Using Capture-Recapture Methods to Estimate Local Influenza Hospitalization Incidence Rates

Authors:

GK Balasubramani, PhD,[1] Mary Patricia Nowalk, PhD, RDN,[2] Lloyd G. Clarke, BS Hons,[3] Jason A. Lyons,[1] Klancie Dauer,[1] Fernanda Silveira, MD,[4] Donald B. Middleton, MD,[5] Mohamed Yassin, MD,[6] Richard K. Zimmerman, MD, MPH, MA[2]

Affiliations:

[1]University of Pittsburgh Department of Epidemiology, Suite 600 Schenley Place, 4420 Bayard St., Pittsburgh, PA 15260 USA

[2]University of Pittsburgh Department of Family Medicine, Suite 520 Schenley Place, 4420 Bayard St., Pittsburgh, PA 15260 USA

[3]UPMC Health System Department of Pharmacy, Division of Infectious Diseases/ Pharmacy Department – AMP 5TH Floor Falk Medical Building, 3601 Fifth Ave; Pittsburgh PA USA

[4]University of Pittsburgh Department of Medicine, 3601 Fifth Avenue, Suite 5B, Pittsburgh, PA 15213 USA

[5]Department of Medical Education, UPMC St. Margaret, 815 Freeport Rd., Pittsburgh PA 15215

[6]Infection Control Department UPMC Mercy, 1400 Locust Street Suite 10550 B Building Pittsburgh, PA, USA 15219

Corresponding author:

Mary Patricia Nowalk, PhD, RD
Department of Family Medicine
4420 Bayard St., Suite 520
Pittsburgh, PA 15260

tnowalk@pitt.edu
412-383-2355

Word counts:

Abstract – 266; Text – 2103; Tables – 4; Figures – 1; Supplemental Tables – 4

Funding:

This work was supported by the Centers for Disease Control and Prevention (CDC) [5U01IP001035-02] and by National Institutes of Health (NIH) [UL1TR001857]. This work represents the views of the authors and not the CDC or NIH. It is subject to CDC's public access policy.

Conflict of Interest:

Drs. Nowalk and Balasubramani, and Mr. Lyons have grant funding from Merck & Co., Inc. for an unrelated project. Dr. Zimmerman has grant funding from Sanofi Pasteur and Merck & Co., Inc. Dr. Silveira has grant funding from Shire, Ansun, Novartis for unrelated projects. Dr. Middleton reports personal fees from Seqirus, grants and personal fees from Pfizer, personal fees from Sanofi Pasteur. Dr. Yassin and Mr. Clarke have no conflicts to report.

Abstract

Background: Accurate population estimates of disease incidence and burden are needed to set appropriate public health policy. The capture-recapture (C-R) method combines data from multiple sources to better estimate prevalence than is possible using single sources. This study used the C-R method to estimate influenza cases using research and administrative databases to calculate county-wide influenza hospitalization burden.

Methods: Data were derived from a database of clinical virology test results and research data from an influenza vaccine effectiveness study from seasons 2015-2016 to 2018-2019. Missed influenza cases were estimated using C-R method. These estimates were used to calculate disease burden using the multiplier method to correct for underreporting due to curtailing data collection before the end of influenza circulation.

Results: Over all seasons, 422 influenza cases were reported in the administrative database and 382 influenza cases in the research database. Seventy-five cases (18%) reported in the administrative database were not captured in the research database, and 35 (9%) cases in the research database were not captured in the administrative database. Completeness of the influenza hospitalization was estimated to be 76%. Influenza hospitalizations were higher among unvaccinated (32%) than vaccinated (22%) in the current season and among unvaccinated (28%) than vaccinated (23%) in the previous year. The incidence rates for influenza hospitalizations varied by age and season and averaged 421 cases/100,000 population annually.

Conclusion: The capture-recapture method offers a more accurate method for estimating influenza hospitalization than relying on a single data source. Using the multiplier method with adjustments improves the detection of influenza disease burden through a matched database. The incidence rates are consistent with national estimates.

Introduction

Policy makers and planners need accurate estimates of the incidence or prevalence of diseases and health conditions to anticipate, prevent, and mitigate the effects of those diseases. The decentralized nature of U.S. health care makes overall population burden estimates difficult to calculate. Thus, policy makers must rely on population sampling for estimates, leaving true population burden unknown. The accuracy of population estimates depends largely upon the quality of sampling, which in turn, is dependent upon many factors including constancy of the population being sampled across all capture occasions,[1] and having contact with the healthcare system to allow for enumeration. Cases of unreported disease are more difficult to count. Estimates of influenza burden for example, may be inaccurate because influenza exhibits a broad spectrum of severity, ranging from cases that do not warrant medical care, but can cause work or school presenteeism or absenteeism, to hospitalization to death.

Statistical methods to improve population disease incidence and prevalence detection include the capture-recapture (C-R) method. C-R uses the overlap of subjects from two or more data sources to more accurately estimate true population burden and can be used to estimate regional or local disease burden. This study used available data from a single health system, C-R calculations, and adjustments for market share and other factors to estimate influenza hospitalization burden in Allegheny County in Western Pennsylvania.

Methods

Data

Data used for this analysis were collected from two sources: 1) a local health system's clinical surveillance software system (Theradoc®), which extracts virology test results from the electronic medical record (EMR); and 2) research data from selected hospitals participating in the Hospitalized Adult Influenza Vaccine Effectiveness Network (HAIVEN). The study was approved by the University of Pittsburgh IRB.

An IRB-approved honest broker extracted a data list from Theradoc® of a cohort of Allegheny County residents who received an inpatient clinical respiratory viral panel (RVP) test at two to five (depending upon the season) general acute care hospitals in the health system during the study period that included the 2015-2016 through 2018-2019 influenza seasons. Specific dates for each season are shown in Supplemental Table 1. This list also contained basic demographic data of race, sex and age and is henceforth called the "administrative" database.

The "research" database was derived from participants who were recruited from the hospitals during the 2015-2016 through 2018-2019 influenza seasons for the HAIVEN study. Detailed study methods for the HAIVEN study have been described elsewhere.[2] Briefly, patients aged ≥18 years admitted with an acute respiratory infection (ARI) including cough or worsening symptoms of a respiratory illness beginning within 10 days were enrolled. Patients who had been enrolled in the prior 14 days were ineligible. Following informed consent, study staff collected respiratory specimens (nasal and throat swabs from patients) for influenza virus testing (including virus type and subtype) by reverse-transcription polymerase chain reaction (RT-PCR), or used

results from a RVP test, if available. Demographic data were obtained from interview. Vaccination status was based on documented receipt of each year's influenza vaccine from the Pennsylvania Statewide Immunization Information System (PA-SIIS).

Statistical analyses

The administrative database was prepared for analysis by limiting it to data from hospitals in which research enrollments were taking place for each influenza season. For example, in 2015-2016, two hospitals were enrolling participants, whereas in 2018-2019 there were five participating hospitals. Secondly, data were limited to the periods during which research enrollments were taking place. Thirdly, patients <18 years of age were eliminated. Finally, patients were separated into influenza cases and non-cases. Market share data for the UPMC Health System were provided by the Pennsylvania Health Care Cost Containment Council (PHC4).

Summary statistics of the demographic and clinical characteristics were determined for the patients found in both the administrative and research databases (matched database). Number and percent of influenza cases in the databases were calculated and the C-R method was used to estimate influenza incidence.[3]

$$\mathrm{CR-Population\ estimate\ (total\ influenza)} = \frac{\left(Observed\ \ administrative\ cases\ (M)\right) * \ (Observed\ research\ Cases\ (n))}{(Observed\ cases\ from\ matched\ database\ (m))}$$

Using the observed rates provides Petersen's estimate of (N) which is,

$$\hat{\mathrm{N}} = \frac{(M) * \ (n)}{m} \qquad \ldots\ldots (1)$$

This population estimate assumes that the probability of being captured by one source does not affect the probability of being captured by the other source.[12] Calculation of

completeness of reporting by the two sources of C-R method is defined in Supplemental Table S2 and one example in Supplemental Table S3. From the above equation (1) with simple algebraic manipulation, we get the estimate for administrative $(\hat{\mathrm{M}})$, research $(\hat{\mathrm{n}})$, and matched $(\hat{\mathrm{m}})$.

$$Administrative\ (\hat{\mathrm{M}}) = \frac{(N) * (m)}{n}$$

$$Research\ (\hat{\mathrm{n}}) = \frac{(N) * (m)}{M}$$

$$Matched\ database\ (\hat{\mathrm{m}}) = \frac{(M) * (n)}{(N)}$$

The variance and 95% confidence intervals (CIs) were calculated using the formula defined in (1) which is as follows:

$$Variance\ (\hat{\mathrm{N}}) = \frac{M * n * N_1 * N_2}{m^3}$$

$$95\%\ \mathrm{CI} = \hat{\mathrm{N}} \pm 1.96 * \sqrt{\mathrm{Variance}(\hat{\mathrm{N}})}$$

N_1: Number of cases reported only in administrative database, and N_2: Number of cases reported only in research database. These values are not included in the tables.

