BMJ Open

Measuring concordance of data sources used for infectious disease research in the USA: a retrospective data analysis

Maimuna S Majumder ,[1,2] Marika Cusick ,[3] Sherri Rose[3]

To cite: Majumder MS, Cusick M, Rose S. Measuring concordance of data sources used for infectious disease research in the USA: a retrospective data analysis. *BMJ Open* 2023;**13**:e065751. doi:10.1136/bmjopen-2022-065751

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2022-065751).

MSM and MC contributed equally.

MSM and MC are joint first authors.

Received 16 June 2022
Accepted 07 February 2023

[1]Computational Health Informatics, Boston Children's Hospital, Boston, Massachusetts, USA
[2]Department of Pediatrics, Harvard Medical School, Boston, Massachusetts, USA
[3]Department of Health Policy, Stanford University, Stanford, California, USA

Correspondence to
Dr Maimuna S Majumder; maimuna.majumder@childrens.harvard.edu

ABSTRACT

Objectives As highlighted by the COVID-19 pandemic, researchers are eager to make use of a wide variety of data sources, both government-sponsored and alternative, to characterise the epidemiology of infectious diseases. The objective of this study is to investigate the strengths and limitations of sources currently being used for research.

Design Retrospective descriptive analysis.

Primary and secondary outcome measures Yearly number of national-level and state-level disease-specific case counts and disease clusters for three diseases (measles, mumps and varicella) during a 5-year study period (2013–2017) across four different data sources: Optum (health insurance billing claims data), HealthMap (online news surveillance data), Morbidity and Mortality Weekly Reports (official government reports) and National Notifiable Disease Surveillance System (government case surveillance data).

Results Our study demonstrated drastic differences in reported infectious disease incidence across data sources. When compared with the other three sources of interest, Optum data showed substantially higher, implausible standardised case counts for all three diseases. Although there was some concordance in identified state-level case counts and disease clusters, all four sources identified variations in state-level reporting.

Conclusions Researchers should consider data source limitations when attempting to characterise the epidemiology of infectious diseases. Some data sources, such as billing claims data, may be unsuitable for epidemiological research within the infectious disease context.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ Direct comparison of infectious disease reporting across publicly available data sources provides insight into their robustness.
⇒ Methods allow for the benchmarking of health insurance claims data reliability for research, which has been challenging to quantify in other contexts.
⇒ While our analysis focused on three infectious diseases (measles, mumps and varicella), our approach may not be generalisable for other diseases.
⇒ We relied on numerous assumptions to identify infectious disease clusters (eg, geographical and temporal constraints).
⇒ Infectious disease research and reporting is not limited to the data sources studied here; it may be worthwhile to investigate other sources.

INTRODUCTION

The COVID-19 pandemic has exposed foundational gaps in government-sponsored public health surveillance across the USA.[1] Most notably, for the first year of the pandemic, the Centers for Disease Control and Prevention (CDC)—which has historically been responsible for reporting population-level situational statistics (eg, cases, hospitalisations and deaths over time during infectious disease outbreaks)—did not efficiently report COVID-19-related statistics. This was due, in part, to lack of prioritisation and underinvestment in local public health surveillance systems.[2] News media organisations such as The Atlantic's COVID Tracking Project partially filled this gap,[3] highlighting the critical role that alternative data sources can play during public health emergencies. Situational statistics are also useful more broadly in infectious disease epidemiology research.

Gaps in government-sponsored public health surveillance have long preceded the pandemic, as has the practice of leveraging alternative data sources. For infectious disease research specifically, case count data obtained from news coverage of outbreaks led to studies that examined the 2014–2015 Disneyland measles outbreak[4] and the 2016 Arkansas mumps outbreak,[5] as well as a broad range of international studies, including Zika[6 7] and dengue[8] in Latin America, H7N9 in China[9] and Ebola in West Africa,[10] among others. News media data have repeatedly demonstrated that usefulness in aggregating case counts, and in each of the aforementioned instances, was implemented to augment otherwise insufficient data from government-sponsored agencies.

In high-income settings such as the USA, insufficiency of government-sponsored public health data is often characterised by

delays in reporting.[11] Although the data may exist and are frequently treated as 'ground truth' statistics, they are reported at a pace that disallows real-time evaluation of emergent crises. For example, even nationally notifiable infectious diseases are only reported once a week by the CDC's National Notifiable Diseases Surveillance System (NNDSS)[12]—a pace that is too infrequent for real-time monitoring and mitigation of highly contagious infectious disease outbreaks. Moreover, NNDSS data—unlike news media data—are only reported at the state level, which is an inadequate geographic resolution in the event of localised (ie, county level or zip level) outbreaks. The CDC also prepares Morbidity and Mortality Weekly Reports (MMWR), which include detailed government reports on notable infectious disease outbreaks.[13] However, these reports are challenging to rely on for emergent crises, as there are no clear inclusion criteria for an MMWR reportable outbreak, inconsistencies exist in the reported level of geographic resolution, and they are often published well after the outbreak.

