

## Generation of Prediction Intervals to Assess Data Quality in the Distribute System Using Quantile Regression

Ian Painter\*   Julie Eaton†   Don Olson‡   Debra Revere§   Bill Lober¶

### Abstract

The Distribute system is a community-based aggregate data national emergency department syndromic surveillance project for influenza-like illness (ILI) that integrates data from existing state and local public health jurisdiction surveillance systems. Data received by Distribute arrives piecemeal from each jurisdiction, data for a particular encounter date accrues over several days before the data becomes complete. We investigated the use of quantile regression to assess the accuracy of ILI ratios calculated using the incomplete, partially accrued data, using models including the number of days lag in the data and indicators for gross changes in data quality. These models were used to calculate 90% prediction intervals for the error in any particular measurement based on partially accrued data, and superimposed on time series graphs of the ILI ratio. The width of the prediction intervals decreased rapidly with increasing number of days since the event date, the median width of the intervals reduced to less than 10% of the complete data ILI ratio after a period of 5 days.

**Key Words:** Syndromic surveillance, Data Quality, Quantile regression

### 1. Introduction

Distribute is a community-based national emergency department syndromic surveillance project for influenza-like illness (ILI) that integrates data from existing state and local public health department surveillance systems. The Distribute project provides both comparisons of ILI-related clinical visits across public health jurisdictions and a national picture of ILI. The Distribute project was started in 2006 by the International Society for Disease Surveillance as a simple experiment to allow public health jurisdictions to share ILI surveillance data. Support for Distribute from the Markle Foundation and the United States Center for Disease Control and Prevention (CDC) has enabled the Distribute system to grow to currently include 42 jurisdictions, covering over 50% of the US population [1].

Unlike other surveillance systems, which typically depend on individual level patient data, Distribute is designed to work solely with aggregate count data, collected from existing surveillance systems at the participating public health jurisdictions. These data consist of daily counts of total emergency department (ED) visits and ED visits falling into two ILI syndromes (a broad and a narrow syndrome) within each jurisdiction, stratified by age group and geographic region (three digit patient or facility zip codes). Each jurisdiction uploads these data to the Distribute system via `sftp` or `https` post, ideally daily but in practice with varying frequency (see Table 1.) Upload processes at each jurisdiction may be automatic or manual. Within each jurisdiction the frequency at which data are received from ED facilities and the delay in receiving data from the facilities typically varies between facilities. As a result, total visit and syndrome counts for a particular encounter date are typically acquired by each jurisdiction piecemeal, accruing over several days. To account

\*University of Washington, Box 359442, Seattle WA 98195

†University of Washington Tacoma, Box 358436, Tacoma, WA 98402

‡International Society for Disease Surveillance, 26 Lincoln Street #3, Brighton, MA 02135

§University of Washington, Box 357266, Seattle WA 98195

¶University of Washington, Box 357266, Seattle WA 98195

**Table 1:** Data completion summary statistics. The mean and standard deviation for the within data feed mean times to first data upload, and to 50%, 80% and 95% data completion.

	Mean	sd
Mean number of days to first data upload	1.89	1.43
Mean number of days to 50% data completion	2.14	1.45
Mean number of days to 80% data completion	2.33	1.56
Mean number of days to 95% data completion	2.81	1.74

for this accrual, the data uploaded to the Distribute system each day contain counts for multiple encounter dates, typically the previous 21 days. Until all of the data are received from a jurisdiction for a particular encounter date, the data are considered incomplete for that date. The time it takes for all of the data for an encounter date to be received from a jurisdiction varies between jurisdictions, as seen in Table 1. For all jurisdictions the data are considered complete after 21 days (as this is longer than the observed maximum time lag to receive complete data). For any particular encounter date it is not possible to know if the data are complete until 21 days after the date.

For comparison purposes a single ILI syndrome is used, however each jurisdiction selects which of the two syndromes is most useful for ILI monitoring for that jurisdiction. This syndrome is termed the preferred ILI syndrome. The Distribute system makes the preferred syndrome data available to the participating jurisdictions via a publicly viewable website (which shows the proportion of visits within each jurisdiction that fall into the preferred syndrome, aggregated by week) and a restricted website available only to participating jurisdictions (which also shows daily syndrome and total visit counts, and allows data to be stratified by age group).

