GOOGLING GREY LITERATURE TO ESTIMATE BURDEN OF PNEUMOCOCCAL DISEASE: AN EXAMPLE FROM THE AGEDD PROJECT

Katie Gorham, Hope L. Johnson, Cristina R. Garcia, Orin S. Levine, Maria Deloria-Knoll, Katherine L. O'Brien
on behalf of the AGEDD Pneumococcal Burden Study Team
Johns Hopkins University Bloomberg School of Public Health

INTRODUCTION

- The Adult Global Estimation of Disease Burden and Distribution of Serotypes of Serious Pneumococcal and Meningococcal disease (**AGEDD**) **project identified relevant data to estimate the burden of invasive pneumococcal disease** (IPD) among older children and adults globally primarily **through systematic searches of the published literature**.
- **National surveillance data published on** national health agency **websites** is often **identified by** search engines like **Google**.
- **We characterized the availability, advantages, and quality of this supplemental data.**

METHODS

- Implemented targeted Google searches for known surveillance websites to identify national IPD data and compared the availability and usefulness of data identified from the web vs. systematic searches of the published literature.

RESULTS

IPD data is unexpectedly abundant and useful in Googled grey literature

- **IPD Data from 90 countries** (n=16 Africa, n=30 Europe, n=42 Americas, and n=2 Western Pacific) **identified from 24 websites**
- **Most data was from developed countries** with well-established surveillance infrastructure; **limited data was found for developing countries** (**Figure 1**)

FIGURE 1. Countries with IPD surveillance data available via websites

- **Types of data available**:
  - **IPD incidence/case data** for 44 countries
  - **Serotype data** for 49
  - **Only pneumococcal meningitis data available for Africa** (also available from literature search)
- **Efficiency in Identifying and abstracting surveillance data from websites**:
  - WHO, European Center for Disease Control (ECDC) and Pan American Health Organization(PAHO) report data also found on national sites
  - **Web-based surveillance data** containing nationally representative data for several decades **abstracted in hours vs. weeks for abstraction of published articles with the same data for a given country**
- **Web-based surveillance data often more detailed than peer-reviewed literature**:
  - Unconstrained report/webpage sizes allowed for more and further stratified data (e.g. IPD data by year, syndrome, specimen type)– often not possible for published literature due to journal guidelines for article length.
- **Limitations**:
  - Methods were sometimes less clear, complicating data quality assessment.
  - Challenges with lack of standardized reporting format, elaborated below.

IPD surveillance data formats vary by source

- Available IPD surveillance web-based data characterized as (**Table 1**):
  1. static pages (n=18) such as downloadable PDFs or static web pages; often difficult to abstract due to format requiring manual abstraction (i.e. cannot use copy/paste functions during abstraction) and possibly prone to more error;
  2. semi-static (n=1) downloads (e.g. Excel files) that could be manipulated somewhat and abstracted relatively easily;
  3. "queryable" databases (n=5) return easily abstractable datasets stratified for specific outcomes of interest.
- **Preference for queryable databases**
  - Databases especially useful in abstracting data by stratifications of interest (e.g. IPD incidence by syndrome and age group) which was impossible to do with many of the static and semi-static resources

"Googling" grey literature can supplement traditional systematic searches of the literature and is time efficient

- Quicker, easier, and often less error-prone abstraction of surveillance data saved time and resources:
  - Abstracted all available surveillance data and then filled in data gaps with published data
  - **IPD data for 23 countries was only identified from the web** (i.e. no published data identified in systematic search of >17,000 articles)
  - **75% of web-based surveillance data also had data available from the published literature** (i.e. peer-reviewed)
  - **Among 116 countries with IPD data identified from the published literature, no web-based surveillance data were identified for 47%** of countries (n=55).
  - Strategic **abstraction of web-based surveillance data saved us the abstraction of 91% of the identified published articles** (n=3541), although some outcomes of interest may be missing (**Figure 2**)

FIGURE 2: IPD surveillance data supplements traditional literature

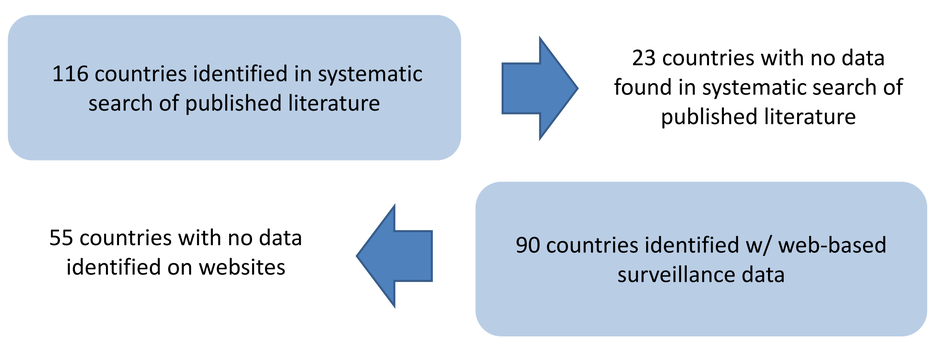

TABLE 1. Characteristics of IPD surveillance data

Format	Characteristics
Static	• PDFs or static web pages • Data often unable to be copy/pasted – more difficult to abstract, more opportunity for error • No opportunity for modification (i.e. multi-stratified abstraction) • Websites Identified by AGEDD search: 18
Semi-static	• Excel/CSF downloads • Data able to be copy/pasted • Opportunity for modification • Websites Identified by AGEDD search: 1
Queryable Database	• Fully functional, queryable databases • Data often able to be copy/pasted • Much opportunity for modification to match outcomes/strata of interest • Websites Identified by AGEDD search: 5

CONCLUSION

- **"Googling" websites increased the amount of IPD data available and was more efficient than abstraction of the published literature;**
- **Supplementing the traditional literature search** with data obtained from web-based surveillance data **reduced the resources required to abstract outcomes of interest .**
- **Queryable databases proved to be the most efficient and comprehensive** – as well as the least error-prone – **source of web-based data.**
- **Improvements in data formats and availability on websites could facilitate improved data capture for disease burden estimates, and thereby allow for more efficient timelines and budgets in similar projects.**

IVAC International Vaccine Access Center

Evidence ➤ Policy ➤ Access

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH