Vaccine 37 (2019) 1325–1331

Contents lists available at ScienceDirect

Vaccine

journal homepage: www.elsevier.com/locate/vaccine

Use of three summary measures of pediatric vaccination for studying the safety of the childhood immunization schedule

Stanley Xu [a,b,*], Sophia R. Newcomer [c], Martin Kulldorff [d], Matthew F. Daley [a,e], Bruce Fireman [f], Jason M. Glanz [a,b]

[a] Institute for Health Research, Kaiser Permanente Colorado, United States
[b] University of Colorado Denver, School of Public Health, United States
[c] University of Montana, School of Public and Community Health Sciences, United States
[d] Brigham and Women's Hospital and Harvard Medical School, Division of Pharmacoepidemiology and Pharmacoeconomics, United States
[e] University of Colorado Denver, School of Medicine, Department of Pediatrics, United States
[f] Kaiser Permanente Northern California, Division of Research, Vaccine Study Center, United States

ARTICLE INFO

Article history:
Received 17 July 2018
Received in revised form 10 January 2019
Accepted 14 January 2019
Available online 29 January 2019

Keywords:
Summary measures
Pediatric vaccination
Immunization safety
Collinearity
Variance inflation factor

ABSTRACT

Background: Summary measures such as number of vaccine antigens, number of vaccines, and vaccine aluminum exposure by the 2nd birth day are directly related to parents' concerns that children receive too many vaccines over a brief period. High correlation among summary measures could cause problems in regression models that examine their associations with outcomes.
Objectives: To evaluate the performance of multiple regression models using summary measures as risk factors to simulated binary outcomes.
Methods: We calculated summary measures for a cohort of 232,627 children born between 1/1/2003 and 9/31/2013. Correlation and variance inflation factors (VIFs) were calculated. We conducted simulations (1) to examine the extent to which an association can be detected using a summary measure other than the true risk factor; (2) to evaluate the performance of multiple regression models including true and redundant risk factors; (3) to evaluate the performance of multiple regression models when all three were risk factors; (4) to examine the performance of multiple regression models with incorrect relationship between risk factors and outcome.
Results: These summary measures were highly correlated. VIFs were 7.14, 6.25 and 2.17 for number of vaccine antigens, number of vaccines, and vaccine aluminum exposure, respectively. In simulations, an association would be detected if a summary measure other than the true risk factor was used. The power to detect the association between the true risk factor and outcome significantly decreased if redundant risk factors were included. When all three were risk factors, multiple regression model was appropriate to detect the stronger risk factor. Correctly specifying the relationship between risk factors and the outcome was crucial.
Conclusions: Multiple regression models can be used to examine the association between summary measures and outcome despite of high correlation among summary measures. It is important to correctly specify the relationship between risk factors and outcome.

1. Introduction

The current immunization schedule recommended by the U.S. Advisory Committee for Immunization Practices (ACIP) [1] has children receiving vaccines to protect against a total of 14 diseases before the 2nd birthday. Although the public health benefit of childhood immunization is evident, some parents have concerns that children receive too many vaccines over a brief period. In response to these concerns, the Institute of Medicine (IOM) recently issued a report on the safety of the current immunization schedule and concluded that available scientific evidence supported the safety of the recommended schedule. However, the IOM report also recommended further observational studies and stated that research networks such as the Vaccine Safety Datalink (VSD) were the best available resources for such research [2].

The VSD is a collaborative project between the Immunization Safety Office of the Centers for Disease Control and Prevention (CDC) and nine health maintenance organizations (HMOs) to

* Corresponding author at: Institute for Health Research, Kaiser Permanente Colorado, 2550 S Parker Rd, Suite 200, Aurora, CO 80014, United States.
E-mail address: stan.xu@kp.org (S. Xu).

https://doi.org/10.1016/j.vaccine.2019.01.040
0264-410X/

conduct post-marketing surveillance of vaccine safety using electronic health record (EHR) data [3,4]. Vaccination data collected through VSD allow researchers to characterize patterns of early childhood vaccination in numerous ways [5–7]. Among them, summary measures such as cumulative exposure to vaccine doses, vaccine antigens, and aluminum in vaccines are directly related to parents' concerns that children receive too many vaccines over a brief period of time. Safety studies that examine the association between these measures and various outcome of interest to parents can address parental concerns of vaccination schedule safety [7]. For example, studies can examine whether the cumulative number of vaccine antigens that a child received at certain age is associated with experiencing an outcome of interest. However, the high correlation among these summary measures could potentially cause problems in regression models that include these summary measures, namely collinearity [8]. Consequently, if a summary measure appears to be associated with the outcome of interest, it may not be possible to detect which risk factors are true or redundant in the multiple regression model and the resulting coefficient estimates may not be valid [9,10]. In a multiple regression model, an independent variable (e.g., summary measure) that is truly associated with an outcome of interest is called the true risk factor while the other highly correlated independent variables (other summary measures) are called redundant risk factors.

Thus, our objectives of this paper were, through simulation, (1) to examine the extent to which an association can be detected using a summary measure other than the true risk factor if the association between the outcome and the true risk factor exists; (2) to compare the performance of multiple logistic regression models including true and redundant risk factors to regression models with only the true risk factor using estimation bias and empirical power; (3) to evaluate the performance of multiple logistic regression models with all three measures when one is a strong risk factor and the other two are weak risk factors; (4) to examine the performance of multiple logistic regression models with all three measures when the relationship of the true risk factor with the outcome is not correctly specified.

