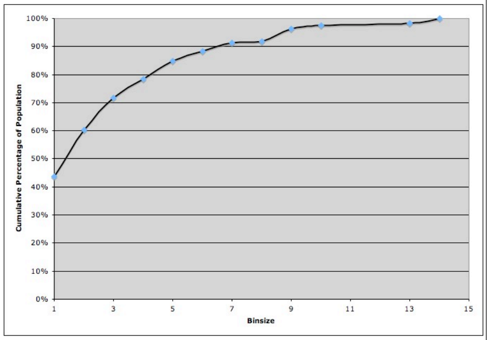
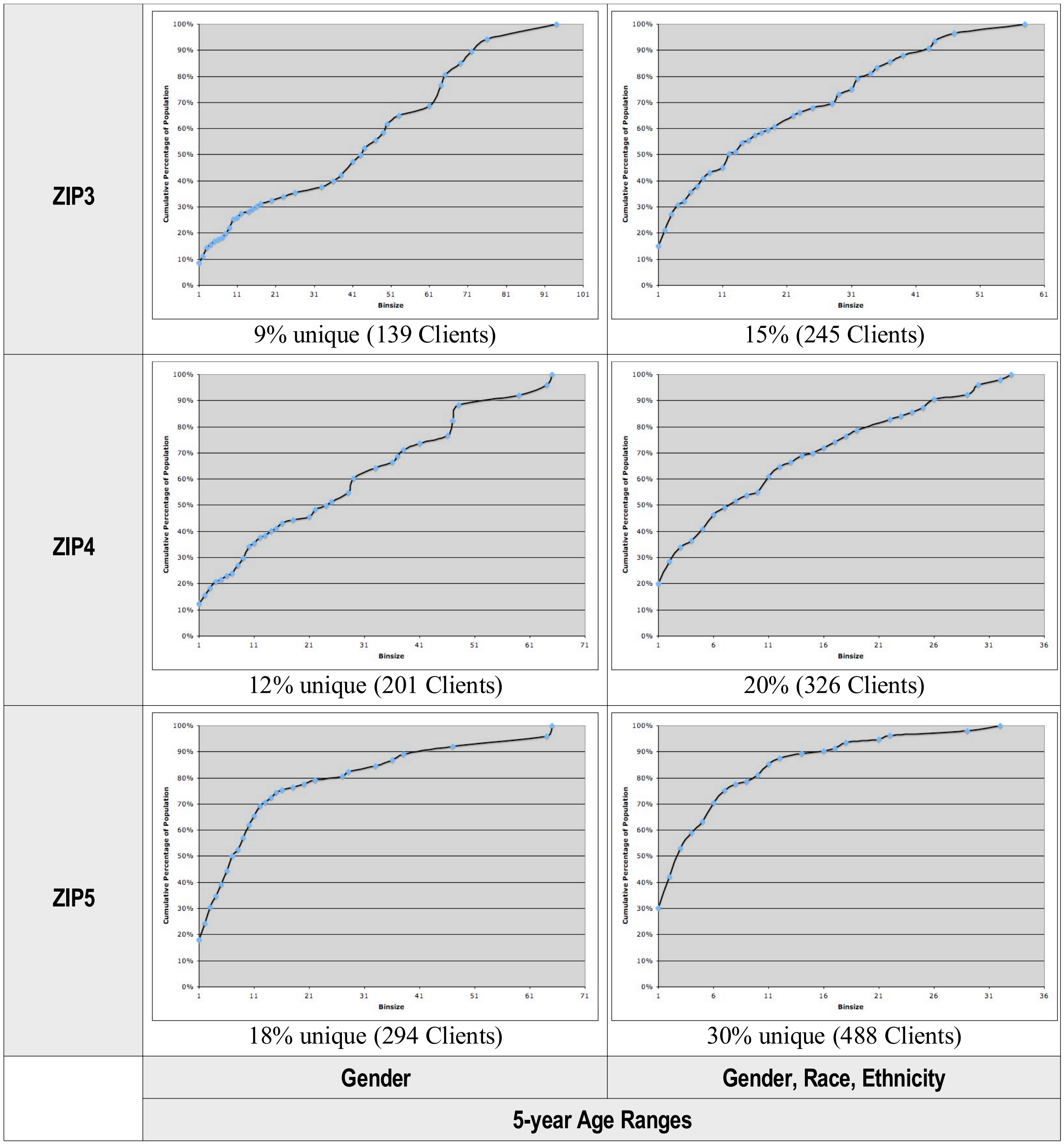
ZIP3	 12% unique (193 Clients)	 21% (346 Clients)
ZIP4	 17% unique (267 Clients)	 28% (455 Clients)
ZIP5	 26% unique (423 Clients)	 44% unique (704 Clients)
	Gender	**Gender, Race, Ethnicity**
	Age	

