RESEARCH ARTICLES

Using open source data to estimate the global epidemiology of pertussis

Haley Stone[1], Aye Moa[1], Raina MacIntyre[1], Abrar Chughtai[2]

[1] Biosecurity Program, Kirby Institute, University of New South Wales, Sydney, Australia
[2] School of Public Health and Community Medicine, University of New South Wales, Sydney, Australia

Abstract

Introduction: Pertussis is a highly infectious disease that remains endemic despite rising vaccination rates globally. Due to the lack of global surveillance data for pertussis, the unconventional use of open-source data gives a glimpse into global outbreaks, compensating for the lack of national reporting systems in some countries. The objective of the study is to describe the global reporting of pertussis through open source data.

Methods: An open-source database, EpiWATCH was used to analyse global outbreaks of pertussis. Data was retrieved on pertussis and analysed on multiple epidemiological factors from 2016 to 2019. In addition, incidence rates were calculated for each country and compared to the World Health Organization's (WHO) public domain data on global reported cases.

Results: A total of 96 reports were collected globally between the years 2016 to 2019. Of those reports, 95.8% (92/96) were from high-income countries. Data from the United States comprised 59.3% (57/96) of the total reports. In addition, prevalence rates were calculated for each country and compared to the WHO's public domain data on global reported cases. An outbreak report was identified in Papua New Guinea, which was not reported in the WHO's surveillance.

Discussion: Open-source data gives insight and analysis on pertussis outbreaks globally, given there is no formal global surveillance system for pertussis. There is a bias toward reports from high income countries in open source data. However, the timeliness of reporting coupled with assisting countries with lacking national reporting systems are benefits of open source data.

Introduction

Pertussis is a respiratory disease caused by the bacterium *Bordetella pertussis.* In 2018, the global annual incidence rate for pertussis was estimated to be 2.17 per 100,000 persons [1]. This estimate is based on global reported case numbers. However, there may be an underreporting of cases due to weak health systems and poor surveillance infrastructures in many countries. Mortality rates are similarly difficult to estimate as the Civil and Vital Registration Systems (CVRS) in many low-to-middle-income countries (LMIC) are extremely limited [2, 3]. Also, the time between infection and the onset of classic clinical features in children in countries with high comorbidities and concomitant illnesses, may result in pertussis being undetected as a cause of death. In high-income countries (HIC), deaths and hospitalisations due to pertussis is linked to children under the age of eight weeks [3]. As the first dose of vaccination for pertussis is not administered until two months of age, this is the most vulnerable age for children.

Classic and more severe clinical manifestations, such as the defining "whoop" cough, also known as the paroxysmal cough stage, often do not present until two weeks after the onset of symptoms in children [4, 5]. These symptoms may be absent, atypical or dampened in adolescents and adults [5, 6]. Symptoms can persist from one to six weeks, depending on the severity of the paroxysmal stage. Complications can range from pneumonia to neurological disorders, with infants under the age of 6 months being the most susceptible [4]. Due to adolescents and adults demonstrating atypical or asymptomatic presentations of the disease, under-reporting, and under-diagnosis are common [6]. With the absence of classic symptoms, adolescents and adults represent the primary source of infection to infants and children [6].

Transmission commonly occurs when an infected individual's respiratory droplets from a cough or sneeze come into contact with the mucous membranes of an uninfected individual [4]. The disease is highly infectious with an estimated range of reproductive numbers (r_0) of 12 to 17, which is variable depending on age-specific and locality data [7]. Second attack rates range from 80% to 100% in susceptible households [7]. Pertussis is a cyclical endemic disease that reoccurs every 2 to 5 years despite high coverage of the vaccination [5-7]. This high degree of recurrence indicates vaccination has little impact on the circulation of the disease [6].

Following vaccination, the risk of pertussis varies, dependent on the type of vaccine and efficacy. The pertussis vaccine is included in a combination vaccine,

which contains antigens for three diseases: diphtheria, tetanus, and pertussis (DTP). The Expanded Program on Immunization (EPI) program calls for three doses of DTP in an infant's first six months of life [8]. The vaccination uses selected antigens of the pertussis pathogen to induce an immune response in the host. There are two primary variations of the DTP vaccine, including the wholesale (DTPw) vaccine and the acellular vaccine (DTPa). The DTPw is an inactivated vaccination, while the DTPa is a subunit vaccination [7]. In 2018, global coverage for three doses of DTP was estimated to be 86%, which has remained at this level since 2016 [9].

