Mining the Gap: Pathways towards an Integrated Water, Sanitation and Health Framework for Outbreak Control in Rural India

By

Xiaoyuan "Charlene" Ren

B.A. Physics
Vassar College, 2013
S.M. Civil and Environmental Engineering
Massachusetts Institute of Technology, 2016

Submitted to the Institute for Data, Systems, and Society
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN TECHNOLOGY AND POLICY

AT THE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2017

© 2017 Massachusetts Institute of Technology. All Rights Reserved.

Author: Signature redacted

Institute for Data, Systems, and Society
February 22, 2017

Certified by: Signature redacted

Chintan Vaishnav
Senior Lecturer, MIT Sloan School of Management
Thesis Supervisor

Accepted by: Signature redacted

Munther Dahleh
William A. Coolidge Professor, Electrical Engineering and Computer Science
Director, Institute for Data, Systems, and Society
Acting Director, Technology and Policy Program

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
MAR 02 2017
LIBRARIES
ARCHIVES

Mining the Gap: Pathways towards an Integrated Water, Sanitation and Health Framework for Outbreak Control in Rural India

by

Xiaoyuan "Charlene" Ren

Submitted to the Department of Civil and Environmental Engineering
On February 22, 2017, in partial fulfillment of the requirements for the
Degree of Master of Science in Technology and Policy

Abstract

The scientific connection between sanitation, water quality and health is well established. However, in the present Indian scenario, monitoring and governance of the three sectors is handled separately. At present, the need to integrate sanitation, water quality, and health is felt during waterborne disease outbreaks such as large-scale diarrhea, typhoid or cholera. Despite the general interest shown for a cross-sector integrated framework in outbreak control, numerous administrative and technical gaps exist preventing the implementation of this framework. This study attempts to address these implementation barriers through the analysis of governing institutions and data integration of large public databases for the selected districts of Gujarat, India.

Interagency collaboration barrier is analyzed through a comprehensive institutional analysis on the water, sanitation and health monitoring sectors. The lack of administrative incentive due to the narrow definition of monitoring targets is identified as the primary barrier for collaboration. Districts that already achieved 100% open-defecation-free status are identified as key entry points for potential pilot implementation of an integrated framework. National Informatics Center and Water and Sanitation Management Organization (WASMO) are considered key nodal points for building channels of interagency connections.

Data integration and utilization barriers are analyzed through habitation-level matching of the 3 separate monitoring databases – namely, Swatch Bharat Mission (SBM) database for sanitation, Integrated Management Information System (IMIS) database for rural drinking water quality and Integrated Disease Surveillance Programme (IDSP) for outbreak data. The most critical data barrier is the discrepancy between administrative units across the databases, resulting in 25% mismatched habitation data and variables with 30% contradictory data entries. Quality concerns over inconsistent and missing data are also raised, especially for data collected by grassroots workers.

A decision support model based on the integrated database is constructed through a Driver-Pressure-State-Exposure-Effect-Action (DPSEEA) framework. A significant correlation is observed between chains connecting sanitation initiatives and water quality. Significant risk factors associated with outbreak occurrence cannot be identified at the current stage. Even though implementing this model is within reach, and doing so promises to offer an efficient tool for integrated governance of the three sectors, incomplete datasets is currently the key barrier to a comprehensive assessment of model effectiveness.

Thesis Supervisor: Chintan Vaishnav

Title: Senior Lecturer

Acknowledgements

First and foremost, I would like to express my gratefulness to my advisor, Prof. Chintan Vaishnav. His wisdom, patience, guidance and, of course, humor lit my path along the way. With all the complexities of studying very real issues in the world, I would have been lost if it weren't for his unfailing and selfless support. Thank you, Chintan, for giving me chance after chance to see potentials rather than limitations. Thank you for making me realize all marvelous possibilities in the world.

I would also like to thank MIT Abdul Latif Jameel World Water and Food Security Lab for supporting the research and my various explorations in India. I am so lucky to have been given the chance to work on this project. I would also like to thank all other members of the research team, Emily, Sydney, Mike and Prof. Rohit Karnik, for sharing this wonderful journey with me.

Much of the local fieldwork would not have been possible without support from Tata Water Mission, UNICEF India, WASMO, GJTI and the local water labs. I am also thankful for Professor James Wesocat and Marianna for their rich experiences in Gujarat, especially in getting the most out of every interview conversation. To everyone that I have interviewed (as detailed in Chapter 4), thank you so much for taking the time and bearing with my endless list of questions, allowing me to understand the complexities of the sectors within a short span of time.

I owe much of the data preprocessing work to support from volunteers at Ahmedabad University - Aashima, Kavi, Mayank, Rishabh, Jayraj and Sharvil. Anmol and Netra, in particular, worked passionately and tirelessly with me to weed out issues with every single dataset. I am also fortunate enough to receive consistent guidance from Gopal, Donghao, Hao and Xinkai – my data-savvy friends who often spend hours helping me and answering my random questions. None of the data analysis would have been possible without your generosity with time and knowledge.

To the TPP administration, especially Barbara and Frank, thank you for being there for me throughout my experience. Your patience and understanding supported me through the most challenging moments, and for that, I will always be grateful.

I would also like to thank the book "What Makes You Not a Buddhist" for supporting me through my most difficult thesis writing moments. It gave me compassion, peace and mission, and made even the hardest moments a surprisingly rich and beautiful experience.

Last but not least, I would like to thank my family and friends for always being there for me. Mom and dad, your consistent support has always been the rock that anchors me. I am not doing the most traditional work, but your ability to understand and support my pathways forward has always been amazing. Xiawei, Summer, Ziqi, Laura, Mengqi, and many many more of my dearest friends, thank you for writing with me, talking with me and just always being there for me. Thanking you for holding my hands along the way. Your encouragement made this thesis develop, and made me develop into a better person along with it. You are the best friends anyone can ever ask for. I cannot begin to express how thankful and blessed I feel to have every single one of you in my life.

I love you all so much.

TABLE OF CONTENTS

1 INTRODUCTION

This chapter introduces the background of current outbreak control management practices and the possible improvements to the process with the introduction of an integrated water, sanitation and health (WaSH-health) decision support framework. Section 1.1 outlines the current framework for outbreak control, and its four key stages of prevention, detection, investigation and case summary. Section 1.2 introduces the possible benefits that an integrated system can bring to the different stages of outbreak control. Section 1.3 describes the current disconnect among the three sectors in the context of India and the potential causes for such disconnect, motivating the research question in Section 1.4 to understand the critical barriers to an integrated WaSH-health system in the context of India. An outline of the thesis is given at the end of the chapter.

INTRODUCTION

Inadequate drinking water management and poor sanitation are still among the leading causes of preventable morbidity and mortality in the world (OECD and WHO, 2003). Despite clear advances in water management and sanitation throughout the 20th century, waterborne diseases are still frequent. Diarrheal diseases alone have caused around 2.2 million of the 3.4 million water-related deaths per year (OECD and WHO, 2003). Disease and outbreaks associated with drinking water are relatively common even in affluent nations such as United States, which reported over 400,000 cases of infectious illness and almost 100 instances of outbreaks linked to drinking water across 1991-1998 (Hunter et al. 2003).

Outbreak, which is generally defined as "the occurrence of more cases of disease than expected in a given area or among a specific group of people over a particular period of time," are of special concern due to its scale and severity, affecting countries at all levels of economic development[1]. For example, the outbreak of cryptosporidiosis in 1993 in Milwaukee, Wisconsin, US resulted in over 400,000 suffering from gastrointestinal symptoms(Medema *et al.*, 2003). An outbreak involving E. coli O157:H7 occurred in Walkerton, Ontario, Canada in 2000 and resulted in over 2300 cases and 6 deaths (Medema *et al.*, 2003). Even for more recent years such as 2009-2010, waterborne outbreaks continued to occur - 33 was reported in the US, with 9 deaths as a result (CDC 2013). Despite progress in water, sanitation and healthcare, the number of such outbreaks happening across the world and their resulting consequences remain significant. Are there potential improvements in the process of battling outbreaks that can be made?

To answer this question, we must first look at the traditional framework for combating outbreaks.

1.1 Current framework for outbreak control

1.1.1 Outbreak prevention

Epidemiologic surveillance systems are entities set in place for the purpose of collecting, analyzing and interpreting health data for planning, executing and evaluating public health practices (Gerstman 2003). Outbreak-related heath data are collected at local health facilities and interpreted by these surveillance units to identify any trends. This information can be used to conduct educational campaigns to raise awareness on potential diseases or outbreaks and advise on positive action to prevent them.

1.1.2 Detection

It is generally expected that detection of an outbreak may usually start with the recognition of an aggregation of cases in a given geographic region over a particular period through epidemiologic surveillance systems. If the surveillance system is setup effectively, the aggregation would readily

[1] https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson6/section2.html

raise alarms and field investigator would likely follow through to check and see the cluster of cases is indeed an outbreak (CDC website).

However, it has been noted that surveillance systems are at times limited in their scope and accuracy. Lack of consistent and detailed reporting formats have caused such methods to be insensitive to subtle changes in disease occurrences, which makes it difficult to evaluate case of increases to effectively confirm instances of disease outbreak. According to Andersson and Bohan (2001), very few examples of outbreaks are actually first revealed by surveillance systems. Hence, apart from early warnings from epidemiologic surveillance systems, individuals who are directly or indirectly affected by the outbreak (e.g. cases, caregivers, relations) also frequently bring specific evidence forward to the attentions of public health authorities to suggest an imminent outbreak that is worth investigating (Gerstman 2003).

This is especially true for waterborne diseases. The official definition of a waterborne (or foodborne, which is also covered by the same definition) is "when two or more persons experience a similar illness after ingestion of the same type of food or water from the same source and when the epidemiological evidence implicates the food or the water as the source of the illness" (Hunter *et al.*, 2003). However, during earlier stages of waterborne outbreaks, the cases are far and few in between and may not be immediately detectable against the general background of infection, and it is usually unclear whether these changes are related to water contamination issues.

According to the CDC Working Group, earlier detection of outbreaks can be achieved in the following ways (Buehler *et al.*, 2004):

- Prompt review and investigation of disease case reports, along with timely communication between physicians, health-care facilities, laboratories and public health departments
- Improved pattern recognition through better analytical systems, which increases predictive values of data to determine the likelihood of the start of an outbreak.
- Collection of new types of data that can signify outbreak at an earlier stage, such as data on healthcare purchase, school or work absences and so on.

These improvements in outbreak detection are much needed, especially for waterborne outbreaks.

1.1.3 Investigation and Hypothesis testing

A general outline of standard outbreak investigation procedure is shown in Table 1-1. The first step of any outbreak investigation would be "defining the problem" - confirmation of the diagnoses to conclude an apparent outbreak, where all possible causes of error need to be considered and excluded (Gerstman, 2003; Hunter *et al.*, 2003). After an outbreak is positively confirmed, the next step would be the derivation of a "case definition," including epidemiology information on characteristics of affected persons, key symptoms, lab results, geographical locations and date of onset (or in other words, the epidemiologic variables of time, place and person), all of which are necessary to identify how to include or exclude cases in the outbreak analysis (Gerstman, 2003; Hunter *et al.*, 2003). With this case definition, additional cases may be searched for, identified and reported systematically (CDC website).

Once a concrete case definition is available after sufficient information collection, a preliminary hypothesis regarding the outbreak can then be generated and corresponding remedial controls may

be suggested (Hunter *et al.*, 2003). The hypothesis may address the source of the outbreak, the mode of transmission or exposures resulting in the cluster of diseases, or all of the above (CDC website).

Further epidemiologic and environmental investigations are required to test and confirm the hypothesis. In the case of waterborne outbreaks, there are generally three parts to this investigation, including (Hunter *et al.*, 2003):

- Epidemiological investigations, including case-control or retrospective cohort studies of groups to determine how many cases are likely correlated with a certain exposure;
- Further microbiological analysis of human and environmental samples for further characterization;
- Sanitary inspection of the water system (when waterborne disease is suspected).

The separate pieces of evidence collected through the investigation procedure can all effectively complement each other, which makes it possible to come to a conclusion on the likely causes of outbreak. There are still discrepancies among the different classification schemes for the strength of evidence. Many classifications give more weight to epidemiological studies (in the cases of US and UK evidence classifications system), although recent studies have shown that bias are likely to exist when the possible outbreak causes have been made public, which can result in false identification of drinking water as the outbreak cause (Hunter *et al.*, 2003). Hence, even though pathogens are not required to be detected in the water supply to confirm an outbreak cause, any extra data supplementing epidemiological data would be useful to improve the strength of an outbreak conclusion (Hunter *et al.*, 2003). Retrospective review of routine water quality data analysis or other registration records of failures in the water system are consequently also of strong importance for the hypothesis testing process.

Once the outbreak causes are known, control should be targeted at the weakest link along the chain of infection, such as source control (e.g. remove pathogen from water), transmission interruption (e.g. shutdown water source) or modification of host response to exposure (e.g. improve sanitation conditions) (Hunter *et al.*, 2003).

1.1.4 Case summaries

A final but crucial step of outbreak investigation include dissemination of findings to the appropriate parties. Oral briefings to local authorities along with written reports are the usual formats of case summaries (CDC website), all of which would provide insights for future prevention and investigation of similar types of outbreak.

Table 1-1 Standard components of outbreak investigations (Gerstman 2003)

1. *Define the Problem*
   - Confirm diagnoses
   - Show that an epidemic exists (observed number of cases is significantly greater than expected)
2. *Describe the Epidemiology of the Outbreak*
   - *Time:* determine dates and times of onset; draw epidemic curve; determine attack rates over time
   - *Place:* draw spot map of cases; consider environments of home, work, recreational, and special meeting places
   - *Person:* calculate attack rates by age, sex, occupation, ethnic group, and other personal factors; consider rates of infection, disease and death; note possible means of transmission; address both common denominators and notable exceptions
3. *Formulate Hypotheses*
   - Source of infection
   - Method of contamination and spread
   - Possible control mechanisms
4. *Test Hypotheses*
   - Conduct special epidemiologic, laboratory, and environmental investigations
5. *Draw Conclusions and Devise Practical Applications*
   - Long-term surveillance
   - Prevention

1.2 An interconnected water quality, sanitation and health system

Global access to safe water, increased sanitation conditions and proper hygiene practices have been shown to effectively reduce illness and death. According to a report by UNICEF India, unimproved hygiene, inadequate sanitation, insufficient water and unsafe drinking water account for approximately 7% of total disease burden and 19% of child mortality worldwide (Cairncross *et al.*, 2010). As shown in Figure 1-2, a number of diseases, especially infectious diarrhea, are highly preventable with better WaSH conditions.

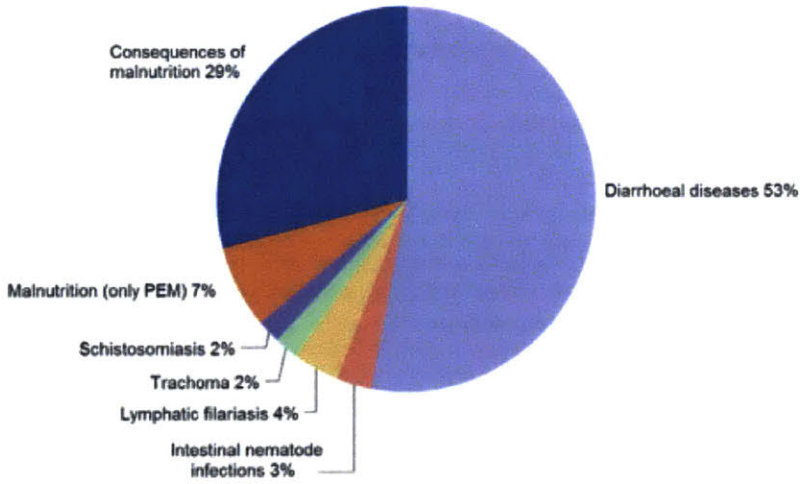


Figure 1-1: Contributions in Disability-Adjusted Life Years of individual diseases to the total burden of diseases preventable by WaSH improvements *(Bartram and Cairncross, 2010)*.

This connection is more prominent in developing countries where the majority of rural population are still drinking from untreated groundwater, which is even source of concern in developed countries. 30.3% of the 818 drinking water outbreaks reported to US CDC between 1971 and 2008 were related to issues with untreated groundwater (Wallender et al. 2014).

While outbreak detection and investigation gave much more weight to epidemiological evidence, the connection between outbreak occurrence and the corresponding water and sanitation information is undeniable. It would be impossible to study waterborne outbreak phenomenon independent of the existing water environment that likely contributed significantly to the risk factors that eventually triggered the outbreak. Many of the limits identified in the outbreak combat framework, such as insufficient means to predict and prevent outbreaks or lags in detecting outbreaks, is likely to be improved if water and sanitation monitoring system are more readily connected with disease and outbreak reporting systems.

As Andersson and Bohan (2001) pointed out in a WHO report, a good surveillance system should include much more than just strong epidemiological and laboratory inputs – environmental factors are just as critical. It is important to go beyond the connection between host and agent, and to try to identify the environmental causes of an outbreak, which can more readily enable relevant interventions.

The benefit of including additional WaSH information in the process of outbreak control is detailed in the sections below.

1.2.1 Prevention: understanding casual links to prevent disease outbreaks

While epidemiologic surveillance systems might be able to prevent the disease from spreading, the system alone has limited capacity to actually prevent the outbreak from occurring in the first place. The occurrence of waterborne outbreaks is essentially a product of an integrated WaSH (water quality, sanitation and hygiene) ecosystem. When proactive interventions are introduced, waterborne diseases such as acute diarrhea has been shown to decrease drastically (Figure 1-2). On the other hand, isolating the outcome in this process would only result in a reactive paradigm where measures

can only be taken after cases occur. If we look at the interconnection among the water, sanitation and health systems, a new "due diligence" paradigm emerges where all reasonable measures can be taken in advance to prevent the occurrence of negative health consequences (Medema *et al.*, 2003).

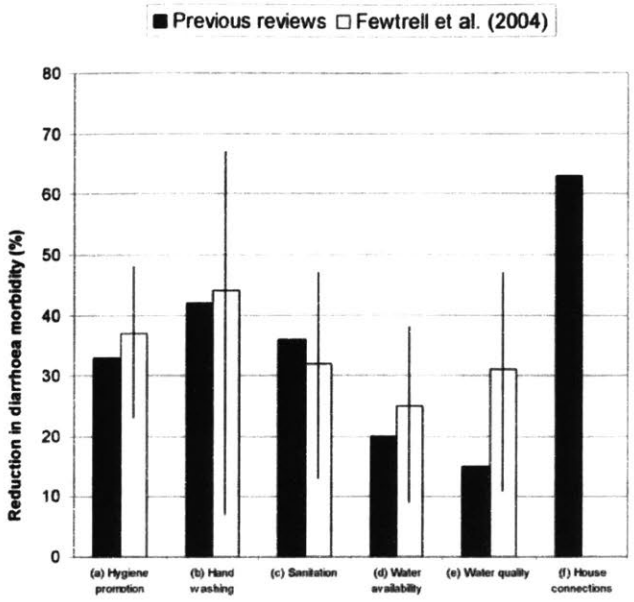


Figure 1-2 Results of reviews of the effect on diarrhea of Water, Sanitation and Hygiene interventions *(Fewtrell and Colford, 2004; Bartram and Cairncross, 2010).*

A casual chain framework is frequently used to organize the interconnections among the different WaSH sectors, as defined by Niemeijer and de Groot (2008):

"In the causal chain, social and economic developments are considered driving forces that exert pressure on the environment, leading to changes in the state of the environment. In turn, these changes lead to impacts on human health, ecological systems and materials that may elicit a societal response that feeds back on the driving forces, pressures, or on the state or impacts directly."

Common variations of the casual chain frameworks include DPSIR (Driving forces-Pressures-State-Impact-Response), PSR (Pressure-State-Response) and DPSEEA (Driver-Pressure-State-Exposure-Effect-Action), as shown by an example Figure 1-3 where the connection between drinking water quality and waterborne diseases are documented through a DPSEEA framework (Khan *et al.*, 2007; Gentry-Shields and Bartram, 2014; Schwemlein, Cronk and Bartram, 2016). An example of the pressure-state section of the framework is "Water Safety Plans", which are management plans developed for many central water supply systems detailing actions to be undertaken from normal conditions to extreme events along the water supply system, to determine whether the water supply chain as a whole is operating properly and can provide water of the appropriate quality (Medema et al. 2003). Effective executive of such water safety plans can continue to act along the casual chain to create positive human health impact.

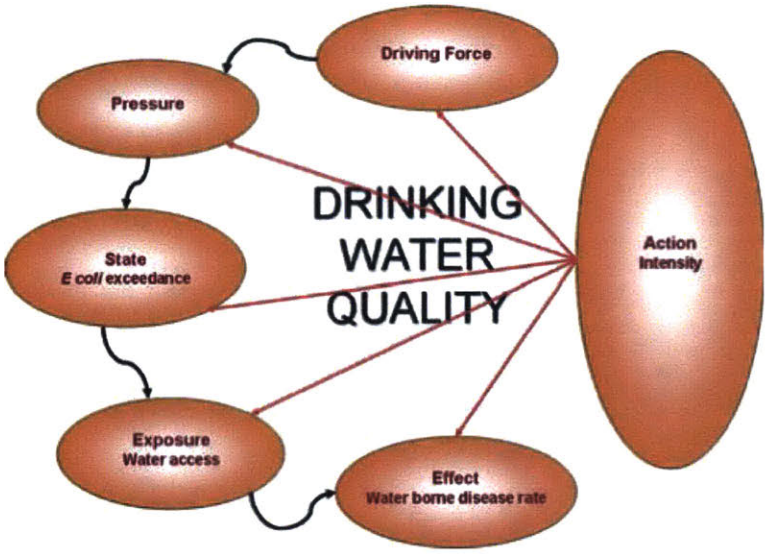


Figure 1-3 DPSEEA framework sample for drinking water quality and its impact on health *(Khan et al., 2007)*

While these frameworks are not yet applied directly to outbreak prevention, they provide intuitive understandings on factors that may lead to health impacts and waterborne disease outbreaks, which in the long run may lead to corrective actions.

1.2.2 Detection: Increasing sensitivity in determining occurrence

Early and rapid recognition of the possibility of outbreaks would result in a timely start to the outbreak investigation, which can greatly increase the likelihood of outbreak cause determination (Andersson and Bohan 2001).

While public health surveillance provides more direct evidence for identifying waterborne outbreaks, it is a comparatively slow route which would take one to two weeks before significant changes can be observed. The delay can be caused by a number of people who may not consult a doctor immediately when they show symptoms (Andersson and Bohan 2001). This is especially true for rural communities in developing countries where the population are less likely to consult a doctor when encountering similar symptoms, which would delay the identification of the outbreak even further. There is even a study which concluded that hospital incidence data from Hyderabad may underestimate waterborne disease cases by a factor of approximately 200 (Mohanty, 1997).

Apart from case reports and specific evidence of clearly affected individuals, a variety of other WaSH triggers may effectively provide the earliest practical warning of the possibility of unsafe water and potentially imminent outbreak. These triggers may include deviations in (Medema et al. 2003):

- water processing indicators (e.g. failure in treatment plants)
- chance events (e.g. spillage of hazardous substance)
- changes in water quality parameters (e.g. increase in turbidity or E.coli concentration)

Hence, if such water quality and sanitation related information can also be incorporated in the early detection of waterborne outbreaks, outbreak detection sensitivity may increase significantly. In fact, John Hopkins University Applied Physics Laboratory have been building a module for the Electronic Surveillance System for the Early Notification of Community-based Epidemics (ESSENCE) syndromic surveillance system which includes both water quality and health indicators for the early detection of water contamination related disease outbreak events (National Homeland Security Research Center, 2012). This system utilizes Bayesian network analysis to synthesize disparate types of data into a warning system that outputs the likelihood of waterborne disease outbreak occurrence based on varying combinations of data anomalies (Burkom *et al.*, 2011). Even though this prototype system has only been tested with simulated data at three pilot locations across water and public health departments, it has shown high promises in the improved detection of outbreaks.

1.2.3 Investigation and hypothesis testing: Increasing evidence support

Environmental information is a crucial part of hypothesis testing during outbreak investigation, especially if water sources are a suspected cause. Microbial results from water supplies that successfully identifies the causative pathogen can provide one of the best evidence for linkage between water system and an outbreak disease (Hunter et al. 2003). To increase environmental evidence collection, enhanced monitoring of water supply systems is also encouraged, including increasing sampling frequency, sampling in other locations along the distribution pathway or carrying out non-regular microbial analysis (Hunter et al. 2003).

In the case of waterborne disease surveillance in the US, public health departments are primarily in charge of detecting and surveying waterborne outbreaks, but additional information on water quality and treatment can be obtained from the state's drinking water agency whenever necessary (Andersson and Bohan, 2001). The official classification of evidence in hypothesis testing for UK and the US include both epidemiological data and water data, as shown in Figure 1-4.

Classification	*Epidemiological data*	*Water quality/water treatment data*
I	Adequate: data provided about exposed and unexposed persons & water implicated & the relative risk > or = 2 (or $P < 0.05$)	Adequate: historical information or laboratory analysis supports association (e.g., chlorinator malfunctioned or a water main broke, no chlorine residual, or coliform bacteria were present)
II	Adequate	Not provided or inadequate
III	Limited: epidemiological evidence provided that did not meet the criteria for adequate or the claim was made that ill persons had no exposures in common except for water but no data were provided	Adequate
IV	Limited	Not provided or inadequate

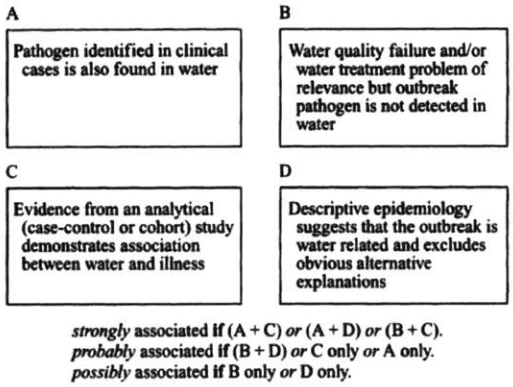


Figure 1-4: Categorization of Levels of Evidence in Confirming an Outbreak Hypothesis *(Tillett, Louvois and Wall, 1998; Andersson and Bohan, 2001)*

1.2.4 Case summary: Follow-up actions to prevent future occurrences

There are frequent examples of interagency collaboration in the creation of waterborne outbreak case summaries. Since 1971, CDC, the US EPA (Environmental Protection Agency) and the Council of State and Territorial Epidemiologists have maintained a close collaborative Waterborne Disease and Outbreak Surveillance System (WBDOSS) for collecting and reporting data related to the occurrences and causes of waterborne disease (Andersson and Bohan, 2001; Yoder *et al.*, 2008), as show in Figure 1-5. These surveillance data, published in Morbidity and Mortality Weekly Reports (MMWR) approximately every two years, include data on the types of water systems, their deficiencies and the etiologic agent related to the outbreak, all of which proves to be useful in evaluating the adequacy of current water systems, as well as identifying potential hazards and improvement plans (Andersson and Bohan, 2001).

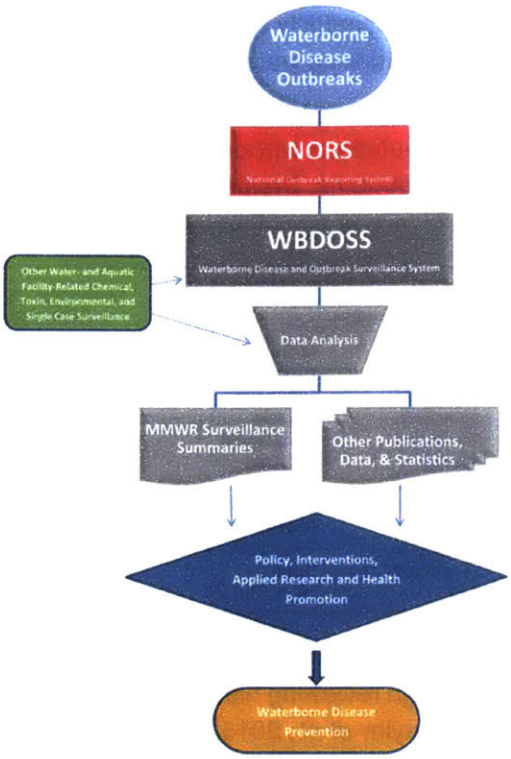


Figure 1-5: Functions of the collaborative US Waterborne Disease and Outbreak Surveillance System[2]

This integrated case summary system has proven to be successful in many occasions. For example, in the US, after an outbreak of cryptosporidiosis in Milwaukee in 1993, CDC and EPA conducted the case summary that led to a more stringent standard for acceptable turbidity values in drinking water (Andersson and Bohan, 2001). This standard came into effect in all states across the US and likely have contributed to the zero-reporting of Cryptosporidium-related water outbreaks in the next few years (Andersson and Bohan, 2001).

1.3 Disconnect among water quality, sanitation & health monitoring systems in India

While a lot of the cases above show the clear benefits of having an integrated system for outbreak surveillance, most of the applications of these interagency connections are restricted to developed countries such as the US and UK.

For developing countries like India, which is burdened with water, sanitation and health issues, such interagency collaboration to improve outbreak combat processes should theoretically be providing even more benefits. According to UNICEF's studies in 2004, India was among the top two nations with the largest populations lacking access to an improved water source and improved sanitation (Rheingans, Dreibelbis and Freeman, 2006), as shown in Table 1-2. In India, around 37.7 million

[2] https://www.cdc.gov/healthywater/surveillance/

people are affected by waterborne diseases annually and over 1.5 million children are estimated to die of diarrhea alone, likely due to such dire drinking water and sanitation situations (Khurana and Sen 2008).

Table 1-2 Countries with the largest populations without access to improved water source and improved sanitation (Rheingans, Dreibelbis and Freeman, 2006)

Water*		Sanitation**	
Country	Population lacking access to improved water source	Country	Population lacking access to improved sanitation
China	298 million	India	735 million
India	147 million	China	725 million
Ethiopia	54 million	Indonesia	104 million
Nigeria	48 million	Nigeria	75 million
Indonesia	48 million	Bangladesh	75 million
Bangladesh	36 million	Pakistan	69 million
Dem. Rep of Congo	28 million	Ethiopia	65 million
Vietnam	22 million	Vietnam	47 million
Afghanistan	20 million	Brazil	44 million
Brazil	19 million	Dem Rep. of Congo	36 million
Total	720 million	Total	1.98 billion
% of Total*	67.8%	% of Total**	76.8%

While many academic research studies use integrated water, sanitation and health data from India, most of the data are either extracted through reviewing historical articles or collected for the purposes of the studies only (George *et al.*, 2015; Taylor *et al.*, 2015; Katakwar, 2016). There is little integrated effort to support timely decision-making at the governance level. Even when attempts are made to conduct such assessments to facilitate governmental decision, the results have been relatively weak. For example, a study commissioned by the Department of Drinking Water and Sanitation attempts to assess whether an initiative within the Total Sanitation Campaign decreased waterborne outbreaks. Only the following results in Figure 1-6 are concluded through household self-reporting, a very unreliable source of information when not backed up by data from health surveillance (CMS India, 2011). Interagency data acquisition was not attempted to improve the rigor of the conclusions by the Department of Drinking Water and Sanitation.

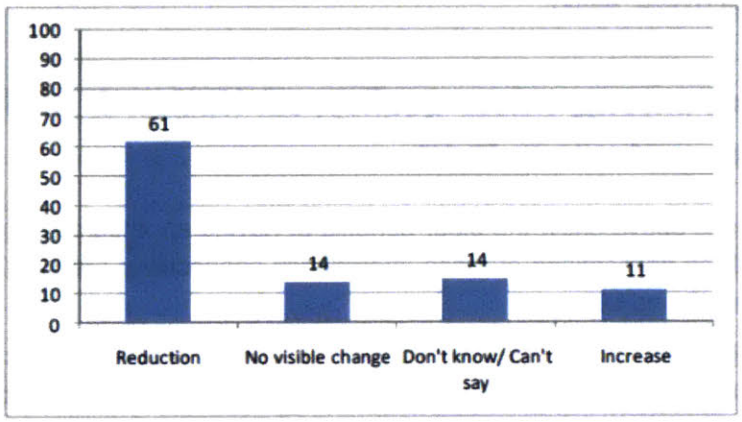


Figure 1-6: Household report on percent changes in the occurrence of waterborne disease as a result of the initiative *(CMS India, 2011)*

For developing countries, one key limitation preventing an integrated approach might be the lack of sufficient data to form an effective connection. As early as 1992, agencies in the US (including EPA, CDC and Agency for Toxic Substances and Disease Registry and so on) have already come together to create an inventory of existing data systems on exposure to toxicants (which includes contaminated water) in the environment and their possible adverse health effects, with the intention to generate new and innovative uses of these data (Eastern Research Group Inc., 1992). In comparison, resources in developing countries may be much too limited to create such comprehensive systems that can more effectively facilitate interagency collaborations. However, this is the not likely the case for India.

To begin with, while there are three separate governmental agencies system monitoring water quality, sanitation and health conditions, each has a comprehensive database collecting the related information. For example, the Integrated Management Information System (IMIS)[3] under the National Rural Drinking Water Programme (NRDWP) is a database that focuses on the conditions of the water sources, with details on community water system attributes and performance in this database are at a level that is not available even for the US. Similarly, the database under Swachh Bharat Mission-Gramin (SBM)[4] specifically focuses on sanitation data, with details on individual household latrine systems and village open-defecation status. Lastly, Integrated Disease Surveillance Project (IDSP)[5] specifically holds a database that focuses on detecting outbreaks and initiating timely responses.

Hence, the infrastructure supporting the interagency collaboration clearly exists. Yet currently the three databases, each with its own political history and focus, are disconnected. At the central government level the National Center for Disease Control oversees outbreak monitoring, while the Ministry of Drinking Water and Sanitation oversees both water quality and sanitation monitoring with occasional data sharing between the two for national level analysis and planning (IMIS website). However, at the state level and below, all three monitoring systems are operating independently with little integration.

Firsthand interview records with water quality lab indicate that they actively work in collaboration with the local disease surveillance units only when there is a waterborne outbreak, such as large-scale diarrhea or cholera (Ms Trivedi, in-person communication). However, the current fragmentation of the monitoring systems offers limited tools for the decision maker to act effectively during and even beyond such emergencies. A decision support tool to facilitate integrated view of water quality, sanitation and health seems to be in high demand.

If there is such a demand, and if the positive effects have already been demonstrated, why is a WaSH-integrated approach towards waterborne disease and outbreak control not yet part of the governmental agenda for India, even though it may benefit even more in compared to developed countries?

1.4 Research question

Many research efforts exist to bring India WaSH-health sectors closer together to increase overall health performance. However, if the barriers to an integrated system cannot be understood fully and

[3] http://indiawater.gov.in/imisreports/
[4] http://sbm.gov.in/sbmreport/home.aspx
[5] http://idsp.nic.in

if pathways towards solution are unclear, research will only exist on a theoretical level but fail to implement in reality. Hence, this study aims to understand the following questions:

- What are the barriers to implementing a data-driven decision support system, which integrates sanitation, water quality, and health data, for controlling waterborne disease outbreaks?
- What are the positive effects of implementing such integrated decision support system?
- How can the barriers be overcome so that the desired effects of such an integrated approach can be realized?

To answer these questions, a thorough literature review is conducted in Chapter 2 on existing frameworks on inter-agency connections in the process of preventing, detecting and investigating outbreaks. The review focuses on summarizing existing evaluation on the implementation of these integrated approaches, including the benefits for outbreak control along with the barriers of implementation. Key gaps for these analyses are identified, and would be subsequently explored in the context of an integrated decision support systems for Gujarat, India.

An overview of the methods used in analyzing the barriers and positive effects of the integrated approach is then outlined in Chapter 3. The different agencies involved in the three sectors are described in detail in Chapter 4 through an institutional analysis. The institutional barriers to implementing an integrated scheme are also analyzed in the chapter. The three separate databases are then introduced in Chapter 5. The challenges of utilizing the database in an integrated outbreak control approach are first described separately for each of the database. An integrated database and model is created in the final section of Chapter 5. The challenges in the integration process, as well as the current limitations of the resulting database are discussed at the end of the chapter. Chapter 6 focuses on the decision support effectiveness of the current WaSH-health integrated model for outbreak control. Considering the many limitations in the current model, the effects are only briefly analyzed through statistical analysis, and conclusions on the current as well as potential benefits of the model is drawn in the chapter. Constraints and barriers that prevented a comprehensive cost-effectiveness analysis are also described at the end of Chapter 6. Conclusions on all key identified barriers and pathways to overcome these barriers are made in Chapter 7. Future work along this pathway towards the adoption and implementation of a WaSH-health integrated framework is laid out at the end of Chapter 7.

2 LITERATURE REVIEW

This chapter reviewed literature on existing integrated WaSH-health frameworks in the process of outbreak control. Literature on WaSH-health integrated approach to different stages of outbreak control is detailed in Section 2.1. Barriers to the implementation of these approaches along with their positive effects are summarized based on literature discussions in Section 2.2. Gaps in the barrier analysis and effectiveness analysis are identified, especially considering the developing country context. The overall research objectives are subsequently defined.