Population burden estimates were made using the following equations:

$$\mathrm{Burden\ estimate} = \frac{Adjusted\ number\ of\ cases\ from\ CR\ estimate}{Adult\ population\ in\ Allegheny\ County}$$

$$Adjusted\ number\ of\ cases\ from\ CR\ estimate = Number\ of\ cases\ from\ CR\ estimate * \left(\frac{1}{Proportion\ of\ ARI\ enrolled\ based\ on\ multiplier\ to\ ARI\ (\frac{B}{C})}\right) * \left(\frac{1}{ARI\ Market\ share\ of\ study\ hospitals\ (\frac{C}{D})}\right) * \left(\frac{1}{Proportion\ of\ study\ period\ influenza\ detections\ compared\ to\ the\ year\ (E/F)}\right) \quad \ldots\ldots(2)$$

To correct for underdetection of influenza hospitalization burden, we adjusted the rate of viral detections for all, for each age group, and other measures listed in Table 1 by the proportion of subjects tested for influenza during the study period and ran sensitivity analyses for the entire year influenza testing. The overall level of detection was summarized using the multiplier method in Equation (2), that is the expected number of true influenza hospitalizations per reported hospitalization in Allegheny County in those seasons. In sensitivity analyses, we estimated influenza cases and disease burden with 95% confidence intervals for the 2017-2018 and 2018-2019 seasons for age groups, race category, sex, season, vaccination status, and prior vaccination status using the same methods. Data were analyzed using SAS version 9.4 (SAS Institute, Cary, NC, USA).

Results

The final analytic databases are shown in Figure 1. The administrative database consisted of 8,440 patients, the research database consisted of 1,765 patients and 1,825 patients were matched in both databases. Demographic characteristics of the patients found in the matched database are shown in Table 1. The highest proportion of the group was 50-64 years old (34%) with the remainder approximately equally divided among patients who were 18-49, 65- 64 and 75+ years old. The patients in this group were predominantly white (70%), female (63%) and vaccinated ≥14 days prior to illness onset (58%). Half of them had been vaccinated in the previous season and 25% were influenza cases.

Table 2 shows the observed and estimated influenza hospitalizations among persons hospitalized with a cough illness in the administrative, research and matched databases, overall seasons, by season and by other factors. Among hospitalized persons with a cough illness, the observed influenza incidence over all seasons was 25% ranging from 19.6% to 34.1% across the four seasons. Influenza hospitalizations were slightly higher in the 65-74-year age group compared with others and among blacks than whites. Larger differences in influenza hospitalizations were seen between unvaccinated (30.8%) and vaccinated patients (21.8%) and between those vaccinated in the previous year (27.6%) and those not previously vaccinated (22.5%). The estimated influenza hospitalizations did not differ from observed rates by more than a percentage point or two across all groups.

Table 3 shows the proportion of observed influenza cases captured and estimated influenza cases captured (coverage) overall, and for each of the subgroups. Again, the observed and estimated values were remarkably similar. Notably, the lowest

proportion captured in both observed and estimated values were among groups from the research database (18-49-year-olds and 2015-2016 influenza season) with the smallest sample sizes (<400).

The estimated influenza incidence values based on C-R method were then used to determine the burden of influenza hospitalization across all health system hospitals in Allegheny County shown in Table 4. Over all four influenza seasons, the average incidence rate for hospitalized influenza was 421/100,000. The lowest rate was 282/100,000 in 2018-2019 and the highest rate was 630/100,000 in 2017-2018, an especially severe influenza season.

In sensitivity analyses, burden estimates were made for two seasons with sufficient numbers by age group shown in Supplemental Table 4. Influenza hospitalization burden was highest in those ≥75 years in both seasons. Burden was considerably higher in the 2017-2018 A/H3N2 dominated season, than in 2018-2019 that was characterized by nearly equal incidence of A/H1N1 and A/H3N2.

Discussion

This study used capture-recapture methods to calculate influenza incidence among hospitalized patients using two data sources – an administrative database and a research database. C-R has been adapted from ecological studies and used in a wide array of health-related studies including Alzheimer's disease, heart attack, HIV infection, gun injury, pediatric disease surveillance, gastric cancer, and norovirus infections.[4-9] Direct C-R uses two databases and is believed to result in better estimates than indirect C-R, which uses three or more databases. In this study, estimates were nearly identical between observed and estimated incidence. By comparison, in a C-R study of norovirus cases, the combined databases yielded incidence at a rate 2.5 times the level of the rate of the highest individual database.[9] Moreover, in the current study, the overlap between databases was large, resulting in high disease coverage estimates.

This study met three of the four assumptions needed for confidence in the reliability of the outcomes. 1) The populations could be considered closed, that is, there was little chance of loss of cases due to outmigration or death. In a disease of short incubation and duration such as influenza, loss to outmigration is minimal. 2) High overlap of cases improves reliability of estimates because of low missed cases. In this study, there was a large overlap of cases between the databases allowing for matching of pairs of cases (same person, both databases). 3) Databases should be homogeneous. The two databases in this study are believed to be homogeneous, in that cases in both have a nearly equal probability of being identified. 4) The data sources should be independent to prevent over- or underestimation of missing cases. Although they were identified using different methods, the cases in the research database were all hospitalized in the same hospitals from which the administrative

database was drawn. Thus, independence of the samples was decreased. An additional assumption is that large databases should be normally distributed within log-linear models. The differences observed in proportion captured in the two population subgroups with smaller sample sizes may be a reflection of non-normal distributions.

Because of our confidence in the hospitalization estimates, county-wide incidence rates were calculated. Influenza hospitalization incidence was highest in 2017-2018, an influenza A/H3N2-dominated year and lowest in 2015-2016 an influenza A/H1N1 dominated year.[10] This finding is not unexpected, given influenza A/H3N2's higher severity than influenza A/H1N1. Using C-R methods, influenza hospitalization incidence estimates among children have ranged from 240/100,000 in 2003-2004[12] and 860/100,000 in 2004-2005[13] to 89/100,000 for the 2009 A/H1N1 influenza pandemic.[14] Among adults, during the 2009 A/H1N1 pandemic, influenza related hospitalizations were 178/100,000 for adults 18-49 years and 76/100,000 for those ≥50 years of age.[14]

Strengths and Limitations

This study is one of only a few capture-recapture papers for influenza hospitalization rates among adults and the only one of which we are aware for these seasons. It has the advantage of also including vaccination status. The results are consistent with national estimates and validate the population burden estimates that Pittsburgh contributes to CDC's HAIVEN study.[15]

Limitations include conducting the study in only one county, although county-specific population burden is a key outcome. Assumptions on independence of the captures was not made. The sources of the positive or negative dependence may lead to underestimation or over estimation of the population size. However, this difficulty can

be verified through a modelling process by fitting various models to the data to handle dependence among sources and make adjustments by including interaction terms in the model.[11] Log-linear models were not conducted because the intent of the paper is to estimate the population size of influenza cases and influenza disease burden. Finally, funding and time constraints limited the duration of active surveillance during the influenza season for the research database.

Conclusions

Influenza illness is associated with significant costs that include lost productivity due to absenteeism and presenteeism, lost wages, and costs of medical care. Understanding the burden of influenza hospitalization is important for policy makers to allocate resources for the prevention and treatment of influenza. This study validates the influenza population burden estimates that our site contributes to the CDC's HAIVEN study.

Acknowledgement

The Pennsylvania Health Care Cost Containment Council (PHC4) is an independent state agency responsible for addressing the problem of escalating health costs, ensuring the quality of health care, and increasing access to health care for all citizens regardless of ability to pay. PHC4 has provided data to this entity in an effort to further PHC4's mission of educating the public and containing health care costs in Pennsylvania.

PHC4, its agents, and staff, have made no representation, guarantee, or warranty, express or implied, that the data—financial, patient, payor, and physician

specific information—provided to this entity, are error-free, or that the use of the data will avoid differences of opinion or interpretation.

This analysis was not prepared by PHC4. This analysis was done by the University of Pittsburgh. PHC4, its agents and staff, bear no responsibility or liability for the results of the analysis, which are solely the opinion of this entity.