Figure 96. Comparison of cumulative binsize distributions for combined demographics, age and ZIP aggregations in Iowa de-duplicated results. Counts based on 1614 clients in the De-duplicated Demographics Database. Categories of ZIP are 5-digits (ZIP5), the first 4 digits (ZIP4) and the first 3 digits (ZIP3). Age computed as a 2-year range using year of birth.



9% unique (139 Clients)

15% (245 Clients)

12% unique (201 Clients)

20% (326 Clients)

18% unique (294 Clients)

30% unique (488 Clients)

Figure 97. Comparison of cumulative binsize distributions for combined demographics, 5-year age ranges and ZIP aggregations in Iowa de-duplicated results. Counts based on 1614 clients in the De-duplicated Demographics Database. Categories of ZIP are 5-digits (ZIP5), the first 4 digits (ZIP4) and the first 3 digits (ZIP3).

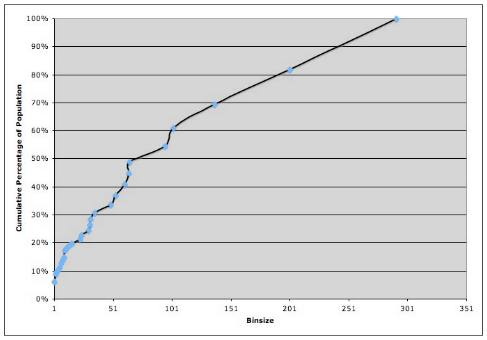
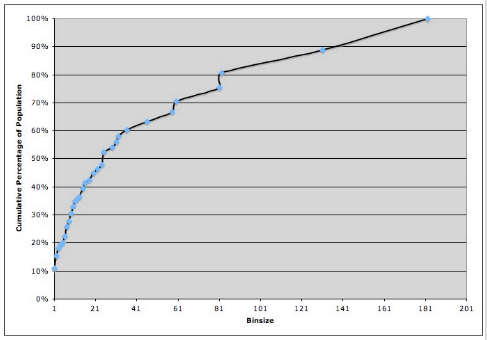
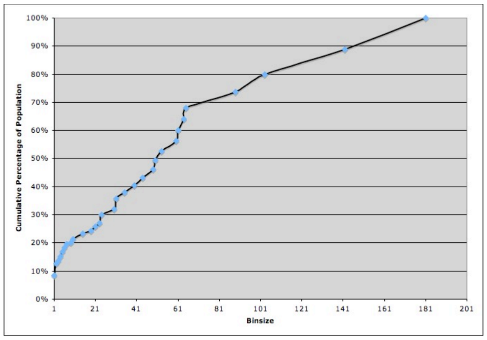
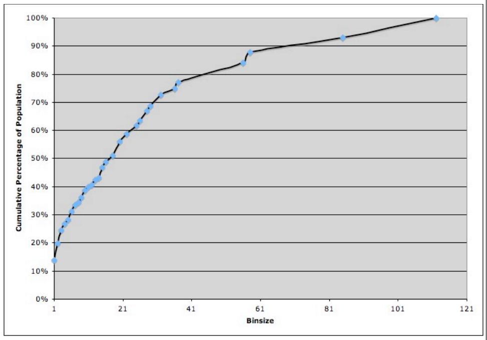
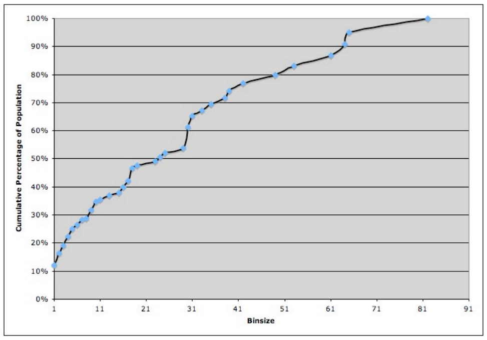
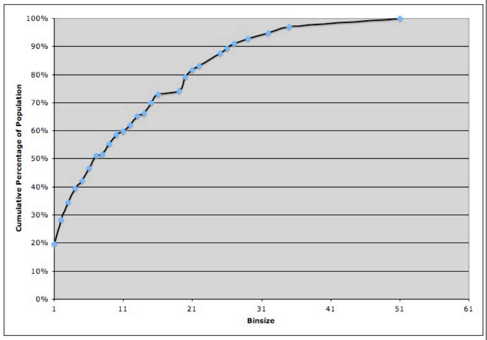
ZIP3	 6% unique (100 Clients)	 11% (178 Clients)
ZIP4	 8% unique (136 Clients)	 14% (223 Clients)
ZIP5	 12% unique (196 Clients)	 20% unique (317 Clients)
	Gender	**Gender, Race, Ethnicity**
	AHAR Age Ranges	