LITERATURE REVIEW

While there are a number of studies focusing on the interconnection among water, sanitation and health, very few focused specifically on the cost-effectiveness of such interconnections. Evaluation of the barriers to implementing their studies for decision-making in future outbreak control circumstances is generally only mentioned as a side note in the discussions. To effectively synthesize these scattered evaluations, literature on WaSH-health integration are categorized and reviewed by the outbreak control stage (prevention, detection, investigation, case summary) that the integrated approach intends to influence.

2.1 Existing WaSH-health framework at various stages of outbreak control

2.1.1 Outbreak Prevention

The majority of studies focusing on the integration of water, sanitation and health approach are aimed at preventing waterborne disease and outbreak and increasing the overall health of the population. Generally, these studies analyze data across the different sectors to connect water and sanitation performances and practices with the ultimate health outcome. Based on the results of these analysis, action plans to prevent waterborne disease are then suggested.

As suggested in Chapter 1, a casual chain framework is adequate when analyzing water-sanitation-health connection for outbreak prevention, because it is able to provide linkage connections between WaSH conditions and the ultimate health outcome in a clear manner that effectively facilitates decision-making (Niemeijer and de Groot, 2008; Gentry-Shields and Bartram, 2014). An early casual framework was the "Pressure-State-Response" (PSR) model, which was later extended into the Driving Force-Pressure-State-Impact-Response (DPSIR) model to take into account societal factors and director modifiers of the environmental state (Gentry-Shields and Bartram, 2014). However, as Gentry-Shields and Bartram (2014) pointed out, intervention points are missing from the model. A model developed by WHO is proposed by the authors as a more action-driven framework: the Driving force–Pressure–State–Exposure–Effect–Action (DPSEEA) model which allows interventions to be targeted throughout the casual chain, as outlined in Figure 2-1. The purpose of the DPSEEA adaption of the original framework aligns with the purpose of cost-effectiveness evaluation, because it is only possible to evaluate the effect on outbreak prevention when action interventions are taken into account. Hence, for literature focused on outbreak prevention, a DPSEEA framework would be utilized to understand the interconnection of factors across water, sanitation and health sectors.

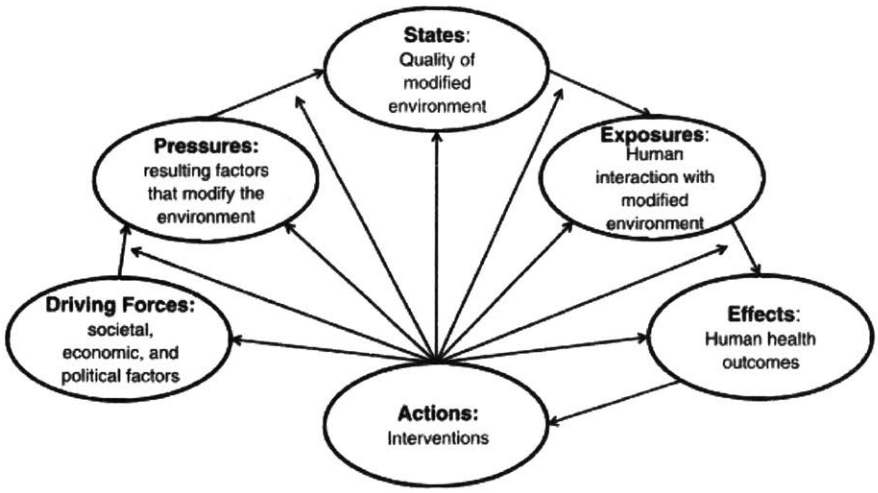


Figure 2-1: Outline of a DPSEEA framework *(Gentry-Shields and Bartram, 2014)*

Based on the DPSEEA framework, integrated analysis for outbreak prevention can generally be categorized into two types:

- **Correlation analysis**, where various WaSH-health indicators are known (symbolized by the round circles in Figure 2-1), and the strength of association among these indicators is studied (symbolized by the arrows in Figure 2-1);
- **Risk assessment**[6], where the strength of association between the WaSH-health indicators of interest are known, and how changes in some indicators can eventually affect others (usually a state or effect indicator with a target level of performance) is studied.

Correlation analysis

Correlation analysis include studies using statistical correlation methods, as well as studies identifying general trends and connections via case summaries.

To begin with, many of the studies focusing on WaSH-Health intersection correlation analysis collected their own data for the purpose of the study (Escamilla *et al.*, 2011; Hlaing, Mongkolchati and Rattanapan, 2016). While this is expected for academic studies, carefully designed small-scale data collection process becomes highly impractical when we are considering a long-term integrated system for decision-making in outbreak control. Hence, the highly scientific and rigorous data generated through these studies are frequently a luxury that national-level governmental data cannot afford. Instead, governmental data frequently include missing data or erroneous data that are collected through varying methods by varying agencies (Strosnider *et al.*, 2014). These studies lack a crucial cost that is common for interagency collaboration, especially in the cases where data sharing is involved. Other studies that utilized metadata through literature reviews or governmental data are more

[6] "Risk assessment" termed in this study is loosely referring to any method attempting to characterize the effect that certain risk factors may have on a performance indicator of interest (which in the case would be waterborne disease or outbreak occurrence). The process can be quantitative or qualitative.

relevant in terms of estimating the data collection cost factor. Broad areas of implementation barriers generally include issues with the data, efforts to integrate them, as well as the capacities of the agencies collecting them.

As for the evaluation of effectiveness, it's important to note that most of these analyses do not directly yield positive effects on outbreak management if the "action" link is missing in the framework. Instead, they point to correlative relationships between data and indirectly offer action steps that can be deduced from these relationships. For example, Cronin et al. (2008) showed statistical evidence that quality and service gap in any one of the water, sanitation, nutrition and health sectors in a refugee camp would have an impact on the another, and came to the action step that integrated approaches must be better planned to tackle issues across all such sectors. However, the correlation between this intervention and other factors in the framework requires additional validation, adding another layer of barrier before the actions can possibly be considered. Hence, these recommendations for action steps would not directly contribute to disease prevention. Only studies that focused more on evaluating the impact of "action" on the rest of the DPSEEA framework offer direct information on the effectiveness evaluation (Fewtrell and Colford, 2004; Khan *et al.*, 2007).

Risk assessment

In comparison literature based on correlation analysis, studies on risk analysis is more proactive as it is focused on distinguishing risks and being preventive (Rizak and Hrudey, 2007). As in the DPSEEA framework, for typical risk assessments, incidence or likelihood for an adverse "effect" caused by "exposure" to risk factors (e.g. risk-inducing "states" resulted from "pressures" by certain "drivers") is estimated. With such assessment of risk, benefits of introducing interventions can also be estimated.

As shown in Figure 2-2, risk management is an expected follow-up action from risk assessment. It corresponds to the "action" item in the DPSEEA framework, that can in turn directly or indirectly impact the ultimate desired "state" or "effect" outcome. This is the missing step in studies based on correlation analysis. As opposed to correlation studies, risk assessments that integrate data across sectors are actually calculating effectiveness by pointing out the chances of outbreak if the assessments are not in place and risk factors are not identified.

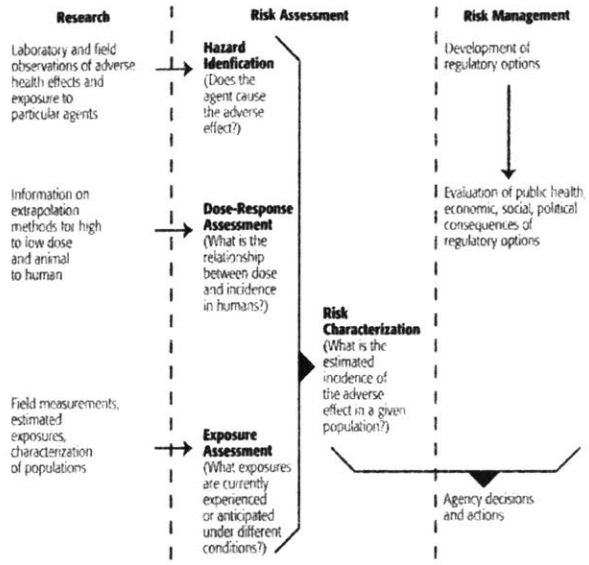


Figure 2-2: Framework for general risk assessment by U.S. National Research Council *(National Academy Press, 1983)*

Moreover, a growing international consensus claims that a preventive risk management that captures sources of risk across WaSH sectors is the most effective means of assuring drinking water quality and ultimately the protection of public health (Rizak and Hrudey, 2007). Such views raise the credibility of risk assessments and in turn improve general perception of their effectiveness.

Implementation barriers on data quality and data collection and integration processes are similar to those for correlation studies. However, there is an even stronger need for a comprehensively integrated database in the case of risk assessment, because the strength of association between WaSH indicators of interest need to be pre-determined in order to carry out the assessment. Risk assessments of outbreaks would be almost impossible without a database containing the most up-to-date information on human exposure to environmental agents and their health effects.

In the US, EPA, CDC and the Agency for Toxic Substances and Disease Registry have taken time and effort to create such a database. Considering the interdisciplinary nature and the required rigor of such a database, many standard operating protocols are required before the database can become effective, such as the following (Sexton *et al.*, 1992):

- standardized procedures for data collection, storage, analysis and reporting
- data retrieval methods that allows easy manipulation of data for model building and testing
- coordination among different entities regarding the design and maintenance of an active information system.

Each additional standard operating requirement above constitute an extra cost factor for an integrated risk assessment system for the prevention of outbreaks.

More recently, there is a similar effort at a nationwide risk assessment database construction by CDC. A National Environmental Public Health Tracking Program is created in order to provide integrated

health, environmental hazard and exposure data (Strosnider *et al.*, 2014)[7]. The tracking network is expected to improve understanding of connections between environmental exposures and public health outcomes, ultimately informing actions to improve health status of communities (Wolff *et al.*, 2008; Strosnider *et al.*, 2014). The results have proven to be fruitful as the Tracking Program was successful in supporting decisions in a number of public health intervention implementations regarding air pollution (Strosnider *et al.*, 2014). The availability and quality of centralized and electronic data available for integration has been thoroughly analyzed before such a Tracking Program can be implemented. Through collaboration between the Tracking Program and academia, critical data gaps were identified via a data assessment framework, and specific data issues associated with community water systems are also discussed at length (Wolff *et al.*, 2008; Strosnider *et al.*, 2014).

2.1.2 Detection

According to CDC, new pattern recognition methods and new types of relevant data may help signal an early onset of outbreak (Buehler *et al.*, 2004). Through collaboration between EPA and Johns Hopkins University Applied Physics Laboratory, increased efforts have been made at exactly these two aspects: utilizing Bayesian Network (BN) analysis as a new pattern recognition and anomaly detection model, and including new data on water quality in the early detection of outbreaks (Babin *et al.*, 2008). This is essentially a hierarchy of networks developed as a module for the Electronic Surveillance System for the Early Notification of Community-based Epidemics (ESSENCE) to quantify the likelihood of a waterborne outbreak via fusing information on the population's health-care-seeking behaviors and drinking water quality in a timely, prospective manner (Burkom *et al.*, 2011). Each component is a probabilistic hierarchy whose inputs are results from statistical alerting algorithms applied to data streams from individual health or water quality indicators, which combines to ultimately improve detection of outbreaks.

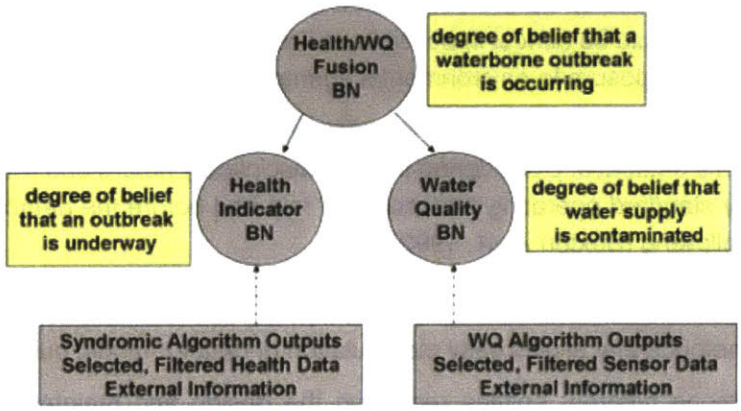


Figure 2-3: A summarized top-level design for BN to detect waterborne outbreaks *(Burkom* et al., *2011)*

The cost-effectiveness of such an integrated system and its capacity for scale-ups have been evaluated at length in EPA's report (National Homeland Security Research Center, 2012). The highest

[7] http://www.cdc.gov/nceh/tracking/

cost factor is the historical data and local expert knowledge required to estimate the conditional dependencies that links connected nodes in BN system (National Homeland Security Research Center, 2012). Other secondary costs are around database construction and maintenance as well as transfer protocols across agencies. On the other hand, data costs may also be reduced considering the perceived advantages of the BN approach – the ability to accommodate of a variety of data that can be traditionally challenging to fuse, as well as a transparent and visual logic system that reduces the reluctance for agencies to use automated decision-support tools (National Homeland Security Research Center, 2012).

The more concrete effect of improving accurate detection of outbreaks (especially the management of false positives) can be statistically demonstrated in Figure 2-4. Due to the lack of multivariate data from known waterborne outbreaks, the effectiveness of the outbreak detection system was only evaluated through a simulated contamination scenario (Burkom *et al.*, 2011).

While US CDC reports only 13-14 drinking water related outbreaks per year, which affects around 1000 people annually, the India state of Gujarat alone reports more than 20 cases of drinking water-related outbreaks a year[8] (Burkom *et al.*, 2011). This also factors into the benefit evaluation of implementing such a system in different country contexts, and may be the reason why the BN module is still in prototype stage and not yet widely adapted in the US.

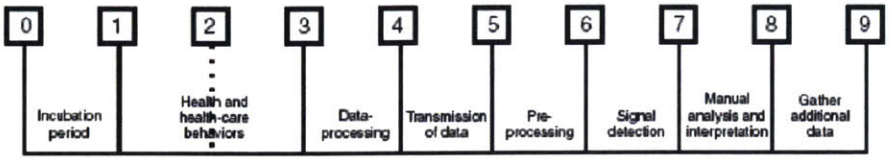
	A	Outbreak occurring	
	B1	Syndromic outbreak evidence	
	B2	Diagnostic outbreak evidence	
Prior Outbreak Degree of Belief		Pr(A)	Pr(~A)
		0.01	0.99
		Conditional Probability Tables	
Conditional Probability Tables		cond Pr(B1)	cond Pr(~B1)
	A	0.7	0.3
	~A	0.01	0.99
		cond Pr(B2)	cond Pr(~B2)
	A	0.5	0.5
	~A	0.001	0.999
Degree of Belief Given One Evidence Source			
Pr(B1,A) (syndromic)	Pr(B1,~A)	Pr(B1)	Pr(A|B1)
0.00700	0.00990	0.01690	0.41420
Pr(B2,A) (diagnostic)	Pr(B2,~A)	Pr(B2)	Pr(A|B2)
0.00500	0.00099	0.00599	0.83472
Degree of Belief Given Two Evidence Sources		Pr(A,B1,B2)	0.00350
		Pr(~A,B1,B2)	0.00001
		Pr(B1,B2)	0.00351
		Pr(A|B1,B2)	0.99718

Figure 2-4: Demonstration of the effect of how two corroborating data sources may increase the degree of belief that an actual event is detected *(Burkom* et al., *2011).* This applies similarly to the fusion of water quality and health evidence in increasing the degree of belief for outbreak detection results. Consider A to be the probability of an outbreak occurrence, and B1 and B2 as

[8] Data summarized from India's IDSP database

probabilities of anomaly detection through different data inputs. Using the Bayes' Theorem, an outbreak likelihood of over 99.7% can be confirmed based on quantified corroboration from fusion of both anomalies, as opposed to 41% and 83% when the two sources of evidence are considered separately.

Apart from the improvement in the validity of outbreak detection through reduction of false negatives, the effectiveness of outbreak detection systems can also evaluated based on their detection timeliness along the timeline milestones for early outbreak detection, as shown in Figure 2-5 (Buehler *et al.*, 2004).



0. Onset of exposure
1. Onset of symptoms
2. Onset of behavior
3. Capture of data
4. Completion of data processing
5. Capture of data in surveillance system
6. Application of pattern recognition tools/algorithms
7. Generation of automated alert
8. Initiation of public health investigation
9. Initiation of public health intervention

Figure 2-5: Timeline milestones for the evaluation of outbreak detection timeliness (Buehler *et al.*, 2004)

2.1.3 Investigation

Interagency collaboration during outbreak investigation processes has been consistently encouraged, and even discussed at length by governmental officials during workshops on waterborne disease outbreak in the US (Craun *et al.*, 2001). As mentioned in Chapter 1, environmental factors such as water quality contamination contributes to the different levels of evidence when confirming an outbreak in a number of countries.

Evaluation of the cost-effectiveness of interagency collaboration during an outbreak investigation process has mostly been through case studies. Case study approach was more appropriate because of the miscellaneous details and large number of variables that changes from case to case (Gelting *et al.*, 2005). A waterborne Novovirus outbreak in Wyoming has been used as an example for studies by both Cassady et al. (2006) and Gelting et al. (2005) as a successful implementation of the integrated approach, with the conceptual model of the case shown in Figure 2-6. A waterborne outbreak in Vuorela, Finland was also used as a success case for integrating novel microbial and water system spatial statistical methods to identify source of infections (Jalava *et al.*, 2014). In both cases, by inclusion of an environmental health assessment component, the underlying environmental antecedent that led to the water supply contamination was efficiently and successfully identified.

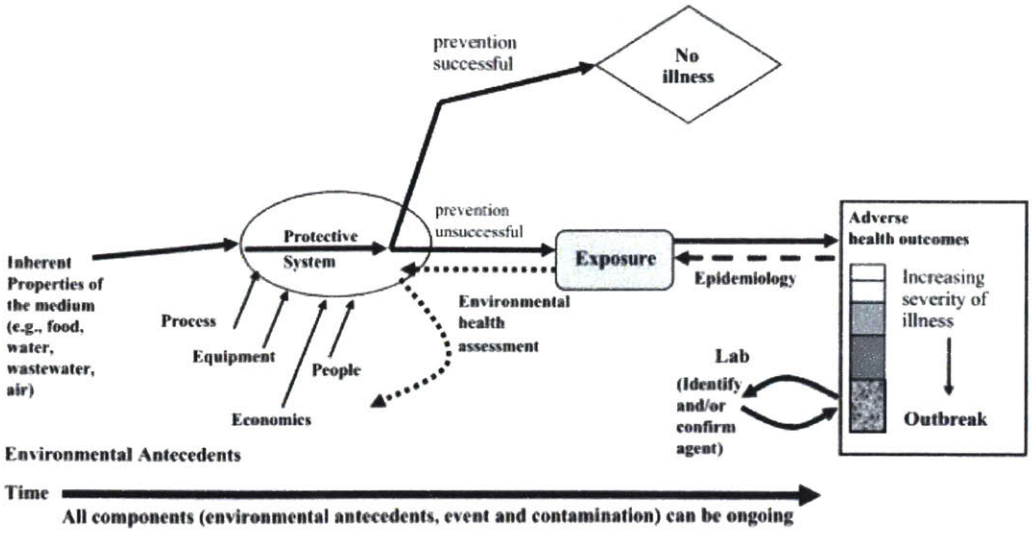


Figure 2-6: Conceptual model of an outbreak investigation utilizing a systems-based approach. As demonstrated by the dashed lines, while epidemiologic investigation can identify the connection between exposure and health effects, environmental assessments are crucial in tracing further backwards in time to discover the underlying environmental antecedents that resulted in the change of the contaminated medium (Gelting *et al.*, 2005).

Barriers to implementing such integrated assessments are also identified in these studies. The inclusion of extra environmental health specialist or environmental engineers requires additional resources (Gelting *et al.*, 2005). The inclusion of professionals from different fields in the investigation team also increase the communication cost between disciplines (Gelting *et al.*, 2005). Standard methods for outbreak-related environmental assessments are not readily available and the development process would also add to the cost (Gelting *et al.*, 2005). However, the effect is also obvious with the increased successes in identifying outbreak causes. As the outbreak unfolded in Wyoming, all states and federal agencies dealing with water and wastewater systems in Wyoming have come together to develop a task force specifically focused on an "integrated, system-based approach" to the response and prevention of waterborne outbreaks, and further interagency collaboration efforts and guidelines for standard practices ensued (Cassady *et al.*, 2006). Such evidence in literature suggest that WaSH-health collaboration in outbreak investigation has been cost-effective.

2.1.4 Case summaries

While case summary is a crucial step in the chain of outbreak control, it acts more as a research method to uncover the interconnection among water, sanitation and outbreak cases, rather than a segment of the outbreak management process where the value of integrated approach needs to be assessed independently.

For example, Schuster et al. (2001) conducted a detailed overview on waterborne outbreaks occurring between 1974 and 2001 in Canada. CDC also works with EPA to compile a report on waterborne disease and outbreaks biannually (CDC 2013; Yoder et al., 2008). These case reviews analyze the connections between outbreak occurrence and the corresponding etiologic agent, water sources or

other environmental factors. They identify correlations to understand trends in waterborne outbreaks for the purpose of outbreak prevention. Hence, these studies are incorporated into the section on correlation analysis for outbreak prevention.

2.2 Pathways towards an integrated water, sanitation and health system

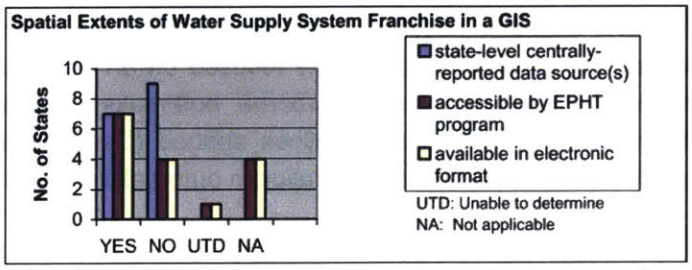
2.2.1 Implementation barriers

Commonly identified implementation barriers of an integrated WaSH-health system for outbreak control processes are the cost of constructing an integrated data system, as well as the cost of inter-agency collaboration and creation of collaborative protocols.

Challenges for an Integrated Data System

Commonly identified challenges to data integration include the inconsistent data quality, agency-dependent procedures for data collection, missing data and so on. For example, the scientific capacity of personnel collecting water quality data is frequently a costly barrier to accurate data, especially considering the high contamination risk during E.coli testing procedures (Khan et al. 2007).

Based on these concerns, Wolff et al. has created a data assessment framework (Figure 2-7) to assess barriers for an integrated Environmental Public Health Tracking (EPHT) Program network. The cost of database integration is assess through a yes/no binary evaluation of the availability and accessibility of centrally-reported data sources and whether there is an electronic format (Wolff et al., 2008). The completeness, as well as the quality and accuracy of the dataset are also rated in a semi-quantitative format (Wolff et al., 2008).



	Min	Max	UTD	NA
Completeness of this data element	0	100	2	7
Quality/accuracy of this data element	D	A	2	7

Figure 2-7: Sample data assessment process for the development of an integrated EPHT Program Network *(Wolff et al., 2008)*

In addition to evaluating the barriers to integration-friendly data, there is a separate set of criteria for evaluating barriers to useful data. For data used in public health surveillance processes, CDC has a specific set of evaluation criteria for an effective data system, as shown in Table 2-1 (Niskar, 2007).

The WaSH-health integrated system is also expected to satisfy all these criteria, and the current gaps are additional data costs to consider.

Table 2-1: Data evaluation criteria for use of public health surveillance *(adapted from CDC, 2001).* The evaluation for early detection of outbreaks is a slightly modified version of the table below, but the essential components remain the same *(Buehler* et al., *2004).*

Evaluation criteria	Details
Simplicity	Simple data system structure that is also easy to operate
Flexibility	Adaptive capacity to regulation changes and new information with little additional time, personnel or allocated funds. Systems using standard data formats can be easily integrated and would also be considered flexible
Data Quality	Completeness and validity of the data recorded in the system
Acceptability	Willingness of persons and organizations to participate in the surveillance system
Sensitivity	Whether trends and violations can be effectively detected
Predictive Value	Whether non-compliant data actually reflects a key problem in the system of concern
Representativeness	Ability to generalize data to the majority of cases applied to the population at large
Timeliness	Prompt actions between steps in the system
Stability	Reliability (i.e., the ability to collect, manage, and provide data properly without failure) and Availability (the ability to be operational when it is needed)

Lastly, for the purpose of risk assessment and outbreak detection, knowledge on the theoretical relationship between variables is crucial to deduce the risk of an outbreak, or the likelihood that an outbreak has already happened. Hence, the availability of historical data and expert knowledge is also an important cost factor.

Barriers to interagency collaboration

In comparison to data barriers, institutional barriers are not analyzed as extensively in literature. Most literature mention the need for collaboration protocols for communication across disciplines, but the barriers to such types of protocol are not mentioned. Since most of the reviewed literature on WaSH integration is focused on the US or Canada, this is likely because disease control agencies and water quality agencies (e.g. US EPA and CDC) already has a long history of collaborative work since the 1970s and the protocols for collaboration simply evolved in the process (Schuster *et al.*, 2001; Centers for Disease Control and Prevention, 2013).

However, assessment of the cost of interagency collaboration is critical. As Jalba et al. (2010) mentions in her study, water, sanitation and health agencies frequently have distinct objectives. Each with its own goal, culture and training processes, the agencies have little incentive to work together on a proactive level unless there is a clear indication of health risk. Hence, while most countries can overcome institutional barrier during catastrophic events (such as waterborne outbreak investigation), interagency efforts on risk management before and after such events remain challenging yet highly

necessary (Gelting *et al.*, 2005; Jalba *et al.*, 2010). Six critical institutional relationship components are identified by Jalba et al. as deficient in past outbreak control practices (Jalba *et al.*, 2010).

Table 2-2 Components of institutional barriers to a collaborative effort in outbreak control (Jalba *et al.*, 2010)

Areas of Institutional Barrier	Details
Proactivity	Insufficient preparation for incidents, and general regulatory passivity for preventative activities during normal times
Communication	Lack of regular lines of communication between agencies, potentially resulting lack of trust and contradicting actions
Training	Lack of interdisciplinary WaSH-heath training, and skills related to risk management, communication and collaborative incident management
Sharing expertise	Limited exchange of information and expertise between stakeholders, potentially resulting in missing actions along the management chain
Trust	Lack of trust between agencies at the institutional level and personal level, frequently due to misguided and narrow-minded attempts to protect reputation, sometimes resulting in unilateral or hostile actions at risk control
Regulation	Gaps in regulatory requirements and oversight role definitions, resulting in confusion over responsibility and delayed (or evening missing) action steps

Gaps for implementation barrier analysis

Overall, as shown in summary Table 2-3, apart from studies where data is collected by the research team, data barriers and institutional barriers documented in existing literature are generally applicable to analyzing the cost of an integrated WaSH-health approach to outbreak control. While frameworks for data-related barriers are quite comprehensive, only Jalba et al. brought forward a framework to analyze institutional barriers and it is yet to be applied to other research. Hence, most studies only scored 2 in the cost analysis feasibility rating. On the other hand, the EPHT program and cross-agency outbreak investigation teams have already been successfully and widely implemented, suggesting both data and institutional costs have already been calculated and overcome. For studies under these categories, the cost analysis feasibility is given a 3.

It is important to note that while the framework to analyze data barriers is quite comprehensive in literature, two of these frameworks are established by CDC specifically for the US. Much of the challenges regarding rural developing communities are unaccounted for in these cost analysis (CDC, 2001; Wolff *et al.*, 2008). For example, in rural India, spellings of village names are frequently inconsistent across databases, as most of the social workers reporting data at the village level are not aware of the standard spelling of these names (if a standard spelling even exists). Similarly, the framework for studying institutional barrier is also very restricted to the developed countries' context as it only used cases from developed nations including UK, Australia, US, Sweden and Canada are used (Jalba *et al.*, 2010).

Lastly, while analysis of barriers offer insights into the potential costs to implementation, there is a lack of proposed solutions to these barriers. The cost of these solutions would be more directly

representative of the implementation cost. Some literature focus on a detailed review of a specific type of solution – for example, Liu *et al.* (2016) proposed a data interpolation method informed by spatial data and expert knowledge to deal with missing or inconsistent data for the purpose of evaluating waterborne disease potential. However, there is no proposed framework through which solutions and their costs may be evaluated.

2.2.2 Effectiveness factors

The effectiveness of an integrated approach for outbreak control should be evaluated at the different stages of the control process that the approach is targeted towards. The key stages of interest are outbreak prevention, detection and investigation.

Outbreak prevention

The direct effect of outbreak prevention is a decrease in outbreak occurrence. However, it is generally impossible to attribute such decrease to any one specific prevention approach. Outbreaks that are actually prevented are very hard to account for, resulting in considerable difficulty for evaluation of outbreak detection efforts. Fortunately, despite the lack of direct statistical proof for benefits, risk assessment procedures are still globally recognized for its ability to reduce outbreaks (Strosnider *et al.*, 2014).

As for indirect measures of effectiveness, disease incidence calculated by risk assessment procedures may be considered a proxy for the prevented disease or outbreaks. Similarly, for correlation studies that connect intervening actions to other arenas of the DPSEEA framework, it is also possible to approximate the prevention of outbreaks by estimating the potential change in health effects in correlation with positive intervention. Considering that proxy factors are used, the feasibility for the effectiveness estimation are rated at 2 for these studies, as shown in Table 2-3 (correlation analysis studies that included an intervention component are marked by italics). Among these studies, the risk-based approaches are more well-recognized for WaSH system management and decision-making, and would be more readily implementable once proven cost-effective. Hence, the cost-effectiveness evaluation of integrated risk assessment approach to outbreak prevention is likely of more practical value and would be prioritized in this study.

For correlation studies that did not incorporate any "action" component, there is not a clear way to approximate outbreak prevention resulting from the studies. Without a clear correlation between intervention and effect, it is impossible to capture the amount of positive effect on health status that interventions based on such studies can lead to. Hence, all correlations studies without an "action" component are rated 1 for the feasibility of effectiveness estimation.

Outbreak detection

Evaluation on the effectiveness of outbreak detection through WaSH-health integration is straightforward. Timeliness and validity are the key targets for any outbreak early detection mechanism. For the BN module, while the timeliness is not directly evaluated, the calculated improvements on the validity of detection and control of false positives are sufficiently demonstrating the effectiveness of

such an integrated approach (Burkom *et al.*, 2011). However, the only drawback is that the study only used simulated data to prove the validity of detection. Limited test runs have been conducted with real-world scenarios (National Homeland Security Research Center, 2012). Hence, the feasibility of effectiveness estimation for the studies on the BN module for outbreak detection are rated in between 2 and 3.

Outbreak investigation

Effectiveness of integrated approaches during an outbreak investigation is also straightforward. While the details of every outbreak vary greatly and individual case studies are commonly used to understand the investigation procedure, there is a consistent expectation to identify the etiologic agent of the outbreak as well as the WaSH-related antecedents that brought about the contamination agent during an investigation. For all cases reviewed in this section, the outbreak causes are successfully uncovered through collaborative investigation efforts, directly demonstrating the effectiveness of such integrated approaches to outbreak investigation. Even in developing countries, active engagement of water and sanitation sectors during outbreak investigation is practiced more and more widely. For example, through interview with the lab director of Vadodara Zonal Lab in Gujarat, Ms Trivedi, it was learned that a recent outbreak in cholera was dealt with in the following process:

- The cases first came to the health department, which then involved the concerned municipalities and informed water department to start tanker water supply;
- Super chlorination was carried out throughout the system by water department;
- Concerned agencies were brought together to start searching for the source of contamination;
- Once the issue is resolved, the labs rechecked the sites of contamination and restarted the supply once everything is confirmed safe.

Even though these are generally still ad-hoc efforts with no reliably consistent operational protocol, it still shows that interagency efforts are valued in this context. This suggests that effectiveness of a collaborative investigation effort has been widely evaluated, and it generally seem to outweigh the cost. Consequently, the feasibility of effectiveness analysis for outbreak investigation-related studies are rated at 3.

Gaps for effectiveness analysis

Apart from WaSH-health correlation analysis studies that does not involve the "action" component within the DPSEEA framework (e.g. analyzing how a "state", types of water sources, can impact the "effects" - health status of the concerned population), it is reasonably viable to evaluate the effectiveness of WaSH-health integrated approaches. However, effectiveness evaluation for risk assessment and BN outbreak detection is still indirect at best. Risk assessments are generally not followed up with how the risks are actually addressed and whether outbreaks are actually prevented as a result. BN modules are yet to be tested with real data.

2.2.3 Summary

Following summary in Table 2-3, the key literature gap that this study attempts to resolve can be identified.

WaSH-health integrated correlation studies cannot be proved effective for outbreak prevention unless an intervention component is accounted for in the study. Studies that do not incorporate an intervention component might not be relevant to our goal. As for integrated outbreak investigation approaches, they are already applied across the world, suggesting that its cost-effectiveness has already been quite clearly demonstrated. These two types of WaSH-health integrated approaches would not be the focus of this study.

While existing literature on risk assessment for outbreak prevention and BN modules for outbreak detection show strong positive evidence, but gaps towards implementation still exist before these approaches can be deemed cost-effective, especially in the context of a developing country like India. Reasons are outlined below.

First of all, frameworks to analyze barriers to data integration require adjustment to suit the context of rural communities in a developing country. It is also important to note that hygiene and sanitation data integration was rarely mentioned, because they are no longer risk factors in the context of a developed country. However, sanitation is still a critical challenge in India, and integration of sanitation data must also be considered.

In addition, while institutional barriers analysis framework is available, it is yet to be applied to cost assessments for WaSH-health integrated approaches. As institutions are the final implementers of such integrated approaches, institutional barrier analysis is essential for planning implementation roadmaps.

For the effectiveness analysis, adjustments are also needed for the rural developing community context. For example, the intervention of treating contamination along a piped water distribution network would impact health effects at a very different exposure scale compared to treating contamination at a local community well. Moreover, the effectiveness analysis for both risk assessment and BN modules are indirect at best and still lack validation. Risk assessment results are not realized until the intervention of concern is adopted by governmental agency. BN early detection results are not realized unless the module is successfully applied during an actual outbreak case. Validation steps are essential for an accurate estimation for the effectiveness of WaSH-health integrated outbreak control approaches.

To resolve these gaps, the cost and effectiveness of the integrated approaches would be evaluated for the rural India. While both risk assessments and BN early detection are of interest, the time scope of this study only allows detailed analysis for one type of integrated approach. Over the course of the study, multiple attempts have been made to obtain detailed health surveillance data through the Health Department but no data have been shared. Without any public health surveillance information on syndromes and patient cases, the BN module would be missing one of its most critical inputs, which severely limits its effectiveness in detecting outbreaks. Hence, the BN early detection module is left for future studies when more data become available.

Additionally, while effectiveness analysis is essential, considering the barriers that not possible to be bridged over the course of this study, an accurate evaluation of framework effectiveness at this stage

is challenging. Instead, an attempt is made to evaluate the possible effects of implementing such an approach, and through the analysis, pathways towards a more reliable effectiveness evaluation will be identified.

Consequently, this study would focus on the implementation barrier analysis of an integrated assessment framework that connects across water, sanitation and health sectors to characterize and ultimately prevent risks for waterborne disease and outbreak. Solutions to these barriers are also concluded so that a clear implementation pathway forward can be outlined. Effectiveness will also be briefly evaluated at a limited scale. Pathways towards more comprehensive cost-effectiveness demonstrations are laid out for future studies.

The India context also sheds a special light this study. New governmental initiatives in India to increase WaSH practices and WaSH-heath data collection started only around 2009-2010. Many drastic changes in water quality, sanitation and health interventions have occurred in the past few years, which are likely to offer interesting data that can readily assist in the understanding of a WaSH-health integrated risk assessments approach to outbreak prevention.

Table 2-3: Summary and categorization of literature reviewed, along with feasibility rating of the cost and effectiveness analysis of the integrated approaches documented in literature

Outbreak Control Stage	Citation	Variables analyzed	Source of data	Key analysis method	Feasibility of Implementation Barrier Analysis (1-3*)	Feasibility of Effectiveness Analysis (1-3*)
	Cronin et al. 2008 Carlton et al. 2012 Teschke et al. 2010 Schuster et al. 2001 WBDOSS biannual case report *Khan et al. 2007*	Health outcomes - water and sanitation indicators	Governmental data	Correlation analysis	2	1/2 (depends on whether intervention is included)
	Taylor et al. 2015 Murphy et al. 2014 Traore et al. 2013 Wallender et al. 2014 Gundry, Wright and Conroy 2004 *Fewtrell and Colford 2004*	Health outcomes - water and sanitation indicators	Metadata	Correlation analysis	2	1/2 (depends on whether intervention is included)
Prevention	Hlaing et al. 2016 Escammilla et al. 2011	Health outcomes - water and sanitation indicators	Collected by research team	Correlation analysis	1	1
	George et al. 2015 Liu et al. 2016	Disease/outbreaks risk - water and sanitation indicators	Collected by research team	Risk assessment	1	2
	Summerscales and McBean 2010	Disease/outbreaks risk - water and sanitation indicators	Governmental data	Risk assessment	2	2
	Strosnider et al. 2014 Wolff et al. 2008 Sexton et al. 1992	Environmental hazard, human exposure and health effects surveillance	Governmental data	Risk assessment	3	2

Investigation	Cassady et al. 2006 Craun et al. 2001 Gelting et al. 2005 Jalava et al. 2014	Outbreak cases - environmental investigation	Governmental data collected by the outbreak response team	Case studies	3	3
Detection	Burkom et al. Babin et al. National Homeland Security Research Center 2012	Outbreaks - Water quality, disease syndromes	Collected or simulated by research team	Bayesian Network Analysis (based on risk assessment principles)	2	2/3 (depends on whether simulated data or real data are used)

***Note on rating scale:**
1 - highly challenging or not applicable to the focus of this study;
2 - possible to carry out, although thoroughly documented in literature;
3 - very clearly documented in literature.

