Appendix E1.

Probabilistic linkage of data sets.

Overview of Record Linkage

The purpose of record linkage is to combine multiple data sets into one database for analysis. Record linkage involves the comparison of common data fields across 2 different files; for example, name, sex, date of birth, and social security number. The comparisons of multiple data fields lead to a judgment that 2 records refer to the same (ie, match) or different (ie, nonmatch) persons or events.

Deterministic linkage usually involves subjective linkage of records. The simplest form of deterministic linkage involves exact ("all or nothing") agreement between one or more selected data fields. Another approach is hierarchic, comparing multiple variables in successive "passes" of the data.

In contrast, *probabilistic linkage* combines information from multiple data fields to estimate the probability of a match or nonmatch.[1-4] Probabilistic linkage incorporates information such as the size of the data sets, the number of expected matches, and the reliability and specificity of linkage variables. By using the information contained in each variable, probabilistic linkage also weights agreement differently for each linkage variable; for example, 2 records that match on social security number are more likely to represent the same person than 2 records that match on sex. Similarly, rare values are more likely to match than common values. Ties (multiple records in one file matching to a single record in another) are less likely in probabilistic than deterministic linkage. Probabilistic linkage can account for data subcomponents (eg, month, day, and year of date), tolerances (eg, time±15 minutes), and dependencies (eg, first name "Mary" likely also has sex field "female").

A range of medical research studies have used probabilistic linkage.[5-11] For this study, we performed record linkage using the software Linksolv, version 6 (Strategic Matching Inc.).

Data Sets

This study involved the linkage of 3 data sets: Pennsylvania Emergency Medical Services Patient Care Report Data Set (PAEMS), Pennsylvania Healthcare Cost Containment Council Hospital Discharge Data Set (PHC4), and the Pennsylvania Death Data Set (PA Death) (Figure e1;). After conducting a self-match to remove duplications in PAEMS, we conducted 3 2-way matches: PAEMS-PHC4 (EMS data to hospital discharge data), PAEMS-PA Death (EMS data to death data), and PHC4-PA Death (hospital discharge data to death data).

PAEMS Unduplication

The PAEMS data file consisted of 33,117 patients receiving tracheal intubation (ETI). To identify duplicate patients and events, we used the following variables: date and time of call, county of call, latitude and longitude of the PAEMS station where the call originated, receiving facility, age and sex of the patient, and injury-related event.

Originally, we attempted the linkage by using only the county of call and receiving facility as location identifiers. However, because of very large urban areas in Pennsylvania, the areas of Philadelphia and Pittsburgh received too little weight to appropriately identify duplicates. We therefore added latitude and longitude of the EMS agency. Because there was strong overlap between select matching variables, we reduced the match weights for county, receiving facility, and latitude and longitude by 65%. In addition, we allowed match tolerances of ±5 minutes on dispatch time and ±10 miles on latitude and longitude radius values. We classified pairs with greater than 0.9 match weights as duplicates. We removed 319 (<1%) duplicates.

PAEMS and PHC4 Linkage

For matching the 32,797 unique PAEMS ETI patients to 983,117 PHC4 hospital discharge patients, we used the variables patient age and sex, date of EMS call, date of hospital admission, time of EMS arrival at hospital, time of hospital admission, receiving facility or hospital identifier, the latitude and longitude of the EMS agency and receiving hospital, injury-related admission, and mechanical ventilation during hospitalization. We allowed match tolerances of ±3 years for age, ±15 miles for latitude and longitude, and ±3 hours for EMS dispatch and hospital admission times. Because of the likelihood of greater than 15-mile transports in rural areas, we did not assign full disagreement weights for EMS and hospital latitudes and longitudes.

A customary practice in probabilistic linkage is to retain only record pairs with predicted match weights over an a priori fixed threshold (eg, match probability >0.90).[12] However, this approach often results in low match rates and may inadvertently exclude true matches just below the defined threshold. To avoid this outcome, we used a multiple imputation procedure that creates a series of linked data sets based on the probability distribution of match weights.[13] We created 5 probability samples from the matched pair distribution, generating 5 imputed set with 14,447, 14,431, 14,403, 14,418, and 14,543 respective matched pairs. The average PAEMS-PHC4 linkage rate was 44%.

PAEMS and PA Death Linkage

We next linked the 32,797 unique PAEMS ETI patients to 389,667 PA Death records. We used the variables date, time, county, hospital, patient age and sex, hospital and EMS agency latitude and longitude, incident minor civil division, a flag indicating whether the EMS destination was a hospital, flag indicating whether the death occurred in the hospital, and a flag indicating whether the EMS and death events were injury related. We allowed match tolerances of ±3 years for age and ±15 miles for latitude and longitude. If the death occurred within 30 minutes of dispatch, we considered the times to agree. If the death occurred on the day after PAEMS dispatch, we considered the dates to agree. We created 5 probability samples containing 20,546, 20,487, 20,497, 20,592, and

Table E1. Probabilistic linkage results by imputation.

Characteristic	Imputed Data Set 1	2	3	4	5	Mean
Total matches	25,237	26,139	25,229	25,979	26,082	25,733
PAEMS-PHC4 match only	6,137	6,017	6,146	5,954	6,062	6,063
PAEMS–PA Death match only	14,657	14,366	14,648	14,378	14,340	14,478
PAEMS–PA Death–PHC4 triplet match	4,443	5,756	4,435	5,647	5,680	5,192
No match or duplicate	7,862	6,960	7,870	7,120	7,017	7,366
Total ETI	33,117	33,117	33,117	33,117	33,117	33,117
Match rate, %	76.2	78.9	76.2	78.4	78.8	77.7

PAEMS, Pennsylvania Emergency Medical Services Patient Care Report Data Set; *PHC4*, Pennsylvania Healthcare Cost Containment Council Hospital Discharge Data Set; *PA Death,* Pennsylvania Death Data Set; *ETI*, endotracheal intubation.

20,516 respective matches, for an average PAEMS-PA Death linkage rate of 63%.

PHC4 and PA Death Linkage

We linked the 983,117 PHC4 hospitalizations with 389,667 PA Death records. We used the variables patient age, sex, ethnicity, race, hospital discharge date, death date, hospital county, death county, hospital facility identifier, latitude and longitude of the hospital and death, and injury-related event. Because hospital discharge and death certificate data were likely to match, we reduced the latitude and longitude error tolerance to ±15 miles and required exact matches for other variables. Because of strong overlap between hospital identifier, county identifier, and latitude and longitude, we reduced agreement weights on these fields by 65%. We generated 5 imputed matched sets containing 69,976, 69,932, 69,989, 70,048, and 69,883 matches, respectively, for an average linkage rate of 7%.

Triple Match Procedure

Because of the overlapping data sets, one patient may have appeared as up to 3 successful record linkages: PAEMS-PHC4, PAEMS-PA Death, or PHC4-PA Death. We conducted a probabilistic triple match to identify these potential overlapping matches. This procedure uses identified agreements and disagreements to determine the probability that 3 records refer to the same person and event. Variables used in the triple match included patient age and sex, hospital facility, and dates, times, counties, and latitude and longitude of EMS agency, hospital, and death.

Summary of Linkage Results

For each of the 5 imputed data sets, successful record linkage ranged from 79.1–79.5% (Table E1). Mean record linkage was 77.7%.

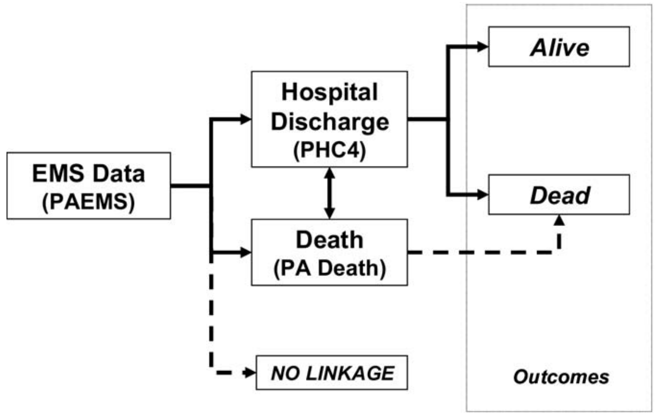

Figure E1. Overview of linkage between data sets. *PAEMS,* Pennsylvania Emergency Medical Services Patient Care Report Data Set; *PHC4,* Pennsylvania Healthcare Cost Containment Council Hospital Discharge Data Set; *PA Death,* Pennsylvania Death Data Set.

