

Supplementary Appendices

Appendix A1: Cluster Size Distributions Across Candidate Values of k for First and Last Name

k (cluster size)	Populated clusters	Largest cluster	Median cluster	Clusters < 5	Within-cluster comparisons (billions)
<i>First Name</i>					
250	250	414,935	28,413.50	0	327.19
500	500	170,414	15,381.50	0	136.27
1,000	1,000	65,006	7,568.50	0	69.16
2,000	2,000	38,865	3,536.00	0	39.08
4,000	3,999	38,670	1,603.00	4	26.16
<i>Last Name</i>					
250	250	378,311	27,359.50	0	327.3
500	500	156,251	13,704.50	0	155.14
1,000	1,000	90,911	6,867.00	0	82.84
2,000	2,000	89,553	3,185.50	2	54.5
4,000	3,986	89,543	1,390.50	29	43.59

Note. Clusters were generated by applying MiniBatchKMeans to character 2-gram count vectors, separately for first and last names. Populated clusters may be fewer than k because k-means can return empty clusters. Clusters < 5 counts sparsely populated clusters, which risk assigning records for the same individual to different clusters. Within-cluster comparisons is the number of record pairs requiring evaluation if the cluster identifier were the sole blocking key ($\sum n(n-1)/2$), and indexes computational cost; because cost scales quadratically with cluster size, it is dominated by the largest clusters.



Appendix A2: Blocking Passes Used for Record Linkage Between DOE and Medicaid Data

Standard Pass Blocking Variables

1. name first 1, gender, dob year, dob month, dob day
2. cluster first, cluster last, gender, dob year, dob month
3. name first 2, name last 2, dob year, dob month, dob day
4. cluster last, gender, dob year, dob month, dob day
5. name last 1, gender, dob year, dob month, dob day
6. cluster first, cluster last, dob year, dob month, dob day
7. name first 1, name last 1, gender, dob year, dob day
8. name first 1, name last 1, gender, dob month, dob day
9. name last soundex 3, name first soundex 3, gender, dob year, dob day
10. name last soundex 3, name first soundex 3, gender, dob month, dob day
11. name last soundex 3, name first soundex 3, dob year, dob month, dob day
12. cluster first, gender, dob year, dob month, dob day
13. name first 2, gender, dob year, dob month, dob day
14. name first 1, name last 1, gender, dob year, dob month
15. name last soundex 3, gender, dob year, dob month, dob day
16. name first 2, name last 2, gender, dob year, dob day
17. name first 2, name last 2, gender, dob month, dob day
18. cluster first, cluster last, gender, dob year, dob day
19. name last soundex 3, name first soundex 3, gender, dob year, dob month
20. cluster first, cluster last, gender, dob month, dob day
21. name first soundex 3, gender, dob year, dob month, dob day
22. name first 2, name last 2, gender, dob year, dob month
23. name first 1, name last 1, dob year, dob month, dob day
24. name last 2, gender, dob year, dob month, dob day

Switched First and Last Name

25. first 1, gender, dob year, dob month, dob day
26. cluster first, cluster last, gender, dob year, dob month
27. name first 2, name last 2, dob year, dob month, dob day
28. cluster last, gender, dob year, dob month, dob day
29. name last 1, gender, dob year, dob month, dob day
30. cluster first, cluster last, dob year, dob month, dob day
31. name first 1, name last 1, gender, dob year, dob day
32. name first 1, name last 1, gender, dob month, dob day



Continued

Appendix A2: Continued

-
33. name last soundex 3, name first soundex 3, gender, dob year, dob day
 34. name last soundex 3, name first soundex 3, gender, dob month, dob day
 35. name last soundex 3, name first soundex 3, dob year, dob month, dob day
 36. cluster first, gender, dob year, dob month, dob day
 37. name first 2, gender, dob year, dob month, dob day
 38. name first 1, name last 1, gender, dob year, dob month
 39. name last soundex 3, gender, dob year, dob month, dob day
 40. name first 2, name last 2, gender, dob year, dob day
 41. name first 2, name last 2, gender, dob month, dob day
 42. cluster first, cluster last, gender, dob year, dob day
 43. name last soundex 3, name first soundex 3, gender, dob year, dob month
 44. cluster first, cluster last, gender, dob month, dob day
 45. name first soundex 3, gender, dob year, dob month, dob day
 46. name first 2, name last 2, gender, dob year, dob month
 47. name first 1, name last 1, dob year, dob month, dob day
 48. name last 2, gender, dob year, dob month, dob day
-