Beyond news media data, insurance billing claims data are also a potential alternative data source for characterising infectious disease epidemiology in the USA. These data experience more considerable delays in reporting, with data released after months or more.[14 15] However, unlike the population-level situational statistics that are obtainable from news media data and government-sponsored public health surveillance systems, insurance claims provide patient-level data. Historically, these patient-level data have enabled important advances in monitoring chronic illness in both individuals and populations, but their utility within the context of acute infectious disease surveillance remains largely untested. Given recent interest in using insurance claims data to study COVID-19,[14] validating the quality of these data for other infectious diseases—those that predate the pandemic—is urgently needed.

In this retrospective study, we evaluate case count data for the years 2013 through 2017 from the news media platform HealthMap and the Optum insurance claims database against two government-sponsored data sources (NNDSS and MMWR) for three infectious diseases: measles, mumps and varicella (chickenpox). Because these three diseases are nationally notifiable, healthcare providers are obligated to report cases of them to state health agencies and state health agencies are obligated to report them to the CDC—thus ensuring a high degree of completeness for government-sponsored data.

METHODS

We compared infectious disease case counts for each disease across all four sources during the period 2013–2017 (online supplemental table 1). Our main outcomes of interest were yearly counts of cases and Micropolitan and Metropolitan Statistical Area (MSAs) clusters at both the national and state level. Clusters are defined as a group of cases interrelated according to both time and geography.

Patient and public involvement statement

Patients and public were not involved in the development of the research questions or design of the analysis in this study.

Data sources

Health insurance claims data

Optum Clinformatics Data Mart Database is a deidentified database derived from a large claims data warehouse.[15] The claims submitted have been adjudicated to the appropriate enrollee, adjusted and deidentified prior to inclusion in the database. Claims are subject to adjustment after initial adjudication due to delays in reporting and additional visit information.

The database includes approximately 15–20 million annual covered enrollees for a total of roughly 83 million unique enrollees from 2006 to 2018. During the 2013–2017 period of our study, there were approximately 39 million unique enrollees in commercial and medicare plans. The Optum Clinformatics Data Mart contains enrollee-level information on demographics (age and documented sex) and geography at the ZIP code level. Individual enrollee medical claims include data on the date of service, as well as associated diagnoses, procedures, laboratory tests, prescriptions and providers.

Using a set of International Classification of Diseases (ICD-9 and ICD-10) codes, we identified enrollees with diagnoses for measles, mumps and varicella (see online supplemental table 2 for ICD codes). Given the nature of these infectious diseases, we assumed that enrollees could only have each disease once during the 5-year period. We identified service dates and ZIP codes associated with the enrollee's first diagnosis.

US Department of Housing and Urban Development United States Postal Service CrossWalk files were used to map patient ZIP codes to the core-based statistical areas (CBSAs) for MSAs as defined by the Office of Management and Budget in February 2013.[16] The Optum Clinformatics Data Mart protects against reidentification by associating enrollee with multiple different ZIP codes if they live in a ZIP code with a small number of people. In this case, we used the first identified ZIP code–MSA pairing. Further details on cleaning and processing data from the Optum Clinformatics Data Mart are provided in the online supplemental appendix.

Enrollees without CBSA and state-level information were not included in the cluster and state-level portion of the analysis. However, enrollees without this granular location information were included in overall national case counts. Descriptive statistics of the enrollees for each disease cohort are in online supplemental table 3.

Online news surveillance data

HealthMap surveillance data aggregates online informal news sources for disease outbreak monitoring and public

health surveillance. Since September 2006, HealthMap has offered free access to their automated database, and many national and international organisations have used these data for surveillance activities.[17] For each HealthMap alert (eg, news article), the database contains the disease of interest, article date, associated latitude and longitude coordinates of the location (which can be used to assign MSA or state), number of confirmed cases and number of confirmed deaths.

We used QGIS—a software application for geographic information systems, to conduct spatial joins between the latitude and longitude coordinates associated with each HealthMap alert to MSAs and states.[18] All HealthMap alerts without granular location information (eg, only at the state level or country level) were removed from the analysis.

Many HealthMap alerts are duplicate entries of the same disease cluster. To avoid overestimating the number of cases reported from this source, we identified clusters within this database according to time (serial intervals) and spatial (MSA) constraints. The start and end of an MSA-level cluster was determined by two consecutive serial intervals, the time between successive cases in transmission, of zero new cases. We assumed the total number of cases associated with each MSA-level cluster was the highest number of confirmed cases reported among all associated HealthMap alerts.

Official government reports

MMWR contain scientific records of public health information and recommendations.[13] For major disease outbreaks, the CDC will publish a conclusive MMWR, describing key information such as the date of identification, locations affected and total number of cases. We manually reviewed all MMWR that related to measles mumps, and varicella and extracted cluster identification dates, confirmed case counts and corresponding MSA locations to allow comparison against the other data sources considered in our analysis. The online supplemental appendix contains detailed information on all MMWR.

Government case surveillance data

The CDC conducts mandatory disease reporting and surveillance for our three diseases of interest. We used data from NNDSS, which provides weekly tables of disease counts.[12] The data contain the number of cases reported during the current week, as well as the number of cumulative cases reported over a given year. If there is a delay in reporting, the case will only appear as a part of the cumulative count. NNDSS only provides case counts at the state level; a more granular geographic resolution is unavailable for public use.

Analyses

Standardised national yearly case counts

For each disease, we reported source-specific national yearly case counts standardised to 100 000 persons. Optum data was standardised to the total number of eligible Optum enrollees during the years 2013–2017. Data from NNDSS, HealthMap and MMWR were standardised to the US population as per census bureau national population estimates.[19] While Optum and MMWR are not meant to capture case counts in ways that are nationally representative, values are standardised to this population for comparability across data sources.

National cumulative case counts

For each disease and each data source, we reported cumulative incidence of cases over the entire study period. Due to Optum data privacy requirements, we display the cumulative case count once the national case counts are at least 16 cases for this data source.

State-level cases

For each disease, we reported yearly state-level case counts for Optum, NNDSS, HealthMap and MMWR. State information was ascertained from each data source. In Optum, we translated patient ZIP code information to state-level information using the pyzipcode python module.[20] We used available NNDSS state-level information directly. Confirmed cases from each HealthMap MSA-level cluster were allocated to corresponding states. In the case of multistate MSAs, we allocated cases to states according to the relative proportion of HealthMap alert-associated states within the cluster. Finally, based on the identified MSA location from the MMWR, we allocated cases to each state. As per Optum privacy constraints, we do not report any state-level cases counts smaller than 16 cases.

State-level clusters

For each disease, we reported the yearly number of MSA-level clusters in a given state according to Optum, HealthMap and MMWR. The start and end of an MSA-level cluster was determined by two consecutive serial intervals of zero new (ie, incident) cases. Serial interval periods differ based on the disease of interest: measles (12 days), mumps (18 days) and varicella (14 days).[21] We report MSA-level clusters with at least 16 cases due to Optum privacy constraints and then comparability across all data sources. Further details regarding cluster identification are provided in the online supplemental appendix. Because NNDSS does not provide granular geographical data beyond the state-level, we did not use this source to identify MSA-level clusters.

In the event of multistate MSAs, we assigned the MSA-level cluster to a single state for each of the sources. In Optum, we assigned the MSA-level cluster according to the most frequent patient-reported state. In HealthMap, we assigned the MSA-level cluster to the most commonly reported state among the associated HealthMap alerts. Finally, for MMWR, we assigned the MSA-level cluster to states by extracting the state from the available text information, as further specified in the online supplemental appendix.

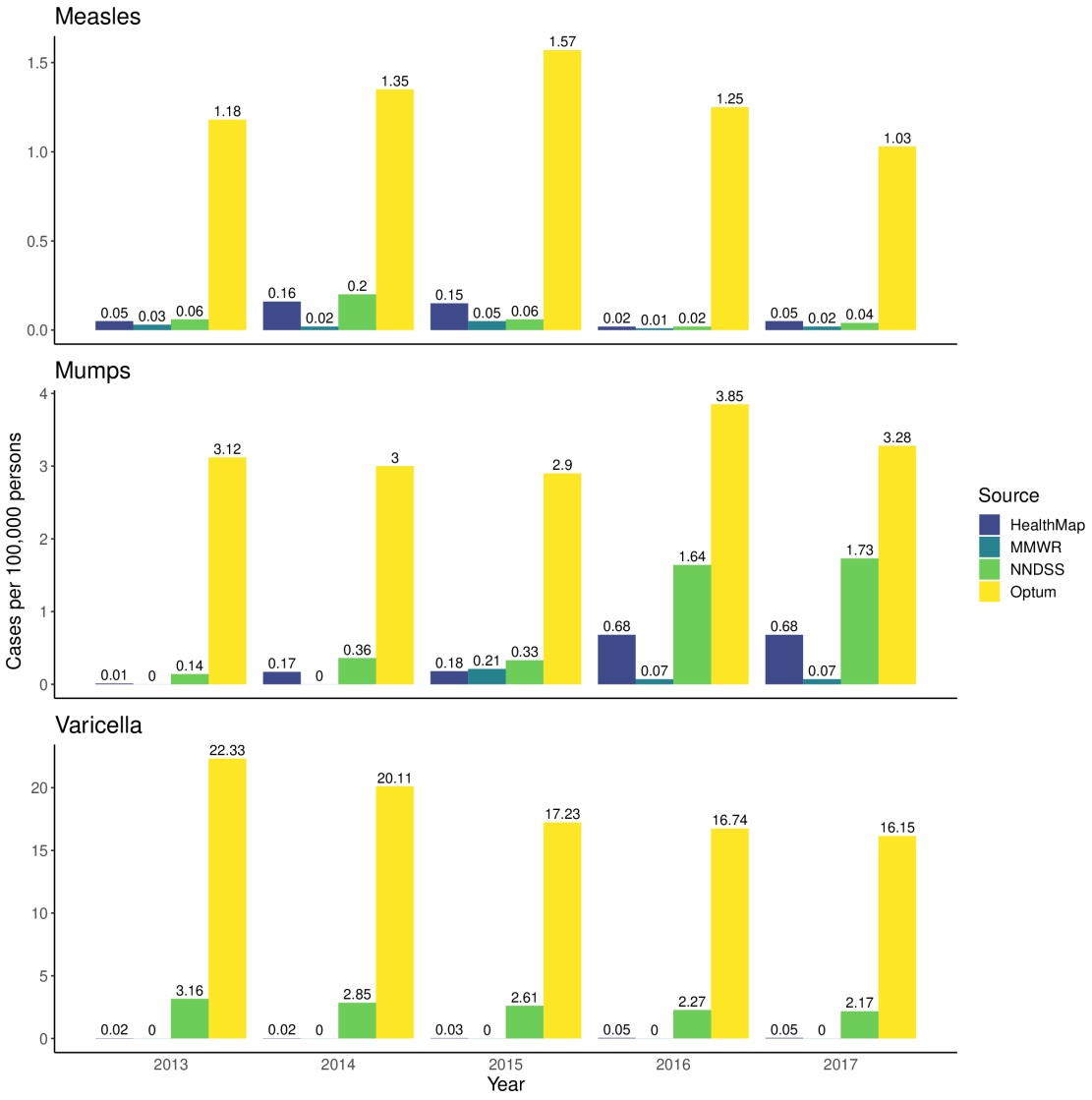

Figure 1 Standardised yearly national case counts. MMWR and Optum are not designed to capture the entire US population; values are standardised to this population for comparability across data sources. MMWR, Morbidity and Mortality Weekly Reports; NNDSS, National Notifiable Diseases Surveillance System.

RESULTS

National standardised incidence for all three diseases is substantially higher for Optum data in comparison to other sources (figure 1) and implausibly large in magnitude. Case counts from MMWR are the lowest, although this is expected as MMWR are only generated for key clusters across the USA. While HealthMap reports slightly higher case counts in comparison to NNDSS for measles and mumps, there were fewer cases reported in varicella, suggesting that varicella is less 'newsworthy'. Unstandardised yearly case counts are provided in online supplemental figures 1–4.

Examining state-level geographic trends, Ohio had the highest number of measles cases during the study period according to HealthMap and NNDSS (figure 2). In comparison, Optum reported the highest number of cases in New York and New Jersey. California had the highest case counts according to MMWR. All states with MMWR were also captured as having measles cases in both HealthMap and NNDSS.

For mumps, there were few MMWR on outbreaks during the study period (figure 3). Of the states with clusters identified by MMWR, all other sources reported cases for these states as well. There was a high concentration of mumps cases in the Midwestern region (Iowa, Illinois, Missouri, Indiana and Ohio) according to HealthMap, yet this concentration was not reflected as clearly in NNDSS and Optum data. NNDSS reported a substantial number of mumps cases in Arkansas, yet there was no MMWR on these cases.

Nearly all states reported varicella cases in the Optum data (figure 4). According to NNDSS, Texas and Florida reported the highest numbers of varicella cases, which was also reflected in the Optum data, as these states also had higher numbers during the study period. Very few varicella cases were reported in HealthMap and MMWR.