Table E2. Multivariable generalized estimating equations (GEE) model of patient outcome (survival) versus rescuer cumulative ETI experience: cardiac arrest ETI only. Rescuer ETI experience reflects cumulative number of procedures performed during 2000 to 2005. Outcomes analysis based on 2003 to 2005 ETI patients and adjusted for patient age, sex, major injury/trauma, bystander-witnessed arrest, bystander CPR, EMS automated external defibrillator use, response time, ECG rhythm, rescuer total patient contacts, EMS agency population setting, and year. ORs reflect estimates from 5 probabilistically linked sets combined using Rubin's method.[14,15]

Variable	**Cardiac Arrest, OR (95% CI)**
Rescuer cumulative ETI experience (2000–2005), No.	
1–10	Referent
11–25	1.02 (0.91–1.15)
26–50	1.13 (0.98–1.31)
>50	1.48 (1.15–1.89)
Patient age, y (ordinal)	
≤6	Referent
7–17	0.92 (0.54–1.58)
≥18	1.42 (1.00–2.01)
Sex	
Male	Referent
Female	0.86 (0.79–0.94)
Major injury/trauma	
No	Referent
Yes	0.94 (0.80–1.12)
Bystander-witnessed cardiac arrest	
No	Referent
Yes	1.25 (1.12–1.40)
Unknown	1.03 (0.89–1.20)
Bystander CPR	
No	Referent
Yes	1.13 (1.01–1.26)
Unknown	1.19 (1.03–1.37)
EMS automated external defibrillator use	
No	Referent
Yes	0.98 (0.85–1.15)
ECG rhythm	
Nonshockable rhythm	Referent
Shockable rhythm	1.33 (1.18–1.51)
Unknown	1.43 (1.30–1.59)
Response time, min	
0–3	Referent
4–6	0.94 (0.84–1.05)
7–10	0.88 (0.78–0.99)
>10	0.64 (0.56–0.74)
Rescuer cumulative total patient contacts (2000–2005), No.	
≤1,000	Referent
1,001–2,000	0.94 (0.84–1.05)
2,002–4,000	1.00 (0.78–0.99)
>4,000	1.01 (0.84–1.21)
EMS agency population setting	
Nonurban	Referent
Urban	1.79 (1.64–1.96)
Air medical	1.47 (0.79–2.71)
Year	
2003	Referent
2004	0.95 (0.86–1.04)
2005	0.92 (0.83–1.02)

Table E3. Multivariable generalized estimating equations (GEE) model of patient outcome (survival) versus rescuer cumulative ETI experience: medical and trauma nonarrest ETI only. Rescuer ETI experience reflects cumulative number of procedures performed during 2000 to 2005. Outcomes analysis based on 2003 to 2005 ETI patients and adjusted for patient age, sex, pulse, systolic blood pressure, Glasgow Coma Scale score, rescuer total patient contacts, EMS agency population setting, and year. ORs reflect estimates from 5 probabilistically linked sets combined using Rubin's method.[14,15]

Variable	Medical Nonarrest, OR (95% CI)	Trauma Nonarrest, OR (95% CI)
Rescuer cumulative ETI experience (2000–2005), No.		
1–10	Referent	Referent
11–25	1.16 (0.97–1.38)	0.92 (0.67–1.26)
26–50	1.29 (1.04–1.59)	1.25 (0.85–1.85)
>50	1.55 (1.08–2.22)	1.84 (0.89–3.81)
Patient age, y (ordinal)		
≤6	Referent	Referent
7–17	4.04 (1.16–14.1)	1.23 (0.47–3.25)
≥18	0.92 (0.41–2.07)	0.51 (0.23–1.13)
Sex		
Male	Referent	Referent
Female	1.02 (0.87–1.19)	1.01 (0.77–1.31)
Pulse, beats/min		
≤40	Referent	Referent
41–80	0.76 (0.57–1.00)	0.62 (0.39–1.00)
>80	1.26 (0.95–1.66)	1.18 (0.74–1.87)
Systolic blood pressure, mm Hg		
≤60	Referent	Referent
61–100	1.37 (1.09–1.72)	1.20 (0.80–1.79)
101–140	2.07 (1.68–2.55)	2.55 (1.72–3.78)
>140	2.17 (1.66–2.83)	2.37 (1.53–3.68)
Glasgow Coma Scale score		
≤8	Referent	Referent
9–12	1.13 (0.90–1.41)	2.76 (1.78–4.26)
13–15	1.68 (1.36–2.08)	2.12 (1.42–3.16)
Rescuer cumulative total patient contacts (2000–2005), No.		
≤1,000	Referent	Referent
1,001–2,000	0.98 (0.79–1.21)	0.83 (0.58–1.18)
2,002–4,000	0.91 (0.74–1.14)	0.63 (0.40–0.98)
>4,000	0.99 (0.74–1.32)	0.58 (0.34–0.99)
EMS agency population setting		
Nonurban	Referent	Referent
Urban	0.87 (0.75–1.01)	0.86 (0.59–1.25)
Air medical	1.34 (0.93–1.96)	1.95 (1.28–2.96)
Year		
2003	Referent	Referent
2004	1.11 (0.94–1.31)	1.10 (0.80–1.51)
2005	1.05 (0.90–1.24)	1.05 (0.77–1.43)

