Title
Using Search Engine Data as a Tool to Predict Syphilis

Journal
Epidemiology, 29(4)

ISSN
1044-3983

Authors
Young, Sean D
Torrone, Elizabeth A
Urata, John
et al.

Publication Date
2018-07-01

DOI
10.1097/ede.0000000000000836

Peer reviewed

HHS Public Access

Author manuscript

Epidemiology. Author manuscript; available in PMC 2019 July 01.

Published in final edited form as:

Epidemiology. 2018 July ; 29(4): 574–578. doi:10.1097/EDE.0000000000000836.

Using Search Engine Data as a tool to Predict Syphilis

Sean D. Young, PhD, MS[1], **Elizabeth A. Torrone, PhD, MSPH**[2], **John Urata, MS**, and **Sevgi O. Aral, PhD, MS**[2]

[1]University of California Institute for Prediction Technology, Department of Family Medicine, University of California, Los Angeles, California, USA

[2]Division of STD Prevention, Centers for Disease Control and Prevention, Atlanta, Georgia, USA

Summary

Background—Researchers have suggested that social data (e.g., social media and online search data) might be used as a tool to monitor and predict syphilis and other sexually transmitted diseases; however, little research has explored this question. Based on the hypothesis that people at risk for syphilis would seek sexual health and risk-related information on the internet, we sought to investigate associations between internet state-level search query (e.g., Google Trends) data and state-level reported weekly syphilis cases.

Methods—Weekly counts of reported primary and secondary (P&S) syphilis for 50 states from 2012 to 2014 were obtained from the Centers for Disease Control and Prevention. Weekly internet search query data regarding 25 risk-related keywords from 2012 to 2014 for 50 states were collected using Google Trends; 155 weeks of Google Trends data were joined (by week and state) with one week lag to the weekly syphilis data, for a total of 7,750 data points. Using the least absolute shrinkage and selection operator, three linear mixed models were trained on the first ten weeks of each year. Models for 2012 and 2014 were validated for the following 52 weeks; the 2014 model was validated for the following 42 weeks.

Findings—The models, consisting of different sets of keyword predictors for each year, accurately predicted 144 weeks of P&S syphilis counts for each state, with an overall average R^2 of 0·89 and overall average RMSE of 4·9.

Interpretation—Google Trends internet search data from the prior week might be used to predict cases of P&S syphilis in the following weeks for each state. This study suggests researchers should further explore this topic.

Funding—National Institute of Mental Health (5R01MH106415).

Syphilis is a nationally notifiable condition in the United States with increasing public health significance.[1] After reaching a historic low in 2001, rates of reported primary and secondary

Corresponding Author: Sean Young, PhD, MS, University of California, Los Angeles, Department of Family Medicine, 10880 Wilshire Blvd., Ste. 1800, Los Angeles, CA 90024, *Telephone*: 310-794-8530, *Fax*: 310-794-2768, sdyoung@mednet.ucla.edu.

Declaration of interests

The authors declare no conflict of interest.

Disclaimer: The findings and conclusions in this manuscript are those of the authors and do not necessarily represent views of the Centers for Disease Control and Prevention

(P&S) syphilis (the most infectious stage of syphilis) in the United States increased almost every year, with 2015 showing the highest recorded increase from the previous year at 19.4%.[2]

Syphilis is traditionally monitored through case reporting and prevalence monitoring.[3] Although successful in identifying syphilis cases, current methods are subject to several limitations. One limitation requires people to interact with a health provider and be diagnosed in order for the case to be identified. As such, many infections are missed and case reports underestimate burden of disease. Although case reports reflect syphilitic infections diagnosed in all provider settings, the majority of prevalence monitoring occurs in high prevalence settings, such as jails and STD clinics, and prevalence estimates may not be representative of the whole population. Another limitation includes a non-trivial investment of time and funds to conduct testing, surveys, interviews, as well as to process, analyze, and report results. In addition, accurate diagnoses and staging of new syphilis infections requires a blood draw as well as an individual's treatment and rapid plasma reagin titer history, which may or may not be readily available. New methods are needed to help supplement current strategies and better monitor syphilis.