Figure 98. Comparison of cumulative binsize distributions for combined demographics, AHAR age ranges and ZIP aggregations in Iowa de-duplicated results. Counts based on 1614 clients in the De-duplicated Demographics Database. Categories of ZIP are 5-digits (ZIP5), the first 4 digits (ZIP4) and the first 3 digits (ZIP3). AHAR age ranges: Under 1, 1 to 5, 6 to 12, 13 to 17, 18 to 30, 31 to 50, 51 to 61, and 62 and above.

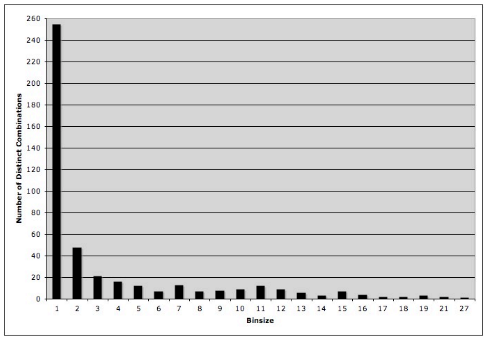
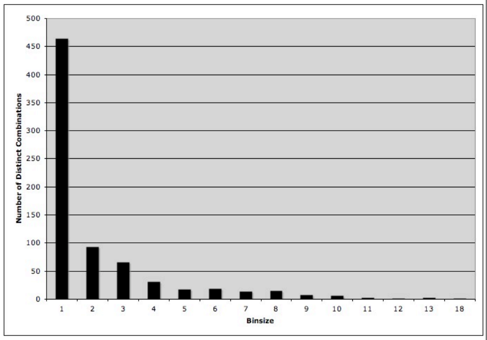
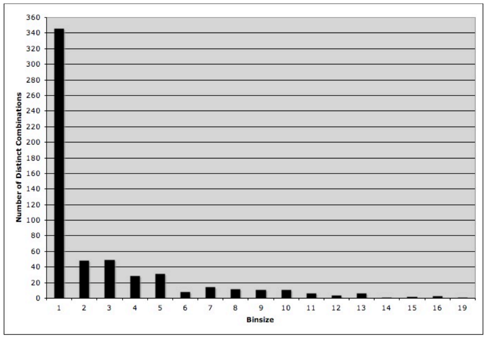
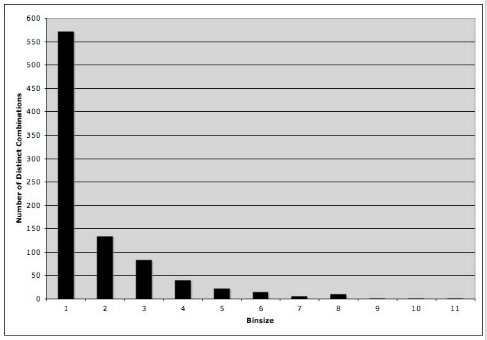
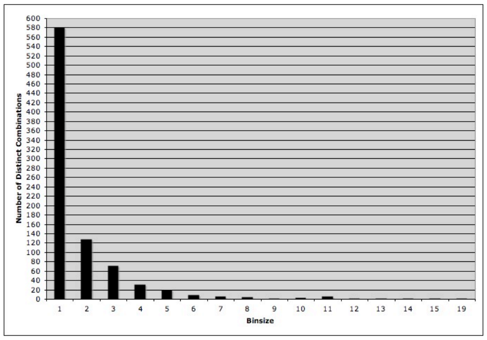
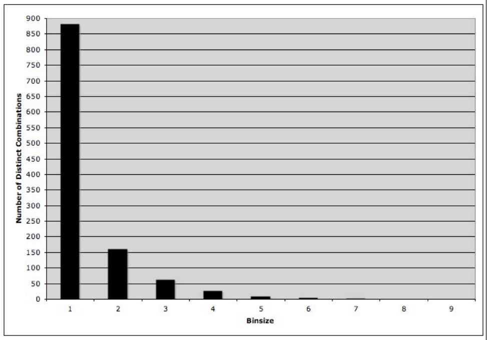
ZIP3	 16% unique (255 Clients)	 29% (464 Clients)
ZIP4	 21% unique (346 Clients)	 35% (571 Clients)
ZIP5	 36% unique (580 Clients)	 55% unique (882 Clients)
	Gender	**Gender, Race, Ethnicity**
	Year of Birth	

Figure 99. Comparison of binsize distributions for combined demographics, year of birth and ZIP aggregations in Iowa de-duplicated results. Counts based on 1614 clients in the De-duplicated Demographics Database. Categories of ZIP are 5-digits (ZIP5), the first 4 digits (ZIP4) and the first 3 digits (ZIP3).

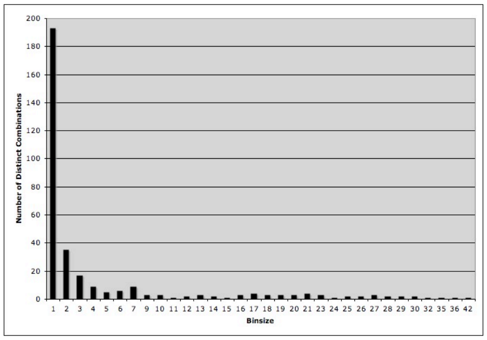
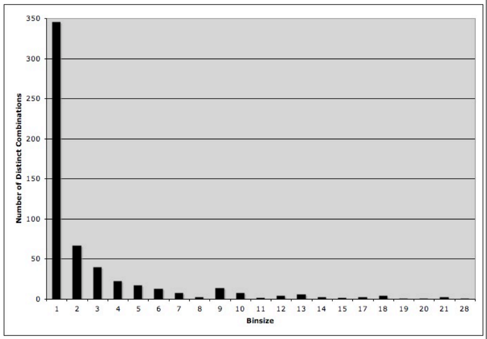
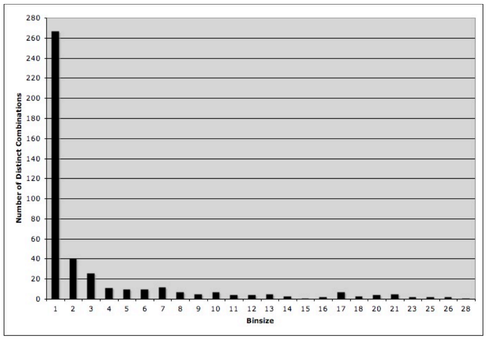
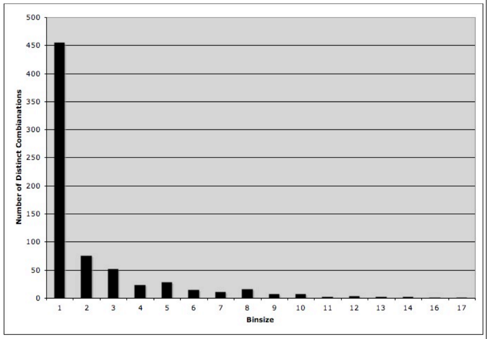
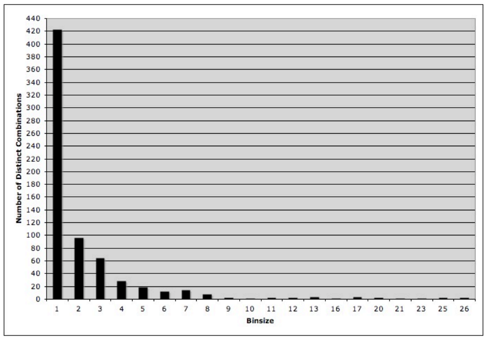
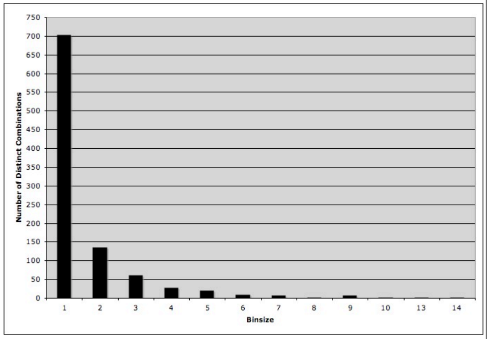
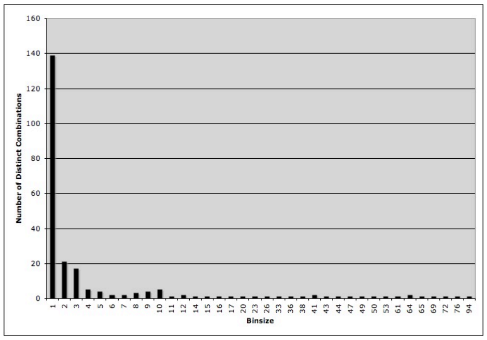
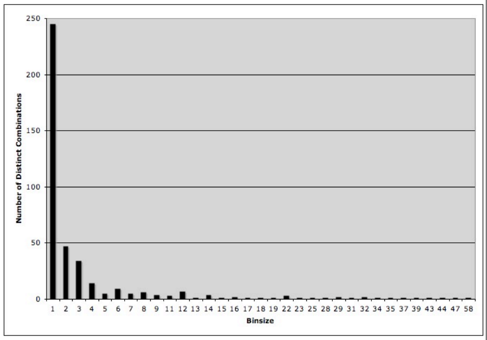