References

1. Hwang W-H, Huggins R. An examination of the effect of heterogeneity on the estimation of population size using capture-recapture data. . *Biometrika.* 2005;92(1):229–233.
2. Ferdinands JM, Gaglani M, Martin ET, et al. Prevention of influenza hospitalization among adults in the United States, 2015–2016: results from the US Hospitalized Adult Influenza Vaccine Effectiveness Network (HAIVEN). *The Journal of infectious diseases.* 2018.
3. Ackman DM, Birkhead G, Flynn M. Assessment of surveillance for meningococcal disease in New York State, 1991. *American journal of epidemiology.* 1996;144(1):78-82.
4. Waller M, Mishra GD, Dobson AJ. Estimating the prevalence of dementia using multiple linked administrative health records and capture–recapture methodology. *Emerging themes in epidemiology.* 2017;14(1):3.
5. Razzak JA, Mawani M, Azam I, Robinson C, Talib U, Kadir MM. Burden of out-of-hospital cardiac arrest in Karachi, Pakistan: Estimation through the capture-recapture method. *J Pak Med Assoc.* 2018;68(7):990-993.
6. Poorolajal J, Mohammadi Y, Farzinara F. Using the capture-recapture method to estimate the human immunodeficiency virus-positive population. *Epidemiology and Health.* 2017;39.
7. Post LA, Balsen Z, Spano R, Vaca FE. Bolstering gun injury surveillance accuracy using capture–recapture methods. *Journal of behavioral medicine.* 2019;42(4):674-680.

8. Knowles RL, Smith A, Lynn R, Rahi JS. Using multiple sources to improve and measure case ascertainment in surveillance studies: 20 years of the British Paediatric Surveillance Unit. *Journal of public health.* 2006;28(2):157-165.

9. Hardstaff J, Clough H, Harris J, Lowther J, Lees D, O'Brien S. The use of capture-recapture methods to provide better estimates of the burden of norovirus outbreaks from seafood in England, 2004–2011. *Epidemiology & Infection.* 2019;147.

10. Centers for Disease Control and Prevention. National and Regional Level Outpatient Illness and Viral Surveillance. http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html. Accessed February 17, 2016.

11. Chao A, Tsay P, Lin SH, Shau WY, Chao DY. The applications of capture-recapture models to epidemiological data. *Statistics in medicine.* 2001;20(20):3123-3157.

12. Grijalva CG, Weinberg GA, Bennett NM, Staat MA, Craig AS, Dupont WD, Iwane MK, Postema AS, Schaffner W, Edwards KM, Griffin MR. Estimating the undetected burden of influenza hospitalizations in children. Epidemiol Infect. 2007 Aug;135(6):951-8. doi: 10.1017/S095026880600762X. Epub 2006 Dec 7. PMID: 17156502; PMCID: PMC2870647.

13. Grijalva CG, Craig AS, Dupont WD, Bridges CB, Schrag SJ, Iwane MK, Schaffner W, Edwards KM, Griffin MR. Estimating influenza hospitalizations among children. Emerg Infect Dis. 2006 Jan;12(1):103-9. doi: 10.3201/eid1201.050308. PMID: 16494725; PMCID: PMC3372368.

14. Jules A, Grijalva CG, Zhu Y, Talbot KH, Williams JV, Dupont WD, Edwards KM, Schaffner W, Shay DK, Griffin MR. Estimating age-specific influenza-related hospitalization rates during the pandemic (H1N1) 2009 in Davidson Co, TN. Influenza Other Respir Viruses. 2012 May;6(3):e63-71. doi: 10.1111/j.1750-2659.2012.00343.x. Epub 2012 Feb 23. PMID: 22360812; PMCID: PMC3773818.

15. Ferdinands JM, Gaglani M, Martin ET, Middleton D, Monto AS, Murthy K, Silveira FP, Talbot HK, Zimmerman R, Alyanak E, Strickland C, Spencer S, Fry AM; HAIVEN Study Investigators. Prevention of Influenza Hospitalization Among Adults in the United States, 2015-2016: Results From the US Hospitalized Adult Influenza Vaccine Effectiveness Network (HAIVEN). J Infect Dis. 2019 Sep 13;220(8):1265-1275. doi: 10.1093/infdis/jiy723. PMID: 30561689; PMCID: PMC6743848.

Table 1. Demographic characteristics of patients identified in the overlapping database (N=1,825)

Variable	Total N (%)
Age group	
18-49 years	357 (19.6)
50-64 years	615 (33.7)
65-74 years	418 (22.9)
75+ years	435 (23.8)
Race	
White	1,274 (69.8)
Black	502 (27.5)
Other, unknown	49 (2.7)
Sex	
Female	1,141 (62.5)
Male	684 (37.5)
Season	
2015-2016	342 (18.7)
2016-2017	422 (23.1)
2017-2018	498 (27.3)
2018-2019	563 (30.9)
Vaccination Status	

Unvaccinated	623 (34.1)
Vaccinated ≥14 days prior to illness onset	1,062 (58.2)
Vaccinated < 14 days prior to illness onset	140 (7.7)
Prior year vaccination (total for all seasons)	
No	905 (49.6)
Yes	920 (50.4)

Table 2. Estimated population influenza hospitalizations using the capture-recapture method

	Observed influenza cases			Observed influenza cases	Estimated influenza cases			Estimated influenza cases (95% CI)	Estimated percent influenza cases
	Admini-strative (M)	Research (n)	Matched (m)	N (%)	$\widehat{M}$	$\widehat{n}$	$\widehat{m}$	$\widehat{N\ (95\%\ CI)}$	$\widehat{(\%)}$
Overall	422	382	347	457 (25.0)	415	376	353	465 (458, 471)	(25.5)
Age group									
18-49 years	82	66	60	88 (24.7)	80	64	61	90 (87, 94)	(25.2)
50-64 years	134	121	110	145 (23.6)	132	119	112	147 (144, 151)	(23.9)
65-74 years	106	95	87	114 (27.3)	104	94	88	116 (113, 119)	(27.8)
75+	100	100	90	110 (25.3)	99	99	91	111 (109, 113)	(25.5)
Race									
White	276	245	221	300 (23.5)	271	240	225	306 (300, 312)	(24.0)
Black	131	122	112	141 (28.1)	129	121	113	143 (140, 146)	(28.5)
Sex									
Female	260	237	213	284 (24.9)	255	233	217	289 (284, 296)	(25.3)
Male	162	145	134	173 (25.3)	160	143	136	175 (172, 179)	(25.6)
Season									
2015-2016	65	18	16	67 (19.6)	60	17	17	73 (63, 84)	(21.3)

2016-2017	102	102	95	109 (25.8)	102	102	95	110 (108, 111)	(26.1)
2017-2018	157	157	144	170 (34.1)	156	156	145	171 (169, 174)	(34.3)
2018-2019	98	105	92	111 (19.7)	97	104	93	112 (110, 114)	(19.9)
Vaccination Status									
Unvaccinated	177	152	137	192 (30.8)	173	149	140	196 (192, 201)	(31.5)
Vaccinated	213	204	186	231 (21.8)	211	202	188	234 (230, 237)	(22.0)
Prior vaccination									
No	234	205	189	250 (27.6)	231	202	192	254 (249, 258)	(28.1)
Yes	188	177	158	207 (22.5)	185	174	161	211 (206, 214)	(22.9)

Table 3. Observed and estimated proportion of cases captured, using the capture-recapture method

	Observed proportion of cases captured (%)			Estimated proportion of cases captured (%)		
	Administrative	**Research**	**Matched**	**Administrative**	**Research**	**Matched**
Over all 4 seasons	92.3	83.6	75.9	89.2	80.9	75.9
2015-2019						
Age group						
18-49 years	93.2	75.0	68.2	88.9	71.1	67.8
50-64 years	92.4	83.4	75.9	89.8	81.0	76.2
65-74 years	93.0	83.3	76.3	89.7	81.0	75.9
75+	90.9	90.9	81.8	89.2	89.2	82.0
Race						
White	92.0	81.7	73.7	88.6	78.4	73.5
Black	92.9	86.5	79.4	90.2	84.6	79.0
Sex						
Female	91.5	83.5	75.0	88.2	80.6	75.1

Male	93.6	83.8	77.5	91.4	81.7	77.7
Season						
2015-2016	97.0	26.9	23.9	82.2	23.3	23.3
2016-2017	93.6	93.6	87.2	92.7	92.7	86.4
2017-2018	92.4	92.4	84.7	91.2	91.2	84.8
2018-2019	88.3	94.6	82.9	86.6	92.9	83.0
Vaccination Status						
Unvaccinated	92.3	99.2	71.4	88.3	76.0	71.4
Vaccinated	92.2	88.3	80.5	90.2	86.3	80.3
Prior vaccination						
No	93.6	82.0	75.6	90.0	79.5	75.6
Yes	90.8	85.5	76.3	87.7	82.5	76.3