Currently, there is no global epidemiological data on pertussis. Challenges in estimating the global pertussis disease burden are linked to three factors. First, there are limited surveillance systems established in many countries, with few resources being allocated to improve the coverage and accuracy of these systems [2]. The lack of surveillance systems impacts the timely collection of data and leads to underreporting the number of cases. HIC countries typically have a higher number of reported cases globally, with minimal cases reported in LMIC. As pertussis is not a notifiable disease in many countries, the case numbers are often under-reported [3]. In LMIC, incidence and attributable mortality rate data can be problematic to obtain accurately due to poor infrastructure and a lack of coverage in their civil registration systems. Where civil registration systems are non-existent or minimal, often the data is obtained from previous census data collections or sentinel surveys, neither of which provide an accurate picture of the burden of disease for pertussis.

The second challenge is access to adequate laboratory infrastructure and pertussis tests, especially in sparsely populated LMIC. An estimated 49.7% of LMIC populations live in rural areas [10]. In these rural areas, access to health systems, coupled with the timeliness of specimen collection and lengthy transport to laboratory centres, proves to be a significant challenge [2].

The third challenge for collecting accurate data on pertussis lies with health professionals [2]. Health professionals are not always aware of the full clinical manifestations of each age group. As such, cases are often not reported. Even if the health professional is aware, less severe and asymptomatic manifestations in adults and adolescents are often not detected.

Estimating the global burden of pertussis is quite difficult, driven by minimal global data on case numbers. The WHO reports on global pertussis case numbers, compiled from the individual country's official reports [1]. However, as illustrated by the challenges listed above, the case numbers do not reflect a comprehensive view of the burden of disease. For example, from 2016 to 2018, nearly 46% of LMIC did not have reported cases in the WHO dataset [1]. Of those LMIC countries reporting cases, a large percentage reported less than 100 cases per year [1]. The WHO data also lacks age-specific delimiters and does not provide any details on the location of the outbreak other than the country. In addition, the WHO and UNICEF report on global DTP 3 dose coverage [11].

Information from these coverage reports is combined with case numbers from given countries to understand opportunities for improving medical outcomes. However, for LMIC countries with limited health systems, the poor quality of case data makes analysis much more difficult.

Aims

The aim of the study is to describe the global epidemiology of pertussis using open source data.

Methods

Pertussis outbreaks globally were analysed for the timeframe of 2016 until the end of September 2019 using EpiWATCH outbreak data. EpiWATCH is an outbreak alert database [12]. The data is collected by monitoring, scanning, and critically analysing global outbreaks from open-source data. The data included is all publicly available information accessed through various means such as search engines, websites, and social media. The EpiWATCH database contains over 8000 report entries on a diverse range of infectious diseases gathered using this publicly available information from 2016 to 2019.

EpiWATCH collects data on pertussis using keywords "pertussis", "whooping cough," and "Bordetella". Geolocation tags are also retrieved and categorised into the dataset. News items that are not related to this topic and duplicates of the same event with identical information were excluded. In order to mitigate potential conflicting or overlap of case numbers for a given outbreak, the case total for an outbreak with multiple reports assigned was based on the latest reports total case count.

Within the collected database, data is further analysed and filtered on the keywords for pertussis dated between 2016-2019. For the analysis, all reported cases are grouped according to the country of reported cases, disease, and the time in which they occurred. Descriptive epidemiologic analysis of the outbreaks was conducted, including the size of outbreaks and mortality (if any reported). Additional public domain data from governments or the WHO was sought to compare with EpiWATCH data [13]. Lastly, prevalence was calculated by the number of cases gathered for a given country using EpiWATCH in year x divided by the total annual population using the formula:

$$Px = \frac{Annual\ Cases}{Annual\ Population}(100{,}000).$$

This was then repeated for reported the WHO's cases and mapped out in figure 2 and 3 respectively.

Results

Report Data

In the time frame of August 2016 to October 2019, 96 reports were gathered on pertussis in 12 countries globally: Australia, Canada, Colombia, United Kingdom, New Zealand, Denmark, Nigeria, Panama, Papua New Guinea (PNG), South Africa, United States and Vietnam. High income countries account for 95.8% (92/96) of the

pertussis reports in EpiWATCH, which correlates to those countries with the perceived highest burden of the disease. However, a small percentage, 4.2% (4/96) of reports, were gathered in LMIC.

Since open source data is not often specific on age of cases, an analysis of number of reports specifying school outbreaks was performed. Of the 96 reports, 52 (52.17%) of the reports were of school outbreaks. Many HIC, minus Panama and Denmark, reported greater than 25% of the outbreaks within a school, with the highest attributed to United Kingdom (100% - 1/1), Australia (77.8% - 7/9), and the United States (63.2% - 36/57).

Reporting

In Figure 2, the EpiWATCH pertussis reporting prevalence rate is graphed and compared to the WHO data mapped in Figure 3 for that year. While the obtained prevalence rates are directly comparable between EpiWATCH and the WHO data, the data obtained from EpiWATCH highlights areas where outbreaks are reported, which tends to reflect with an increase in WHO data. In addition, EpiWATCH was able to pick up cases that the WHO organization did not report. For example in 2018, the WHO identified Papua New Guinea (PNG) at zero cases [1]. However, EpiWATCH contradicted this data and established an incident rate of 0.29 per 1000,000 for PNG. Lastly, although the US comprised 59.3% of the reports, there was an overall low prevalence rate for the country in both EpiWATCH and the WHO. EpiWATCH reported US annual prevalence rates for 2016-2019 were 0.33, 0.25, 0.11, and 0.8 per 100,000 persons respectively. The WHO reported the US prevalence rate for 2016-2018 as 5.56, 5.84, and 4.11 per 100,000 persons, respectively. In both the EpiWATCH and WHO data, this seems to be an indication of under-reporting nationally in the US.

Timeliness

Timeliness of reporting outbreaks helps communities and national governments better prepare and deal with disease outbreaks. Timeliness of reporting is especially relevant to pertussis. EpiWATCH, in the context of pertussis, is quick to identify reports of outbreaks globally and work in tandem with other global reporting systems. In order to measure the timeliness of reporting, outbreaks were extracted from the United States EpiWATCH Pertussis dataset and compared to national and state public health alert reporting systems. Public alert news was scattered and often not consistent from state to state. However, each state has a Health Alert Network (HAN) system, which partners with the CDC HAN system. Most states limit access to HAN systems and post press releases to the general public on major outbreaks that are occurring within the state.

After comparing the cases with each state public health alert system in the USA in Table 1, a noticeably small number of individual outbreaks were reported by states. In many cases the states only reported when the outbreaks were in large numbers or affected multiple counties or parts of the state. Timeliness varied in the US, with some outbreaks detected earlier by open source reporting compared to official state reports.

**Figure 1.** Country Percentage of Reports. Cumulative from 2016 to 2019.
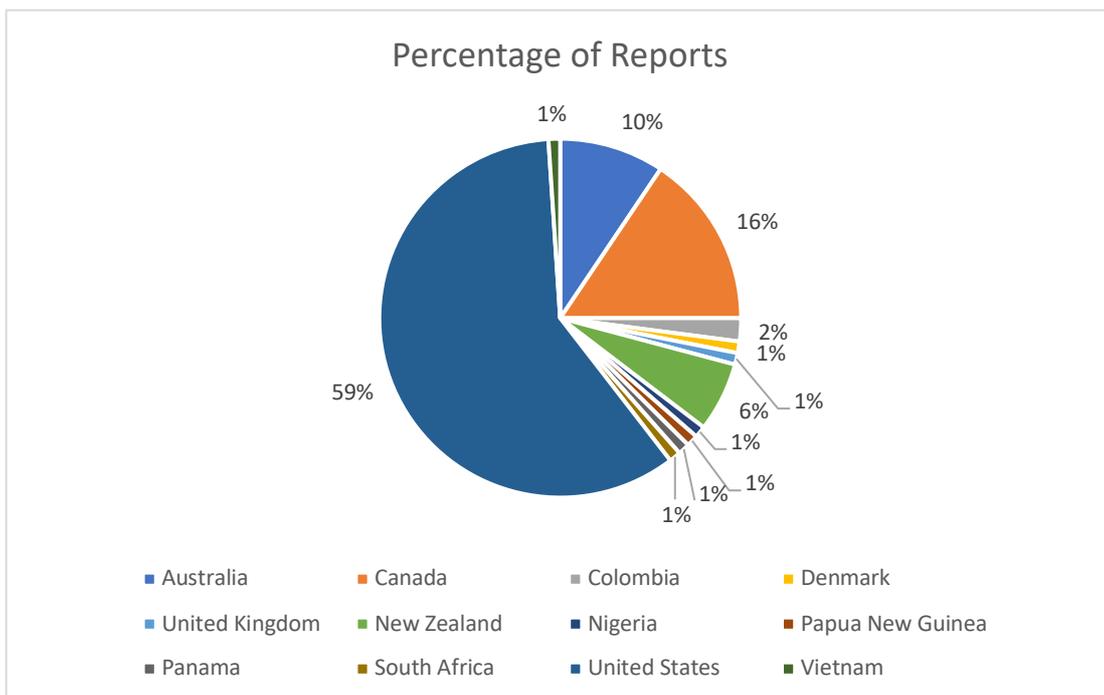
Figure 2. EpiWATCH Calculated Average Annual Prevalence rate (per 100,000).