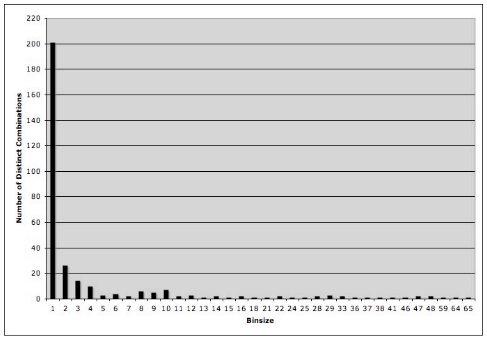
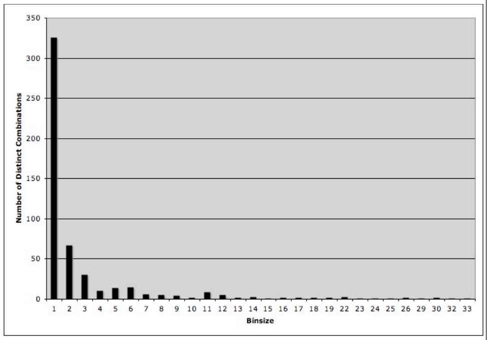
ZIP3	 12% unique (193 Clients)	 21% (346 Clients)
ZIP4	 17% unique (267 Clients)	 28% (455 Clients)
ZIP5	 26% unique (423 Clients)	 44% unique (704 Clients)
	Gender	**Gender, Race, Ethnicity**
	Age	

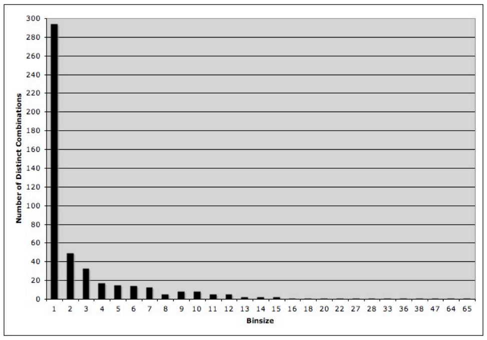
Figure 100. Comparison of binsize distributions for combined demographics, age and ZIP aggregations in Iowa de-duplicated results. Counts based on 1614 clients in the De-duplicated Demographics Database. Categories of ZIP are 5-digits (ZIP5), the first 4 digits (ZIP4) and the first 3 digits (ZIP3). Age computed as a 2-year range using year of birth.

