Use of Multiple Imputation to Estimate the Proportion of Respiratory Virus Detections Among Patients Hospitalized With Community-Acquired Pneumonia

Catherine H. Bozio, Emory University
W Dana Flanders, Emory University
Lyn Finelli, Centers for Disease Control and Prevention
Anna M. Bramley, Centers for Disease Control and Prevention
Carrie Reed, Centers for Disease Control and Prevention
Neel R. Gandhi, Emory University
Jorge Vidal Graniel, Emory University
Dean Erdman, Centers for Disease Control and Prevention
Min Z. Levine, Centers for Disease Control and Prevention
Stephen Lindstrom, Centers for Disease Control and Prevention

Only first 10 authors above; see publication for full author list.

Journal Title: Open Forum Infectious Diseases
Volume: Volume 5, Number 4
Publisher: Oxford University Press (OUP) | 2018-04-01, Pages ofy061-ofy061
Type of Work: Article | Final Publisher PDF
Publisher DOI: 10.1093/ofid/ofy061
Permanent URL: https://pid.emory.edu/ark:/25593/t0ksr

Final published version: http://dx.doi.org/10.1093/ofid/ofy061

Copyright information:
Published by Oxford University Press on behalf of Infectious Diseases Society of America 2018. This work is written by (a) US Government employee(s) and is in the public domain in the US.

Accessed September 29, 2018 5:00 AM EDT
Use of Multiple Imputation to Estimate the Proportion of Respiratory Virus Detections Among Patients Hospitalized With Community-Acquired Pneumonia


Departments of 1Epidemiology and 2Graduate Program in Molecules to Mankind and 3Global Health, Emory University, Atlanta, Georgia; 4Centers for Disease Control and Prevention, Atlanta, Georgia; 5Emory University School of Medicine, Atlanta, Georgia; 6University of Utah Health Sciences Center, Salt Lake City, Utah; 7Le Bonheur Children’s Hospital, Memphis, Tennessee; 8University of Tennessee Health Science Center, Memphis, Tennessee; 9Vanderbilt University School of Medicine, Nashville, Tennessee; 10Vanderbilt Vaccine Research Program, Nashville, Tennessee; 11St. Jude Children’s Research Hospital, Memphis, Tennessee; 12Northwestern University Feinberg School of Medicine, Chicago, Illinois

Background. Real-time polymerase chain reaction (PCR) on respiratory specimens and serology on paired blood specimens are used to determine the etiology of respiratory illnesses for research studies. However, convalescent serology is often not collected. We used multiple imputation to assign values for missing serology results to estimate virus-specific prevalence among pediatric and adult community-acquired pneumonia hospitalizations using data from an active population-based surveillance study.

Methods. Presence of adenoviruses, human metapneumovirus, influenza viruses, parainfluenza virus types 1–3, and respiratory syncytial virus was defined by positive PCR on nasopharyngeal/oropharyngeal specimens or a 4-fold rise in paired serology. We performed multiple imputation by developing a multivariable regression model for each virus using data from patients with available serology results. We calculated absolute and relative differences in the proportion of each virus detected comparing the imputed to observed (nonimputed) results.

Results. Among 2222 children and 2259 adults, 98.8% and 99.5% had nasopharyngeal/oropharyngeal specimens and 43.2% and 37.5% had paired serum specimens, respectively. Imputed results increased viral etiology assignments by an absolute difference of 1.6%–4.4% and 0.8%–2.8% in children and adults, respectively; relative differences were 1.1–3.0 times higher.

Conclusions. Multiple imputation can be used when serology results are missing, to refine virus-specific prevalence estimates, and these will likely increase estimates.

Keywords. community-acquired pneumonia; multiple imputation; respiratory viruses.

Community-acquired pneumonia (CAP) is a common cause of hospitalization in the United States [1, 2]. Bacterial and viral pathogens can cause pneumonia and can be detected using a broad range of diagnostic tests. However, there is no true “gold standard” for etiology determination, because currently available diagnostic tests for respiratory pathogens have well-known limitations [3–5]. Real-time polymerase chain reaction (PCR) assays are most commonly performed on respiratory specimens in etiology studies, because they have improved sensitivity and turnaround time for respiratory virus detection relative to culture [6–10]. Although not timely, serology may capture infections even if a virus is not sufficiently present to be detected by PCR in respiratory specimens; thus, acute and convalescent (paired sera) serology results can increase the diagnostic yield for respiratory viruses in CAP studies [3, 4, 11–13]. However, even in the research setting, paired sera are not always available from each patient, particularly because convalescent serology requires specimen collection posthospital discharge or death. Multiple imputation, an analytic method used to assign values for missing data, can be applied to pneumonia etiology studies when test results are missing. We applied multiple imputation to missing serology results to improve estimates of respiratory virus detection in the Centers for Disease Control and Prevention (CDC) Etiology of Pneumonia in the Community (EPIC) study, a multicenter active surveillance study of the incidence and etiology of CAP among hospitalized US adults and children [14, 15].

METHODS

Enrollment, Specimen Collection, and Laboratory Methods
The CDC EPIC study was a prospective, multicenter, population-based, active surveillance study, as previously described in detail [14, 15]. In brief, from January 1, 2010 to June 30, 2012, children <18 years of age were enrolled in 3 pediatric hospitals...
in Memphis, Tennessee, Nashville, TN, and Salt Lake City, Utah, and adults were enrolled in 3 hospitals in Chicago, Illinois, and 2 hospitals in Nashville, TN. Patients admitted to a study hospital were eligible for enrollment if they resided within the hospital catchment area and had evidence of acute respiratory infection and radiographically confirmed pneumonia within 72 hours of admission. Patients were included in the etiologic analysis if they (1) met final radiographic criteria for CAP based on a study radiologist review and (2) had specimens available for both bacterial and viral testing. Clinical, demographic, and epidemiologic information were collected through interviews and medical chart review. Informed consent was obtained, and the study protocol was approved by institutional review boards at each institution and the CDC.

Whole blood, acute serum, and nasopharyngeal/oropharyngeal (NP/OP) swabs were collected as soon as possible after enrollment; urine was only collected in adults [14, 15]. For adults with a productive cough, sputum was also collected; results from only high-quality (≤10 epithelial cells/low-power field [lpf] and ≥25 white blood cells/lpf) sputum specimens were included [14]. Pleural fluid, endotracheal aspirates, and bronchoalveolar-lavage specimens, when collected for clinical care, were included in the analysis. Convalescent serum was collected 3–10 weeks after enrollment.

Real-time PCR was performed on naso-oropharyngeal (NP/OP) swabs for detection of respiratory viruses (adenoviruses [AdV], coronaviruses, human metapneumovirus [HMPV], influenza A and B viruses, parainfluenza virus types 1–3 [PIV], rhinoviruses, and respiratory syncytial virus [RSV]) and atypical bacteria (Mycoplasma pneumoniae and Chlamydia pneumoniae) [14, 15]. Serologic testing for select viruses (AdV, HMPV, influenza A and B, PIV, and RSV) was performed on available paired acute and convalescent serum specimens. Influenza serology used both hemagglutination inhibition and microneutralization assays to define positive results [14, 15], and serology for other viruses was based on CDC-developed indirect immunoglobulin (Ig)G enzyme immunoassays [11, 12]. Bacterial culture was performed on specimens from blood, sputum, pleural fluid, endotracheal aspirates, and bronchoalveolar lavage. To test for specific bacteria, PCR assays were performed (1) on all available pleural fluid specimens from adults and children and (2) on whole blood specimens from children only. Urine antigen testing for Legionella pneumophila and Streptococcus pneumoniae was performed in adults [14].

**Categorization of Variables Based on Diagnostic Test Results**

A positive real-time PCR test for AdV, coronaviruses, HMPV, influenza A/B viruses, PIV, RSV, or rhinovirus was defined as a PCR cycle threshold value <40 from the NP/OP specimen. A positive serologic test for AdV, HMPV, influenza A/B viruses, PIV, or RSV was defined as a ≥4-fold rise in agent-specific IgG antibody titer between paired acute and convalescent sera; for influenza viruses, a minimum titer of 1:40 on the convalescent sample was also needed. Final determination of influenza serology results accounted for timing of influenza vaccination status relative to serum collection (Supplementary Data). For each virus, we categorized viral test results dichotomously as either positive or negative using the PCR and serologic definitions described above. Presence of each virus was defined by a positive PCR or serologic result.