Table 4. Influenza hospitalization incidence rates per 100,000 Allegheny County population

Variable	Estimated influenza cases in study hospitals during study period from C-R (A)	ARI* cases in study hospitals during study period in matched databases (B)	ARI* cases in study hospitals during study period in all databases (C)	ARI* cases in all county hospitals† over the entire year (D)	Influenza detections in study hospitals during study period (E)	Total influenza detections in study hospitals over the entire year (F)	Burden/100,000 Allegheny County population (95 % CI)
Overall	465	1,825	12,030	62,004	1,777	1,885	1,682 (1,657 – 1,709)
Annual average (overall/4)							421 (414 – 427)
Age group							
18-49 years	90	357	2,625	7,623	458	483	203 (197 – 213)
50-64 years	147	615	3,567	14,994	478	510	384 (376 – 394)
65-74 years	116	418	2,655	13,705	370	393	406 (395 - 416)
75+ years	111	435	3,178	25,682	469	497	697 (685 – 710)
Race							
White	306	1,274	8,592	52,996	1,217	1,290	1,354 (1,328 – 1,381)
Black	143	502	3,083	8,518	499	533	260 (255 – 266)

Season							
2015-2016	73	342	1,284	15,554	164	179	364 (314 – 419)
2016-2017	110	422	2,229	16,171	377	386	433 (425 – 437)
2017-2018	171	498	3,378	16,353	642	718	630 (623 – 641)
2018-2019	112	563	5,139	13,926	594	602	282 (277 -287)

*ARI=acute respiratory illness

†All health system hospitals in Allegheny County

Supplemental Tables

Table S1. HAIVEN enrollment period dates

Season 2015-2016	12/14/2015 to 4/28/2016
Season 2016 2017	11/14/2016 to 4/27/2017
Season 2017-2018	10/31/2017 to 3/28/2018
Season 2018-2019	11/01/2018 to 4/30/2019

Table S2. Calculation of completeness of reporting by the two independent sources Capture-Recapture method

Research database	Administrative Database		Total
	Cases with clinical test	Cases missed	
Cases enrolled	*m*	N_2	*n*
Cases missed	N_1	*X*	
Total	*M*		$\hat{N}$

m = number of cases identified by both data sources; N_2: Number of cases reported only in research database; n is the number of cases identified in research database source; N_1: Number of cases reported only in administrative database; $X\ is\ the\ number\ cases\ not\ reported\ in\ either\ databases. It\ is\ estimated\ from\ both\ database\ X = (N_1 * N_2)/m$; M is the number cases identified in administrative database source; and $\hat{N}$ is the estimated total number of cases and is defined in equation (1).

Table S3. Example of Capture-Recapture estimate

Research database	Administrative Database		Total
	Cases Clinical Test	Cases Missed	
Cases Enrolled	*347 (m)*	*35 (N_2)*	*382 (n)*
Cases Missed	*75 (N_1)*	*8 (X)*	
Total	*422 (M)*		*465 ($\hat{N}$)*

$\hat{N} = 465\ (95\%\ CI\ 458, 471)$

Table S4. Burden estimates by age group for seasons 2017-2018 and 2018-2019 per 100,000 persons

Variable	Estimated Influenza cases in study hospitals during study period based on CR method (A)	ARI* cases in study hospitals during study period in matched databases (B)	ARI* cases in study hospitals during study period in all databases (C)	ARI* cases in all county hospitals† over the entire year (D)	Influenza detections in study hospitals during study period (E)	Total influenza detections in study hospitals over the entire year (F)	Burden/ 100,000 Allegheny County population (95% CI)
2017-2018 season	171	498	3,378	16,353	642	718	630 (623 – 641)
Age group							
18-49 years	29	95	717	1,955	149	164	66 (64 – 68)
50-64 years	48	157	952	3,838	153	179	138 (135 – 144)
65-74 years	44	124	761	3,644	128	141	143 (140 – 146)
75+ years	49	122	945	6,916	210	232	308 (302 – 321)
2018-2019 season	112	563	5,139	13,926	594	602	282 (277 – 287)
Age group							
18-49 years	20	116	1,086	1,725	172	174	30 (29 – 32)
50-64 years	43	207	1,540	3,360	188	191	71 (70 – 73)
65-74 years	27	129	1,143	3,299	124	126	70 (68 – 73)
75+ years	21	111	1,369	5,542	110	111	106 (101 – 111)

Figure 1

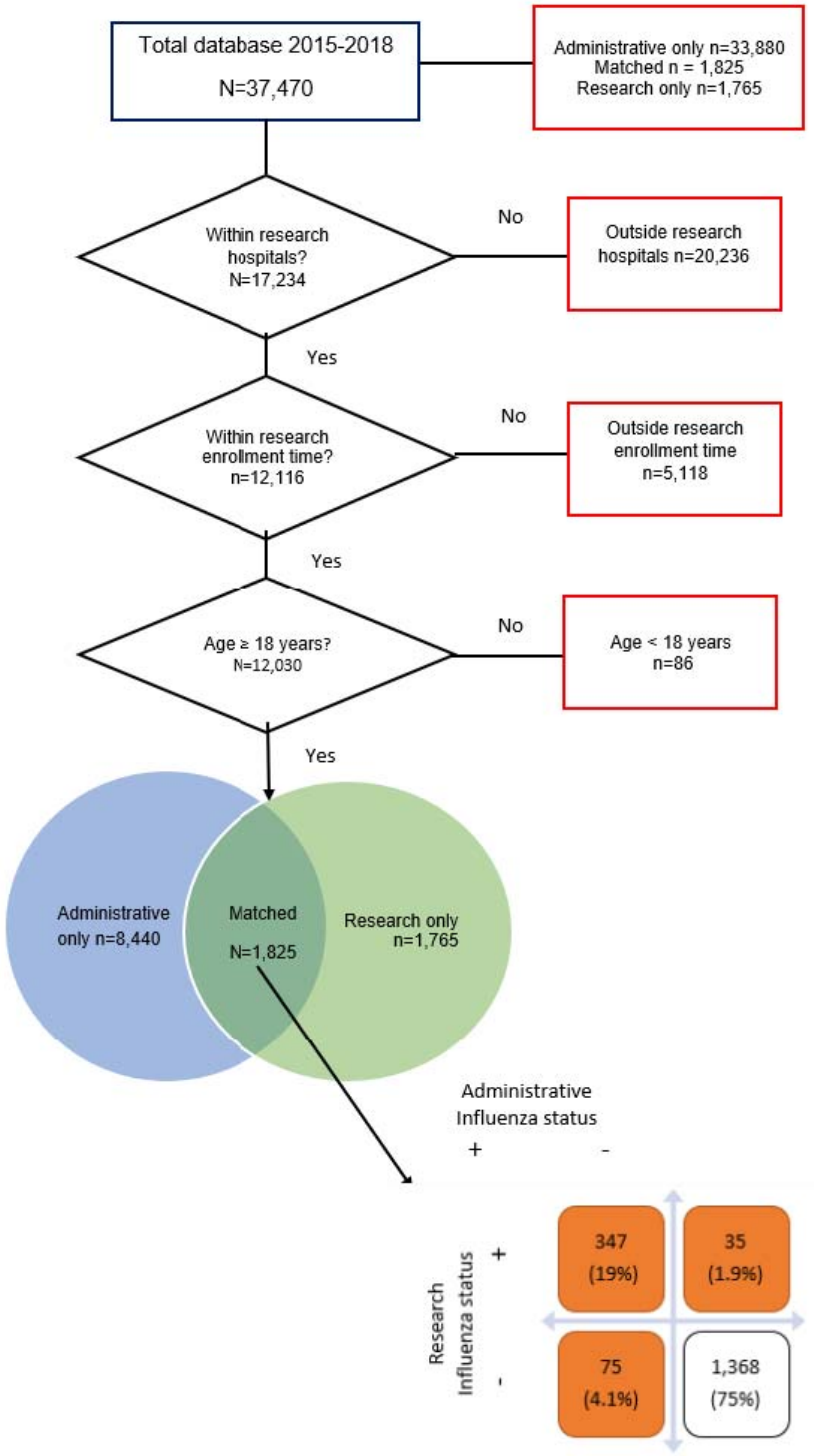

Figure 1 Legend. Flow Chart Capture – Recapture Method of Estimating Population Size