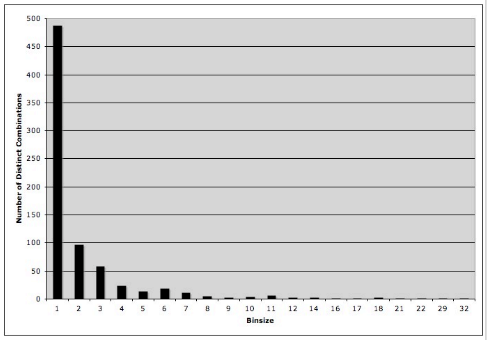
<table>
<tr>
<td><b>ZIP3</b></td>
<td>


9% unique (139 Clients)
</td>
<td>


15% (245 Clients)
</td>
</tr>
<tr>
<td><b>ZIP4</b></td>
<td>


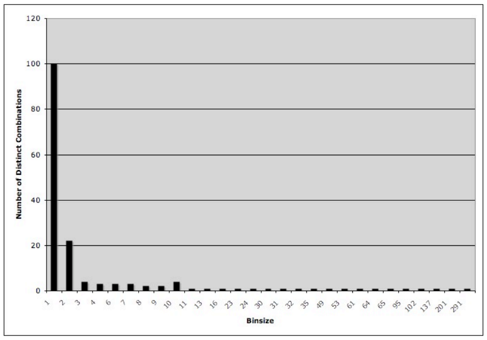
12% unique (201 Clients)
</td>
<td>


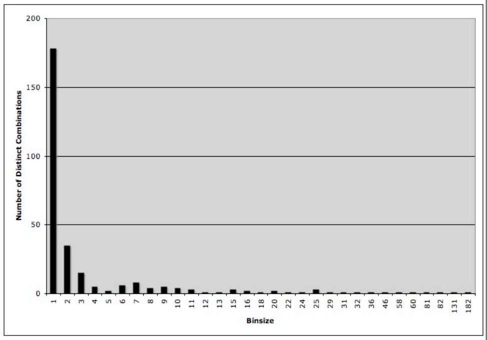
20% (326 Clients)
</td>
</tr>
<tr>
<td><b>ZIP5</b></td>
<td>


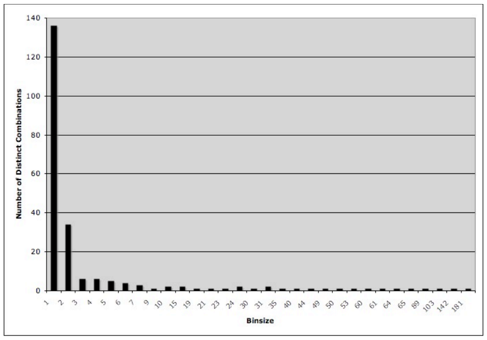
18% unique (294 Clients)
</td>
<td>


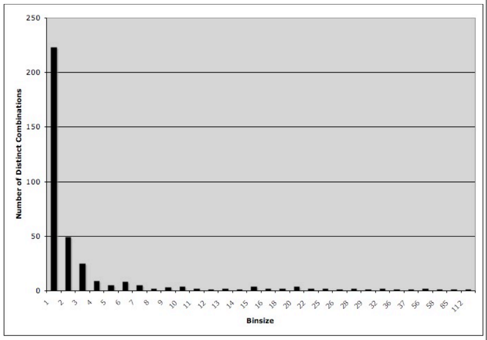
30% unique (488 Clients)
</td>
</tr>
<tr>
<td rowspan="2"></td>
<td><b>Gender</b></td>
<td><b>Gender, Race, Ethnicity</b></td>
</tr>
<tr>
<td colspan="2"><b>5-year Age Ranges</b></td>
</tr>
</table>