3 METHODS

Methods on barrier and effectiveness analysis for the WaSH-health integrated framework are described in Chapter 3. Section 3.1 describes the selection of Gujarat as the study site. Section 3.2 and 3.3 outlines the analysis method for interagency collaboration barriers and data integration barriers respectively. Method for evaluating the effectiveness of the resulting integrated framework is described in Section 3.4.

METHODS

3.1 Site selection

While the Ministry of Drinking Water & Sanitation monitors WaSH[9] status at the central level, ultimately water quality management in India is a state-based issue. Each state adapts the guidelines and forms their own practice. Hence, it would be more effective to focus the research on one specific state to consistently explore the interagency collaborations. The state of Gujarat is selected as the subject of this research.

Gujarat is situated in the west coast of India (Figure 3-1). A basic profile of Gujarat can be shown below in Table 3-1. With a population of 6 crores, approximately 4.99% of total Indian population, Gujarat has 7.3% of India's GDP and 40% more per capita income than India average (Directorate of Economics and Statistics, 2012). As of 2012, the state also has a literacy rate of 78% in comparison to India's total literacy rate of 73%. From the health perspective, the IMR of rural Gujarat is only at an average level. In comparison, Kerela state has already achieved an IMR of 12[10].

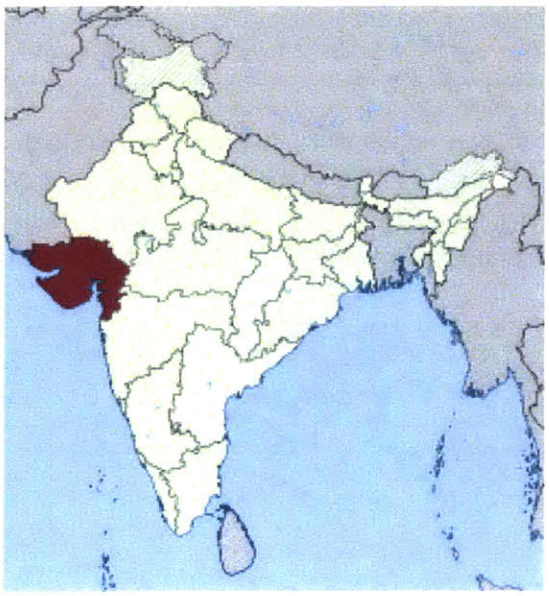


Figure 3-1: The location of Gujarat within India[11]

[9] Note that in the context of India's sanitation monitoring, hygiene practices are not recorded due to the challenge in recording individual practices at the household level. However, this research still loosely refers the water and sanitation monitoring efforts in India as WaSH monitoring, as increasing efforts are being made to include behavioral data in the sanitation monitoring process.
[10] http://niti.gov.in/content/infant-mortality-rate-imr-1000-live-births
[11] http://wikitravel.org/en/Gujarat

Table 3-1: Basic profile comparison between Gujarat and India (Directorate of Economics and Statistics, 2012)

Parameter	Gujarat	India	Ratio
Towns	348	7925	4.4%
Villages	18225	640867	2.8%
Area (km2)	196244	3287469	6.0%
Total Population (Millions)	60.44	1210.57	5.0%
Rural Population (Millions)	34.69	833.46	4.2%
Urban Population	25.75	377.11	6.8%
GDP (Billion Rs.)	6117.67	83534.95	7.3%
Per Capita Income	89668	61564	
Literacy Rate	78%	73%	
Rural Literacy	72%	68%	
Urban Literacy	86%	84%	
Infant Mortality Rate (IMR, per 1000 births)	41	44	
Urban IMR	27	29	
Rural IMR	48	48	

Overall, Gujarat presents an interesting case because while it showcases an impressive economic growth, the health and social development indicators remain unsatisfactory (Department of Health & Family Welfare, 2014).

Extensive work on water issues has already been done in Gujarat by MIT colleagues, and close rapport has been established with the water monitoring agencies in Gujarat (Novellino, 2015; Wescoat, Fletcher and Novellino, 2016). Hence, with the pre-existing connections to the government of Gujarat, selecting Gujarat as the subject of study would allow for potential long-term connection and higher possibility of carrying out further pilot runs of the integrated framework if this research project comes to fruition.

3.2 Barriers to interagency collaboration

As suggested in Chapter 2, two main barriers to the implementation of a WaSH-health integrated approach to outbreak prevention are the challenge of interagency collaboration and data integration.

A clear understanding of the cost of inter-agency collaboration can be established through effective institutional analysis. Governmental documents were reviewed to create a comprehensive picture of the structure of the three institutions monitoring rural drinking water, sanitation and health. To be more specific, while diseases and outbreaks monitoring are part of health monitoring, they differ from routine health statistics monitoring such as records of infant mortality or cancer rates. Outbreak and outbreak related disease, which is the subject of interest for this study, should be noted separately from here on, as opposed to the overarching term of "health monitoring".

In the context of India, the sanitation agencies primarily monitor household latrine construction. Disease-related agencies primarily conduct surveillance on outbreaks as well as symptoms or diseases related to possible outbreaks. Rural drinking water agencies monitor a much wider variety of

information, including rural demographics, water supply, water quality, personnel training and so on. Generally, water quality, especially microbial contamination of the water, is more directly related to waterborne diseases. Thus, among all the parameters in the water monitoring database, water quality, especially parameters related to biological contamination, is of the most interest. Although there are diseases caused by a lack of water access (e.g. no water to wash hands), broadly termed "water-washed diseases", they are typically not cause for large scale and would not be the primary focus of this study (Gentry-Shields and Bartram, 2014). Chemical contamination is more likely to cause chronic diseases (e.g. fluoride contamination – dental fluorosis), which are also not the focus of this study. In conclusion, the subject of WaSH-health monitoring can be narrowed down to water quality, sanitation and disease/outbreak monitoring.

Outlines of the institutional hierarchies were then created for institutions focusing on water quality, sanitation and disease/outbreak monitoring, based on literature and modified through direct communication with these institutions.

In addition, informational interviews were conducted with key personnel at these institutions. These interview results are also analyzed to understand each agency's existing monitoring practices, decision-making processes and overall challenges. Questions on their interest for interagency connections and their perception of barriers were also posed. Outline for the key interview questions is listed below.

Table 3-2: Sample agency interview outline

Topics	Sample Questions
Responsibility	What types of responsibilities are there regarding water quality, sanitation and health monitoring?
	What are the general roles of agencies at the national, state, district and community levels?
Capacity	What is the general distribution of workforce?
	How is the workforce trained?
	Are their specific programs in place to increase capacity?
	Are there collaborations with other entities?
Scale	What is the geographic coverage of the monitoring conducted?
	What is the quantity of specific monitoring activities conducted?
Data Collection	Who collects the monitoring data?
	Are there validation of the data entries?
	What other variables would you like to collect in support of these variables?
Data-triggered Action	Is the data currently being used in any way?
	What decisions are based on this data?
	What other data or information is necessary for these decisions?
Data Analysis	Is the data analyzed or mapped in any way, by whom?
	What decisions are based on the analysis of these data?
	What type of analysis on the data would help with decision making?
	What other data or information would be helpful for these analyses?
Interagency Connection	Are there current interagency collaboration efforts? What about for WaSH-health related purposes?
	Are monitoring data by your agency related to data in other databases? What about for WaSH-health related purposes?
	Would agency's decisions benefit from such interagency connection?

	If agency would benefit, ideally how should it be connected to other agencies and their data systems?
	Would it be possible for these connections to be done?
	What are barriers preventing them from being done?
Improvement	What are some of the most successful practices?
	What are some of the most difficult challenges?
	What is the key challenge for a comprehensive WaSH-health monitoring system?

Three districts were visited for informational interviews, including:

- Bhavnagar: local connections available via our key contact, Tata Water Mission;
- Dang: the smallest district with almost 100% tribal population, with makes for an interesting case to explore;
- Narmada: the only district as of summer 2016 that achieved 100% latrine construction status, and the implications to water quality and health would be interesting to explore.

After summarizing the practices and expectations for each separate monitoring institution, an interagency collaboration barrier evaluation is conducted based on the components laid out in the framework by Jalba et al. – proactivity, communication, training, sharing expertise, trust and regulation. Referring to the framework, we will similarly evaluate the institutional barriers are evaluated through the following aspects:

- Incentive (proactivity has generally not been there, so instead the motivation to be proactive in the future is evaluated);
- Existing connections (consolidating "communication", "sharing expertise" or "training" to look at all existing channels of connection together);
- Trust;
- Regulation.

3.3 Barriers to data integration

3.3.1 Variable assessment within each database

To understand the barriers to integration across the water quality, sanitation and disease/outbreak database, each of the database and its variables of interest are explored. The historical development of the database, and the data collection and validation processes are also reported.

Following suggestions by Burkom et al. on parameter selection for outbreak management, an inventory of variables necessary for outbreak risk assessment is created.

We first assess challenges that each individual variable may pose to a successful integration. Wolff et al. and Buehler et al.'s framework for the data assessment is adapted for assessment at the individual variable level. Each of the variable is evaluated by the following attributes:

Data characteristics:

- Accessibility: Ease of access for analysis and data integration purposes
- Simplicity: Simple data definition and data storage structure
- Uniformity: Uniformity of data entry formats
- Completeness: Whether data is available across all existing units (e.g. water sources, habitations) where data are expected to be measured, and across all years that the database has been available
- Quality/Accuracy: Validity of the data recorded in the system
- Integration viability: Ease of integration into a schema, which requires primary key for each dataset table and standard reporting formats across datasets.

Data Utility:

- Acceptability: Willingness of workers and organizations to engage in the data collection and monitoring process
- Sensitivity: whether trends and violations are effectively reflected via the variable
- Predictive value: whether non-compliant data is positively reflecting a key violation in the system of concern
- Timeliness: Prompt data entry and validation processes and appropriate data collection frequencies that allows for timely recognition of issues in the system.

Both qualitative assessments and quantitative analysis are used when possible to evaluate uniformity, completeness, quality, integration viability and overall utility of the data. Other factors are only evaluated qualitatively.

3.3.2 Database integration assessment

After analysis at the individual variable level, an effort is made to connect across the water quality, sanitation and outbreak databases using a database schema. Through the process of implementing the schema and constructing the database, data integration barriers are evaluated. This assessment framework in Section 3.3.1 can be adapted again for assessment at the database level. Consequently, the WaSH-health integrated database can be evaluated by the following attributes:

- Simplicity: simple database structure with ease of operation
- Flexibility: adaptive capacity to regulatory or other changes
- Uniformity: consistency of data reporting formats and data results of similar variables across the databases
- Completeness: whether data entries for all administrative units where WaSH-health data are all expected to be collected are available, and whether all variables essential to the integrated outbreak prevention framework are available
- Integration viability: ease of carrying out the schema through connecting unique identifiers
- Stability: reliability in collecting, managing and providing data without failure and availability to operate smoothly at any given occasion.

The completeness of records across sectors can be evaluated quantitatively. Uniformity and integration viability can also be evaluated semi-quantitatively by observing the scale and magnitude of discrepancies across databases. Other criteria would be evaluated qualitatively.

The overall predictive value and sensitivity are evaluated separately as the effectiveness of the integrated system in the following section.

3.3.3 Color coding

Considering that most analysis in this section are done across the 3 sectors, the data tables and charts for the three sectors are color coded separately for ease of differentiation:

- Yellow: tables related to IMIS and water quality monitoring
- Blue: tables related to SMB and sanitation monitoring
- Black: tables related to IDSP and disease/outbreak monitoring.

3.4 Effectiveness for outbreak prevention

After creating the integrated database, it would be adapted to a decision support model framework based on the DPSEEA casual chain framework for waterborne disease control. Regression analysis is conducted to assess the correlation and correlative significance between critical components along the casual chain framework.

Through interpreting the correlative relationships between the different WaSH-health components along the DPSEEA chain, the current capacity of the model is analyzed, and estimations on future decision support potentials are also concluded. Current limitations that prevented a more comprehensive effectiveness analysis are also outlined to suggest future pathways forward.

4 INSTITUTIONAL ANALYSIS

Chapter 4 focus on analyzing the three separate institutions monitoring water quality, sanitation and diseases. The general hierarchical structures of three institutions and other associated agencies in the field are outlined in Section 4.1-4.3 respectively. Their data collection and utilization practices are also described, and their motivation for interagency collaboration is analyzed. An overall evaluation on the interagency collaboration motivations and challenges are summarized in Section 4.5.

INSTITUTIONAL ANALYSIS

Institutional analysis is conducted based on governmental document reviews and interviews. Interviewees are selected from various important agencies across WaSH-health sectors, with the attempt to bring forward a variety of voices across different entities and different administrative levels (as defined in Table 4-1) to gain a more comprehensive and transparent view of WaSH-health monitoring practices and collaboration motivations across rural India.

Table 4-1: Definitions of the different hierarchy in administrative levels for rural India, listed in top-down order (IMIS, no date)

Administrative Level	Definitions
State and District	India is a federal union of states comprising twenty-eight states and seven union territories. The states and territories are further subdivided into districts.
Block	A block is an administrative entity that districts are further divided into. The jurisdiction is generally limited to rural parts of a district.
Gram Panchayat (GP)	Gram panchayats, local self-governments at the village or small town level, are further divisions within blocks. Frequently two or more villages are clubbed together to form a group - gram panchayat especially when the population of the individual villages is less than 300.
Village	A village is a clustered human settlement or community with the population ranging from a few hundred to a few thousand (sometimes tens of thousands). Each gram panchayat may contain one or more villages.
Habitation	Villages are further divided into habitations, which are usually a group of families living in proximity to each other within a village.

Unless otherwise cited, results in this chapter are collected through interviews.

4.1 Water quality monitoring institutions

4.1.1 General Information

The National Rural Drinking Water Program (NRDWP) was set up in 2009 as an ongoing program ensure sustainable rural water supply in India (Ministry of Drinking Water and Sanitation, 2013a). NRDWP is implemented at different administrative levels according to Figure 4-1.

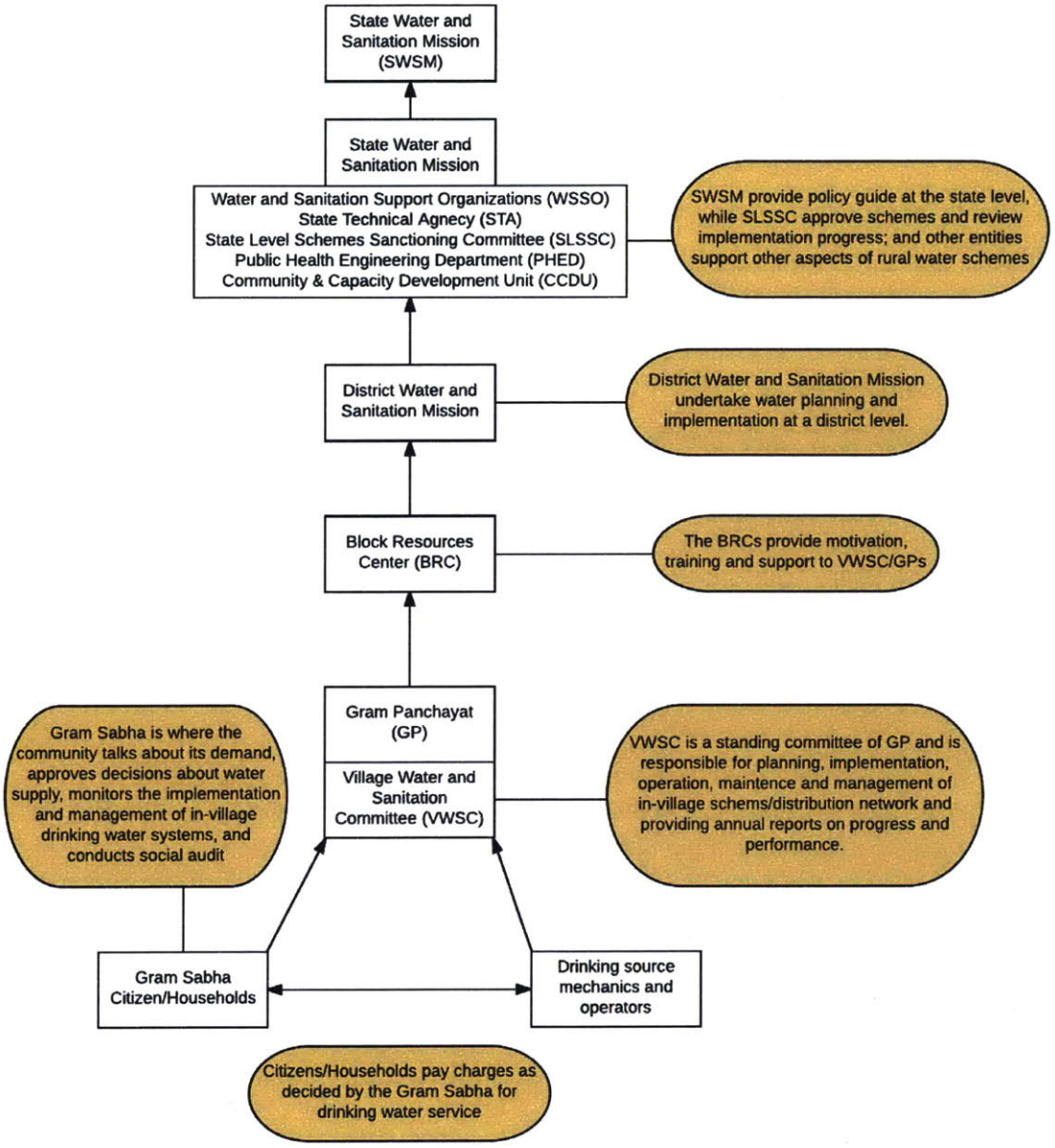


Figure 4-1: NRDWP Institutional Arrangement (adapted from the 2011-22 Strategic Plan by MDWS 2011)

For water quality monitoring, there is a specific set of entities that operates in conjunction with the NRDWP's overarching institutional arrangements - local labs which are specifically set up for the purpose of rural drinking water quality surveillance. Since 2006, the decentralized Water Quality Surveillance and Monitoring (WQMS) Programme proposed testing of all drinking water sources using sanitary inspection methods and field test kits (FTK), and samples suspicious of contamination shall be referred to District or Sub-district water quality labs for further investigation. In the Uniform Drinking Water Quality Monitoring Protocol, mandates on routine water quality data collection include the following (MDWS, 2013):

- Basic minimum parameters include total coliforms and E. coli;
- All sources should be tested once a year for chemical parameters and twice a year for bacteriological parameters;
- Baseline status should be established for all parameters once pre-monsoon and once post-monsoon, with GPS registration and groundwater depth recording;
- Discrete monitoring is required for calamities, especially monitoring of residual chlorine;
- District and sub-district level labs are expected to test around 3000 samples/year.

The 2013 Protocol also recognized that more labs were required because even with the ideal test load, only about 50% of the sources can be covered. Almost 300 new labs have been established since the then to fill the gap (IMIS website).

In the context of Gujarat, two key agencies are working towards the improvement of water quality in Gujarat: the Gujarat Jalseva Training Institute (GJTI), a unit of Gujarat Water Supply & Sewage Board (GWSSB), and the Water and Sanitation Management Organization (WASMO). GJTI works at a central level to carry out the WQMS scheme through the water testing labs, while WASMO provides support through FTK testing and educational campaigns at the village level. A general overview of the water quality monitoring setup in Gujarat is shown in Figure 4-3. WASMO works with villages for the collection of water quality data through FTK. Subdivisional labs, district labs and zonal labs conduct more rigorous water quality testing. Issues or concerns of water quality are reported up the hierarchy chain from Subdivisional labs to state labs.

GWSSB and WASMO operate independently but are very closely integrated. Annual trainings for both staff are conducted together, and district WASMO office and district water quality labs are often located very close to each other for timely communication.

The following stakeholders in Table 4-2 are interviewed at different administrative levels within GJTI and WASMO.

Table 4-2: Interviewees at water quality monitoring institutions in Gujarat

Interviewee	Admin Level	Position
Mr Shukla	State	Geologist and water quality expert in Gujarat Jalseva Training Institute Geologist by training
Mr Tripathi and team	State	Administrator of Water and Sanitation Management Organization (WASMO), and other team members from WASMO
Ms Trivedi	Zone[12]	In charge of Vadodara Zonal Water Quality Lab Microbiologist by training
Team	District	Bhavnagar District Water Quality Lab and WASMO team
Team	District	Dang District Water Quality Lab
Team	District	Narmada District WASMO team

[12] Non-standard administrative unit between state and district

According to interviews with WASMO, as of Jan 2016, Gujarat has already achieved 76% piped water supply coverage, almost twice the national average. Many water supply schemes are created to deliver water from one central source to multiple habitations via pipelines and delivery points, as shown in Figure 4-2. The original source may be from groundwater or surface water coupled with filter tanks. A lot of mini piped water supply schemes are also available where 20-25 households would share a piped system to retrieve groundwater from a central source.

Such a significant piped coverage progress came from the fact that Gujarat had a history of drought and salinity intrusion, which required extra effort to provide habitants with the appropriate quantity and quality of water. Despite the high piped source coverage, which is generally considered safer for consumption, waterborne outbreaks are still discovered monthly.

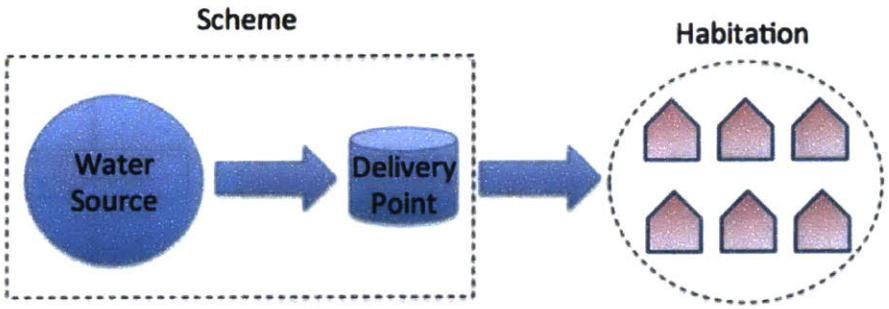


Figure 4-2: Demonstration of a water supply scheme (Novellino, 2015)

To ensure that water is free from microbial contamination, the VWSC is expected to chlorinate central water sources for water supply schemes on a daily basis. Chlorination may also be conducted through the pump operator who turns on the sump at the water source on a daily basis to pump and supply water to the delivery points. Sanitary surveys are also expected to be carried out at the central water source for water supply schemes, but not at the individual delivery points.

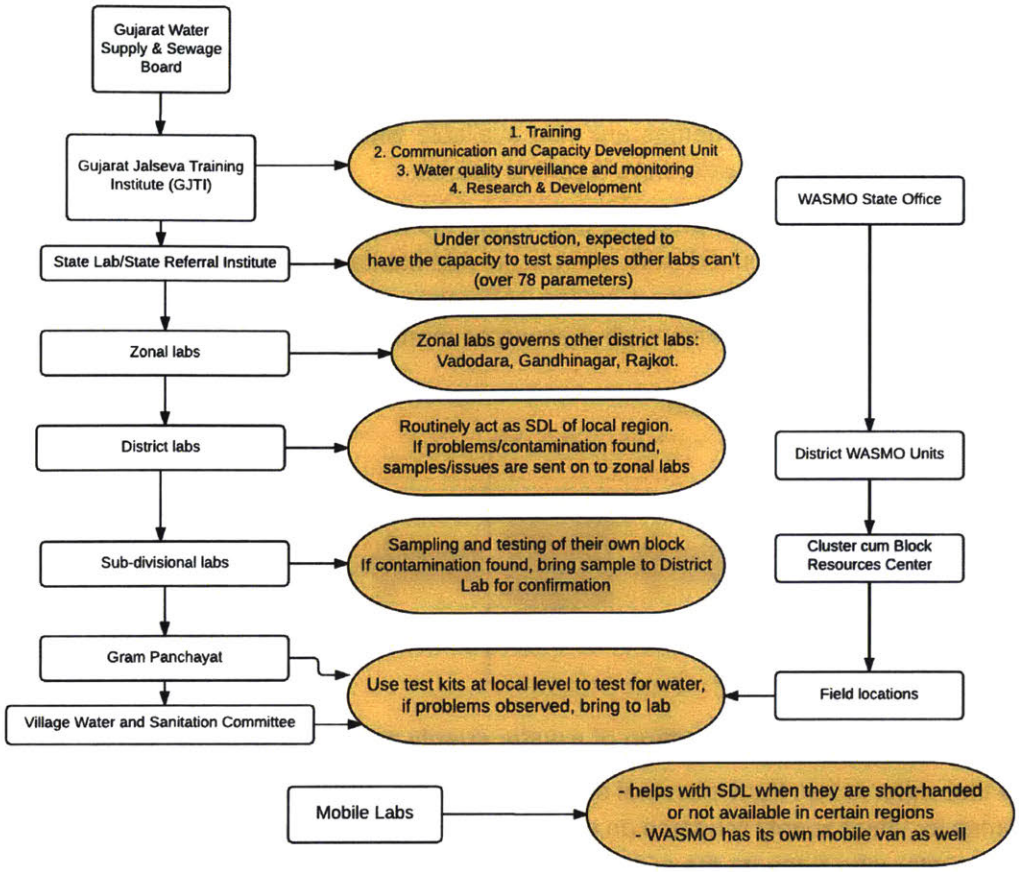


Figure 4-3: Water quality monitoring institutional arrangements in Gujarat

4.1.2 Data utilization

According to the WASMO team, water quality results are collected through both lab tests and FTKs. Groundwater sources are expected to be monitored by the labs two times a year: pre- and post-monsoon. Bacterial contamination is still a challenge in Gujarat. The contamination levels generally rise after monsoon, and once contamination is identified, WASMO works to raise local awareness and also inform local ASHA workers (Accredited Social Health Activist – governmentally instituted local health workers) to distribute chlorine tablets while GWSSB notifies local operators to add chlorine to the central water source. TCl powder and chlorine tablets are also distributed by the health department workers on a regular basis.

Sanitary surveys are also expected to be carried out pre- and post-monsoon by WASMO. However, as Ms Trivedi suggested, a lot of sanitary survey records are in hard copy and not yet entered online. Sanitary survey results may prompt actions based on the risk level identified through the process, but there is not a protocol on actions regarding these survey results. Results can also hint at risks of microbial contamination, but survey data and water quality have not been analyzed together.

According to WASMO, FTK testing are scheduled to be conducted every 15 days for key water sources in the village (which is the ideal case but not fully implemented), and positive samples will be taken to the lab for confirmation and instructions on remediation. However, many of these results are not reported through IMIS unless the sample is contaminated and has been passed on to local labs. The FTK result will then be uploaded to IMIS by the local lab.

As for further data utilization, Mr Tripathi mentioned that comprehensive analysis is carried out to review performances across regions and to set goals for future water supply planning. Overall, however, planning and decision procedures are very decentralized in Gujarat. Villages are expected to take the initiative to implement decisions based on data, while WASMO only plays an assistance role. Similarly, if bacteriological pollution is identified, WASMO district office will hold meetings in the villages to help the villages put remedial plans forward. While a bottom-up community-based water management approach is highly valuable, on some level it decreased the need to analyze data for decision-making at a central level because most decisions are made at the community level on a case-by-case basis.

4.1.3 Incentive for interagency connection

Limited interest has been shown by WASMO for interagency collaboration. Their primary focus is empowerment at the community level, and much of their educational campaigns also include sanitation related initiatives, so they consider themselves already part of the SBM program. There does not seem to be a strong need to evaluate the health impacts of their community capacity building efforts. However, while not actively in need of information from other agencies, the Bhavnagar WASMO district team mentioned that their water supply information should be useful for agency carrying out latrine construction, because they have encountered quite a number of defunct latrines due to lack of water availability in the vicinity.

Interest in interagency collaborations were shown by state or zonal level GJTI and water lab staff during the interviews. Mr Shukla, as a geologist by training, focused strongly on chemical contamination of water and its hydrogeological associations. He is interested in more irrigation and agricultural data that can help understand the causes of chemical contamination and how they may be dealt with. Ms Trivedi, as a microbiologist by training, is interested in more information from the health sector to understand impacts of microbial contamination. Specifically, she mentioned that her labs work closely with the health department only during outbreaks (raising a recent cholera incident as an example where the two departments worked together to identify the cause and prevent further spreading of the cases), but a more routine collaboration and data sharing process may enhance their outbreak management capacity. According to Ms Trivedi, GJTI and the water labs have planned to monitor the National database for Communicable Diseases and observe its likely connection with water quality, but this is still yet to apply. In addition, Ms Trivedi is also interested in gaining more sanitation information, especially during monsoon season when runoff from latrines may significantly increase biological contamination of water sources.

At district labs and below, there is limited interest in collaborating with other sectors. Their key focus was only on getting accurate water quality results and notifying the labs above when results are abnormal.

At the GP/village level, the understanding of microbial contamination is limited, even for members of the Panchayat leadership. Largely unaware of the connection between E. coli, latrine construction and diarrhea, the demand for inter-sector collaboration is low.

4.1.4 The Millennium Development Goal Implications

Despite some strong interest at the state-level testing agencies, likely stemming from scientific curiosity, the actual incentive to push for such interagency connections is lacking. To understand this better, we need to first understand the context of international drinking water and sanitation monitoring. In 1990, WHO and UNICEF combined efforts to form the "Joint Monitoring Programme for Water Supply and Sanitation" (JMP), which set the time-bound 1990-2015 MDG targets for drinking water and sanitation progress as show in Figure 4-4. However, as shown in Table 4-3, the performance of drinking water has been evaluated based only on increasing coverage of "improved sources", rather than the actual water quality measurements. As a consequence, there is a strong push on water scheme construction across India, and many of our interviewees state the piped coverage of Gujarat as a proud achievement but was uncertain when asked about water quality progress. Water quality collection is much more challenging and resource-intensive, and during the 1990 baseline year for the evaluation of MDG goals, very limited accurate water quality information was available globally. Hence the MDG progress decided not to target water quality but focused on the more tangible source categorization (Bartram *et al.*, 2014). This has been criticized in a range of studies, which claim that national safe water coverage level can be reduced up to 40% if microbial contamination is also incorporated (Godfrey *et al.*, 2011; Bain *et al.*, 2012). As India works towards goals under JMP, it is reasonable that the priority may not be on water quality targets, not to mention its further connections with sanitation, disease and outbreak. Direct water contamination issues, along with its causes and consequences, are left out of the JMP-MDG picture.

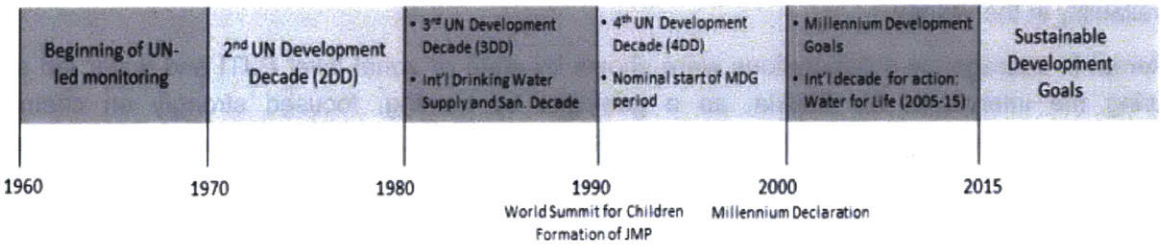


Figure 4-4: Timeline of international targets and actions related to drinking water and sanitation (Bartram et al., 2014)

Table 4-3: JMP categorization of drinking water and sanitation coverage (Bartram et al., 2014)

(1a) Drinking Water		
Drinking water core questions (1)		
What is the main source of drinking water for members of your household? Where is that water source located? How long does it take to go there, get water, and come back? Who usually goes to this source to collect the water for your household? Do you do anything to the water to make it safer to drink? (Introduced 2005) What do you usually do to make the water safer to drink? (Introduced 2005)		
MDG Categorisation of Households (2)	JMP Disaggregated Categorisation of Households	Underlying Questionnaire Responses
Not using an improved drinking water source	Collection of water from a surface water source	Surface water (river, dam, lake, pond, stream, canal, irrigation channel)
	"Other unimproved sources"	Unprotected dug well Unprotected spring Cart with small tank or drum Tanker truck (3) Bottled water where other water source is classified as unimproved (4)
Using an improved drinking water source	"Other improved sources"	Public tap or standpipe Tubewell or borehole Protected spring Rainwater collection Bottled water where other water source is classified as improved (4)
	Piped drinking water into dwelling, plot or yard	Piped water into dwelling, yard or plot

Even when water quality is taken into consideration as more and more data are being collected across India, the evaluation criteria are limited. Uniform Water Quality Monitoring Protocol requirements focus on lab capacity with target of 3000 samples per lab per year (MDWS, 2013). Ms Trivedi has repeatedly mentioned to us that the majority of labs under the management of her Vadodara Zonal Lab are reaching the 3000 target and covering all required sources. However, it is not within her responsibility to ensure that all contamination can be taken care of. She merely reports the issues to GWSSB for it to be further addressed there. The ultimate health performance resulting from the local water quality is not part of the responsibilities of the labs. In fact, it is not even part of the responsibility for GWSSB. Their corrective measures only target at removing contamination, and potential health effects are not considered under their jurisdiction unless an outbreak has already happened due to the contamination (in which case it will be directly reported to the health department and the responsibility is again transferred). Without a clear water quality performance target, contamination is likely to be dealt with on a case-by-case basis, rather than evaluated centrally for their implications and impacts. Initiatives to integrate health and sanitation data for a better understanding of water quality is at best an interest without a motivation for these agencies.

Water quality monitoring are being included in standard JMP household surveys on a trial basis post-2015, with the intention to ultimately modify the coverage status to account for water quality variables (Bartram *et al.*, 2014). As global focus shifts toward ensuring safe water in addition to just safe delivery

of water, there are positive prospects for a more integrated approach for analyzing water quality and its implications.

4.2 Sanitation monitoring institutions

4.2.1 General Information

According to the standards of JMP, sanitation evaluation is based on toilet facility details of communities and households, while hygiene evaluation is based on handwashing practices. In India, only sanitation-related data are collected by governmental agencies at a national level. Hence, this study focuses on the efforts of monitoring toilet construction and open defecation-free (ODF) statuses in rural India, carried out mainly under the SBM-G (Swachh Bharat Mission - Gramin[13]) initiative.

SBM-G was established to improve the general quality of life in rural India through promoting cleanliness, hygiene and eliminating open defecation, with the ultimate goal to achieve Swachh Bharat ("clean India") by Oct 2, 2019 – as a tribute to the 150th Birth Anniversary of Mahatma Gandhi. This is carried out through motivating communities to adopt sanitary facilities and sanitary practices through awareness raising campaigns, implementing appropriate technologies for ecological safe and sustainable latrine facilities, and developing community managed sanitation systems with scientific solid and liquid waste management systems (Ministry of Drinking Water and Sanitation Government of India, 2014).

The 1981 census revealed a rural sanitation coverage of only 1% in India (MDWS 2014). Following the International Decade for Drinking Water and Sanitation during 1981-90 (Figure 4-4), a demand-driven approach to sanitation have been introduced to India in 1999 - the "Total Sanitation Campaign", which focused strongly on Information, Education and Communication (IEC) and Human Resource Development (HRD) activities and increased the capacity of the communities to choose their own appropriate sanitary facilities based on their conditions (MDWS 2014). Financial incentives were awarded to Below Poverty Line (BPL) households for the construction of individual household latrines (IHHL). As challenges started to emerge during TSC, especially the large number of dysfunctional toilets that was built only to superficially meet requirements but never used or even connected to water systems, TSC further evolved into the "Nirmal Bharat Abhiyan" (NBA), launched in 2012 to also incorporate technology and community behavioral change requirements in sanitation progress. Incentives for latrine coverage was enhanced by provision of awards to best performing GPs. To further the efforts on sanitation, the Prime Minister of India launched the Swachh Bharat Mission on Oct 2, 2014 under the Ministry of Drinking Water and Sanitation with two focuses SBM (Gramin) and SBM (Urban), with the ambitious target to achieve an open defecation free, clean and sanitized India by 2019. The focus of this study is solely on SBM(G) – rural sanitation.