Presence of a bacterial pathogen was defined by the following: detection of *C pneumoniae* or *M pneumoniae* in an NP/OP swab by a PCR assay; culture of bacteria (eg, *Staphylococcus aureus*, *Haemophilus influenzae*, *S pneumoniae*, and *Streptococcus pyogenes*) from blood, endotracheal aspirate, bronchoalveolar-lavage specimen, or pleural fluid; detection of bacteria in pleural fluid by PCR; detection of *S pneumoniae* or *S pyogenes* in whole blood by PCR (children only); detection of *S pneumoniae* or *L pneumophila* by the urine antigen test (adults only); or culture of bacteria from high-quality sputum (adults only). Based on these definitions, a summary variable of any bacterial detection was created.

**Statistical Analysis**

**Multiple Imputation**

Serology results were considered missing if either one or both of the paired serum specimens were missing. We performed multiple imputation to calculate missing values for AdV, HMPV, influenza, PIV, and RSV serology tests (the viruses for which both PCR and serology tests were performed). Multiple imputation uses a multivariable regression model [16–18] that is fit using data from patients for whom the outcome (serology result) is available, to impute a value for patients whose result is missing; this method may be useful when patients had a negative PCR result for a given virus but had a missing serology result. Multiple imputation is based on the assumption that serology results are missing at random [16, 17]. We assessed the assumption of data missing at random by comparing demographic and clinical characteristics among patients with and without available serology results using χ² or Fisher’s exact tests, considering statistical significance (*P* < .05). To create the multivariable logistic regression model used for imputation, among patients with available serology results, we first determined which variables were significantly associated with positive serology results in bivariate analyses. Second, we constructed “a priori” models with positive serology as the outcome using age (ordinal variable), sex, any bacterial detection, and the corresponding viral PCR result as predictors (Supplementary Data). We also considered age as a categorical variable, which did not meaningfully change our results. Third, we identified which of the variables that were significant in bivariate analyses remained significant when added to the multivariable a priori model. Variables considered included symptoms, underlying medical conditions, radiographic findings, intensive care unit admission, and influenza season (Supplementary Data). Fourth, we used the area under the curve (AUC) to determine whether
we would include a variable in the final regression model: the variable was included in the final model only if the AUC was higher for the a priori model when the variable was added to the model compared with the model without that variable. The AUC ranged from 0.5 to 1, with higher values indicating better discrimination of the model at predicting patients with positive versus negative serology results. In addition, we evaluated model fit using the Hosmer-Lemeshow goodness-of-fit test (models with $P < .05$ were rejected). Separate models were constructed for children and adults for each virus. Twenty imputed datasets were created for improved efficiency [19] and pooled (Supplementary Data).

We created a binary outcome for each patient defined as “positive” if either the observed PCR or observed or imputed serology result was positive and as “negative” if the PCR was negative and the imputed serology result was negative. We then estimated the proportion of patients with at least 1 positive result for each virus and the associated 95% confidence interval (CI) [20], referred to as the proportion of CAP with a given virus detected. We performed a sensitivity analysis in which the models for multiple imputation included variables that were significantly different between patients with and without available serology results, to account for these potentially systematic differences. We also performed a second sensitivity analysis in which we excluded patients for whom NP/OP specimens were collected greater than 7 days after illness onset, because the sensitivity of PCR can decrease with a longer interval between illness onset and specimen collection [6, 10].

For each virus, we calculated the absolute difference by subtracting the observed proportion from the imputed proportion, and we calculated the relative difference by dividing the imputed proportion by the observed proportion. To assess precision of multiple imputation estimates, we calculated a range of the absolute and relative differences by subtracting or dividing the observed proportion by the upper and lower limits of the imputed proportion, respectively [21]. All analyses were conducted using SAS 9.3 (Cary, NC).

RESULTS

Children

Of 2638 enrolled children, 2358 (89.4%) met the final radiographic criteria for CAP, among whom 2222 (94.2%) had samples available for both bacterial and viral diagnostic tests. Median age was 2 years (interquartile range [IQR], 1–6); 50.9% of children had an underlying medical condition, including 33.4% that reported asthma [15]. Among these 2222 children, 2196 (98.8%) had NP/OP specimens and 961 (43.2%) had paired serum specimens available. When comparing children with and without paired serology, children with paired serology results were older and more likely to be non-Hispanic white (Table 1). For 2190 (98.6%) children who had an NP/OP swab and illness onset data available, the median time between illness onset and specimen collection was 4.6 days (IQR, 2.8–7.4).