Notes: Annual Prevalence was based on population estimates for the given country for the year. In addition, some countries had reported case numbers in one year and did not have reported cases in another. As the legend indicates, the darker the hue of red, the higher the annual prevalence, but varying depending on the year. The following ranges of prevalence are reported below: For 2016, the prevalence ranges from 1.6 to 0.33 per 100,000. For 2017, the prevalence ranges from 28.7 to 0.02 per 100,000. For 2018, the prevalence ranges from 1.5 to 0.03 per 100,000. In 2019, the prevalence ranges from 22.5 to 0.01 per 100,000. The countries in grey data were not gathered for comparison.

Figure 3. WHO Calculated Annual Prevalence Rates (per 100,000).
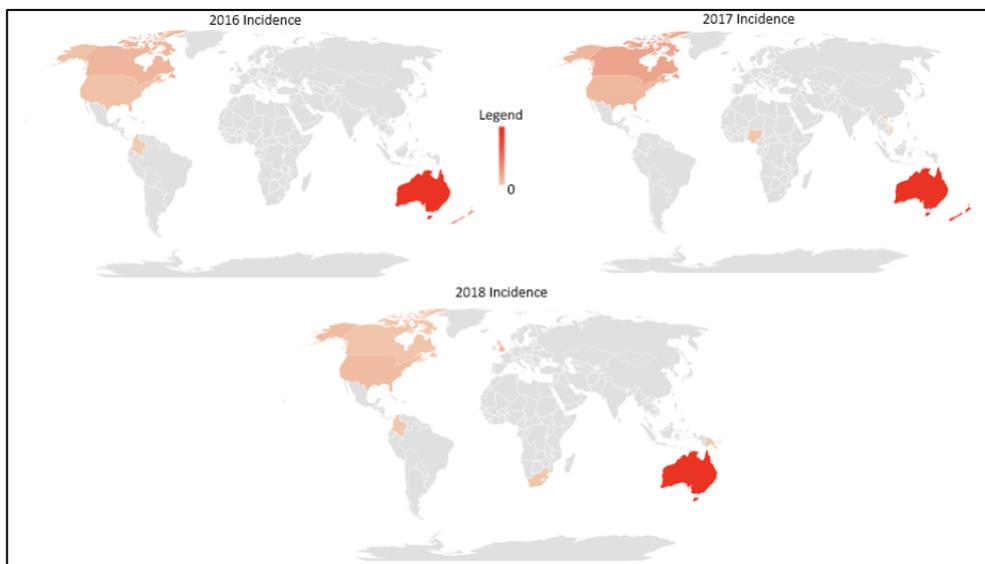
Notes: Annual Reporting Prevalence was based on population estimates for the given country for the year.
Countries chosen for 2016-2018 are based on the countries who reported cases in EpiWATCH for that year. Case data for 2019 from WHO was not published and was therefore unavailable for comparison. As the legend indicates, the darker the hue of red, the higher the annual prevalence, but varying depending on the year. The following ranges of prevalence are reported below: For 2016, the prevalence ranges from 82.8 to 1.1 per 100,000. For 2017, the prevalence ranges from 49.2 to 0.6 per 100,000. For 2018, the prevalence ranges from 50.2 to 0.0 per 100,000. The countries in grey data were not gathered for comparison.

Table 1. Summary of reported cases and outbreaks reported by State Governments or Centre of Disease Control United States 2016-2019.

EpiWATCH Data			State Reported Outbreaks		
Date	**State**	**Case Number**	**Date**	**State**	**Case Number**
19-Aug-16	Minnesota	-	22-Dec-16	Minnesota[1]	-
30-Sep-16	Missouri	3	-	-	-
22-Oct-16	Utah	-	-	-	-
1-Nov-16	Virginia	1			
4-Nov-16	Michigan	31	-	-	-
22-Nov-16	Wisconsin	31	-	-	-
23-Dec-16	New Mexico	20	20-Dec-16	New Mexico[2]	20
28-Dec-16	Ohio	3			
2-Feb-17	Kentucky	10	-	-	-
9-Feb-17	Pennsylvania	1	-	-	-
10-Feb-17	Kentucky	25	-	-	-
22-Mar-17	Wisconsin	9	-	-	-
30-Mar-17	Nevada	5	-	-	-
30-Mar-17	Washington	55	-	-	-
7-Apr-17	Maine	5	-	-	-
7-Apr-17	Utah	12	-	-	-
5-May-17	Alabama	6	4-May-17	Alabama[3]	6
9-May-17	Washington	2	-	-	-
20-May-17	Washington	21	-	-	-
21-May-17	New York	2	-	-	-
23-May-17	Alabama	19	22-May-17	Alabama[4]	19
16-Jun-17	California	14	-	-	-
16-Jun-17	Maine	1	-	-	-
23-Jun-17	Indiana	3	-	-	-
29-Jul-17	Indiana	136	27-Jul-17	Indiana[5]	136
14-Aug-17	Indiana	1	-	-	-
4-Sep-17	Alabama	5	-	-	-
4-Sep-17	Montana	5	-	-	-
14-Sep-17	Pennsylvania	1	-	-	-
15-Sep-17	Oregon	2	-	-	-
23-Jan-18	North Carolina	3	-	-	-
22-Mar-18	New Mexico	8	20-Mar-18	New Mexico[6]	15
17-Jul-18	Wisconsin	21	-	-	-
6-Aug-18	Alabama	31	3-Aug-18	Alabama[7]	9
20-Aug-18	Idaho	32	-	-	-
24-Aug-18	Delaware	97	27-Aug-18	Delaware[8]	97
22-Nov-18	Colorado	13	-	-	-
13-Dec-18	Indiana	-	-	-	-
9-Jan-19	Washington	14	-	-	-
25-Feb-19	California	-	-	-	-
28-Feb-19	California	12	-	-	-
23-May-19	Washington	6	-	-	-
21-Sep-19	California	5	-	-	-
21-Sep-19	California	2	-	-	-
25-Sep-19	Illinois	1	-	-	-
28-Sep-19	California	40	-	-	-

Note: Dashes refer to no outbreaks or case numbers reported. Individual outbreak data was pulled from the United States EpiWATCH data. Citations: 1.[14] 2.[15] 3.[16] 4. [17] 5.[18] 6.[19] 7.[20] 8.[21]

Discussion

In the absence of global epidemiological data on pertussis, EpiWATCH provides a snapshot of global outbreaks of pertussis. The overview data from EpiWATCH is not representative of the total prevalence and is an underestimate of the burden of disease and has a bias toward high income countries. While pertussis still occurs in LMIC, detection and surveillance methods might be lacking, but also other illnesses, such as measles, might have higher priority for the available resources within these countries [22-26]. However, using open source reporting provides early warnings of outbreaks, often as soon as the outbreak becomes known in the community. The purpose of this system is not to replace other systems of surveillance, but rather work in tandem with the national surveillance systems in the given countries and provide early warning. It may be particularly useful in small Pacific Island nations which

do not have rapid surveillance systems [27]. Half of the stakeholders in epidemic responses report lacking access to timely surveillance data, yet 90% do not utilise available open source systems such as HealthMap [27].

Pertussis outbreak response and preparedness vary by country and is a challenge to manage globally. Case definitions, as reported by the WHO, lack age-specific information when describing clinical manifestations, which can lead to underreporting. In addition, in HIC, studies have been published that note underreporting is universally prevalent [28-30]. The clinical manifestations of those above the age of the at-risk group (>1) are dampened in severity or asymptomatic and are often rationalised as not ill enough to present to a health system.

In HIC, National Government reporting is often delayed due to many factors, including case confirmations. Also, the system for validating notifications in HIC can be time consuming, which also delays the reporting of outbreaks. For example, in the United States, the CDC receives case information from the individual state's health department. However, as these data are often collected from the county level in each state, the process to collect and aggregate the data may delay release to the CDC. When an outbreak occurs with small numbers, data is not necessarily captured in national reporting.

While this is an issue in HIC, the absence of robust surveillance programs in LMIC lead to even longer delays. The lack of reporting from LMIC is a concern and may reflect a higher priority given to other communicable diseases, and lack of diagnostic tests in the community setting. Timeliness of reporting is dependent on availability of testing, laboratory confirmation, investigations and documentation of cases.

The strengths of these data vary between HIC and LMIC countries. In HIC, the timeliness of reporting is the largest benefit. While there is routine surveillance established in these countries, the outbreaks are delayed in official reporting on a state or national level. In LMIC, resources required for reporting outbreaks and surveillance systems are not as robust as those in HIC. Also, some countries do not have any available surveillance available at all for pertussis. For those countries, such as Papua New Guinea, EpiWATCH provides data on outbreaks when other data may be unavailable. For example, for PNG, the WHO reported no cases of pertussis in 2018. However, an epidemic in the Southern Highlands Province was identified by EpiWATCH, comprising of 26 cases and two deaths following an earthquake.

Limitations of this study are the potential biases linked to the EpiWATCH collection and data source. For example, all reports gathered in this study, with the exception of two, were in English, meaning the full scope of the disease globally might not be attained. There is also a bias towards reporting from HIC. With the addition of languages to the dataset, the increase in outbreak awareness globally should also increase. Open-source data gives a timely response with notification but is not meant to replace existing, validated surveillance systems, but to provide early warnings.

In summary, open-source data was used to give a summary of pertussis outbreaks globally. While many reports were located in HIC, these data were able to identify outbreaks not captured by other systems, whether national or global. In the absence of surveillance systems in these countries, open-source data can be used. In addition, open-source data is often more timely than national reporting systems and can be used to augment individual country's established surveillance systems.

References

1. World Health Organization. *Reported incidence time series*. 2019 [cited 2019; Available from: https://www.who.int/immunization/monitoring_surveillance/data/en/.
2. He, Q. and J. Mertsola, *Factors contributing to pertussis resurgence*. Future Microbiol, 2008. **3**(3): p. 329-39.
3. United Nations Statistical Division. *Demographic and Social Statistics* 2018 [cited 2019; Available from: https://unstats.un.org/unsd/demographic-social/crvs/index.cshtml.
4. Center of Disease Control. *Immunology and Vaccine-Preventable Diseases – Pink Book – Pertussis.* 2019 [cited 2019; Available from: https://www.cdc.gov/vaccines/pubs/pinkbook/downloads/pert.pdf.
5. Cherry, J.D., et al., *Defining pertussis epidemiology: clinical, microbiologic and serologic perspectives*. The Pediatric infectious disease journal, 2005. **24**(5): p. S25-S34.
6. Hegerle, N. and N. Guiso, *Epidemiology of whooping cough & typing of Bordetella pertussis*. Future Microbiol, 2013. **8**(11): p. 1391-403.
7. Kilgore, P.E., et al., *Pertussis: Microbiology, Disease, Treatment, and Prevention.* Clin Microbiol Rev, 2016. **29**(3): p. 449-86.
8. World Health Organization. *Routine EPI schedule*. 2019 [cited 2019; Available from: who.int/countries/eth/areas/immunization/routine/en/index3.html.
9. World Health Organization. *Pertussis.* 2018 [cited 2018; Available from: https://www.who.int/immunization/monitoring_surveillance/burden/vpd/surveillance_type/passive/pertussis/en/.
10. World Bank. *Rural Population (% of total population)* 2019 [cited 2019; Available from: https://data.worldbank.org/indicator/SP.RUR.TOTL.ZS.
11. WHO & Joint Reporting Form (JRF). *Immunisation Coverage GAVI* 2019; Available from: https://www.gavi.org/programmes-impact/country-hub.

12. NHMRC Centre for Research Excellence Integrated Systems for Epidemic Response. *EpiWATCH*. 2019 [cited 2019; Available from: https://iser.med.unsw.edu.au/epiwatch.
13. World Health Organization. *Data, statistics and graphics.* 2019 [cited 2019; Available from: https://www.who.int/immunization/monitoring_surveillance/data/en/.
14. Minnesota Department of Health. *Health Advisory: Pertussis Increase in Minnesota.* 2016 [cited 2019; Available from: https://www.health.state.mn.us/communities/ep/han/2016/dec22pertussis.pdf.
15. New Mexico Heath. *Pertussis outbreak Mckinley County* 2016; Available from: https://nmhealth.org/news/alert/2016/12/?view=522.
16. Alabama Public Health. *ADPH investigates pertussis outbreak in East Alabama.* 2017 [cited 2019; Available from: https://www.alabamapublichealth.gov/news/2017/05/04.html.
17. Alabama Public Health. *ADPH continues pertussis investigation in East Alabama.* 2017; Available from: https://www.alabamapublichealth.gov/news/2018/08/03.html.
18. Indiana State Department of Health. *STATE HEALTH OFFICIALS INVESTIGATING INCREASE IN WHOOPING COUGH.* 2017 [cited 2019; Available from: https://calendar.in.gov/site/isdh/event/state-health-officials-investigating-increase-in-whooping-cough/.
19. New Mexico Heath. *Pertussis Outbreak.* 2018 [cited 2019; Available from: https://nmhealth.org/news/alert/2018/3/?view=661.
20. Alabama Public Health. *Multiple Alabama counties experience pertussis outbreak.* 2018 [cited 2019; Available from: https://www.alabamapublichealth.gov/news/2018/08/03.html.
21. Delaware Health Alert Network. *UPDATE TO PERTUSSIS OUTBREAK IN KENT COUNTY, DELAWARE.* 2018 [cited 2019; Available from: https://www.dhss.delaware.gov/dhss/dph/php/alerts/dhan392.html.
22. Barger-Kamate, B., et al., *Pertussis-Associated Pneumonia in Infants and Children From Low- and Middle-Income Countries Participating in the PERCH Study.* Clinical Infectious Diseases, 2016. **63**(suppl_4): p. S187-S196.
23. Guiso, N. and F. Taieb, *Pertussis in Low and Medium Income Countries: A Pragmatic Approach.* 2019.
24. Maertens, K., K. Edwards, and E.E. Leuridan, *Pertussis*, in *Maternal Immunization.* 2020, Elsevier. p. 169-190.
25. Muloiwa, R., et al., *The burden of pertussis in low- and middle-income countries since the inception of the Expanded Programme on Immunization (EPI) in 1974: a systematic review protocol.* Systematic Reviews, 2015. **4**(1): p. 62.
26. Muloiwa, R., et al., *Pertussis in Africa: findings and recommendations of the Global Pertussis Initiative (GPI).* Vaccine, 2018. **36**(18): p. 2385-2393.
27. Aurysia, H., et al., *Epidemic intelligence needs of stakeholders in the Asia−Pacific region.* Western Pacific Surveillance and Response Journal, 2018. **9**(4).
28. Crowcroft, N.S., et al., *Under-reporting of pertussis in Ontario: A Canadian Immunization Research Network (CIRN) study using capture-recapture.* PloS one, 2018. **13**(5).
29. Shaikh, R., et al., *Underreporting of pertussis deaths in the United States: need for improved surveillance.* Pediatrics, 1998. **101**(2): p. 323-323.
30. van den Brink, G., et al., *Bordetella pertussis: an underreported pathogen in pediatric respiratory infections, a prospective cohort study.* BMC infectious diseases, 2014. **14**(1): p. 526.