Unlike the water quality monitoring labs, there is not a separate governmental entity to monitor sanitation progress in Gujarat. Instead, it is taken on as a top priority initiative by existing rural administrative agencies, and carried out with the assistance of additional consultants such as specialists from UNICEF. The overall monitoring and implementation of the SBM(G) scheme is shown in Figure 4-5. A 2012-13 Baseline Survey was conducted by the Government of India based on

[13] Hindi word for "rural." There is also a separate SBM-U for urban sanitation.

national guidelines issued by MDWS, covering aspects related to toilet access, functionality, access to toilets in Aanganwadi Centers and schools, access to water supply for households and institutions, availability of human resources and village-level partners working in sanitation. Adapting from the Baseline Survey, key components identified now include IEC and other triggering activities for behavior change, construction of toilets, usage of toilets and creation of ODF communities. For the state of Gujarat, the Commissionerate of Rural Development is in charge of SBM(G), and the mission is implemented through existing administrative level governments at the district and block level, as shown in Figure 4-6 . There are SBM commissioners at the state level to track the overall progress, and additional SBM personnel are being recruited to work specifically on SBM. While the SBM(G) Guidelines state that a separate SBM entity formed at each district and block level, many of the posts are not yet filled and monitoring activities are still carried out through existing personnel such as the development officers at the district and block-level administrations, or the directors of rural development agencies. Additionally, three organizations – UNICEF, Tata Water Mission and World Bank are extending technical support to governmental efforts in Gujarat. Each of the three organizations takes care of 1/3 of the districts (11 out of 33) to assist in ODF achievement.

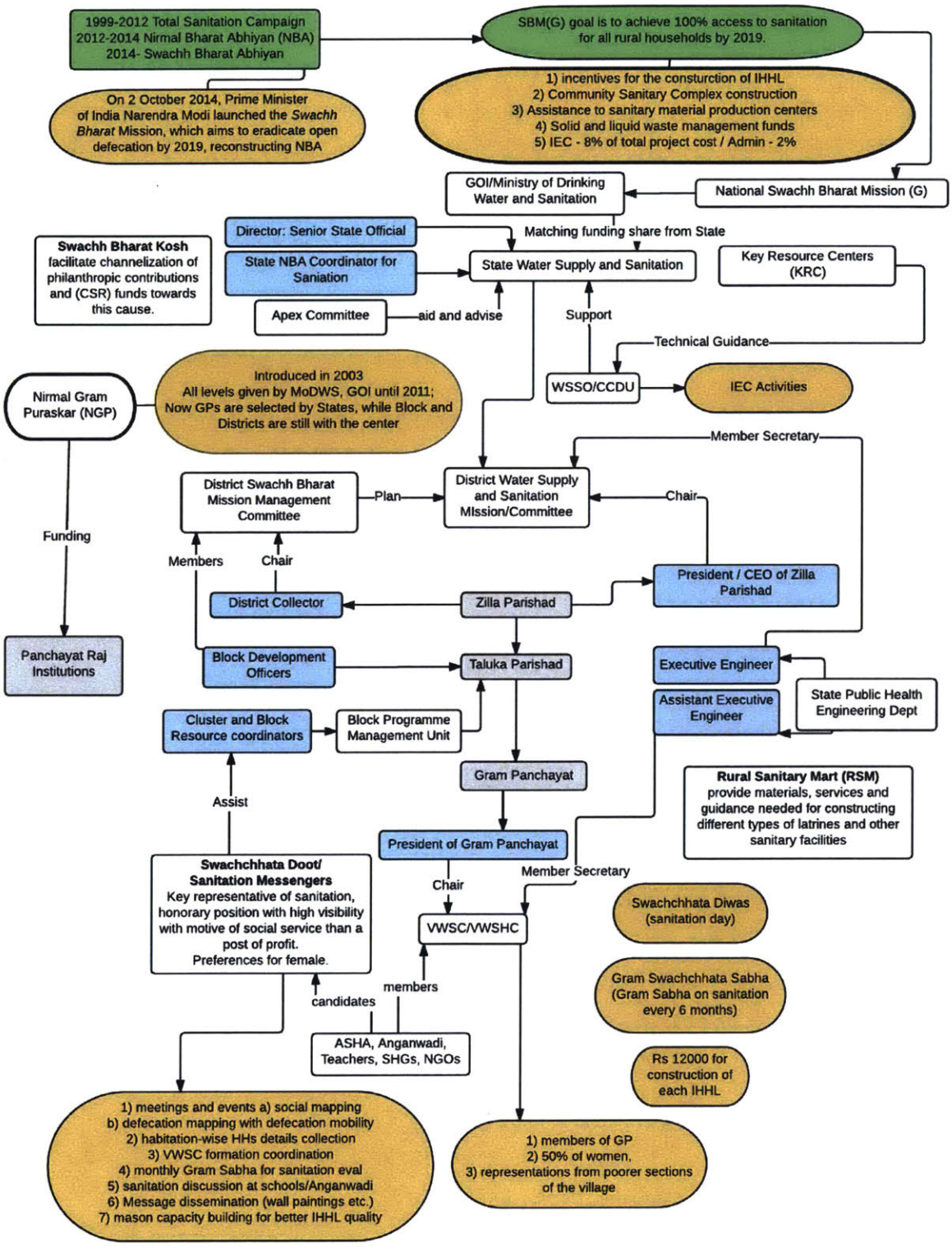


Figure 4-5: SBM(G) Institutional Arrangement

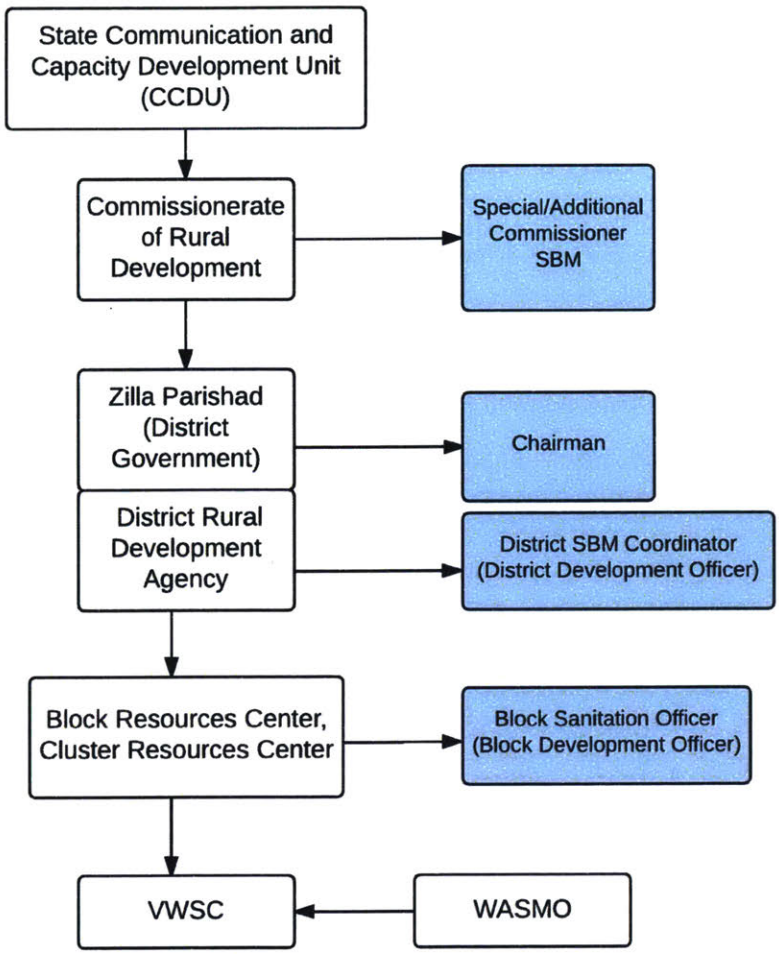


Figure 4-6: Sanitation monitoring institutional arrangements in Gujarat

The following stakeholders in Table 4-4 are interviewed at different administrative levels within the SBM program, as well as from consultant organizations supporting the program. Special focus was given to Narmada because it is the first district ready to declare ODF. UNICEF assists in the SBM implementation in Narmada and has piloted interactive games for community behavioral change, which was met with great success (Commissionerate of Rural Development, 2016).

Table 4-4: Interviewees at sanitation monitoring institutions in Gujarat

Interviewee	Admin Level	Position
Mr Astik	State	Assistant Administrative of SBM in Rural Development Agency
Mr Amit Wajpe	State	WaSH Consultant, Tata Water Mission
Mr Shyam Dave	State	WaSH Specialist, UNICEF Gandhinagar, India
Mr Aayush Oak	District	District Development Officer in Bhavnagar
Mr Singh	District	Director of District Rural Development Agency in Dang District
Mr Hetal Pathak	District	UNICEF WaSH Consultant for Narmada District
Mr Rifakat	District	Solid and Liquid Waste Management Consultant for Narmada District

Sarpanch of Fariyadka Village, Bhavnagar	GP	Gram Panchayat leader, took on SBM initiatives to construct household latrines

4.2.2 Data utilization

Data in the SBM database are widely used for planning purposes, especially considering that the entire SBM program is very clear on its target key performance indicator – 100% IHHL and ODF. According to Gujarat's SBM Annual Implementation Plan, by March 2016 there are still 1945549 rural households without toilets, of which 62% are planned to be covered during 2016-17, so that by the end of the year Gujarat will have at least 7 districts, more than 100 blocks and 8000 GPs that are ODF.

In the Baseline Survey, characteristics of each of the households are recorded. These characteristics facilitate decisions on the level of support provided for household IHHL construction. Incentives are only provided to households that are BPL, or identified as SCs/STs[14], small and marginal farmers, landless laborers with homestead, physically handicapped or women-headed households for APL (Above Poverty Line) households. Incentives up to Rs. 12,000 are provided for each unit of IHHL upon construction. Generally, GP would work with local households to construct toilets, and a form would be sent to sanitation mission at the block level when construction is completed. Once around 10-15 forms are collected at the block level, officers in charge would visit the households to confirm the construction and usage of toilets, input the data and funding can then be dispensed. The IHHL status is frequently reviewed at the district level, by the government as well as the SBM consultants. For example, the UNICEF SBM team gathers consultants from all 11 districts together every few months for a discussion on latrine construction progress, during which each consultant would report specifically on the progress of their district and gather feedback from all others.

There is a stringent process for ODF status verification. After the GP makes a ODF declaration (suggesting that all households, health centers, schools, Aganwadi Centers have latrine facilities), a cross-block verification by non-SBM functionary personnel of agencies or organizations in other blocks would be carried out within 3 months of self-declaration. Toilets construction, general cleanliness of the village, and practices of the villagers are all inspected in the process. Once inspected and confirmed, then the GP would be recorded as verified ODF in the database. After 6 months, another cross-district validation is done to 10% of the verified ODF GPs by other districts to ensure the sustainability of sanitation practices. This extra verification step is set in place considering that toilet usage is a relatively significant habit change for many rural villagers, and many past sanitation schemes have failed at sustainable ODF status due to villagers reverting to old defecation habits. On the other hand, many of our interviewees noted that while behavioral change is an essential part of SBM(G), the data collection on toilet usage is very limited and no information is collected on hygiene practices. While the cross-verification processes for ODF check on behavioral factors, many interviewees consider this insufficient and hope for more routine behavioral data collection. Aga Khan

[14] The Scheduled Castes (SCs), also known as the Dalit, and the Scheduled Tribes (STs) are two historically disadvantaged groups of people that are given express recognition in the Constitution of India (IMIS website).

office has already employed a mobile data collection system to record local hygiene and sanitation awareness and practices, but the system is yet to be applied at a large scale.

Apart from latrine construction and awareness building, solid and liquid waste management (SLWM) is also a key section under SBM(G). However, due to the priority to achieve 100% ODF, almost no data has yet been collected for SLWM across the state of Gujarat. According to the SLWM consultant of Narmada District, the only district that achieved 100% IHHL construction as of Aug 2016, progress on solid and liquid waste management is slowly starting to pick up after the district finished latrine construction progress. With 100% IHHL as the most prioritized target, other agenda items may have been pushed aside only until ODF can be declared. Considering how waste management from latrines may have critical impact on water sources, this prioritization may lead to long-term issues.

4.2.3 Incentive for interagency connection

At the state level, there is an interest in observing changes in diarrheal diseases as more ODF is being achieved across the state. However, from the administrative point of view, health data is not essential because only latrine data and ODF data are required for the 2019 clean India target. The effects of these targets are of less concern from an administrative point of view. On the other hand, Mr Dave, UNICEF state consultant addressed the issue from more technical points of view – stating that toilet construction is very much reliant on water availability and data on water would ensure that latrines can be established and used effectively. Mr Dave was also in favor of a new definition of ODF that also incorporated functional and safe water sources in the GP because this would strongly incentivize a definition of "clean" India that encapsulates all WaSH aspects. Because the overarching goal of UNICEF cuts across all WaSH-health aspects, Mr Dave was very much interested in a database that cuts across water, sanitation and health sector and suggested that the State Public Health Department should take on such an initiative to connect all data together geographically to understand the health effects of WaSH activities. However, despite his strong interest, he mentioned that a policy decision at the state level is required for initiatives to fall into place. While UNICEF may publish reports on cross-sector data analysis, their role in the end is still technical, which limits their capacity to set up new collaboration frameworks.

At the district level, the District Development Officers at Bhavnagar and Narmada (Officer was away during visit to Dang), who monitor SBM progress but also oversee the overarching progress in all three aspects of water, sanitation and health, are also interested in a more integrated data system. They want to have a comprehensive understanding of all aspects of the district for planning and decision-making. However, time and resources were cited as key constraints. There are large amounts of data for the district, but each District Development Officer only stays in the post for 2 years – so while they would benefit significantly from a collaborative system, there is insufficient time and energy for them to build such a system from scratch. Moreover, their jurisdiction is beyond just WaSH-health topics, so it is unrealistic to expect them to develop an integrated approach solely for WaSH systems within the 2 years of their residency.

In addition to the District Development Officers, directors of District Rural Development Agency (DRDA) also work on SBM initiatives. DRDA works more directly on SBM, in comparison to the overarching perspective of the District Development Officer. As the director of Narmada DRDA, Mr. Singh showed strong interest in verifying the health and safe water effects of Narmada's 100% ODF status. He

indicated that once their declaration of ODF are verified, Narmada would focus on health data collection for impact evaluation. This is confirmed by UNICEF consultant Ms Ranjan, who mentioned an ongoing proposal for new ODF+ status definition, which would include water quality status and health status to confirm the long-term sustainability of ODF, was set to be piloted in Narmada as it is the first one to achieve ODF. Hence, for district SBM coordinators, only when the ODF status is achieved for the 2019 target would more initiatives around WaSH-health impacts follow.

At the GP/village level, there is an interest in using water information to improve sanitation coverage, but mostly for educational purposes. By showing that open fecal matter may easily get into nearby open wells, villagers can become more aware of the consequences of their open defecation practices. However, even for villages that have achieved ODF, there is not a strong demand to understand the water quality and health implications, likely due to the fact that the connection is not well understood at this level.

4.3 Outbreak monitoring institutions

4.3.1 General Information

The Integrated Disease Surveillance Program (IDSP) was launched in 2004 with assistance from World Bank to strengthen the disease and outbreak surveillance system in India (IDSP website). It is a decentralized, state-based program intended to increase laboratory-based and IT-enabled surveillance for epidemic-prone diseases to monitor trends and to detect outbreaks during its early phases to help initiative effective response in a timely manner (IDSP website; Kumar *et al.*, 2014). IDSP is administered by National Center for Disease Control, and it is also a part of the National Rural Health Mission as of 2007-2008 (IDSP website).

As shown in Figure 4-7, under IDSP, data are collected on a weekly basis in three specific formats – "S" (suspected cases), "P" (presumptive cases) and "L" (laboratory-confirmed cases) forms, which are filled out by Health Workers, Medical Officers and Lab Technicians respectively.

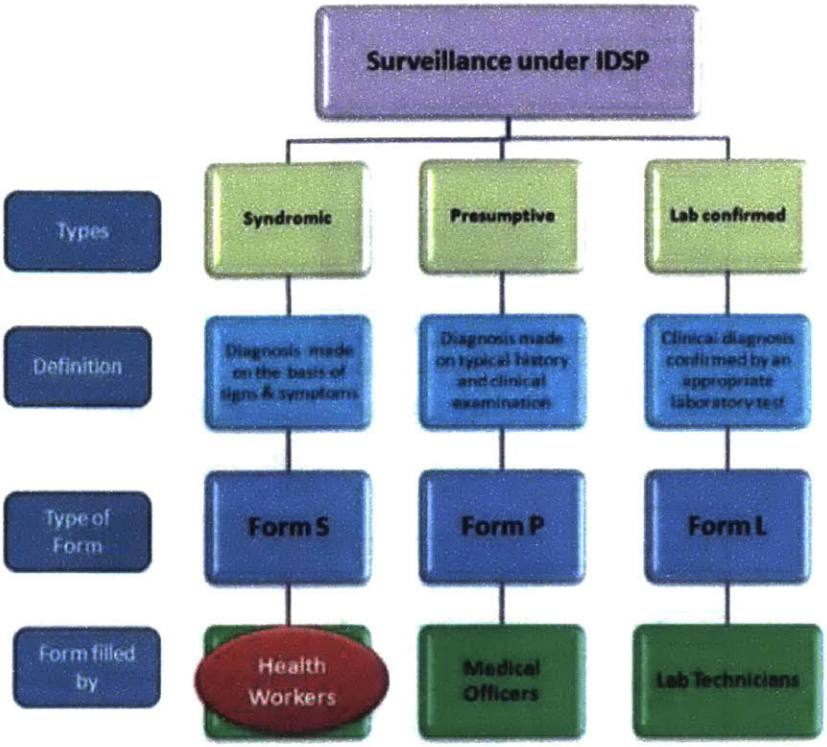


Figure 4-7: Type of Surveillance under IDSP (IDSP, 2015)

As shown in Figure 4-8, Health Workers are generally expected to collect weekly syndromic surveillance data on the various syndromes listed in IDSP through routine visits to the survey area, media reports or from key informants locally (IDSP, 2015). After collecting data for the S form and entering it into the register, copies are submitted to the Medical Officer at the PHC every Monday of the week. In case of unusual events, the Health Workers are expected to report immediately to the Medical Officer in addition to the routine Monday reporting, and take the steps needed to respond to severe cases. For example, if a diarrhea outbreak has been reported, a standard response protocol is outlined in Figure 4-9.

Afterwards, the Medical Officer is expected to analyze information from form S and send copies along with its own P form from outpatient records to the CHC/BPHC by Tuesday. The CHCs and BPHCs are then expected to review all S and P forms and send them along with its own L forms to DSU by Wednesday (IDSP, n.d. b). Units that have failed to reported will be checked by CHCs/BPHCs on Thursday, and all the weekly data entry would be expected to be done by Friday.

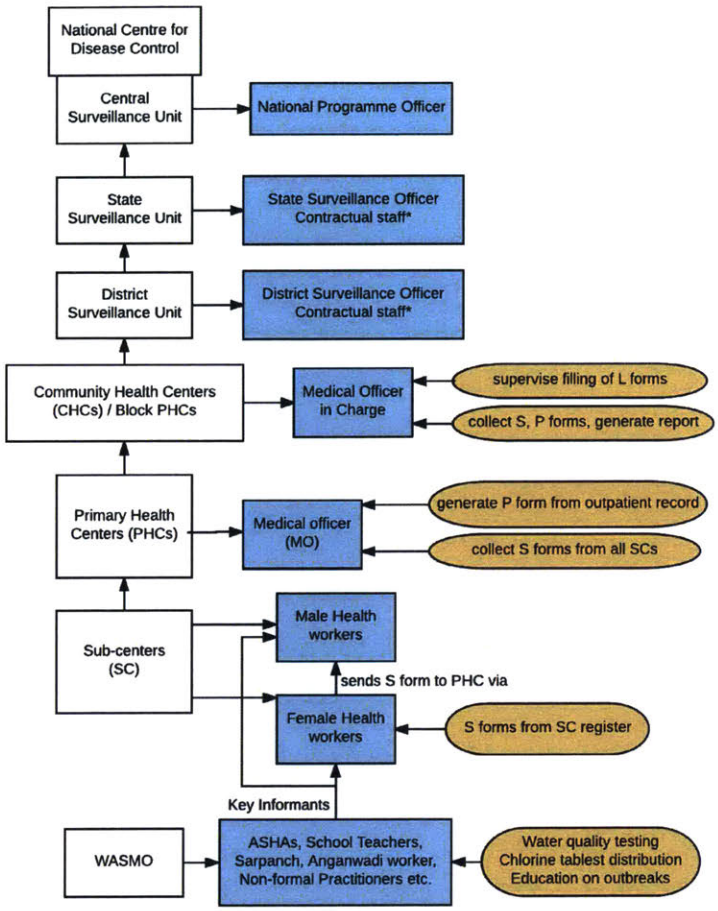


Figure 4-8: IDSP Institutional Arrangement (*key contractual staff include administrators, data managers, epidemiologists, microbiologists, entomologists etc.)

- Case Management
- Inform MO PHC
- Epidemiological investigation
  - Active search for all new cases in that area
  - Line listing of cases by name, age, sex
- Prevention of further cases/deaths
  - Provision of safe drinking water
    - Distribute chlorine tablets
    - Chlorinate all water sources in community
    - Inform VHNSC to provide properly treated or safe water
    - Orthotoludine testing of drinking water sources to check for residual chlorine level.
    - Collect water sample and send it to PHC for H_2S testing and to district labs for MPN count.
    - Check TCL stock.
  - IEC to promote food/personal hygiene & train local person about water chlorination
  - Distribution of ORS packets

Figure 4-9: Steps to be taken in the field if diarrhea outbreak is reported (IDSP, 2015)

The weekly data would generally show time series trend for outbreak-prone illness, and any rising numbers would trigger an investigation by the Medical Officers and Rapid Response Teams (RRT) to diagnose and control the outbreak (IDSP, n.d.). On average 30-40 outbreaks are reported weekly to the Central Surveillance Unit (IDSP website).

In addition, IDSP engages in a number of innovative initiatives in outbreak control, including establishing a Media Scanning and Verification Cell to detect early warning signals or unusual health events through media in 2008 and piloting a Referral Laboratory Network for diagnostic support involving Medical Colleges and other high-level public and private laboratories in 2009 (IDSP website).

The interest to engage the health sector came relatively late to this study, so there was not a diverse range of health-related personnel that were interviewed on the topic of outbreak management in Gujarat, so a Gujarat-specific institutional arrangement is not charted. However, based on the personnel interviewed as shown in Table 4-5, it generally follows the typical IDSP institutional arrangement. Outbreak monitoring practices and interagency collaboration interests can still be gathered and evaluated from these interviews.

Table 4-5: Interviewees at health monitoring institutions in Gujarat

Interviewee	Admin Level	Position
Dr. Dodhi	District	District Child Survival Officer, Department of Health and Family Welfare at Dang District In charge of WaSH in health facilities
Dr. Vegada	District	District Health Officer, Narmada District
Dr. Kashyap	District	Epidemic Medical Officer, Narmada District

According to Dr. Dodhi, apart from reporting syndromic surveillance results, Multi-Purpose Health Workers in Gujarat are working closely with the local communities on outbreak prevention, especially through monitoring water quality and distributing chlorine tablets. Under the monitor of these Health Workers, ASHA workers routinely administer chlorination to potable water at the household level, and the tablets/powder are provided by both the DRDA and the Public Health Department. During monsoon season, where outbreaks are more prominent, TCI powder or chlorine tablets are offered to the household for preventative measures. Education on the causes of water contamination and outbreaks during monsoon are also conducted.

In addition, ASHA workers are expected to test for microbial water quality with FTK at all central sources monthly, while the Health Workers randomly collect water samples to verify the testing results of ASHA workers. All results from the Health Workers and ASHA workers are consolidated and sent to local PHCs. Results are expected to eventually reach the district surveillance units and uploaded in IDSP. WASMO conducts all water-related training to ASHA workers, and the agency also has a copy of all the available FTK water quality test results from ASHAs.

4.3.2 Data utilization

As mentioned in the previous sections, reporting units in IDSP are generally from the following categories as shown in Table 4-6. Depending on the size of each state and district, there is generally a fixed number of reporting units that cover the population. As of March 2012, Gujarat has 7274 Sub-

centers, 1158 Public Health Centers and 318 Community Health Centers (Department of Health & Family Welfare, 2014).

The key data that each S/P/L form collect through the various reporting units are listed out in Table 4-7. In comparison to the water and sanitation institutions which have specific key performance indicators to target, the outbreak monitoring institutions are instead looking for trends and abnormality. They are much more reliant on data analysis and effective algorithms of pattern recognition, and a robust data system is essential.

Table 4-6: Reporting Units in IDSP (IDSP, n.d. b)

	Unit	Population covered	Location	Focal point for IDSP
1.	Sub Centre	5000	Village	Multi Purpose Health Worker (Male or Female)
2.	Primary Health Centre (One MO PHC)	30 000	Big Village	Medical Officer
3.	Community Health Centre (CHC)/ Block PHC	100 000	Semi Urban Area	Medical Officer in charge
4.	Dispensary/Hospital /Private Hospital	No Covered Population as they provide only facility data	Urban Area	Medical Officer i/c or Medical Superintendent

Table 4-7: Reporting Formats from the Reporting Units (IDSP, n.d. b)

	Unit	Format	Major Data looked for
1.	Sub Centre	S	a. Fever b. Diarrhea c. Fever, Cough and Cold d. Jaundice
2.	Primary Health Centre (One MO PHC)	P	a. Malaria b. Dengue c. Chikungunya d. Measles e. Acute Encephalitis Syndrome f. Pyrexia of Unknown Origin g. Meningococcal Meningitis h. Leptospirosis i. Diphtheria j. Typhoid k. Influenza l. Cholera m. Bacillary Dysentery n. Hepatitis o. Acute Flaccid Paralysis p. (Unusual Syndromes)
3.	Community Health Centre (CHC)/ Block PHC	P and L	*P form classification as above* *L form* a. Lab confirmed Malaria b. Lab confirmed Tuberculosis
4.	Dispensary/Hospital /Private Hospital	P and L	As above

The goal during an outbreak is on identifying key cases and proposing the critical time and place for intervention. The goals of analysis in IDSP before the occurrence of the outbreak focus on (IDSP, no date c):

- identifying outbreaks or potential outbreaks through recognition of abnormal trends

- identifying high-risk population groups for targeted interventions
- predicting changes of disease rates over time
- comparing regional differences for improving surveillance and increasing collaboration.

To ensure that these goals can be met effectively, trend analysis are required in weekly to monthly summaries for outbreak control decisions (Figure 4-10), timely report completion rates are given for all reporting units for capacity development decisions, and linkages between different S/P/L form results are reviewed for surveillance design improvement decisions (IDSP, no date c). Analysis is expected to be carried out at all administrative levels, including Health Workers at Sub-centers, health inspectors at PHCs, Medical officer at PHC/CHC and the Data Manager at the District level, while the degree of analysis depend on the capacity of the personnel (IDSP, no date c). IDSP has provided many supportive systems to ensure a robust foundation for these analysis, including establishing a country-wide Satellite Broadband Hybrid Network connecting all the states, districts, major medical colleges and the Central/State/District Surveillance Unit to improve issues with insufficient bandwidth and ensure speedy data transmission, for which Gujarat was one of the pilot states to be fully covered (IDSP website). Multiple detailed data reporting, analysis and management manuals are issued to support Data Operators and Health Workers in their routine. They are not in place for sanitation and water quality monitoring.

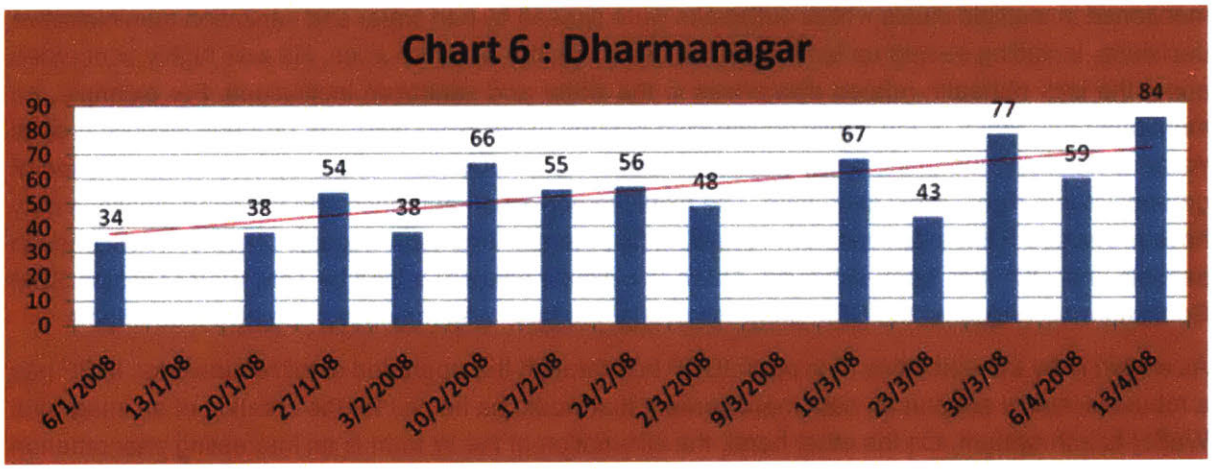


Figure 4-10: Example of a rising trend of presumptive diarrhea cases (form "P") per lakh population in Dharmanager that could qualify for a preliminary outbreak investigation

Nevertheless, the system still has limitations. Studies suggest that Health Workers are not all aware of the trigger events for outbreak surveillance, and only less than 40% of reports are sent on time by Sub-centers, indicating the surveillance system may be much less alert than it was prescribed to be (Kumar *et al.*, 2014). This has been verified through our interview with Dr. Dodhi, who mentioned that many community health workers and volunteers has only around 10th-grade education and do not have sufficient training. IDSP has recently created a "Revised Training Manual for Health Worker" in July 2015, which may be a response to these concerns. We were not able to interview Health Workers or Medical Officers at Sub-centers or PHCs due to schedule constraints to verify the implement of the latest IDSP framework in Gujarat.

4.3.3 Incentive for interagency connection

As Mr. Dodhi mentioned, water quality FTK data and chlorination data are available to the ASHAs and the local PHCs. However, much of the information is not uploaded through IDSP. While a "W" form used to exist in IDSP for water quality reporting alongside S/P/L forms in the past, it has been eliminated from IDSP. The defunct W form used to require water quality testing at sources supplying water to a large population, and the frequency of testing and reporting was mandated by the Medical Officer at the PHC. While the form is no longer available, this practice is still carried out and water data are still available for referencing in the district-level disease and outbreak reports. Since these data are also reported to WASMO, most of it may have also been uploaded through IMIS. While water quality data has been used as reference information for disease surveillance, there are no protocols on decision-making based on the water data. The ASHAs also work to advocate SBM activities alongside WASMO, but no sanitation data are collected or used for disease surveillance purposes.

At the district level, there is a clear demand for more information on sanitation and water quality, because they are critically connected to waterborne disease prevention as well as outbreak etiology identification efforts conducted by the District Surveillance Unit. Mr. Dodhi was well-aware of the benefit of an interconnected system and how it can help identify environmental indicators of diseases. The current data does not allow for preventative problem-solving. Similarly, Mr. Kashyap has mentioned in multiple cases where outbreaks were caused by bad water and sanitation administrative decisions, including setting up latrines very close to central water sources. He was highly concerned about the lack of health-related awareness in the water and sanitation institutions. For example, he mentioned that many of the newly constructed latrines under SBM were not properly disposing of its waste (potentially due to the delay of SLWM initiative until ODF has achieved), and contamination of groundwater is highly likely to happen. The highly target-driven approach and the short timelines may have prevented SBM initiatives from reviewing its potential long-term impacts on other WaSH-health aspects until ODF has been achieved. There is a similar concern raised by Tata Water Mission SBM Consultant Mr Wajpe on the SBM progress that is "too fast," and may have negative side-effects.

As shown in by the extensive IT support IDSP has through the upgraded satellite networks, IDSP has a robust technical support on data management that would be helpful for the creation of an integrated WaSH-health system. On the other hand, the elimination of the W form is an interesting phenomenon to further explore. The elimination happened around the establishment of NRDWP and IMIS, as well as the a restricted funding extension from World Bank that was only available for nine States (domestic funds were used for the rest of the States). Both the establishment of a separate institution and the change in funding may have resulted in an administrative decision to drop the water-related responsibilities. Despite the critical resources that IDSP may be able to offer, a more thorough understanding of this elimination decision can help identify potential barriers for a reversion to the original integrated approach.

4.4 Inter-agency collaboration evaluation

After reviewing the structure for all three sectors, this section gives a combined view of their current stance on interagency collaboration based on the evaluative criteria listed in Chapter 3.

4.4.1 Incentive

According to the analysis in the previous sections, the following two types of incentives are charted in Figure 4-11:

- Administrative incentive - the demand to have this approach for administrative decision support or necessary administrative processes
- Technical/scientific incentive - the demand to have this approach because of awareness of an integrated WaSH system and interest its scientific implications

The size of circle is also a general indication of the comparative amount of resources of each entity may have to initiative or support WaSH collaborations. As we can see, the water sector generally has technical interests, but without policy requirements dictating either the prevention of contamination or evaluation of contamination consequences, their administrative motivation is highly limited. For the sanitation sector, there is a large division depending on whether the entity achieved ODF or not. ODF districts are underway to pilot the new ODF+ plan, motivating them to understand the water quality and health effects of ODF. Thus, they are at the top right corner of the comparison chart. Non-ODF districts are still focused on toilet construction and community awareness campaigns, with limited interest in considering other topics. For the health sector, there is moderate-high motivation from both administrative and technical sides, especially considering water quality data are already routinely collected by health workers. Additionally, the surveillance units share a rigorous data management system with high technical capacity that can support potential WaSH-health integration. However, IDSP's decision to eliminate water quality data from its database suggest a barrier to integration that is worth further exploration.

Overall, the incentive for collaborative effort is the strongest with the SBM team (DRDA and consultants) of ODF-districts, and moderately strong for the disease surveillance units and other SBM consultants. Key barriers are the lack of administrative incentives for the water quality monitoring institutions and the lack of general collaborative motivation for non-ODF districts.

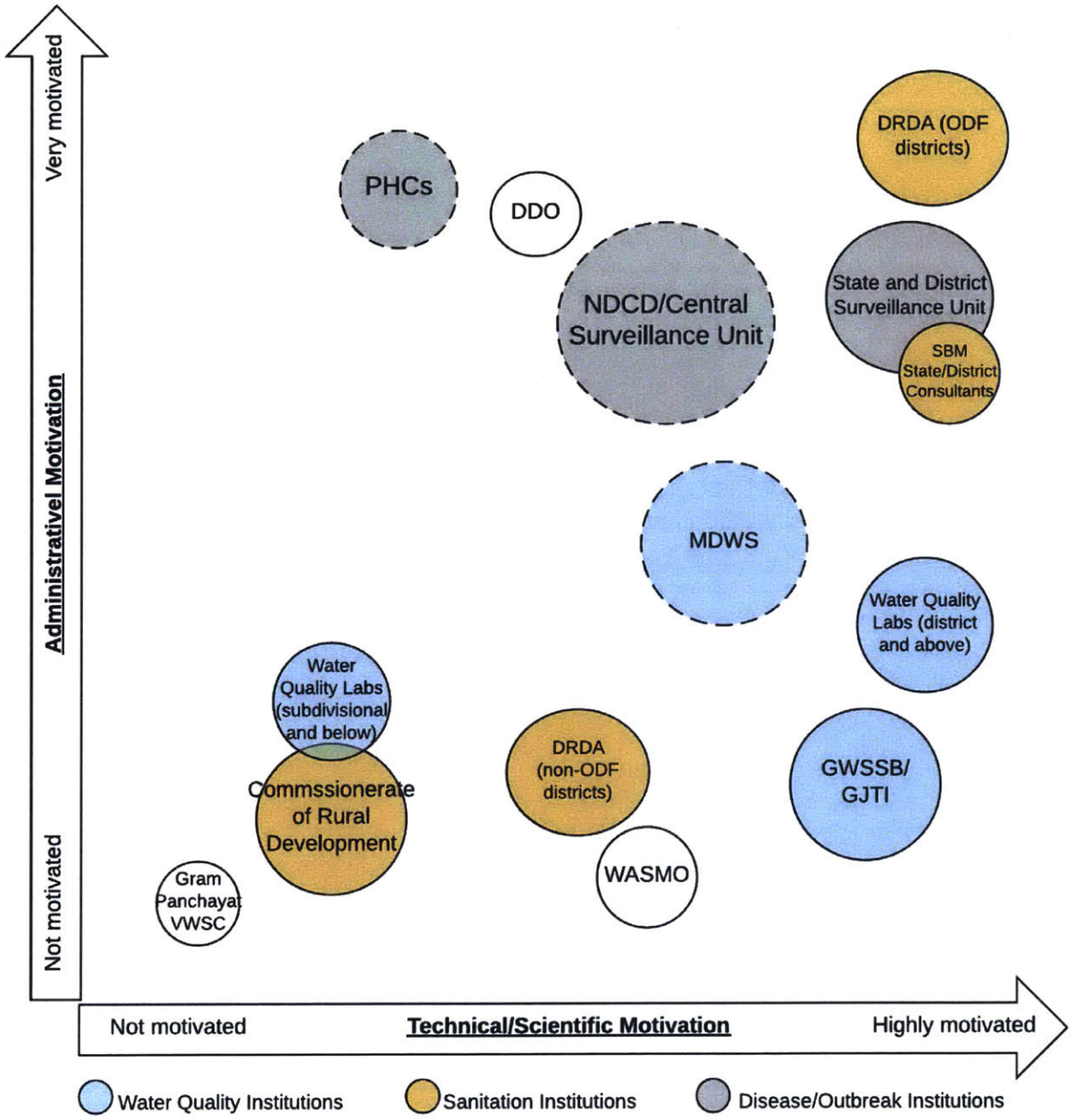


Figure 4-11: Comparison of the Incentvie for an integrated approach among different agencies in the water, sanitation and health sectors. Overlapping circles indicate similar positions. Dashed lines indicate agencies that were not interviewed, and the position is derived from literature and other interviews.