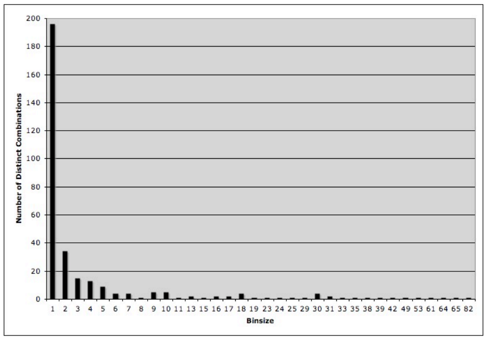
Figure 101. Comparison of binsize distributions for combined demographics, 5-year age ranges and ZIP aggregations in Iowa de-duplicated results. Counts based on 1614 clients in the De-duplicated Demographics Database. Categories of ZIP are 5-digits (ZIP5), the first 4 digits (ZIP4) and the first 3 digits (ZIP3).

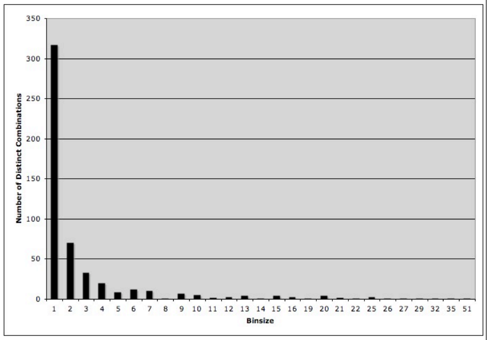
<table>
<tr>
<td>**ZIP3**</td>
<td>

6% unique (100 Clients)</td>
<td>

11% (178 Clients)</td>
</tr>
<tr>
<td>**ZIP4**</td>
<td>

8% unique (136 Clients)</td>
<td>

14% (223 Clients)</td>
</tr>
<tr>
<td>**ZIP5**</td>
<td>

12% unique (196 Clients)</td>
<td>

20% unique (317 Clients)</td>
</tr>
<tr>
<td rowspan="2"></td>
<td>**Gender**</td>
<td>**Gender, Race, Ethnicity**</td>
</tr>
<tr>
<td colspan="2">**AHAR Age Ranges**</td>
</tr>
</table>

Figure 102. Comparison of binsize distributions for combined demographics, AHAR age ranges and ZIP aggregations in Iowa de-duplicated results. Counts based on 1614 clients in the De-duplicated Demographics Database. Categories of ZIP are 5-digits (ZIP5), the first 4 digits (ZIP4) and the first 3 digits (ZIP3). AHAR age ranges: Under 1, 1 to 5, 6 to 12, 13 to 17, 18 to 30, 31 to 50, 51 to 61, and 62 and above.

11.3 Re-identification of Universal Data Elements

As stated earlier, PrivaMix only provides guaranteed privacy protection for UID creation and use in de-duplicating. The privacy of the data elements associated with the UIDs are beyond the scope of the PrivaMix Demonstration System. However, the next section (Section 12) examines some possible ways for PrivaMix to provide privacy protection to Universal Data Elements by expanding its post processing. In the absence of such a remedy, the Universal Data Elements themselves must be altered to thwart data linking (Section 4.2).

This section briefly discusses three re-identification strategies: (1) linking on demographic data; and (2) trail re-identification by HMIS.

11.3.1. Data linkage using demographic fields

Section 4.5 examined the identifiability of the Universal Data Elements. These fields currently include the full month, day and year of birth and the 5-digit ZIP. Figure 13 shows the identifiability of combinations of aggregations of these values. It reports that {date of birth, gender, 5-digit ZIP} uniquely identifies 97% of the U.S. Population. Even changing date of birth to year of birth drastically reduces the identifiability. Figure 13 shows that {year of birth, gender, 5-digit ZIP} uniquely identifies 0.04% of the U.S. Population.