Bivariate analyses informed the multivariable model for imputing missing serology results. Headache was significantly associated with positive AdV serology results; abdominal pain, diarrhea, and any underlying medical condition were significantly associated with positive HMPV serology results; and illness during the influenza season was associated with positive influenza serology results. Abdominal pain, headache, and myalgia were significantly associated with positive PIV serology results, and abdominal pain, chest retraction, chills, diarrhea, headache, myalgia, wheezing, consolidation, pleural effusion, and infiltrates were significantly associated with positive RSV serology results (Supplementary Table S1). After each variable was added to the a priori model (age, sex, any bacterial detection, and the corresponding viral PCR result) for each virus, none of the additional variables remained significant in the models for AdV, PIV, and RSV; thus, the a priori model was used to impute serology results for these viruses. For HMPV and influenza, each of the variables that were significant in the bivariate analysis was added to the a priori model, and only diarrhea and any underlying medical condition remained significant in the multivariable a priori model for HMPV, as did illness during the influenza season for influenza (Table 2). The AUC for the final multivariable models was 0.75 for AdV, 0.88 for HMPV, 0.69 for influenza, 0.74 for PIV, and 0.91 for RSV.

For all 2222 children, the imputed proportions of CAP with AdV (14.4%), HMPV (15.0%), influenza (11.1%), PIV (8.9%), and RSV (29.9%) detected were 1.6%–4.4% higher by absolute difference and 1.1–1.7 times higher by relative difference compared with the observed results for each virus (Table 3). For the first sensitivity analysis, these models were adjusted by variables that significantly differed between children with and without available serology, and the imputed proportions were within ±0.3% of the primary imputed analysis, except for AdV which had a 2.5% lower absolute difference than the unadjusted imputed result (Table 3). For the second sensitivity analysis among 1765 (79.4%) children who had NP/OP specimens collected within 7 days of illness onset, the imputed proportions for AdV, HMPV, influenza, PIV, and RSV were within 0.1%–2.0% of the imputed proportions that included all 2222 children (Supplementary Table S2).

Adults

Of 2488 enrolled adults, 2320 (93.2%) met the final radiographic criteria for CAP, among whom 2259 (97.3%) had specimens available for bacterial and viral diagnostic testing. Median age was 57 years (IQR, 46–71); 78.2% of adults had an underlying medical condition [14]. Among these 2259 adults, 2248 (99.5%) had NP/OP specimens and 846 (37.5%) had paired serum specimens available. When comparing adults with and without available serology, adults with serology results available were more likely to have coronary artery disease and less likely to require invasive mechanical ventilation or to die (Table 4). For 2246 (99.4%) adults that had a NP/OP swab and illness
onset data available, the median time between illness onset and specimen collection was 4.8 days (IQR, 2.8–8.8).

Bivariate analyses informed the multivariable model for imputing missing serology results. Illness during the influenza season, diarrhea, and abdominal pain were associated with positive influenza serology results; wheezing was significantly associated with positive PIV serology results; and chills and wheezing were significantly associated with positive RSV serology results.

Table 1. Characteristics of Hospitalized Children with Community-Acquired Pneumonia: Comparison of Children With and Without Available Serology Data

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Children With Available Paired Serology Data (n = 961) No. (%)</th>
<th>Children Without Available Paired Serology Data (n = 1261) No. (%)</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age groups</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;2 years</td>
<td>359 (37.4%)</td>
<td>621 (49.2%)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>2–4 years</td>
<td>241 (25.1%)</td>
<td>318 (25.2%)</td>
<td></td>
</tr>
<tr>
<td>5–9 years</td>
<td>217 (22.6%)</td>
<td>191 (15.2%)</td>
<td></td>
</tr>
<tr>
<td>10–17 years</td>
<td>144 (15.0%)</td>
<td>131 (10.4%)</td>
<td></td>
</tr>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>546 (56.8%)</td>
<td>880 (53.9%)</td>
<td>.175</td>
</tr>
<tr>
<td>Female</td>
<td>415 (43.2%)</td>
<td>581 (46.1%)</td>
<td></td>
</tr>
<tr>
<td><strong>Race/Ethnicity</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-Hispanic White</td>
<td>439 (45.7%)</td>
<td>433 (34.3%)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Non-Hispanic Black</td>
<td>235 (24.5%)</td>
<td>530 (42.0%)</td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>213 (22.2%)</td>
<td>201 (15.9%)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>74 (7.7%)</td>
<td>97 (7.7%)</td>
<td></td>
</tr>
<tr>
<td><strong>Underlying Medical Conditions</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None reported</td>
<td>484 (50.4%)</td>
<td>606 (48.1%)</td>
<td>.281</td>
</tr>
<tr>
<td>Asthma</td>
<td>314 (32.7%)</td>
<td>429 (34.0%)</td>
<td>.505</td>
</tr>
<tr>
<td>Congenital heart disease</td>
<td>80 (8.3%)</td>
<td>79 (6.3%)</td>
<td>.062</td>
</tr>
<tr>
<td>Preterm birth in children under 2 years</td>
<td>64/359 (17.8%)</td>
<td>141/621 (22.7%)</td>
<td>.070</td>
</tr>
<tr>
<td>Intensive care unit admission</td>
<td>214 (22.3%)</td>
<td>249 (19.8%)</td>
<td>.144</td>
</tr>
<tr>
<td>Mechanical ventilation</td>
<td>64 (6.7%)</td>
<td>85 (6.7%)</td>
<td>.940</td>
</tr>
<tr>
<td>Death</td>
<td>0 (0%)</td>
<td>3 (0.2%)</td>
<td>.263</td>
</tr>
</tbody>
</table>