2. Methods

2.1. Study population and covariates

A cohort of 232,627 children born between 1/1/2003 and 9/31/2013 with at least two years of continuous enrollment in two HMO sites (Kaiser Permanente Colorado and Kaiser Permanente Northern California) within the VSD project was assembled. Using EHR data from both sites, we created an analytic dataset containing covariates such as gender, premature status at birth, chronic condition count index (CCCI), and well-care visit (WCV). Premature status was defined as gestational age less than 37 weeks, supplemented with International Classification of Diseases, 9th edition (ICD9) diagnosis code "765.xx" (excluding 765.29) and birth weights less than 2500 g. The Pediatric Medical Complexity Algorithm (PMCA) [11] was used to identify the body systems in which the child had a diagnosed chronic condition from birth through the 2nd birthday, and this number was summed to create CCCI. WCV was defined as the sum of well-care visits in the first two years of life for each cohort member.

2.2. Three summary measures of pediatric vaccination for studying immunization schedule safety

Records of vaccination between birth and the 2nd birthday for each cohort member were obtained from EHR data including vaccine type, and dates of administration. We considered the following vaccines in this study: hepatitis B (HepB), diphtheria-tetanus-acellular pertussis (DTaP), *haemophilus influenza*e type b (Hib), pneumococcal, polio, rotavirus, measles-mumps-rubella (MMR), varicella, seasonal influenza, 2009–2010 pandemic H1N1 influenza, hepatitis A, and any combination of the above vaccines. Records of other immunizations were excluded from the calculation of summary measures. We used these vaccination records to create the following three summary measures of early childhood immunization schedule exposure by the 2nd birthday.

(1) *Number of vaccines* received by the 2nd birthday. This measure was quantified as the cumulative number of individual vaccine doses that a child received by the 2nd birthday. Combination vaccines (for example, DTaP-IPV-HepB) are considered as one vaccine dose in this measure. Both oral (for example, rotavirus) and injected vaccines were counted.
(2) *Number of vaccine antigens* received by the 2nd birthday. Vaccine antigen exposure was quantified as the number of immunogenic proteins and polysaccharides in each vaccine as described in Glanz et al. [12]. We calculated *number of vaccine antigens* contained in each vaccine dose that a child received and summed these antigens across all vaccines received by the 2nd birthday.
(3) *Vaccine aluminum exposure* (milligrams). Some vaccines include aluminum as an adjuvant, whereas others do not. This information is available on package inserts. The information of aluminum content in early childhood vaccines can be found in Table 2 in Glanz et al. [13]. We used the maximum amount of aluminum per dose in milligrams (mg) to calculate *vaccine aluminum exposure* before children's 2nd birthdays.

For those children with no vaccination records before the 2nd birthday, we set the above three summary measures equal to zero.

2.3. Rescaling three summary measures, correlations and variance inflation factors

These three original summary measures had different scales with means (standard deviation) for *number of vaccines*, *number of vaccine antigens*, and *vaccine aluminum exposure* of 18.63 (3.66), 245.01 (49.26), and 3.95 (0.66) for this study population, respectively. Coefficients estimated from statistical models with different summary measures as risk factors would not be directly comparable, making it hard to achieve the objectives described earlier. To create comparable estimated coefficients, we employed the range value (i.e., the maximum minus the minimum) to rescale the three summary measures using formulas (1):

$$\text{Range} - \text{rescaled measure} = \frac{\text{original measure}}{\text{range of original measure}} \quad (1)$$

We generated descriptive statistics and evaluated the correlations among the three summary measures and rescaled summary measures. We used range-rescaled measures for subsequent analyses.

Inclusion of all three measures in a multiple regression model may result in collinearity due to their high correlation. To measure the impact of collinearity on parameter estimation, we regressed each measure (response variable) against the other two in a linear regression model to obtain R-squared (R^2) – the percentage of the response variable variation that is explained by a linear model. We then calculated variance inflation factor (*VIF*) as follows [14]:

$$VIF_{(X)} = \frac{1}{1 - R_X^2} \quad (2)$$

where $VIF_{(X)}$ is the *VIF* for the *X* summary measure, R_X^2 is the R^2 where the *X* summary measure is regressed on the other two measures. Note that these *VIF*s are specific to this study population because these summary measures were not simulated; they were derived from the study population described above.

Let $\widehat{\beta}_X$ denote the estimated coefficient for the association between risk factor *X* (a summary measure) and a clinical outcome (*y*), and $\widehat{\sigma}_{\widehat{\beta}_X}$ denote the standard error of $\widehat{\beta}_X$ from a regression model with *y* as the dependent variable and only *X* as the independent variable. If the redundant factors are completely redundant, inclusion of the other two redundant risk factors will not change the point estimate $\widehat{\beta}_X$, but will have impact on the estimated $\widehat{\sigma}_{\widehat{\beta}_X}$. Let $\widehat{\sigma^*}_{\widehat{\beta}_X}$ represent the estimated standard error of $\widehat{\beta}_X$ from a regression model with the true and redundant risk factors, theoretically

$$\widehat{\sigma^*}_{\widehat{\beta}_X} = \sqrt{VIF_X}\,\widehat{\sigma}_{\widehat{\beta}_X} \quad \text{or} \quad \sqrt{VIF_X} = \frac{\widehat{\sigma^*}_{\widehat{\beta}_X}}{\widehat{\sigma}_{\widehat{\beta}_X}} \tag{3}$$

Thus, the z-value, the ratio of a point estimate and its standard error, decreases by $\sqrt{VIF_X}$ fold.

2.4. Simulation schemes

We used these range-rescaled summary measures to simulate the true association between an immunization schedule exposure summary measure and an outcome. First, we simulated the effect of a range-rescaled measure (X) on an outcome (y). The outcome occurred after the 2nd birthday. Model (4) was used to simulate the outcome.

$$\text{prob}(y=1) = \frac{\exp(\beta_0 + CCCI * \beta_{CCCI} + gender * \beta_{gender} + premature * \beta_{premature} + wcv * \beta_{WCV} + X * \beta_X)}{1 + \exp(\beta_0 + CCCI * \beta_{CCCI} + gender * \beta_{gender} + premature * \beta_{premature} + wcv * \beta_{WCV} + X * \beta_X)} \tag{4}$$

Model (4) also included other covariates such as gender, premature status at birth, CCCI and WCV; the distribution of gender, premature status at birth, comorbidity and well-care visits were from the study population; β_{gender}, $\beta_{premature}$, β_{CCCI}, and β_{WCV} are their coefficients.

We chose $\beta_{gender} = 0.2$, $\beta_{premature} = 0.4$, $\beta_{CCCI} = 0.2$, and $\beta_{WCV} = 0.2$. We also set $\beta_0 = -5.0$. For the main effect, we chose $\beta_X = 0.1$, $\beta_X = 0.2$ and $\beta_X = 0.3$. For each scenario 2000 datasets with a different random seed were simulated.

To examine the extent to which an association can be detected using a summary measure other than the true risk factor if the association between the outcome and the true risk factor exists, we analyzed the simulated datasets with three range-rescaled summary measures separately.

To compare the performance of a multiple logistic regression model including the true and redundant risk factors to the one with only the true risk factor, we analyzed the simulated data (1) with only the true risk factor and other covariates, and (2) with three measures and other covariates in the model. To assess the impact of redundant risk factors, we also examined how much β_X must be increased to yield approximately 90% of empirical power from a regression model with all summary measures.

To evaluate the performance of the multiple logistic regression model with all three measures when one is a strong risk factor and the other two are weak risk factors, we simulated outcome data with $\beta_X = 0.2$ and $\beta_X = 0.3$ for the strong risk factor and a coefficient of 0.1 for the other two weak risk factors. We analyzed the data using a multiple logistic regression model (1) with each of the three risk factors only and (2) with both the strong and weak risk factors.

To examine the performance of the multiple logistic regression model with all three measures when the relationship of the true risk factor is not correctly specified, we assumed that only those with a measure value greater than its 95th percentile had an elevated risk of an adverse outcome, thus the relationship between the true risk factor and the logit of the probability of an outcome is nonlinear. In this dataset, the 95th percentiles for rescaled *number of vaccine antigens*, *number of vaccines* and *vaccine aluminum exposure* were 0.71, 0.72, and 0.84, respectively. We simulated outcome data with $\beta_X = 0.1$, $\beta_X = 0.2$ and $\beta_X = 0.3$ for the true risk factor with the true nonlinear relationship. We then analyzed the data with three models: (1) with only the true risk factor and its correct nonlinear relationship with the outcome; (2) with only the true risk factor but an incorrect linear relationship; (3) with the true risk factor and two redundant risk factors but all having an incorrect linear relationship with the outcome.

From 2000 replica, we reported the mean (standard error) estimates of coefficients for the risk factors. Empirical power for detecting the association between the outcome and risk factors was calculated as the proportion of 2000 replica where a statistical test correctly rejected a false null hypothesis when simulations were performed under the true alternative. We also calculated the average ratios of standard errors of $\widehat{\beta}_X$ from the model with three measures and from the model with only the true risk factor and compared them to the square root of *VIF*s when one measure had a true linear relationship with the outcome.

3. Results

Table 1 shows the descriptive statistics of the original and rescaled summary measures. *Number of vaccine antigens* had the largest mean and standard deviation (245.01 and 49.26) while *vaccine aluminum exposure* had the smallest mean and standard deviation (3.95 and 0.66). By design, the range-rescaled measures had the same range (0–1).

Number of vaccine antigens and *number of vaccines* had the highest correlation with a Pearson correlation coefficient equal to 0.92. The correlations between *vaccine aluminum exposure* and the other two measures were slightly lower, 0.73 and 0.69, respectively. All Pearson correlation coefficients had a p-value less than 0.0001. Note that the rescaling did not change the correlations among the three measures due to a linear transformation of the variables.

The *VIF*s were 7.14, 6.25 and 2.17 when the range-scaled *number of vaccine antigens*, *number of vaccines*, and *vaccine aluminum exposure* were each regressed against the other two measures, respectively. If the value of *VIF* is greater than 5, then multicollinearity is considered high. The square roots of *VIF*s were 2.67, 2.50 and 1.47 and can be used to measure how much larger the standard error of the coefficient for the true risk factor ($\widehat{\sigma^*}_{\widehat{\beta}_X}$) was compared to the same measure in a model that did not include highly correlated redundant risk factors [14]. These values were later compared to ratios of standard errors from the model with

Table 1
Descriptive statistics of original and rescaled summary measures.

	Original measures			Range-rescaled measures		
	Number of vaccine antigens	*Number of vaccines*	*Vaccine aluminum exposure* (mg)	*Number of vaccine antigens*	*Number of vaccines*	*Vaccine aluminum exposure* (mg)
Mean (std)	245.01 (49.26)	18.63 (3.66)	3.95 (0.66)	0.59 (0.12)	0.58 (0.11)	0.74 (0.12)
median	255.00	19.00	4.15	0.62	0.59	0.78
Min-max	0–414.00	0–32	0–5.35	0–1	0–1	0–1

Table 2
Impact of using summary measures other than the true risk factor on coefficient estimates and empirical power when the association between the outcome and the true risk factor exists.

Trueβ_X	Evaluation measures	*Number of antigens* as the true risk factor			*Number of vaccines* as the true risk factor			*Vaccine aluminum exposure* as the true risk factor		
		Number of antigens	Number of vaccines	Aluminum (mg)	Number of antigens	Number of vaccines	Aluminum (mg)	Number of antigens	Number of vaccines	Aluminum (mg)
0.100	Mean $\widehat{\beta}_X$(std)	0.103 (0.090)	0.099 (0.093)	0.069 (0.087)	0.090 (0.090)	0.105 (0.093)	0.063 (0.087)	0.074 (0.089)	0.074 (0.092)	0.102 (0.087)
	Power (%)	22.0	17.9	13.0	17.0	18.9	11.0	12.8	11.5	22.3
0.200	Mean $\widehat{\beta}_X$(std)	0.202 (0.090)	0.192 (0.092)	0.135 (0.086)	0.176 (0.089)	0.203 (0.092)	0.122 (0.086)	0.143 (0.086)	0.141 (0.089)	0.201 (0.085)
	Power (%)	64.8	56.9	36.8	53.7	60.9	30.0	39.4	35.8	67.1
0.300	Mean $\widehat{\beta}_X$(std)	0.302 (0.087)	0.286 (0.089)	0.201 (0.083)	0.262 (0.087)	0.303 (0.090)	0.181 (0.083)	0.212 (0.083)	0.208 (0.086)	0.301 (0.083)
	Power (%)	94.4	91.0	69.1	87.6	93.1	60.0	73.6	69.3	96.2

three summary measures and the one with only the true risk factor in the simulation study.

In the simulations, when *number of vaccine antigens* was truly associated with the outcome (Table 2), as expected, the model with *number of vaccine antigens* yielded unbiased estimates and can be considered the gold standard. Using *number of vaccines* as the main exposure slightly underestimated the association while using *vaccine aluminum exposure* significantly underestimated the association. Consistently, the empirical power for detecting an association between the measure and outcome only slightly decreased if *number of vaccines* was used, while the empirical power decreased at a greater rate if *vaccine aluminum exposure* was used (Table 2). For example, when the coefficient for the association between *number of vaccine antigens* and the outcome was 0.20, the empirical power was 64.8% if *number of vaccine antigens* was used in the model, 56.9% if *number of vaccines* was used in the model, and 36.8% if *vaccine aluminum exposure* was used in the model.

When *number of vaccines* was truly associated with the outcome, we observed comparable results where the coefficient was slightly underestimated and yielded slightly lower empirical power when *number of vaccine antigens* was modeled. If *vaccine aluminum exposure* was modeled, the coefficient was underestimated, and the empirical power decreased significantly. When *vaccine aluminum exposure* was truly associated with the outcome, either using *number of vaccine antigens* or using *number of vaccines* underestimated the coefficient and yielded significantly low empirical power.

The multiple logistic regression model with the true risk factor and two redundant risk factors yielded a comparable point estimate for the true risk factor but significantly larger standard error (Table 3). For example, when the coefficient for the true association between *number of vaccine antigens* and the outcome was 0.20, the model with *number of vaccine antigens* only gave a point estimate of 0.202 with a standard error of 0.090, while the multiple logistic regression model with three measures yielded a comparable point estimate (0.199) but a more than doubled standard error (0.218). The same phenomena were observed across different true risk factors and different levels of association except that the impact was less if the true risk factor was *vaccine aluminum exposure* due to its relatively lower correlation with the other two summary measures. The average ratios of standard errors from the model with all summary measures and the one with only the risk factors were 2.43, 2.30, and 1.35 for *number of vaccine antigens*, *number of vaccines*, and *vaccine aluminum exposure*, respectively. They were only slightly lower than the corresponding square roots of *VIF*s (i.e., 2.67, 2.50 and 1.47).

Coefficient estimates for redundant risk factors in the multiple logistic regression model with three measures were also provided in Table 3. As expected, the point estimates for redundant risk factors were close to zero for all scenarios with very large standard errors regardless of increasing association between the true risk factor and the outcome. The results suggested that high correlation among the three measures would not falsely detect the association between redundant risk factors and the outcome in multiple logistic regression models with three measures as covariates even when the association between the true risk factor and the outcome was strong.

Table 4 showed that empirical power for detecting the association between the true risk factor and the outcome decreased significantly when redundant risk factors were included in the multiple logistic regression model due to the overestimation of standard deviation. For example, when the true coefficient was 0.2 for the association between the *number of vaccine antigens* and the outcome, the model with the true risk factor only had an empirical power of 64.8%, but the empirical power dropped to 15.3% if the other two redundant risk factors were also included in the multiple logistic regression model. The coefficient for the true association had to more than double to achieve about 90% empirical power if the multiple logistic regression model included redundant risk factors when the true risk factor was either *number of vaccines* ($\beta_X = 0.63$) or *number of vaccine antigens* ($\beta_X = 0.65$). Again, the impact was less if the true risk factor was *vaccine aluminum exposure* due to its relatively low correlation with the other two measures.

Table 3
Impact of including redundant risk factors on coefficient estimates (standard error) in simulations where only one of three summary measures (true risk factor) is associated with the outcome.

True risk factor	True coefficient for the true risk factor	Estimated coefficients				Ratio of standard errors*
		Model only with the true risk factor	Model with the true risk factor and two redundant risk factors			
		Number of vaccine antigens	*Number of vaccine antigens*	*Number of vaccines*	*Vaccine aluminum exposure (mg)*	
Number of vaccine antigens	0.1	0.103 (0.090)	0.095 (0.221)	0.010 (0.216)	0.000 (0.120)	2.45
	0.2	0.202 (0.090)	0.199 (0.218)	0.004 (0.212)	0.000 (0.116)	2.42
	0.3	0.302 (0.087)	0.298 (0.211)	0.005 (0.206)	0.000 (0.112)	2.43
	0.65	0.650 (0.080)	0.644 (0.194)	0.008 (0.192)	0.003 (0.106)	2.43
		Number of vaccines	*Number of vaccines*	*Number of vaccine antigens*	*Vaccine aluminum exposure* (mg)	
Number of vaccines	0.1	0.105 (0.093)	0.108 (0.216)	−0.004 (0.223)	0.001 (0.119)	2.32
	0.2	0.203 (0.092)	0.203 (0.212)	−0.001 (0.216)	0.001 (0.116)	2.30
	0.3	0.303 (0.090)	0.305 (0.206)	−0.003 (0.210)	0.001 (0.112)	2.29
	0.63	0.631 (0.083)	0.637 (0.191)	−0.008 (0.193)	0.002 (0.104)	2.30
		Vaccine aluminum exposure (mg)	*Vaccine aluminum exposure* (mg)	*Number of vaccine antigens*	*Number of vaccines*	
Vaccine aluminum exposure (mg)	0.1	0.102 (0.087)	0.100 (0.119)	−0.004 (0.221)	0.007 (0.216)	1.37
	0.2	0.201 (0.085)	0.200 (0.115)	−0.002 (0.213)	0.005 (0.210)	1.35
	0.3	0.301 (0.083)	0.300 (0.111)	−0.002 (0.206)	0.004 (0.203)	1.34
	0.35	0.351 (0.082)	0.352 (0.109)	−0.004 (0.201)	0.005 (0.200)	1.33

* Ratio of standard errors: the ratio of the estimated standard error of $\widehat{\beta}_X$ from a regression model with the true and the redundant risk factors ($\widehat{\sigma^*}_{\widehat{\beta}_X}$) and the estimated standard error of $\widehat{\beta}_X$ from a regression model with only the true risk factors ($\widehat{\sigma}_{\widehat{\beta}_X}$).

Table 4
Impact of including redundant risk factors on empirical power in simulations where only one of three summary measures (true risk factor) is associated with the outcome.

True risk factor	True coefficient for the true risk factor (β_X)	Empirical power (%) for detecting the association between the outcome and the true risk factor	
		Model with the true risk factor only	Model with the true risk factor and two redundant risk factors
Number of vaccine antigens	0.1	22.0	6.1
	0.2	64.8	15.3
	0.3	94.4	29.8
	0.65	100.0	90.7
Number of vaccines	0.1	18.9	7.2
	0.2	60.9	16.1
	0.3	93.1	31.7
	0.63	100.0	90.8
Vaccine aluminum exposure (mg)	0.1	22.3	14.3
	0.2	67.1	40.5
	0.3	96.2	77.7
	0.35	99.7	89.6

In a vaccine safety study, it is also possible that all three summary measures have association with the outcome, for example, one has a strong association and the other two have a weak association with the outcome. Results in Table 5 show that the coefficients increased slightly, and the empirical power also increased, if only one risk factor was included in the model compared to the multiple logistic regression model with all three risk factors. In the multiple logistic regression model, if *number of vaccines* was a weak risk factor, this measure was less likely to be detected (lower empirical power) than the other weak risk factor in the multiple logistic regression model that included both strong and weak risk factors. For example, when *number of vaccine antigens* was the strong risk factor with a true coefficient of 0.2 and the other two summary measures had a true coefficient of 0.1, the model with *number of vaccine antigens* only with yielded a coefficient of 0.21 and an empirical power of 97.7%; the model with all three measures yielded a coefficient of 0.20 and an empirical power of 96.2% for *number of vaccine antigens*, a coefficient of 0.11 and an empirical power of 10.8% for *number of vaccines*, and a coefficient of 0.10 and an empirical power of 57.1% for *vaccine aluminum exposure*.

The impact of specifying an incorrect linear relationship between a true risk factor and the outcome when the true relationship is nonlinear is presented in Table 6. Since the form of the relationships are different across simulations, the coefficients from the models with the correct and incorrect specifications for the form of the relationships are not comparable between models. The empirical power for detecting the association between the true risk factor and the outcome significantly decreased in models specifying an incorrect relationship. In multiple logistic regression models with all three summary measures and specifying an incorrect linear relationship, not only the empirical power for detecting the true risk factors decreased significantly, it was also possible to detect the redundant risk factors as significant risk factors. In particular, *vaccine aluminum exposure* could be detected as a protective factor. For example, when the true risk factor was *number of vaccine antigens* with a true coefficient of 0.3, there was a 38.9% chance that *vaccine aluminum exposure* could be detected as a significant protective factor in the multiple logistic regression model with all three summary measures and all having incorrect relationships specified with the outcome. In addition, the empirical power to detect the true risk factor (*number of vaccine antigens*) was only 14.2%.

4. Discussion

Many factors can influence children's receipt of vaccines (thus the three summary measures) and outcomes. These factors included, but not limited to, health insurance, family income, parental care-seeking preferences, etc [7]. Directed acyclic graphs (DAG) has been used for causal analyses and was recommended

Table 5
Performance of multiple logistic regression models in simulations with both strong and the weak risk factors*

Strong risk factor	True coefficient	Estimated coefficients (standard deviation), empirical power (%)					
		Model with only one risk factor			Model with the strong risk factor and two weak risk factors		
		Number of vaccine antigens	*Number of vaccines*	*Vaccine aluminum exposure (mg)*	*Number of vaccine antigens*	*Number of vaccines*	*Vaccine aluminum exposure (mg)*
Number of vaccine antigens	0.2	0.21 (0.05), 97.7	0.12 (0.12), 12.9	0.11 (0.05), 67.5	0.20 (0.05), 96.2	0.11 (0.12), 10.8	0.10 (0.05), 57.1
	0.3	0.31 (0.05), 100	0.13 (0.12), 13.8	0.12 (0.05), 74.1	0.30 (0.05), 100	0.11 (0.12), 10.8	0.10 (0.05), 57.6
		Number of vaccines	*Number of vaccine antigens*	*vaccine aluminum exposure (mg)*	*Number of vaccines*	*Number of vaccine antigens*	*vaccine aluminum exposure (mg)*
Number of vaccines	0.2	0.21 (0.12), 39.9	0.11 (0.05), 56.8	0.11 (0.05), 67.0	0.21 (0.12), 36.7	0.10 (0.05), 48.0	0.10 (0.05), 59.8
	0.3	0.32 (0.12), 74.6	0.11 (0.05), 60.1	0.11 (0.05), 71.0	0.31 (0.12), 72.4	0.10 (0.05), 51.9	0.10 (0.04), 63.3
		vaccine aluminum exposure (mg)	*Number of vaccine antigens*	*Number of vaccines*	*vaccine aluminum exposure (mg)*	*Number of vaccine antigens*	*Number of vaccines*
vaccine aluminum exposure (mg)	0.2	0.21 (0.05), 99.6	0.12 (0.05), 61.3	0.12 (0.12), 12.9	0.20 (0.05), 99.4	0.10 (0.05), 46.3	0.11 (0.12), 10.8
	0.3	0.31 (0.04), 100	0.13 (0.05), 68.6	0.13 (0.12), 14.5	0.30 (0.04), 100	0.10 (0.05), 46.3	0.11 (0.12), 10.8

* True coefficient for the weak risk factors was 0.1.

Table 6
Impact of specifying an incorrect linear form of a relationship between summary measures of immunization schedule exposure and outcome on coefficient estimates (standard error) and empirical power when only one of three summary measures is a true risk factor.

True risk factor	*True coefficient for the true risk factor*	Estimated coefficients, empirical power (%)				
		Model including only the true risk factor with correct and incorrect relationships		Model with the true risk factor and two redundant risk factors all with incorrect relationship with the outcome		
		Correct relationship*	Incorrect relationship+			
		Number of vaccine antigens	*Number of vaccine antigens*	*Number of vaccine antigens*	*Number of vaccines*	*Vaccine aluminum exposure (mg)*
Number of vaccine antigens	0.1	0.1 (0.04), 67.4	0.07 (0.09), 12.2	0.06 (0.23), 5.6	0.06 (0.22), 6.1	−0.06 (0.12), 8.6
	0.2	0.2 (0.04), 100	0.14 (0.09), 34.5	0.13 (0.23), 8.1	0.13 (0.22), 9.2	−0.13 (0.12), 18.5
	0.3	0.30 (0.04), 100	0.22 (0.10), 67.8	0.20 (0.23), 14.2	0.19 (0.22), 13.6	−0.20 (0.12), 38.9
		Number of vaccines	*Number of vaccines*	*Number of vaccines*	*Number of vaccine antigens*	*vaccine aluminum exposure (mg)*
Number of vaccines	0.1	0.1 (0.03), 85.1	0.12 (0.10), 24.2	0.22 (0.22), 17.3	−0.06 (0.23), 6.1	−0.07 (0.12), 9.1
	0.2	0.2 (0.03), 100	0.25 (0.10), 75.3	0.44 (0.22), 50.1	−0.11 (0.22), 8.1	−0.14 (0.12), 19.8
	0.3	0.30 (0.03), 100	0.39 (0.10), 97.2	0.68 (0.22), 87.1	−0.16 (0.22), 12.6	−0.21 (0.12), 43.0
		vaccine aluminum exposure (mg)	*vaccine aluminum exposure (mg)*	*vaccine aluminum exposure (mg)*	*Number of vaccine antigens*	*Number of vaccines*
Vaccine aluminum exposure (mg)	0.1	0.10 (0.04), 64.0	0.05 (0.09), 9.2	0.08 (0.12), 9.7	−0.02 (0.23), 5.8	−0.01 (0.22), 5.2
	0.2	0.20 (0.04), 99.8	0.11 (0.09), 24.4	0.16 (0.12), 25.1	−0.03 (0.23), 5.8	−0.04 (0.22), 5.3
	0.3	0.30 (0.04), 100	0.17 (0.09), 49.7	0.25 (0.13), 53.4	−0.04 (0.23), 5.9	−0.07 (0.22), 5.8

* The relationship between the true risk factor and the logit of the probability of the outcome was nonlinear;
+ The relationship between the true risk factor and the logit of the probability of the outcome was linear.

for examining the causal pathway in vaccine schedule safety analyses. The current study addressed an important part of DAG analyses: examine the association between outcome and vaccine schedule after possible confounders were identified.

Studying the association between summary measures of pediatric immunization, such as cumulative *number of vaccine doses*, *number of vaccine antigens* and *vaccine aluminum exposure*, and clinical outcome of interests is challenging because these measures are highly correlated. The high correlation may make it difficult to identify the true risk factor if there is an association with only one of the three measures. Through simulation, we demonstrated that the association could be underestimated if a false risk factor was used as the sole risk factor, especially for *number of vaccine antigens* and *number of vaccines* because they had higher correlation with each other than with *vaccine aluminum exposure*. Thus, it is very likely that an association will be detected if a summary measure other than the true risk factor is used. We also demonstrated that point estimates for a true risk factor were unbiased when the other two measures (redundant risk factors) were included in multiple logistic regression models. However, the standard error of the point estimates of the true risk factor from multiple logistic regression models with the redundant risk factors increased dramatically compared to the one from regression models with only the true risk factor; consequently, empirical power decreased significantly, which may lead to false negative findings. Researchers should consider using the values of *VIF*s in sample size and power calculations if they plan to include all three measures in regression models. Methods for adjusting sample size and power calculation by a *VIF* can be found elsewhere [15].

We also examined two more complex scenarios. First, when all three risk factors were associated with the outcome (e.g., one strong risk factor and two weak risk factors), the multiple logistic regression model yielded unbiased coefficient estimates for all three risk factors, but *number of vaccines* was less likely to be

detected if it was one of the two weak risk factors. Second, specifying incorrect relationships between risk factors and the outcome not only decreased empirical power to detect true risk factors, but also increased the chance to detect *vaccine aluminum exposure* as a protective factor.

To examine the associations between these three summary measures and outcomes of interest, we recommend starting with three separate regression models (univariate analyses); each regression will be fully adjusted for other potential confounders to test the association of outcome with one of the three summary measures. The form of the relationship between a risk factor and outcome must be investigated thoroughly as our simulation results showed that specifying an incorrect relationship decreased empirical power to detect the true risk factor and raised the possibility of detecting redundant risk factors. If the null hypothesis cannot be rejected in any of the separate models, then conclusions can be drawn as to the lack of an association between the vaccine exposures (*number of vaccines, number of vaccine antigens,* and *vaccine aluminum exposure*) and outcome of interest. However, if there is evidence that one or more of the measures is associated with outcome of interest, then multiple logistic regression models be fit with the significant and marginally significant summary measures identified in univariate analyses.

The methods we present here can complement existing epidemiologic and statistical methods for causal analysis. For example, directed acyclic graphs (DAGs) should be used as a guide for examining potential causal pathways between vaccine schedule exposures and outcomes. In our modeling, we did consider other potential confounders; however, for actual vaccine schedule safety studies, a DAG may help elucidate other confounders that merit consideration. In addition, principal components and factor analyses are other methods that are often considered when dealing with correlated data. However, those methods typically deal with higher dimensional data and a larger number of correlated variables. For vaccine schedule safety research, our methods may be preferable since the number of summary measures is limited.

This study has limitations. First, the correlation among the three measures may differ in a different cohort than the cohort from the two HMOs. For a different cohort, the conclusion may differ slightly depending on the correlation strength. Second, since our sample size was fixed, the results may differ for different cohorts with different sample sizes, especially the statistical power. Third, we did not review methods to correctly specify the relationship between a risk factor and the outcome. Fourth, we did not consider unmeasured confounders in simulation. If a great proportion of the dependent variable variance is not explained by the independent variables, in general, the empirical power will decrease, and coefficient estimates may be biased. Fifth, we simulated a binary outcome and did not consider the time between vaccination summary measures (2nd birth day) and outcome. The time distance between vaccination measures and outcome is an important factor for evaluating biological plausibility of vaccination effect on outcomes of interest. The longer this time distance is, it may be less plausible that the outcome of interest is due to vaccination effect. In a real vaccine safety study, we recommend that vaccine safety researcher consider strategies such as sensitivity analyses to address the biological plausibility of vaccination effect on outcomes.

We conclude that (1) three summary measures of pediatric immunization – *number of vaccines, number of vaccine antigens, and vaccine aluminum exposure* - are highly correlated; (2) a multiple logistic regression model with the true risk factor and two redundant risk factors significantly reduced the empirical power for detecting the association between the true risk factor and the outcome, and estimated *VIF*s can help researchers to estimate the loss of statistical power and plan studies accordingly; (3) if one measure has a strong and the other two have a weak association with the outcome, a multiple logistic regression model with the three summary measures will be appropriate for detecting the strong risk factor; (4) correctly specifying the form of the relationship between risk factors and the outcome is crucial.

Conflict of interest

The authors declared that there is no conflict of interest.

Acknowledgement

This study was funded by a grant from the National Institutes of Health, National Institute of Allergy and Infectious Diseases (NIH R01AI107721, ‘‘Methods for Safety Evaluation of Vaccine Schedules”, Principal Investigator: Martin Kulldorff, PhD). Xu was also Supported by NIH/NCATS Colorado CTSA Grant Number UL1 TR002535.

References

[1] Robinson CL. Advisory committee on immunization practices recommended immunization schedules for persons aged 0 through 18 years — United States, 2016. MMWR Morb Mortal Wkly Rep 2016;65(4):86–7.
[2] Institute of Medicine. Committee on Assessment of Studies of Health Outcomes Related to the Recommended Childhood Immunization Schedule and Board on Population Health and Public Health Practice, Childhood immunization schedule and safety: stakeholder concerns, scientific evidence, and future studies; 2013. Available from <http://www.iom.edu/Reports/2013/The-Childhood-Immunization-Schedule-and-Safety.aspx> [accessed 30.09.14].
[3] Baggs J, Gee J, Lewis E, Fowler G, Benson P, Lieu T, et al. The Vaccine Safety Datalink: a model for monitoring immunization safety. Pediatrics 2011;127 (Supp 1):S45–53.
[4] McNeil MM, Gee J, Weintraub ES, Belongia EA, Lee GM, Glanz JM, et al. The Vaccine Safety Datalink: successes and challenges monitoring vaccine safety. Vaccine 2014;32(42):5390–8.
[5] Luman ET, Barker LE, Shaw KM, McCauley MM, Buehler JW, Pickering LK. Timeliness of childhood vaccinations in the United States: days undervaccinated and number of vaccines delayed. JAMA 2005;293 (10):1204–11.
[6] Glanz JM, Newcomer SR, Narwaney KJ, Hambidge SJ, Daley MF, Wagner NM, et al. A population-based cohort study of undervaccination in 8 managed care organizations across the United States. JAMA Pediatrics 2013;167(3):274–81. CCCID: 23338829.
[7] Glanz JM, Newcomer SR, Jackson ML, Omer SB, Bednarczyk RA, Shoup JA, DeStefano F, Daley MF. Contributors; Subject Matter Experts; Centers for Disease Control and Prevention. White Paper on studying the safety of the childhood immunization schedule in the Vaccine Safety Datalink. Vaccine 2016;34(Suppl 1):A1–A29. https://doi.org/10.1016/j.vaccine.2015.10.082.
[8] Kulldorff M. Study designs for the safety evaluation of different childhood immunization schedules. Washington, D.C.: Institute of Medicine of the National Academies; 2012.
[9] Farrar DE, Glauber RR. Multicollinearity in regression analysis: The Problem Revisited. Rev Econ Stat 1967;49(1):92–107.
[10] Kumar TK. Multicollinearity in regression analysis. Rev Econ Stat 1975;57 (3):365–6.
[11] Simon TD, Cawthon ML, Stanford S, et al. Center of Excellence on Quality of Care Measures for Children With Complex Needs (COE4CCN) Medical Complexity Working Group. Pediatric Medical Complexity Algorithm: a new method to stratify children by medical complexity. Pediatric 2014;133(6): e1647–54. https://doi.org/10.1542/peds.2013-3875.
[12] Glanz JM, Newcomer SR, Daley MF, et al. Association between estimated cumulative vaccine antigen exposure through the first 23 months of life and non–vaccine-targeted infections from 24 through 47 months of age. JAMA Pediatr 2018;319:906–13. https://doi.org/10.1001/jama.2018.0708.
[13] Glanz JM et al. Cumulative and episodic vaccine aluminum exposure in a population-based cohort of young children. Vaccine 2015;33:6736–44.
[14] Kutner MH, Nachtsheim CJ, Neter J. Applied linear regression models. 4th ed. McGraw-Hill Irwin; 2004.
[15] Hsieh FY, Bloch DA, Larsen MD. A simple method of sample size calculation for linear and logistic regression. Stat Med 1998;17:1623–34.