Section 11.3 revealed a somewhat substantial number of unique combinations of values appearing in demographic fields. Figure 89 shows that 36% of Clients in the De-duplicated Demographics Database had unique combinations of {year of birth, gender, 5-digit ZIP}. At first glance, this seems to contradict the identifiability rate mentioned above of 0.04%. That's because having a unique combination of demographic values does not necessarily make the Client identifiable. Successful re-identification requires another dataset on which to link to actually re-identify the Client. If a re-identification attempt only has access to data on the general population, such as a voter list, then the likelihood of re-identification using {year of birth, gender, ZIP} is 0.04%. However, if a re-identification attempt has access to the HMIS, the likelihood of a re-identification using {year of birth, gender, ZIP} is 36%. The HMIS seems to hold sufficient data that the percentage of uniquely occurring combinations of demographic values approximates their likelihood of unique re-identification of Clients.

11.3.2. Data linkage using exact service dates

Another strategy for an HMIS to re-identify Clients in de-duplicated data containing the Universal Data Elements is to exploit the service dates, which can uniquely combine with even the most general demographics, to re-identify Clients.

When a Client receives a service not limited to domestic violence homeless shelters, the Client's explicitly identifying information appears alongside that record in the HMIS. When the HMIS de-duplicates with Shelters, the record of a Client at a Shelter and a record of that same client receiving a non-DV service are related.

If the HMIS has access to the de-duplicated results, the demographic, entry date, and exit date fields may combine sufficient to reliably re-identify Clients by matching service dates. Because the non-DV record has the Client's name, the HMIS learns the client's DV information.

A possible remedy is to provide number of days of service or time ranges (e.g., overnight, a week or less, a month or less, more than a month) and not the actual dates of service.

11.3.3. Trail re-identification using exact service sequence

Another strategy for an HMIS to re-identify Clients in de-duplicated data containing the Universal Data Elements is to exploit the sequence of services received. The service dates provide a longitudinal record of services received by a Client. This longitudinal record poses a linkage threat.

Because the exact entry and exit dates appear in the Universal Data Elements, the Client has a longitudinal record of services over time. The sequence of provided services is likely unique to each Client.

A possible remedy is to provide number of days of service or time ranges (e.g., overnight, a week or less, a month or less, more than a month) and not the actual dates of service.

11.4 Changes to Universal Data Elements

Below are recommendations related to demographics appearing in the Universal Data Elements.

Recommendation #34: The AHAR does not require the demographic specificity currently found in the Universal Data Elements. More general values can be shared without any loss to reporting ability. Therefore, the Universal Data Elements should be revised to reduce the likelihood of recognition by the intimate stalker and/or data linkage threats by using the most general values possible.

Recommendation #35: The date of birth field should minimally be an age range. In fact, a Client may have more than one kind of age range specification. For example, there may be a data element related to 5-year age ranges, and another related to AHAR ranges (under 1, 1 through 5, 6 through 12, 13 through 17, 18 through 30, 31 through 50, 51 through 61, and 62 and over), enabling more reporting uses of the resulting data.

Recommendation #36: The ZIP of last residence field should be changed to either report the first 3 digits of ZIP, or even better, be changed to be a boolean flag denoting whether the Client's last residence was within the geography covered by the Planning Office or not. If the first 3 digits of ZIP are used, then only those values local to the Planning Office need be recorded. Clients from outside the local area would just have a special value, like 999, in order to prevent them appearing as unique outliers.

Recommendation #37: *PIN should be removed. The Shelter should not provide its internal unique number. Instead, the Shelter should maintain an exact copy of the data provided so that records can be referred to in discussion with the Planning Office by the place (or row) in which the record appears.*

Recommendation #38: *Consider removing Race and Ethnicity. Experimental results showed that the addition of these fields increase risks to re-identification.*

Recommendation #39: *Shelters should consider renumbering Household identification numbers from 1 to the last household, prior to forwarding the information to the Planning Office. This makes sure the household identification number itself cannot be the basis for linking.*

Recommendation #40: *Replace the exact service dates (Program Entry Date and Program Exit Date) with number of days of service or with time periods (e.g., overnight, 2-14 days, 15-30 days, 30 plus days).*

Recommendation #41: *More sensitive data elements (such as first name, Social Security number, or full date of birth) may still be collected by Shelters in order to produce a useful UID. However, those values should continue to not be forwarded to the Planning Office as part of the Universal Data Elements.*