*Underlying medical conditions included asthma; congenital heart disease, coronary artery disease; preterm birth (defined as gestational age <37 weeks at birth in children under 2 years old); diabetes mellitus; chronic kidney disease; chronic liver disease; immunosuppression; any cancer (excluding skin cancers); neurological disorders (including seizure, cerebral palsy, scoliosis); and chromosomal disorders (including Down’s syndrome).

Table 2. Multivariable Models to Determine Covariates Independently Associated With Select Respiratory Virus Positive Paired Serology Results Among Children Hospitalized With Community-Acquired Pneumonia With Available Paired Serology Results (n = 961)

<table>
<thead>
<tr>
<th>Covariates</th>
<th>Adenoviruses</th>
<th>Human Metapneumovirus</th>
<th>Influenza Viruses</th>
<th>Parainfluenza Viruses</th>
<th>Respiratory Syncytial Virus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>aOR</td>
<td>95% CI</td>
<td>aOR</td>
<td>95% CI</td>
<td>aOR</td>
</tr>
<tr>
<td>Age*</td>
<td>0.67</td>
<td>(0.49–0.92)</td>
<td>0.88</td>
<td>(0.69–1.13)</td>
<td>1.10</td>
</tr>
<tr>
<td>Sex†</td>
<td>0.80</td>
<td>(0.42–1.55)</td>
<td>1.03</td>
<td>(0.60–1.78)</td>
<td>0.98</td>
</tr>
<tr>
<td>Any bacterial detection</td>
<td>0.52</td>
<td>(0.15–1.75)</td>
<td>0.86</td>
<td>(0.36–2.08)</td>
<td>0.94</td>
</tr>
<tr>
<td>NP/OP result for the given virus</td>
<td>5.59</td>
<td>(2.82–11.11)</td>
<td>52.82</td>
<td>(30.37–91.84)</td>
<td>29.92</td>
</tr>
<tr>
<td>Any underlying medical conditions</td>
<td>—</td>
<td>—</td>
<td>0.52</td>
<td>(0.29–0.91)</td>
<td>—</td>
</tr>
<tr>
<td>Diarrhea</td>
<td>—</td>
<td>—</td>
<td>1.95</td>
<td>(1.10–3.47)</td>
<td>—</td>
</tr>
<tr>
<td>Influenza season</td>
<td>—</td>
<td>—</td>
<td>2.04</td>
<td>(1.08–3.88)</td>
<td>—</td>
</tr>
</tbody>
</table>

Abbreviations: aOR, adjusted odds ratio; CI, confidence interval; NP/OP, nasopharyngeal/oropharyngeal.

*Ordinal variable with age groups as <2, 2–4, 5–9, 10–17 years.
†Reference category is male sex.
No additional variables were significant for positive AdV and HMPV serology results; thus, the a priori model was used to impute serology results for these viruses. For PIV, wheezing was not significant in the a priori model so the a priori model was used to impute serology results. Abdominal pain remained significant in the multivariable a priori model for influenza, as did wheezing for RSV (Table 5). The AUC for the final multivariable models was 0.80 for AdV, 0.83 for HMPV, 0.86 for influenza, 0.82 for PIV, and 0.91 for RSV.

For all 2259 adults, the imputed proportions of CAP with AdV (4.2%), HMPV (4.7%), influenza (7.9%), PIV (4.0%), and RSV (4.0%) detected were 0.8%–2.8% higher by absolute
difference and 1.2–3 times higher by relative difference than the observed results for each virus (Table 6). For the first sensitivity analysis, models were initially adjusted by variables that significantly differed between adults with and without serology; however, CIs were unstable due to few deaths, and thus models were adjusted for coronary artery disease and mechanical ventilation. These imputed proportions were within an absolute difference of 0.5% from those in the unadjusted imputed analyses, except for PIV, which was 7.8% higher (Table 6). For the second sensitivity analysis among 1600 (70.8%) adults who had NP/OP specimens collected within 7 days of illness onset, the imputed proportions of AdV, HMPV, influenza, PIV, and RSV were within 0.4%–1.1% of the imputed proportions that included all 2259 adults (Supplementary Table S4).

### DISCUSSION

Using prospective clinical and microbiological data from patients hospitalized with CAP from a large multicenter study, we estimated the proportion of patients with AdV, HMPV, influenza, PIV, and RSV detected, using multiple imputation to account for missing serology results. Our multiple imputation estimates were 0.8%–4.4% higher by absolute difference and 1.1–3.0 times higher by relative difference than the observed results. Virus-specific prevalence of CAP will likely be underestimated if missing serology results are treated as negative. Multiple imputation can help refine virus-specific prevalence when results are missing and will likely increase the proportion of viral etiologies.

### Table 5. Multivariable Models to Determine Covariates Independently Associated With Select Respiratory Virus-Positive Paired Serology Results Among Adults Hospitalized With Community-Acquired Pneumonia With Available Paired Serology Results (n = 846)

<table>
<thead>
<tr>
<th>Covariates</th>
<th>Adenoviruses</th>
<th>Human Metapneumovirus</th>
<th>Influenza Viruses</th>
<th>Parainfluenza Viruses</th>
<th>Respiratory Syncytial Virus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Adjusted aOR</td>
<td>95% CI</td>
<td>aOR</td>
<td>95% CI</td>
<td>aOR</td>
</tr>
<tr>
<td>Age†</td>
<td>2.44</td>
<td>(0.92–6.48)</td>
<td>1.14</td>
<td>(0.64–2.04)</td>
<td>0.75</td>
</tr>
<tr>
<td>Sex†</td>
<td>0.47</td>
<td>(0.12–1.94)</td>
<td>0.62</td>
<td>(0.22–1.76)</td>
<td>1.08</td>
</tr>
<tr>
<td>Any bacterial detection &lt;0.001</td>
<td>1.71</td>
<td>(0.42–7.03)</td>
<td>0.15</td>
<td>(0.02–1.26)</td>
<td>1.30</td>
</tr>
<tr>
<td>NP/OP result for the given virus</td>
<td>344.1</td>
<td>(41.37–&gt;999.99)</td>
<td>85.85</td>
<td>(30.04–245.37)</td>
<td>70.10</td>
</tr>
<tr>
<td>Abdominal pain</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>2.65</td>
</tr>
<tr>
<td>Wheezing</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

### Table 6. Multiple Imputation Estimations of an Imputed Proportion of Community-Acquired Pneumonia with Detections of Specific Respiratory Viruses Among Adults Hospitalized With Community-Acquired Pneumonia (n = 2259)