Table E4. Sensitivity analysis. Multivariable generalized estimating equations (GEE) model of patient outcome (survival) versus rescuer cumulative ETI experience: cardiac arrest ETI only. Model reflects use of lowest ETI procedural experience where the data set attributed the ETI to more than 1 rescuer. Rescuer ETI experience reflects cumulative number of procedures performed during 2000 to 2005. Outcomes analysis is based on 2003 to 2005 ETI patients. Cardiac arrest models adjusted for patient age, sex, major injury/trauma, bystander-witnessed arrest, bystander CPR, EMS automated external defibrillator use, response time, ECG rhythm, rescuer total patient contacts, EMS agency population setting, and year. Medical and trauma nonarrest models adjusted for patient age, sex, pulse, systolic blood pressure, Glasgow Coma Scale score, rescuer total patient contacts, EMS agency population setting, and year. ORs reflect estimates from five probabilistically linked sets combined using Rubin's method.[14,15]

Variable	Cardiac Arrest, OR (95% CI)	Medical Nonarrest, OR (95% CI)	Trauma Nonarrest, OR (95% CI)
Rescuer cumulative ETI experience (2000–2005), No.			
1–10	Referent	Referent	Referent
11–25	1.05 (0.94–1.19)	1.16 (0.97–1.39)	0.96 (0.70–1.32)
26–50	1.21 (1.04–1.40)	1.28 (1.03–1.59)	1.24 (0.84–1.83)
>50	1.48 (1.15–1.89)	1.58 (1.10–2.27)	1.82 (0.89–3.73)

Table E5. Sensitivity analysis. Multivariable generalized estimating equations (GEE) model of patient outcome (survival) versus rescuer cumulative ETI experience, stratified by urban, nonurban, and air medical patients. Rescuer ETI experience reflects cumulative number of procedures performed during 2000 to 2005. Outcomes analysis based on 2003 to 2005 ETI patients. Cardiac arrest models were adjusted for patient age, sex, major injury/trauma, bystander-witnessed arrest, bystander CPR, EMS automated external defibrillator use, response time, ECG rhythm, rescuer total patient contacts, and year. Medical and trauma nonarrest models were adjusted for patient age, sex, pulse, systolic blood pressure, Glasgow Coma Scale score, rescuer total patient contacts, EMS agency population setting, and year. ORs reflect estimates from 5 probabilistically linked sets combined using Rubin's method.[14,15]

Variable	Cardiac Arrest, OR (95% CI)	Medical Nonarrest, OR (95% CI)	Trauma Nonarrest, OR (95% CI)
Urban: rescuer cumulative ETI experience (2000–2005), No.			
1–10	Referent	Referent	Referent
11–25	1.03 (0.89–1.19)	1.21 (0.90–1.64)	0.70 (0.33–1.47)
26–50	1.11 (0.93–1.33)	1.16 (0.80–1.68)	0.78 (0.36–1.71)
>50	1.43 (1.06–1.92)	1.28 (0.79–2.09)	0.97 (0.29–3.19)
Nonurban: rescuer cumulative ETI experience (2000–2005), No.			
1–10	Referent	Referent	Referent
11–25	1.03 (0.87–1.22)	1.19 (0.94–1.52)	1.30 (0.71–2.39)
26–50	1.17 (0.92–1.49)	1.50 (1.05–2.14)	2.32 (1.03–5.26)
>50	1.56 (1.02–2.38)	2.05 (1.17–3.60)	5.91 (1.38–25.3)
Air medical: rescuer cumulative ETI experience (2000–2005), No.			
1–10	N/A*	N/A*	Referent
11–25	N/A	N/A	0.84 (0.53–1.33)
26–50	N/A	N/A	0.99 (0.51–1.94)
>50	N/A	N/A	0.87 (0.10–7.21)

*The air medical cardiac arrest and medical nonarrest models did not converge because of the small numbers of patients in these subsets.

Table E6. Rescuer tracheal intubation experience versus systolic blood pressure, nonarrest medical cases.

Cumulative ETI Experience (2000–2005)	Systolic Blood Pressure, mm Hg, No. (%)					
	0–60	61–100	101–140	>140	Unknown	Total
1–10	396 (16.0)	389 (15.7)	736 (29.7)	769 (31.0)	192 (7.74)	2,482
11–25	547 (16.4)	553 (16.6)	907 (27.1)	1,121 (33.5)	214 (6.4)	3,342
26–50	332 (17.0)	298 (15.3)	530 (27.1)	729 (37.3)	64 (3.3)	1,953
>50	63 (16.4)	61 (15.8)	108 (28.1)	146 (37.9)	7 (1.8)	385

Table E7. Rescuer tracheal intubation experience versus Glasgow Coma Scale score, nonarrest medical cases.

