Opportunities and Challenges in Public Health Data Collection in Southern Asia: Examples from Western India and Kathmandu Valley, Nepal

Amruta Nori-Sarma, Anobha Gurung, Gulrez Shah Azhar, Ajit Rajiva, Dileep Mavalankar, Perry Sheffield and Michelle L. Bell

1 School of Forestry and Environmental Studies, Yale University, New Haven, CT 06511, USA; amrutasri.nori-sarma@yale.edu (A.N.-S.); ajit.rajiva@yale.edu (A.R.)
2 Department of Civil, Architectural and Environmental Engineering, University of Texas at Austin, Austin, TX 78712, USA; anobha.gurung@utexas.edu
3 International Centre for Mountain Development, Kathmandu 44700, Nepal
4 Pardee RAND Graduate School, Santa Monica, CA 90401, USA; gulrez_azhar@rand.org
5 Public Health Foundation of India, Indian Institute of Public Health—Gandhinagar, Gandhinagar, Gujarat 382042, India; dmavalankar@iiphg.org
6 Department of Environmental Medicine and Public Health and Department of Pediatrics, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; perry.sheffield@mssm.edu

Received: 24 May 2017; Accepted: 19 June 2017; Published: 26 June 2017

Abstract: Small-scale local data resources may serve to provide a highly resolved estimate of health effects, which can be spatially heterogeneous in highly populated urban centers in developing countries. We aim to highlight the challenges and opportunities of health data registries in a developing world context. In western India, government-collected daily mortality registry data were obtained from five cities, along with daily hospital admissions data from three government hospitals in Ahmedabad. In Nepal, individual-level data on hospital admissions were collected from six major hospitals in Kathmandu Valley. Our process illustrates many challenges for researchers, governments, and record keepers inherent to data collection in developing countries: creating and maintaining a centralized record-keeping system; standardizing the data collected; obtaining data from some local agencies; assuring data completeness and availability of back-ups to the datasets; as well as translating, cleaning, and comparing data within and across localities. We suggest that these “small-data” resources may better serve the analysis of health outcomes than exposure-response functions extrapolated from data collected from other areas of the world.

Keywords: capacity building; data collection; hospital admissions; India; mortality; public health; Nepal; South Asia

1. Introduction

The World Health Organization (WHO) and others have demonstrated a large, increasing health burden in developing countries, some of which is attributable to environmental factors [1,2]. However, corresponding difficulties exist in collecting high-quality health data to estimate locally-derived relationships between environment and health in regions without centralized systems for health records. Understanding how the environment affects human health in the present day is important, and may inform our understanding of the public health burden of many environmental exposures (e.g., air pollution, extreme weather conditions, infectious disease) which are anticipated to increase under a changing climate.
Existing datasets in hospitals and municipal bodies present vast potential resources in terms of cost compared to prospective studies and direct measurements. US studies delineating opportunities and limits of epidemiological research in hospital settings date back to 1965 [3]. However, developing countries do not have historical precedents for research using already collected health data. The WHO, the World Bank, and other global agencies have published guides regarding the importance of proper statistical training and the development of systems for health data collection in developing countries, for the purpose of epidemiologic and other research [4,5].

Secondary data sources on health endpoints, including hospital and registry data, have been powerful tools for research in developed countries in North America and Europe for both acute and chronic health outcomes across a wide range of exposures. Examples include studies at the community level, such as the use of union health records to understand the impact of chronic asbestos exposure on neoplasia outcomes among insulation trade workers in New York and New Jersey [6], as well as large studies across the US examining the impacts of ambient air pollution on health [7,8]. Health data in developing regions are needed for research aimed at understanding the local determinants of health. We believe that greater centralization of data and increased access to data within the scientific community is critical to epidemiological studies in these regions. This will require increased uniformity in definitions for health outcomes, as well as reporting procedures across communities. The benefits of such data collection may extend far beyond those of the original research.

Medical and public health research communities have recently emphasized “big data” and the development of methods to process and analyze large, complex datasets, with a growing number of research articles on bioinformatics, computational methods, and genomics [9–11]. Additionally, projects such as the Million Deaths Study in India use qualitative sampling and verbal autopsy to model the numbers of premature deaths for all causes among 14 million Indians residing in 2.4 million households in cities across the country [12]. The use of massive, high-dimension datasets can provide dramatic advances for public health. However, we encourage a closer look at “small data”, meaning local data resources collected by governments and health agencies in developing societies. Despite the challenges and limitations inherent in data collected locally in developing countries, such data present a potentially rich and relatively inexpensive source for health-based studies.

The objective of this paper is to present our collaboration with government officials, hospital record keepers, and others to concatenate and present locally collected health data in Gujarat and Rajasthan, India and Kathmandu Valley, Nepal. We discuss the challenges to setting up the necessary collaborations to collect and format these existing data, as well as the opportunities for health data collection in developing country settings. We further present the way we acquired and formatted our health data, and the overall quality of the finished products. We describe the availability of the data and potential uses and challenges of these and similar data in a public health context. The long-term goal of this work is to highlight the potential value of such scientific resources that could be used innovatively to develop highly localized analyses of health outcomes, to address environmental health conditions and improve sustainability for the present day and under climate change.

Below, we summarize the current state of record keeping for specified health outcomes in western India and Kathmandu Valley, Nepal (Figure 1). In western India, daily total mortality data were collected for five cities, three in Gujarat (Ahmedabad, Himatnagar, and Idar) and two in Rajasthan (Jaipur and Churu). In addition, daily hospital admissions data from the three hospitals that fall under the jurisdiction of the Ahmedabad local government were collected. In Nepal, individual-level data on hospital admissions were collected from six major hospitals in the Kathmandu Valley. Among these cities, Ahmedabad, Jaipur, and Kathmandu are major urban areas with rapidly growing populations. Himatnagar, Idar, and Churu are smaller cities, which still experience rapid urbanization and population increase.
The Indian city governments collect information on deaths occurring within the municipal area, in different formats depending on the city. Registration of births and deaths in India is conducted according to the Registration of Births and Deaths Act (1969). Civil registration of deaths must occur within 21 days of the event. In Ahmedabad and Jaipur, death is recorded as per the Act by hospital authorities if the death happened in hospital or by a relative of the deceased if it occurred elsewhere. In rural areas, deaths are hand-recorded on a daily basis by administrative officials as they are reported. For the morbidity data from Ahmedabad hospitals, monthly reports from each of the government hospitals are sent to the Ahmedabad Municipal Corporation, containing daily summaries of the numbers of patients admitted to the hospital, number of deaths, number of surgeries conducted, inpatients versus outpatients, number of child deaths, and specific infectious disease cases that are proscribed by the city government. The Office of the Registrar General and Census of India currently estimates mortality rates through the Sample Registration System, and provides annual and semi-annual estimates of mortality rates by cause [14].

The Nepalese government collects only monthly hospital admission summary data, although detailed information (e.g., International Classification of Diseases (ICD) code) is not included and not all hospitals participate. Each hospital has its own system for archival data. Most hospitals have recently started to computerize records; hence most hospitals’ archives are paper-based. To the best of our knowledge, prior to this project no computerized dataset existed for hospitalizations in Nepal, other than within some individual hospitals. We obtained data from each of six individual hospitals.

2. Materials and Methods

Original Data Collection

Currently, India does not have a health statistics cadre in training, and health departments do not routinely employ statisticians or epidemiologists, which can hinder data management, analysis, and interpretation of large-scale public health studies [15,16]. Ongoing training in health data collection and analysis, which is taking place at the Indian Institutes of Public Health (IIPH), will generate interest in the uses for public health data locally.

Figure 1. India and Nepal (Study sites indicated in red). Source: (Global Administrative Areas, GADM Database) [13]. Note: Map generated pre-2014 split of Andhra Pradesh into two states.
One challenge to multi-city analyses in this region is the lack of standardization in the collection of health data. For hospitals in Ahmedabad and Kathmandu, data were recorded by doctors from patient charts, and reported to a registrar within-hospital for aggregation. In Ahmedabad, these data were tabulated by the medical records department into aggregated daily data for reporting to the Ahmedabad Municipal Corporation (AMC), using different reporting forms. We obtained mortality data in Ahmedabad and Churu from respective city-level government offices, which require death certificates before a body can be cremated or buried; as such, the data can be considered fairly complete. Other cities such as Jaipur have no such requirement, which may lead to data losses [17].

Another consideration is the medium in which records are stored. In some cities, records are paper-based, while in others records are computerized. In Ahmedabad, death certificates are originally issued on paper; the Registrar of Births and Deaths then concatenates information electronically. The hospital admissions data for Ahmedabad are often collected on paper, but currently are not computerized. In Himatnagar and Idar, data are collected in a hand-written register of deaths, and are not computerized. In Churu, more recent data on deaths occurring outside of the hospital are collected by the Nagar Parishad (urban political unit for Class 2–3 towns of 15,000–25,000 inhabitants) office and housed in a storage facility, where the records are electronically available within a data storage program. In Kathmandu, computerized hospital records were available only from mid-2004 onwards for some hospitals, with earlier records in paper format. For three hospitals in Kathmandu, data were transcribed to electronic format, whereas for the three remaining hospitals all records are in paper format.

3. Results

3.1. Description of Data

Table 1 summarizes the collected data and illustrates the differences in data storage methods and formats. The type of administrative area included in the datasets may vary by city. Himatnagar, a city in northern Gujarat, counts deaths within the district and thereby includes peri-urban and some rural deaths, while data from Churu in Rajasthan and Ahmedabad in Gujarat only include deaths within the city limits (including non-residents). For Nepal and Ahmedabad, hospital records data correspond to admissions at a specific hospital that may have patients from multiple administrative areas.

3.2. Acquiring the Data

The WHO has outlined the importance of collecting and maintaining good health records for medicinal purposes and research [18]. However, obtaining high-quality records from regional management systems can be challenging. To collect our health data, researchers met in-person with government officials, hospital workers, and partners in Nepal and India. These data come from multiple sources, including hospital administrators and city governments.

One obstacle to data collection was a lack of familiarity among record keepers of recent trends in data utility. Unlike many developed countries with established pathways for access to detailed health data (e.g., Centers for Disease Control-managed National Hospital Discharge Survey [19]), the use of health data for research beyond census record keeping is very new and not considered within the purview of the relevant authorities. Therefore, common practices for collecting health data may not be well-known or universally applied in lower- and middle-income countries. Requests for data on health can be misinterpreted as efforts to identify problems or be overly critical of the community. In small towns, local customs may make it difficult for health officials to openly discuss ongoing public health efforts. Leaders and institutions valued a clear description of the intention to use the data to promote good health.
Table 1. Data description.

<table>
<thead>
<tr>
<th>Study Area</th>
<th>State/Country</th>
<th>Census Population in Area (Year)</th>
<th>Health Outcomes</th>
<th>Timeframe of Interest</th>
<th>Number of Events</th>
<th>Source of Data</th>
<th>Original Format of Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ahmedabad urban area</td>
<td>Gujarat/India</td>
<td>6.06 million (2011)</td>
<td>Daily totals for all-cause mortality; daily totals for hospital admissions</td>
<td>2005–2013 (deaths); 2007–2012 (hospital admissions)</td>
<td>283,628 deaths; approx. 10,000 outpatient admissions/month (avg)</td>
<td>Municipal Corporation Registrar; government hospitals</td>
<td>Computerized (mortality); paper (hospital)</td>
</tr>
</tbody>
</table>

1 (Government of India, 2011) [20]; 2 (National Population and Housing Census 2011, 2012) [21].
Given the lack of precedent for how data acquisition for research should take place, multiple Institutional Review Board (IRB) approvals were obtained so that oversight was distributed among local and international sites, along with data sharing agreements. This can be difficult to coordinate, as different agencies will require different protocols for IRB approval. In addition to the US agencies (Icahn School of Medicine at Mount Sinai and Yale University), IRB approvals were obtained from the Indian Institute of Public Health—Gandhinagar approval board, Nepal Health Research Council, and each individual hospital, with different processes and paperwork requirements. As some organizations had not previously provided data for research purposes, they had no procedure in place for data acquisition or a structure to authorize approval of data collection. Changes in personnel occasionally resulted in fluctuating approval processes and reversals of previous decisions on the sharing of data.

Researchers from abroad collecting data should be prepared to discuss frankly the objectives for data collection in small villages and towns. Scientists even from different parts of the same country can be questioned regarding their interest in a population other than their own, especially in India, which presents different languages and customs in different states [22]. Our research also showed that communities themselves might be unable to access to the final results of studies based on the data collected; sustainable development of local data resources should allow communities to accrue benefit from health studies in their region.

3.3. Post-Collection Challenges for Health Analysis

Given issues with data collection, storage, and sharing, the quality of the data obtained across different sites and sources can vary widely. For example, paper records available in Idar and Himatnagar are hand-written in Gujarati, whereas records in Ahmedabad, Jaipur, and Churu are recorded electronically in English. Hospital data in Ahmedabad are collected and summarized in Gujarati to the city government, and must be translated. We observed that handwritten data registers obtained in western India can have data points that are difficult to accurately translate, as well as missing data. In addition, converting paper records to electronic files and translating the data into a common language are important steps of data cleaning, which can impact the quality of the data.

Finally, the completeness of the data record is a further limitation. Especially in places that use paper records, the data may have gaps due to occasional purging of records or accidental data loss (e.g., fire). In Kathmandu, no data were available for entries before mid-2004 as all hospitals discarded hard copies of records after a certain time period. The lack of back-up data can make replicability of health-based studies difficult. In Ahmedabad, paper copies of data records from one hospital were not transferred when the facility moved locations, and paper records from the previous facility were discarded.

Despite these challenges, the datasets we have collected will help to inform the discussion around health outcomes associated with environmental impacts or other exposures or risk factors. As opposed to developed countries such as the US or many in western Europe, developing countries have a dearth of studies on health. Potential paths forward to fill this existing gap include capacity building through partnerships with academic institutions to facilitate data collection and translation from indigenous languages, as well as statistical analyses to fill in data gaps and combine data from neighboring locations.

4. Discussion

Our initial goal in investigating potential sources of health data in this region was to evaluate the health impacts of specific environmental exposures; however, the data can also be used to investigate trends in health outcomes over different areas across time, to improve health outcomes among local populations in the long term. Temporal tracking is of substantial importance for research regarding climate change and sustainability, such as studies of adaptation. Potential further uses for this data that extend beyond the current planned studies include the identification of major public health issues of concern (which causes of death are most common, etc.) and the evaluation of the effectiveness of
policies and interventions aimed at improving public health. Evaluation can continue in real-time as interventions are staged if data collection continues.

In international studies, such as the Global Burden of Disease (GBD), exposure-response relationships derived from developed areas are applied, sometimes with modifications, to developing countries [1]. Relationships derived in developed countries with greater data availability may be a good proxy to estimate the burden of disease in developing countries that lack data on exposure and health. However, such extrapolation introduces uncertainties, and the exposure-response relationship even within developed country settings may be drastically different due to fundamental differences in the population, environment, levels of exposure, and adaptive capacities [23]. These differences may be exacerbated in developing areas.

One broad challenge to data collection in some developing countries is a sensitivity of hierarchy, gender issues, and skepticism directed towards non-locals. Systems in these countries can have a strictly hierarchical system of authority, with positions of authority traditionally assigned to men. As a result, social science research and fieldwork conducted by female scientists has been historically viewed as contradictory, and the utility of short-term stints in the field has been questioned in anthropological and ethnographic literature [24,25]. Knowledge of the culture and long-term collaboration with local data collection agencies as well as administrators are essential.

Availability of health data in countries such as India and Nepal may also provide insight into population-specific exposure-response curves at high levels of exposure, where existing literature is sparse. Recent extreme air pollution levels in western Europe provide a good example of how data from developing countries may be useful in estimating potential health impacts of unusually high air pollution levels in more developed countries [26]. Findings can further be useful for extrapolating to other rapidly developing areas as a more locally relevant proxy than studies conducted in industrialized regions.

One of the main observations from our work is that more data exists in these places to be collected. In India, paper and computer records were found at government facilities that detail age at time of death, location where the subject was born and raised, ethnicity, religion, and information on occupation and income during the lifetime of the subject, etc. In Nepal, paper and computer records were collected from each hospital with a unique identification number for each patient, admission and discharge date, age, gender, cause of admission, and home location. These and other relevant morbidity and mortality data can serve to further inform the important exposure-health analyses discussed above. The data are potentially available, but a suitable and sustainable methodology for data collection and management should be developed, keeping in mind the complexities of collecting health data on the ground.

Another key implication of our work is that these data can be used to implement practical, local solutions for current health-relevant issues. In Ahmedabad, for example, the Ahmedabad Heat and Climate Study group (working in conjunction with the Ahmedabad Municipal Corporation) used the existing registry data on mortality outcomes to develop a heat wave early warning system, the first such system to be implemented in South Asia [27]. This early warning system has been successful at improving health outcomes in the city during times of extreme temperature in subsequent heat wave seasons. Another example of the utility of registry data for studying infectious disease epidemics is found in the work of Mavalankar et al., around the chikungunya epidemic in the region in 2006 [15].

The tools that have been outlined for the collection of existing health data can be more systematically collected in developing country settings, provided access to adequate resources and interest in data collection for public health purposes on the part of local stakeholders. Validation of the data at this stage is difficult, as the processes for data collection are new. However, as data collection becomes more sustainable in developing regions in the future, internal data validation methods can be developed and applied.

Key efforts in the near future are needed to capture health data that will otherwise be lost, so that such data can be used to understand the determinants of health in a range of societies. The lessons
learned can be used to increase the capacity to monitor health outcomes in developing country settings. Through collaborations with local scientists, government officials, community leaders, and health practitioners, we can begin to further understand and characterize health outcomes in the areas where current knowledge is limited.

Acknowledgments: The authors acknowledge the Ahmedabad Heat and Climate Study Group, Ahmedabad Municipal Corporation, Jaipur Municipal Corporation, and municipal governments of Idar, Himmatnagar, and Churu for their cooperation, and the Bhoruka Charitable Trust for support in data collection. This project was supported by a Fulbright Program grant sponsored by the Bureau of Educational and Cultural Affairs of the United States Department of State and administered by the Institute of International Education, as well as by the US Environmental Protection Agency through a Science to Achieve Results (STAR) fellowship and the Yale Institute for Biospheric Studies Pilot Grant fellowship.

Author Contributions: A.N.-S. and A.G. conceived the collection of data and the format of the paper. A.N.-S., G.S.A. and A.R. collected the data from the cities in India. A.G. collected the data from Kathmandu, Nepal. A.N.-S. and A.G. analyzed the data. A.N.-S. and M.L.B. wrote the manuscript. M.L.B., A.G., D.M., G.S.A., A.R. and P.S. provided edits and comments on the manuscript. M.L.B. formatted the manuscript for publication.

Conflicts of Interest: The authors declare no conflict of interest.

References


