Vaccine Decision Information System, Concept note for SAGE, 24 September 2018

Wilbert van Panhuis, MD PhD
Assistant Professor of Epidemiology and Biomedical Informatics
University of Pittsburgh
wilbert.van.panhuis@pitt.edu
+1 412 624 7693

Introduction

The amount and resolution of data collected in global health is rapidly increasing, creating new opportunities for data-driven decision making on disease prevention and control strategies. For example, national and international health agencies now collect vaccination coverage data at local level, for districts, municipalities, and schools, and infectious disease case data at the individual level. The unprecedented resolution of key health information currently available can give detailed insight in progress of health programs and can help technical advisory committees to make detailed strategy recommendations. For example, high-resolution vaccination coverage and case data can enable monitoring of locations, instead of countries, at high risk for disease outbreaks. International and national technical advisory committees and program managers can now make strategy recommendations to address, e.g., gaps in vaccination coverage rates in very specific locations. The increased amount and resolution of data in global health does not automatically translate in the effective use of data and comprehensive information management and analytical approaches will be required to efficiently and effectively use all available data for global health decisions.

Strategies for vaccination programs are often made at the national or international level, but implementation occurs at the local level. Targets for vaccination programs are often monitored at the national level, e.g., using a nationwide critical vaccination rate of 95% for measles to enable elimination of the disease. This critical vaccination fraction assumes that vaccination coverage and population mixing are distributed homogeneously throughout a country [1]. This assumption of homogeneity is not always realistic, as recently found in in Sub Saharan Africa [2]. Spatial heterogeneity of vaccination coverage can increase the critical vaccination fraction required for herd immunity and can delay disease elimination, as illustrated by continued measles outbreaks, even in countries with high average nationwide vaccination coverage rates [3].

Subnational vaccination coverage rates and disease surveillance data are now available for many countries. In the United States, many states have recently made school-level vaccination coverage data publicly available and at the World Health Organization (WHO), district-level vaccination coverage data are being collected since 2016. Disease surveillance data are also becoming more detailed. For example, data systems have been developed at WHO for individual case line listings for polio (Polio Information System, POLIS) and measles.

Subnational data are more complex compared to national summary statistics. For example, countries can have thousands of districts or municipalities that are often named differently in different data sources. Also, files can be organized in many different ways when representing a large number of datapoints, compared to a few summary statistics. A comprehensive and standardized information management system will be required to efficiently process, standardize, and integrate subnational vaccination coverage and case data.

From an analytical perspective, researchers have advanced methods for spatial interpolation of

areas without observations using machine-learning algorithms trained on observed data [4,5], and multiple analytical approaches have been used for estimating the risk of infectious disease outbreaks from spatial data ranging from clustering methods to mathematical pathogen transmission models [2,6,7]. Models can enable estimation of the proportion of the population susceptible to infection. Monitoring the susceptible population at the subnational level would provide detailed information about the risk of outbreaks at the local level that could support targeted recommendations on vaccination strategies.

Goals and specific aims

Our long-term goal is to improve the use of high-resolution information for global population health. The use of information can be improved by data science principles for data management and integration. The FAIR (Findable, Accessible, Interoperable, and Reusable) data principles are being advanced by the US National Institutes of Health and the European Science Cloud to improve integration and use of data in the health sciences [8]. Key elements of the FAIR data principles are: (1) assigning a unique identifier to each dataset that resolves to an online location where that dataset can be accessed, (2) rich metadata represented according to a standard vocabulary or ontology, (3) access to the dataset through a known and documented access protocol, possibly with authentication requirements, (4) data representation according to standards widely used in the community. **Our objective is to establish a Vaccine Decision Information System (VADIS) at WHO, guided by the FAIR data principles, to enable the effective use of high-resolution, subnational data for risk analysis and vaccine policy making.**

Our **specific aims** are to:
(1) Catalogue datasets with subnational vaccination coverage, disease surveillance, and other datasets available at the WHO Initiative for Vaccine Research (IVR) and its partners and to create standardized, machine-interpretable metadata for each dataset;
(2) Standardize datasets using commonly used vocabularies and ontologies, including annotation of the data collection methods and quality indicators;
(3) Integrate datasets into a database and country-specific dashboards that present relevant information in a format easily digested by vaccine advisory committees and analysts; and
(4) Design semi-automated information processing pipelines to efficiently collect, annotate, standardize, and integrate new datasets into the vaccine decision information system.

Dr. Wilbert van Panhuis, MD, PhD will lead the development of VADIS. Dr. Van Panhuis is an epidemiologist at the University of Pittsburgh (Pitt) and a data scientist in the NIH Big Data to Knowledge (BD2K) program. Dr. Van Panhuis has a decade-long, unique track record of improving access to global health data, best evidenced by his role as PI of Project Tycho (www.tycho.pitt.edu), a repository of all US national notifiable disease reports since 1888, which currently has over 3600 registered users [9]. Dr. Van Panhuis also has the proven ability to design and successfully lead large-scale collaborations. For example, he personally launched and led an international partnership between eight governments and 29 investigators from Southeast Asia for sharing and analyzing dengue surveillance data [10]. Building on these examples of his leadership, Dr. Van Panhuis will proactively work with WHO and other partners to develop VADIS.

Pilot Project: Subnational data for vaccine strategy recommendations

Objective: We collaborated with the WHO Initiative for Vaccine Research (IVR) during August and September 2018 to explore and visualize data on vaccination coverage and reported cases at subnational resolution for measles, yellow fever, and diphtheria. Our main goal was to assess the opportunities and requirements for using subnational data for risk analysis to strengthen

vaccination strategy recommendations at the subnational level by technical advisory committees.

Methods: We used a variety of datasets from different original sources and in a range of different formats available from IVR and partners. We explored the content of each dataset to determine the variables included and roughly the amount of missing data. Based on initial screening, we included case data for measles and diphtheria from surveillance systems (anonymized line listings), and for yellow fever from a HealthMap outbreak listing. We included DTP3 and yellow fever vaccination coverage data from the WHO Expanded Programme on Immunization (EPI) and vaccination coverage data from the Demographic and Health Surveys (DHS) for measles, diphtheria, and yellow fever. Most included datasets covered years as recent as 2017. We excluded six datasets (Yellow fever outbreak list and case line listings, national-level vaccination coverage rates, and non-routine immunization data) due to discordant time period (1980-2013) with other datasets, lack of subnational data, or non-routine immunization programs. We lacked time to standardize and use three datasets of which the content likely overlapped with some of the included datasets (AFRO vaccination coverage, yellow fever district-level vaccination coverage, and WHO-UNICEF joint reporting forms). For each dataset, we selected example countries for which data were available for most locations at the second administrative level (admin2), which correspond to the district level in most countries.

We used a standard process to prepare each dataset for visualization. We removed any variables not of immediate interest, aggregated data by first or second administrative level (admin1 or admin2) by taking the mean (vaccination coverage) or sum (reported cases), and standardized location names per the ISO-3166 and Geonames vocabularies. We matched the standardized location names to the Global Administrative Boundary (GADM) database that provides open access to spatial boundary files for locations. Many location names did not match to names in the standard vocabularies or GADM: 70-80% of admin1 names matched a standard, but only 20-30% of admin2 names.

Results: We visualized measles case data for Nigeria 2002-2009 at the admin2 level (**Fig. 1**). We were not able to match every district to a standard name or to spatial boundary definitions. The maps of Nigeria measles cases per district do not show any clear areas at risk, but illustrate the potential value of district-level maps as tool for subnational outbreak risk assessment.

We visualized yellow fever vaccination coverage data for Angola 2015-2016 at the admin2 level from EPI and the DHS (**Fig. 2**). We also visualized yellow fever case data for Angola 2016 at the admin1 level from the HealthMap outbreak database (**Fig. 2C**). The Angola maps also illustrate the potential value for subnational data from multiple sources for subnational vaccination strategy recommendations. The EPI vaccination coverage data for yellow fever show coverage below 50% in many districts in Southwestern Angola, the same area for which most of the yellow fever cases were detected in 2016 by HealthMap.

We visualized DTP3 vaccination coverage and diphtheria case data for the Philippines in 2017 at the admin1 level (**Fig. 3**). Although the maps indicate many areas of the country with DTP3 coverage rates below 75%, the vaccination coverage rates do not correspond well with the case data that show outbreaks in a few districts in the Manila area and in the South. No case data could be visualized for many admin1 areas due to mismatches between location names in the data and the standard vocabularies.

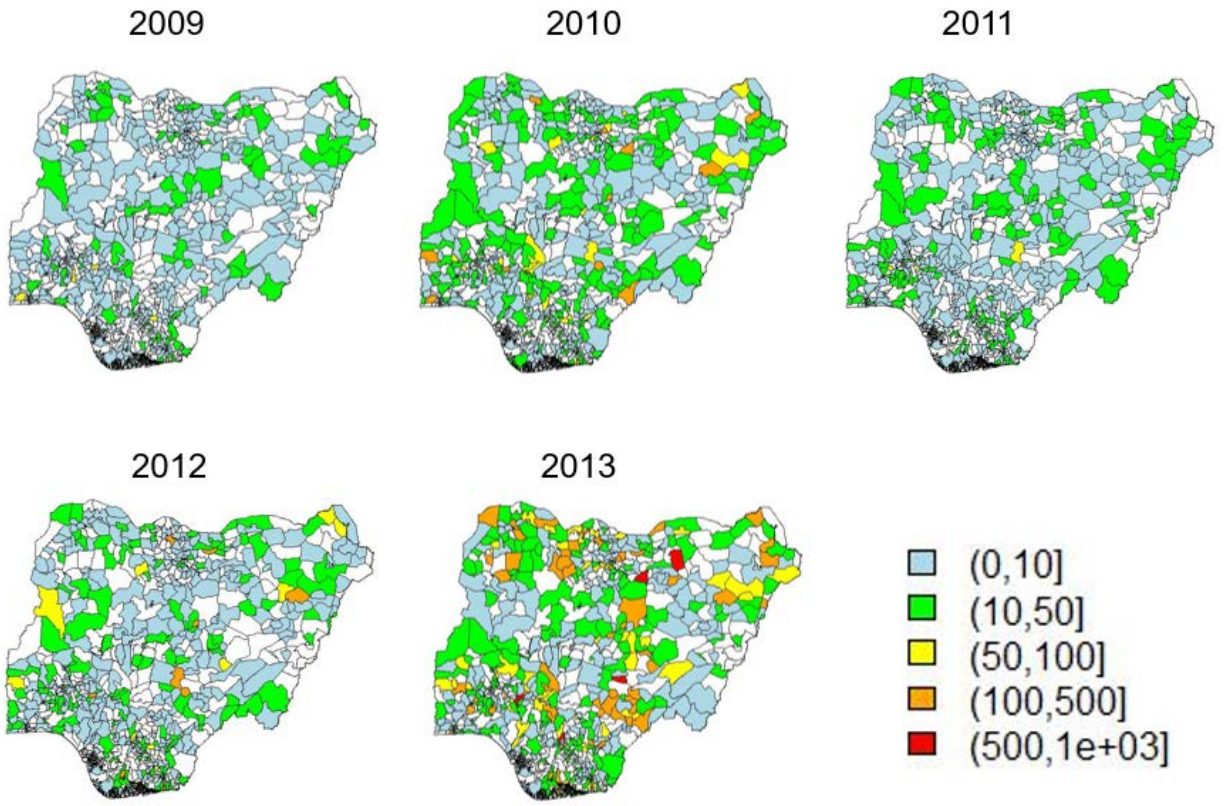

Figure 1, Measles cases per district for Nigeria: 2002-2009

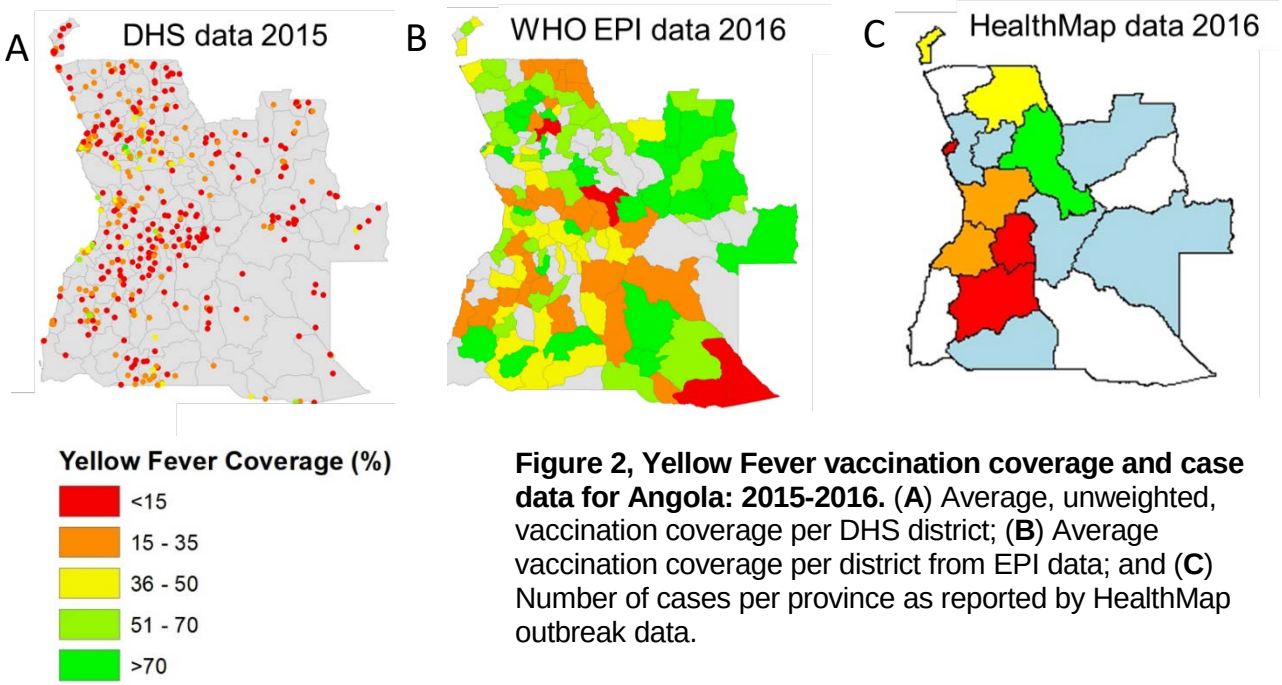

Figure 2, Yellow Fever vaccination coverage and case data for Angola: 2015-2016. (**A**) Average, unweighted, vaccination coverage per DHS district; (**B**) Average vaccination coverage per district from EPI data; and (**C**) Number of cases per province as reported by HealthMap outbreak data.

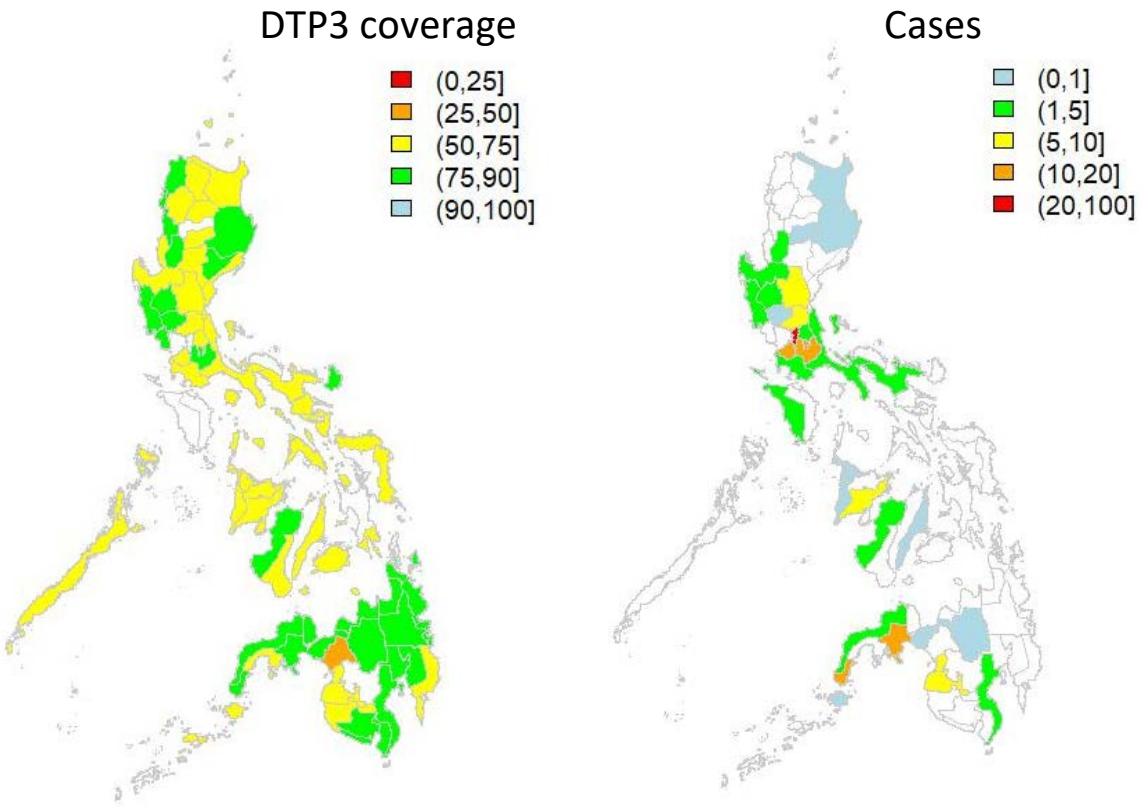

Figure 3, Diphtheria vaccination coverage and cases by district for Philippines: 2017

Lessons learned from the pilot project

(1) A large variety of data sources exist that are relevant for vaccine decision making. The same information could be derived from multiple sources, requiring source evaluation, selection and prioritization;
(2) The format and content of individual data files are often not standardized and can only be discovered from opening files and studying file content, requiring extensive dataset exploration, annotation, and cataloguing;
(3) The name for the same location can vary widely among datasets and location names and administrative levels in the data often do not correspond to widely-used international standards, such as ISO-3166 or Geonames, requiring extensive location standardization;
(4) Datasets often don't include complete codebooks or metadata explaining the meaning of variables, requiring additional information from data producers/owners about methods used to collect the data and about the variable definitions;
(5) Various partners maintain datasets that contain human subject identifiers, such as names, addresses, or birth dates, requiring ethical approval and safeguards of private information or anonymization before such datasets can be used;
(6) Many large datasets exist that can be useful for vaccine decision making, requiring a cloud service for data transfer/hosting/backups.

Conclusion

Valuable opportunities exist to use subnational vaccination coverage data, diseases surveillance datasets, and related data, such as migration data, to monitor gaps in immunity and susceptible populations at a high spatial resolution. Spatial interpolation and risk modeling methods have been advanced recently and could be systematically applied to support vaccination programs. To

effectively use subnational data, a comprehensive information management approach is needed including a review of data sources, data collection methods, and data quality and completeness. Datasets should be standardized and annotated systematically using semi-automated data processing pipelines. Once datasets have been standardized, a wide range of analytical methods can be applied by researchers and other partners to develop risk estimates at the subnational level that can be operationalized to improve spatial monitoring and targeting of vaccination programs.

References

1 Feng Z, Hill AN, Smith PJ, *et al.* An elaboration of theory about preventing outbreaks in homogeneous populations to include heterogeneity or preferential mixing. *J Theor Biol* 2015;**386**:177–87. doi:10.1016/j.jtbi.2015.09.006

2 Brownwright TK, Dodson ZM, Van Panhuis WG. Spatial clustering of measles vaccination coverage among children in sub-Saharan Africa. *BMC Public Health* 2017;**17**. doi:10.1186/s12889-017-4961-9

3 Zipprich J, Winter K, Hacker J, *et al.* Measles outbreak--California, December 2014-February 2015. *MMWR Morb Mortal Wkly Rep* 2015;**64**:153–4.

4 Utazi CE, Thorley J, Alegana VA, *et al.* High resolution age-structured mapping of childhood vaccination coverage in low and middle income countries. *Vaccine* 2018;**36**:1583–91. doi:10.1016/j.vaccine.2018.02.020

5 Pigott DM, Deshpande A, Letourneau I, *et al.* Local, national, and regional viral haemorrhagic fever pandemic potential in Africa: a multistage analysis. *Lancet (London, England)* 2017;**390**:2662–72. doi:10.1016/S0140-6736(17)32092-5

6 Shearer FM, Longbottom J, Browne AJ, *et al.* Existing and potential infection risk zones of yellow fever worldwide: a modelling analysis. *Lancet Glob Heal* 2018;**6**:e270–8. doi:10.1016/S2214-109X(18)30024-X

7 Pons-Salort M, Grassly NC. Serotype-specific immunity explains the incidence of diseases caused by human enteroviruses. *Science* 2018;**361**:800–3. doi:10.1126/science.aat6777

8 Wilkinson MD, Dumontier M, Aalbersberg IjJ, *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016;**3**:160018. doi:10.1038/sdata.2016.18

9 van Panhuis WG, Grefenstette J, Jung SY, *et al.* Contagious Diseases in the United States from 1888 to the Present. *N Engl J Med* 2013;**369**:2152–8. doi:10.1056/NEJMms1215400

10 Van Panhuis WG, Choisy M, Xiong X, *et al.* Region-wide synchrony and traveling waves of dengue across eight countries in Southeast Asia. *Proc Natl Acad Sci U S A* 2015;**112**. doi:10.1073/pnas.1501375112