<table>
<thead>
<tr>
<th>Estimates</th>
<th>Adenoviruses</th>
<th>Human Metapneumovirus</th>
<th>Influenza Viruses</th>
<th>Parainfluenza Viruses</th>
<th>Respiratory Syncytial Virus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed virus test result by PCR and serology (%)</td>
<td>1.4% (0.9%–1.9%)</td>
<td>3.9% (3.1%–4.7%)</td>
<td>5.8% (4.8%–6.8%)</td>
<td>3.0% (2.3%–3.7%)</td>
<td>3.0% (2.3%–3.7%)</td>
</tr>
<tr>
<td>Primary analysis: multiple imputation estimate* (%)</td>
<td>4.2% (−13.2% to 21.6%)</td>
<td>4.7% (3.7%–5.7%)</td>
<td>7.9% (6.3%–9.5%)</td>
<td>4.0% (3.0%–4.9%)</td>
<td>4.0% (2.9%–5.0%)</td>
</tr>
<tr>
<td>Absolute difference†, % (range)</td>
<td>2.8% (−14.6% to 20.2%)</td>
<td>0.8% (−0.2% to 1.8%)</td>
<td>2.1% (0.5%–3.7%)</td>
<td>1.0% (0.0%–1.9%)</td>
<td>1.0% (−0.1% to 2.0%)</td>
</tr>
<tr>
<td>Relative difference‡, % (range)</td>
<td>3.0 (−9.4 to 15.4)</td>
<td>1.2 (0.9–1.5)</td>
<td>1.4 (1.1–1.6)</td>
<td>1.3 (1.0–1.6)</td>
<td>1.3 (1.0–1.7)</td>
</tr>
<tr>
<td>Sensitivity analysis: multiple imputation estimate* accounting for coronary artery disease, mechanical ventilation (%)</td>
<td>4.7% (−3.9% to 13.3%)</td>
<td>4.8% (3.6%–6.0%)</td>
<td>7.9% (6.3%–9.5%)</td>
<td>11.8% (−15.3% to 39.0%)</td>
<td>4.0% (3.0%–5.0%)</td>
</tr>
</tbody>
</table>

Abbreviations: aOR, adjusted odds ratio; CI, confidence interval; NP/OP, nasopharyngeal/oropharyngeal
†Ordinal variable with age groups: 19–49, 50–64, and ≥65 years.
‡Reference category is male sex.
*Confidence intervals for the multiple imputation estimate were derived as part of the multiple imputation procedure in SAS.
†Absolute difference: subtraction of the observed proportion from the imputed proportion.
‡Relative difference: division of the imputed proportion by the observed proportion.
The absolute and relative increases found in the viral-specific detections with multiple imputation were modest among both children and adults. It is notable that the largest absolute increase was for influenza among children (4.4%), and the largest relative increase was for AdV among adults (3-fold); however, the CI for this imputed AdV estimate was wide. The reasons for the differences in the multiple imputation results compared with observed results are likely multifactorial. Serology has been shown to increase the diagnostic yield for AdV, HMPV, PIV, and RSV by 3.0%–86.7% [11, 12, 22] and for influenza viruses by 15%–130% [3, 4, 11, 12] when compared with PCR across all ages. The diagnostic yield of one method versus another may differ by age, because we observed in the EPIC study that the diagnostic yields for AdV, HMPV, PIV, and RSV were 11.8%–48.9% among children and adults [22]. These differences indicate that serology increases diagnostic yield for most viruses but differs by specific virus and age of the patient. In addition, serology may capture detections missed by PCR because of late collection of NP/OP specimens relative to illness onset, inadequate sample collection or preservation, or collection of an NP/OP specimen before onset of viral shedding.

Because of potential differences in viral dynamics, the timing of illness onset and collection of NP/OP and serum specimens may partially explain the different multiple imputation estimates for each virus. However, we did not observe a meaningful change in the virus-specific prevalence when excluding patients for whom NP/OP specimens were collected after 7 days from illness onset (Supplementary Tables S2 and S4). Ultimately, the increases in virus-specific proportions were due to the increase in serologic detections estimated through multiple imputation. For patients who had a positive PCR result and missing serology, imputing missing serology would not have increased the virus-specific proportion because the positive PCR result would have already contributed to the proportion. Thus, an increase in virus-specific proportion would have been observed only if an observed PCR result is negative and an imputed serology result is positive. Our analysis quantified these increases in virus-specific detections by using multiple imputation, demonstrating the utility of this method when data are missing.

Multiple imputation can be useful in situations where convalescent sera are not uniformly available [16, 17], and it is important to assess whether these data were missing at random. More than half of enrolled children and adults were missing serology results, largely due to the lack of a convalescent specimen. Although we observed differences between patients with and without available serology results, it does not necessarily indicate a major violation of the assumption. Systematic differences can occur between patients with and without missing data even when data are missing at random; these differences can be explained by other observed variables [23]. After accounting for these differences through a sensitivity analysis, the multiple imputation estimates were overall similar to those from the main analysis; the only exceptions were AdV among children and PIV among adults, but the difference in these assessments may be partly explained by sparse data within strata.