Switched Part of Name (e.g., name first 1/name first 2)

-
49. name first 1, gender, dob year, dob month, dob day
 50. name first 2, gender, dob year, dob month, dob day
 51. name last 1, gender, dob year, dob month, dob day
 52. name first 1, name last 1, gender, dob year, dob month
 53. name first 1, name last 1, gender, dob year, dob day
 54. name first 2, name last 2, gender, dob year, dob day
 55. name first 2, name last 2, dob year, dob month, dob day
 56. name first 1, name last 1, gender, dob month, dob day
 57. name first 2, name last 2, gender, dob year, dob month
 58. name first 2, name last 2, gender, dob month, dob day
 59. name first 1, name last 1, dob year, dob month, dob day
 60. name last 2, gender, dob year, dob month, dob day
-

Switched Month and Day

-
61. name first 1, gender, dob year, dob month, dob day
 62. cluster first, cluster last, gender, dob year, dob month
 63. name first 2, name last 2, dob year, dob month, dob day
-



Appendix A2: Continued

-
64. cluster last, gender, dob year, dob month, dob day
 65. name last 1, gender, dob year, dob month, dob day
 66. cluster first, cluster last, dob year, dob month, dob day
 67. name first 1, name last 1, gender, dob year, dob day
 68. name first 1, name last 1, gender, dob month, dob day
 69. name last soundex 3, name first soundex 3, gender, dob year, dob day
 70. name last soundex 3, name first soundex 3, gender, dob month, dob day
 71. name last soundex 3, name first soundex 3, dob year, dob month, dob day
 72. cluster first, gender, dob year, dob month, dob day
 73. name first 2, gender, dob year, dob month, dob day
 74. name first 1, name last 1, gender, dob year, dob month
 75. name last soundex 3, gender, dob year, dob month, dob day
 76. name first 2, name last 2, gender, dob year, dob day
 77. name first 2, name last 2, gender, dob month, dob day
 78. cluster first, cluster last, gender, dob year, dob day
 79. name last soundex 3, name first soundex 3, gender, dob year, dob month
 80. cluster first, cluster last, gender, dob month, dob day
 81. name first soundex 3, gender, dob year, dob month, dob day
 82. name first 2, name last 2, gender, dob year, dob month
 83. name first 1, name last 1, dob year, dob month, dob day
 84. name last 2, gender, dob year, dob month, dob day
-

Appendix A3: Blocking Diagnostics for DOE and Medicaid Record Linkage

Measure	Computation	Interpretation
DOE De-duplication	<u>17,729,937</u>	Reduces candidate pairs by
Reduction Ratio (RR)	$(21,078,606)^2$	~399 million times
Medicaid De-duplication	<u>13,929,606</u>	Reduces candidate pairs by
Reduction Ratio (RR)	$(19,256,218)^2$	~376 million times
Cross-dataset Link	<u>31,898,140</u>	Reduces candidate pairs by
Reduction Ratio (RR)	$(13,929,606 \times 17,729,937)^2$	~12.9 million times
Candidate Recall	<u>2,479,742</u>	Matches ~7.8% of candidate
	31,898,140	pairs; sufficient coverage

Note: Reduction ratios (RR) quantify the extent to which blocking reduces the number of record pairs that must be compared. A smaller RR indicates a more efficient blocking strategy. Candidate recall measures the proportion of true matches retained after blocking; maintaining sufficient recall ensures that computational gains from blocking do not come at the cost of missing valid links.

Appendix A4: Feature importance for matching