Internet technologies have the potential to mitigate some of the shortcomings of current health monitoring systems and might be used to supplement existing methods. The CDC has already created guidelines on best practices for using the internet and social media as an intervention tool for sexually transmitted diseases (STDs).[4] Because a large percentage of people use the internet to search for health information, other internet technologies, such as search engine data might be used as a tool for syphilis monitoring and prediction.[5] Recent research already suggests that internet search behavior might be useful for predicting public health-related events.[6]

Google provides national and state-level search engine query data for free via a web-based dashboard called Google Trends (https://www.google.com/trends/). In 2009, a study demonstrated that it was possible to predict influenza based on geography and time using Google's search query data;[7] subsequent research improved modeling and addressed limitations of this initial study.[8,9] A limitation of the subsequent research is that both studies focused on geographic scales that were relatively coarse (at the USA nation-level or 9 USA influenza regions).[8,9] Other researchers have explored ways of using search query data to monitor infectious diseases[10–14] and search terms used in STD-related searches.[15] Additional studies have explored the use of internet technologies such as Yahoo[16] and databases[17] for illness surveillance. However, there is a lack of studies focusing on using Google Trends data to predict syphilis cases and at geographically finer state-level scale.

The primary purpose of the present study was to investigate whether Google Trends search data could be used to predict weekly state-level syphilis case reports in the United States. In addition, this study explored whether search data trained on the first ten weeks of each year could be used predict the remaining weeks of the year of P&S syphilis new cases.

Methods

Weekly relative search volume data for 25 selected keywords were collected from January 1, 2012, to December 31, 2014, for all 50 states using Google Trends. The 25 keywords consisted of drug related and sexual-risk related keywords which may be associated with transmission and infection. Specific keywords are not included as per internal review board-related concerns about using this type of information to identify participants.[18] However, examples of similar keywords and phrases used include ‘sex;’ ‘STD help,’ ‘find sex,’ ‘sex without a condom;’ ‘do I have an STD;’ ‘symptoms of STD’s;’ and ‘how to find sex right now.’ The weekly relative search volume for each keyword is based on a random sample of all searches within the timeframe for each state and was normalized and then scaled from 0 to 100 by Google relative to the keyword with the highest weekly peak search volume.

This study focuses on cases of P&S syphilis which represent recent infections.[1] Weekly county-level P&S syphilis data for 2012 through 2014 were obtained from the CDC and aggregated to state-level. The data were then merged by state and joined with the weekly Google Trends data; the processed dataset included 155 data points for each state, for a total of 7,750 data points.

In order to assess whether Google Trends data could predict cases of reported syphilis by state, we implemented a linear mixed-effects model (LMM). LMM is a regression method that accounts for repeated measurements not only between groups (in our case states), but also within groups.[19] Least absolute shrinkage and selection operator (LASSO) was implemented as a supervised machine learning method in order to select the model input predictors from a larger set of possible predictors. LASSO shrinks the coefficients of less important features and sets the most insignificant features to zero by minimizing the usual sum of squared errors via the regularization parameter lambda.[20] Whereas subset selection methods can produce different models from small data changes and result in lower prediction accuracy, Lasso is more robust to small changes in data by shrinking coefficients toward zero which produces more stable models.[21] Additional advantages of LASSO over other variable subset selection methods such as stepwise Forward and backward propagation include faster computational speeds, and estimating predictor effects at the same time as variable selection.[20]

Model iterations were trained using varying lambda values, with lambda values closer to zero usually yielding models with more predictors than models with larger lambda values.[22] Three models were trained on ten-week periods starting from the beginning of each of the three years (weeks 1–10; 53–62; 105–114) to assess how the models might change over time and to ascertain how well a short training period can predict a longer period of time. Parameters for each of the three models included training by method using Maximum Likelihood, standardized equal to true, and grouping by both state and week. The three models were each tuned by iterating between lambda parameter values of 1 to 100 and selecting the lambda associated with the model with the lowest the Bayesian Information Criterion (BIC), which yields better fitting models.[23] The lambda parameter values with lowest BIC for the 2012, 2013, and 2014 models were 56, 53, and 98, respectively. The two models for 2012 and 2013 were assessed by validating on the 52 weeks following the

training weeks. Due to limited data, the model for 2014 was assessed by validating on 41 weeks following the training weeks.