Cumulative incidence of measles and mumps cases over the study period follows similar general patterns in HealthMap and NNDSS (figure 5). Disease clusters are evident as case counts rise rapidly and then are stagnant. In

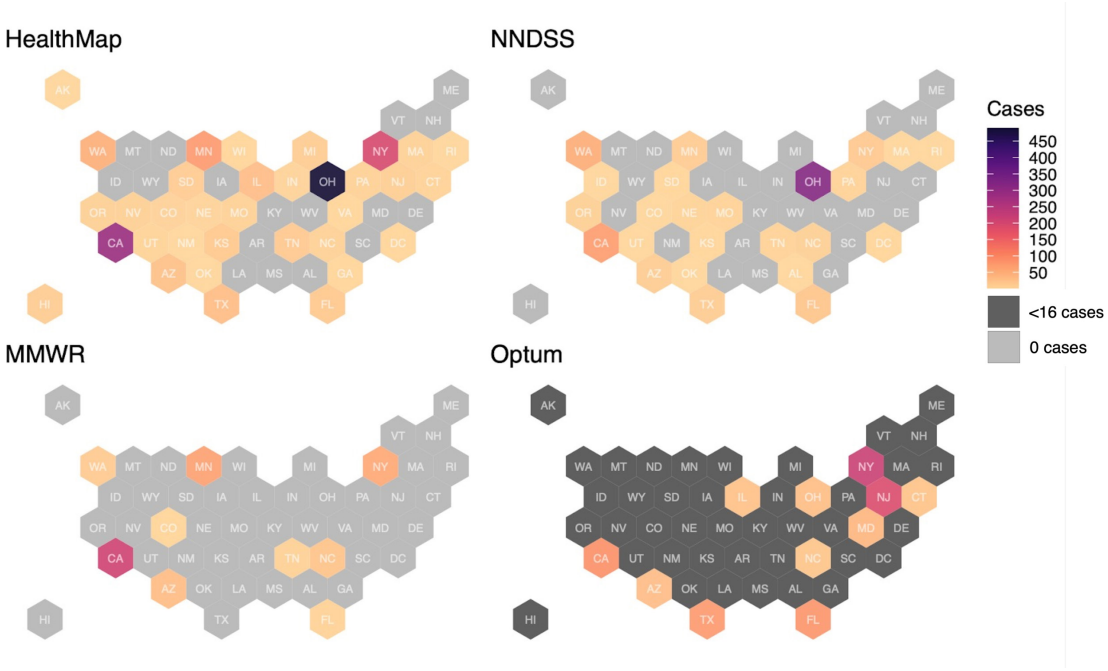

Figure 2 State-level measles cases (2013–2017). Optum data are presented for states with at least 16 cases during the study period. MMWR, Morbidity and Mortality Weekly Reports; NNDSS, National Notifiable Diseases Surveillance System.

comparison, in the Optum data, measles and mumps case counts rose constantly over time. Incidence of varicella cases were constant over time in all data sources.

DISCUSSION

To our knowledge, this is the first study to examine the concordance of infectious disease case counts across multiple disparate sources, including news media, insurance claims and government-sponsored data. We found wide variation in the number of reported cases for measles, mumps and varicella across these data sources, with implausibly high volumes of standardised cases reported by Optum that far exceed the other sources considered. Because these three highly infectious diseases are nationally notifiable and thus must be reported both to state health agencies and to the CDC, it is highly unlikely that Optum would capture cases that were not reported by NNDSS.

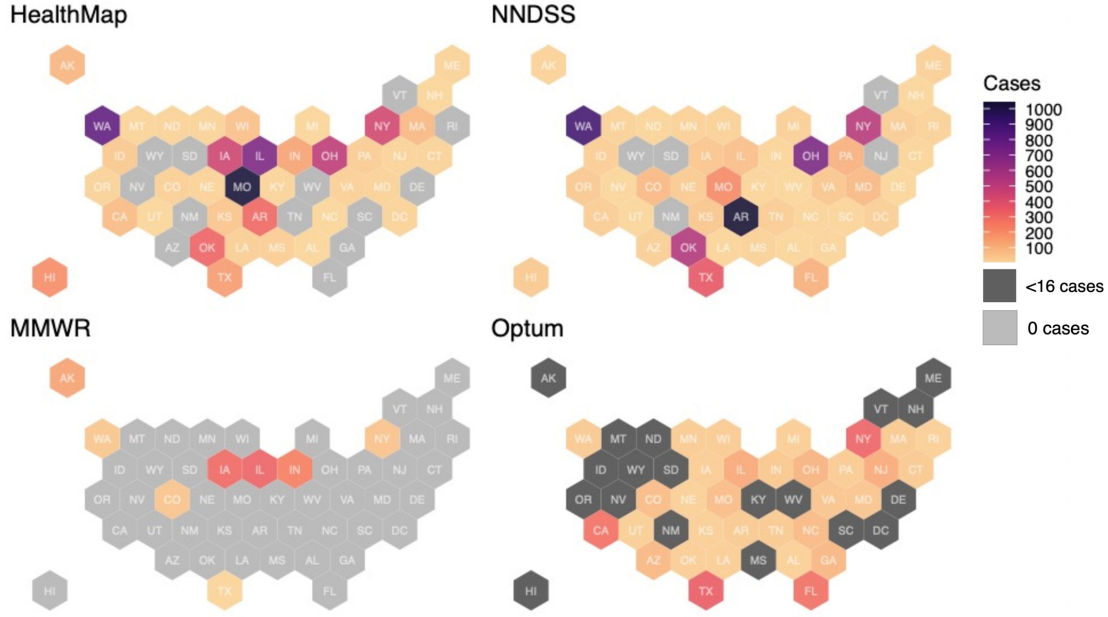

Figure 3 State-level mumps cases (2013–2017). Optum data are presented for states with at least 16 cases during the study period. MMWR, Morbidity and Mortality Weekly Reports; NNDSS, National Notifiable Diseases Surveillance System.

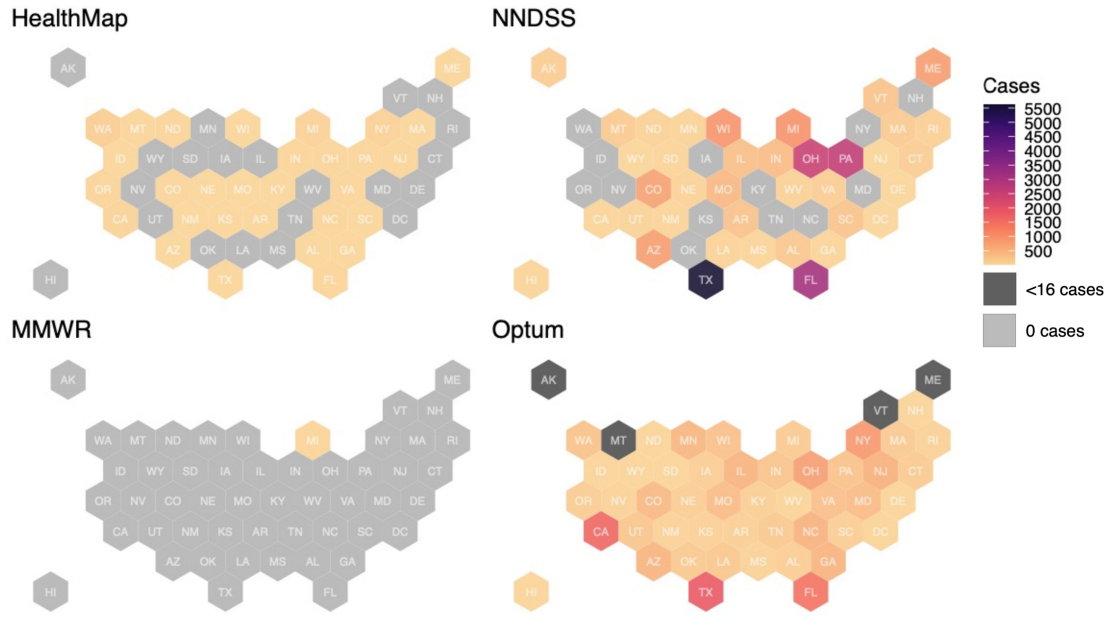

Figure 4 State-level varicella cases (2013–2017). Optum data are presented for states with at least 16 cases during the study period. MMWR, Morbidity and Mortality Weekly Reports; NNDSS, National Notifiable Diseases Surveillance System.

Overcounting may be due to the coding of likely cases, as perceived by providers, rather than laboratory confirmed diagnoses. However, laboratory results in claims data are typically incomplete as many test results are not recorded[22]; thus, analyses that include only laboratory confirms cases produce severe undercounts (as presented in the online supplemental appendix)

Notably, evidence of overbilling for conditions such as measles and mumps may contribute to the rise in medical expenditures and patient healthcare spending. Using Optum data on reported total paid charges, we estimated wasted expenditures from suspected overbilling of measles and mumps cases to be roughly US$ 396 000 for the 5-year period among Optum enrollees alone (online supplemental appendix, online supplemental table 4). While the use of insurance claims data to characterise infectious disease epidemiology might appear appealing due to the availability of additional individual-level information, these analyses may lack credibility given the erroneous coding issues we identified here.

While there are well-known gaps in government-sponsored data sources, NNDSS compared favourably to other sources, capturing a larger scope of the mumps outbreak in 2016–2017 as well as more varicella than HealthMap or MMWR. We also saw that HealthMap may produce similar case counts to NNDSS in non-outbreak years for measles and mumps. This is advantageous as HealthMap does not have the same delays in reporting as NNDSS and is also available at a more granular geographic resolution. However, HealthMap is not likely to be a reliable source for case counts of less 'newsworthy' diseases such as varicella.

Our study focused on three unique sources of data. However, infectious disease research and reporting is not limited to these sources, and it is critical to investigate the reliability of other data in the future, including electronic health records, social media and wastewater data.

Before using a particular data source to characterise the epidemiology of a given infectious disease, researchers should consider conducting qualitative interviews to understand the underlying data generation processes that led to the creation of the data and how these processes may impact reliability. Our study illustrated that health insurance billing claims data may not have reliable estimates of measles and mumps in the USA. These issues

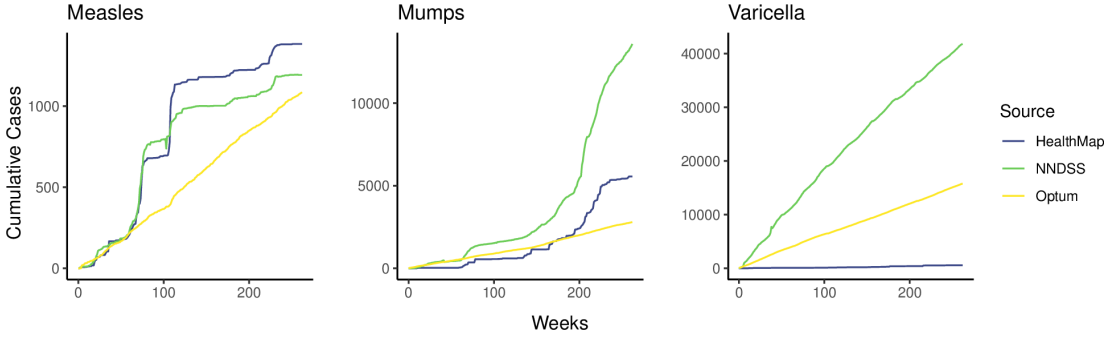

Figure 5 Cumulative incidence during study period. NNDSS, National Notifiable Diseases Surveillance System.

likely arise due to disease under-coding and misclassification, lack of population representativeness and lagged reporting, as previously shown in other infectious disease and chronic disease settings.[14 23 24] Data sources with identified reliability issues may not be suitable for research questions that are contingent on reliable reporting of situational statistics—including those that pertain to the epidemiological properties of COVID-19 and other infectious diseases.

Acknowledgements Health insurance claims data for this project were accessed using the Stanford Center for Population Health Sciences Data Core. The PHS Data Core is supported by a National Institutes of Health National Center for Advancing Translational Science Clinical and Translational Science Award (UL1TR003142) and from Internal Stanford funding. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. We would like to acknowledge and thank David Scales for using QGIS to do the spatial joins between HealthMap geographic coordinates and MSAs. We would also like to acknowledge and thank the HealthMap team for providing the HealthMap data.

Contributors MSM and SR conceptualised the research question and methodology. MC curated the data and conducted the analysis. All authors contributed to the final draft in writing, reviewing, and editing. SR, as guarantor, accepts full responsibility for the work, had access to the data, and controlled the decision to publish.

Funding This project is supported in part through the NIH Director's New Innovator Award DP2-MD012722. MC is supported by the T32HS026128 grant from the Agency for Healthcare Research and Quality.

Map disclaimer The inclusion of any map (including the depiction of any boundaries therein), or of any geographic or locational reference, does not imply the expression of any opinion whatsoever on the part of *BMJ* concerning the legal status of any country, territory, jurisdiction or area or of its authorities. Any such expression remains solely that of the relevant source and is not endorsed by *BMJ*. Maps are provided without any warranty of any kind, either express or implied.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data may be obtained from a third party and are not publicly available. All data relevant to the study are included in the article or uploaded as online supplemental information. All data in the present study are available online with the exception of the Optum Clinformatics Data Mart.

ORCID iDs
Maimuna S Majumder http://orcid.org/0000-0002-3986-4303
Marika Cusick http://orcid.org/0000-0002-1750-5246

REFERENCES

1 Data heroes of covid tracking project are still filling U.S. government void. bloomberg.com [internet]. 2020 nov 20 [cited 2022 mar 13]. n.d. Available: https://www.bloomberg.com/news/features/2020-11-20/covid-tracking-project-volunteers-step-up-as-u-s-fails-during-pandemic
2 Kondilis E, Papamichail D, Gallo V, *et al*. COVID-19 data gaps and lack of transparency undermine pandemic response. *J Public Health (Oxf)* 2021;43:e307–8.
3 Analysis & updates | giving thanks and looking ahead: our data collection work is done [internet]. the COVID tracking project. 2022. Available: https://covidtracking.com/analysis-updates/giving-thanks-and-looking-ahead-our-data-collection-work-is-done
4 Substandard vaccination compliance and the 2015 measles outbreak | infectious diseases | JAMA pediatrics | JAMA network [internet]. [cited 2022 mar 13]. 2022. Available: https://jamanetwork.com/journals/jamapediatrics/article-abstract/2203906
5 Majumder MS, Nguyen CM, Cohn EL, *et al*. Vaccine compliance and the 2016 arkansas mumps outbreak. *Lancet Infect Dis* 2017;17:361–2.
6 McGough SF, Brownstein JS, Hawkins JB, *et al*. Forecasting Zika incidence in the 2016 Latin America outbreak combining traditional disease surveillance with search, social media, and news report data. *PLoS Negl Trop Dis* 2017;11:e0005295.
7 JMIR public health and surveillance - utilizing nontraditional data sources for near real-time estimation of transmission dynamics during the 2015-2016 colombian zika virus disease outbreak [internet]. 2022. Available: https://publichealth.jmir.org/2016/1/e30/
8 Hoen AG, Keller M, Verma AD, *et al*. Electronic event-based surveillance for monitoring dengue, Latin America. *Emerg Infect Dis* 2012;18:1147–50.
9 Salathé M, Freifeld CC, Mekaru SR, *et al*. Influenza A (H7N9) and the importance of digital epidemiology. *N Engl J Med* 2013;369:401–4.
10 Majumder MS, Kluberg S, Santillana M, *et al*. 2014 ebola outbreak: media events track changes in observed reproductive number. *PLoS Curr* 2015;28.
11 Ghosh S, Chakraborty P, Nsoesie EO, *et al*. Temporal topic modeling to assess associations between news trends and infectious disease outbreaks. *Sci Rep* 2017;7:40841.
12 National notifiable diseases surveillance system | CDC [internet]. 2022. Available: https://www.cdc.gov/nndss/index.html
13 Morbidity and mortality weekly report (MMWR) | MMWR [internet]. 2022. Available: https://www.cdc.gov/mmwr/index.html
14 Health care claims data may be useful for COVID-19 research despite significant limitations | health affairs [internet]. [cited 2022 mar 13]. 2022. Available: https://www.healthaffairs.org/do/10.1377/forefront.20201001.977332/full/
15 Stanford center for population health sciences. optum ZIP5 (v5.0) [internet]. redivis; p. 4949603231427 bytes. 2022. Available: https://redivis.com/datasets/5c1s-bvewzf4td?v=5.0
16 HUD USPS ZIP code crosswalk files | HUD USER [internet]. n.d. Available: https://www.huduser.gov/portal/datasets/usps_crosswalk.html#codebook
17 Freifeld CC, Mandl KD, Reis BY, *et al*. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. *J Am Med Inform Assoc* 2008;15:150–7.
18 Welcome to the QGIS project! [internet]. 2022. Available: https://www.qgis.org/en/site/
19 Bureau UC. National population totals and components of change: 2010-2019 [internet]. census.gov. 2022. Available: https://www.census.gov/data/datasets/time-series/demo/popest/2010s-national-total.html
20 Gheem NV. Pyzipcode: query zip codes and location data [internet]. 2022. Available: https://github.com/vangheem/pyzipcode
21 Vink MA, Bootsma MCJ, Wallinga J. Serial intervals of respiratory infectious diseases: a systematic review and analysis. *Am J Epidemiol* 2014;180:865–75.
22 Schneeweiss S, Rassen JA, Glynn RJ, *et al*. Supplementing claims data with outpatient laboratory test results to improve confounding adjustment in effectiveness studies of lipid-lowering treatments. *BMC Med Res Methodol* 2012;12:180.
23 Johnson EK, Nelson CP. Values and pitfalls of the use of administrative databases for outcomes assessment. *J Urol* 2013;190:17–8.
24 Ellis RP, Martins B, Rose S. Chapter 3 - risk adjustment for health plan payment. In: McGuire TG, vanRC, eds. *Risk Adjustment, Risk Sharing and Premium Regulation in Health Insurance Markets [Internet]*. Academic Press, 2018: 55–104. Available: https://www.sciencedirect.com/science/article/pii/B9780128113257000038