State-level transport flows are predictive of the dynamics of porcine epidemic diarrhea virus

Eamon B. O'Dea[1*], Harry Snelson[2], Shweta Bansal[1,3]
1 Department of Biology, Georgetown University, Washington, District of Columbia, United States
2 American Association of Swine Veterinarians, Perry, Iowa, United States
3 Fogarty International Center, National Institutes of Health, Bethesda, Maryland, United States
*** E-mail: odea35@gmail.com**

Abstract

More than a year after the emergence and rapid spread of porcine epidemic diarrhea virus (PEDV) in the U.S. swine herd, the extent to which the virus has spread through pathways associated with the transportation of swine remains unclear. We analyze counts of state-level, laboratory-confirmed infections to better discern the pathways by which the virus has propagated. In particular, we aim to establish and quantify any large-scale association of swine movements with the spread of PEDV. To that end, we find that the similarity of the dynamics of cases in a pair of states increases with transport flows. We find with stability selection that balance sheet variables and the number of farms in a state are likely be relevant predictors of PEDV burdens. Fitting a time series susceptible–infected–recovered model by maximum likelihood, we reject the hypothesis that flows have no effect on the transmission rate. We show with simulation how our state-level analyses may be affected by farm-level variation in risk relations. Overall, the results are consistent with the common belief that transmission is associated with swine movement and provide quantification of the strength of association.

Author Summary

In the spring of 2013, a virus to which the U.S. swine population was highly susceptible emerged and began causing high-mortality outbreaks. The virus is highly contagious but exactly how often it spreads by any particular route between farms remains unknown. We find the similarity in the dynamics of infected farms between states to increase with the flow of live swine between them, which supports the hypothesis of transportation-associated spread. Looking at cumulative burdens, we find that swine inventories are more likely to be relevant predictors than climatic variables or the number of farms in a state. The association of inventories with flow volumes may explain why. Consistently, the time series data support the inclusion of flow volumes in models of disease spread. Thus overall we find evidence for transportation-associated spread, which supports the emphasis placed on transportation biosecurity in the control of this and similar pathogens. We also submit that more conclusive findings could be obtained from the next such event with incremental improvements in surveillance.

Introduction

The 2013 emergence of porcine epidemic diarrhea virus (PEDV) [1] in the United States has provided an example of both the economic hardships livestock diseases can cause and our limited understanding of how such diseases spread. This virus acutely infects the intestine and causes severe diarrhea and vomiting [2]. The first confirmed U.S. outbreak occurred in April [3], and in less than a year PEDV outbreaks were confirmed in 27 states [4], states that together produce 95 percent of the U.S. pig crop [5]. Farms experiencing outbreaks have suffered 90 percent and higher losses of unweaned pigs [3]. The time it takes for a farm to return to stable production is highly variable but on the order of weeks, leading to great expenses in infection control costs and production losses alike.

Losses are also apparent on a national economic scale. Producers had for the previous 8 years been making steady increases in the average litter size of about 0.16 head per year [6]. By November 2013, the average litter size had begun an abnormal downturn [6], dropping 0.66 head by March 2014 [7]. The virus also affected swine production in other parts of America and Asia, as reviewed by Refs. [8] and [9].

With no vaccines of proven efficacy yet on the market, producers wishing to protect their herds from PEDV must find effective practices to prevent exposure of their herds, making research into the mechanisms by which PEDV is spreading a priority. Transportation-associated transmission of PEDV has been supported by the observation at harvest facilities that it spreads among trailers used to transport swine [10], and some experts believe that current resources of livestock trailers, trailer-washing facilities, and transport personnel are insufficient to allow for a standard 3-hour trailer cleaning between every load [11]. With such concerns in mind, some states have responded to PEDV by requiring that imported swine are from PEDV-free premises. Nevertheless, in many of the outbreaks occurring on U.S. farms, the source of infection remains open to speculation; no animals were introduced to the farm in the weeks leading up to the outbreak and no personnel, feed, or equipment provide a clear link to another outbreak [3, e.g.]. Research is in progress to evaluate the importance of transportation-independent mechanisms of spread such as airborne particles [12] and contaminated feed [13–16].

Most of the research on PEDV involves detailed investigations on a small scale. For example, there have been epidemiological investigations of infected farms in North Carolina and a cluster of infected farms in Oklahoma and adjacent states [17]. Such work is effective for determining the biological plausibility of different routes, but the risk-factors identified in a small-scale study may be specific to the small area of the study. Modeling studies based on large-scale surveillance data, such that of bovine tuberculosis in Ref. [18], can thus be a valuable complement to other work by providing evidence that a transmission route is consistently used across a large population.

Here we analyze cases of PEDV-infected farms to evaluate at the national level evidence for transportation-based pathways of spread. We consider whether states that exchange large numbers of swine have similar PEDV dynamics, what predictors have the most robust association with PEDV burdens, and whether incorporating flows improves the fit of a mechanistic model of PEDV spread.

Results

A few preliminary facts pertain to all our results. First, all of the contiguous 48 states share some portion of the nation's swine but the Midwest and North Carolina are areas of major concentration (Fig. 1A), holding some 88 percent of the inventory [5].

Second, PEDV data are available at the state level in the form of weekly counts of

the number of diagnostic case submissions that tested positive for PEDV. These counts, reported as positive accessions, appear to reflect swine inventories fairly well (Figs. 1A and 1B), but in fact are likely to be informative of the number of infective farms because each infected farm will submit a limited number of samples for testing. We refer to positive accessions as cases and work with the assumption that they are correlated with the number of PEDV-positive farms. We analyze cases from the year 2013, which has the advantage that farms protected by their immunity rather than a lack of exposure were most likely rare during this initial period of spread. As of June 2014, data on the number of infected farms is available. In support of our assumption that counts of positive accessions and positive farms are correlated in 2013, we find that they have a Spearman rank correlation of 0.74 with data from June 2014 to February 2015 [19].

Third, as a proxy variable for all pathways of spread involving shipment of live swine, we use estimates of swine transport flows (i.e., the total number of swine moved between pairs of states each year for purposes other than slaughter). These flows vary greatly in size but generally the larger ones move swine into the Midwest (Fig. 1C). While they may seem to be a crude proxy, previous phylogeographic analysis [20] has found evidence that they were associated with movement of H1 influenza A virus among swine. A detailed description of the data we have analyzed appears in S1 Text.

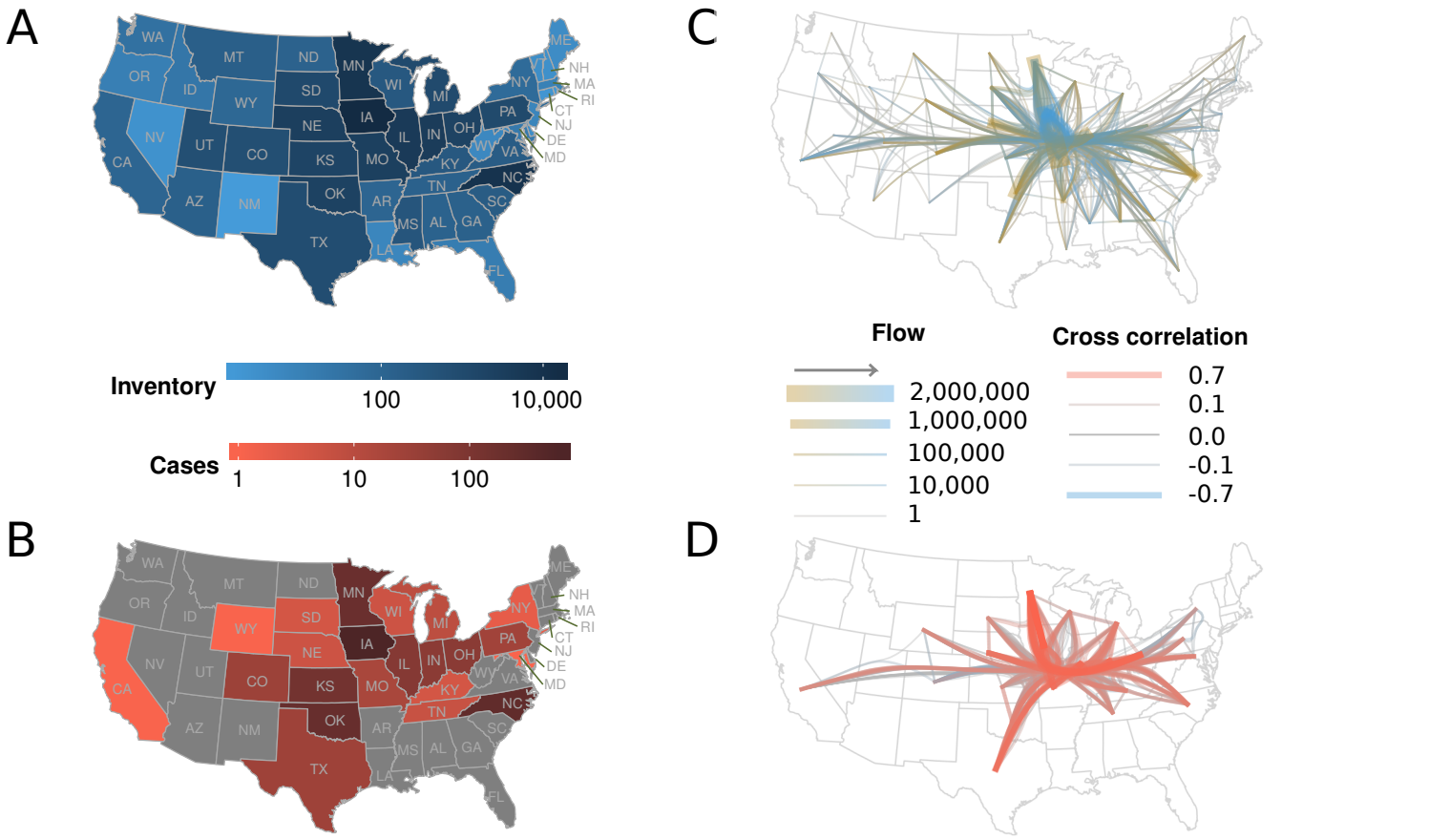

Figure 1. Spatial structure in the PEDV outbreak reflects that of swine production. (A) State-level swine inventory estimates in thousands of head. (B) Cumulative cases in each state for 2013. (C) Network of estimated annual interstate flows of head of swine. The arrow in the key indicates the direction of flow along the color gradient. (D) Network of cross correlation in weekly cases between states reporting cases in 2013. In both (C) and (D), edges with similar origins and destinations are bundled together to make spatial patterns in relationships visible.

Flows and cross correlations of cases

To measure the coupling of disease dynamics, we computed the cross correlation with a lag of 1 week between all pairs of states reporting cases (Fig. 1D). A simple metapopulation model predicts that cross correlations will increase with increased contact rates between populations [21]. Consistent with this prediction, we found that cross correlations were correlated with the logarithm of transport flows. This relation held whether flows and cross correlations were treated as directional (Fig. 2A), were averaged over both directions (S1 Figure), or were ranked (S2 Figure and S3 Figure). This correlation seemed to be driven in part by concentration of both high cross correlations (Fig. 1D) and large flows (Fig. 1C) in Midwestern states. The cross correlations of these states results from the presence of a small wave of cases early in the outbreak and a much larger wave toward the end of our observations (Fig. 2B, left column). To describe a second point of support, Kansas and Oklahoma share a distinctive period of high cases in the middle of the time series and fairly large flows (S4 Figure and Fig. 2B).

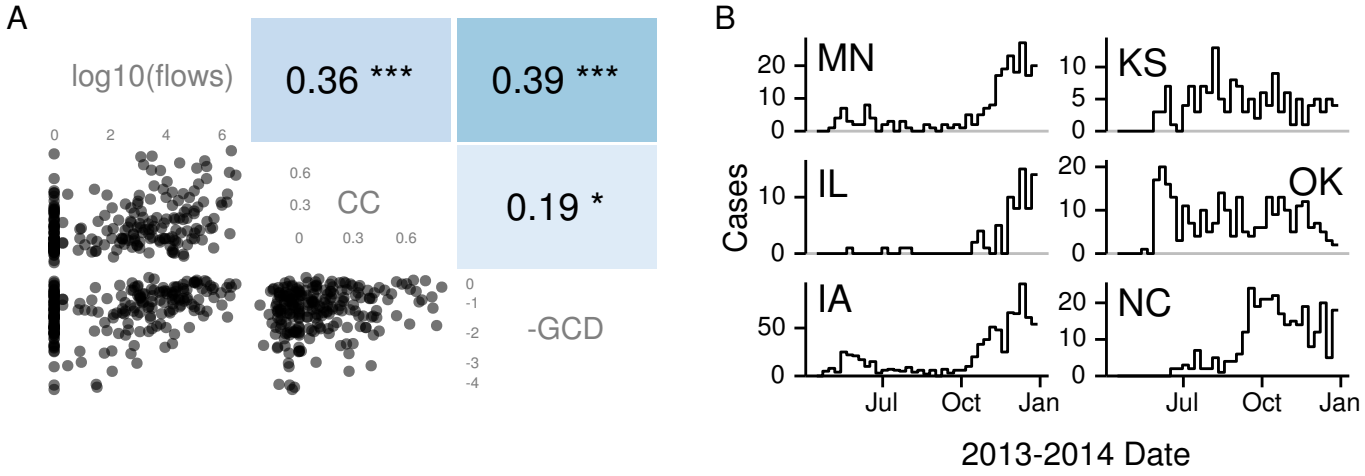

Figure 2. Similarity in case dynamics correlated more strongly with transport than with distance. (A) Scatter plots and Pearson correlations between shipment flows, similarities in PEDV dynamics, and spatial proximity. These state-to-state relationships are measured as annual transport flows, cross correlations (CC) of cases, and negative centroid-to-centroid great circle distances (-GCD). Stars next to the correlations indicate the significance level from a Mantel test as follows: $*$, $P < 0.05$; $***$, $P < 0.001$. -GCD is in units of 1000 km. (B) PEDV dynamics for selected states. Distinct shapes are apparent in the time series of the Midwestern states (MN, IL, IA), Kansas and Oklahoma, and North Carolina.

Now, the flows were themselves correlated with the geographic distance between states, and these distances were in turn correlated with cross correlations (Fig. 2A). Thus we also examined the partial correlation of flows and cross correlations, controlling for geographic distance. This partial correlation equaled about 0.31 whether directed or undirected relationships were used, and thus controlling for distance does not greatly diminish the correlation. As the limited strength of the correlations in Fig. 2A also indicates, there is not a clear linear relationship between the sizes of flows and cross correlations. For example, North Carolina's time series did not correlate strongly with many other states in spite of North Carolina being a major source of swine for many other states (S4 Figure).

Predictors of cumulative burdens

The previous analysis did not account for many potential confounding variables. To address that limitation, we performed variable selection on a panel of candidate variables to identify those with the most robust associations with cumulative burdens of PEDV. Candidate variables were chosen based on availability and expected effects on either reporting rates or risk. In addition to transport-associated risk, we considered risk dependent on climate, as PEDV is an enveloped virus, and on farm density and geographic distance, as spatial clusters of infection have been reported. Table 1 contains a brief description of all of the variables considered.

Table 1. Variables available for selection in regression model of cumulative cases. This table summarizes the variables used by describing groups of one or more variables that were closely related.

Variable group	Description
Number of operations	Count of livestock operations in state with 25 or more swine.
Balance sheet	Dec. 2012 swine inventory and 2011–2012 pig crop, inshipments, and marketings.
Farm resource region	Proportion of swine farms in each state in each region, indicative of climate.
Nearby cases	Weighted average of cases nearby in flow network or geographically.
Farm density	Summary statistics for each state of the number of farms in each county per km^2.

We considered cumulative burdens to be an appropriate response variable because many of the candidate variables were not time-varying. Also, cumulative measures of burden may be more robust measures of risk. Using the more detailed data available after June 2014 [19], we found the Spearman rank correlation between positive accessions and positive farms to equal 0.91, as compared to 0.74 for the weekly counts.

We used absolute burdens rather than prevalence as the response variable because of uncertainty in the correct denominator for calculation of prevalence. Our analysis of the cases by age class, available in S2 Text, indicates that sampling of cases may be highly biased toward farms with suckling pigs, which is reasonable because such farms would likely observe the most mortality in an outbreak [22]. However, we did not attempt to correct for this bias, as it was not clear that such a correction would be accurate.

Most of our predictors were correlated with other predictors as well as with the total cases in each state, making it unclear which variables were likely to be the best predictors. We used elastic net regression [23] with stability selection [24] to identify a set of best predictors. Further details on the modeling approach are in Materials and Methods.

With this approach, we found that the number of farms in a state was the only variable selected as a reliable predictor of whether it reported any cases. Among those states reporting cases, swine inventory and marketings were selected as predictors of the total number of cases.

Because estimates of average pig litter size are available and PEDV has high mortality among newborn pigs, we considered percent decrease in pig litter size as a second cumulative burden. We fit a model for the probability that a state's decrease exceeded 2 percent, which split the decreases into two loose clusters. For this model, swine inventory was the only variable selected.

Flows in a model of transmission

One explanation for the robust association of inventory with cumulative burdens is that transport flows increase with inventories, and that farm-to-farm contact rates increase with transport flows. One explanation for the comparatively weak association of nearby cases with cumulative burdens is that a more detailed model which includes time is necessary to see their effect. To flesh out these explanations, we fitted the case data to

time series susceptible-infected-recovered models. See the Materials and Methods for the derivation of these models.

Fig. 3 displays the predicted and observed marginal relationships between flows and cases for one of the models, and suggests that flows have an effect. Comparing the likelihoods of models with and without a term for within-state flows, we found evidence against the hypothesis that flows had no effect on transmission rates ($\chi^2_1 = 12.9$, $p < 0.001$). Among those models containing flows, undirected models, which assumed that flows increased contact rates in both source and destination states, fit best, and directed models, which assumed that flows increased contact of susceptible farms in the destination state to infective farms in the origin state, fit worst (Table 2). However, the parameter estimates were generally similar for all of these models, with flows having an appreciable effect (Fig. 4).

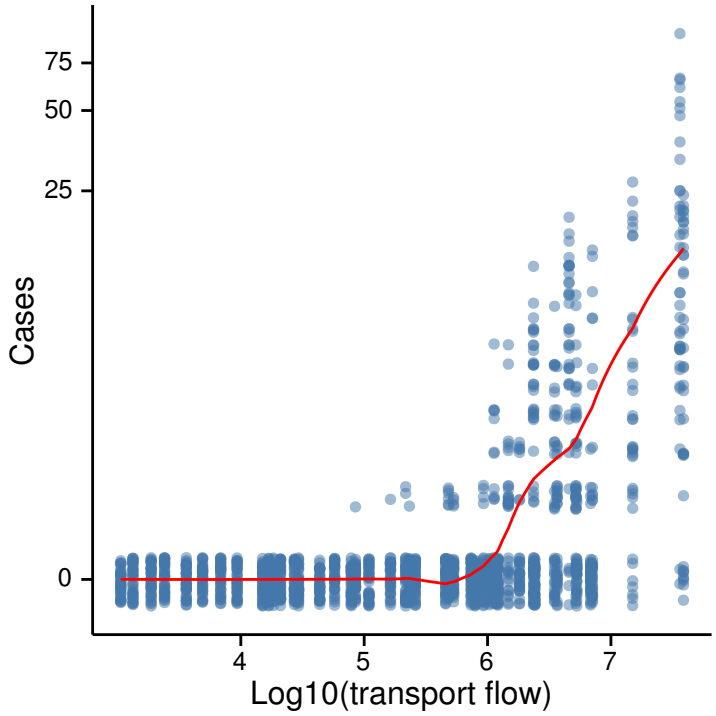

Figure 3. Transport flows were predictive of the number of new cases. The line is R's [25] LOESS smoother of predicted values from the undirected model, where predictions were calculated for each observation by adding fixed effects and conditional modes of random effects. The points are the original data. To display their density, they have been made semi-transparent and jittered along the y axis. The y axis was transformed using $y = \log(\text{Cases} + 1)$.

Effects of farm-level heterogeneity

We have analyzed state-level data with the aim of better understanding farm-level relationships. This ecological approach made the large scale and scope of this study possible but such an approach raises the question of how individual-level variation around group averages may be affecting the results. Thus we conducted simulations to clarify how farm-level variation within states could affect the apparent relationship between state-level flows and cross correlations.

Table 2. Summary of models. The models chiefly differ by how contact is assumed to depend on flows. In the null model, denoted by none, contact was independent of flows. In the internal model, contact was a function of within state flows. In the directed model, contact was a function of flows moving into a state and within-state flows. In the undirected model, contact was a function of both flows into and out of a state. The column "Fit η" indicates whether we estimated the value of η, which corresponds to risk that is independent of the number of infective farms. The symbol θ denotes the dispersion parameter of the negative binomial response. The symbol σ denotes the standard deviation of the random effect of State on transmission rates. The abbreviation d.f. is for degrees of freedom (i.e., the number of parameters estimated). ΔAIC gives the AIC (Akaike information criteria) of a model minus the lowest AIC of all models.

Flow term	Fit η	Intercept	Intercept	$\hat{\sigma}$	d.f.	Log lik.	ΔAIC
undirected	yes	−4.5	2.21	1.36	8	−999.8	0.0
directed	yes	−4.7	1.94	1.52	8	−1017.9	36.3
internal	yes	−4.0	2.20	1.42	8	−1005.4	11.3
internal	no	−3.8	2.17	1.10	7	−1005.6	9.7
none	no	−3.9	2.16	1.52	6	−1012.1	20.6

Our model tracked the infection status of individual farms and included the full set of relationships by which farms could infect each other. Farms were represented as nodes in a network, and the set of transmission routes as undirected edges. A pair of states was represented by placing each of the nodes into one of two disjoint sets.

To obtain a simple model of the dynamics of PEDV, we use a stochastic susceptible–infected–susceptible (SIS) model in which the states of the vertices are either susceptible to infection from any of its neighbors (i.e., other vertices that share an edge with it) or infective and able to spread infection to any of its neighbors. Full model specifications appear in Materials and Methods.

The use of an SIS model is a simplification that does not include the immune state farms are likely to experience following an outbreak, in which no clinical signs are visible and farms may be less infectious. However, the eventual return to a susceptible state in our model is supported by reports [26] of farms experiencing two PEDV outbreak within about one year of each other. Such reinfections are to be expected because, unless controlled oral exposure or vaccines are regularly used, a farm's immunity will wane as animals that were exposed to the virus are replaced.

If we consider sets of vertices in each of two partitions, any edges that link vertices in each of the partitions are members of what is called the *edge cut set* of those partitions—removing those edges would cut off all paths between them. We refer to this set of edges as the cut set for brevity. Our study consists in calculating the cross correlation (with a lag of 1 time step) of the number of cases in two states with varying cut sets. The goal is to see how the cross correlation in cases between a pair of states depends on both data we have (the flows of swine) and data we do not have (how the flow is distributed among pairs of farms). Clearly, the total flow, size of the cut set, and the number of vertices incident to edges in the cut set are all important variables. We use three schemes to tune these variables systematically to provide insight into how they work together to determine the observable coupling of two subpopulations. These schemes, fully described in Materials and Methods, cover extreme scenarios in the distribution of cut set edges among nodes and thus allow for the full range of possible outcomes.

As seen in models that assume homogeneous mixing [27], the largest cross correlations typically occur when the population of infected nodes is near subsistence

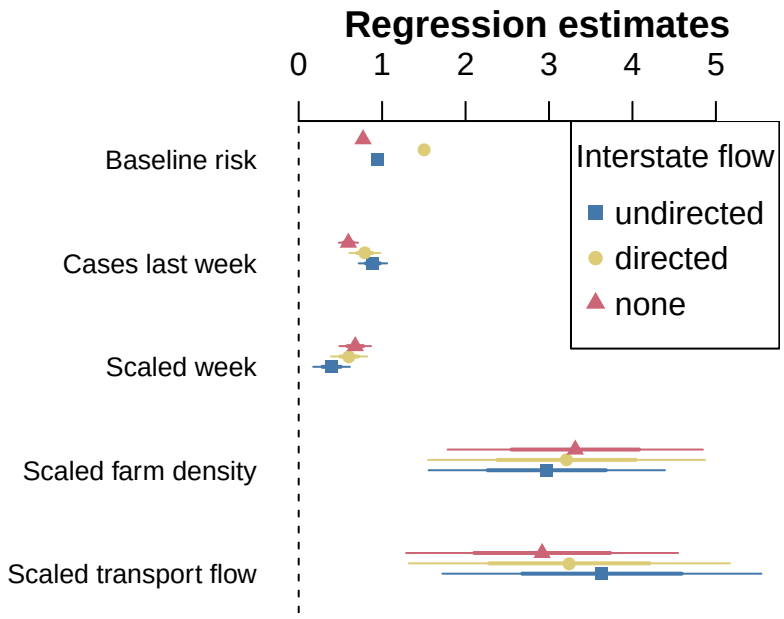

Figure 4. Parameter estimates for the transmission model. The estimates are not sensitive to the choice of interstate flow model, and flows have similar effect sizes to farm densities. Baseline risk refers to the parameter η, which determines the risk of infection when no infectives are present. The error bars represent 50- and 95-percent Wald confidence intervals. The scaled variables were divided by the interquartile ranges to make their effect estimates comparable. The interquartile ranges for week and the logarithm of farm density were were 19.0 and 4.5. Those for the logarithm of undirected, directed, and internal (none interstate) transport flows were 4.5, 4.3, and 3.8.

levels and transient flare-ups in the number of infected nodes occasionally occur in one population and move to the other. In such cases, the R_0 baseline parameter (defined in Materials and Methods) is near 1, and Fig. 5A shows a decrease in the cross correlation as this parameter changes from 1 to 2.

In a similar manner, intermediate values of the *capacity factor*, which we define as the average weight of edges between the subpopulations, lead to the largest cross correlations (Fig. 5B). The capacity factor is representative of the transport flows between states, and the increase in cross correlations as the capacity factor moves from 0.5 to 1 confirms the intuition that cross correlations should increase with flows if flows are transmission pathways. The decrease in the cross correlations as the capacity factor moves from 1 to 2 is counter-intuitive and may explain the weakness of the correlation between cross correlations and flows in Fig. 2A.

In contrast to the previous two variables, the cross correlation is a non-decreasing function of the number of edges in the cut set (Fig. 5C). It seems that a certain threshold number of edges is necessary for large cross correlations to occur and that this threshold depends on the wiring scheme of the network. The controlling parameter for this threshold appears to be closely rated to the vertex connectivity of the network (Fig. 5D). The vertex connectivity of a network may be defined as the number of vertices that must be removed to disconnect part of the network, and it has a close

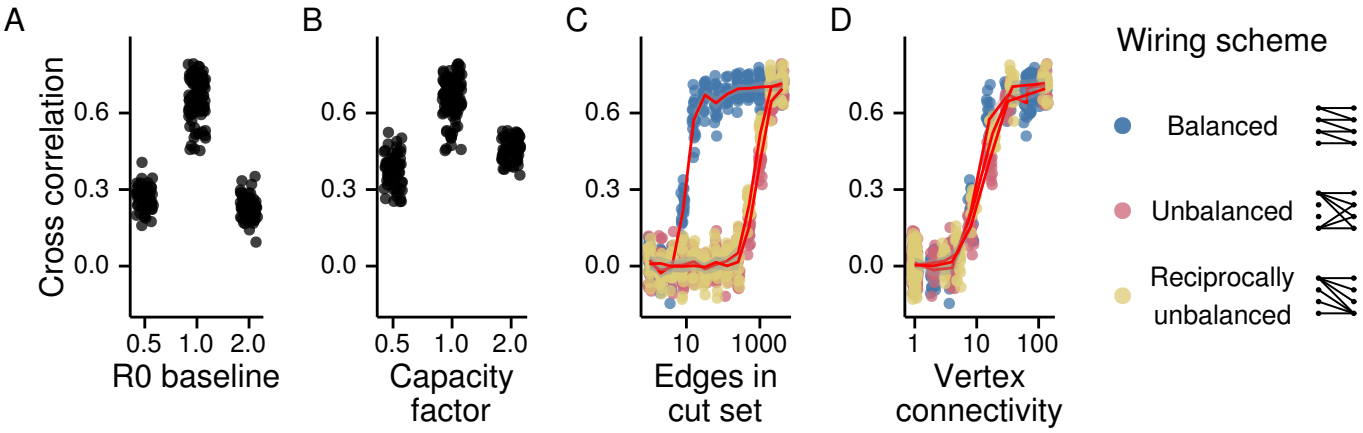

Figure 5. Dependence of the cross correlation of cases on the transmission rate, between-subpopulation flows, and individual-level contact network structure. These relationships are shown in terms of (A) R_0 baseline, (B) the capacity factor, (C) edges in the cut set, and (D) vertex connectivity. For (A) and (B), results from subpopulations of size 16 and 32 are combined in each plot and the maximum number of edges have been added such that the network is fully connected. In (A), the capacity factor is fixed at 1 as R_0 baseline varies. In (B), R_0 baseline is fixed at 1 as the capacity factor varies. For (C) and (D), the subpopulations were of size 64, R_0 baseline was set to 1, and the capacity factor was set to 1. The red lines and semi-transparent bands interpolate through sample means and 95% confidence intervals for each of the schemes of edge addition. The network diagrams on the right give an example of a network with seven between-subpopulation edges. In each diagram, the two vertically aligned columns of dots represent the nodes in each subpopulation. Within-subpopulation edges are not shown to keep the diagrams simple. For all panels, the expected number of infections introduced from outside was set to 1 per time step, and points have been jittered and made semi-transparent to illustrate densities.

connection to the number of vertex-independent paths between pairs of non-adjacent vertices [28]. The implication is that the large cross correlations observed between some pairs of states (Figs. 1C and 2A) are not as informative of the number of edges connecting farms in either state as much as of the number of independent pathways spread in the contact network of the farms in the two states.

Discussion

We have evaluated evidence for transport-associated transmission routes of PEDV at the state level. We first found that the disease dynamics of a pair of states became more similar as transport flows increased. To address some possible confounding, we screened several candidate predictors of the cumulative burden of PEDV and found the relevant ones to be total swine inventory, total number of farms, and marketings. Fitting a mechanistic model, we found that including undirected or internal flows in the transmission term significantly improved likelihoods. Since inventories are closely related to flows, this result illustrates one interpretation of the relevance of inventory for predicting cumulative burdens. A simulation study illustrated how the coupling of disease dynamics at the state level may depend not only on the interstate flow data we analyzed but also on how close the rate of disease spread is to the rate required for long-term subsistence and on farm-level heterogeneity in numbers of interstate contacts.

We also found that suckling cases are highly overrepresented among all age groups. Although reporting bias toward high mortality outbreaks could explain this finding, it may also be the result of real differences in risk. Assuming that each time a trailer arrives for a pick-up there is a similar risk of infection, and that pigs typically spend

about one month on sow-farms being weaned versus three months on finishing farms being fed to market weight, a sow farm of a certain size inventory would have a time-averaged risk 3-fold greater than a finishing farm of the same size inventory.

Our result that transmission pathways are associated with transport flows is consistent with the findings of Ref. [20]. Now at least two pathogens, influenza and PEDV, have been found to have their movement predicted by these flow data. The flow data were generated with such a use in mind [29], and these findings thus provide some support of the concept. Also, this study provides parameters for indirect transmission rates that are based on field data. Such parameters have a strong effect on the output of outbreak simulations, but estimates are rare enough that modelers typically must rely on expert opinion to set them [30, 31].

Our simulation results suggest a refined interpretation of the cross correlation of disease dynamics as indicative of the number of independent pathways of spread in the contact network linking those areas in addition to the flow of infective material between those areas. This refinement may explain a few apparent contradictions in the literature. For example, gravity models replicate the coupling of seasonal influenza dynamics in the United States fairly well [32] in spite of some problems in replicating the commuter flows [33]. Likewise, Bharti and coauthors [34] found it necessary to artificially increase the coupling of coastal towns to the rest of England and Wales by 50 percent to fit the observed persistence of measles in coastal towns in spite of not finding any evidence of increased flows to the region. Given that the seaside was a popular vacation destination at the time [35] and that crowded lodging or recreational areas could lead to large numbers of contacts for children, it seems natural to attribute this increased coupling to increased numbers of independent pathways of spread. The general implication is that we cannot expect any particular relationship between coupling strength and flows without making some assumption about the number of independent paths in the underlying contact network.

We must acknowledge the limitations of our data. The accuracy of the case data and the flow estimates are unknown. One source of error in the flows as a measure of contact rates may be that they exclude transport to harvest plants, and Ref. [10] observes that such movements can result in the contamination of trailers. This limitation may in part explain the weakness of the correlation of flows with cross correlations. At any rate, the common assumption that movements of animals to slaughter represent a dead end for disease spread may well be invalid for PEDV, and more data about these movements and the sharing of trailers among farms could allow for estimation of any associated risks.

Undirected flows were used both in our statistical models and simulations, which deserves some explanation. We suggest interpreting directed flows as a model of transmission by movement of live animals and undirected flows as a model of transmission where the trucks, trailers, and transport personnel act as mechanical vectors of PEDV. For example, if a truck regularly moves animals from one group of farms to another, they might pick up the virus where they drop off the animals and drop off the virus where they pick up the animals just as frequently as the other way around. If trucks generally serve farms in one or two states, then each such mechanical-vector contact from one state to another must generally be reciprocated if we suppose that the number of trucks in each state is stable over time. Given a constant average shipment size, the rate of these contacts will be proportional to our undirected flows. We do not suggest interpreting the better fit of the undirected model (Table 2) as strong evidence that it is in general a better model for PEDV spread because we were only able to differentiate between directed and undirected flows between states, while the bulk of all flow is within states.

Flows are correlated with several other variables, and we cannot rule out that these

other variables are the true drivers of the observed effect of flows. Flows increase with inventory, both because more swine are moving through the farm and because they often have shorter residence times, as larger farms tend to specialize on specific production stages. Perhaps larger farms did not experience more PEDV outbreaks but only reported them more frequently, providing a false signal that flows are associated with risk. This possibility may be reduced in more recent data because the USDA issued a Federal Order [36] that requires all detected outbreaks of PEDV to be reported, actively managed, and monitored.

With the more active and detailed surveillance currently taking place, retrospective analyses may allow the route of infection to farms to be more frequently identified by simple considerations of the order of events in time and the biological plausibility of implicated pathways of spread. However, such analyses may be inconclusive because many hypothetical explanations for most infections may be consistent with the data. In such a situation, assembly of further data that would allow the hazard of infection by multiple routes to be modeled could allow for valuable further insight. For example, one might collect data on the trailer-cleaning protocols and air-filtration systems in place for different farms. Both of these biosecurity measures can be costly, so there are practical benefits to understanding which one may be more effective and when. By further quantifying the risk of spread by movement of animals from infected farms, we could also quantify the benefits of current movement restrictions from PEDV-infected farms. Most generally, the insights into the biology of PEDV and the pragmatics of data collection could likely transfer to other livestock diseases and result in more effective management.

Materials and Methods

Data

The data on transport flows comes from a study by the USDA Economic Research Service [29], and the PEDV case data are provided by the American Association of Swine Veterinarians. Both data sets are open and publicly available.

Percent decrease in pig litter was calculated as the difference in litter size between the 2013 and 2014 December through February estimates [7] divided by the 2013 estimates and multiplied by -100.

Sixteen states had individual pig litter estimates, and a group average is reported for the other states. We assumed that the decreases for states in that group were close to the group average, which was 1 percent, and thus that those states had decreases less than 2 percent.

Relevant predictors of cumulative burden

Many states had no confirmed cases (Fig. 1B) such that the case counts appear to be a mixture of zeroes and a right-skewed distribution of counts. Thus we chose to fit the data to a hurdle model in which the probability of a state having a confirmed case and the number of cases, given that there is at least one case in the state, are described by separate regression models.

Elastic net regression is a penalized likelihood approach to data-driven variable selection, and such an approach is less likely to lead to overfitting than alternatives such as stepwise model selection or the use of Bayes factors [37, p. 59]. Stability selection uses subsampling of the data to identify variables in a way that allows the probablity of selecting noise variables to be limited [24].

The elastic net penalty includes a tuning parameter, denoted by α, that determines the extent to which groups of correlated variables are selected together. We set α to 0.8

to allow for highly correlated variables to be grouped while still keeping the total number of selected variables small.

The choice $\alpha = 0.8$ was made subjectively, but we checked that the results were not sensitive to this choice by also looking at the results with $\alpha \in \{0.01, 0.2, 0.5, 1\}$. For $\alpha \neq 1$, only additional balance sheet variables were selected for all models. When $\alpha = 1$, inventory and resource region 4 were selected as predictors of both litter rate decrease and total cases, and no variables were selected as predictors of whether any cases occurred. We consider these aberrations likely to be an artifact of correlations among predictors, as single members of correlated groups can be selected somewhat arbitrarily when $\alpha = 1$.

For stability selection, we used 1000 subsamples of 63.2 percent of the full data sets (the same percentage that would appear in large bootstrap samples of a data set). The set of selected variables was chosen by using a threshold parameter π_{thr} of 0.6 and choosing the regularization parameter λ to select as many variables as possible while keeping the per-comparison error rate (i.e., the probability that any one variable is incorrectly selected) below 0.05. Predictors were put onto the same scale by dividing by standard deviations.

Transmission models

We assume that the expected number of infectives in state i at week $t+1$, $\mathrm{E}(I_{i,t+1})$, follows

$$\mathrm{E}(I_{i,t+1}) = \beta_{i,t}(\textstyle\sum_j w_{i,j} I_{j,t} + \eta)^{\alpha} S_{i,t}, \tag{1}$$

where $\beta_{i,t}$ is the transmission rate for state i at time t, $w_{i,j}$ is the weight for the influence of infectives in state j on susceptibles in state i, η is parameter that determines the influence of other sources of infection, α determines the power by which the expected number of transmissions grows with these risks, and $S_{i,t}$ is the number of susceptibles in state i at week t. We set $S_{i,t} = N_i - \sum_{k=0}^{t-1} I_{i,t}$, where N_i is the number of farms in state i from the 2002 Census of Agriculture [38]. This model is a variant of the time series SIR (susceptible–infective–recovered) model introduced in Ref. [39]. S3 Text discusses some of the assumptions and data we used for this model.

Our calculation of $S_{i,t}$ assumes that all farms were susceptible to infection at the beginning of the epizootic and that farms pass on to an immune state following infection. The assumption of complete susceptibility seems reasonable for the United States given the absence of previous reports of PEDV and the high frequency of high-mortality outbreaks that followed the first reported outbreak [9]. Although PEDV has been observed to reoccur on a farm [26], that observation was a newsworthy event [40] and it followed a 6-month interval of normal operations. Thus the assumption of immunity over the 38 week period that we analyze seems reasonable.

Our transmission rate $\beta_{i,t}$ in Eq. 1 takes the form

$$\beta_{i,t} = \exp(c_0 + Z_i + c_1 t)(N_i^2 d_i)^{c_2} f_i^{c_3} N_i^{-2}, \tag{2}$$

where the c_i are unknown parameters that we estimate, Z_i represents state-level random effects, d_i is a state-level summary statistic of the county-level farm density from the 2007 Census [41], and f_i is value characterizing the average flow of swine through individual farms in state i. c_1 allows the transmission rate to vary seasonally, which has been proposed as an explanation for why most cases occurred in the fall and winter. For the summary statistic d_i, we used the median county-level density among counties with any farms in the state. The results were not sensitive to using this statistic versus others such as the overall median or mean. d_i is multiplied by N_i^2 because that led to the greatest correlation between the density and flow terms on the logarithmic scale, and we wished to as much as possible separate the estimated effects of flows with those

of farm density. It also allowed us to see whether density-dependent transmission [42] is suggested by the data, which would have corresponded to estimates $(\hat{c}_2, \hat{c}_3) \approx (1, 0)$.

The characteristic flows f_i in Eq. 2 and the weights w_i in Eq. 1 are calculated in various ways to model the rate of contact of a susceptible farm with infected farms in various scenarios. We make the derivations assuming $\alpha = 1$, and values of α below 1 (as usual) break the precise mechanistic interpretation of the model but can be understood as capturing the effects of infective farms being clustered together in the contact network. Let $F_{i,j}$ be the number of swine shipped to farms in state i from farms in state j per year. In the *directed model,* only farms receiving animals are at risk for infection. Then, omitting the time subscripts for simplicity, susceptible farms in state i are infected at a rate proportional to $\sum_j F_{i,j}(N_i N_j)^{-1} I_j$, or $f_i N_i^{-2} \sum_j w_{i,j} I_j$, where $f_i = \sum_j F_{i,j}$ and $w_{i,j} = N_i N_j^{-1} F_{i,j} f_i^{-1}$. In the *undirected model,* both farms sending and farms receiving animals may be at risk, and susceptible farms in state i are infected at a rate proportional to $\sum_j (F_{i,j} + F_{j,i})(N_i N_j)^{-1} I_j$, which implies that $f_i = \sum_j F_{i,j} + F_{j,i}$ and $w_{i,j} = N_i N_j^{-1}(F_{i,j} + F_{j,i}) f_i^{-1}$.

In the *internal model,* both farms sending and receiving animals may be at risk, but transmission associated with flows only occurs within a state. Thus susceptible farms in state i are infected at a rate proportional to $2F_{i,i} N_i^{-2} I_i$, which implies that $f_i = 2F_{i,i}$ and $w_{i,j} = \delta_{i,j}$, Kronecker deltas. Comparison of the fit of this model with the directed or undirected models allows any effects of between-state transmission to be seen. The internal model also includes in the case that $c_3 = 0$ a null model which has no flows in it, which we use in a likelihood ratio test of the hypothesis that flows have no effect on transmission rates.

The values of $F_{i,j}$, when $i \neq j$, come directly from the estimates of interstate flows in Ref. [29]. We estimated within-state flows in two ways. In the first, a demand for pigs was calculated for state i from 2002 sales [38] finish-only and nursery operations plus the deaths reported in the 2001 balance sheet [43]. Internal flow, $F_{i,i}$, was estimated as the this demand less imports, $\sum_{j,j\neq i} F_{i,j}$. In the second method, $F_{i,i}$ was estimated as the combined sales of farrow-to-wean, farrow-to-feeder, and nursery operations less exports, $\sum_{j,i\neq j} F_{j,i}$. For most states with large inventories, the logarithms of these two estimates were similar relative to estimates from other states, and we averaged the log-transformed estimates to generate a single estimate. For the other states, one of the estimates was negative, and we simply used the positive estimate. We suspect the negative estimates and the difference between the positive estimates stem in part from us not being able to use 2001 sales data or to account for internal supplies of and demand for breeding animals. Coarse as these estimates may be, it still seems reasonable to us that they will permit detection of large, state-level associations.

To fit the model, we form a linear predictor of $\log \mathrm{E}(I_{i,t+1})$ by substituting Eq. 2 into Eq. 1 and taking logarithms to obtain

$$\log \mathrm{E}(I_{i,t+1}) = c_0 + Z_i + c_1 t + c_2 \log(N_i^2 d_i) + c_3 \log f_i + \alpha \log(\textstyle\sum_j w_{i,j} I_{i,t} + \eta) + \log S_{i,t} - 2 \log N_i. \quad (3)$$

We fit this model to data from all 50 states with the assumption that the observed cases $I_{i,t+1}$ have a negative binomial distribution with an unknown, but constant, dispersion parameter which we denote with θ. This parameter is related to the variance by $\mathrm{Var}(I_{i,t+1}) = \mathrm{E}(I_{i,t+1})[1 + \mathrm{E}(I_{i,t+1})/\theta]$. We assume that the random effect Z_i is normally distributed. Then the likelihood is fully specified. We calculate marginal likelihoods with the Laplace approximation and numerically find the parameters that maximize it. In some cases we fixed η to 0.5, which allowed the model to be fully fit with both the lme4 [44] and glmmADMB [45] packages in R [25]. To make sure our results were not sensitive to $\eta = 0.5$, we used R's optimize function to find the value of η in $[0, 5]$ with highest likelihood.

We performed several diagnostic checks of our fits, including checking for signs of nonlinearity with partial residual plots and for signs of temporal autocorrelation in the residuals. We also verified that the flows term is significant in models lacking random effects, and after excluding any data points with dfbetas [37, e.g.] above 0.2.

Simulation model

We use a discrete-time model in which the probability that a susceptible vertex i avoids becoming infected at the next time step is equal to $\exp\left(-\beta \sum_{j \neq i} A_{ij} I_j\right)$, where β is the transmission rate, A_{ij} is the weight of the edge pointing to vertex i from vertex j, and I_j is an indicator variable equal to one if vertex j is infected and equal to zero otherwise. The edge weight A_{ij} represents the amount of infective material that vertex i may receive from vertex j and the transmission rate is the expected number of infections per unit of this infective material. In the case of livestock diseases, we might think of the edge weight as proportional to the number of animals moved from one farm to another and the transmission rate as the rate at which the probability of a farm avoiding infection decreases for a given number of animals introduced from an infective farm.

We set the transmission rate in terms of an R_0 *baseline* parameter, which we define as $\beta(N-1)$, where N is the number of vertices in the network. Thus we do not change the transmission rate as the number of edges is changed, which we find makes the results easier to interpret.

For simplicity, we assume infective vertices recover in one time step. To allow for highly stochastic dynamics without extinction, we assume that all vertices have some constant probability of infection from vertices outside of the simulated population. We describe this parameter in terms of the expected number of introduced infections, which is equal to the number of vertices in the network times the per-time-step probability of any one of them being infected from an external source.

We calculate the cross correlation using a window of 500 time steps that follows a warm-up period of 500 time steps that allowed the model to reach a stationary distribution. All simulations began with a completely susceptible population at the beginning of the warm-up.

Wiring schemes

The vertex sets corresponding to each of the subpopulations are kept fully connected to emphasize the community structure of the network. Also, because fully connected networks are highly symmetric, the sets of topologically unique cut sets are reduced from more general cases and thus easier to systematically explore.

The wiring schemes differ in which edges are added as we increase the size of the cut set, which is most easily described in terms of non-zero elements of the adjacency matrix A of the network. We begin with a block-diagonal adjacency matrix where the the blocks (submatrices) on the diagonal contain weights of within-subpopulation edges and the complementary blocks contain the weights of edges in the cut set. We consider only undirected networks so a particular cut set can be described in terms of one of the cut-set blocks. In the *balanced* scheme, the degree distribution (the probablity mass distribution for the number of neighbors of each vertex) of the two subpopulations is kept as balanced as possible. Thus cut-set edges are added by forming bands on the diagonal of the block of increasing width. In the *unbalanced* scheme, the degree distribution is kept as unbalanced as possible. Thus cut-set edges are added by filling in the cut-set block columwise. Consequently, one of the subpopulations contains vertices with many cut set members incident to them, which we refer to as hubs. In the *reciprocally-unbalanced* scheme, cut set edges are added by filling in columns and rows

in an alternating manner. Thus hubs appear in each subpopulation in an alternating manner.

In all schemes, we distribute the total weight of cut-set edges evenly among them. The total flow is set to n^2c, where n is the size of each of the subpopulations and c is a tuning parameter we refer to as a capacity factor. The weight of edges outside of the cut set was fixed at 1. Thus when varying the number of edges in the cut set, the modularity statistic Q [28] remains unchanged. This invariance makes our schemes similar to the different mixing styles introduced by Ref. [46].

Software

We used R [25] for most of this work. The key contributed packages used were c060 [47], igraph [48], glmmADMB [49], glmnet [50], ggplot2 [51], lme4 [52], and vegan [53]. We performed the edge bundling for Figs. 1C and 1D using JFlowMap [54]. Code to reproduce the results is archived on the web [55], and has been developed to run in Docker [56] containers for enhanced reproducibility. Thus, after installing one open-source software package on their personal computer, interested readers may quickly repeat our analysis, examine intermediate results, perform their own diagnostics, and extend this work.

Supporting Information

S1 Text

Detailed data description. This text describes the original sources, rationale for inclusion, and preparation of various data for analysis.

S2 Text

Age-specific reporting bias. This text describes the analysis that indicates an age-specific reporting bias.

S3 Text

Transmission model details. This text provides further details on assumptions and data used for the transmission model.

S1 Figure

Scatter plots and Pearson correlations between pair-averaged (i.e., undirected) transport flows, similarities in PEDV dynamics, and spatial proximity. These state-to-state relationships are measured as annual transport flows, cross correlations (CC) of cases, and negative centroid-to-centroid great circle distances (-GCD). Stars next to the correlations indicate the significance level from a Mantel test as follows: **, $P < 0.01$; ***, $P < 0.001$. -GCD is in units of 1000 km.

S2 Figure

Rank scatter plots and Spearman correlations between transport flows, similarities in PEDV dynamics, and spatial proximity. These state-to-state relationships are measured as annual transport flows, cross correlations (CC) of cases, and negative centroid-to-centroid great circle distances (-GCD). Stars next to the

correlations indicate the significance level from a Mantel test as follows: *, $P < 0.05$; ***, $P < 0.001$. -GCD is in units of 1000 km.

S3 Figure

Rank scatter plots and Spearman correlations between pair-averaged (i.e., undirected) transport flows, similarities in PEDV dynamics, and spatial proximity. These state-to-state relationships are measured as annual transport flows, cross correlations (CC) of cases, and negative centroid-to-centroid great circle distances (-GCD). Stars next to the correlations indicate the significance level from a Mantel test as follows: *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$. -GCD is in units of 1000 km.

S4 Figure

Comparison of swine shipment flows and coupling of case dynamics. (A) Annual swine shipment flows from source states to destination states. (B) Cross correlations of PEDV cases per week. Cross correlations are calculated as the correlations between cases in the leading state with those in the lagging state in the previous week. Within-state values of flows and cross correlations are not included in the analysis and appear as white squares. In both panels, rows and and columns are arranged to cluster together states with similar shipment flows.

Acknowledgments

We thank John Korslund for useful feedback on veterinary and swine industry subject matter.

References

1. USDA APHIS (2013). PED technical note. Available: `http://www.aphis.usda.gov/animal_health/animal_dis_spec/swine/downloads/ped_tech_note.pdf`. Accessed 24 March 2014.

2. Jung K, Wang Q, Scheuer KA, Lu Z, Zhang Y, et al. (2014) Pathology of US porcine epidemic diarrhea virus strain PC21A in gnotobiotic pigs. Emerg Infect Dis 20: 662–665.

3. Stevenson GW, Hoang H, Schwartz KJ, Burrough ER, Sun D, et al. (2013) Emergence of porcine epidemic diarrhea virus in the United States: clinical signs, lesions, and viral genomic sequences. J Vet Diagn Invest 25: 649–654.

4. USDA APHIS VS (2014). Porcine epidemic diarrhea virus (PEDv) testing data from NAHLN laboratories. Available: `http://www.aasv.org/pedv/PEDV_weekly_report_140108.pdf`. Accessed 14 January 2014.

5. USDA NASS (2013). Meat Animals Production, Disposition, and Income 2012 Summary. Available: `http://usda01.library.cornell.edu/usda/current/MeatAnimPr/MeatAnimPr-04-25-2013.zip`. Accessed 31 July 2013.

6. USDA NASS (2013). Quarterly hogs and pigs (December 2013). Available: `http://usda.mannlib.cornell.edu/usda/nass/HogsPigs//2010s/2013/HogsPigs-12-27-2013.pdf`. Accessed 6 May 2014.

7. USDA NASS (2014). Quarterly hogs and pigs (March 2014). Available: http://www.nass.usda.gov/Publications/Todays_Reports/reports/hgpg0314.pdf. Accessed 8 April 2014.

8. Hill C, Raizman E, Snider T, Goyal S, Torremorell M, et al. (2014) Emergence of porcine epidemic diarrhoea in North America. FOCUS ON 9: 1–8.

9. EFSA AHAW Panel (2014) Scientific opinion on porcine epidemic diarrhoea and emerging pig deltacoronavirus. EFSA Journal 12: 3877.

10. Lowe J, Gauger P, Harmon K, Zhang J, Connor J, et al. (2014) Role of transportation in spread of porcine epidemic diarrhea virus infection, United States. Emerg Infect Dis 20: 872–874.

11. Thomas PR, Karriker LA, Ramirez A, Zhang J, Ellingson JS, et al. (2015) Evaluation of time and temperature sufficient to inactivate porcine epidemic diarrhea virus in swine feces on metal surfaces. J Swin Heal Prod 23: 84–90.

12. Alonso C, Goede DP, Morrison RB, Davies PR, Rovira A, et al. (2014) Evidence of infectivity of airborne porcine epidemic diarrhea virus and detection of airborne viral RNA at long distances from infected herds. Vet Res 45: 73.

13. Dee S, Clement T, Schelkopf A, Nerem J, Knudsen D, et al. (2014) An evaluation of contaminated complete feed as a vehicle for porcine epidemic diarrhea virus infection of naïve pigs following consumption via natural feeding behavior: proof of concept. BMC Vet Res 10: 176.

14. Pasick J, Berhane Y, Ojkic D, Maxie G, Embury-Hyatt C, et al. (2014) Investigation into the role of potentially contaminated feed as a source of the first-detected outbreaks of porcine epidemic diarrhea in Canada. Transbound Emerg Dis 61: 397–410.

15. Opriessnig T, Xiao CT, Gerber PF, Zhang J, Halbur PG (2014) Porcine epidemic diarrhea virus RNA present in commercial spray-dried porcine plasma is not infectious to naïve pigs. PLoS One 9: e104766.

16. Bowman AS, Krogwold RA, Price T, Davis M, Moeller SJ (2015) Investigating the introduction of porcine epidemic diarrhea virus into an Ohio swine operation. BMC Vet Res 11: 38.

17. USDA (2014). Summary of PEDV Actions. Available: http://www.usda.gov/documents/pedv-summary-actions.pdf. Accessed 10 May 2014.

18. Brooks-Pollock E, Roberts GO, Keeling MJ (2014) A dynamic model of bovine tuberculosis spread and control in Great Britain. Nature 511: 228–231.

19. USDA APHIS VS (2015). Swine enteric coronavirus disease (SECD) situation report – Feb 12, 2015. Available: http://www.aphis.usda.gov/animal_health/animal_dis_spec/swine/downloads/secd_sit_rep_02_12_15.pdf. Accessed 14 Februrary 2014.

20. Nelson MI, Lemey P, Tan Y, Vincent A, Lam TTY, et al. (2011) Spatial dynamics of human-origin H1 influenza A virus in North American swine. PLoS Pathog 7: e1002077.

21. Rozhnova G, Nunes A, McKane A (2012) Phase lag in epidemics on a network of cities. Phys Rev E 85: 051912.

22. Jung K, Saif LJ Porcine epidemic diarrhea virus infection: etiology, epidemiology, pathogenesis and immunoprophylaxis. Vet J doi:101016/jtvjl201502017 .

23. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol 67: 301–320.

24. Meinshausen N, Bühlmann P (2010) Stability selection. J R Stat Soc Series B Stat Methodol 72: 417–473.

25. R Core Team (2014) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.

26. Ackerman MA (2014). PEDv recurrence. Available: `https://www.pig333.com/clinical-case-of-the-world/pedv-recurrence_9260/`. Accessed 25 January 2015.

27. O'Regan SM, Drake JM (2013) Theory of early warning signals of disease emergence and leading indicators of elimination. Theor Ecol 6: 333–357.

28. Newman MEJ (2010) Networks: An introduction. Oxford: Oxford Univ. Press.

29. USDA ERS (2003). Interstate livestock movements. By D. Shields and K. Mathews. Available: `http://www.ers.usda.gov/publications/ldpm-livestock,-dairy,-and-poultry-outlook/lpdm10801.aspx#.U26fN1Qt5Mk`. Accessed 14 November 2013.

30. Martínez-López B, Ivorra B, Fernández-Carrión E, Perez A, Medel-Herrero A, et al. (2014) A multi-analysis approach for space–time and economic evaluation of risks related with livestock diseases: The example of FMD in Peru. Prev Vet Med 114: 47–63.

31. McReynolds SW, Sanderson MW, Reeves A, Hill AE (2014) Modeling the impact of vaccination control strategies on a foot and mouth disease outbreak in the Central United States. Prev Vet Med 117: 487–504.

32. Viboud C, Bjørnstad ON, Smith DL, Simonsen L, Miller MA, et al. (2006) Synchrony, waves, and spatial hierarchies in the spread of influenza. Science 312: 447–451.

33. Truscott J, Ferguson NM (2012) Evaluating the adequacy of gravity models as a description of human mobility for epidemic modelling. PLoS Comput Biol 8: e1002699.

34. Bharti N, Xia Y, Bjornstad ON, Grenfell BT (2008) Measles on the edge: Coastal heterogeneities and infection dynamics. PLoS One 3.

35. Walton JK (2000) The British Seaside: Holidays and Resorts in the Twentieth Century. Manchester University Press.

36. USDA (2014). Federal order: Reporting, herd monitoring and management of novel swine enteric coronavirus diseases. Available: `http://www.aphis.usda.gov/newsroom/2014/06/pdf/secd_federal_order.pdf`. Accessed 3 July 2014.

37. Harrell FE (2001) Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. New York: Springer.

38. USDA NASS. 2002 Census of Agriculture Query Tool. Available: http://www.agcensus.usda.gov/Publications/2002/Download_Data_Query_Application/agcensus2002.zip. Accessed 23 December 2014.

39. Bjørnstad ON, Finkenstädt BF, Grenfell BT (2002) Dynamics of measles epidemics: Estimating scaling of transmission rates using a time series SIR model. Ecol Monogr 72: 169–184.

40. Polansek T (2014). Exclusive: Deadly pig virus re-infects U.S. farm, fuels supply fears. Available: http://www.reuters.com/article/2014/05/28/us-pig-virus-immunity-idUSKBN0E811N20140528. Accessed 10 December 2014.

41. USDA NASS (2009). 2007 Desktop Data Query Tool 1.02. Available: http://www.agcensus.usda.gov/Publications/2007/Online_Highlights/Desktop_Application/dataquery.zip. Accessed 14 January 2013.

42. Begon M, Bennett M, Bowers RG, French NP, Hazel SM, et al. (2002) A clarification of transmission terms in host-microparasite models: numbers, densities and areas. Epidemiol Infect 129: 147–153.

43. USDA NASS (2002). Meat Animals Production, Disposition, and Income 2001 Summary. Available: http://usda.mannlib.cornell.edu/usda/nass/MeatAnimPr//2000s/2002/MeatAnimPr-04-26-2002.zip. Accessed 23 December 2014.

44. Bates D, Mächler M, Bolker B, Walker S (2014). Fitting linear mixed-effects models using lme4. Preprint. Available: arXiv:1406.5823. Accessed 9 March 2015.

45. Fournier DA, Skaug HJ, Ancheta J, Ianelli J, Magnusson A, et al. (2011) AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. Optim Methods Softw 27: 233–249.

46. Min Y, Jin X, Ge Y, Chang J (2013) The role of community mixing styles in shaping epidemic behaviors in weighted networks. PLoS One 8: e57100.

47. Sill M, Hielscher T, Becker N, Zucknick M (2014) c060: Extended inference with lasso and elastic-net regularized Cox and generalized linear models. J Stat Softw 62: 1–22.

48. Csardi G, Nepusz T (2006) The igraph software package for complex network research. InterJournal Complex Systems: 1695.

49. Skaug H, Fournier D, Bolker B, Magnusson A, Nielsen A (2014) Generalized Linear Mixed Models using AD Model Builder. R package version 0.8.1.

50. Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33: 1–22.

51. Wickham H (2009) ggplot2: Elegant Graphics for Data Analysis. Springer New York.

52. Bates D, Maechler M, Bolker B, Walker S (2014) lme4: Linear mixed-effects models using Eigen and S4. URL http://CRAN.R-project.org/package=lme4. R package version 1.1-7.

53. Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, et al. (2015) vegan: Community Ecology Package. URL http://CRAN.R-project.org/package=vegan. R package version 2.2-1.

54. Boyandin I, Bertini E, Lalanne D (2010) Using flow maps to explore migrations over time. In: Proceedings of Geospatial Visual Analytics Workshop in conjunction with The 13th AGILE International Conference on Geographic Information Science (GeoVA). Guimaraes (Portugal).

55. O'Dea E (2015). 2015pedv: Files associated with Feb. 17 draft. doi:10.5281/zenodo.15541.

56. Boettiger C (2015) An introduction to Docker for reproducible research. SIGOPS Oper Syst Rev 49: 71–79.