The primary indicator of interest produced by the Distribute system is the proportion of ED visits that fall into the preferred ILI syndrome. As the data used to calculate this indicator accrue over several days, the incomplete data indicator (the indicator calculated from the data currently at hand) may not be accurate, that is to say, it may differ from the final data indicator value. To be of use for near real-time surveillance purposes, it is necessary to understand the accuracy of the incomplete data indicator. The focus of this research is on using prediction intervals generated by quantile regression to assess the accuracy of the incomplete data indicator.

## 2. Methods

### 2.1 Data

The data used in this research consist of, for each data upload to Distribute, the observed counts of the preferred ILI visits and total ILI visits for each encounter date contained in that upload, for each data feed received by Distribute. Note that some jurisdictions send more than one data feed; a total of 52 data feeds are received by the Distribute system from the 42 jurisdictions. Due to the great variability between data feeds we model the data separately for each feed.

Let  $X_{ij}$  and  $W_{ij}$  be the cumulative ILI visit count and total visit count for encounter date  $i$  received by  $j$  days after the encounter date. We assume that  $W_{ij} > 0$  for  $j > 0$ . In practice this may not be the case, therefore prediction intervals are calculated only after at

least some data have been received for a particular encounter date. Let

$$Y_{ij} = X_{ij}/W_{ij} \tag{1}$$

be the incomplete ILI ratio calculated for encounter date  $i$  using the cumulative counts received by  $j$  days after the encounter date and let  $Y_i$  be the complete data ILI ratio, that is, the ILI ratio calculated when all data have been received. Let

$$E_{ij} = \log(Y_{ij}) - \log(Y_i) \tag{2}$$

be the log relative error for the ILI ratio calculated from incomplete data. Inspection of the log relative errors over sites showed a monotonic relationship between the lag ( $j$ ) and the variance of the errors (larger variance for smaller lags). The observed quantiles showed the same monotonic relationship with lag. Because ILI ratios change substantially over the course of a year and because the variance of the counts increases with the mean count, the log relative error was used rather than the absolute error.

Figure 1 shows two representative jittered scatter plots of the log relative error in the incomplete ILI ratio as a function of lag for two jurisdictions. These plots present two interesting features: possible asymmetry in the distribution of errors and bias in the median error.

Large errors in the prediction intervals periodically occur, typically when a facility fails to report data. This results in a drop in the total counts from that jurisdiction, and potentially a large change in the observed ILI ratio if the baseline ILI ratio for the dropped facility is different from the mean ratio over the remaining facilities. Let  $U_{ij}$  be the standardized difference between the observed total count and a running mean of total counts, so that

$$U_{ij} = \frac{W_{ij} - \overline{W_{ij}}}{\sigma_{W_{ij}}} \tag{3}$$

where  $\overline{W_{ij}} = \sum_{i-n \leq d < i} W_{dj}/n$  is the mean of the total counts at the same lag for the previous  $n$  days and  $\sigma_{W_{ij}}$  is the standard deviation of these same values. Similarly, define

$$V_{ij} = \frac{Y_{ij} - \overline{Y_{ij}}}{\sigma_{Y_{ij}}} \tag{4}$$

as the standardized difference between the incomplete ILI ratio and a running mean of the incomplete ILI ratio. We define two indicator variables in terms of these standardized differences:

$$I_{ij} = \begin{cases} 1 & \text{if } -U_{ij} > c_u \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

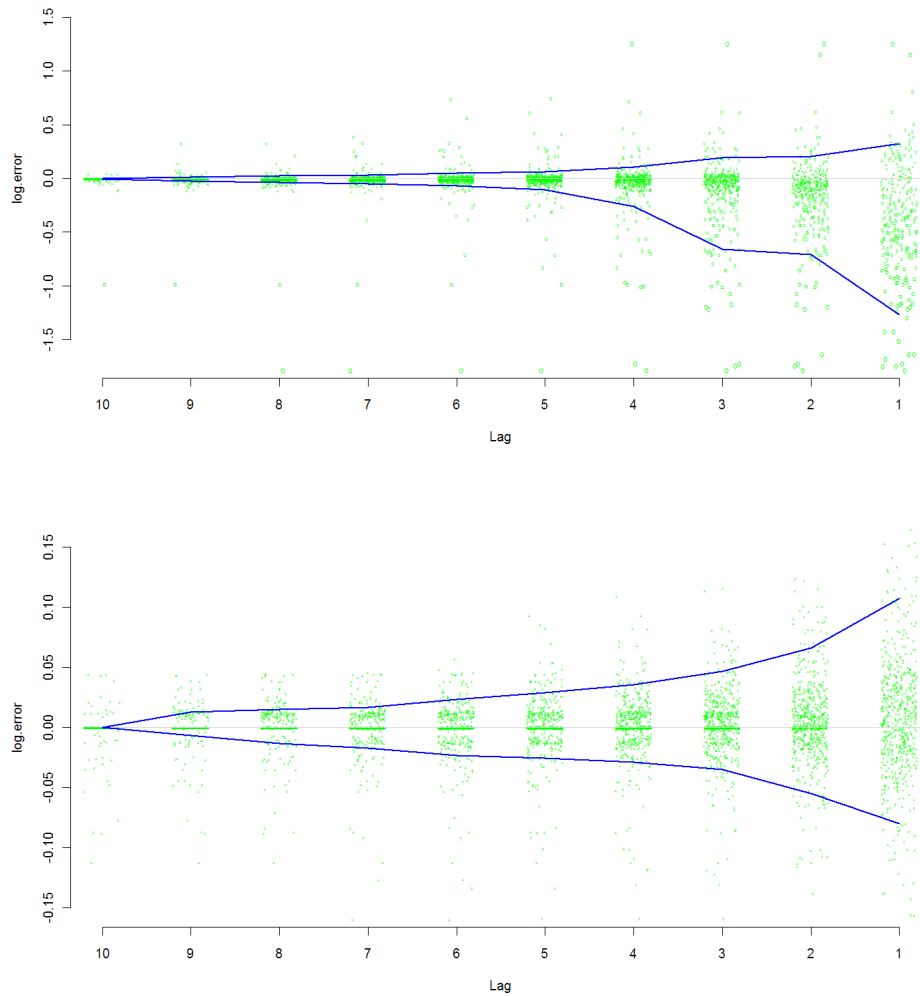
and

$$J_{ij} = \begin{cases} 1 & \text{if } |V_{ij}| > c_v \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

where  $c_u$  and  $c_v$  are cutoff values discussed in the next section. We also define a three-state indicator variable as follows:

$$K_{ij} = \begin{cases} 1 & \text{if } I_{ij} = 1 \text{ and } J_{ij} = 0 \\ 2 & \text{if } I_{ij} = 1 \text{ and } J_{ij} = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

The indicator defined in equation 7 captures the scenarios in which there is either an unexpected change in the total counts but no unexpected change in the ILI ratio indicator or an unexpected change in both the total counts and the ILI indicator. This indicator does not, however, include the scenario in which there is an unexpected change in the ILI ratio indicator but not in the total counts. This latter scenario is what we would expect if there were a real but unexpected change in the indicator.



**Figure 1:** Jittered scatter plots of the log relative error in the incomplete ILI ratio as a function of lag for two representative jurisdictions. The solid lines are the 0.10 and 0.90 quantiles.

## 2.2 Modeling Error Quantiles

To borrow strength across lag times and to allow for the inclusion of covariates, quantile regression models ([2]) were considered. First, a cubic relationship between each quantile and the lag was used to model the relationship between lag and error quantiles, called model 1:

$$Q_{E_{ij}}(\tau) = \alpha(\tau) + \beta_1(\tau)j + \beta_2(\tau)j^2 + \beta_3(\tau)j^3 \tag{8}$$

The indicator function based on the standardized difference between the observed and expected total counts ( $I$ , defined in equation 5) was considered as a covariate that incorporates an estimate for unusually incomplete data. The indicator function based on the standardized difference between observed and expected indicator values ( $J$ , defined in equation 6) was considered as a covariate that indicates sudden changes in the expected indicator value indicative of potential ‘spiky’ errors. A model incorporating each of these indicators was

considered in the second and third models:

$$Q_{E_{ij}}(\tau) = \alpha(\tau) + \beta_1(\tau)j + \beta_2(\tau)j^2 + \beta_3(\tau)j^3 + \beta_I(\tau)I_{ij} \quad (9)$$

and

$$Q_{E_{ij}}(\tau) = \alpha(\tau) + \beta_1(\tau)j + \beta_2(\tau)j^2 + \beta_3(\tau)j^3 + \beta_J(\tau)J_{ij} \quad (10)$$

Finally, a fourth model including the three-state covariate  $K$  (defined in equation 7) was considered:

$$Q_{E_{ij}}(\tau) = \alpha(\tau) + \beta_1(\tau)j + \beta_2(\tau)j^2 + \beta_3(\tau)j^3 + \beta_K(\tau)K_{ij} \quad (11)$$

Separate models were fit for the 0.05 and 0.95 quantiles for each jurisdiction using the function (`nlsrq`) in the package (`quantreg`)[3] using R 2.12. [4]. Nested analysis of variance was used to assess whether models with indicator covariates had better model fit.

### 3. Results

Examination of lagged time series of the standardized deviations  $U$  and  $V$  suggested a cutoff value of 3.0 for both  $c_u$  and  $c_v$  in equations 5 and 6. Results did not vary substantially for small changes in these cutoff values, and appeared worse for larger values. Separate models were fit for the 0.05 and 0.95 quantiles.

Estimates for the coefficients in the four models converged for 37 of the 52 data feeds received by the Distribute system. Data feeds for which estimates failed to converge corresponded to jurisdictions with insufficient data or with haphazard upload patterns. These jurisdictions were not considered further in this analysis. Model 1 was statistically significant for all 37 of the jurisdictions. Analysis of variance comparisons of each of models 2, 3 and 4 with model 1 were statistically significant at the 0.05 level for 21 data feeds for model 2, 13 data feeds for model 3 and 24 data feeds for model 4. For 11 of the data feeds all three of the models were not statistically significant when compared to model 1. Of the 26 data feeds which showed statistical significance, 24 were significant for model 4.

We used the estimated coefficients from model 4 to generate 90% prediction intervals for the log relative error in the incomplete data indicator for a particular encounter date using the data collected a set number of days lag after the encounter date by estimating the 0.05 and 0.95 quantiles based on the observed values for lag ( $j$ ) and the three state indicator variable  $K$  for the particular lag and encounter date. These intervals were then converted into prediction intervals for the complete data indicator value and superimposed on time series of indicator values. Figure 2 shows two such time series for two representative data feeds. Table 2 summarizes the predicted accuracy for the 37 data feeds as a function of lag, when no unexpected deviations are observed (i.e.  $K = 0$ ).

Figure 3 illustrates the effects of unexpected deviations on the prediction intervals, keeping the lag time constant. In the illustrated time series an unexpected jump in the observed ratio is observed on 2011-02-14, and a corresponding increase in the width of the prediction interval occurs.

### 4. Discussion

Distribute uses a novel method of data collection to provide valuable surveillance information based on the aggregate data that health agencies and hospitals feel comfortable sharing, and may indeed report publicly. This offers advantages in developing widespread voluntary surveillance as the privacy and confidentiality concerns, as well as business process concerns, associated with patient-level data are reduced. At the same time, from a DQ perspective, Distribute faces a number of challenges:

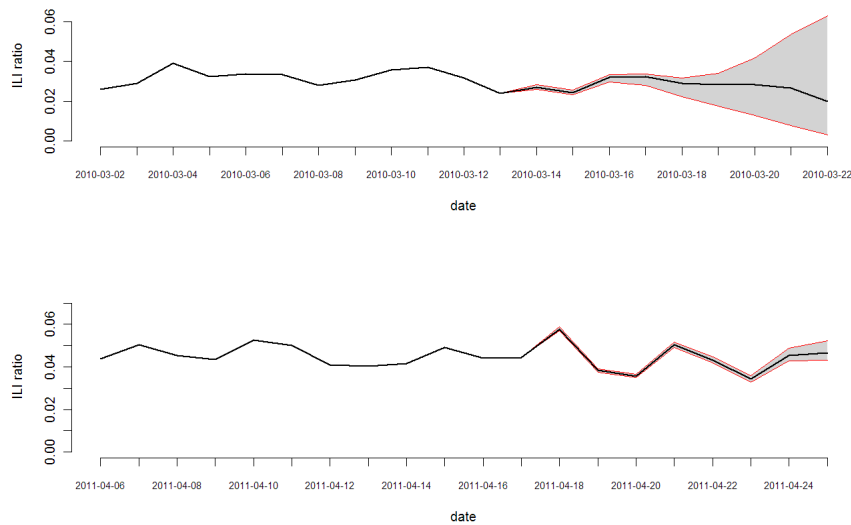
**Table 2:** Mean width of prediction intervals for ILI ratio for lags of 0 to 9; in absolute terms, relative to the mean ILI proportion and relative to the sd of the ILI proportions

Lag in days	0	1	2	3	4	5	6	7	8	9
Mean width of 90% prediction intervals for ILI proportion										
mean	0.155	0.030	0.018	0.012	0.009	0.007	0.006	0.007	0.004	0.003
lower quartile	0.020	0.002	0.002	0.001	0.001	0.001	0.001	0.000	0.000	0.000
median	0.045	0.014	0.008	0.005	0.003	0.002	0.002	0.002	0.001	0.001
upper quartile	0.139	0.031	0.023	0.015	0.009	0.007	0.007	0.006	0.005	0.003
Mean width of 90% prediction intervals relative to mean ILI proportion										
mean	1.454	0.638	0.399	0.255	0.179	0.135	0.134	0.119	0.080	0.072
lower quartile	0.827	0.123	0.077	0.062	0.030	0.023	0.023	0.019	0.016	0.010
median	1.096	0.479	0.264	0.199	0.117	0.071	0.073	0.062	0.055	0.025
upper quartile	1.753	1.100	0.559	0.342	0.220	0.157	0.162	0.140	0.121	0.071
Mean width of 90% prediction intervals relative to standard deviation of ILI proportions										
mean	2.265	1.061	0.671	0.442	0.324	0.254	0.248	0.212	0.150	0.135
lower quartile	1.138	0.266	0.175	0.109	0.068	0.048	0.036	0.040	0.028	0.020
median	1.820	0.708	0.579	0.362	0.215	0.167	0.154	0.135	0.113	0.067
upper quartile	2.744	1.613	1.020	0.637	0.423	0.377	0.328	0.275	0.240	0.145

- It receives aggregate data only, and there is no direct access to the raw data that aggregates are derived from (due to data sharing restrictions necessitated to allow sharing of any data).
- The data that it does receive is usually aggregated by the data source from multiple sub sources (facilities).
- There is wide heterogeneity in the timeliness of the data, in both terms of the timeliness of data sources sending data to distribute, and in terms of the timeliness of the data that the data source receives from its sub-sources.
- There is limited meta-data available, in particular meta-data about the components (facilities) than the data feeds are composed from. This is also a result of restrictions of what data jurisdictions are willing to share.
- Timeliness of the Distribute system is an important factor in the usefulness of the system; waiting for the data to be complete degrades the usefulness of the system.

The data quality challenges faced by Distribute are documented in few other fields that we are aware of. Fields with similar problems that we know about include network traffic monitoring [5] and vaccine safety adverse event reporting [6].

The volume of data received in network monitoring is several orders of magnitude higher than is the case for Distribute. This enables automated machine learning approaches to be used that would be difficult to use in Distribute due to insufficient training data. Data quality issues in vaccine safety adverse event reporting have many of the same characteristics as the in the Distribute system. In [7] rigorous methods for directly testing for changes in adverse event rates in the presence of partially accrued data are developed using sequential probability ratio tests.

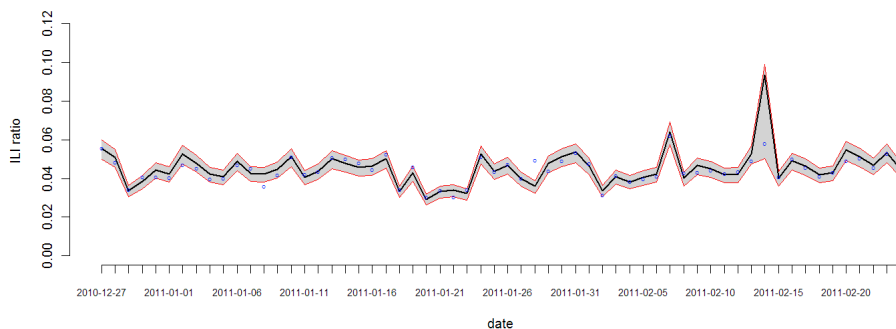


**Figure 2:** Time series of observed ILI proportions with 95% prediction intervals imposed for two representative data feeds.

The emphasis of the research we present is on estimating the accuracy of a measurement (in this case the ILI ratio) based on partially accrued data. Note that this differs from the problem of estimating what the complete data measurement will be at a future date. Methods focused on predicting future values would normally include estimates of trends in the complete data indicator, here we consciously want to avoid including trend information in our model. For this reason we directly model the error in the incomplete data indicator rather than model the complete data indicator as the outcome variable, and we use models which fix the lag time across covariates.

Overall our fourth model provided the best fit for the majority of the data feeds. This model corresponds to separate estimates of the error quantiles for three scenarios based on the unexpected change indicators. The first scenario is when there is an unexpected drop in the total number of ED visits (when compared to the previous encounter dates but keeping the lag fixed) but no unexpected change in the indicator. This could correspond to the situation where a data quality problem, as manifested by a drop of volume, does not result in a misleading value for the incomplete data indicator. However it could also correspond to a situation where there is an actual change in the indicator value (for example, at the outbreak of an epidemic) but this change is masked by a data quality problem. The second scenario is when there is an unexpected drop in the total number of ED visits and an unexpected change in the indicator. In this situation the incomplete data indicator value should be considered unreliable, and we would like our prediction intervals to represent this. However if the indicator value changes unexpectedly but there is no evidence of a data quality problem (that is to say, no corresponding drop in total ED visits), an increase in the prediction interval has the risk of downplaying the significance of a possibly important change in the indicator. For this reason we do not include as a separate value the situation where the indicator value changes unexpectedly but the total visit count does not.

The models considered in this paper deal with short-term changes in data quality. However in some situations a jurisdiction loses or adds a facility to their data feed, resulting in a long-term change in the characteristics of the data feed. In this situation, better estimates



**Figure 3:** Time series of observed ILI proportions with a fixed lag of 2 days with superimposed 95% prediction intervals. The points show the complete data ILI proportion for each day.

of the incomplete data error might be obtained by using methods that incorporate these long-term changes in the data characteristics. These changes are manifested typically as step-like changes in the total visit counts, and a logical approach for incorporating these changes would be to include indicators based on change point methods (for example, using only data subsequent to the most recent change point in the estimating procedure) or using state change models.

## 5. Acknowledgments

We would like to thank the Markle Foundation and the Centers for Disease Control for providing funding for the Distribute project.

## REFERENCES

1. Olson DR, Paladini M, Lober WB, Buckeridge D, ISDS Distribute Working Group (2011), “Applying a New Model for Sharing Population Health Data to National Syndromic Surveillance for Influenza: The DiS-TRIBuTE Project Proof of Concept, 2006 to 2009,” *PLoS Currents: Influenza* 2011, Aug 2;3: RRN1251. <http://www.ncbi.nlm.nih.gov/pubmed/21894257>
2. Koenker, Roger (2005) “Quantile Regression”, Cambridge University Press. ISBN 0-521-60827-9
3. Koenker, Roger (2009). “quantreg: Quantile Regression. R package version 4.71.” <http://CRAN.R-project.org/package=quantreg>
4. R Development Core Team (2003): “R: A language and environment for statistical computing”, <http://www.R-project.org>.
5. Flip Korn, S. Muthukrishnan, Yunyue Zhu: “Checks and Balances: Monitoring Data Quality Problems in Network Traffic Databases”. *VLDB 2003, Proceedings of 29th International Conference on Very Large Data Bases, September 9-12, 2003, Berlin, Germany*. Morgan Kaufmann 2003, ISBN 0-12-722442-4536-547 pp 536-547
6. Greene SK, Kulldorff M, Lewis EM, Li R, Yin R, Weintraub ES, Fireman BH, Lieu TA, Nordin JD, Glanz JM, Baxter R, Jacobsen SJ, Broder KR, Lee GM. (2010) “Near real-time surveillance for influenza vaccine safety: proof-of-concept in the Vaccine Safety Datalink Project.” *American Journal of Epidemiology*. 2010 Jan 15;171(2):177-88. Epub 2009 Dec 4. PubMed PMID: 19965887; PubMed Central PMCID: PMC2878099.
7. Greene SK, Kulldorff M, Yin R, Yih WK, Lieu TA, Weintraub ES, Lee GM. (2011) “Near real-time vaccine safety surveillance with partially accrued data.” *Pharmacoepidemiology and Drug Safety*. 2011 Jun;20(6):583-90. doi: 10.1002/pds.2133