4.4.2 Existing connections

Monitoring of water, sanitation and diseases operates independently under different jurisdiction in Gujarat. The information from separate sectors are consolidated for nodal entities such as the District

or Block Development Officers, but considering that WaSH-health is only a small section of their overall work, this "connection" is unlikely to result in an integrated system. There are a few other lines of connection within the WaSH sector that are well worth exploring.

The community-level initiative for all three sectors are support a common agency – WASMO. They train ASHAs on water quality testing, they collect FTK results and deliver contaminated samples to local water quality labs, and they also conduct SBM educational campaigns with local communities. Even though WASMO focuses on advocacy and community empowerment and does not see an administrative need for inter-agency collaboration, the organization may become a critical alliance that helps bring all three sectors together.

ASHAs are delivering water quality data to both WASMO and PHCs. District Surveillance Units are consolidating water quality into their reports. Despite that the data is not used effectively by the health sector and that IDSP eliminated water quality entries, the availability of water quality data within the health institution is a great advantage. In comparison to other barriers, this path may have the least resistance and become a great entry point for further collaboration.

Lastly, all databases are managed by NIC. The three databases are very similar in their structure and data uploading mechanisms. The frequent loading errors that IMIS and SBM websites frequently encounter suggest that they may not benefit from an enhanced bandwidth and improved data transmission networks as IDSP does. Nevertheless, a shared data management system still decreases the barrier for future data collaboration.

4.4.3 Trust

Since the agencies have not yet worked closely on a routine basis, there is not a clear manifestation of trust issues. However, there are frequent hints at lack of confidence between technical personnel and administrative personnel. For example, consultants have shown dissatisfaction at governmental agency's understanding of WaSH-health issues and their lack of initiative to push forward SLWM and health evaluations. Medical officers have also expressed discontent with the GPs that have constructed toilets close to water sources, as well as with the development officers that approved of such structure and failed to identify the evident risks. The interconnection of WaSH-health is obvious to anyone with a technical background, and the failure of administrators to take this into consideration has caused frustration, and eventually a lack of confidence on their capability.

This lack of confidence would likely become a barrier when administers and technical personnel need to work closely together to create the integrated framework.

4.4.4 Regulation

As analyzed in Section 4.2, regulation – specifically the target definition for sanitation and water monitoring, is a critical barrier limiting the administrative motivation for interagency collaboration. These targets generally stemmed from WHO/UNICEF JMP definitions of clean water and sanitation, focusing solely on improved water supply and toilet construction. These action-oriented targets are much more tangible than a performance target such as a certain percentage outbreak decrease. However, setting them as the ultimate target makes people forget that they were originally only meant to be means to an end – the end of improved health status. Without this extra layer of intention, the

collaboration motivation for water and sanitation sector is very limited. Consequently, the ODF+ initiative mentioned by UNICEF consultants would be a very effective extension of the MDG. If it actually becomes a governmental initiative, the barriers to collaboration would diminish significantly.

As for disease surveillance regulations, concrete action plans are only available for suspected or confirmed cases of outbreak. Action plans for decreasing overall waterborne disease are vague and limited. For general disease reporting, the surveillance unit, much like the water quality labs, are evaluated by their capacity to survey and report results of the surveillance, rather than improvements in the results themselves. Hence, the administrative motivation for an integrated WaSH-health system is limited to outbreak investigation processes.

4.4.5 Summary

In conclusion, the key barrier to interagency collaboration include the following:

- A general lack of administrative motivation to collaborate due to lack of regulatory enforcement. Regulation mandates effective data collection as targets by water and disease sectors, and ODF as targets by sanitation sectors. The focus of these targets narrowed their scopes and limited demand for interagency collaboration.
- A lack of existing direct connections channels between GJTI, Rural Development Commissionaire and the Surveillance Units.
- A general mistrust between technical consultants/staff and administrative staff.

The pathways to overcome these barriers to establish interagency outbreak control system are suggested as the following:

- Consider districts with ODF+ targets a good entry point.
- Consider actions related to the existing water quality data within the disease surveillance units a good entry point.
- Effectively utilize the nodal agencies including WASMO and NIC to initiative a collaboration process.

5 DATA SUMMARY AND INTEGRATION ASSESSMENT

Chapter 5 examines the water quality, sanitation and disease databases to evaluate the data characteristics and data utility, concluding on pathways an integrated database. Section 5.1 – 5.3 reviews each of the three databases separately and evaluates all variables relevant to outbreak control within the database. An integrated database is created in Section 5.4, and its limitations and potentials are assessed.

DATA SUMMARY AND INTEGRATION ASSESSMENT

For each of the water, sanitation and health monitoring institutions, a separate database exists to host the data at a central level. As described previously in Chapter 1, these databases are referred to as IMIS (Integrated Management Information System) for drinking water information, SBM (Swachh Bharat Mission Management Information System, hereafter referred to as SBM) for sanitation information, IDSP (Integrated Disease Surveillance Programme database) for diseases and outbreak information.

All the data across the sectors are available at certain administrative levels, so the all the tables would be connected to the rural administrative hierarchy. This hierarchical structure is set up as a database relational structure as shown in Figure 5-1, which would be the basis for creating database schema across all the tables.

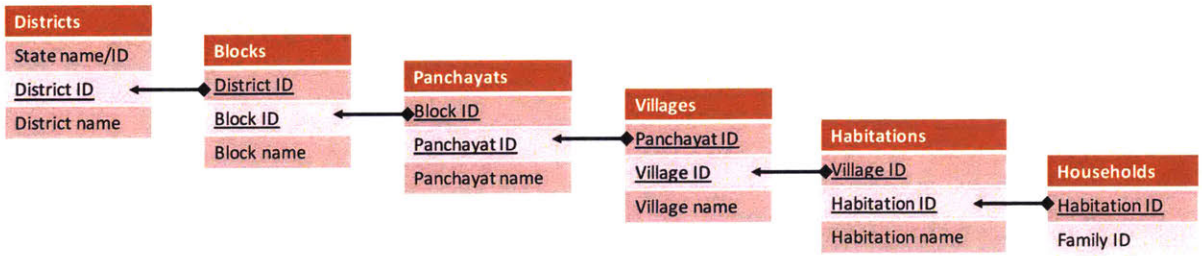


Figure 5-1: Basis for database schema: India rural administrative levels. Arrows connecting the tables indicate one-to-many relationships between the sets, where the pointed arrow end corresponds to the "one" and the diamond arrow end corresponds to the "many" (e.g. under the same district ID there will be many block-level tables). The same representation would be used for all following schema.

SBM database was only constructed on Oct 2014, so apart from the baseline year data, SBM data only date back to 2015-2016. To offer a valid comparison across the tables over the same time span, we have limited our evaluation to 2015-16.

While we attempted to analyze data integration for the entire state of Gujarat, there was data accessibility and availability issues for some of the key tables of interest. For the key tables related to outbreak control processes, the availability and accessibility across different districts for IMIS, SBM and IDSP data as shown in Table 5-1 (IMIS) and Table 5-2 (SBM and IDSP)[15]. The 0 entries are marked in grey, while inaccessible data sheets were marked in light red. For IDSP outbreak numbers, districts with multiple cases of outbreak are also prioritized so that WaSH data can be better analyzed with outbreak data.

[15] Values are aggregation results from all available entries, apart from values for water supply scheme sources, delivery point sources and FTK. Due to the challenge of scraping all possible entries for these datasets, numbers for them are copied from the IMIS aggregated values online, which may not accurately reflect the actual number of total entries.

Based on the criteria above, we have selected Kheda, Navsari and Surat as case examples (see rows highlighted in green in the comparison tables), where all dataset entries are available and more than 1 outbreak cases are reported. For these districts, all datasets from IMIS, SBM and IDSP outlined in Table 5-1 and Table 5-2 are described and evaluated in details in the following sections.

To be consistent with database terminology, datasets with a collection of variables are hereafter referred to as database "tables" with a collection of "columns."

Table 5-1: Data entry comparison across the different districts (IMIS). Zero entries for an entire dataset are marked in grey, and inaccessible datasets are marked in light red.

IMIS	Habitation Details				Water sources			Sanitary Survey	FTK test		Lab test	Training			
15-16 entry records	Blocks	Panchayat	Village	Habitation	Water supply Scheme	Delivery points	Public & private sources	Sanitary surveyed sources	Tested sources (C25)	Tested sources (E20)	Lab tested sources	Training sessions	District Officers	Block Officers	GRWs
AHMADABAD	9	472	483	640	3119	3831	6347	348	1483	1015	7380	516	4	0	556
AMRELI	11	607	611	641	1126	2209	12190	0	2651	1164	7512	459	3	184	1945
ANAND	8	341	350	946	3452	2677	10187	1044	3119	1390	4745	368	1	30	770
ARAVALLI	6	300	685	1351	2608	3278	0	0	2304	1132	4203	60	1	0	63
BANAS KANTHA	14	775	1242	1727	4770	6098	6235	0	1575	1462	5793	259	14	0	226
BHARUCH	9	544	658	885	1747	2264	9902	486	488	339	6476	179	2	67	783
BHAVNAGAR	10	678	679	687	3818	4723	22877	468	994	573	6150	366	10	0	2372
BOTAD	4	187	187	190	1213	1383	11798	0	747	358	6612	361	4	220	2956
CHHOTAUDEPUR	6	323	872	1154	3205	3077	14836	0	6616	2898	6487	33	1	0	201
DANG	3	68	303	318	2117	2523	8276	2958	6307	2357	4478	118	1	324	2169
DEVBHOOMI DWARKA	4	238	238	267	1807	1821	6502	693	1611	663	3941	409	1	0	1075
DOHAD	8	450	692	3140	3328	4299	31816	0	7989	3667	10096	22	2	0	25
GANDHINAGAR	4	299	309	496	2126	2500	3325	0	328	287	5829	216	33	0	693
GIR SOMNATH	6	338	373	374	1452	2146	6294	891	2158	747	6177	250	2	0	869
JAMNAGAR	6	409	409	481	2751	2840	8171	1190	3053	1491	4913	429	10	0	1289
JUNAGADH	9	524	524	524	2128	2892	7827	552	2584	1504	3415	268	2	791	1279
KACHCHH	10	739	892	1086	2611	5307	19426	1053	3469	1338	9801	55	13	231	2743
KHEDA	10	476	519	1708	2637	3075	15003	726	3523	1794	7174	743	2	0	7177
MAHESANA	11	580	621	873	3314	4109	6445	700	3311	1240	3323	4	3	0	36
MAHISAGAR	6	290	700	1503	2752	2914	18808	0	5633	2671	9203	139	3	0	396
MORBI	5	324	337	363	1509	1849	3301	0	714	342	3248	49	3	0	308
NARMADA	4	215	542	717	2079	2326	8686	508	5665	2273	5709	47	3	0	377
NAVSARI	6	368	394	2085	6220	5287	25848	199	13786	8955	12650	496	7	176	1366
PANCH MAHALS	7	425	600	1373	3118	4234	18617	0	1840	927	4720	140	2	52	1820
PATAN	9	476	516	646	1954	3892	8050	0	1626	738	4765	130	1	150	598
PORBANDAR	3	178	179	179	425	658	3789	466	827	349	2725	317	2	0	960
RAJKOT	11	575	577	583	3188	3735	13139	1279	2549	900	11442	75	9	0	264
SABAR KANTHA	8	401	675	1080	2282	3193	954	0	1950	1058	4033	521	2	288	5434
SURAT	9	569	757	1652	3606	4607	18673	71	9167	4945	8005	727	10	0	1258
SURENDRANAGAR	9	541	572	594	3298	2219	11810	0	1544	692	6521	2783	5	0	3990
TAPI	5	259	421	1505	1688	2137	15494	5769	5645	3879	7347	35	2	0	307
VADODARA	8	521	651	982	3228	3713	10480	37	2605	1177	4500	18	26	0	137
VALSAD	6	353	478	4094	7910	8119	34273	46	7640	3790	10499	49	54	64	559

*

Table 5-2: Data entry comparison across the different districts (cont., with SBM and IDSP)

15-16 entry records			SBM: Household data						SBM: ODF declarations			IDSP
IMIS name	SBM name	IDSP name	Block	Panchayat	Village	Habitation	Household with details	Total Surveyed Households	Block	Panchayat	Village	Outbreaks
AHMADABAD	AHMEDABAD		10	514	520	592	189204	245028	9	473	488	1
AMRELI			11	612	621	637	202325	229350	11	605	614	0
ANAND			8	347	351	470	332381	330422	8	350	362	7
ARAVALLI		Arvalli							6	290	671	2
BANAS KANTHA		Banaskantha	12	773	1148	1248	236354	475191	14	777	1242	2
BHARUCH			8	541	642	704	207736	199178	9	541	669	1
BHAVNAGAR			11	766	770	771	297676	299245	10	649	657	0
BOTAD									4	174	177	0
CHHOTAUDEPUR		Chhota Udepur							6	320	868	1
DANG	DANGS		1	68	272	273	23330	50339	3	69	307	0
DEVBHOOMI DWARKA									4	235	235	0
DOHAD	DAHOD		7	471	686	2316	248686	296615	8	476	718	0
GANDHINAGAR		Gandhi Nagar/ Gandhinagar	4	284	291	334	114684	183385	4	299	318	7
GIR SOMNATH									6	334	334	1
JAMNAGAR			10	659	661	685	124545	216059	6	413	417	1
JUNAGADH			14	808	808	809	173142	343783	9	490	493	0
KACHCHH			10	544	666	729	81672	296582	10	580	754	0
KHEDA			10	532	580	1305	269448	272890	10	468	509	8
MAHESANA	MEHSANA	Mahesana/Mehsana	9	581	611	698	317859	320806	10	597	656	2
MAHISAGAR									6	290	701	0
MORBI									5	336	349	1
NARMADA			4	217	473	510	100104	106699	5	217	544	1
NAVSARI		Navasari/Navsari	5	365	392	1889	199669	209774	6	365	392	4
PANCH MAHALS			11	640	1062	1647	221034	420878	7	449	628	0
PATAN			7	455	482	512	171299	230622	9	454	498	1
PORBANDAR			3	151	152	152	57398	80390	3	149	150	0
RAJKOT			14	840	845	848	334830	348001	11	586	589	1
SABAR KANTHA		Sabarkantha	13	698	1233	1675	320241	439258	8	412	692	5
SURAT			9	560	676	905	266994	270632	9	562	686	8
SURENDRANAGAR			10	610	638	655	197620	290254	10	540	571	0
TAPI			5	273	371	586	56603	175244	7	281	434	0
VADODARA			12	847	1430	1626	363377	468212	8	533	665	1
VALSAD			5	334	429	2399	103689	230342	6	346	467	1

5.1 Water quality database

5.1.1 Database background

The Integrated Management Information System (IMIS) (Figure 5-2) is launched by Ministry of Drinking Water and Sanitation (MDWS) to support the National Rural Drinking Water Program (NRDWP) by providing up-to-date information on drinking water status of rural habitations of India (IMIS website). The National Informatics Center (NIC) developed the database and assists in its management and update. Reports provided by IMIS are in a drill-down format to view rural water related information in the order of administrative hierarchies - National, State, District, Block, Gram Panchayat (GP), Village and Habitation (as defined in Table 4-1).

IMIS started the first habitation survey of rural drinking water status in 2003, and employed verification of data through third parties since 2006 (Novellino, 2015). However, it was only after NRDWP was launched on April 1, 2009 that national standardization and guidelines are established for a renovated version of IMIS (Novellino, 2015). Data from 2009-2010[16] onward are considered more reliable as the monitoring has been systemized (Ministry of Drinking Water and Sanitation, 2016). IMIS has been consistently updating water quality data across all the 17 lakhs[17] habitations (as of 2016) within its directory (MDWS, 2016). In addition, in 2013, the Uniform Drinking Water Quality Monitoring Protocol has also been introduced, marking the progress towards a more standardized community-based rural water quality monitoring system (Ministry of Drinking Water and Sanitation, 2013b). While the availability of a public water census like IMIS offers a unique opportunity for managing community water resources across India that even developed countries lack, presently this database is grossly under-utilized in government decision-making (Parsai and Rokade, 2016; Wescoat, Fletcher and Novellino, 2016).

Water quality and sanitary survey data, as well as the addition of new water schemes and sources, are entered on a regular basis whenever new results come in. Sub-district level results are then approved of at a district level (Novellino, 2015). These data constitute part of the Monthly Progress Report for the physical progress created at the district level. The Monthly Progress Report has to be entered by the 15th of every month and sent on to the state for approval (Novellino, 2015).

[16] For India, a financial year spans the period from April 1 of a year to March 31 of the next year.
[17] Numerical units used in India, same as a hundred thousand.

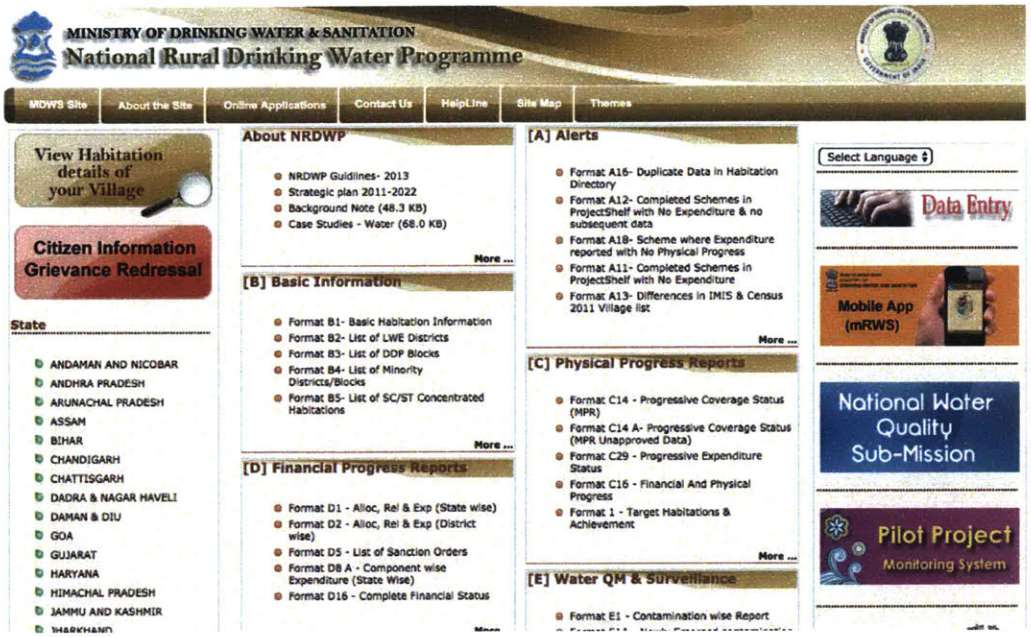


Figure 5-2: Demonstration of the front page for the IMIS database

5.1.2 Variable summaries

Variables selected from IMIS are based on their association with potential outbreaks, with a focus specifically on biological contamination of water or other factors that may be related biological contamination. While water quality data are the primary variables of interest, we also included factors related to the action component of the DPEESA framework, such as the number of water management trainings conducted, which may relate to water quality control and outbreak control.

The list of relevant columns and tables that they belong to are shown in Table 5-3.

Table 5-3: List of relevant IMIS variables for the integrated database

Table Description	Relevant Columns	Location in IMIS database
Habitation details and status	Administrative region (habitation-level) Population under each category (Total, ST, SC, General[18]) Number of households Liters per capita per day (LCPD) House connection	B1
Source details	Administrative region (habitation-level) Scheme Name + Type Source Category Type of Source Source location and ID	B6

[18] Similarly, this shows the distribution of historically disadvantaged ST/SC population.

Water quality lab results	Administrative region (habitation-level) Source location and ID Lab name Testing date Parameter name and value	E1-6
Water quality FTK results	Administrative region (habitation-level) Source location and ID Type of Test (bacteriological/chemical) Testing Date Positive for contamination (yes/no) Contaminant name	E20/C25
Sanitary Survey results	Administrative region (habitation-level) Source location and ID Source sanitary category Date of visit Survey personnel and agency Recommendation and measure remark Risk score	E27
Trainings	Administrative region (panchayat-level) Block training participation + date Panchayat training participation + date Number of trainees through panchayat-level training	E18

These tables are all connected with each other through the schema in shown in Figure 5-3. The top row in the schema is the relevant segment from the administrative-level base schema in Figure 5-1. Different types of water sources within each habitation are separated into three different tables. These water sources are associated with sanitary survey results and water quality testing results (lab test and FTK test). Independently, training records for local leaders and workers exist at the panchayat level.

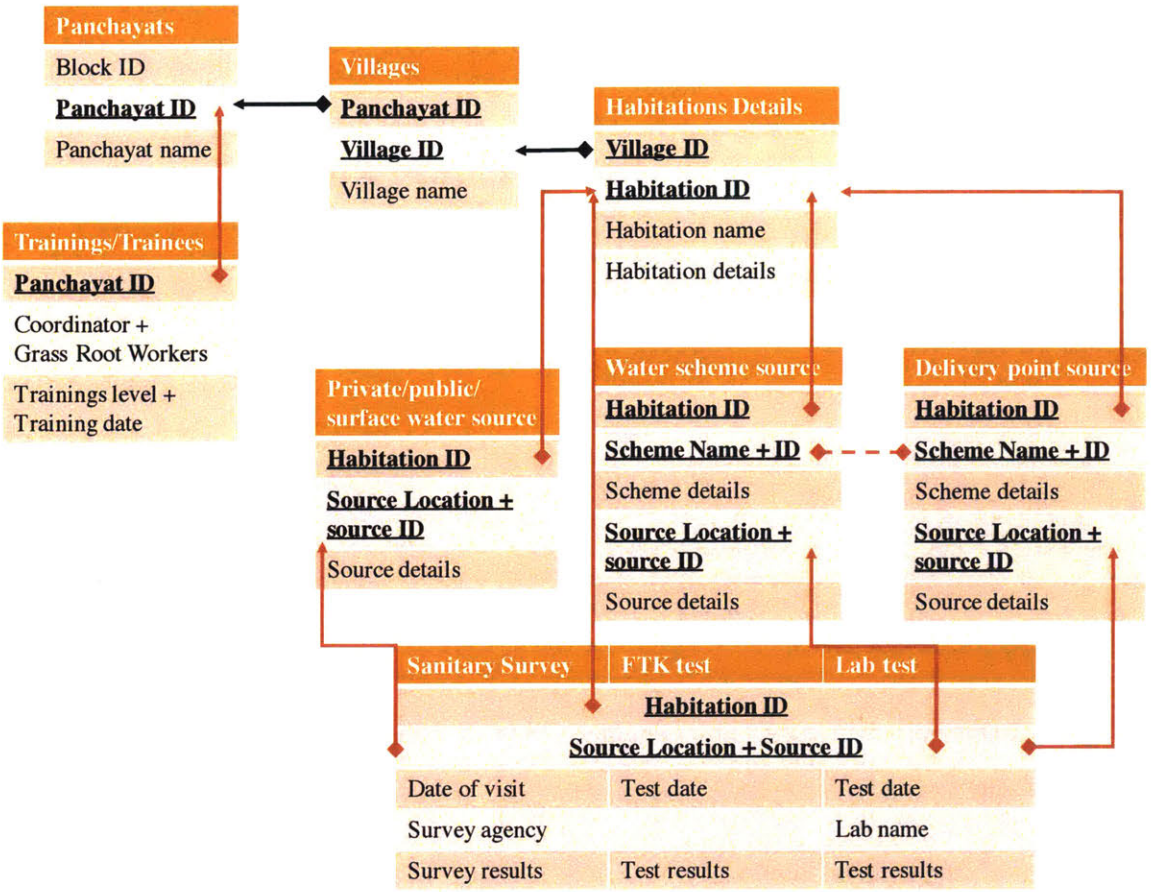


Figure 5-3: Schema connecting tables of interest within the IMIS database

We go on to individually evaluate each table for their cost of integration, as outlined in Chapter 3.

Habitation Information

The basic habitation information data table includes basic population and water access details at the habitation level, including details on (IMIS website):

- Population characterization: the total population of the habitation, as well as the number of population characterized by Scheduled Tribes, Scheduled Castes;
- Total household numbers: the number of households within the habitation;
- LPCD: water access described by average liters per capita per day;
- Water supply coverage status: "fully covered," where the average supply of drinking water is equal to or more than 40 lpcd; "partially covered," where the average supply of drinking water is less than 40 lpcd but equal to or more than 10 lpcd. Other statuses also include "Not Covered" with habitations less than 10 lpcd, as well as "Quality Affected" where samples tested indicated levels of chemical contamination (limited to arsenic, fluoride, iron, nitrate and salinity) for all sources in the habitation.

- House connections: the number of connections to the households.

A general summary of the data evaluation results is shown below in Table 5-4.

Table 5-4: Evaluation summary for IMIS habitation details data

IMIS	Data Availability	Total Entry	Columns	Data Type	Missing Data	Abnormal Data	Simplicity	Uniformity	Processing
B1	2009-2010 until now	5445	Administrative region	Text	0	0 (1 for 16-17 cycle)	Yes	No	No
			Population Characterization	Integer	0	0	Yes	Yes	No
			Households	Integer	0	0	Yes	Yes	No
			Coverage Status	Text: 2 categories	0	0	Yes	Yes	No
			LPCD	Numeric	50	0	Yes	Yes	No
			House Connection	Integer	202	0	No	Yes	No

Analysis on the data characteristics is carried out through the framework in Chapter 3:

- Accessibility

This table is readily accessible and can be downloaded for an entire district at once. Data can be dated back to 2009-2010, and is updated annually.

- Simplicity

Most variables are relatively straightforward. However, there is not a clear definition for "house connection." The number of house connections is in many cases larger than the household number (one habitation has 6 households but 176 house connections). This variable clearly requires stronger definition.

- Uniformity

The entries are generally consistent with no extra processing required.

For the different administrative levels, spelling issues occasionally occur when "Falia" is erroneously typed as "Fal;ia." Some errors are corrected in the 2016-17 cycle, but there is still one error entry with "VANJAR FAL;IA." In addition, there does not seem to be a uniform standard for the upper/lowercase of the text. For 2015-16 data, all administrative level names are in capital letters, but many of the new 16-17 entries have lowercase names. This may be standardized at the end of the year in March 2017. There are also habitation names entered with numbers in front of the names, such as "16 GALA" or "42-GALA." Although these habitation entries are consistent across the different tables, their format is clearly not standard.

The text entries for coverage status are well categorized.

- Completeness

There are 50 habitations missing LPCD entries for 2015-2016, and 202 habitations missing House Connection number entries. For all habitations missing LPCD entries, they are still classified as "Fully Covered," which raises concern over the accuracy of their coverage status entry.

- Quality/Accuracy

Data for this table are theoretically accurate, because they are part of the annual data entry process where the data is validated at both the state and central level. Target habitations are also identified through this process and funding allocation is also contingent upon these basic habitation data.

However, the missing entries still raise question about data accuracy, especially missing LPCD when habitation is identified as "Fully Covered." Additionally, in the past two years, there is a significant increase in habitation entries (Figure 5-4). Since it is unlikely that over 1000 new habitations were constructed in the past year, there may be significant ongoing data reconfiguration between last year and the current year. This raises more concern about the accuracy of habitation information for the 2015-2016 cycle.

Moreover, for the 2016-17 cycle, there is one repeat entry of Sayan habitation in Surat district. For 2015-2016, there are also repeat entries in other districts of Gujarat. There are also over 173 habitations that are missing an upper level village or panchayat name across Gujarat. These do not occur in our 3 sample districts.

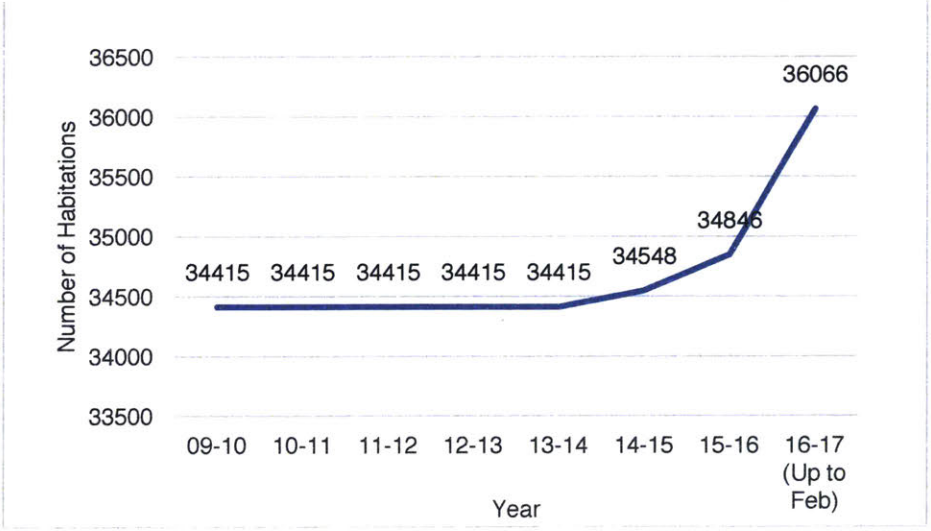


Figure 5-4: Increase in the total habitation entries in IMIS for the state of Gujarat

- Integration Viability

The habitation details data are at the base level of the schema. Other variables are all joined to the habitation table or other administrative aggregation based on the habitation table and their integration viability is analyzed according to the results. Although there are IMIS IDs from state to habitation listed sporadically in a few tables IMIS, the IDs are not consistently presented and are not available to any of our tables of interest. Joins have to be conducted through matching administrative region names. The same habitation and village name may indicate different regions across different panchayats and blocks, due to the likely repeated usage of the same name across different regions. Hence, the primary

key or unique identifier for the habitation table spans all five columns of district, block, panchayat, village and habitation. This clearly creates more challenges for data integration.

For Navsari district, there is often an empty space after the district name, which causes issues during data integration. This can be relatively easily fixed through space removal before and after the text, but it may be still worth noting and does add extra cost factors to the integration process.

Analysis on the data utility is carried out similarly:

- Acceptability

Considering that the data entry is related to state and central funding, there should be strong incentive to collect data on habitation details at the agency level. Motivation for data collection process at the local level is unclear.

- Sensitivity

The variables that reflect issues with water accessibility include coverage status, LPCD and house connections, all of which may be indirectly connected with health status. Coverage status also contains the "Quality Affected" status, but this variable only reflects serious chemical contamination. Water quality data have shown various cases of chemical and bacteriological contamination across Gujarat, yet no "Quality Affected" habitation is identified. Overall, this set of variable are only weakly effective at reflecting potential trends and violations within the system.

- Predictive value

The coverage status reflect a clear target, which is currently set at 40 lpcd, and expected to be raised to 55 lpcd by 2017, 70 lpcd by 2022 (Ministry of Drinking Water and Sanitation, 2011). This status is reviewed to define priorities and funding distributions, hence the "Partially Covered" should effectively indicate required improvement in water access.

On the other hand, since there is no clear target for house connection, its predictive value is low.

- Timeliness

Since this data is validated annually, it is not a timely reflection of outbreak-related hazards within the water system.

Water Source

The water source information data table includes basic details on all drinking water sources. Three different tables of sources are included in the dataset, including supply scheme sources, delivery point sources, public and private sources. As shown in Figure 4-2, supply scheme sources and delivery points with the same scheme ID belong to the same supply scheme, where sources at the supply scheme origin is considered water supply scheme sources, while distribution points (likely connected through pipelines) are considered delivery point sources. Other individual sources are more generally categorized as public and private sources, depending on whether they are privately or publicly owned. There are also surface water sources listed in a separate table in IMIS, but none of them are in the 3 districts of interest.

Variables for Water Supply Scheme Sources table include:

- Administrative region (habitation-level)
- Scheme ID: identification number for the water supply scheme
- Scheme details, including scheme name, sanction year, estimated scheme cost, reported expense, commencement data and estimated completion date
- Scheme type: 9 types including combined supply, individual supply, regional supply, hand pumps, simple well, tapping, tubewell power pump, treatment plant and sustainability scheme.
- Source type: there are 21 source type with some of the main types outlined in Table 5-5.
- Source category: 5 categories including groundwater, surface water, rain water, traditional and others (the text would directly state "others").
- Source ID: identification number for the source at the supply scheme site
- Source location: description of the supply scheme source location

Variables for the Delivery Point Source table include many of the same variables that match with the water supply scheme they are connected with, including scheme ID and all scheme details. In addition, there are also:

- Administrative region detail (habitation level, not necessarily same as the corresponding scheme source)
- Source ID: identification number for the delivery point
- Source location: description of the delivery point location

Variables for the Public & Private Source table include:

- Administrative region detail (habitation level)
- Source ID: identification number for the source
- Source location: description of the source location
- Source category and source type: same as defined above for water supply scheme sources

Table 5-5: Source type categorizations in IMIS (categorization hierarchy order from left to right)

Source Types			
Pipe Water Supply			Ground water based
			Surface water based
Other	Ground Water Schemes	Open Well	
		Tube Well	Shallow
			Deep
		Infiltration Gallery/Well	
	Surface Water Schemes	Pond	
		River	
		Lake	
		Stream	

		Canal
		Spring
		Treated Surface Water
		Rivulet Naula Gadhera
	Rain Water Schemes	Rooftop Structures (Community, Individual)
		Ground Collection (Community, Individual)
	Traditional	Khadins/ Nadis/ Tankas/ Ponds/ Wells/ Ooranis
	Other	Non-Conventional

A general summary of the data evaluation results for the variables above is shown in Table 5-6.

Table 5-6: Evaluation summary for IMIS water source description data

IMIS list	Data Availability	Total Entries	Column	Data Type	Missing Data	Abnormal Data	Simplicity	Uniformity	Processing/ Table Joins
B6- Supply Scheme Sources	only current version	12463 (11969 unique schemes)	Administrative region	Text	0	0	repeated		100% Joined to Habitation table
			Scheme ID	Integers	0	0	Yes	Yes	277 joined to multiple schemes
			Scheme name	Text	0	0	-	No	
			Sanction Year	Year	2209	0	Yes	Yes	
			Est Scheme Cost	Numeric	0	2177	Yes	Yes	
			Reported Expense	Numeric	0	3768	Yes	Yes	
			Commencement Date	Date	2135	0	Yes	Yes	
			Est. Completion date	Date	2118	0	Yes	Yes	
			Scheme Type	Text: 9 categories	0	0	No	No	Category reconfiguration
			Source Type	Text: 22 categories	0	0	No	No	Category reconfiguration
			Source Category	Text: 5 categories	0	0	Yes	Yes	
			Source ID	Integer	0	0	Yes	No	
			Source Location	Text	5637	0	-	No	
B6- Delivery Point Sources	only current version	12969	Administrative region	Text	0	0	repeated		100% Joined to Habitation table
			Scheme ID	Integers	0	0	Yes	Yes	31 not joined to any schemes
			Source ID	Integer	0	0	Yes	Yes	
			Source Location	Text	1		-	No	Text String Extraction
B6- Public and Private sources	only current version	59767	Administrative region	Text	0	1	repeated		100% Joined to Habitation table
			Source Type	Text: 22 categories	84	0	repeated		
			Source Category	Text: 5 categories	23	0	repeated		
			Source ID	Integer	0	0	Yes	Yes	
			Source Location	Text	15	0	-	No	

Analysis on the data characteristics is summarized below.

- Accessibility

Data are accessible and can generally be downloaded at the district level. However, only the most updated data are available - there is no data snapshot by year. Snapshot data are important because it relates to the total number of sources to be tested in a given year. The estimated completion date variable may help recreate the snapshot, but it is an approximate and it would increase the workload for conducting time-series related WaSH-health analysis.

Apart from the source tables in B6, attempts are made to access the actual tables for water supply schemes that some of these sources belong to. However, none of the water supply schemes tables can be accessed on IMIS. There is also more detailed information for each water supply scheme in IMIS such as functionality status and contamination status of its connected scheme sources and delivery point sources. However, these data are not in the IMIS-B6 table or other easily accessible formats. They are only available at each individual scheme level when clicking on the scheme profile (Figure 5-5). This makes accessing the data challenging, which requires scraping and consolidating over 10,000 schemes. These data, while relevant to our analysis, are not included in this study due to the same reason.

Scheme Profile

State: **GUJARAT** District: **SURAT**

Scheme Name: **Afva Hand pump scheme (PWS)** Scheme Type: **INDIVIDUAL WATER SUPPLY SCHEME(PWS)**

Scheme Category (For QA/SCDominated/STDominated/Schools/DDP/LWE/PilotProject/General): **Scheme for ST dominated habs**

Percentage of schemes compeletion (Physically): **0%**

*** All amount in Lakh Only**

Sanction Year :	2002-2003	**Estimated Cost :**	5	**Revised Cost :**	0
GOI Share :	2.5	**State/Other Share :**	2.5	**Community Share(Cash+kind+labour) :**	0
Expenditure(Till Date) :	1.75	**Commencement Date (dd/mm/yyyy):**	01/01/1900	**Completion Date/Tentative Date (dd/mm/yyyy) :**	31-10-2015
Service Level (LPCD) :	40	**Population (Proposed to be Covered) :**	82	**Implementing Agency :**	GWSS Board
Programme (Scheme Covered Under) :	MNP (State Funds)	**Contamination Found :**	None	**Agency Responsible for O & M :**	Gram Panchayat / PRI / Village Committee
Sustainability measure taken :	None	**Use of Non-Conventional Energy :**	None	**Waste Water Managment :**	None

Scheme Expenditure Reported

Financial year	Total	Apr	may	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar
2015-2016	1.75485	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	1.75485	0.00000	0.00000	0.00000	0.00000	0.00000

No of Quality Affected Habitations Covered

Arsenic	Fluoride	Others	Total
0	0	0	0

Name of Quality Affected Habitations Covered

District Name	Block Name	Panchayat Name	Village Name	Habitation Name	Quality Description
-	-	-	-	-	-

Details of Sources in Scheme :

Sl.N.	DistrictName	BlockName	panchayatName	VillageName	HabitationName	Total Population	Location / Name	Source Category	Source Type	Functional	Location Details on Map
1	SURAT	BARDOLI	AFVA-KHALI-GOTASA-TAJPOR	AFVA	NISAL FALIA	344	In habitation	Ground Water	Deep Tubewell	Yes	--

Details of Delivery Point / House Connections :

Sl.N.	DistrictName	BlockName	panchayatName	VillageName	HabitationName	Total Population	Location / House Connections	DateOfCommissioning	Functional	Location Details on Map
1	SURAT	BARDOLI	AFVA-KHALI-GOTASA-TAJPOR	AFVA	NISAL FALIA	344	whole village	01/01/1900	Yes	--

Figure 5-5: Scheme profile with relevant additional information(highlighted) on water sources that are not easily accessible

- Simplicity

The scheme and source IDs are relative straightforward identifiers for water supply schemes. Source details include expenses and construction time are in standard numeric and date/time formats. The definition for scheme types and source types were broad with overlapping categories. For example, both the supply system type of categories (individual/combined/regional) and source related categories (hand pumps, simple well) are categories under the same variable of "scheme types." While major source types are more clearly laid out in Table 5-6, there are still minor source types such as "filter points" whose definition is not clear. Source category was well-defined with a built-in "other" category. As for scheme names and source locations, there are no definition at all, and the entries are relatively random and varied widely.

- Uniformity

For scheme names and source locations, the entries had no consistency because there is no definition of these variables. For the source location in the Delivery Point Sources table, there is also an extra "(Delivery Point)" at the end of each location entry, which need to be removed to extract the actual text for source location. For the scheme types and source types, the consistency was also low. Quite a few of the same category have different phrasing (e.g. "Individual" / "Individual Water Supply Scheme", or "Hand Pump" / "Hand Pumps") that need to be consolidated.

Source IDs are generally consistent, and can be considered the primary key to identify sources in the Delivery Point Source table and the Public & Private Source table. However, Source ID is not the primary key for the Supply Scheme Source table, where there are 277 source IDs that are repeated. For all supply scheme sources with repeated IDs, they are all in the same habitation and have similar scheme names and details (except 1 with a different habitation name from others but still share the exact same scheme name). They are very likely to be the same source that has been duplicated, likely due to repeated updates to the water supply schemes or errorenous duplicates. To ensure that source IDs can be consistently used to uniquely identify all sources, repeated entries may require consolidation.

- Completeness

Missing data on scheme details are calculated based on unique scheme entries rather than scheme source entries. Around 18% (2133 out of 11969) of supply schemes do not have commencement date, completion date entries. A similar 18% (2177 out of 11969) also have abnormal entry of 0 for the estimated cost. Around 28% (3272 out of 11659) of the sources that were estimated to be completed before the 2015-2016 cycle still reported 0 expense. 5637 source location entries are missing for the supply scheme sources.

For the public and private sources, there were a few (<0.2%) source types, source categories and source locations missing. For delivery points, there were no source type or category columns (likely because it should just be the same as the connecting supply scheme sources) and only one source was missing the location entry.

- Quality/Accuracy

Based on observing the data tables, the quality of data on delivery points and public/private sources are generally acceptable. On the other hand, many inconsistencies exist within the water supply schemes and scheme sources. The same source corresponds to multiple schemes. Many schemes with very similar or sometimes even identical details are assigned different scheme IDs. There are

ongoing efforts working on consolidating the schemes in Format A within IMIS, including listing out schemes that are likely duplicates, and schemes where expenditure and physical progress are missing. More updates are expected for the supply scheme sources before the data quality becomes ideal.

- Integration viability

All habitation columns in the 3 source tables can 100% be joined with the current 2016-2017 habitation details.

A many-to-many relationship is expected between the scheme ID of water scheme sources and delivery point sources, because theoretically they are both matched by one-to-many relationships to the same supply scheme as shown in Figure 5-6. Since Water Supply Scheme tables are not accessible, we generated a supply scheme aggregation table summarized through water supply scheme sources. Due to the process of aggregation, all water supply schemes are matched with 1 or more water scheme sources. 277 sources are matched with multiple schemes.

As for delivery point sources, all but 31 delivery points were matched to water supply schemes. Among these 31 delivery points, there are two unmatched schemes. One is a scheme with a scheme source that is in Narmada, which is outside of the 3 districts' data that we collected. The other is a relatively old scheme built in 1991 that does not have a water scheme source listed. There are also 1649 water supply schemes that are not matched with delivery point sources. Most of these are schemes supplying water to schools or other community complexes via water scheme sources – these sources may not be considered "delivery points," which are generally at the household level.

The source IDs are expected to be matched with water quality results and sanitary survey results. For delivery point sources and public/private sources where the source ID is the primary key, the integration should be smooth. However, for water supply scheme sources, there are still a few source ID that may correspond to multiple entries. The source ID and scheme ID column need to combine together to satisfy the uniqueness criteria for the primary key. Consequently, to match water test and survey results with water supply scheme sources, it would be ideal if both source ID and scheme ID are available, which would require more effort in the process of data collection.

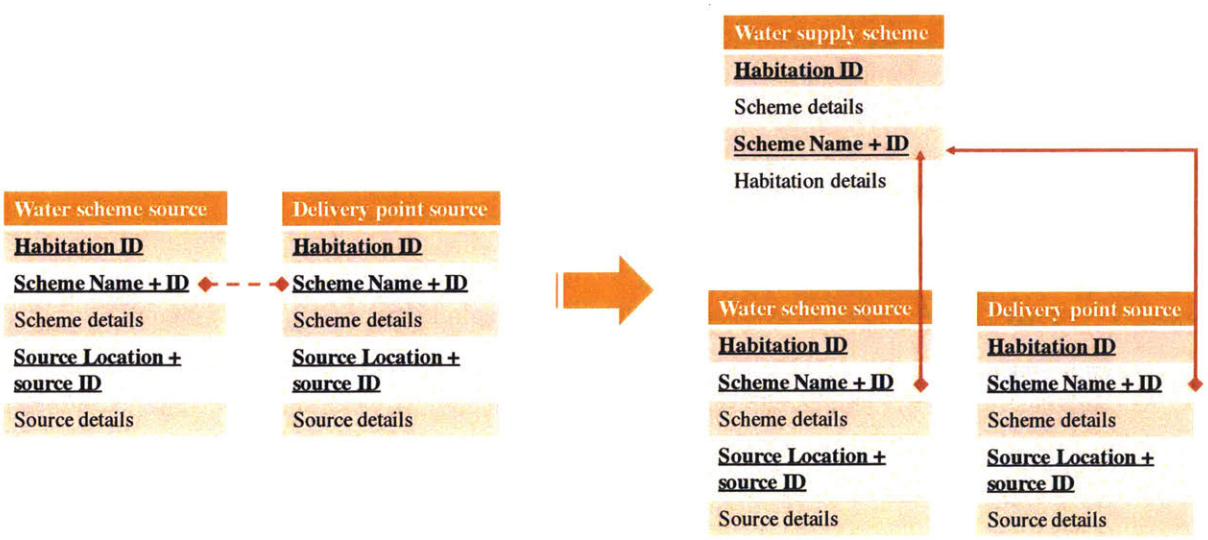


Figure 5-6: Details on the many-to-many relationship between the Water Supply Scheme Source table and the Delivery Point Source table

Analysis on the data utility is summarized below.

- Acceptability

Data entry on water sources are largely conducted through GWSSB and WASMO. There seems to be a higher incentive to update data on household delivery point sources and public/private sources. These entries may be more directly related to the number of house connections and overall water coverage LPCD, so there are more incentives to complete the entries. On the other hand, with over 20% data missing and over 200 entries that require consolidation, there seem to be less requirements and motivation on reporting progress of water schemes and water scheme sources.

- Sensitivity/Predictive value/Timeliness.

The sources table do not contain information regarding quality or functionality compliance of the sources, so the sensitivity and predictive value of the data cannot be evaluated. The information is available at the individual scheme level, which was not available for this study. Scheme entries and updates are conducted on a regular basis, same as water quality tests. In theory, contamination should be reflected timely. However, only fluoride and arsenic are clearly labeled in the scheme profile (Figure 5-5), so it is unclear if biological contamination is reported at the scheme level.

Water Quality Lab Test Results

The Water Quality Lab Result table include laboratory test results of drinking water sources. Details on the water source are listed together with the test results. A variety of parameters are tested in the water labs, but only parameters related to potential outbreaks, as suggested by variables selected in the Bayesian Network outbreak detection algorithm (Burkom *et al.*, 2011), are selected for this table, including:

- Total coliforms and E. coli, both of which are indicators of fecal contamination. Coliform bacteria are naturally occurring and usually non-pathogenic, but can generally be used as an indicator for contamination. E. coli is a particular species of coliform bacteria, and while typically non-pathogenic, there are strains that can be harmful to human health. E. coli presence likely indicates more serious health concerns. Both are expected to be non-detectable in safe drinking water sources.
- Residual chlorine: a low level of chlorine remaining in water after the initial chlorination disinfection application, which serves to continuously safeguard water from bacteriological contamination. A decrease in chlorine level may indicate higher risk for contamination.
- Total Dissolved Solids (TDS), Turbidity, pH: all of which can be indirect indictors of pathogens. Turbidity and TDS can indicate suspended or dissolved solids that might include pathogens, while abnormal pH may indicate unnatural substances in water, which may relate to pathogenic contamination.

The values for these 6 parameters are available for each entry and they are categorized as either "above permissible" or "below permissible." Hence both a numeric and a Boolean column is available for each water quality parameter entry.

For each test record, there is an adjusted alphanumeric source ID containing letters that identify one of the three tables where the source belongs to, and numbers that correspond to the original source ID in the source tables. This alphanumeric ID is separated into a source list column containing the letters, and the source ID column which is the original source ID.

The resulting table include the following variables:

- Administrative region (habitation-level)
- Source ID and location: same as the ID and locations entered in the source tables
- Source list: the letters separated from the alphanumeric ID which indicate the relevant table that the source belongs to, i.e. supply scheme sources, delivery point sources, public sources, private sources or surface water sources table.
- Lab name: the name of the lab where the test is conducted
- Test date: the date that the test is conducted
- Parameters: the numeric value of the test result and whether it exceeds permissible standard.

There are 111 duplicate records where all entries including test results and test dates are the exact same for the same source. While these may be repeat or parallel samples, there are insufficient details justifying the duplicates. They are removed for subsequent analysis.

A general summary of the data evaluation results for the variables above is shown in Table 5-7.

Table 5-7: Evaluation summary for IMIS water quality lab results data

IMIS list	Data Availability	Total Entries	Column	Data Type	Missing Data	Abnormal Data	Simplicity	Uniformity	Processing/ Table Joins
E3 – Water Quality Lab Test Results	2006-2007 until now	27718 entries (excluding 111 repeats) 16741 sources	Administrative region	Text	0	0	repeated		100% joined to Habitation Table
			Source Location	Text	33		-	no	Source ID, list and location extracted from text string; 100% joined to Source Table
			Source ID	Integer	0	0	yes	yes*	
			Source List	Text: 5 categories	0	0	yes	yes	
			Source Type	Text: 22 categories	0	0	repeated		
			Lab Name	Text: 86 Labs	0	0	yes	no	Category reconfiguration
			Testing Date	Date	0	0	yes	yes	
			TDS	Numeric/Boolean	2387	0	yes	yes	Test result extraction from text string
			pH	Numeric/Boolean	2388	2	yes	yes	
			Turbidity	Numeric/Boolean	15888	0	yes	yes	
			Residual Chlorine	Numeric/Boolean	13180	0	no	yes	
			E.coli	Numeric/Boolean	13288	8	yes	yes	
			Total Coliforms	Numeric/Boolean	13288	8	yes	yes	

*apart from the uniqueness issue with water scheme sources, the IDs are uniform and correspond 1-to-1 to water source entries within the respective source table.

Analysis on the data characteristics is summarized below:

- Accessibility:

The data are relatively easy to access and can be downloaded at the district and state level. Data are available since 2006-2007 since the Guidelines for National Rural Drinking Water Quality Monitoring and Surveillance Programme was first published (Rajiv Gandhi National Drinking Water Mission, 2006), and total number of results have increased drastically over the past 10 years from only around 1000 samples to the 220,000 samples tested in Gujarat.

- Simplicity:

For the source types and source location entries, the lack of clear definition is the same as described in the previous section. For all other variables, the definition is straightforward. The standards for classifying water quality results are shown in Table 5-8. Permissible limits, which is a lower standard used in the absence of an alternate source, are used in IMIS, rather than the actual requirement that is recommended. There is confusion regarding residual chlorine entries because the standard actually requires a minimum value of chlorine rather than maximum. Hence while for all other parameters, above permissible means contamination, for residual chlorine being above permissible level is the requirement. Additionally, 1 mg/L was used as the permissible limit cutoff for residual chlorine. However, 0.2 mg/L is actually the more relaxed standard. Since relaxed standards are used for all other parameters, it is only reasonable that 0.2 mg/L is used as the permissible standard for Residual Chlorine instead of 1 mg/L. The Boolean variable for Residual Chlorine is reconfigured to better align with other parameters for indicating water safety, and the cutoff has been adjusted to 0.2 mg/L.

Table 5-8: Water quality standards (Bureau of Indian Standards, 2012)

Parameter	Unit	Requirement	Permissible limit
Coliform	maximum value, MPN/100mL	0	**0**
E-Coli	maximum value, MPN/100mL	0	**0**
pH	acceptable range	6.5-8.5	**6.5-8.5**
Residual Chlorine	minimum value, mg/L	**(0.2)**	(1)
Total Dissolved Solids (TDS)	maximum value, mg/L	500	**2000**
Turbidity	maximum value, NTU	1	**5**

- Uniformity

A number of extra processing is needed to ensure consistency of the data entries. The source list and source IDs needed to be extracted from the source location string and separated into the letter component and the number component. The water quality test results are initially listed in text strings as well (e.g. "Nitrate[0.260 mg/l],Fluoride[0.050 mg/l]"), and the variable name, value and units also need to be extracted to create separate columns for each parameter. After the initial processing, the entries extracted are generally consistent.

The lab name entries are not uniformly presented. There are 86 operating labs in Gujarat, but the entries for the same lab varies widely with different spellings, abbreviations and capitalizations used. All other variables are generally uniform in formats.

- Completeness

16,741 sources are tested out of over 80,000 sources listed in the sources table. The lack of snapshots may have exaggerated the number of sources available in 2015-16, but even if we exclude the 1000+ sources associated with ongoing schemes or non-functional schemes, there is still more than 80,000 sources left that should theoretically be tested, out of which only around 20% are.

Among the 16741 sources tested for at least one parameter in the lab (including other parameters that are not among the six selected for this analysis), around 2300 are not tested for TDS and pH parameters, while over 13000 are not tested for biological parameters including Residual Chlorine, E.coli and Total coliforms.

While the table includes data on water quality compliance, there is no information on follow-up actions. Apart from 1 source (public source ID: 3198310) that showed a repeat test record indicating water safety through chlorination after the initial contamination record, there are no repeat tests for the other 26 E. coli or Total Coliforms-positive entries to show improved water safety. It is unclear how the contamination is dealt with.

- Quality/Accuracy:

There are a few abnormal entries of water quality results including two extremely low pH values, and 8 very high Total coliforms (all of which are exactly the same at 1600 MPN/100mL) and E.coli values. The 8 extremely high biological contamination values all seem to be for raw water, and it is not clear whether these "raw" sources are directly consumed or not. The 2 samples with pH < 4 are from tube wells and it is uncertain why the acidity is extremely high.

Among all the parameters, the biological contamination parameters may have the most concerns with accuracy, considering the general challenges with sample collection, transportation and storage which would largely affect bacteria test results. Nevertheless, the lab staff are generally well-qualified. There is also a designated sample collector in most of the labs. With the enhanced focus that the Uniform Drinking Water Monitoring Protocol has placed on developing the capacity of water quality labs, the results are expected to be increasingly reliable.

There is also one water quality entry with a public source (source ID: 3150032) which is listed in Machhivad habitation instead of Borsi - the habitation for all other water quality lab test entries in the this table, and the habitation listed in the Sources table. This is likely an entry error, which also raises question on how water source information is entered for water quality entries, and why the data table allowed for the same source ID to correspond to different administrative locations.

- Integration viability

All water quality test records are matched with the latest administrative regions at the habitation level. Even though the water quality lab test table is a snapshot at the end of 2015-16, its administrative region column is updated according to the current 2016-17 habitation in the Habitation Details table.

All water quality test records are matched with source details from the 3 Source tables. Both delivery point sources and public/private sources can be matched through source ID, while supply scheme sources need to be matched through both source ID and source location. There are no scheme ID or scheme name columns for the water test entries, so source ID and source location along cannot

guarantee 1-1 matches. For the 47 supply scheme sources tested in 2015-16, this issue was not encountered.

Analysis on the data utility is summarized below:

- Acceptability

Increasing lab capacity is observed through IMIS and through interviews with the Vadodara Regional Lab. The 3000 samples annual target is monitored through IMIS, and labs are motivated to effectively conduct data collection. However, there is a concerning discrepancy between the number of biological and chemical test records, especially considering that the Uniform Drinking Water Monitoring Protocol recommends bacteriological tests be conducted twice a year and chemical tests once. The comparatively weaker motivation for bacteriological tests is likely due to the time and meticulous practice they require.

- Sensitivity

Out of over 80,000 sources available in the 3 districts, only 3453 sources (~4%) have been tested for direct indicators (E. coli, Total coliforms) that can most immediately reflect a health concern within the drinking water system.

Other parameters, while useful at flagging abnormality potentially associated with pathogens in the drinking water, are not the most effective reflections of bacteriological contamination. Considering these indirect parameters as well, 16740 sources (~20%) are covered. The low lab test coverage level also decreases the ability of these test records to effectively reflect health concerns. However, as we have noted before, the total number of sources may not actually reflect existing numbers of operating sources. Defunct sources and sources still under construction are also listed, and there are also cases where multiple test record entries for the same source are mistakenly entered as multiple sources (according to interview with Ms Trivedi from Vadodara lab). The lab test coverage level may be much higher, but it cannot be accurately determined from the database.

On the other hand, 5242 habitations out of 5445 (~96%) have at least one source tested (for any of the parameters) in 2015-2016. Parameters may be more effective reflect hazards at the habitation level. Nevertheless, the connections among different sources within the same habitation is not noted in IMIS, and test results from one source may not reflect the safety of other sources. In addition, if we only consider E. coli or Total Coliforms test, only 929 out of the 5445 (~17%) habitations have any sources that have been tested from direct bacteriological parameters.

Overall, the sensitivity of lab tests results for health concerns in all drinking water systems is relatively low.

- Predictive value

Typically, non-compliant data for E. coli and Total Coliforms should reflect health concerns. However, half of the non-compliant results seem to be for raw water / headworks water as the source location description suggests. Despite the many categories of source types, the source information does not specify whether the it is for direct consumption or further treatment. Hence, it is unclear whether these data actually reflect health concerns.

Other parameters are only suggesting increased likelihood of biological contamination. 1624 out of the 1722 non-compliant (~94%) Residual Chlorine records are accompanied with E. coli or Total Coliforms confirmations, and none of the follow-ups found bacteriological contamination. None of the 4 non-compliant records for pH and Turbidity have bacteriological test as a follow-up. As a result, these parameters alone are ineffective in predicting health concerns within the drinking water system.

- Timeliness

For sources that have been tested, the frequency distribution for Total Coliforms and pH tests over 2015-16 is shown in Figure 5-7. E. coli and Residual Chlorine are often tested concurrently with Total Coliforms, while TDS is frequently tested with pH, so they share similar frequency histograms. As shown, around half of the sources that are tested for the parameter of concern are tested only once across the 2015-16 cycle, while most of the other half is tested 2-3 times.

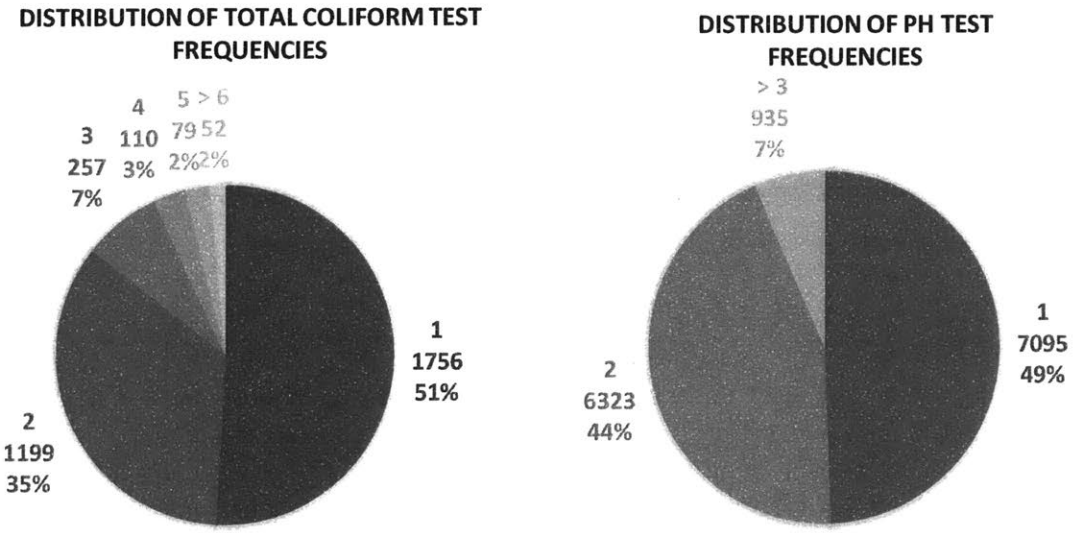


Figure 5-7: Distribution of the lab test frequency for each source (with the number of sources in each category): Coliform and pH (2015-2016)

For half of the sources which have been tested more than once, the distribution of all intervals between any two consecutive Total Coliform tests and TDS tests at the same source is plotted in Figure 5-8. A number of Total Coliforms tests are conducted repeatedly within 50 days of the initial testing. Many of these sources were continuously monitored for a number days, likely for specific occasions such as the installation of a new system. In addition, both Total Coliforms and TDS testing intervals show a peak region around 200 days, which is likely the interval between the required pre-monsoon (April-May) and post-monsoon (Oct-Nov) water quality testing.

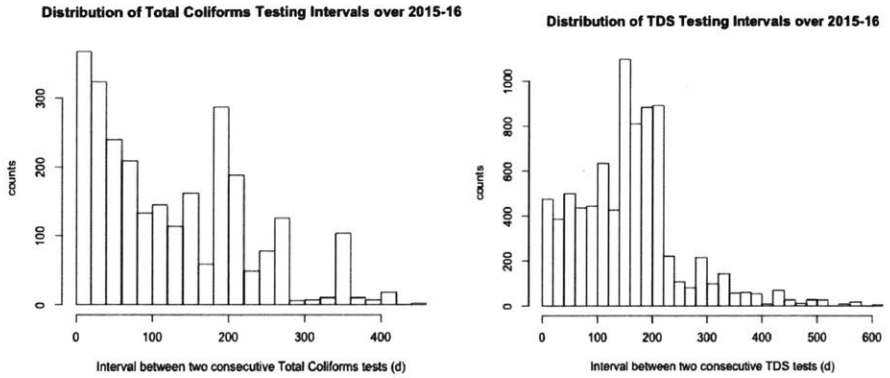


Figure 5-8: Distribution of intervals between two consecutive tests at each source: Total Coliform and pH (2015-16)

Overall, while quite a few sources have been repeatedly tested for bacteriological contamination at an interval of less than 3 months, the majority of sources are either tested only once or twice – with once pre-monsoon and once post-monsoon. Timely detection of health hazards is unlikely, but general trends can be observed especially when both pre-monsoon and post-monsoon test results are available, since monsoon is likely to have an impact on water quality.

Water quality FTK test results

The water quality field test kit (FTK) results table include water quality results of drinking waters sources obtained through field test kits. Details on the water sources are listed together with the results. For each entry, a source is identified as either safe or contaminated, and for the latter, the contaminants found through FTK are listed. The contaminants of interest are extracted from the list and separated into a separate column as a Boolean data. Similar to parameters selected for the lab tests, TDS, Total Coliform and E. coli are selected for this table. Turbidity, pH and Residual Chlorine are not listed in the FTK test results in IMIS. In contrast to lab test results, parameters tested to be negative are not listed, so there are no "false" Boolean entries. Null values are entered if the FTK found no contamination, because the exact tests conducted are unknown. Only the type of test – chemical or biological – is listed.

Similar to lab records, source ID, location and source list columns need to be extracted from a text string containing an alphanumeric ID. For FTK results, the supply scheme name is also included in the text string for all supply scheme sources and delivery point sources, and it is also extracted as a separate scheme name column.

In addition, both E20 and C25 datasets on IMIS contain very similar FTK test results. However, C25 records are much more comprehensive – it generally contains all E20 records with a number of more entries. This can be observed through the entry numbers in Table 5-1 as well. E20 tables are frequently missing data for certain blocks or months. In addition, C25 records also contain the test type column. Consequently, C25 datasets are selected for this table. However, data for water quality entries in May 2015 is inaccessible for Chikhali block of Navsari District. For this specific segment of data, E20 entries

were used. Only entries positive for contamination are accessible for this block in E20, so 33 entries with coliform contamination are inserted into the C25 dataset.

The resulting table includes the following variables:

- Administrative region (habitation-level)
- Source details, including source location, source ID, supply scheme name and source list
- Test type: chemical, biological or both
- Test date: date of FTK test
- Contaminated or not: Boolean variable indicating whether any contaminant is found during the FTK test
- Parameters: Boolean variable indicating whether the parameter of interest is found as a contaminant. There are no "false" entries because only positive results are listed. All other entries are null.

A general summary of the evaluation results for the variables above is shown in Table 5-9.

Table 5-9: Evaluation summary for IMIS water quality FTK results data

	Data Availability	Total Entries	Column	Data Type	Missing Data	Abnormal Data	Simplicity	Uniformity	Processing/Table Joins
C25 (E20)-FTK results	C25: 2010-2011 until now (E20: 2005-2006 until now)	17948 entries	Administrative region	Text	4	0	Repeated		E20: fill empty habitation entries; 100% joined to Habitation table
			Source Location	Text	155				Source ID, list, location and scheme name extracted from text string; 12 not joined to source table
			Source ID	Integer	0	0			
			Source List	Text: 5 categories	0	0			
			Scheme Name	Text	2	0			
			Test Type	Text: 3 categories	33 (E20)	0	yes	yes	
			Testing Date	Time	0	0	yes	yes	
			Contaminated or not	Boolean	2409	0	yes	yes	
			TDS	Boolean	-	0	no	yes	need to split the contaminants list into binary variables
			Total Coliforms	Boolean	-	0	no	yes	
			E. coli	Boolean	-	0	no	yes	

Analysis on the data characteristics is summarized below:

- Accessibility

There are many inaccessible links within the C25 and E20 datasets. For most districts, while download links exist at the district level, it is inaccessible, potentially due to server load issues. A number of the datasets are only accessible monthly at the block level, which requires month-by-month scraping and consolidation. A few datasets are completely inaccessible, including May testing results from Chikhali as mentioned above, and the majority of results from Gandhinagar districts.

While the data before 2010-2011 were listed in C25, none of them were accessible. Data from E20 are incomplete and frequently missing entries in comparison to C25, but they are accessible up to 2005-2006.

Overall, the accessibility of this table is low. C25 and E20 also requires consolidation to ensure consistent results.

- Simplicity

Most definition and storage structure for the variables in this table is straightforward, apart from the parameter results. The current format only records contaminants, missing all compliance data. The test type variable can make up for some of the missing information, where sources marked as safe with biological FTK are likely negative for either e. Coli or Total Coliforms. Nevertheless, the parameter compliance data format can be improved. All parameters tested should be recorded.

- Uniformity

After extracting all source details from the text strings, the data entries are in relatively consistent formats. On the other hand, it is concerning that the same FTK data set are displayed quite inconsistently across two different tables – E20 and C25 – on IMIS, with the test type and many entries missing in E20.

- Completeness

4 habitation entries are missing in the 33 E20 records – this can be filled by matching with the Water Source tables.

Among the 14323 sources that have been tested with FTK, 2409 of the sources were not tested for biological contaminants.

In comparison to the water lab records, the FTK test table lacks any quantitative records on the parameters tested, even though a number of the test kits demonstrated by WASMO can show semi-quantitative values for certain parameters. In addition, the FTK test table do not have information on the agency conducting the test.

Similarly, there is a lack of information on follow-up actions – whether the samples are sent on to the labs, and whether the sources are safe after treatment. Among the 601 records of biological contamination or TDS contamination, only 23 are accompanied by same-day sanitary survey results, and only 25 have a follow-up FTK biological test record within 30 days. 8 of these follow-up FTK results still showed biological contamination, and it remains unclear whether the safety hazards are removed or not.

- Quality/Accuracy

Many of the FTK tests are conducted by the local community members, members of WASMO or ASHA workers. While these agents are generally trained by WASMO, most of them have minimal educational background. Without details on exact agency conducting the field test, the accuracy of the test results might not be guaranteed.

- Integration viability:

All test records can be joined to the Habitation table.

12 records cannot be joined to the 3 Sources table by the source list, source ID and scheme names (scheme sources only). Among these records, 11 of them can be joined successfully after adjustments to the scheme name or the source ID (Table 5-10). The majority of the mismatched records from the FTK results table have source IDs with an extra 1 or 2 digits at the end. This may potentially be errors

resulting from the data storage system where integers larger than 4 bytes are out of the "integer" datatype range. There is still one delivery point record with source ID 210019784 that cannot be matched with any reasonable adjustments to the ID. This is also the only entry among all FTK entries without any source location or scheme name information, so there are no extra details to match this source by.

These discrepancies question the source entry system and why sources that are not in the Source table can be recognized as a valid entry. It also shows the importance of having better definition and non-null requirement for the source and scheme name columns, which can help match the sources when there are erroneous source IDs.

Table 5-10: Updates on the source entries of Water Quality FTK Results table

Entry in Water Quality FTK Results Table				Entry in Source Table
Source Table	**Source Location**	**Source ID**	**Scheme Name**	**Updated Scheme Name/ Source ID**
Water Supply Scheme Sources	TW In Patel Faliya	4609882	Lakadbari - Chikar Faliya/patel Faliya WSS	**Lakadbari - Chikar Faliya/Patel Faliya WSS**
	CHOWKI FALIYA	2100041401	HANDPUMP -1	**210004140**
		48683405	Lakadbari (5 Habi.)	**4868340**
Delivery Point Sources	VANIYA FALIYA	21000673313	HANDPUMP-13	**210006733**
	at village level sump	63818471	Borsad degadia rwss part 1	**6381847**
	HALPATIVAS FALIYA	21000664345	HANDPUMP-45	**210006643**
	Dungri Falia	56209185	Lakadbari (5 Habi.)	**5620918**
	Near Gram Panchayat	56148522	Sitapur (2 Habi. / H.C.)	**5614852**
		210019784		**No reasonable revisions**
	In Village Sump	50097628	HANDPUMP IN 28 GALA FALIA	**500976**
	near panchayat	668653112	12th Finance	**6686531**

Analysis on the data utility is summarized below.

- Acceptability

With strong community-level mobilization support from WASMO, Gujarat has been one of the best performing states in terms of field test kit practices. ASHA workers and other community volunteers trained by WASMO are in charge of collecting FTK data, and since these people are likely also members of the communities that the water sources belong to, they have a stronger incentive to understand the water quality of these sources. On the other hand, most of the community volunteers are may not be scientifically-trained enough to understand the implications of the test results. Their willingness and motivation to engage in the monitoring process may largely dependent on the effectiveness of WASMO's training and management routines.

- Sensitivity

11914 sources have been tested for biological contamination via FTK. While this is only about 15% of all sources listed, it is still 3.5 times more coverage than lab tests on direct biological contamination.

4316 habitations (79%) are covered by FTK for biological tests. While this is lower than the 96% coverage by lab tests for all 6 parameters, it is significantly higher than the 929 habitations covered with E. coli and Total Coliforms lab tests, which are more comparable to the biological FTK tests. Overall, while FTK are still limited in their scope, they cover more drinking water supplies and are more likely to identify health concerns in the system compared to lab tests.

- Predictive value

TDS results are only indicative of potential biological contaminations, so they would require follow-up. Out of the 63 positive TDS records, only 15 are also tested for biological contamination with 3 days, where 5 of them were found positive.

Considering the comparatively lower accuracy of FTK and higher chances of operational errors by the community volunteers, results from FTK tests are expected to be reconfirmed by lab results (Ministry of Drinking Water and Sanitation, 2013b). However, among the 538 test results that have found biological contamination or TDS contamination, only 5 have a next-day follow-up E. coli and Total Coliform records at water quality labs.

If we consider the overall water quality results across 2015-2016 for each source, sources that are tested for biological parameters both in lab and with FTK show the following distribution in Table 5-11. Only 8 sources have positive lab test results and none of the 8 sources tested positive for FTK. Among all hundreds of sources showing FTK contamination, only 6 of them had any accompanying biological lab test records. As a result, these two water quality records show little correlation with each other.

Table 5-11: Source biological water quality FTK results and lab results correspondence. Any source that have been tested positive for Total Coliforms or E. coli are considered positive for biological contamination.

No. of sources	No contamination (FTK)	Contamination (FTK)
No contamination (lab)	825	6
Contamination (lab)	8	0
Contamination ratio	1%	0%

Overall, positive TDS, E. coli and Total Coliform FTK results are indicative of potential hazards with drinking water supply, but without lab confirmation, the predictive power is considerably weakened due to potential false positives caused by the FTKs themselves or operational errors.

- Timeliness

For all sources that have been tested for biological contamination with FTK, the frequency distribution is shown in Figure 5-9. As shown, the majority of the sources are tested only once over the year. Only around 16% of the tested sources have more than one entries. If we compare this to the frequency of Total Coliform lab test in Figure 5-7, it can be observed that similar numbers of sources have been tested for more than once, but significantly more sources are tested once with FTK. Hence, while FTK has more coverage, the timeliness of FTK and lab tests are similar.

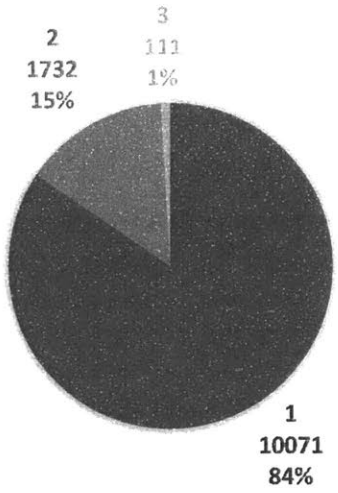


Figure 5-9: Distribution of Biological FTK Testing frequencies over 2015-2016 (with source numbers)

For sources tested more than once, the distribution of intervals between consecutive biological FTK tests is shown in Figure 5-10. Compared to lab test intervals in Figure 5-8, much fewer sources are repeated within 50 days of the initial testing. The significant peak around 200 days also indicate that for the majority of the sources that tested once pre-monsoon and once post-monsoon. Similar to lab tests, FTK tests are also only good for trend identification in the long run, and timely detection of issues in system is unlikely.

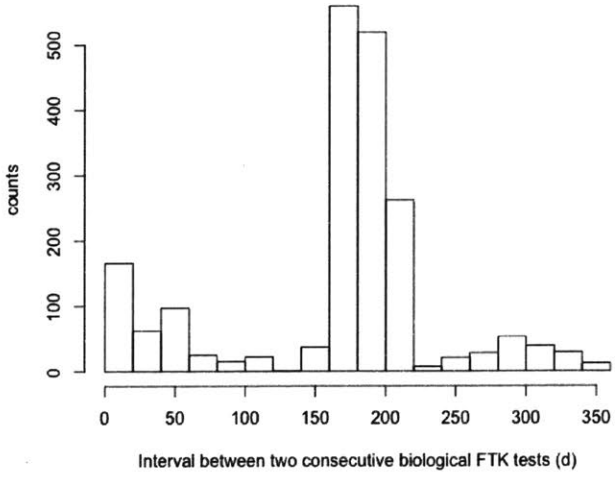


Figure 5-10: Distribution of Biological FTK Testing Intervals over 2015-16

Chemical FTK tests are not analyzed for timeliness because there is no available information on the parameters for the chemical tests. For most cases, it would be unclear whether TDS tests are included, and limited conclusions can be drawn.

Sanitary Survey Result

The Sanitary Survey Result table include results from sanitary inspections on the drinking water sources. As outlined in the Appendix of the Uniform Drinking Water Quality Monitoring Protocol (adapted from WHO), sanitary inspection is an on-site inspection of water supply facility to identify actual and potential risks of contamination with the drinking water source. Physical structures, operations and environmental factors (for example, pipe leaks, human excreta in the vicinity, permeability of the floor near the source) are all evaluated through inspections. An ultimate risk score is obtained by adding up all risk factors. A low-high risk ranking is then concluded. Communities can carry out their own routine frequent inspections for risk prevention purposes, while two annual inspections by local surveillance agencies are expected to check the reliability of reporting by local communities.

Similar to water quality results, details on the water source are also listed together with the test results. The same text string variable is present - source ID, source list and source location extraction is required.

Variables in this table include:

- Administrative region (habitation-level)
- Source details: Source ID, list, location and source type
- Source sanitary category: This is a separate category that is only used for the purpose of sanitary surveys. These categories are broadly defined in the Uniform Drinking Water Quality Monitoring Protocol, and different sanitary survey forms are used from each of the different category or sources because the risk factors vary by the drinking water facilities.
- Agency: The agency conducting the survey, which includes "Water Supply Agency" and "Surveillance Agency" for the 3 districts.
- Survey Done By: The person conducting the inspection, which includes "Community Representative" and "Water Authority" for the 3 districts.
- Recommendation: general statements of conclusion following the sanitary inspection, consisting of 3 broad categories which include comments on "keeping water sources clean", or statements that the water is "free from sanitary risk" or is "unfit".
- Measure remark: Suggestions following the sanitary inspection, which are mostly statements indicating "regular chlorination recommended". However, a number of the remarks are also conclusions that the water source is "free from sanitary risks". These are the 2 categories in the remark column, which seem to partially overlap with the recommendation column.
- Risk Level: the risk ranking concluded from the sanitary inspection form. Risk levels vary from "Very Low" to "High." Both "intermediate" and "medium" are used to categorize the risk level, and it is uncertain whether these two terms are used interchangeably.

There are 31 entries that are exactly the same, and they were excluded from the analysis.

A general summary of the data evaluation results for the variables above is shown in Table 5-12.

Table 5-12: Evaluation summary for IMIS sanitary survey results data

	Data Availability	Total Entries	Column	Data Type	Missing Data	Abnormal Data	Simplicity	Uniformity	Processing/ Table Joins
E27 sanitary survey	2010-2011 until now	965 rows (exclude 31 repeated entries)	Administrative region	Text	0	0	Repeated		100% joined to Habitation table
			Location	text	1	0			Name, source ID and list extracted from text string; 1 source matched with 2 scheme source with same ID
			Source ID	Integer	0	0			
			Source List	Text: 5 categories	0	0			
			Source Type	Text: 22 categories	0	0			
			Source Sanitary Category	Text: 12 categories & other	0	0	yes	no	Category consolidation
			Date of Visit	Time	0	0	yes	yes	
			Survey Done By	Text: 2 categories	0	0	no	yes	
			Agency	Text: 2 categories	0	0	no	yes	
			Recommendation	Text: 3 categories & other	0	49	no	no	Category consolidation
			Measure Remark	Text: 2 categories & other	973	0	no	no	
			Risk Level	Text: 5 categories	0	0	yes	no	

Analysis on the data characteristics is summarized below.

- Accessibility

The sanitary surveys overall have low accessibility, as shown in Table 5-1. Many of the districts have links that were unable to load. Some of the data become accessible only during certain times of the day, or only at the block level and need to be scraped and compiled. However, data for some districts such as Gandhinagar are still inaccessible despite all the attempts above.

While the data is available from 2010-2011, very few states across India were actively collecting sanitary survey information back then. Even for 2015-2016, only 8 states out of 32 have more than 50 records of sanitary survey entries.

- Simplicity

While the risk level is a simple ranking based on the sanitary survey score, the recommendation and measure remarks are much less straightforward. The entries for "recommendations" seem to mostly be conclusions on water safety, with some very broad recommendations such as "keep village clean." On the other hand, the entries for "measure remarks" are actually recommendations on follow-up actions such as chlorination. However, measure remarks also contain some of the same statements as recommendations entries, such as "free from sanitary risks." These two columns have very poor and unclear definitions.

In addition, the definition for survey agency and survey personnel is also vague, resulting in very general entries. While it's likely that WASMO and GWSSB units are conducting these surveys, it is still unclear which agencies "Water Supply Agency" and "Surveillance Agency" are referring to.

- Uniformity

While there are broad categories defined for "recommendation" and "measure remark", the exact entries vary significantly. For example, there are over 30 variations of "free from sanitary risk" and close to 10 variations of "please use chlorination treatment." There seems to be no standard entry formats for either of these two columns. While the risk level is better defined and consistently recorded, the "medium" and "intermediate" ambiguity is a critical issue. The distinction between "Low" and "Very Low" is also concerning because there are only 9 records that used "Very Low." The "Very Low" risk category is only documented for gravity-fed piped supplies and piped water supplies with service reservoir, while "Intermediate" is only documented for dugwells. This may be related to the definition on the sanitary form templates. Different sanitary source categories have different inspection form templates and while some of them had "Very Low," others only have "Low" as the safest level. In this case, it is unclear whether these two safety levels were used interchangeably, or whether some categories of sources are inherently lower in risk level even when no risk factors are observed (e.g. piped water may inherently be of less risk than open dugwells). Without these specifications, it is hard to conclude whether the categorization is consistent or not.

The source category specific for sanitary surveys is loosely based on the sanitary survey templates in the Protocol, but there are a few ambiguously similar categories (e.g. "rainwater collection and storage" or "rain water tank catchment") that may require consolidation. The entries are not consistent.

Overall, the definitions for many variables in the Sanitary Survey Result table are weak to begin with. There are also limited format constraints, resulting in many inconsistent and ambivalent entries. Manual consolidation and interpretation of the categories is required and the process is highly time and labor consuming.

- Completeness:

Out of 996 entries, only 59 rows had information in the measure remarks column. There is no consistent pattern for when a "measure remark" is made – chlorination remarks are offered for sources ranging from "Low" to "Very High" risk levels. All of the chlorination remark correspond to a "source is unfit" recommendation (but not vice versa), which questions the utility of this column if the information is a mere duplicate of the "recommendation" column.

Based on the sanitary inspection templates, detailed risk identification is carried out in each inspection, but the actual risk factors are not recorded in IMIS. Risk management actions only include chlorination, neglecting any direct fixes regarding critical risks such as leaked pipes, faulty drainage channel,

proximity to latrines and so on. These facility and environment related factors are the actual target of sanitary surveys, rather than point of use treatment suggestions such as chlorination.

For the 54 sources that are marked either unfit or medium-high risk, no other follow-up actions are recorded. Same-day FTK results are available for some of these sources, but they do not reflect actions on the risks factors. The current safety status of these risky sources is unknown.

- Quality/Accuracy

194 of the sources are tested by "Community Representatives" and similar concerns to FTK testing personnel are raised. Their credential and training may be crucial in determining the reliability of the results. In addition, a number of the inspections by the community show a "Low" risk level, yet the recommendations concluded that the "source is unfit." The blatant discrepancies further raise concern on the accuracy of their entries. Entries by "Water Authority" are more consistent, where all sources with "High" risk are also marked "unfit" under the recommendation column with "chlorination" as measure remark.

The "Sanitary Done By" column help discern inspections by community members from inspections by authorities, making it possible to weigh the credibility of the results differently.

On the other hand, 48 of the "Water Authority" entries have very abnormal "Recommendation" entries of random letters (e.g. "ljjhgg", "dvdfb"). All of these entries are from the same block – Kapadvanj block in Kheda district. These entries might be randomly typed in to satisfy non-null requirements of the "Recommendation" column, but they largely decrease the credibility of the sanitary survey results, even if they are conducted by water authorities.

- Integration viability:

All administrative entries are matched with entries in the Habitation Details table. All sources are matched with sources in one of the Source tables. One water supply scheme source in the Sanitary Survey Results table with the source ID 5324734 is matched with two separate supply scheme sources in the Source table. After observing the two schemes, it becomes clear that one of the schemes is a pipeline replacement of the previous scheme, and the two sources are the same.

The sanitary survey does not have a scheme name or scheme ID column for the scheme sources, and considering that the Water Supply Scheme Source table has a combined primary key of both Source and Scheme IDs, these table join issues may constantly occur during the process of data integration. They would take effort to fix manually.

Analysis on the data utility is summarized below:

- Acceptability

Compared to the FTK results, sanitary results are fewer and much less consistent. The random entries and the contradicting statements within the entries suggest a low level of motivation among the workers conducting sanitary inspections. 13 out of the 33 districts in Gujarat do not even have any records of sanitary survey results on IMIS. Interviews with WASMO suggest that more inspections are conducted but not uploaded on IMIS. The enforcement on inspection frequency and data entry is weak, and the motivation to collect data is negatively affected.

- Sensitivity

In comparison to water quality FTK and lab results, sanitary inspection is more effective at revealing potential risks which allows for proactive rather than reactive actions. It may reveal non-compliance in the drinking water facility before they manifest into water quality and health hazards, which is the ideal goal for risk assessments in outbreak control.

However, even though sanitary inspections are theoretically beneficial for outbreak prevention, their entries are much fewer compared to FTK and lab results. Only 755 (~14%) of the habitations and 965 sources (~1%) have sanitary survey records. As a consequence, the ability of these limited records to effectively reflect outbreak risks is very low.

- Predictive value

In the ideal case, sanitary survey records have good predictive values and can positively reveal concerns with the drinking water facilities, but for IMIS, the identified risk factors are not reported. In addition, the sanitary inspection procedure is more subjective compared to water quality tests, so the credential of the inspector is critical. Considering the quality of many of the records and the lack of detailed risk factors, the predictive value of the survey results is considerably lowered.

When only considering the risk level and recommendations in the sanitary survey results, the corresponding biological water quality results are shown in Table 5-13 and Table 5-14. The contamination ratio across different risk levels is plotted in Figure 5-11. As shown, water quality contamination is much more prevalent in sources that had higher risk levels during sanitary inspections. It is important to note the absolute number of sources with high and medium risk levels are significantly less than ones with low risk levels.

Fisher's Exact Test showed significance ($p<0.01$) for both contingency tables, suggesting that a significant relationship exists between the sanitary survey results and the actual status of contamination of the source. While this may suggest sanitary surveys can positively reflect water quality issues, this may also result from concurrent FTK tests and sanitary surveys, since they are frequently conducted on the same day, likely by similar WASMO personnel. The FTK results may affect the conclusions of the sanitary surveys.

Table 5-13: Source contamination and source sanitary risk level contingency table. Sources where any lab or FTK test show positive Total Coliform or E. coli over the course of the year are considered "contamination found." Sources where biological lab test and FTK test are conducted and showed negative for contamination are considered "contamination not found."

No. of sources	Sanitary Risk Level		
	Low/Very Low	Medium/Intermediate	High
Contamination not found	643	14	1
Contamination found	32	3	12
Contamination ratio	5%	18%	92%

Table 5-14: Source contamination and source recommendation contingency table. Source contamination definition same as above.

No. of sources	Sanitary Inspection Recommendation	
	Source unfit	Source free from sanitary risk
Contamination not found	4	625
Contamination found	16	30
Contamination ratio	80%	5%

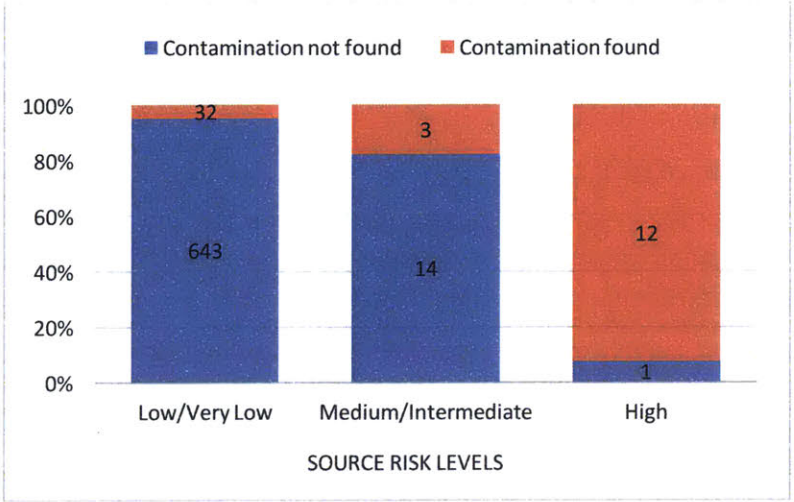


Figure 5-11: Source water quality contamination status percentage by source risk level. Note that only the contamination ratio is plotted.

Overall, predictive value of sanitary surveys can significantly improve if the data quality is higher and the exact risk factors of the water facility or surrounding environment are recorded. For the FTK tests that are conducted on the same day as sanitary surveys, it will be useful to note whether they are conducted independently from sanitary surveys. Without this information, it is challenging to analyze the predictive power of sanitary surveys results.

- Timeliness

Even though there is a recommended inspection frequency of twice a year, all sources are only inspected once throughout 2015-2016. The data are uploaded to IMIS on a regular basis as the surveys are conducted, but delay in entries are reported by WASMO. While the person carrying out the inspection may recognize and rectify risks on the spot, the information may not be timely reflected in the database and not immediately accessible to other agencies.

Overall, the timeliness of sanitary survey results is low.

Training Details

While training is only indirectly related to a better managed water system and safer water quality, it is still an integral section of the DPSEEA framework for outbreak risk assessment, so the Training Details table is still briefly evaluated for its data quality and utility.

The Training Details table records training participation at the level of each individual training sessions. Trainings are conducted at the district, block and gram panchayat level. Only one district-level training in Navsari was recorded among the 3 districts, where all blocks in Navsari are recorded as present. For the purpose of this study, only block and gram panchayat level trainings are considered, where gram panchayats are listed as the participants.

The variables included in this table include:

- Administrative region (gram panchayat-level)
- Training agency: the different types of agencies conducting the training, including ASHA worker, MPHW (Multi-purpose Health Workers), WASMO core team, gram agevan (similar to panchayat leaders) and CCDU/DWSM/WSSO.
- Training level: a column created after consolidating the block-level and panchayat-level training tables to indicate which level the training is conducted at.
- Participating member: the number of members trained. The block-level trainings only listed a total number of participating members, so the number for each panchayat is missing.
- Training type: the type of training is a general description of the training content. There is not a uniform format for this entry.
- Training month and year: The time that the training is conducted
- Number of trainees: the overall number of Grassroots Worker trainees and Coordinator Trainees within each gram panchayat. This variable is retrieved from the comprehensive progress report of each district rather than the training table, and then consolidated with the training information.

A general summary of the data evaluation results for the variables above is shown in Table 5-15.

Table 5-15: Evaluation summary for IMIS training details data

	Data Availability	Total Entries	Column	Data Type	Missing Data	Abnormal Data	Simplicity	Uniformity	Processing/Table Joins
E18/E20 - Training Details	2006-2007 until now	1966	Administrative region (GP)	Text	0	0	yes	yes	GP name extraction from text strings 100% matched to GPs in Habitation table
			Training Agency	Text: 5 categories	0	0	yes	no	Category consolidation

	Training level	Text: 2 categories	0	0	-	-	New column created
	Participating member	integer	1620	0	yes	yes	
	Training type	Text	0	0	-	no	
	Training month	Text/ month	0	0	yes	yes	
	Number of total trainees	Integer	1204	25	yes	yes	Retrieved from E20

Analysis on the data characteristics is summarized below:

- Accessibility

The training data are generally accessible through IMIS. The training details are documented by each single training session where all participating panchayats are listed in a text string. Their names have to be extracted so that the training details can be converted to a panchayat-level format to align with the rest of the water source and quality data. It would be more efficient to have a uniform data reporting unit.

The training data dates back to 2006-2007. Similar to sanitary survey reports, few states were reporting data on block and panchayat level training back then. Even for 2015-2016, only around half of the states have a reasonable number of trainings reported.

- Simplicity

There is no definition on the training type. A variety of different details are reported as a result, including the location of the training, the content of the training or a general overarching topic such as "introduction meeting." Considering that some of the training are focused on collecting contributions from the community or general auditing, having a clear and simple definition of training types would be helpful in determining whether the training is related to water management and outbreak control.

- Uniformity

Due to the lack of definitions for training types, there is no consistency in the entries.

For the 5 different training agencies on recorded, the entries are also slightly inconsistent where the same agency name can be capitalized differently. These categories require consolidation.

- Completeness

As mentioned before, 1620 of the entries are from block-level trainings and thus lack the participating member count. Overall, the training details are relatively complete. All 1413 gram panchayats (as of 2015-2016) have at least one record of training participation, except Vasana panchayat from Kapadvanj block and Kheda district.

Many of the trainee entries from E20 are missing some administrative-level information, similar to the case of FTK results from E20 – the same IMIS dataset. Many of these are filled by matching up with the Habitation Details table. However, 25 of them are actually missing the gram panchayat name,

which cannot be filled. Compared to the mostly complete records on training participation, only 305 gram panchayats (~22%) had information on the number of trainees.

- Quality/Accuracy

There is not much room for errors in the training detail records, and most entries are consistent. There are missing data for the "Number of Trainees" variables, which is likely due to formatting issues with the tables of E20. The quality is not of too much concern with this table.

- Integration viability

All entries in the Training Details table are matched with the gram panchayats listed in the Habitation Details table.

Considering that this table is generally not reflective of immediate violations or risks for outbreaks, the data are not evaluated on their utility. Overall, it is a good reflection of training participation and absence. However, the ambiguity of the training content makes it difficult to recognize trainings on outbreak control related topics, which decreases the overall predictive value of the training details in the table.

5.2 Sanitation database

5.2.1 Database background

SBM-G Management Information System (Figure 5-12), a comprehensive web-based information system enabling monitoring of latrine coverage at the center, state, district, block and panchayat level, was launched on Oct 2, 2014. The database contains household sanitation details of around 18 crore rural families across more than 16 lakh habitations, 6 lakh villages and 2.5 lakh gram panchayats (MDWS 2016). Mobile collection and communication methods are also developed for uploading location and photos of latrines, as well as for SMS communication with beneficiaries on their satisfaction with toilets provided to them under SBM-G (MDWS 2016).

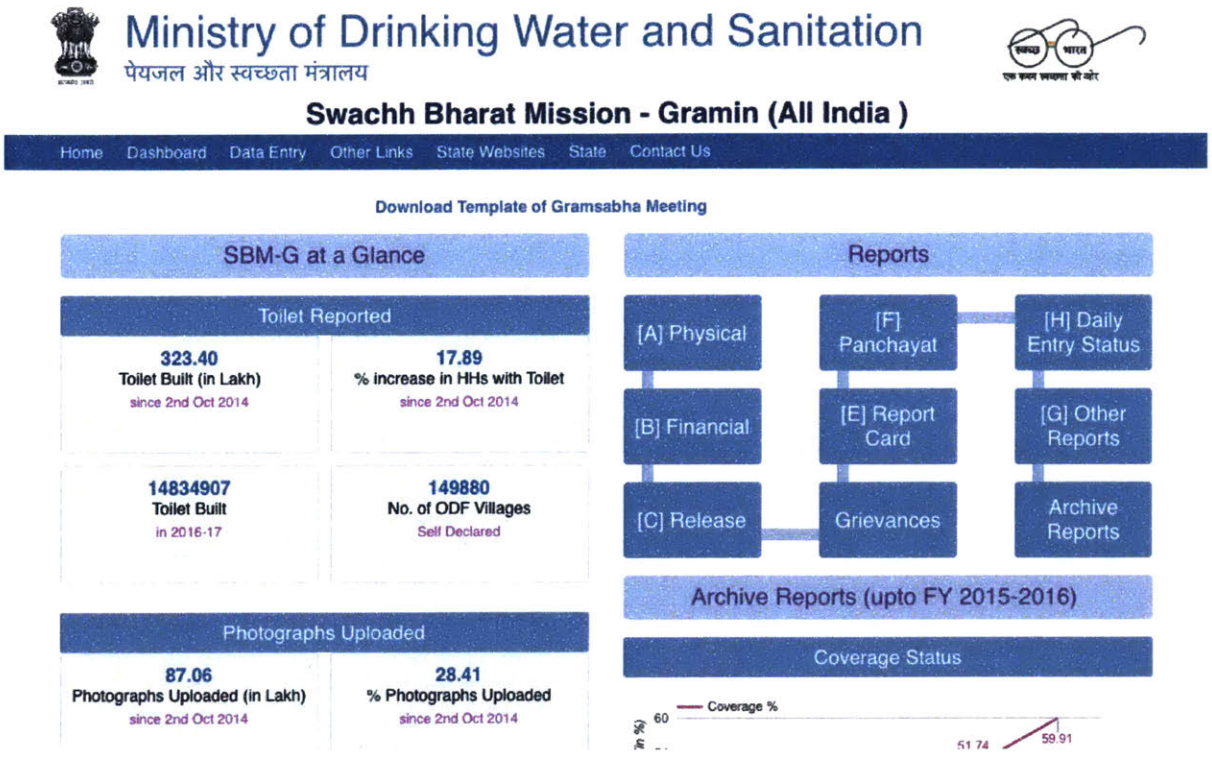


Figure 5-12: Demonstration of the front page for the SBM database

The 2012-2013 Baseline Survey, originally residing in the IMIS database before the SBM-MIS database was constructed, functions as a progress reference. States are permitted to update the Baseline Survey once a year in the month of March-April if additional household details become available. The main focus of the monitoring system is documenting toilet construction, usage with the goal of creating ODF communities, which is a separate report created in SBM that was not available for the Baseline Survey.

Physical and financial progress reports of implementation are reported monthly through the online SBM-MIS database. GP-wise physical and financial progress should be entered with photographs of the toilets by the 10th of the following month by block or district level sanitation missions. The information entered at the block and district level has to then be approved by the state by the 15th of each month before the results are sent forth and finalized by the MDWS (MDWS 2016).

In comparison to IMIS where almost no data analysis on the water quality has been done, graphic displays and target reporting are displayed prominently on the front page of the SBM-G database. Key figures include toilet construction numbers, toilet photograph verification numbers and ODF village numbers, which again demonstrates SBM's strong focus on physical progress.

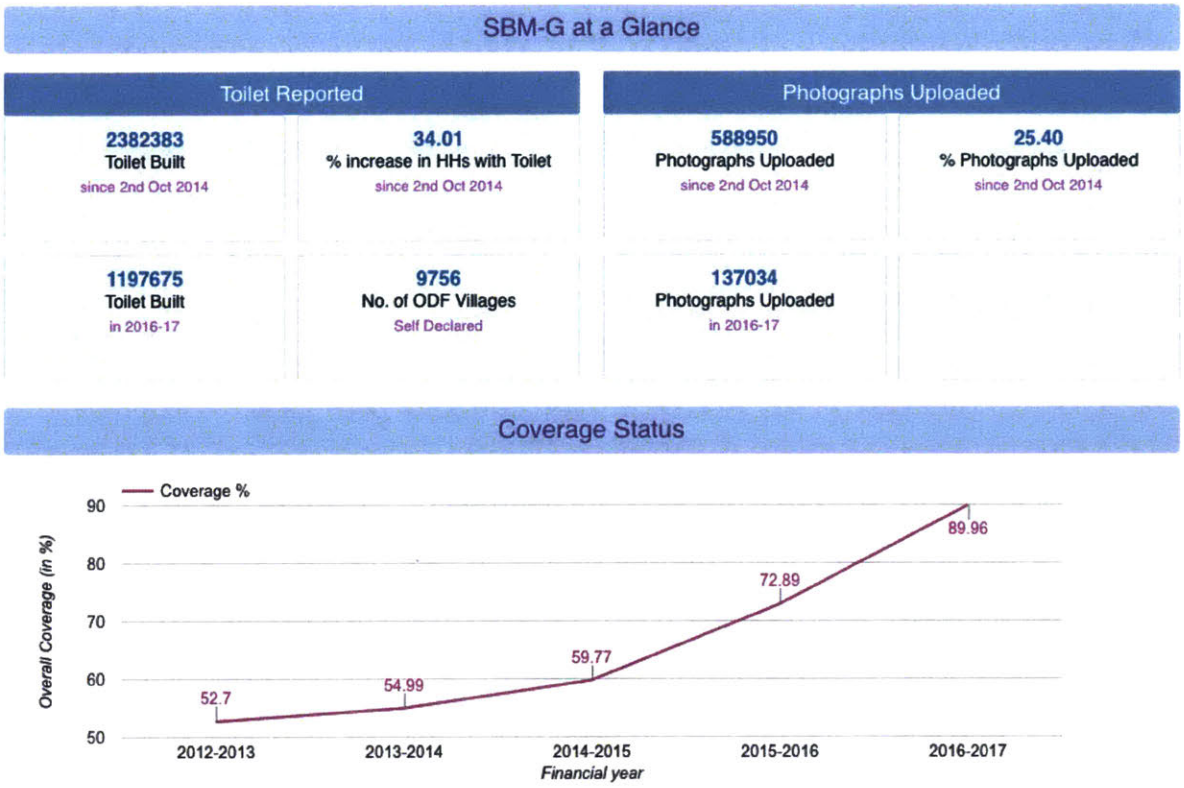


Figure 5-13: A visual representation of Gujarat's SBM progress available on SBM website

5.2.2 Variable summaries

All variables in the SBM database are related to individual household latrine construction progress and ODF declarations. The SBM database is also a basis for funding distribution for the support of toilet construction, so financial progress reporting is also a major component of SBM. For the purpose of creating an integrated approach for outbreak control, this study only focuses on the basic physical progress of latrine construction and ODF verification. While ODF declaration is a reflection of 100% toilet construction, the verification process, as outlined in Chapter 4, takes into account many additional factors and is more rigorous than simply observing the physical latrine structures. Thus, the ODF status and latrine construction status would be considered as two separate variables.

The variables of interest are listed in Table 5-16.

Table 5-16: List of relevant SBM variables for the integrated database

Table Description	Relevant columns of variables	Location in SBM database
IHHL installation (GP-level)	Administrative region (panchayat-level) Total details entered in baseline survey Total APL/BPL households	Format A3

	Additional number of households covered each year after baseline year	
IHHL installation (household-level)	Administrative region (habitation-level) Household information (benefit ID, family head and gender, father or husband name, ID card types and numbers) Household category (APL/BPL) Sub-Category Have toilet or not BLS-2012-only descriptions: Toilet construction scheme (NBA/non-NBA) Toilet functional or not Toilet used or not Water facility availability Monthly Progress Report (MPR) status	Format A3
ODF status	Administrative region (village-level) ODF declaration status and date ODF verification status Total number of household details Total number of households with toilet Number of households accessing community or other toilets Remaining number of households to be covered Coverage percentage	Format F42
Toilet details and Photographs	Administrative region (habitation-level) Household information (same as the records above) Total toilet cost MPR approval status by district and state GPS location (latitude, longitude) Image of beneficiary and toilet Detailed location on map Uploaded data	Format F28 A

These tables are all connected with each other through the schema shown in Figure 5-14. The top row is again the relevant segment from the administrative-level schema in Figure 5-1. Latrine construction status are reported at the household level. For latrine systems constructed before the Baseline Survey in 2012 (BLS-2012), details including toilet functionality, toilet usage and water facility are also reported as part of the baseline survey. For new latrine systems constructed after Oct 2014, which marks the commencement of the reconstructed SBM campaign and the establishment of the SBM database, these toilet details are no longer available, but new variables such as toilet construction dates, GPS location, images of the latrines and the total costs are reported at the household level.

The household details are also aggregated at the panchayat level, which creates the physical progress report of each panchayat. At the panchayat level, household coverage and latrine photo collection progress are reported for each year after the baseline year.

The ODF status, including both declaration and verification, is reported at the village level and GP level. We are only considering village level reports to ensure higher granularity.

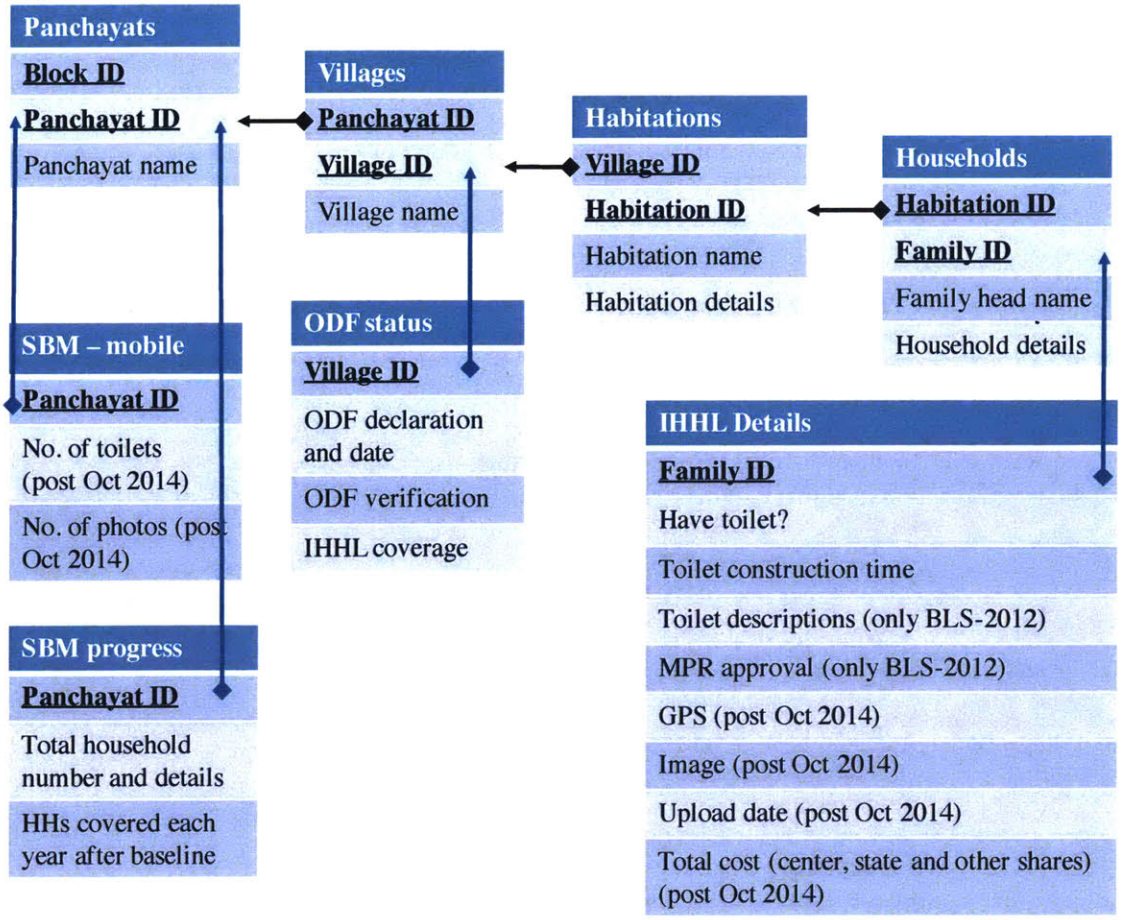


Figure 5-14: Schema connecting tables of interest within the SBM database

We go on to individually evaluate each table for their cost of integration, as outlined in Chapter 3.

Household IHHL Details

2015-2016 archive data for IHHL (individual household latrine) status are available at the household level. Considering that the panchayat-level SBM progress is only an aggregation of the household data, we will only evaluate the household IHHL Details table. Some of the details, such as GPS, image, construction dates and cost shares are only available for newly constructed latrines. In addition, the collection rate for these variables, especially the image and GPS location, is only around 25%. Hence, these variables are not included in the evaluation, but would be useful to consider in the future when completion rates are higher.

The IHHL Details table include house details such as basic household identification information and characteristics that are related to IHHL funding support, and latrine ownership details.

Variables for this table include:

- Administrative region (habitation level)
- Household ID: two different IDs are listed for each household, including a beneficiary ID and

a Card Number (the card type is listed, which is typically Ration Card or APL/BPL card). There are also other identification details of the household including the family head, the husband or father of the family head and so on, which are considered similar to the household ID and not included in the analysis.

- Category: whether the household is considered above poverty level (APL) or below poverty level (BPL).
- Sub-category: 8 categories including small marginal farmers, landless with homestead, women headed HH, SC, ST, physically handicapped, general and others. These categories are defined in the SBM guidelines and determine the level of incentives that households can receive for latrine construction.
- Have toilet: whether the household has a toilet available. Toilets constructed before the 2012 baseline survey are noted separately.
- Toilet access from: date that toilet become available to the household. While this data is available, it is not listed in the aggregated IHHL details and requires clicking on the profile of each toilet to retrieve. This information is only available for toilets constructed after the baseline survey.
- The following columns are available only in the baseline survey. Hence, the variables are only available for toilets that are constructed before BLS-2012. For the new toilets, these functionality and usage information is not recorded in SBM.
- Toilet constructed from: the initiative that supported the toilet construction, such as NBA (Nirmal Bharat Abhiyan, later renamed as SBM) or MGNREGA (Mahatma Gandhi National Rural Employment Gurantee Act). These are the separate programs that funded construction of toilets before the creation of SBM.
- Have functional toilet: for households that have a toilet, whether it is functional
- Used functional toilet: for households that have a functional toilet, whether it is used
- Water facility availability: for households that have a toilet, is there also water facilities that are available for the toilet. The exact definition of "water facility" is unclear. Many of the toilets are still marked as "functional" despite the lack of water facilities.
- MPR report: whether monthly progress reports are delivered.

SBM	Data Availability	Total Entries	Column	Data Type	Missing Data	Abnormal Data	Simplicity	Uniformity
SBM A3 IHHL Details	GP-level summary: 2012 - now Household Details: 2015-2016 until now	736111 (17185 missing entries)	Administrative region	text	0	0	yes	yes
			Household ID	Integer	0	0	yes	yes
			Category (APL, BPL)	Text: 2 categories	0	0	yes	yes
			Sub-category	Text: 8 categories & NA	6	0	no	yes

	Toilet Constructed From (2012 only)	Text: 4 categories & NA	0	0	yes	yes
	Have toilet or not	Boolean	0	0	yes	yes
	Have functional toilet (2012 only)	Boolean	0	0	yes	yes
	Used functional toilet (2012 only)	Boolean	0	1	yes	yes
	Water facility available for toilet (2012 only)	Boolean	0	-	no	yes
	MPR Reported or not (2012 only)	Boolean	0	0	yes	yes

Analysis on the data characteristics is carried out below.

- Accessibility

The IHHL details are accessible only at the gram panchayat level. This requires aggregating gram panchayat data to the blocks and aggregating further to each district.

In addition, the data downloaded from SBM website are data from BLS-2012, rather than the current IHHL data displayed on the webpage. Hence, new toilet coverage data can only be obtained through scraping the webpage.

The household level details contain the most up-to-date information on all toilets structures regardless of the time of construction. There are only two history snapshot data available – one from BLS-2012, and one at the end of 2015-2016. On the other hand, at the gram panchayat level, information on new construction of toilets across each year can be accessed starting from 2013-2014, so it is possible to recreate these snapshot data for the past years. Compared to BLS-2012, datasets for these newly constructed toilets include extra columns of data on the toilet cost and distribution of cost shares, but do not have information on toilet functionality, usage or water facility availability.

- Simplicity

The definition of the columns is generally straightforward.

The subcategories in SBM are defined solely based on SBM funding priorities, and while each household may fit multiple criteria, only one category is recorded. However, considering that these categories also partially overlap with the more standard SC/ST/General characterization in IMIS Habitation Information table, allowing for each household to identify under multiple categories may generate more consistent results across SBM and IMIS and potentially other databases.

Additionally, it is unclear if the water facility is referring to flushing system for the latrines, handwashing stations, or any general types of water source. This may be a key link between the SBM and the water sources information in IMIS. However, the vague definition weakens this potential connection.

- Uniformity

Apart from the variable discrepancy between BLS-2012 and current IHHL details table, the data are generally in consistent formats. The categories are well defined and text entries for each category are uniform.

- Completeness

The data on toilet access time is not incorporated into the IHHL details table - they are only accessible at each IHHL level.

6 households had "-1" as their subcategory entry, likely suggesting missing data.

BLS-2012 had more information on the status of the toilets after they constructed, but these details are no longer reported for new toilets. The new data table focuses primarily on the absence/presence of toilets and their costs. As more physical progress and ODF status is achieved, follow-up surveys on the continued usage and functionality of these toilets may be conducted.

In addition, compared to the total household number recorded in the 2012 Baseline Survey, IHHL details are still missing for 17185 households (~2%). The general latrine ownership status may be known for these households, but no record is available in the IHHL Details table.

- Quality/Accuracy

For BLS-2012 data, many of the column entries are conditional upon other columns. For example, water facility availability and functionality entries are conditioned upon the ownership of a toilet, and usage of the toilet should only be possible with a functional toilet. Considering these constraints, the data entries are still quite consistent. There is only one questionable entry (Beneficiary ID: 128390871) where toilet usage is recorded as "yes" while its functionality reports "no." In addition, 53504 households reported a functional toilet even though the water facility availability is reported as "no." However, due to the lack of definition on "water facility," these 53504 cases might not be erroneous. For example, it is possible for a toilet to be functional without a water facility for handwashing.

Additionally, during 2015-2016 – the beginning of the SBM database setup, a lot of the data are still under reconciliation. 7 districts in Gujarat did not have 2015-2016 archive data. Much of the rest of the data was adapted from BLS-2012, and there are significant changes to the names and jurisdiction after 2015-2016 - even state names have been revised, such as changing from "Mehsana" state to "Mahesana" state. In the 2016-17 records for Kheda district, Balasinor and Virpur blocks from 15-16 records are no longer reported and Galteshwar and Vaso appeared. In fact, both Balasinor and Virpur are now actually blocks under Mahisagar district. While these adjustments do not indicate issues with the accuracy of IHHL data, they do reflect the outdated 15-16 administrative region information, which raise significant challenges during WaSH-health data integration process.

- Integration viability

The challenges with integration between the IHHL details table and the village-level ODF status table are reflections of the data reconciliation process. ODF status data is only available in the up-to-date version, so the administrative region details follow the 2016-2017 version, which had significant revisions from 2015-2016. Out of the 1648 villages reported in the IHHL Details table, 159 are unable to match to any ODF status entries. The majority of this discrepancy is because of the removal of Balasinor block and Virpur block from Kheda District. In addition, a number of panchayats/villages have been moved to another block within the same district or had revisions in the panchayat name,

as shown in Table 5-4. The Sachin panchayat (Chorasi block, Surat district), however, is unable to match to any new panchayats, and none of the households in this panchayat can be found under any new region within the 3 districts. After these adjustments, all but 1 (Sachin village, the only village in Sachin panchayat) of the 159 villages in the Habitation Details table can be matched with ODF status.

Table 5-17: Administrative region adjustment from 2015-16 to 2016-17

District	Original	2016-17 Revisions
Kheda	Matar (block)	Vaso
	Nadiad (block)	Vaso
	Thasra (block)	Galteshwar
Navsari	Chikhali (block)	Khergam
Surat	Pardi Koba (panchayat)	Sayan
	Sachin (panchayat)	-

Analysis on the data utility is shown below.

- Acceptability

As the SBM initiative is currently the priority in the WaSH landscape of India and funding support is closely contingent upon valid SBM data reporting, there is relatively high motivation for local agencies as well as local panchayat leaders to carry out SBM monitoring. SBM is highly target driven, which depends on consistence and high quality in the data collection process. This is demonstrated by the uniform data reporting formats and the rigorous data reconciliation process.

On the other hand, data that are important but not essential to SBM physical progress, such as toilet images and GPS location, are reported much less consistently, with only around 25% completion rate across Gujarat. Overall, the willingness for data collection is highly dependent on the SBM target, which raises concern on post-ODF data collection motivation.

- Sensitivity

Sanitation data can readily reflect the lack of toilets and the likelihood of open defecation possibility. The target-driven nature of SBM allows easy identification of regions that are prone to sanitation risks through IHHL details data. Trends in sanitation status and unsatisfactory levels of toilet coverage are easily observed.

However, if we are considering the ultimate goal of outbreak control, sanitation data are less sensitive to immediate risks for waterborne diseases compared to water quality information. If we consider the DPSEEA framework, sanitation is identified as a "pressure" factor that modified the water environment, resulting in changed "states" of the environment that lead to potential "health effects" through "exposure". Hence, sanitation is indirectly connected to the ultimate health outcome via water quality states. Direct health-related violations may not be immediately reflected in sanitation results.

- Predictive value

The low toilet coverage rate is a clear reflection of sanitation risk due to slow physical progress. While the BLS-2012 data showed behavioral information, the current IHHL only contains toilet construction status. Any other behavior related health risks, such as lack of toilet usage, are not reflected.

Theoretically, IHHL coverage is also related to water quality status, especially biological contamination of water quality due to open defecation. As shown in Figure 5-15, while it is counterintuitive that habitation with positive water contamination results is showing a slightly higher IHHL coverage level, t-test shows that the difference in average values is insignificant. Logistic regression also shows no significance of the IHHL coverage percentage as a predictor for water quality contamination. Hence, IHHL coverage alone cannot effectively predict water quality contamination.

This may be related to the lack of hygiene and behavioral data, which also largely affects water quality performance. Other factors, such as the solid and liquid waste management from the latrines and proximity of sewer to water sources, can also strongly influence water quality. The coverage of physical IHHL alone can hint at the sanitary risk factors, but may not directly reflect key risks in the water and health aspects.

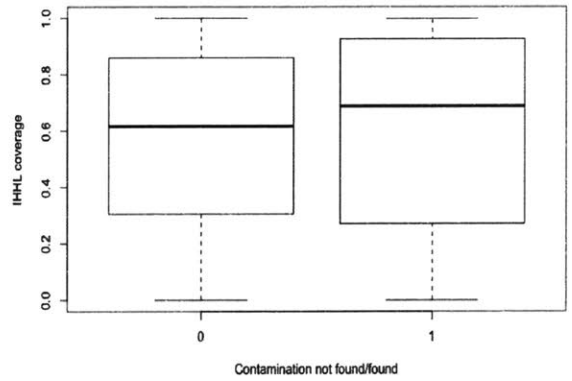


Figure 5-15: Box plot of habitation latrine coverage across different water quality status. The plot excluded all the households that already achieved 100% IHHL coverage. Habitations where any sources have been tested positive for biological contamination, whether via FTK or lab tests, are considered habitations where "contamination found." Habitations where all sources that have been tested for biological contamination showed no positive records are considered habitations where "contamination not found." Habitations where no sources have been tested for biological contaminants are not considered. The same definition is applied to all following analysis.

- Timeliness

IHHL data are reported to the block-level, and confirmed in batch. The target of 100% IHHL overall drives a quick data entry process, so lack of sanitation progress can be easily spotted through updates on SBM. However, with no columns of data on toilet functionality or water facilities for the new 2015-2016 entries, all information after toilet construction is missing. Issues with dysfunctional toilets and lack of usage, cannot be timely reflected in the IHHL data records, although they may still be uncovered during the final ODF status inspection.

ODF status

ODF status is a confirmation of 100% IHHL construction and an overall declaration of clean and sanitary behaviors. It is carried out at the village and panchayat level. This is a new initiative under

SBM, and villages only started declaring ODF status since the end of 2015. As mentioned in Chapter 4, ODF is first declared by the panchayat leaders, and then verified through a cross-block inspection process.

In the ODF status table, the following details are reported:

- Administrative region (village-level)
- ODF declaration: whether the village has declared ODF status or not
- ODF declaration date: the date that ODF status is declared by the village
- ODF verification: whether the village has passed the ODF verification process after its declaration

A general summary of the data evaluation results is shown in Table 5-18.

Table 5-18: Evaluation summary for SBM ODF status data

	Data Availability	Total Entries	Column	Data Type	Missing Data	Abnormal Data	Simplicity	Uniformity	Processing/ Table Joins
SBM F42 ODF status	current version	1590	Administrative region (village)	text	0	3	repeated		30 unmatched to villages IHHL details table
			ODF declaration	Boolean	0	0	Yes	Yes	
			ODF verification	Boolean	0	0	Yes	Yes	
			ODF declaration date	date	0	0	Yes	Yes	Use the date to pick out 2015-16 status

Analysis on the data characteristics is summarized below:

- Accessibility

ODF status of villages are accessible at the block level, and can be downloaded block by block from the website. ODF status is only reported in the most updated format. It is possible to recreate past year snapshots using the ODF declaration date, but without the ODF verification date, this status may lack accuracy.

- Simplicity

While the ODF verification criteria lacks a more uniform and clear standard, the definition of the variables in this status is overall quite simple and straightforward.

- Uniformity

The variables are reported consistently in Boolean values and data values.

- Completeness

There are no verification dates. For ongoing verification or villages that have failed the verification process, there are no columns for status updates. In addition, after ODF has been verified, the 6-month cross-district validation records are not reported, and the sustainability of the ODF status is unclear. Concerns over ODF sustainability and post-ODF plans are raised by UNICEF consultants through interviews. In Chapter 4, we identified DRDA from ODF districts as high-potential partners for the development of the WaSH integrated system. However, with the lack of post-ODF data reporting structures in the SBM database, the interest from DRDA and other SBM-related agencies to continue WaSH data collection may diminish.

- Quality/Accuracy

To begin with, the ODF self-declaration variable is not a reliable reflection of cleanliness achievement. The declaration is generally just a reiteration of the 100% IHHL coverage, but to verify ODF, a lot more behavioral observations across the village are conducted. According to UNICEF consultants, almost 30% self-declared GPs could not pass the cross-verification test. The ODF verification variable would be much more reliable as a confirmation of status.

While the data records in the ODF Status table is overall consistently and straightforward, a few issues still hint at data quality challenges. For example, there are 3 duplicate village entries for Navsari district as shown in Table 5-19, with two of them having conflicting ODF status records.

Table 5-19: Duplicate records for ODF declaration status

District	Block	Panchayat	Village	ODF status Record 1	ODF status Record 2
Navsari	Khergam	Achhavani	Achhavani	Declared (Apr 2, 16) and verified	Not declared
Navsari	Gandevi	Bigri	Bigri	Declared (Apr 2, 16) and not verified	
Navsari	Gandevi	Taliyara	Taliyara	Declared (Apr 2, 16) and verified	Not declared

There are also 156 villages that have declared ODF before 2016-2-1 (around a year ago as of February 2017) but still has a no records of ODF verification, which is generally expected to be carried out within 3 month of ODF declaration. Without a status update, it is unclear if these villages simply failed the verification and is waiting for the next round, if the verification is still yet to happen, or if the data is simply outdated.

The ODF verification process is also considerably subjective. Non-SBM staff are expected carry out the verification to eliminate conflict of interest, but with a limited comprehension of the sanitation criteria, the data quality maybe affected.

- Integration viability

30 villages in the ODF status table are not matched to villages in 2015-2016 IHHL Details Table. More IHHL details became available over 2016-2017 – but even considering these updated habitation information, there are still 26 villages with ODF information but no habitation details entered. Among these 26 villages, 8 of them still declared ODF without any IHHL details entry, although none of them are verified. This goes on to further question the validity of the ODF self-declaration status.

Analysis on the data utility is shown below.

- Acceptability

Considering that 100% ODF is a key target under the SBM initiative, there is a strong incentive for reporting ODF declaration. However, there seems to be delays on the verification side, considering that it requires cross-block engagement of non-SBM personnel who may not have strong incentive to conduct the assessment routinely and timely. In addition, as shown in Figure 5-13, the SBM progress report is focusing on the number of self-declared ODF panchayats and villages. Considering that the declaration is a much less accurate reflection of the village status compared to verified results, more efforts should be placed on increasing motivation for the cross-verification process.

- Sensitivity

Compared to IHHL Details table, sanitary risks due to the lack of latrines are not as effectively reflected via ODF status. ODF declaration is only a binary variable at the end of 100% IHHL construction. It does not reflect progress or households of concern in the granularity that IHHL information can. However, ODF status can reflect additional information on sanitary practices and behavioral risks. This is especially true for villages that declared ODF but have not received verification after the 3-month period. For these villages, since 100% IHHL construction have usually been achieved, there are clearly still risky practices raising sanitation concerns. An ODF verification status reflects more sanitary progress and behavioral change than IHHL physical progress.

Similar to IHHL details, ODF status, as a sanitation variable, is not as sensitive to direct waterborne outbreak concerns compared to direct water quality results.

- Predicative value

Village without ODF declaration is generally reflecting the lack of IHHL completion. Village that declared ODF but without timely verification is reflecting sanitation violations during the verification inspection process, but with no details or identified risk factors, the exact violations preventing ODF verification are unknown.

Nevertheless, a lack of ODF status can still indicate sanitation risks and further water quality and health risks. As Table 5-20 and Figure 5-16 show, water quality biological contamination status varies across habitations under different ODF declaration categories of the village that the habitations belong to.

Habitations in villages that have not declared ODF shows a higher chance of contamination, while habitations in villages with ODF verification shows the lowest. Both Chi-square and Fisher's Exact Test shows significance ($p<0.01$) for the contingency table, suggesting that a significant relationship exists between the ODF categorization and contamination status of habitations. It is important to note that much fewer habitations are in the declared or verified ODF category compared to the numbers that have not declared ODF, so the expected number of contamination found in the verified ODF category may be relatively low and this may violate the goodness of fit test assumptions. Additionally, the habitations selected for this analysis are not selected through random sampling – rather, we only take into account habitations that have full sets of data across water quality and sanitation. This

intentional sampling can affect the results. In addition, not all sources in habitations are tested and it is unclear how sampling of sources is determined at the local level.

Table 5-20: Contingency table between ODF status and water quality contamination status. Contamination status for habitations is similarly defined as before.

# of habitations	verified ODF	declared ODF but not verified	not declared ODF
Contamination not found	272	724	1930
Contamination found	9	54	286
contamination ratio	3%	7%	13%

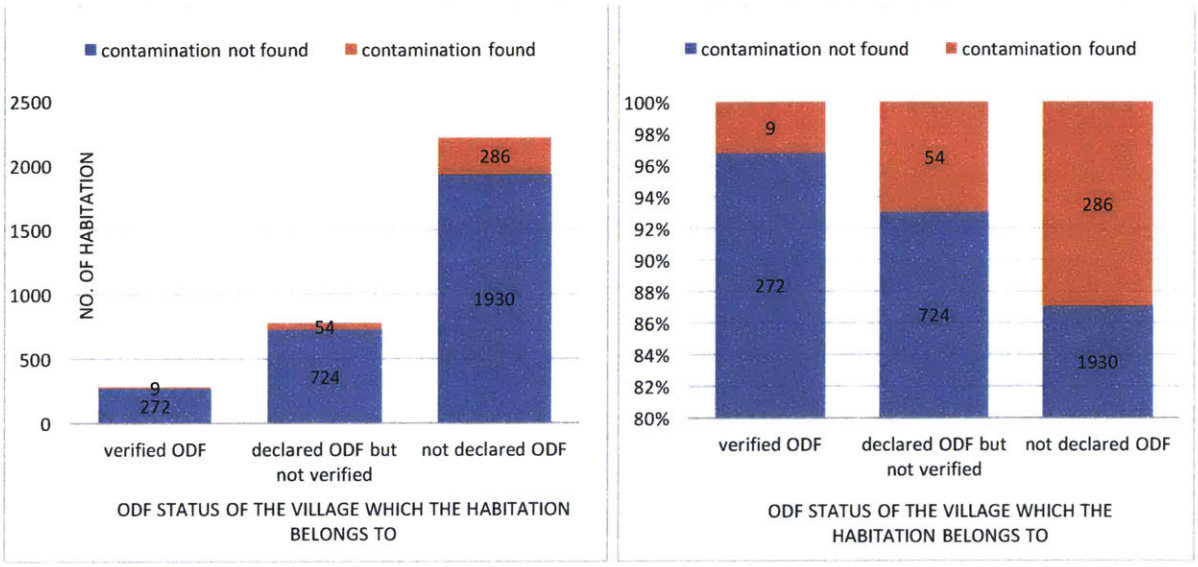


Figure 5-16: Habitation water quality status by source. Both the absolute numbers of habitations (left) and the ratio of habitations (right) are plotted.

- Timeliness

The declaration of ODF is likely reported in a timely fashion, considering the motivation to achieve 100% ODF as soon as possible. However, as indicated in analysis above, ODF verification may not be conducted within the allotted time. Delays in the verification process, coupled with the uncertain state of villages that failed verification, the ODF status table may not be able to timely and accurately identify sanitation concerns.

5.3 Disease and outbreak database

5.3.1 Database background

The Integrated Disease Surveillance Project/Programme (IDSP) was launched by the Minister of Health & Family Welfare in Nov 2004. The early version was in operation until March 2010, after which a major restructure created the IDSP as it is presented now (Figure 5-17). The mission of such a database is to maintain a decentralized state-based disease surveillance system for epidemic-prone diseases to detect early warning signs so that immediate public health actions can be carried out to effectively control health challenges (IDSP, no date a).

Figure 5-17: Screenshot of the IDSP database webpage

IDSP collects information for disease surveillance purpose. Diseases or syndromes that are potentially related to outbreaks are reported at 3 different levels through S/P/L forms, and reported in the IDSP database the Reporting Unit – which is generally a health center, hospital or surveillance unit. Data can then be aggregated at the block level to show the number of different diseases reported via each of the 3 formats. The data reporting for IDSP is done weekly on a Monday-Sunday schedule, and the Data Manager at the district and state surveillance units monitors the reporting from the sub-centers up to the District Surveillance Unit as outlined in Figure 5-18.

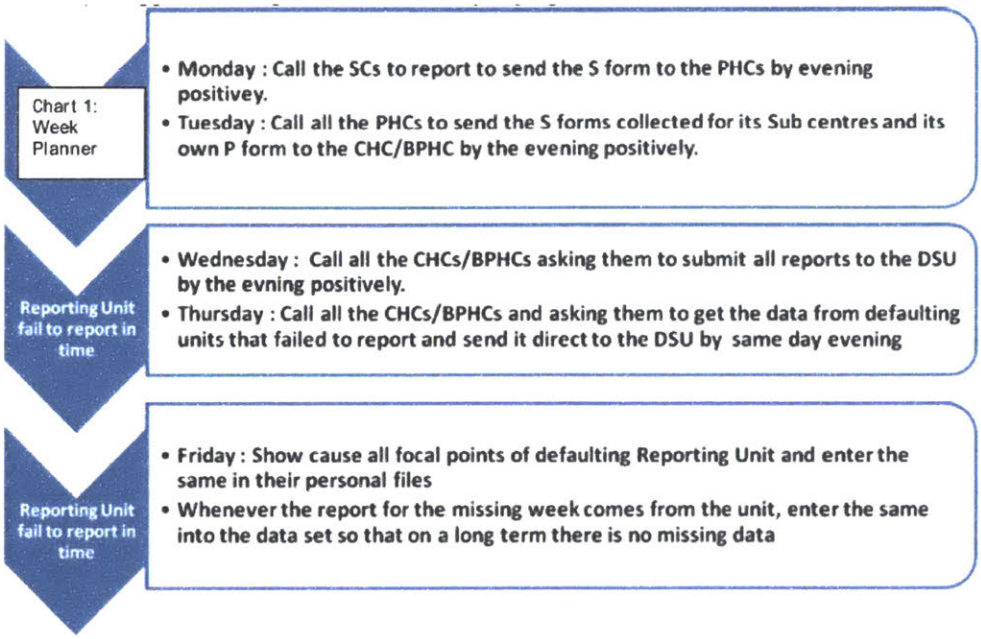


Figure 5-18: Weekly reporting schedule for the Data Manager (IDSP, no date b)

Compared to IMIS and SBM, IDSP has an additional function for graphical outputs generation, such as the cumulative disease records across the years (Figure 5-19). The graphical outputs assist in recognition of trends or abnormality in the disease data to detect potential risks in the system.

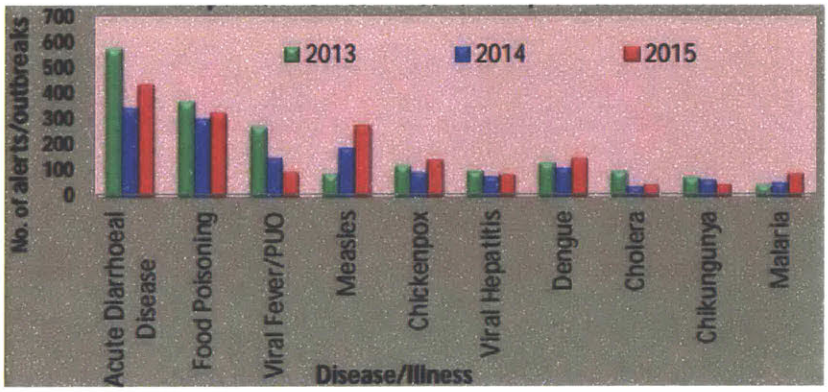


Figure 5-19: Comparison of cumulative no. of disease alerts and outbreaks reported in 2013, 2014 and 2015 (IDSP, 2015)

5.3.2 Variable summaries

Through syndromic surveillance from the health workers, probable surveillance from clinical facilities and laboratory surveillance from government and private labs, cases are detected and reported through S/P/L forms at all different levels of Reporting Units. Early warning signs and instantaneous reporting of outbreaks, as well as outbreak summary reporting, are conducted through the District Surveillance Units or State Surveillance Units.

Symptoms, diseases and outbreak occurrence are at the end of the DPSEEA chain, as an "effect" caused by changing factors of sanitation "pressures", water quality "states" and other factors. The risk to health and outbreaks is the ultimate variable we are interested in predicting through the WaSH integrated approach.

The list of relevant variables of interest and tables that they belong to are shown in Table 5-21.

Table 5-21: List of relevant IDSP variables for the integrated database

Data Table Description	Variables	Location in database
Master Data	Administrative region (block-level) Reporting Units Population	SPL form master data
Form S/P/L	Report Unit Name of Health Worker/Volunteer/Practitioner and Supervisor (S only) Officer-in-charge (L and P only) Name, Reporting Week, Date of Reporting S cases P cases L cases + positive case description	SPL forms
Early warning signals	Administrative region (block-level) Diseases/syndromes Areas affected No. of cases No. of deaths Date of start of the outbreak Total population of affected area Epidemiological observations Lab results Control measures undertaken Present status	Outbreak forms
Outbreak record details	State District Outbreak reference number Outbreak date Outbreak number Outbreak details	Outbreak forms

We were unable to obtain data from S/P/L forms because the data from IDSP is not publicly accessible. Data requests are sent through the health departments and they are still under processing. Hence, we are only able to analyze the data based on the templates as outlined in the IDSP data operator manuals (IDSP, no date c).

While early warning signals of outbreak are reported in IDSP, the exact details are not outlined as clearly in the data manuals compared to the SPL forms. There is insufficient data for analysis. For outbreak records, although the raw data are inaccessible, a weekly compiled report is available on the website in PDF format. The analysis is based on the outbreak records online.

The SPL forms table and the outbreak records table are all joined through the following schema in Figure 5-20. The top row is the relevant schema from the administrative-level schema. While it is possible to interpret the address of the reporting units, reported cases to get a higher granularity of health-related information, the raw data are only aggregated at the block-level through all the different reporting units under the block.

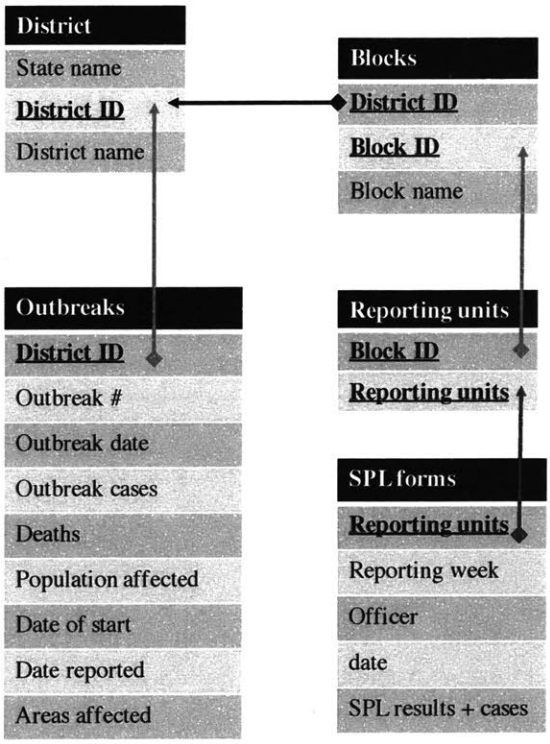


Figure 5-20: Schema connecting tables of interest within the IDSP database

Master data and SPL Form

The master data contains information on districts, blocks and each of the reporting units where diseases and symptoms data are collected. Details include the following:

- Administrative region (block-level)
- Reporting unit: the units reporting outbreak-prone disease and symptom data. The types of reporting units and the forms that they report are listed in Table 5-22.
- Population: only sub-center (SC) type of reporting units can report the number of location population

Table 5-22: Type of Reporting Units (IDSP, no date c)

S. NO	Reporting Unit Type	RU Type Code	Forms Type	Focal Point
1	SC/HSC	SC	S	Health Workers
2	HC/Addl PHC/New PHC	HC	P	Medical Officer
3	CHC/Rural Hospitals	CH	P	
4	Infectitious Disease Hospital (IDH)	ID	P	
5	Govt. Hospital / Medical College*	MH	P	
6	Private Health Center/ Private Practitioners	PC	P	
7	Private Hospitals*	PH	P	
8	Private Labs	PL	L	Lab Incharge
9	Government Laboratories	GL	L	
10	Private Hospitals(Lab.)	LPH	L	
11	CHC/Rural Hospitals(Lab.)	LCH	L	
12	HC/Addl PHC/New PHC(Lab.)	LHC	L	
13	Infectitious Disease Hospital (IDH)(Lab.)	LID	L	
14	Govt. Hospital / Medical College(Lab.)	LMH	L	
15	Private Health Center/ Private Practitioners(Lab.)	LPC	L	

The following basic information are reported across all the SPL forms:

- Name of health worker and supervisor (S form)
- Name of officer-in-charge (L/P form)
- Reporting week and date range: IDSP data are reported on a weekly basis. The year starts in January, as opposed to April 1 in the IMIS and SBM system.
- Date of reporting

Form S:

Key syndromes reported in syndromic surveillance include:

- Fever (< 7 days; > 7 days; with additional symptoms)
- Cough (< 3 weeks; > 3 weeks)
- Loose water stools (with additional symptoms)
- Jaundice case
- Acute flaccid paralysis
- Other unusual symptoms leading to death/hospitalization

The following details are reported for each of the syndromes:

- Number of cases (categorized by female/male and <5 years old or >5 years old)
- Number of deaths (categorized by female/male and <5 years old or >5 years old)

The workers collect these data based on symptoms reported by the patient or signs observed. Cases of these symptoms may be prevalent but may not directly indicate the specificity of the diseases. This dataset is limited to rural areas covered by the health workers.

Form P:

Form P include cases compiled by the pharmacist or medical technician according to the conditions listed under IDSP and based on provisional diagnosis is written by the doctor. These are reported as probable cases, and come from both rural and urban regions. Mild cases may not reach hospital so cases reported in P forms are generally less than S forms.

22 categories of diseases and syndromes (including 2 "others" category that would require more specification) are reported in the P form, some of the ones that may be related to water contamination include Acute Diarrheal Disease, Viral Hepatitis and so on.

The exact number of cases are reported for each of the 22 categories.

Form L:

Laboratory surveillance are finalized diagnosis reported by the lab technician after performing appropriate lab test. All undiagnosed conditions and out-break related conditions require accompanying laboratory test records. Reported diseases that may be related to water include Cholera, Viral Hepatitis and so on. The following details are listed for each of the diseases:

- Number of samples tested
- Number found positive
- Line list of positive cases: where the name, age, gender, address, test type and diagnosis of any positively confirmed cases are reported

Ideally, same cases under the 3 forms should be linked with one another, but with the current reporting format a clear tracking system is not yet in place. Reports across each year can be generated for a certain disease based on the SPL forms at the state, district or block level, and graphical display of the trends can also be automatically generated.

A general summary of the data evaluation results from the variables above is shown in in Table 5-23. There are no accessible data for evaluation, so the data characteristics and data utility are only summarized briefly.

Table 5-23: Evaluation summary for IDSP S/P/L form data

	Data availability	Columns	Column	Data Type	Processing/ Table Joins
IDSP-SPL form	2004 to now	Master Data	Administrative region	text	
			Reporting Unit	text	Challenge to extract administrative levels
			Population	integer	
		Basic Information	Name of Health Worker/Volunteer/Practitioner or Name of Supervisor	text	

		Officer-in-charge (L and P only)	text	
		Reporting Week	integer	Reporting years differs from SBM/IMIS
		Date of Reporting	date	
	S form	Suspected Syndromes	6 categories (with sub categories)	Extract water-related diseases and symptoms
		cases and deaths (by gender and age range)	integer	
	P form	Diseases/Syndromes	20 categories and other	
		No. of cases	integer	
	L form	Diseases	12 categories and other	
		No. of samples tested	integer	
		No. found positive	integer	
		List of positive cases	-	

Analysis on the data characteristics is summarized below.

- Accessibility

Health data contains sensitive information, and it is much more challenging to obtain. Compared to SBM and IMIS where data can be immediately scarped or downloaded from the webpages online, approval must go through the health departments for all symptom and disease related data.

Theoretically, IDSP records should date back to 2004 when the database was first constructed. There is discrepancy between the calendar year in IDSP and SBM/IMIS, but since IDSP data is reported routinely on a weekly basis, the extraction of data from any given period of time should be simple via IDSP.

- Simplicity

The record of tests, cases and deaths are straightforward. There is no category specifying the likely or confirmed origin of these disease, so we can only make assumptions on whether they are waterborne by the nature of the disease. Extra columns of data that assists with extraction of water-related disease and symptoms would support the creation of an integrated WaSH-health system.

- Integration viability

The biggest barrier to data integration is the administrative region of each case. Data are only reported to the block-level, but many of the reporting units, especially rural Sub-centers and Primary Health Centers, only cover a certain number of villages. The confirmed cases at the lab level also records the exact address of the patient. It would be possible to reconfigure the data and obtain health details at a higher granularity, but this may require strong knowledge of each of the reporting units. Spatial approximation may also be used to estimate the administrative region coverage of each reporting units.

- Other factors

According to literature, there is considerable implementation gap at the rural sub-center level with the local health workers, and the quality and completeness of S form data is of concern (Kumar *et al.*,

2014). There is insufficient information to quantitatively evaluate the uniformity, completeness and quality of data.

Analysis on the data utility is summarized below.

- Accessibility

It theory, SPL form data should be entered promptly and the process is supervised closely by the Data Managers. There are designated data entry personnel at each level of reporting, with a relatively clear task description. Consistent and timely data inputs can be motivated through this administrative structure.

There are limited data on the actual monitoring process implementation.

- Sensitivity and predictive value

By definition, outbreak is an aggregation of large-scale occurring diseases of the same cause. Hence, number of outbreak-related disease is directly related to cases of outbreaks. They are also a reflection of contamination in the surrounding environments.

SPL data can allow us to analyze their effectiveness at predicting outbreaks, and the rate of false positives or false negatives. Without data, it is not possible to draw conclusions on whether the SPL disease and symptom numbers can effectively predict outbreaks, and whether all cases of outbreaks are readily reflected through the SPL monitoring.

- Timeliness

Outbreak management is highly time sensitive, and among all three databases, IDSP has the rigorous regulations on timely delivery of data. There is a data table of "Consistency Report" that specifically evaluates whether Reporting Units have delivered data timely each week, as shown in the sample form Figure 5-21. While the exact consistency of each reporting units in the 3 districts are not known, the attentive monitoring of timeliness increases the likelihood of timely reports.

Ministry of Health & Family Welfare(IDSP)
Government of India
Weekly Form - P Submission Status FOR **DISTRICT:** CENTRAL DELHI and **STATE:** DELHI
Reporting Period: 3-1-2011 To 13-3-2011
Date Report Generated :Aug 20, 2011 2:10:10 PM
To view Week Start Date & End Date, place your cursor on week number

S.No	Block Name	RU Name	RU Type	Week No: 1	Week No: 2	Week No: 3	Week No: 4	Week No: 5	Week No: 6	Week No: 7	Week No: 8	Week No: 9	Week No: 10	Consistancy For 10 Week		
														>=80	>=50 & <80	<50 or NULL
1	CENTRAL DELHI	DGD AJMERI GATE	HC	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N
2	CENTRAL DELHI	DGD BALLIMARAN	HC	Y	Y	Y	Y	N	Y	Y	N	N	Y	N	Y	N
3	CENTRAL DELHI	DGD CHAMELIAN ROAD	HC	Y	Y	Y	Y	Y	N	Y	N	Y	Y	Y	N	N
4	CENTRAL DELHI	DGD DUJANA HOUSE	HC	Y	Y	Y	Y	Y	N	Y	N	Y	Y	Y	N	N
5	CENTRAL DELHI	DGD DELHI SACHIVALAY	HC	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	N	N
6	CENTRAL DELHI	DGD GALI SAMAUSAN	HC	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N
7	CENTRAL DELHI	DGD GALI GULIYAN	HC	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N
8	CENTRAL DELHI	DGD MOTIA KHAN	HC	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N
9	CENTRAL DELHI	DGD NABI KARIM	HC	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y	N	N
10	CENTRAL DELHI	DGD PAHADGANJ	HC	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N
11	CENTRAL DELHI	DGD REGARPURA	HC	Y	Y	Y	Y	Y	N	N	Y	Y	Y	Y	N	N
12	CENTRAL DELHI	DGD SUIWALAN	HC	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	N	N
13	CENTRAL DELHI	DGD TANK ROAD	HC	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N
14	CENTRAL DELHI	MCW-DEV NAGAR	HC	N	Y	N	Y	Y	Y	Y	Y	Y	N	N	Y	N
15	CENTRAL DELHI	MCW-KATRA KUSHAL RAI	HC	N	Y	N	Y	Y	N	N	Y	Y	Y	N	Y	N
16	CENTRAL DELHI	MCW-NAUGHERA	HC	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N

Figure 5-21: Consistency report for P forms weekly submission status of Central Delhi district (IDSP, no date c)

Outbreak records

Disease outbreak records are reported by the SSU/DSU. The cases are reported at the district level, but contains details on the exact locations of the outbreaks. Both urban and rural outbreaks are reported. The causes of outbreaks are also updated once investigation is done. While the original outbreak data table is not accessible, weekly summary reports can be retrieved from the IDSP website. The outbreak records are analyzed based on these weekly summaries, which may vary in their format compared to the raw data table.

Based on IMIS/SBM fiscal year, data from the April 2015 – March 2016 are selected. The following variables are in the summary reports:

- Administrative region (district-level)
- Outbreak ID (post-2016 only): ID for each of the outbreak case
- Diseases: a variety of disease outbreak are reported. The comment section offered information on the cause of the disease, and whether water sampling is done during the outbreak investigation process. All water-associated outbreaks in the 3 districts are selected, including Acute Diarrheal Disease (Acute Gastroenteritis), Cholera, Hepatitis E, Jaundice and Dysentery.
- Number of cases: the number of cases affected by the outbreak
- Number of deaths: the number of deaths caused by the outbreak

- Date of outbreak start: the traceable original data that the first case of the outbreak happened
- Date of outbreak reported: the date that outbreak is reported in IDSP
- Status: current status of the outbreak, including "Under Surveillance" or "Under Control." This status is likely to be updated later in the IDSP data table once more investigation is carried out, but the summary report only shows the status snapshot during the reporting week.
- Comment: details on the region of the outbreak occurrence, actions of the rapid response teams and preliminary investigation conclusions.

A general summary of the data evaluation results is show below in Table 5-24.

Table 5-24: Evaluation summary for IDSP outbreak data

IDSP	Data Availability	Total Entries	Column	Data Type	Missing Data	Abnormal Data	Simplicity	Uniformity	Processing/Table Joins
Outbreak	June 2009 until now	20	Administrative region	text	0	1	yes	no	supplemented by details in comments
			Outbreak ID	alphanumeric text	15	0	yes	yes	
			Disease	5 categories	0	0	yes	no	category reconfiguration
			Number of cases	integer	0	0	yes	yes	
			Number of deaths	integer	0	0	yes	yes	
			Date of outbreak start	date	0	0	yes	yes	
			Date of outbreak reported	date	1	0	yes	yes	
			Status	2 categories	0	0	yes	yes	
			Comment	text	0	-	no	no	Extract village/town location of the case

Analysis on data characteristics is carried out as below:

- Accessibility

Outbreak weekly summaries are available since June 2009. However, instead of an accessible data table format, they are printed in PDF and need to be transformed and transcribed for processing. Raw outbreak data from the IDSP database is not accessible publicly and requires approval from the health department for access.

- Simplicity

The definition of the variables is generally straightforward. However, there is no definition as to what exactly is written in the comments. Although outbreak case, location and reactions are usually documented in the comments, there is no standard definition for reporting.

- Uniformity

Case numbers, dates and outbreak IDs are reported uniformly. However, there are frequent misspellings in the administrative level names, where "Navsari" is written as "Navasari" and "Gujarat"

as "Gujrat." Compared to the standardized administrative region inputs in SBM and IMIS, entry errors and variations are much more common in the IDSP reports. The non-standard IDSP district names compared to the other two databases are also demonstrated in the Table 5-2.

Diseases are reported in text format, and the same type of disease are written in many variations. For example, "Acute Diarrheal Disease," "Acute Diarrheal Disorder," "Acute Gastroenteritis" and "Acute GE" are all used interchangeable, which requires category consolidation for data analysis. The lack of definition also resulted in non-standard comments. Extracting any useful information from the comment would require extra process.

- Completeness

The data are overall complete. 15 of the entries are from 2015 and missing the outbreak ID. 1 of the entries is reporting an update on a previous outbreak, so the "date of outbreak reported" is empty.

- Quality/Accuracy

Considering that outbreaks are widely verified and reported, there are unlikely to be any false positive reporting in the outbreak cases observed. On the other hand, there may be undocumented cases, especially in the rural regions. If the Health Workers are not consistently reporting high quality symptoms at the local Sub-centers, clusters of cases can easily go unnoticed if the patients don't actually visit a hospital.

- