Applying Aberration Detection Algorithms to Live Public Health Data:

Lessons from National Syphilis Case Surveillance Data

Erika Martin, PhD, MPH[1,2]; John Angles, MPH[3]; Tracy Pondo, MSPH[4]; Melissa Pagaoa, MPH[4]; Elizabeth Torrone, PhD, MSPH[4]

This work was supported by the Centers for Disease Control and Prevention, National Center for HIV, Viral Hepatitis, STD, and TB Prevention Epidemiological and Economic Modeling Agreement (no. 5U38PS004650). The findings and conclusions are solely the responsibility of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention or the Department of Health and Human Services.

[1] Public Health Accreditation Board

[2] Rockefeller College of Public Affairs and Policy, University at Albany, State University of New York

[3] School of Public Health, University at Albany, State University of New York

[4] Surveillance and Data Science Branch, Division of STD Prevention, Centers for Disease Control and Prevention

Table of Contents

Introduction

Rates of syphilitic infections in the United States have risen dramatically, with a 608% increase in the rate of reported cases of primary and secondary (P&S) syphilis – the most infectious stage – in the 20-year period from 2003 to 2022. Concurrent with the growing incidence among women of reproductive age, there has been a 769% increase in the number of congenital syphilis cases in the same period. Additionally, there are persistent racial and ethnic disparities: in 2022, non-Hispanic American Indian or Alaskan Native and non-Hispanic Black or African American persons had P&S syphilis rates of 67.0 and 44.4 per 100,000, respectively; in comparison, White non-Hispanic persons had a P&S syphilis rate of 10.2 per 100,000.[i,ii]

Timely, accurate, and complete surveillance data are needed to monitor trends in morbidity and disparities, identify hot spots for outbreak response, and deploy public health resources more efficiently. Beyond the delivery of public health services, high quality data are needed to assess progress towards the Sexually Transmitted Infections National Strategic Plan[iii] and other state or local initiatives, support program evaluation and research, and ensure fair distributions of resources in formula-based allocations that rely on morbidity. More broadly, beyond syphilis surveillance data, the COVID-19 pandemic highlighted the need for more robust and responsive public health data systems. That gap catalyzed the Data Modernization Initiative[iv] and the Public Health Data Strategy,[v] which aim to modernize public health data and surveillance at the federal, state, and local levels.

This report provides public health programs with strategies for implementing data aberration detection algorithms to identify meaningful changes in syphilis case surveillance data that indicate potential outbreaks or actionable data quality issues. The target audience includes state, Tribal, local, and territorial (STLT) health departments and Centers, Institutions, and Offices across the Centers for Disease Control and Prevention (CDC). Even if the Federal vision for a modernized public health data system with more complete, timely, and rapidly exchanged information is achieved, there remains a need for public health agency staff to review data routinely to identify shifts in morbidity or data quality issues that can be corrected more quickly before data are finalized. Although the concepts and approaches described here are specific to national case syphilis surveillance data from the National Notifiable Diseases Surveillance System (NNDSS), they are applicable to state and local data as well as other notifiable diseases.

The Syphilis Case Surveillance Data Lifecycle

Implementing data aberration detection algorithms – including identifying specific aberrations to prioritize for detection – first requires understanding the data production lifecycle and differences across jurisdictions that may influence the timeliness, completeness, and accuracy of the data.

Case surveillance data acquisition, preparation, and maintenance. NNDSS is part of a complex data ecosystem comprising partners from STLT health departments, community partners, and Federal agencies (Exhibit 1, Exhibit 2).

Case surveillance for syphilis usually begins when there is laboratory or clinical evidence that an individual has a syphilitic infection, often a reactive serologic test. Because syphilis is a reportable condition in all 50 states and the District of Columbia, laboratories and providers are required by state law, regulation, or statue to report results to their local or state public health authority. These case reports include identifiable information on the patient so that the public health authority can take action to investigate, including assuring treatment and disease investigation, including partner notification. Then, using the laboratory and clinical information provided and data collected during the case investigation, surveillance staff in the STLT health department determine if the infection meets the Council of State and Territorial Epidemiologist (CSTE) surveillance syphilis case definition.[vi]

To facilitate case identification and reporting, STLT health departments enter and store information from syphilis case reports and disease investigations in their public health information systems (PHIS). Jurisdictions choose which PHIS they use, including commercial disease surveillance software platforms as well as "homegrown" or custom systems.[vii]

To support national syphilis surveillance, STLT health departments from all states, territories, the District of Columbia, and New York City transmit de-identified syphilis case notifications from their PHIS to CDC at least weekly. Syphilis case notifications are based on the case information available at the STLT health department and are sent to CDC using data transmission standards provided by NNDSS. Although all PHIS can transmit case notifications for nationally notifiable conditions to CDC, the use of different PHISs across jurisdictions increases the complexity of standardizing data in NNDSS.

At CDC, the Office of Public Health Data, Surveillance, and Technology (OPHDST) receives, processes, and stores case notifications received through NNDSS.[viii] OPHDST provides case data for nationally notifiable sexually transmitted infections (STIs), including syphilis, to the Division of STD Prevention (DSTDP), which reviews data throughout the year, working with STLT health departments to conduct ongoing quality assurance and improvement activities. This review and revision process for syphilis cases provided through NNDSS occurs throughout the year and then is iterated multiple times during a data reconciliation period, resulting in a finalized, annual data set used in national reporting and evaluation.

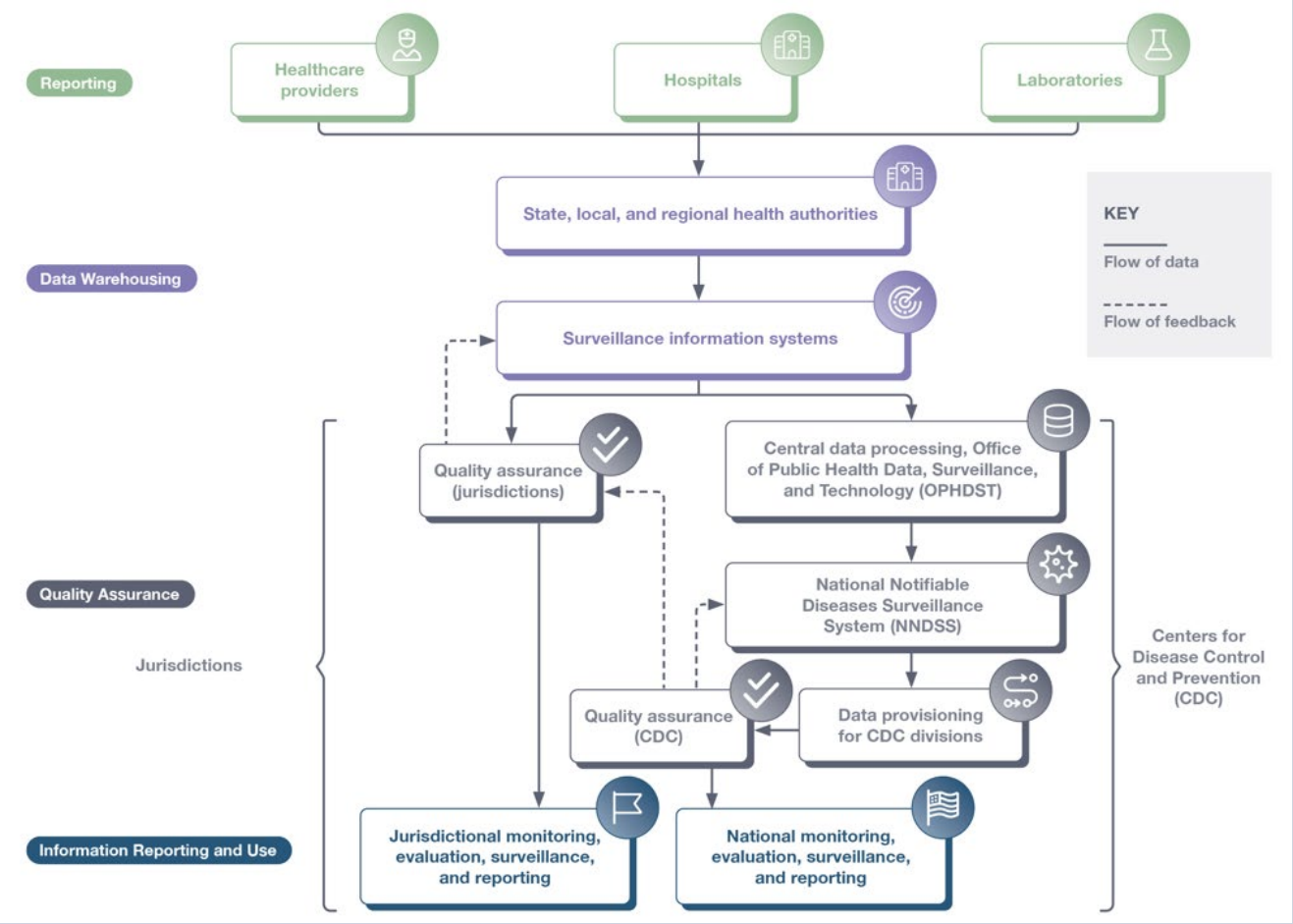
Of note, with each data transmission, jurisdictions not only add notifications for newly identified cases, but they can revise previously submitted case notifications. For example, they may revise values in data fields such as race/ethnicity or remove previously submitted case notifications that were determined to be duplicates. Therefore, prior to annual reconciliation, live partial year data are not only incomplete, but they may on occasion include overreporting (i.e., if there are duplicate case notifications that are later reconciled) or demographic characteristics that are later revised.

Adding to the complexities of interpreting live data, patterns of incompleteness may vary throughout the year.

Although the data undergo multiple steps before annual reconciliation, which requires considerable resources at the national and STLT level, the work is essential to ensure that STI surveillance data are of high quality and thereby actionable for STLT and community partners.

Exhibit 1. Data ecosystem stakeholders

Exhibit 2. Flow of information in surveillance information systems for syphilis



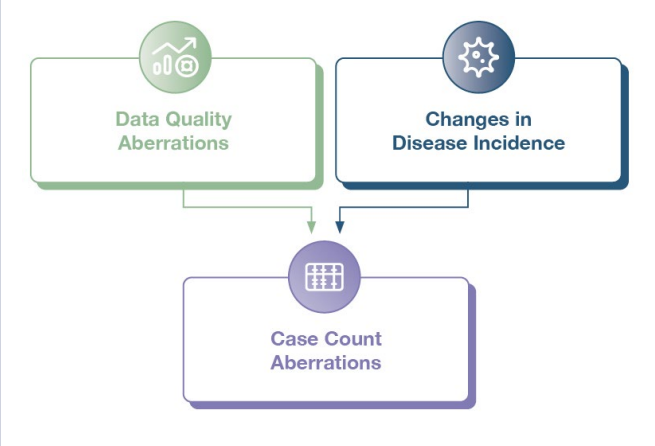
Reproduced from Martin & Angles, 2023 [xii]

Factors that influence the timeliness, completeness, and accuracy of underlying data. With numerous entities collecting and transmitting data to the STLT health departments that subsequently provide syphilis case notification data to CDC through NNDSS, there are substantial differences across jurisdictions that can lead to different data quality considerations. As of December 2023, there are six different PHISs (NBS, PRISM HDS, Clinisys, Maven, STD*MIS, and EpiTrax), along with many additional customized systems, in use by the STLT health departments who send STI case notification data to CDC.[ix] Further, jurisdictions may use different data transmission standards to provide data to CDC. Some have onboarded the Health Level 7 (HL7) message mapping guides,[x] while others still provide data using the National Electronic Telecommunications System for Surveillance (NETSS) standard.[xi] Public health governance structures, and relationships among state, local, and regional health departments, can vary widely, leading to differences in the way case information is initialized in the jurisdiction's PHIS.[xii] Additionally, within a state, local health departments may use a different PHIS from the state, resulting in additional challenges to standardizing data. The number of public health employees per capita also varies widely across reporting agencies, [xiii] potentially resulting in challenges in case investigations for understaffed locations.

Aberration Detection Algorithms for Public Health Surveillance Data

An "aberration" refers to changes in an event, such as the number of observed cases per month, that are significantly different than what is expected based on history.[xiv] Aberrations and outbreaks are related but do not necessarily overlap (Exhibit 3). An ***outbreak investigation*** focuses on unexpected changes in morbidity while an ***aberration detection*** is broader and additionally includes the assessment of anomalies that indicate data quality concerns. Identifying and interpreting aberrations requires both quantitative assessment and qualitative insights based on contextual information; for example, slow-growing outbreaks might not be flagged as an aberration and a sudden increase in cases might represent an anomaly such as a screening campaign, entering a backlog of cases into the PHIS, or random noise.

Exhibit 3. Types of case count aberrations



There are multiple potential statistical approaches to aberration detection. These are summarized below, with a comparison of their strengths and weaknesses in Exhibit 4. Although some methods are stronger with respect to improved sensitivity and specificity to detect aberrations, the selection of an appropriate method depends on the context, ease of implementation, and the target audience. From a statistical standpoint, methods have varying suitability depending on the presence of strong temporal or seasonal trends, whether the condition has high or low morbidity, and the extent to which counts follow an approximately normal distribution. There are also trade-offs between the most suitable statistical approach and ease of interpretation; for example, advanced algorithms might have the highest sensitivity and specificity in certain circumstances, but they require specialized software, technical expertise, and additional computational time. Furthermore, they are more difficult to explain to diverse audiences, which might influence end users' trust in the findings.

Exhibit 4. Comparison of Selected Aberration Detection Methodologies for Use in Case Surveillance[xv]

Methodology	Strengths	Weaknesses
Linear or log-linear regression	• Likely most familiar to audiences • Intuitive to explain • Simple to implement • Does not require specialized data analytic software	• Likely to violate ordinary least squares assumptions such as autocorrelations among observations in a time series • Difficult to interpret outcomes that are not normally distributed such as in low-morbidity jurisdictions or where seasonal trends are present
Historical limits	• Intuitive to explain • Comparing the time period to the same weeks in prior years accounts for seasonality • Credibility established through well-known past use by CDC for aberration detection	• More appropriate for weekly data • Does not account for long-term trends potentially leading to artificially high or low thresholds • Assumes a normal distribution
Pseudo CUSUMs	• Intuitive to explain • Easy to implement • Requires the least amount of training data • Easy to modify key parameters (rolling window of historical data and number of standard deviations for the threshold)	• Key parameters are arbitrary and results are sensitive to those analytic decisions • Jurisdictions with varying morbidity may require different values of key parameters • Negative predicted values might be generated; log transformations or manually setting negative expected values to zero addresses the problem but may be difficult to interpret by users
ARIMA	• Sound statistical approach reflecting best practices for time series • Existing "auto.arima" package in R automatically selects the most appropriate model within the ARIMA class, avoiding data analysts needing to manually parameterize models for each jurisdiction • Allowing jurisdictions to have unique parameters addresses variable morbidity and epidemic trends	• Difficult to explain to non-technical audiences • Negative predicted values might be generated; manually setting negative expected values to zero addresses the problem but may be difficult to interpret by users • Appropriate for continuous data as opposed to right skewed positive integers, making model performance questionable for smaller case counts
Time series generalized linear models	• Sound statistical approach reflecting best practices for time series of count values • Designed for discrete distributions, resulting in no negative predictive values • More suitable than other regression-based approaches for low-morbidity jurisdictions	• Difficult to explain to non-technical audiences • Until an automated approach is developed (e.g., R's "auto.arima" package), data analysts need to make decisions on appropriate parameters including how to adjust them for jurisdictions with varying morbidity and historical trends
Bayesian methods	• Robust body of literature supporting these methods • Capable of multivariate modelling • Can give results in the form of a traditional hypothesis test	• Requires prior knowledge of the distribution of cases. This can be circumvented by using a non-informative prior, but this reduces the utility of a Bayesian approach • Difficult to explain to non-technical audiences
Hidden Markov models	• Incorporates the strengths of other Bayesian methods • Independence assumptions often more flexible than other time series models	• Computationally very intensive • Requires prior knowledge of case distribution and some knowledge of epidemic distribution • Difficult to explain to non-technical audiences
Machine learning methods	• Many methods to choose from with a wide variety of complexities • Many algorithms already available in statistical software packages	• The lack of strict training data makes it impossible to use most machine learning aberration detection methods • Unsupervised models offer additional challenges to fine-tune and validate • Difficult to explain to non-technical audiences

Abbreviations: autoregressive integrated moving average (ARIMA), Centers for Disease Control and Prevention (CDC), pseudo cumulative sum (CUSUM)

Linear or log-linear regression: In this classical regression approach, a trend line is fitted to the data, selecting the trend that provides the smallest difference between observed and expected values. For linear regression, this trend line is determined based on ordinary least squares. The forecasted values are projected from the trend line, with the regression-based confidence interval used to flag aberrations. Techniques such as implementing a log transformation of the outcome can be applied for data that do not follow a normal distribution or meet other regression assumptions. The threshold for flagging aberrations is based on confidence intervals whose range can be modified by users (e.g., 95% or 90%). This statistical approach is commonly used for outbreak detection so is likely familiar to many audiences. Farrington et al., 1996 illustrate how this method can be used for outbreak detection for multiple infections using weekly infectious disease case data from the United Kingdom's Communicable Disease Surveillance Centre, which is analogous to NNDSS.[xvi]

Historical limits: This approach compares the current observed value to the average value for the same period across multiple years. This method is suitable for weekly data. By comparing the time frame of interest to the same window in prior years (e.g., looking at *MMWR* weeks 4 through 8 across years), it explicitly accounts for seasonality. Stroup et al., 1993[xvii] describe an example for measles case reporting, in which the weekly average of a four-week period is compared to historical data from the prior five years. For each historical year, there are three observations to increase the sample size and address potential seasonal effects: the corresponding 4-week period, the preceding 4-week period, and the following 4-week period. The threshold for flagging aberrations is based on confidence intervals whose range can be modified by users (e.g., 95% or 90%). *MMWR* weekly surveillance reports from 1994 through 2017 used historical limits to set thresholds for notifiable case report counts in both their tables and figures displaying comparisons of the number of provisional cases compared to prior time periods.[xviii]

Pseudo-cumulative sum control chart (pseudo-CUSUMs): This technique uses a rolling average to provide an expected value. The expected range is determined using the rolling average plus or minus a few standard deviations prescribed by the analyst (e.g., three or five standard deviations). For example, this method could take the average of the prior three quarters to project the expected value in the fourth quarter. Observations outside the predetermined number of standard deviations are flagged as aberrations. The analyst can prescribe the number of past observations to include in the rolling average. Past examples include the Early Aberration Reporting System that was subsequently integrated into Epi Info to monitor for bioterrorist attacks[xix] and real-time infection disease surveillance at the Olympic village using daily hospital data.[xx]

Auto regressive integrative moving average (ARIMA) models: ARIMA models are a class of time series models that explicitly address autocorrelation and seasonality in time series data. Similar to classical regression models, prior values of a time series predict future values, with regression-based confidence intervals used to flag aberrations. Many analytic and forecasting methods require that time series data are stationary (i.e., absent of trends or seasonality), which is analogous to the independence requirement for cross-sectional data. Violations of these assumptions can yield a prediction that is biased or with incorrect confidence intervals, which could in turn result in observed values incorrectly flagged (or not flagged) as a potential aberration. The ARIMA class includes autoregressive (AR) models, moving average (MA) models, autoregressive moving-average (ARMA) models, and autoregressive integrated moving-average (ARIMA) models. The ARIMA class of models are commonly used in stock valuations and

forecasting in the financial sector. The New York State Department of Health previously used an ARIMA model for predictive modeling of early syphilis and gonorrhea case report data.[xxi]

Time series generalized linear models: These are similar to the ARIMA models described above, except that they use a generalized linear model through the use of a link function and a discrete distribution. These models are appropriate for data that do not follow a continuous distribution, such as counts.[xxii] At this time, R does not have an automated package for this class of models, so the analyst needs to specify the link function (identity or logarithmic), the distribution (Poisson or negative binomial), and the parameterization (the degree of the model and the number of prior observations to include). Compared to the auto ARIMA models described above, this approach avoids negative predicted values and would be appropriate for modeling time series with observations that do not follow a normal distribution or with low case counts.

More complex data aberration detection methods: The utility of other methods that are more technically complex has been demonstrated for disease surveillance and they could potentially prove more accurate than the simpler methods previously described.[xxiii] However, specific challenges with implementing these methods include computational time, required specialized expertise in areas of ongoing research, and difficulty in explaining to non-technical audiences. Bayesian methods are one increasingly popular approach for outbreak detection, as they are well suited to quantifying prediction uncertainty, particularly when the expected case distribution is known. One illustration of a Bayesian approach to outbreak detection is a study that applied this approach to influenza surveillance data for a single county in 2010.[xxiv] Hidden Markov models fall under the Bayesian framework and model each event in a system as a sequence of "hidden" processes, each with distinct probability distributions. An early example of using hidden Markov models for outbreak detection was a Bayesian Markov switch model applied to Spanish influenza data in 2008.[xxv] Machine learning and data mining are broad classes of methods that can be applied to time series to identify aberrations. For example, Kane et al., 2014 applied a machine learning categorization algorithm to Egyptian avian influenza data.[xxvi] In addition to the main challenges of the advanced methods previously described, machine learning methods need "training data" to "teach" and validate the models,[xxvii] and finding appropriately classified data for past outbreaks is challenging.

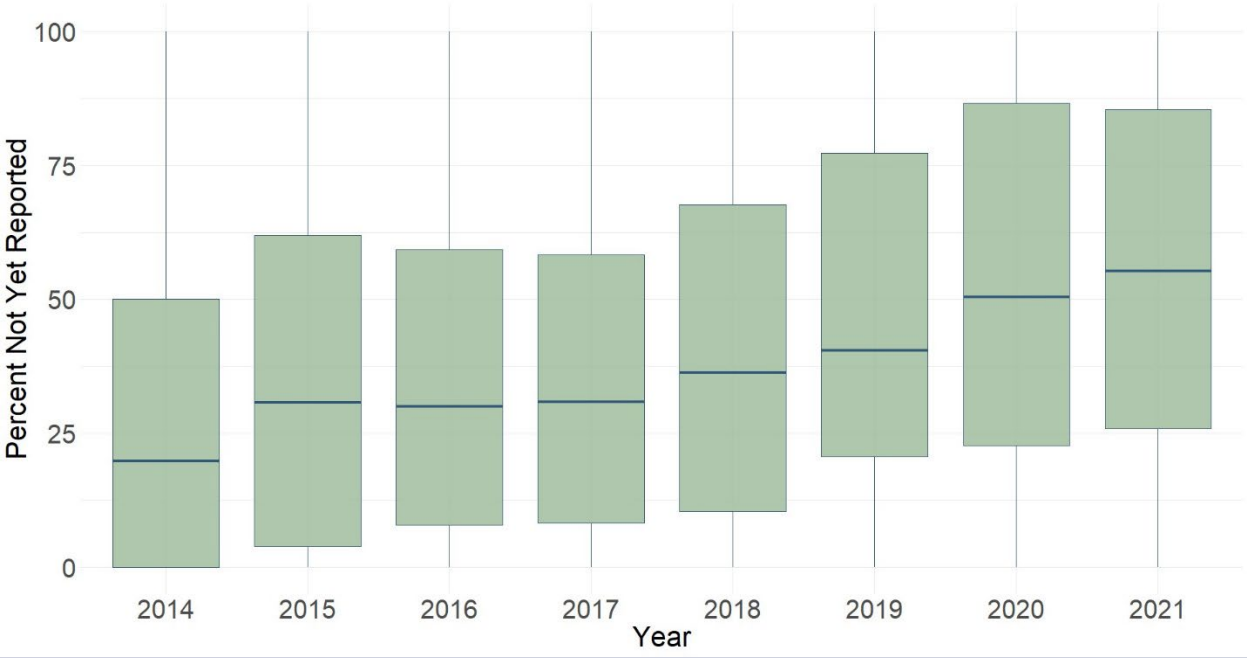
Challenges to Aberration Detection in Syphilis Surveillance Data

Establishing a clear definition for outbreaks. Conceptually and operationally, it is difficult to define a syphilis outbreak. Unlike foodborne outbreaks, STI outbreaks are usually slow-growing and do not have a discrete start date. In 2018, CSTE developed a comprehensive framework to help STI programs determine whether their data indicated a syphilis outbreak. Although the authors included syphilis outbreak definitions for sample states and localities, they did not explicitly recommend specific numerical thresholds.[xxviii] In part, this is because the syphilis epidemic varies across the United States. For example, in a county with low syphilis morbidity, an increase of just a few cases could be considered an outbreak, whereas some jurisdictions have much more substantial syphilis morbidity and a higher threshold may be warranted.

Currently, there is not a single system for jurisdictions to report syphilis outbreaks. CDC's Health Alert Network (HAN) is one way in which jurisdictions share information regarding outbreaks, and some jurisdictions have provided HANs for syphilis outbreaks.[xxix] Although this is a national network, participation is voluntary, jurisdictions determine what level of information is provided in the HAN, and some jurisdictions have their own HAN systems. For example, some syphilis-related HANs provide information about the specific counties or cities affected, while others are for the entire state. Further, HANs may provide information about increases in a specific stage of syphilis (e.g., increases in primary and secondary syphilis) while others just reference syphilis overall, and congenital syphilis outbreaks are occasionally included in the same HAN as a non-congenital syphilis outbreak.[xxx] These reporting differences limit the generalizability of using documented outbreak declaration, such as a HAN, to develop standardized thresholds.

Lag times in reporting and data processing. Syphilis case surveillance has not historically been in real-time because determination of whether an infection meets the CSTE case definition usually requires public health investigation, often including a medical chart review and a patient interview, which can take weeks (or longer). After a public health investigation is complete and surveillance staff determines that the CSTE syphilis case definition is met, data subsequently go through the information flow described in Exhibit 2. Given these processes, extensive delays from initial detection (e.g., initial laboratory testing) to inclusion in NNDSS might be expected; however, based on national data, there is evidence that the majority of syphilis cases are provided to CDC within two months of case identification date. For example, comparing the number of P&S syphilis cases in the live NNDSS weekly data (i.e., partial year data that are not yet finalized) to the number of cases recorded for that period in the annual, finalized data file, the estimated median percentage of cases recorded in the live data within two months of the reported case identification date was 56.1% (interquartile range: 26.3%, 81.8%) during 2014-2021. However, there is considerable variability across jurisdictions in the percent of cases not yet reported after two months in the live weekly data, indicated by the width of the box-and-whiskers plots in Exhibit 5. Additionally, while the range in the percent of cases not yet reported has been somewhat consistent over time (based on the width of the boxes), the level of delayed reporting (based on the medians) has increased.

Exhibit 5. Percent of P&S syphilis cases not yet reported after two months in live weekly files, 2014-2021



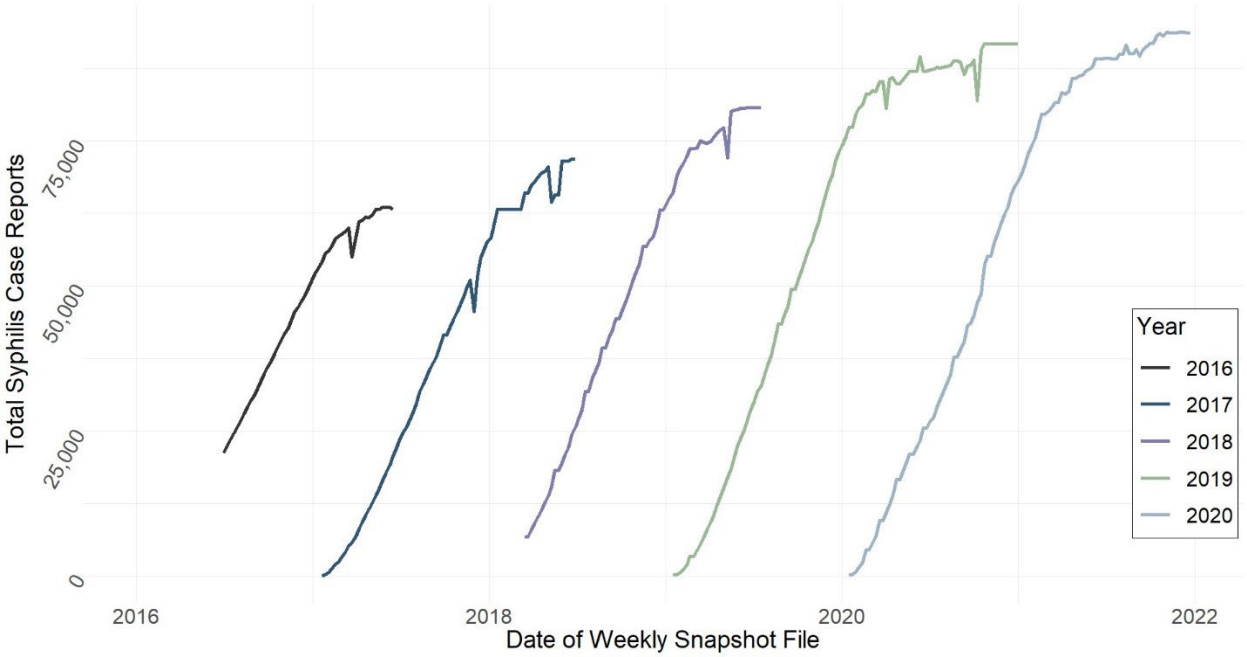
Notes: Data are from the National Notifiable Diseases Surveillance System weekly files and include cases from 50 states and the District of Columbia.[xxxi]

These reporting delays and other data anomalies can be seen in plots of cumulative case counts in live syphilis case notification data (Exhibit 6). In that chart, each line represents the total number of cases provided through NNDSS throughout the year based on information in the live weekly files. The last data point for each line is the value from full case reporting in the reconciled year-end file. Consistent with Exhibit 5, the width of each line (i.e., the number of months from the first recorded case to the last recorded case in the final file with fully reconciled, complete data) increases throughout the period and the upper tail of each curve is notably flatter. That reflects an increasing time to data closeout. The longer time to complete case reporting influences perceptions of morbidity at a given point during the live surveillance year. Additionally, although some spikes in the data might be expected due to disease outbreaks, each curve displays some dips and sudden increases that are not likely attributable solely to outbreaks and other changes in morbidity. The timing of these fluctuations is inconsistent across years; for example, in 2017, there was a sudden drop about halfway through the year. That aberration is likely a data artifact because the trend lines are cumulative case counts and values would be expected to increase or stay the same over the year. This unpredictability makes it difficult to know in real-time whether a sudden shift in the number of cases is the result of a true change in morbidity or a data artifact.

These sudden drops or increases in national data are typically due to changes in a single high-morbidity jurisdiction and sudden shifts in low-morbidity are not likely to be reflected in the national data. Examining trends within each jurisdiction is important for a complete understanding of data

aberrations because reported cases from high-morbidity jurisdictions may mask changes in other areas. However, such examination is time-consuming. An additional challenge with understanding these trends is that incomplete case notification in the live data is likely not missing at random; for example, case investigation may be timelier for persons diagnosed in public STI clinics or persons who have had a prior syphilitic infection.

Exhibit 6. Lag times and fluctuations in reporting and data processing for live syphilis case notification data, 2016-2020



Note: Data are from the National Notifiable Diseases Surveillance System and include all stages of syphilis.[xxxii]

There are several contributing causes of reporting lag times beyond two months and delays in data closeout. One potential explanation for increased lag times is the large increase in P&S syphilis cases, which has increased the workload of Disease Intervention Specialists (DIS) and surveillance staff. This increase can be seen by the increasing height of the cumulative case reporting charts in Exhibit 6. A second potential explanation is the COVID-19 pandemic, which led jurisdictions to redirect DIS to work on COVID-19; this may have resulted in fewer staff available to investigate and provide partner services to P&S syphilis cases. If more senior DIS were redirected to COVID-19, less experienced staff might have taken longer to complete partner services investigations.

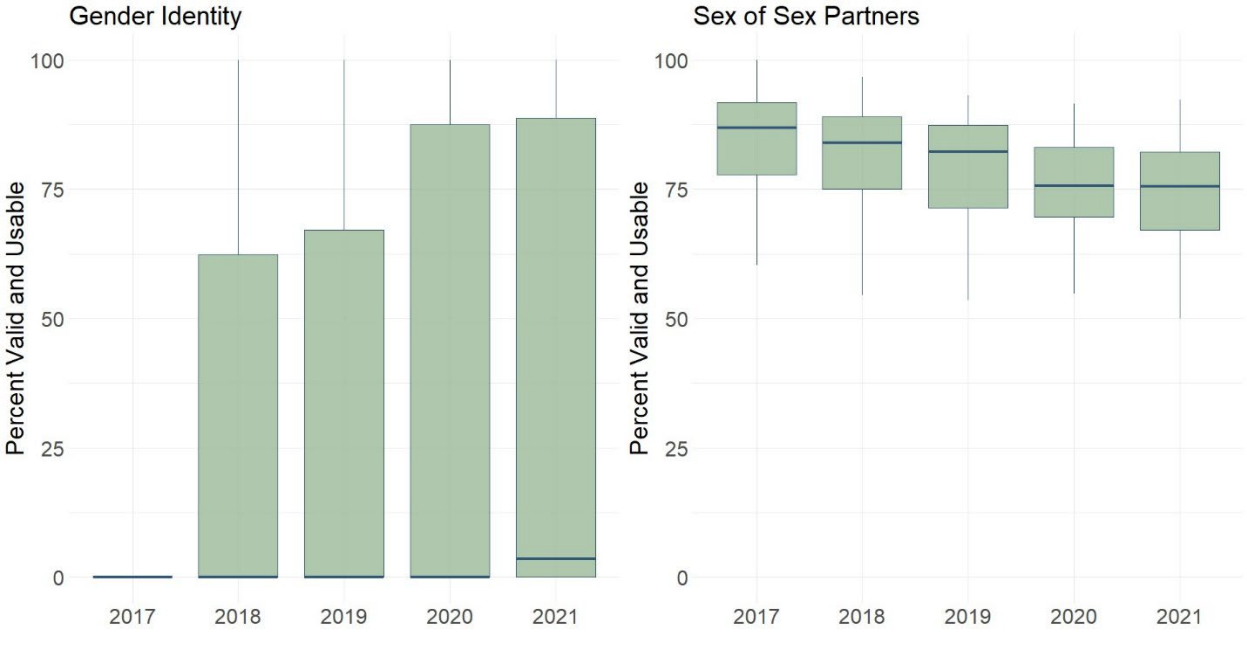
Batch uploads or bulk deletions of data to or from a jurisdiction's PHIS are common reasons for sudden dips and spikes in the NNDSS syphilis data. A jurisdiction may have months without data transmission if there are PHIS changes including

transitioning to a new system or changing data transmission standards (e.g., onboarding message mapping guides). Further, data system errors at the jurisdiction level might result in a temporary period where new cases are not transmitted; these technical issues may be compounded when there are changes in surveillance staff who need time to learn the system.

Incomplete data for demographics and other characteristics. The percent of cases with valid and usable values for core demographic variables varies across jurisdictions and over time (Exhibit 7). This can lead to inconsistencies or gaps in the reporting of syphilis trends by subpopulation. To minimize bias due to missing data when displaying data in national surveillance reports, jurisdictions must meet a criterion of at least 70% of their cases having valid and usable data for a given variable.[xxxiii] For example, in the most recent national STI surveillance report, only 37 states and the District of Columbia had their data included in the analyses of P&S syphilis among men who have sex with men and 23 states had their data included in the analyses of P&S syphilis by gender identity.

The percentage of usable and complete data may vary throughout the year, which poses additional challenges for releasing unreconciled live data. For example, some jurisdictions may wait until the end of the year to do a one-time match with eHARS to identify HIV status for syphilis cases. Additionally, jurisdictions might prioritize different variables for completeness and accuracy depending on the unique features of their local epidemics. A sudden decrease in the percent of cases with valid and usable values for a priority variable could be indicative of a data quality issue, a data transmission issue, or that a program needs support. Some PHIS are unable to store all the variables contained in NNDSS data transmission standards for syphilis (e.g., some PHIS do not capture gender identity), making it impossible for a jurisdiction using one of these PHIS to transmit information for all NNDSS variables. An additional complexity is that jurisdictions may have inconsistencies in the extent to which their response questions for demographic variables such as race and ethnicity align with NNDSS variables.

Exhibit 7. Percent of P&S syphilis cases with valid and usable values for gender identity and sex of sex partners, 2017-2021



Note: Data are from the National Notifiable Disease Surveillance System reconciled yearly files covering 50 states and the District of Columbia[xxxiv]

Complexity of date variables. Assessing data anomalies in time series data requires a clear determination of key dates in the case identification and notification lifecycle. However, this is complex for syphilis case notification data. CDC uses *MMWR* weeks to break up the calendar year. *MMWR* weeks start with the first week in January that has at least four days in the calendar year. Jurisdictions assign an *MMWR* week for each case notification, and concurrently can provide a variety of other dates, including the dates of disease onset, diagnosis, laboratory result, and first report to the state or community health center. CDC provides a hierarchical algorithm that indicates what date should be used to assign the *MMWR* week for syphilis case notifications, prioritizing dates closest to disease onset (e.g., specimen collection date). [xxxv] However, there is often variability in the algorithm's application and there are instances when provided dates conflict with the assigned *MMWR* week (e.g., the assigned *MMWR* week might be three weeks after the date of the laboratory result). An added complexity to the date variables is that the value for the data field representing the date when CDC received the case notification may be updated between the initial notification and when the case data are finalized; examples of situations when that may occur include a jurisdiction transitioning to a new PHIS or a jurisdiction replacing the case data during the final year-end data reconciliation. It is critical to understand the limitations of the various date variables as context for interpreting trends and assessing whether aberrations reflect true changes in morbidity, data quality issues, or variation in how jurisdictions assign dates.

Strategies for Integrating Aberration Detection into the Data Lifecyle

Selecting and implementing a data aberration detection methodology is complex. However, CDC and STLT health departments interested in integrating aberration into their data lifecycle can follow several strategies to make the data project more manageable and ensure that findings can be used for action.

Focus on three key decisions: prioritize aberrations to evaluate, establish thresholds, and establish the target audience and use case. Aberrations include both changes in morbidity that could indicate an outbreak as well as changes in values and counts that reflect data quality issues. As such, there are endless possibilities for what to analyze in a data aberration detection algorithm including case counts, case counts stratified by demographics or geography, and completeness of key variables. Furthermore, the best aberration detection algorithm for syphilis at the national level may not be optimal for identifying aberrations at the local level or for another reportable disease with different epidemiological features. Three guiding questions can help to focus initial thinking. First, use requirements gathering to make decisions about specific aberrations to prioritize based on dimensions such as public health importance, ease of understanding, frequency of occurrence, and whether information is actionable. For syphilis, identifying data quality issues with substance use variables (e.g., methamphetamine) may be actionable but may be a lower priority than ensuring completeness and accuracy of core variables such as sex of sex partners. This can be accomplished through the intentional use of different data visualizations. Second, use exploratory data analyses, CSTE's framework for outbreak detection,[xxxvi] and discussions with program staff who are closest to the data to establish operational definitions for thresholds. Additionally, it is important to note that these thresholds can change over time, may vary between high-morbidity and low-morbidity states, and may be different by prioritized population. Third, decide who is the target audience and establish a clear use case (e.g., a data report that will be emailed to jurisdiction-level surveillance staff versus an interactive dashboard limited to authorized data users). Different users will have varying priorities and preferences for data visualizations; as such, it is critical to be intentional about refining the concept for the product early in the development process.

Do not get bogged down in complexity. Data aberration detection is challenging because identifying and interpreting aberrations requires both a strong technical solution and qualitative insights about the infection and broader context. When selecting a statistical method, think carefully about balancing the tradeoffs between optimizing the tool's sensitivity and specificity versus pragmatic considerations such as the ease of implementation, the ability of end-users to understand and trust the findings, and that results are actionable. All approaches to detecting aberrations come with drawbacks, and there is unlikely to be a single best method for your application. The New York State Department of Health developed an Excel-based heat map that uses a historical limits approach to flag aberration in HIV surveillance data that may indicate issues with data quality or changing morbidity;[xxxvii] this has been adapted for current use with STI surveillance. Although the statistical approach may have lower accuracy than a more complex methodology, the tool is effective in that context because results are intended to be viewed

as exploratory to help surveillance staff identify counties to explore in more detail. Other advantages to the jurisdiction's Excel-based solution are that users can review and easily understand the calculations, and that its interface is accessible to users with varying coding skills. One suggestion for jurisdictions interested in developing data aberration detection tools for syphilis is to start small, such as only focusing on case counts stratified by sex and sex of sex partners, and later adding more complexity such as additional stratifications by other demographics and geographic levels.

Use a human-centered design process to incorporate users throughout the development. Incorporating user feedback throughout the design process helps to ensure the usability of the tool.[xxxviii] Detailed requirements gathering enables a clear understanding of end users, their information needs, and their technical skills. A formal usability evaluation can provide valuable information; however, even informal discussion with end-users around a prototype of the tool can yield critical feedback on the tool's content, design, and clarity.[xxxix] Tool development should not be approached as a statistical problem requiring a technical solution, but rather should be approached from the end-users' perspective on how the tool will be implemented.[xl]

Build a cross-functional development team with complementary expertise. For the data aberration detection project to be successful, do not think about it as a technology problem in need of a software solution. Careful consideration of implementation is critical throughout the design process. This requires establishing a development team with multi-disciplinary expertise and representation across functional units. Epidemiologists, biostatisticians, informaticians, and public health practitioners such as DIS with complementary expertise in the data system, public health surveillance and programs, and statistical methods are critical to ensuring the selected aberration detection approach is appropriate for the data system and context. Additionally, the development team needs members with the skills and resources to ensure a successful implementation so that the aberration detection solution becomes a routine activity after the development project is completed. This includes assessing how the solution will fit into staff workflows, navigating the political environment to secure buy-in from agency leadership and key partners, and recruiting key staff to participate in requirements gathering and usability assessment. Although this work can be done in-house, it can be helpful to include an external partner to provide an outside perspective during the development process, additional expertise, and dedicated time to conduct requirements gathering and other project activities. To ensure sustainability and use, the management of the tool after development is completed should be done internally without external assistance or technical support.

Develop a strategy for discussing aberrations when they are identified. For implementation of the aberration detection tool to succeed, it is important to consider how the results will be communicated. Discussing aberrations as they are detected should be done intentionally and respectfully across units in the jurisdiction, including both surveillance and program staff. The disease surveillance system necessitates strong partnerships between CDC and STLTs, and messaging that implies a partner is at fault can strain this relationship. Including individuals familiar with these challenging discussions in the development team is invaluable in developing the communication strategy. In addition to carefully designing messages, developing a system to track aberrations and communications can help to prevent a STLT partner from repeatedly receiving messages about a known issue. Further, this tracking can provide a means to evaluate how timely aberrations are being resolved.

Conclusions

There are many reasons for aberrations in live syphilis case surveillance data. While outbreaks are commonly identified as a reason for data aberrations, there are also data quality issues that can result in unexpected changes in observed values. Understanding and addressing these data quality issues is important for identifying changes that are attributable to outbreaks. Aberration detection is difficult to execute. The inherent complexity of the data lifecycle with multiple partners and processes to transmit data to CDC necessitates a deep understanding of the data ecosystem. There is no single "best" data aberration detection algorithm, and the most appropriate solution will differ depending on the context. Specific challenges to identifying and understanding aberrations include establishing a clear definition for what constitutes an outbreak, lag times in reporting and data processing, incomplete data for key variables, and the complexity of data variables. Although data aberration monitoring of case surveillance data is challenging, this work can be enhanced through an interdisciplinary developer team to create a customized aberration detection tool considering human-centered design principles and an interdisciplinary review team to inspect suspected aberrations before public health action.

Sources

[i] Centers for Disease Control and Prevention. Sexually transmitted disease surveillance 2022. Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention; 2024. https://www.cdc.gov/std/statistics/2022/default.htm Published January 30, 2024. Accessed June 24, 2024.

[ii] Centers for Disease Control and Prevention. Sexually transmitted disease surveillance 2003. Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention; 2004. https://www.cdc.gov/std/statistics/archive.htm Accessed June 24, 2024.

[iii] US Department of Health and Human Services. 2020. Sexually Transmitted Infections National Strategic Plan for the United States: 2021-2025. Washington, DC. Available at: www.hhs.gov/STI. Accessed June 24, 2024.

[iv] Centers for Disease Control and Prevention. Data modernization initiative. Available at: https://www.cdc.gov/surveillance/data-modernization/index.html. Accessed December 27, 2023.

[v] Centers for Disease Control and Prevention. Public health data strategy. Available at: https://www.cdc.gov/ophdst/public-health-data-strategy. Accessed December 27, 2023.

[vi] Centers for Disease Control and Prevention. Syphilis (*Treponema pallidum*) 2018 case definition. Available at: https://ndc.services.cdc.gov/case-definitions/syphilis-2018/. Accessed August 11, 2024.

[vii] Centers for Disease Control and Prevention. Public health information systems (PHIS). Available at: https://www.cdc.gov/std/informatics/public-health-information-systems.htm. Published December 21,2023. Accessed June 26, 2024.

[viii] Martin EG, Angles JS. Incorporating a health equity lens into surveillance information systems: Opportunities and challenges. J Public Health Manag Pract. 2023; 29(1): 1-4.

[ix] Centers for Disease Control and Prevention. Public health information systems (PHIS). Available at: https://www.cdc.gov/std/informatics/public-health-information-systems.htm. Published December 21,2023. Accessed June 26, 2024.

[x] Centers for Disease Control and Prevention, Office of Public Health Data, Surveillance, and Technology. HL7 message mapping guides & standards. Available at: https://www.cdc.gov/nndss/trc/mmg/index.html. Published March 15, 2024. Accessed June 26, 2024.

[xi] Centers for Disease Control and Prevention. Sexually transmitted infections surveillance, 2022. Available at: https://www.cdc.gov/std/statistics/2022/nndss.htm. Published January 30, 2024. Accessed June 26, 2024.

[xii] Martin EG, Angles JS. Incorporating a health equity lens into surveillance information systems: Opportunities and challenges. J Public Health Manag Pract. 2023; 29(1): 1-4.

[xiii] Cunningham M, Patel K, McCall T, et al. 2022 National profile of local health departments. National Association of County and City Health Officials. 2024. Available at: https://www.naccho.org/resources/lhd-research/national-profile-of-local-health-departments.

[xiv] Stroup DF, Wharton M, Kafardar K, Dean AG. Evaluation of a method for detecting aberrations in public health surveillance data. Am J Epidemiol. 1993; 137(3): 373-380.

[xv] This is not an exhaustive list of methodologies and classification of models into distinct domains is subjective. For example, hidden Markov models use a Bayesian approach and time series generalized linear models are a special case of ARIMA models.

[xvi] Farrington CP, Andrews NJ, Beale AD, Catchpole MA. A statistical algorithm for the early detection of outbreaks of infectious disease. Journal of the Royal Statistical Society. Series A (Statistics in Society). 1996; 159(3): 547-563.

[xvii] Stroup DF, Wharton M, Kafardar K, Dean AG. Evaluation of a method for detecting aberrations in public health surveillance data. Am J Epidemiol. 1993; 137(3): 373-380.

[xviii] Centers for Disease Control and Prevention. Morbidity and Mortality Weekly Report (MMWR), week ending December 30, 2017 (52nd week). Available at: https://www.cdc.gov/mmwr/volumes/66/wr/mm6652md.htm.

[xix] Hutwagner L, Thompson W, Seeman GM, Treadwell T. The bioterrorism preparedness and response Early Aberration Reporting System (EARS). J Urban Health. 2003; 80(2 Suppl 1): i89-96.

[xx] Gundlapalli AV, Olson J, Smith SP, Baza M, Hausam RR, Eutropius LJ, et al. Hospital electronic medical record–based public health surveillance system deployed during the 2002 Winter Olympic Games, Am J Infect Control. 2007; 35(3):163-171.

[xxi] Joshi M, Yuan Y, Miranda W, Chung R, Rajulu DT, Hart-Malloy R. A peek into the future: How a pandemic resulted in the creation of models to predict the impact on sexually transmitted infection(s) in New York State (excluding New York City). Sex Transm Dis. 2021; 48(5): 381-384.

[xxii] Liboschik T, Fokianos K, Fried R. Tscount: An R package for analysis of count time series following generalized linear models. J Stat Softw. 2017; 82(5): 1–51.

[xxiii] Yuan M, Boston-Fisher N, Luo Y, Verma A, Buckeridge D. A systematic review of aberration detection algorithms used in public health surveillance. J Biomed Inform. 2019; 94:103181.

[xxiv] Jiang X, Cooper G. A Bayesian spatio-temporal method for disease outbreak detection. J Am Med Inform Assoc. 2010; 17(4):462-471.

[xxv] Martínez-Beneito MA, Conesa D, López-Quilez A, López-Maside A. Bayesian Markov switching models for the early detection of influenza epidemics. Stat Med. 2008; 27(22):4455-4468.

[xxvi] Kane MJ, Price N, Scotch M, et al. Comparison of ARIMA and random forest time series models for prediction of avian influenza H5N1 outbreaks. BMC Bioinformatics. 2014 15:276.

[xxvii] Zaki M, Meira Jr, W. Part 4: Classification. In: *Data Mining and Machine Learning: Fundamental Concepts and Algorithms 2nd ed.* Cambridge University Press, 2020. 467:585.

[xxviii] Council of State and Territorial Epidemiologists STD Subcommittee. Syphilis outbreak detection guidance. Available at: https://npin.cdc.gov/publication/syphilis-outbreak-detection-guidance Published 2019. Accessed December 27, 2023.

[xxix] Centers for Disease Control and Prevention. Health Alert Network (HAN). Available at: https://emergency.cdc.gov/han/index.asp. Published March 7, 2022. Accessed June 26, 2024.

[xxx] Source: Authors' review of Health Alert Network messages for syphilis.

[xxxi] Data are from the National Notifiable Disease Surveillance System weekly snapshot files and the reconciled end-of-year surveillance files, with each box-and-whiskers plot including 50 states and the District of Columbia. Most plots include twelve observations for each jurisdiction, one for each month although in some instances snapshot files were unavailable for the first months of the year. It is possible to change a case's date after the initial report;

as such, all snapshot file case counts are approximations and for some months, more cases were reported by a jurisdiction in a live file than in the reconciled end-of-year file. In those instances, the calculation resulted in a negative value, which was manually recoded to zero for this figure.

[xxxii] Data are from sampled National Notifiable Disease Surveillance System weekly snapshot files (with 75% of counties included) and cover 50 states and the District of Columbia. Lines indicate the total number of syphilis cases for each MMWR year between 2016 and 2020. Each line begins with the first available snapshot file of the MMWR year and ends at the time of final data reconciliation for that year. Some lines are longer and overlap with the following MMWR year because the length of time until data reconciliation varies by year. The line for 2016 is truncated because snapshot files were unavailable until June 2016. The X-axis extends to 2022 because the 2020 surveillance data were not reconciled until December 2021.

[xxxiii] Centers for Disease Control and Prevention. Sexually transmitted disease surveillance 2022. Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention; 2024. https://www.cdc.gov/std/statistics/2022/default.htm Published January 30, 2024. Accessed June 24, 2024.

[xxxiv] Data are from the National Notifiable Disease Surveillance system reconciled end-of year surveillance file. Each box-and-whiskers plot includes 50 states and the District of Columbia. Five years of data are displayed for each priority variable collected by NNDSS. Sexual orientation and gender identity were not collected by NNDSS before 2018. Jurisdictions are encouraged to include valid and usable values for all variables for all cases, and at least 70% of values must be valid and usable in order for the jurisdictions' data to be included in national reporting.

[xxxv] Centers for Disease Control and Prevention. Guidance on classifying STD case reports into MMWR week. Available at: https://www.cdc.gov/std/program/mmwr-week-guidance-cleared-feb-2021.pdf. Published February 2021. Accessed June 26, 2024.

[xxxvi] Council of State and Territorial Epidemiologists STD Subcommittee. Syphilis outbreak detection guidance. Available at: https://npin.cdc.gov/publication/syphilis-outbreak-detection-guidance Published 2019. Accessed December 27, 2023.

[xxxvii] New York State Department of Health. Using heatmaps to visualize trends in HIV diagnoses. Data Analysis and Research Translation (DART) Presentation of HIV/AIDS Current Topics (PHACT) report #8. Published April 2023. Accessed November 2, 2023. https://www.health.ny.gov/diseases/aids/general/statistics/docs/dart_phact_heatmap.pdf

[xxxviii] Norman D. The Design of Everyday Things. Rev ed. New York, NY: Basic Books; 2013.

[xxxix] Lloyd D, Dykes J. Human-centered approaches in geovisualization design: investigating multiple methods through a long-term case study. IEEE Transactions on Visualizations and Computer Graphics. 2011; 17:2498-2507.

[xl] Ansari B, Martin EG. Integrating human-centered design in public health data dashboards: lessons from the development of a data dashboard of sexually transmitted infections in New York State. J Am Med Inform Assoc. 2024; 31(2):298-305.