Normalized residuals plots for the initial trained models displayed heteroscedasticity and bias, which was improved by transforming the dependent variable (P&S syphilis cases) using the natural log.[24,25] Applying the natural log transformation and retraining the models with LMM-LASSO produced more robust models due to relatively improved residual plots. Natural log transformation in regression introduced statistical bias, thus bias correction was applied as part of the back-transformation.[26] The predicted values were then back-transformed using the exponential function and bias correction. To assess prediction errors, R^2 and Root Mean Square Error (RMSE) were then calculated after completing back-transformation with bias correction. R^2 was calculated by calculating a regression line between the actual and predicted P&S syphilis counts. R^2 values range from 0 to 1 where a number closer to one is good. RMSE is calculated by taking the square root of the mean squared difference between actual and prediction P&S syphilis counts. RMSE values can range from 0 to infinity where a number closer to 0 is considered good.

Role of the funding source

The funder did not participate in the gathering, analysis, or interpretation of the data detailed in the article.

Results

Table 1 displays the R^2 (average and standard deviation) and RMSE (average and standard deviation) for the three validation periods and overall period used to assess each of the three natural log transformed best fitting LMM-LASSO models. The 2012 model, validated for weeks 11 to 62, was tuned to minimize BIC using a lambda value of 56. The model identified two predictors, which included two sexual risk–related keywords. The average R^2 between the predicted and actual count of P&S cases for the 52 validation weeks was 0·896 (SD = 0·11), and the RMSE average was 4·21 (SD = 1·98).

The model for 2013, validated on weeks 63 to 114, resulted in a minimized BIC at a lambda value of 53 and identified two drug related keywords and one sexual risk–related keyword. The validation period for this model resulted in an average R^2 of 0.89 (SD = 0·11); the RMSE average was 6.12 (SD = 2·86). The minimized BIC 2014 model, after validating on weeks 115 to 155 with lambda value of 98 identified three predictors, which included two sexual risk–related keywords and one drug related keyword. This model resulted in an average R^2 of 0.92 (SD = 0·03) and RMSE average of 4·23 (SD = 0·83).

The week-by-week R^2 values of predicted and observed P&S syphilis counts are displayed in Figure 1. The overall average R^2 was 0.90 for the three years (weeks 11–155). The R^2 varied by week, with a relatively small standard deviation of 0·094. The weekly RMSE between the predicted and observed P&S syphilis counts is displayed in Figure 2. The overall average RMSE throughout weeks 11 to 155 was 4.90. RMSE also varied each week, with a standard deviation of 2.31. Weekly P&S syphilis counts by state ranged from 0 to 151 which indicates that the average is small RMSE in comparison, however the percent

difference between predicted and observed values may be greater when predicting states with lower values of P&S syphilis counts.

Two points coinciding with the New Year's holiday (weeks 53 and 105), in comparison with other weeks, resulted in a very low R^2 (> 0·2) and relatively high RMSE. Other holidays did not produce similar spikes in R^2 nor RMSE values.

Discussion

Findings suggest that internet search data from Google Trends can be used to predict cases of syphilis. After identifying a method that used Google Trends data to predict weekly P&S syphilis counts, we applied that model to a new set of syphilis data and found it predicted cases with high accuracy. The relatively high average R^2 and relatively low R^2 standard deviation indicate that an association is likely to exist between Google Trends internet search data and P&S syphilis case counts. Although relatively low, the overall average RMSE resulting from this study has the potential to be lowered further with future work to better predict P&S syphilis cases. The RMSE standard deviation indicates that future work may further improve week-by-week error by reducing variability, resulting in better predictions. Based on the average R^2 and RMSE, similar to the influenza studies, it is possible to predict in advance not only one or two weekly P&S syphilis counts, but up to the following 52 weeks for each state using Google Trends internet search data.

Our findings suggest that Google Trends data could be a new data source for researchers and organizations focused on addressing syphilis. In the U.S., where syphilis is nationally notifiable, case reports of P&S syphilis reflect diagnosed infections and are useful to inform disease intervention (e.g., ensuring patient treatment and partner notification). Supplementing case report data with extant data sources like Google Trends data may help enhance current surveillance by forecasting changes in the epidemic and informing targeted awareness campaigns. In parts of the world without routine syphilis case reporting, social media data may useful as a surveillance platform to help monitor syphilis trends and inform prevention and control activities.

Freely available internet search data provided by Google Trends were used for this study. We used search query behavioral data based on the intuition that people who are at-risk or might have recently contracted syphilis would be likely to search for syphilis-related information online prior to diagnosis. Researchers and public health practitioners can further explore this psychology by accessing and analyzing recent Google Trends data, which could assist with predicting regions of concern for P&S syphilis. Additional research on this topic could help to determine how organizations might use Google Trends data as a tool for prediction and targeted interventions. For example, researchers could explore the psychology behind why people search for syphilis and sexual health information, such as whether they search primarily for information related to experiencing potential symptoms, had heard of an outbreak on the news, or read other information to prompt the search. Additionally, as keyword predictors varied over time, additional research could investigate how searches may relate to current behavioral trends related to syphilis transmission (e.g., the role of drug use in sexual networks).

This study has several limiting factors. One major limitation is that the models were trained on reported cases of P&S syphilis that likely underestimate true burden of disease.[27] That is, since the current model is designed to predict reported cases, and reported cases underrepresent actual infections, it is likely that the current predictive model underestimates true syphilis infections by not accounting for unreported cases. Another limitation is that the participant sample was biased toward internet users who use the Google search engine. However, this bias is mitigated by the fact that as of March 2015, Google accounted for almost two-thirds (64·4%) of all internet search traffic, whereas the next most popular search engine accounted for only 20·1% of traffic during a given month.[28] It is also not known how many syphilis cases used the internet to seek health information prior to visiting a health care provider, and it is possible that search engine data of internet users may not be representative of syphilis cases. Another limitation is that Google Trends search data are only comprehensive at the state or country level. As a result, it is difficult to train P&S syphilis models on smaller regions (e.g., county, city, or zip code), which would be useful for allocating limited resources to areas with the highest infection rates. Another similar limitation is that Google Trends search data are provided pre-normalized (based on geography and time) and pre-scaled (0–100) which prevents aggregating state-level data to perform analysis by regions or urban vs rural areas. Furthermore, events and holidays may not be accurately modeled. The results indicate that it is possible that New Year's week altered users' internet search behavior patterns and/or how P&S syphilis cases were recorded. In addition, the data used in this study did not contain sub-group information. This topic could be explored in future work. Finally, Google Trends data are provided as a random sample of all data from Google. Although this is a large random sample of data from people's searches across the United States, as with other random samples it is unknown whether and how models might change with the full dataset.

Conclusion

We found an association between freely available internet search query data and weekly reported cases of P&S syphilis. Results suggest the need for further exploration on whether and how internet search data can be integrated into public health monitoring systems for sexually transmitted infections. In particular, these methods may be useful in areas outside of the U.S. without access to routine, case-based syphilis surveillance. It is our hope that the approach outlined in this paper, along with the collaboration between academic researchers and the CDC, will encourage additional work investigating the use of internet technologies for health-related research and partnerships between researchers and public health organizations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding: This work was supported by support from the National Institute of Mental Health (NIMH) grant 5R01MH106415 (Young), the National Institute of Allergy and Infectious Diseases grant R56 (Young), and the University of California Office of the President (UCOP).

References

1. Centers for Disease Control and Prevention. [Accessed November 22, 2016] 2016 Nationally Notifiable Conditions. https://wwwn.cdc.gov/nndss/conditions/notifiable/2016/
2. Centers for Disease Control and Prevention. [Accessed December 5, 2016] 2015 STD Surveillance | CDC. https://www.cdc.gov/std/stats15/
3. Centers for Disease Control and Prevention. Recommendations for Public Health Surveillance of Syphilis in the United States. Atlanta, GA: 2003. https://www.cdc.gov/std/syphsurvreco.pdf
4. Centers for Disease Control and Prevention. [Accessed December 5, 2016] CDC Social Media Tools, Guidelines & Best Practices | Social Media | CDC. http://www.cdc.gov/socialmedia/tools/guidelines/index.html
5. Fox, S., Duggan, M. [Accessed December 5, 2016] Health Online 2013. Pew Res Cent Internet Sci Tech. Jan, 2013. http://www.pewinternet.org/2013/01/15/health-online-2013/
6. Weaver JB, Mays D, Weaver SS, Hopkins GL, Eroglu D, Bernhardt JM. Health information-seeking behaviors, health indicators, and health risks. Am J Public Health. 2010; 100(8):1520–1525. DOI: 10.2105/AJPH.2009.180521 [PubMed: 20558794]
7. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009; 457(7232):1012–1014. DOI: 10.1038/nature07634 [PubMed: 19020500]
8. Santillana M, Zhang DW, Althouse BM, Ayers JW. What can digital disease detection learn from (an external revision to) Google Flu Trends? Am J Prev Med. 2014; 47(3):341–347. DOI: 10.1016/j.amepre.2014.05.020 [PubMed: 24997572]
9. Preis T, Moat HS. Adaptive nowcasting of influenza outbreaks using Google searches. R Soc Open Sci. 2014; 1(2):140095.doi: 10.1098/rsos.140095 [PubMed: 26064532]
10. Desai R, Hall AJ, Lopman BA, et al. Norovirus disease surveillance using Google Internet query share data. Clin Infect Dis Off Publ Infect Dis Soc Am. 2012; 55(8):e75–78. DOI: 10.1093/cid/cis579
11. Jena AB, Karaca-Mandic P, Weaver L, Seabury SA. Predicting new diagnoses of HIV infection using internet search engine data. Clin Infect Dis Off Publ Infect Dis Soc Am. 2013; 56(9):1352–1353. DOI: 10.1093/cid/cit022
12. Desai R, Lopman BA, Shimshoni Y, Harris JP, Patel MM, Parashar UD. Use of Internet search data to monitor impact of rotavirus vaccination in the United States. Clin Infect Dis Off Publ Infect Dis Soc Am. 2012; 54(9):e115–118. DOI: 10.1093/cid/cis121
13. Milinovich GJ, Avril SMR, Clements ACA, Brownstein JS, Tong S, Hu W. Using internet search queries for infectious disease surveillance: screening diseases for suitability. BMC Infect Dis. 2014; 14:690.doi: 10.1186/s12879-014-0690-1 [PubMed: 25551277]
14. Johnson AK, Mehta SD. A comparison of Internet search trends and sexually transmitted infection rates using Google trends. Sex Transm Dis. 2014; 41(1):61–63. DOI: 10.1097/OLQ.0000000000000065 [PubMed: 24326584]
15. Johnson AK, Mikati T, Mehta SD. Examining the themes of STD-related Internet searches to increase specificity of disease forecasting using Internet search terms. Sci Rep. 2016; 6:36503.doi: 10.1038/srep36503 [PubMed: 27827386]
16. Polgreen PM, Chen Y, Pennock DM, Nelson FD. Using internet searches for influenza surveillance. Clin Infect Dis Off Publ Infect Dis Soc Am. 2008; 47(11):1443–1448. DOI: 10.1086/593098
17. Santillana M, Nsoesie EO, Mekaru SR, Scales D, Brownstein JS. Using clinicians' search query data to monitor influenza epidemics. Clin Infect Dis Off Publ Infect Dis Soc Am. 2014; 59(10):1446–1450. DOI: 10.1093/cid/ciu647
18. Jernigan C, Mistree BFT. Gaydar. Facebook friendships expose sexual orientation. First Monday. 2009; 14(10) doi:http://dx.doi.org/10.5210/fm.v14i10.2611.
19. Baayen RH, Davidson DJ, Bates DM. Mixed-effects modeling with crossed random effects for subjects and items. J Mem Lang. 2008; 59(4):390–412. DOI: 10.1016/j.jml.2007.12.005
20. Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. J R Stat Soc Ser B Stat Methodol. 2011; 73(3):273–282. DOI: 10.1111/j.1467-9868.2011.00771.x

21. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. J R Stat Soc Ser B In Journal of the Royal Statistical Society. 1994; (58):267–288.

22. Groll, A., Tutz, G. [Accessed December 5, 2016] Variable selection for generalized linear mixed models by L1-penalized estimation | SpringerLink. http://link.springer.com/article/10.1007/s11222-012-9359-z

23. Müller S, Scealy JL, Welsh AH. Model Selection in Linear Mixed Models. Stat Sci. 2013; 28(2): 135–167. DOI: 10.1214/12-STS410

24. Jacqmin-Gadda H, Sibillot S, Proust C, Molina J-M, Thiébaut R. Robustness of the linear mixed model to misspecied error distribution. ResearchGate. 2007; 51(10):5142–5154. DOI: 10.1016/j.csda.2006.05.021

25. Park T, Lee S-Y. Model Diagnostic Plots for Repeated Measures Data. Biom J. 2004; 46(4):441–452. DOI: 10.1002/bimj.200210044

26. Newman, M. [Accessed December 5, 2016] Regression analysis of log-transformed data: Statistical bias and its correction - Newman-- 1993 - Environmental Toxicology and Chemistry - Wiley Online Library. http://onlinelibrary.wiley.com/doi/10.1002/etc.5620120618/abstract

27. Satterwhite CL, Torrone E, Meites E, et al. Sexually transmitted infections among US women and men: prevalence and incidence estimates, 2008. Sex Transm Dis. 2013; 40(3):187–193. DOI: 10.1097/OLQ.0b013e318286bb53 [PubMed: 23403598]

28. comScore. [Accessed December 5, 2016] comScore Releases March 2015 U.S. Desktop Search Engine Rankings - comScore, Inc. http://www.comscore.com/Insights/Rankings/comScore-Releases-March-2015-US-Desktop-Search-Engine-Rankings

Research in context

Evidence before this study

Previous research detected a significant positive relationship between risk-related keywords in social media (i.e., Twitter) postings and HIV prevalence. This study aimed to determine whether the same evaluation process could be used with online search query (e.g., Google) data to predict weekly state-level P&S syphilis case reports.

Added value of this study

Internet technologies have led to faster, easier identification of disease incidence, but the evidence for widespread implementation of these technologies to predict disease is still being gathered. This study shows that freely available internet search data can identify areas of elevated peak activity for syphilis and may be able to predict future cases.

Implications of all the available evidence

Developed countries have established effective reporting mechanisms, yet “surprise” regional changes in trends of infectious diseases such as HIV and influenza have occurred during the past few years. Partnerships between researchers and health departments could be used to explore the potential of social media in forecasting syphilis trends. Automated technologies to handle large amounts of data have been developed, but methods to disseminate information about health events to regional health administrators must be improved and require the assistance of health departments to do so. Many issues prohibit the formal integration of internet search behavior results into official surveillance systems—privacy and the role of media influence, among them—but this study provides evidence that internet search behavioral data might be an effective tool for public health surveillance, especially in global low-income areas without routine, case-based syphilis surveillance.

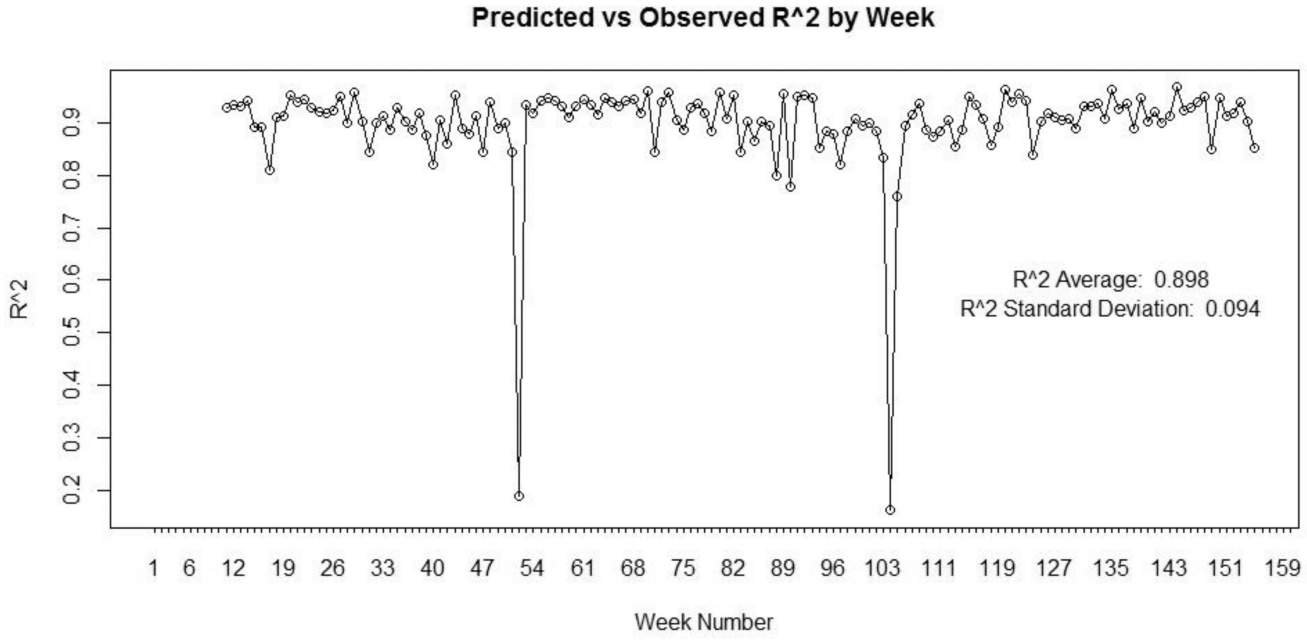

Figure 1.
Weekly R^2 between the predicted and observed P&S syphilis count for 2012 to 2014.

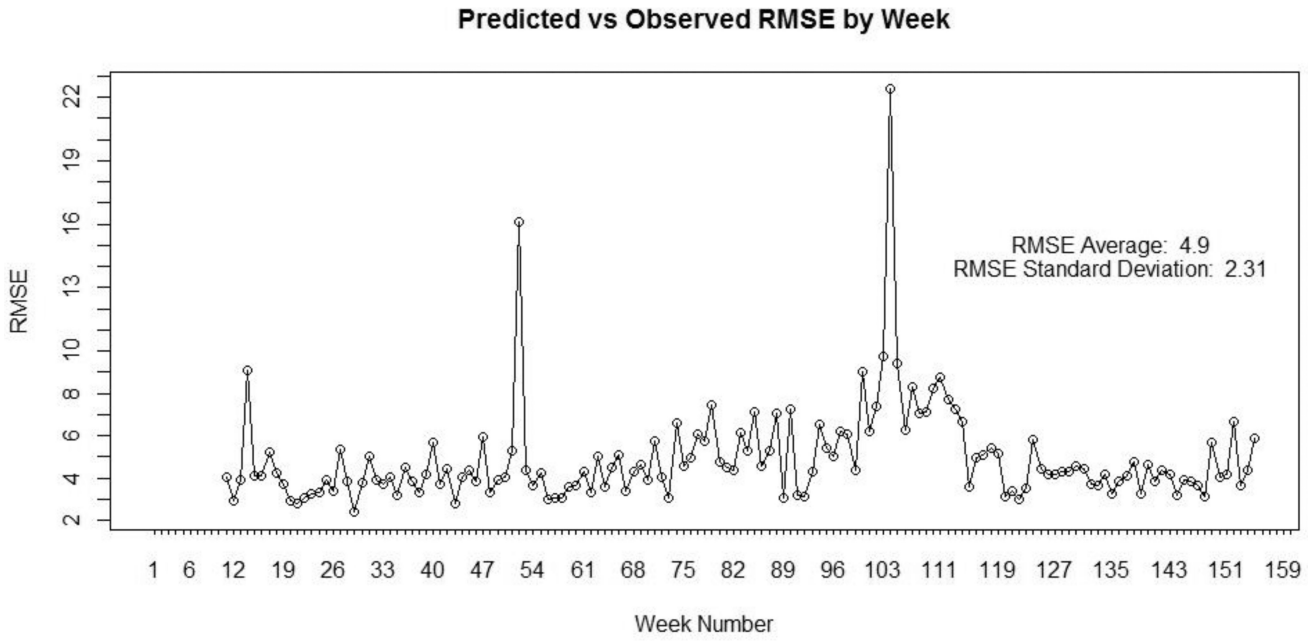

Figure 2.
Weekly RMSE between the model predicted and observed P&S syphilis count for 2012 to 2014.

Table 1

Natural log transformed best fitting LMM-LASSO model R^2 and RMSE for the three validation periods and overall period (weeks 11–155).

Model Year	2012		2013		2014		Overall	
Validation Period	Weeks 11–62		Weeks 63–114		Weeks 115–155		Weeks 11–155	
Model Fit Measure	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
Average	0·896	4·21	0·885	6·12	0·918	4·23	0·898	4·90
Standard Deviation	0·106	1·98	0·112	2·86	0·031	0·838	0·094	2·31