Multiple imputation has been used in other pneumonia studies to handle missing data on covariates when estimating an association between an exposure and an outcome, as an alternative to performing an analysis that only includes subjects with complete data. For example, in a randomized trial of an indoor wood smoke emission intervention to prevent childhood pneumonia, approximately one third of children had an incomplete assessment of various pneumonia outcomes due to missing data on physician diagnosis, respiratory pathogen testing, and/or chest radiograph. By applying multiple imputation, some significant associations were observed for specific pneumonia outcomes, and the authors considered these outcomes based on imputation to be reliable effect estimates [24]. In addition, 2 meta-analyses estimated the global burden of pediatric respiratory infections due to seasonal influenza [25] and RSV [26] using multiple imputation when necessary to account for missing incidence data.

Our analyses have limitations. First, imputing missing data assumed they were missing at random; evaluation of “missingness” demonstrated that differences existed between those with and without serology specimens available, but those differences were minimal. Second, estimates from multiple imputation were sensitive to the independent variables used and the strength of their association with serology results among patients with available data. This concern may be minimal because we used the viral PCR result as an independent variable, which was a strong predictor of the paired serology result. Third, although not a limitation of this analysis, the detection of viruses in NP/OP swabs is not necessarily indicative of the cause of CAP [27, 28] and may represent (1) resolving infection rather than acute infection or (2) an infection limited to the upper tract and not lower tract [29]. Similarly, paired serology could represent an interval infection that occurred after the initial presentation or hospitalization and may not be indicative of the cause of the CAP hospitalization; thus, including viral serology results in any analysis could overestimate the true viral burden of CAP.

Fourth, we combined influenza A and B viruses together and parainfluenza types 1–3 together, although the predictors of positive serology results may differ by each virus's type. Finally, illness onset was self- or caregiver-reported, which introduces potential recall bias. However, our sensitivity analyses indicated that virus-specific prevalence estimates did not change significantly when excluding patients who had NP/OP swabs collected after 7 days from illness onset.

CONCLUSIONS

In conclusion, because of missing serum specimens, the proportion of CAP with AdV, HMPV, influenza, PIV, and
RSV detected using only the observed PCR and serology results likely underestimated the virus-specific prevalence by 0.8%–4.4% in both children and adults. Multiple imputation estimates may help achieve more accurate disease burden estimates, better establish performance measures of the molecular assays, and infer virus-specific prevalence estimates in other studies when convalescent serum specimens are not available for all patients. Our reported higher estimates of viral-specific pneumonia prevalence based on multiple imputation underscore the need for improved respiratory viral diagnostic testing, which would better inform population-based estimates and better guide clinicians and public health policy.

Supplementary Data
Supplementary materials are available at Open Forum Infectious Diseases online. Consisting of data provided by the authors to benefit the reader, the posted materials are not copiededit and are the sole responsibility of the authors, so questions or comments should be addressed to the corresponding author.

Acknowledgments
Financial support. The Etiology of Pneumonia in the Community study was funded by the Influenza Division in the National Center for Immunizations and Respiratory Diseases at the Centers for Disease Control and Prevention through cooperative agreements with each study site and was based on a competitive research funding opportunity. Additional funding was provided in part by the National Institutes of Health: Grants K24AI14444 (to N. R. G., Principal Investigator [PI]), R21AI112768-01A1 (to J. E. V., PI), K23AI104779 (to D. J. W., PI), R01AG043471 (to C. G. G., PI), K23GM110469 (to PI), and K23GM110469 (to W. H. S., PI). K. A. reports receiving grants from CDC during the conduct of the study; other support from GlaxoSmithKline, other support from Cubist Pharmaceuticals outside the submitted work; and collaborating with BioFire Diagnostics, Inc. (formerly Idaho Technology, Inc.) on several NIH grants. D. F. reports receiving personal fees from Biogen Idec, outside the submitted work, and owning shares of Johnson and Johnson. C. G. G. reports receiving grants from CDC during the conduct of the study; other support from Pfizer, other support from Merck, grants from NIH, grants from AHRQ outside the submitted work. D. J. W. reports receiving grants from CDC during the conduct of the study. S. R. A. reports receiving grants from CDC during the conduct of the study; other support from Johnson and Johnson. S. R. A. reports receiving grants from NIH outside the submitted work. W. H. S. reports receiving personal fees from Venaxis, Inc., personal fees from Cempa Pharmaceuticals, personal fees from Ferring Pharmaceuticals, personal fees from BioTest AG, personal fees from Abbott Point of Care, from Gilead Pharmaceuticals, outside the submitted work.

Disclaimer. The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention or the National Institutes of Health.

Potential conflicts of interest. All authors: No reported conflicts of interest. All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest.

References