Cumulative ETI Experience (2000–2005)	Glasgow Coma Scale Score, No. (%)				
	3–8	**9–12**	**13–15**	**Unknown**	**Total**
1–10	1,512 (60.9)	243 (9.8)	596 (24.0)	131 (5.3)	2,482
11–25	2,122 (63.5)	361 (10.8)	740 (22.1)	119 (3.6)	3,342
26–50	1,150 (58.9)	236 (12.1)	523 (26.8)	44 (2.3)	1,953
>50	184 (47.8)	47 (12.2)	135 (35.1)	19 (4.9)	385

Table E8. Rescuer tracheal intubation experience versus systolic blood pressure, nonarrest trauma cases.

Cumulative ETI Experience (2000–2005)	Systolic Blood Pressure, mm Hg, No. (%)					
	0–60	**61–100**	**101–140**	**>140**	**Unknown**	**Total**
1–10	163 (17.8)	119 (13.0)	325 (35.4)	224 (24.4)	87 (9.5)	918
11–25	260 (18.2)	207 (14.5)	466 (32.6)	356 (24.9)	140 (9.8)	1,429
26–50	162 (21.7)	122 (16.3)	236 (31.6)	182 (24.3)	46 (6.2)	748
>50	36 (33.6)	15 (14.0)	36 (33.6)	18 (16.8)	2 (1.9)	107

Table E9. Rescuer tracheal intubation experience versus Glasgow Coma Scale score, nonarrest trauma cases.

Cumulative ETI Experience (2000–2005)	Glasgow Coma Scale Score, No. (%)				
	3–8	**9–12**	**13–15**	**Unknown**	**Total**
1–10	602 (65.6)	117 (12.8)	162 (17.7)	37 (4.0)	918
11–25	985 (68.9)	174 (12.2)	214 (15.0)	56 (3.9)	1,429
26–50	498 (66.6)	91 (12.2)	125 (16.7)	34 (4.6)	748
>50	67 (62.6)	16 (15.0)	17 (15.9)	7 (6.5	107

REFERENCES

1. Jaro M. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *J Am Stat Assoc*. 1989;84:414-420.
2. Jaro MA. Probabilistic linkage of large public health data files. *Stat Med*. 1995;14:491-498.
3. Fellegi I, Sunter A. A theory for record linkage. *J Am Stat Assoc*. 1969;64:1183-1210.
4. Newcombe H, Kennedy J. Record linkage. *Comun Assoc Computing Machinery*. 1962;5:563-566.
5. Cercarelli LR, Rosman DL, Ryan GA. Comparison of accident and emergency with police road injury data. *J Trauma*. 1996;40:805-809.
6. Overpeck MD, Hoffman HJ, Prager K. The lowest birth-weight infants and the US infant mortality rate: NCHS 1983 linked birth/infant death data. *Am J Public Health*. 1992;82:441-444.
7. Henderson J, Goldacre MJ, Graveney MJ, et al. Use of medical record linkage to study readmission rates. *BMJ*. 1989;299:709-713.
8. Goldacre MJ, Simmons H, Henderson J, et al. Trends in episode based and person based rates of admission to hospital in the Oxford record linkage study area. *Br Med J (Clin Res Ed)*. 1988; 296:583-585.
9. Henderson J, Goldacre MJ, Griffith M. Hospital care for the elderly in the final year of life: a population based study. *BMJ*. 1990; 301:17-19.
10. Newman TB, Brown AN. Use of commercial record linkage software and vital statistics to identify patient deaths. *J Am Med Inform Assoc*. 1997;4:233-237.
11. Cook LJ, Knight S, Olson LM, et al. Motor vehicle crash characteristics and medical outcomes among older drivers in Utah, 1992-1995. *Ann Emerg Med*. 2000;35:585-591.
12. Cook LJ, Olson LM, Dean JM. Probabilistic record linkage: relationships between file sizes, identifiers and match weights. *Methods Inf Med*. 2001;40:196-203.
13. McGlincy MH. A bayesian record linkage methodology for multiple imputation of missing links. In: *ASA Proceedings of the Joint Statistical Meetings*. Alexandria, VA: American Statistical Association; 2004:4001-4008.
14. Rubin DB, Schenker N. Multiple imputation in health-care databases: an overview and some applications. *Stat Med* 1991; 10:585-598.
15. Little RJA, Rubin DB. *Statistical Analysis With Missing Data*. 2nd ed. Hoboken, NJ: Wiley; 2002.