A scoping review on data integration in the field of infectious diseases, 2009-2018

Seulgi Kim[1] , Chulwoo Rhee[2] , Su Jin Kang[3] and Sangwoo Tak[3]

1. Graduate School of Public Health, Seoul National University, Seoul, South Korea;
2. Department of Preventive Medicine, College of Medicine, Seoul National University, Seoul, South Korea; 3. Institute of Health and Environment, Seoul National University, Seoul, South Korea.
Corresponding author: Sangwoo Tak, e-mail: tak.sangwoo@gmail.com
Co-authors: SK: ksk8916@snu.ac.kr, CR: rhee275@gmail.com, SJK: socsjk@gmail.com
Received: 15-03-2021, **Accepted:** 07-06-2021, **Published online:** 07-08-2021

Abstract

Background and Aim: Little is known about data integration in public health research and its impact. This study aimed to summarize known collaboration information, the characteristics of the datasets used, the methods of data integration, and knowledge gaps.

Materials and Methods: We reviewed papers on infectious diseases from two or more datasets published during 2009-2018, before the coronavirus disease pandemic. Two independent researchers searched the Medline and Global Health databases using predetermined criteria.

Results: Of the 2375 items retrieved, 2272 titles and abstracts were reviewed. Of these, 164 were secondary reviews. Full-text reviews identified 153 relevant articles; we excluded 11 papers that did not meet our inclusion criteria. Of the 153 papers, 150 were single-country studies. Most papers were from North America (n=47). Viral diseases were the most commonly researched diseases (n=66), and many studies sought to define infection rates (n=62). Data integration usually employed unique national identifiers (n=37) or address-based identifiers (n=30). Two data sources were combined (n=121), and at least one data source typically included routine surveillance information.

Conclusion: We found a growing usage of data integration in infectious diseases, emphasizing the advantages of data integration and linkage analysis, and reiterating its importance in public health emergency preparedness and response.

Keywords: data integration, infectious disease, national policy, public health, scoping review, surveillance.

Introduction

The public health and socioeconomic consequences of the past outbreaks and the current coronavirus disease 2019 (COVID-19) pandemic emphasize that the prevention and mitigation of outbreaks are critical. Outbreaks of West African Lassa fever, Ebola virus, Middle East respiratory syndrome coronavirus, and COVID-19 have shown that humans, animals, and the environment contribute to the development and spread of emerging infectious diseases [1-4].

The "One Health" initiative emphasizes the need for a collaborative and multisectoral approach to improving public health surveillance and risk management. International health organizations, including the World Health Organization and the Centers for Disease Control and Prevention, have long made it clear that cooperation among agencies and ministries in terms of data integration/sharing is of fundamental importance when seeking to increase public health

surveillance capacity [5-7]. Collaboration efforts among governmental public health agencies have shown that no single actor/agency has all the knowledge and capacity about emerging infectious diseases. Such diseases are complex and associated with many uncertainties. Enhancements of national and global surveillance systems that collect and analyze the interconnections between people, animals, and the shared environment are critical when trying to prevent or rapidly respond to infectious diseases. Governments, scholars, and organizations worldwide have paid increased attention to public health surveillance and risk management. Recent advances in informatics have also greatly aided data integration [4,8].

Although environmental factors and zoonoses have become matters of public health concern, requiring multisectoral, the "One Health" initiative, including risk monitoring through public surveillance systems, and surveillance systems remain independent [1,9,10]. Despite the increasing interest in multisector collaboration/data integration worldwide [11,12], little is known about how data are integrated by public health researchers and the impact of such work.

This study aimed to review recent publications on infectious diseases and summarize information on sectors that collaborate the characteristics of the

dataset, the methods of data integration, and knowledge gaps.

Materials and Methods

Ethical approval

This study is based on a literature review and hence, ethical approval is not necessary.

Study period and location

Data were extracted and interpreted from July 2019 to August 2019 at Seoul National University.

Scoping review

A scoping review seeks to define the direction and implications of future research by analyzing existing trends [13]. Such a review identifies research questions, searches for appropriate studies, selects studies using predefined criteria, and gathers and summarizes data [14,15].

Research questions

How has data integration in public health research on infectious diseases been presented? This study specifically explores the characteristics of the datasets used, the methods of data integration, and knowledge gaps.

We systematically searched the Medline and Global Health databases for papers published between January 1, 2009, and December 31, 2018, on data integration using search terms that had been identified by a group of experts in epidemiology and infectious diseases (Table-1).

Inclusion criteria

The use of two or more different data sources to generate new information promoting public health, with a specific focus on infectious disease under the following headings: Public health, infectious disease, collaboration, and study findings.

Exclusion criteria

Incident reports, modeling studies lacking actual data, reviews and perspectives, letters to editors, clinical trials, reports on laboratory procedures, and publications in languages other than English were excluded from the study.

Literature selection

Two researchers independently reviewed the article titles and abstracts of all the initial search results. Articles considered eligible by either reviewer underwent a full-text review. If the two reviewers disagreed on inclusion, a third reviewer made the final decision.

Data charting and synthesis

Excel program, developed with expert consultation from senior researchers, was used for data charting. Then, the following information was extracted: (1) General information (year of publication, country of study, areas of study, and objectives), (2) the research topic (infectious disease classification according to the pathogen agent), (3) database characteristics (numbers and types of data sources), (4) integration methods, and (5) collaboration information. The synthesis includes quantitative analysis (frequency) and qualitative analysis (thematic) to identify topics, objectives, methods, and gaps in data integration and/or linkage in infectious disease research. In addition, to identify and categorize sectors from which data were provided, we assigned datasets containing key information (results) to Group 1 and the other dataset to Group 2. If three or more datasets were integrated, the two principal datasets utilized for or associated with the primary study objective were selected. All datasets were classified into nine categories of sectors (A-I) in terms of data types and combinations. Finally, we used modified questions from the REporting of studies Conducted using Observational Routinely Collected Health Data (RECORD) checklist [16] to evaluate the papers and used four questions in the RECORD.

Results

Of 2375 initially retrieved papers, 2272 titles and abstracts were reviewed (103 excluded: Not in the English language). Of these, 164 were classified as secondary full-text reviews, excluding 2108 papers (1378 single-dataset papers, 347 review articles, 120 modeling studies, 114 laboratory procedures, 69 outbreak investigation reports, 16 clinical trials, 15 case reports, 49 other works including non-communicable disease articles, and study protocols). After excluding 11 papers that did not meet our inclusion criteria, we collected data from 153 articles (Figure-1).

We reviewed 153 papers published from 2009 to 2018; most (n=150) were single-country studies. Table-2 presents the number of studies by country and only includes studies conducted in a single country (n=150). In Table-2, the United States accounted for the highest number of reports (37) followed by the United Kingdom (21) and Canada (10 cases). Reference to the World Bank national income standards [17] showed that high-income countries accounted for 107 reports; only two were from low-income countries.

Table-1: Search strategy (predefined search terms).

Domain	Keywords
Public health	Surveillance, epidemiology*, public health, monitor*, disease control
Infectious disease	Infection*, communicable
Collaboration	Intersectoral, multisectoral, interagency, interdisciplinary, multidisciplinary, cross-sector, transdisciplinary, collaborate*, link*, integrate*
Findings	Prevalence, incidence, morbidity, mortality, coverage, demography*, attack rate, effectiveness, efficacy, odds ratio, relative risk, sensitivity, specificity, disease burden, survival rate, indicator, distribution, attributable rate, validity, determinant*, bias, confound*, mediate*, associate*, relation*, interaction, causation, correlate*, inference

*=An asterisk is used for multiple character searching.
?=A question mark is used for single character searching

Table-3 shows that most studies explored viral diseases (n=66), including infection with the human immunodeficiency virus (HIV). Data integration was applied to estimate incidence and mortality (n=62), identify risk factors (n=49), assess program impact (n=20), and perform spatial analysis (n=17). In most data integration approaches, unique national identifiers (n=34) or address-based identifiers (n=30) were used.

In Table-4, we summarize the types of sectors from which datasets were provided and the pairs of datasets used. Routine surveillance (n=96) was the most frequently used primary source of the dataset (Group 1) followed by physical/clinical data (n=20) and disease registry data (n=12). For the secondary dataset (Group 2), routine surveillance (n=45) again showed the most commonly used data source followed by hospital/clinical data (n=34). A total of 28 papers integrated datasets derived from two different routine surveillance sources, and 25 papers used routine surveillance and hospital/clinical datasets. Table-5 shows the results of the evaluation of selected papers using items from RECORD. Most of the papers described the name of the dataset (n=149), geographic region (n=143), and timeframe (n=130) in the title or abstract; 10 studies did not include information on data integration.

Discussion

We reviewed and analyzed studies that sought to extend current public health knowledge in the field of infectious diseases by integrating two or more data sources. The novel information may not have been attainable because the data sources were analyzed individually. Our work emphasizes the potential utility of data integration and identifies gaps in how research is performed and reported.

Table-2: Number of studies by country and World Bank income[†] classification.

Parameter	Count (%)
Country (Top 10)	
United States	37 (24.7)
United Kingdom	21 (14.0)
Canada	10 (6.7)
Australia	9 (6.0)
Kenya	7 (4.7)
South Africa	6 (4.0)
Norway	5 (3.3)
Taiwan	5 (3.3)
Thailand	5 (3.3)
India	4 (2.7)
World Bank income[†]	
High	107 (71.3)
Upper-middle	22 (14.7)
Lower-middle	19 (12.7)
Low	2 (1.3)
Total[‡]	150

[†]World Bank National Income Standards, fiscal year 2021. [‡]Single-country studies only

Table-3: Number of studies by infectious disease agent, primary objective, and data integration method.

Parameter	Count (%)
Pathogenic agent	
Viral diseases	66 (43.1)
Bacterial diseases	38 (24.8)
Parasites	12 (7.8)
Fungi	1 (0.7)
Multiple/other	36 (23.5)
Primary objectives	
Incidence/mortality/rate estimation	62 (40.5)
Risk factor analysis	49 (32.0)
System/program operation and evaluation	20 (13.1)
Geospatial distribution	17 (11.1)
Economic burden or quality-of-life-related estimation	2 (1.3)
Other	3 (2.0)
Data integration method	
National unique identifier	34 (22.2)
Address-based identifier	30 (19.6)
Epidemiological correlation (not individually linked)	20 (13.1)
Unique study identifier	11 (7.2)
Not mentioned	10 (6.5)
Other	48 (31.4)
Total	153

Table-4: Matrix of data integration scheme.

Group 2[‡]	Group 1[†]								
	A	B	C	D	E	F	G	I	Total
A	28	8	5	1	1			2	45
B	25	3		4	1	1			34
C	7	2	4		1	2		1	17
D	11	3	2						16
E	10			2				1	13
F	3	1			1				5
G	1		1						2
H		1						1	2
I	11	2			2	1	1	2	19
Total	96	20	12	7	6	4	1	7	153

[†]Group 1: Principal dataset, [‡]Group 2: Secondary dataset. A=Routine surveillance, B=Hospital/clinical data, C=Disease registry data, D=Vital statistics, E=Claims data, F=Cohort data, G=Environmental data, H=GIS/address data, I=Other

Table-5: Number of studies using modified RECORD items.

Modified RECORD items	Yes	No	NA
The type of data used should be specified in the title or abstract When possible, the name of the database used should be included	149	4	
If applicable, the geographic region within which the study took place should be reported in the title or abstract	143	10	
If applicable, the geographic timeframe within which the study took place should be reported in the title or abstract	130	23	
If a linkage between databases was in play, this should be clearly stated in the title or abstract	112	31	10[†]

[†]NA indicates that no linkage was apparent in the title or summary, but the linkage was evident in the main text
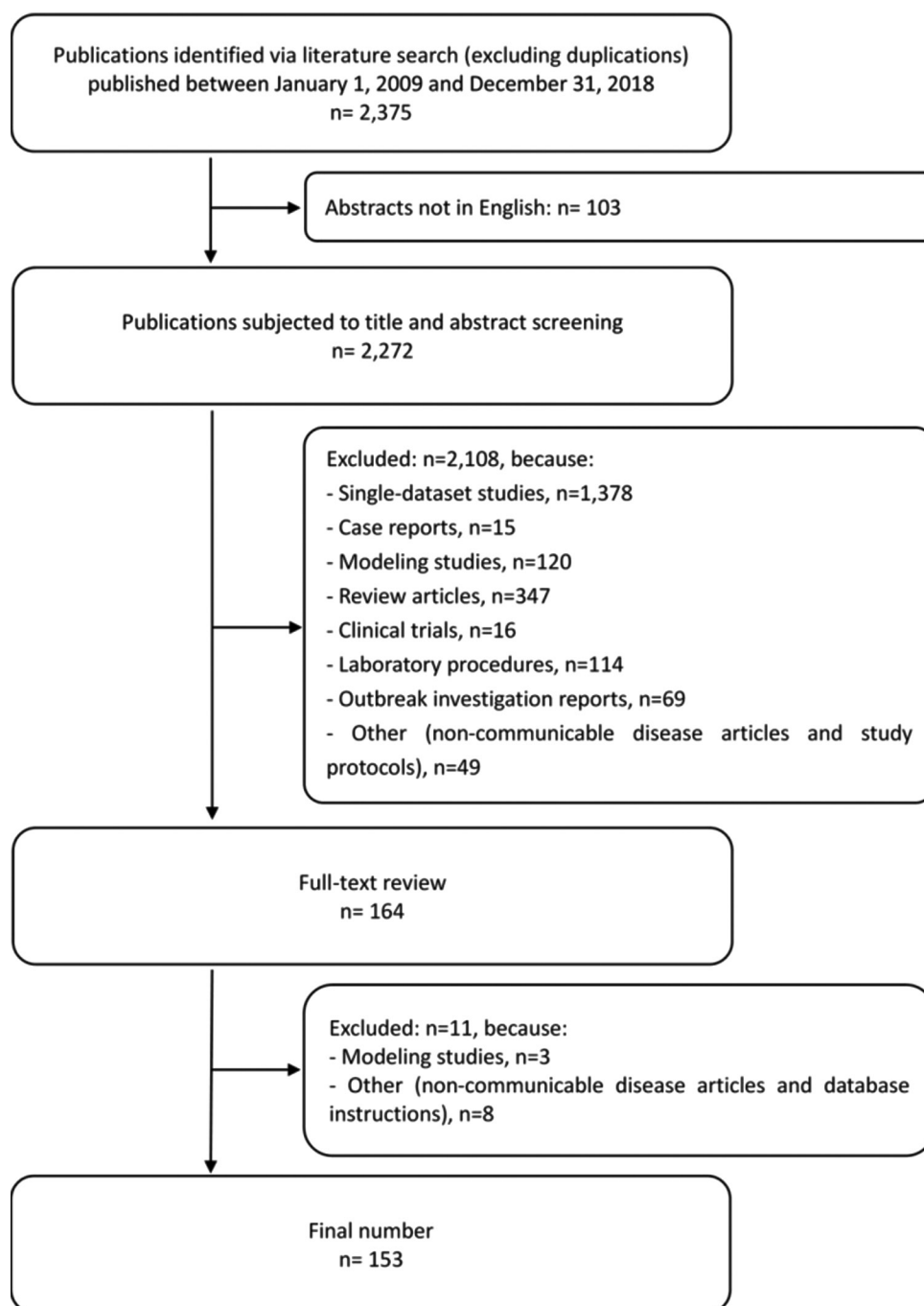
Figure-1: Flowchart (study selection).

First, we found that the use of data integration when studying infectious diseases was very rare in low-income countries that lack resources, deliver suboptimal care, and have poor public health systems. In such countries, the infectious disease burden is high and outbreaks occur frequently. However, it should be noted that data integration for public health surveillance may already exist in these countries, but for internal country use and not necessarily for wider research purposes. Countries with a lack of resources often face barriers and disincentives to publications (e.g., technology, cost, and publishing processes), leading to some publication bias. Given the global mobility of the population today, systematic cooperation among government agencies has become essential [11,18]. Of the studies reviewed, two were conducted in Ethiopia. One linked malaria infection records to local health information, identified differences in transmission patterns, and generated data aiding the development of future strategies [19]. Another study linked temporal data to spatial information when exploring the path of malaria propagation, yielding primary data facilitating appropriate preventative interventions [20]. The findings suggest that examining the current status of establishment and linkage in monitoring systems across low-income countries should be a priority.

Second, most studies that linked data explored viral diseases, including HIV. New human viruses

will continue to emerge in the absence of any vaccine or a known therapeutic intervention. Virus-related research is crucial; we must understand the basic biology of new viral pathogens and human susceptibilities [21,22]. An effective, global surveillance system for novel viruses is urgently needed to track future transmission and help inform vaccine needs and development. In addition, a well-established national surveillance system can help clinicians to promptly detect and respond to these threats and helps establish local vaccination strategies.

Third, in terms of sources used for data integration, country registration numbers served as effective links connecting various data sources through individual numbers, names, and ages. Government-organized data linkages are possible in Australia, the United Kingdom, and North Europe. In Australia, only an "integrating authority" can link data provided by researchers belonging to a data integration partnership, whereas Scotland has a "Health and Social Care Data Integration" framework [23,24].

One of the key advantages of data integration is that it can proceed to levels beyond what is possible with individual data sources. Population sizes are expanded and the denominators increase, enhancing the reliability of the disease-related estimates. The combination of routine surveillance and hospital data can reduce information bias or misclassification bias by including information from laboratory and clinical examinations. For example, in Scotland, mortality data collected over many years were linked to hospital data on *Clostridium difficile* infections, yielding an incidence rate calculated from unbiased population-based data [25]. In addition, data integration helps identify new factors that must be considered when establishing or prioritizing public health policies. For example, in the USA, community risk factors for *Campylobacter* infection, a food-borne disease (which had been ignored in the past), were assessed by integrating *Campylobacter* surveillance data with socioeconomic and environmental information [26].

In connection with the "One Health" initiative, data integration can be utilized for the control and prevention of zoonotic or vector-borne diseases by monitoring or exploring potential risk factors for humans. For instance, a Q-fever outbreak in goats in the Netherlands has led to thousands of human cases. In such cases, the data integrated from veterinary sectors and services could have enabled public health authorities to monitor outbreaks to avoid human cases [27]. This merged data from multiple disciplines and sectors will reduce the time to detect the source of infection and identify the animal host reservoir of human disease or zoonotic disease. In addition, practical resources can be shared and interpreted through data integration to strengthen cross-sector surveillance collaboration.

Finally, data integration not only yields useful insights but also extends the spatiotemporal scope of research. For example, a study from South Africa combined HIV monitoring and demographic data to assess the survival rate of children based on maternal HIV status. Such integrated data are important in terms of HIV treatment planning and adequate resource allocation [28]. Another study from Australia examined geographic clusters of babies born before arrival (BBA) by linking demographic and geographic information, thus identifying relevant geographic BBA status over time [29].

However, if the dataset communication standards vary, it is necessary to verify the accuracy of the data integration. Are the definitions of variables identical in both datasets? Were the principal data adequately identified? As data integration is not automated, researchers must critically examine every step of integration, including implementation and verification. Despite these difficulties, the expansion of research scope, especially that of large-scale epidemiological studies of infectious diseases, yields valuable new information on the pattern and direction of disease spread. Our ethical considerations included the need for patient privacy, data confidentiality and security, and respect for intellectual property.

Limitations

Our study had several limitations. First, we may have missed relevant papers despite our efforts to include broad search terms to find studies featuring data integration or the use of a multisectoral approach. Most of our results are presented within the public and medical health systems. We assume that this may be due to a lack of understanding or knowledge between human and animal or environmental sectors in terms of data features, such as types of surveillance, different entry points, and different search terms. We suggest that future studies include specific search terms or focus on specific diseases. Next, some papers did not adequately identify the sources of the datasets employed; we used our discretion to fill these gaps based on comments in the Discussions or Acknowledgments or the characteristics of the institutions involved (e.g., academia and government public health agencies). In addition, we were not able to look into how data analysis was carried out in the selected papers due to insufficient methodological descriptions in the papers. Future research should assess and report data analysis techniques for data integration. Finally, it is worth noting that there may be diverse routes to improving health security through data integration (e.g., early warning forecasting systems, rapid response, and deployment of resources, preventing spill-over risk, reducing the incidence, reducing cost, etc.). Since these are often not reported in the methods, future research should look into gray literature or government documents/publications.

Conclusion

Our findings identify the growing usage of data integration in infectious diseases, emphasizing the advantages of data integration and linkage analysis, and reiterating its importance in public health emergency preparedness and response. It is necessary to progress beyond the use of traditional public health data sources. Hence, it is recommended that at least two such sources should be integrated through collaboration among stakeholders to enhance public health surveillance. The results may garner continued technical and financial support for the "One Health" initiative; we provide a synopsis of the potential advantages of data integration and linkage in daily practice.

Authors' Contributions

SK, CR, and ST: Conceptualized the review article. SK and CR: Conducted the bibliography research and drafted the manuscript. SK and SJK: Drafted, reviewed, and revised the manuscript. All authors contributed intellectually and approved the final manuscript.

Acknowledgments

Competing Interests

The authors declare that they have no competing interests.

Publisher's Note

Veterinary World (Publisher of International Journal of One Health) remains neutral with regard to jurisdictional claims in published institutional affiliation.

References

1. Mwangi, W., De Figueiredo, P. and Criscitiello, M.F. (2016) One health: Addressing global challenges at the nexus of human, animal, and environmental health. *PLoS Pathog.*, 12(9): e1005731.
2. Taylor, L.H., Latham, S.M. and Woolhouse, M.E. (2001) Risk factors for human disease emergence. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 356(1411): 983-989.
3. Morens, D.M. and Fauci, A.S. (2020) Emerging pandemic diseases: How we got to COVID-19. *Cell*, 182(5): 1077-1092.
4. Ashour, H.M. (2014) One health--people, animals, and the environment. *Clin. Infect. Dis.*, 59(10): 1510-1510.
5. WHO Regional Office for Africa. (2019) Integrated Disease Surveillance and Response Technical Guidelines, Booklet One: Introduction Section. Licence: CC BY-NC-SA 3.0 IGO, Brazzaville.
6. Office of Public Health Scientific Services. Centers for Disease Control and Prevention. (2018). Public Health Surveillance: Preparing for the Future. Centers for Disease Control and Prevention, Atlanta, GA.
7. Lee, B., Martin, T., Khan, A., Fullerton, K., Duck, W., Kinley, T., Stoutenburg, S., Hall, J., Crum, M., Garcia, M.C., Iademarco, M.F. and Richards, C.L. (2018) Modernizing centers for disease control and prevention informatics using surveillance data platform shared services. *Public Health Rep.*, 133(2): 130-135.
8. Dos, S.R.C., Van De Burgwal, L.H.M. and Regeer, B.J. (2019) Overcoming challenges for designing and implementing the one health approach: A systematic review of the literature. *One Health*, 7: 100085.
9. Janati, A., Hosseiny, M., Gouya, M.M., Moradi, G. and Ghaderi, E. (2015) Communicable disease reporting systems in the world: A systematic review article. *Iran J. Public Health*, 44(11): 1453-1465.
10. Bagherian, H., Farahbakhsh, M., Rabiei, R., Moghaddasi, H. and Asadi, F. (2017) National communicable disease surveillance system: A review on information and organizational structures in developed countries. *Acta Inform. Med.*, 25(4): 271-276.
11. Kim, A.J. and Tak, S. (2019) Implementation system of a biosurveillance system in the Republic of Korea and its legal ramifications. *Health Secur.*, 17(6): 462-467.
12. Huff, A.G., Allen, T., Whiting, K., Williams, F., Hunter, L., Gold, Z., Madoff, L.C. and Karesh, W.B. (2017) Biosurveillance: A systematic review of global infectious disease surveillance systems from 1900 to 2016. *Rev. Sci. Technol.*, 36(2): 513-524.
13. Margevicius, K.J., Generous, N., Taylor-Mccabe, K.J., Brown, M., Daniel, W.B., Castro, L., Hengartner, A. and Deshpande, A. (2014) Advancing a framework to enable characterization and evaluation of data streams useful for biosurveillance. *PLoS One*, 9(1): e83730.
14. Nyanchoka, L., Tudur-Smith, C., Thu, V.N., Iversen, V., Tricco, A.C. and Porcher, R. (2019) A scoping review describes methods used to identify, prioritize and display gaps in health research. *J. Clin. Epidemiol.*, 109: 99-110.
15. Arksey, H. and O'malley, L. (2005) Scoping studies: Towards a methodological framework. *Int. J. Soc. Res. Method.*, 8(1): 19-32.
16. RECORD-Statement.Org. (2021) RECORD-Statement. org. RECORD Reporting Guidelines. Available from: http://www.record-statment.org/checklist.php. Retrieved on 23-05-2021.
17. Bank, T.W. (2020) World Bank Country and Lending Groups World Bank Data Help Desk. Available from: http://www.datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups. Retrieved on 23-05-2021.
18. Liu, D.P., Wang, E.T., Pan, Y.H. and Cheng, S.H. (2014) Innovative applications of immunisation registration information systems: Example of improved measles control in Taiwan. *Euro Surveill.*, 19(50): 20994.
19. Sena, L., Deressa, W. and Ali, A. (2014) Dynamics of *Plasmodium falciparium* and *Plasmodium vivax* in a micro-ecological setting, Southwest Ethiopia: Effects of altitude and proximity to a dam. *BMC Infect. Dis.*, 14: 625.
20. Alemu, K., Worku, A., Berhane, Y. and Kumie, A. (2014) Spatiotemporal clusters of malaria cases at village level, Northwest Ethiopia. *Malar. J.*, 13: 223.
21. Parvez, M.K. and Parveen, S. (2017) Evolution and emergence of pathogenic viruses: Past, present, and future. *Intervirology*, 60(1-2): 1-7.
22. Woolhouse, M., Scott, F., Hudson, Z., Howey, R. and Chase-Topping, M. (2012) Human viruses: Discovery and emergence. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 367(1604): 2864-2871.
23. Australia, D.I.P. (2020) Australian Government: Department of the Prime Minister and Cabinet. Available from: http://www.pmc.gov.au/public-data/data-integration-partnership-australia. Retrieved on 23-05-2021.
24. Integration, H.A.S.C. (2020) Public Health Scotland 2020: Public Health Scotland 2020. Available from: http://www.

https://www.isdscotland.org/health-topics/health-and-social-community-care/health-and-social-care-integration. Retrieved on 23-05-2021.

25. Banks, A., Moore, E.K., Bishop, J., Coia, J.E., Brown, D., Mather, H. and Wiuff, C. (2018) Trends in mortality following *Clostridium difficile* infection in Scotland, 2010-2016: A retrospective cohort and case-control study. *J. Hosp. Infect.*, 100(2): 133-141.

26. Goldstein, R.E.R., Cruz-Cano, R., Jiang, C., Palmer, A., Blythe, D., Ryan, P., Hogan, B., White, B., Dunn, J.R., Libby, T., Tobin-D'angelo, M., Huang, J.Y., Mcguire, S., Scherzinger, K., Lee, M.L. and Sapkota, A.R. (2016) Association between community socioeconomic factors, animal feeding operations, and campylobacteriosis incidence rates: Foodborne diseases active surveillance network (FoodNet), 2004-2010. *BMC Infect. Dis.*, 16: 354.

27. Zinsstag, J., Schelling, E., Crump, L., Whittaker, M., Tanner, M. and Stephen, C. (2020) One Health: The Theory and Practice of Integrated Health Approaches. CABI, Wallingford, Oxfordshire.

28. Ndirangu, J., Newell, M.L., Thorne, C. and Bland, R. (2012) Treating HIV-infected mothers reduces under 5 years of age mortality rates to levels seen in children of HIV-uninfected mothers in rural South Africa. *Antivir. Ther.*, 17(1): 81-90.

29. Thornton, C.E. and Dahlen, H.G. (2018) Born before arrival in NSW, Australia (2000-2011): A linked population data study of incidence, location, associated factors and maternal and neonatal outcomes. *BMJ Open*, 8: e019328.

****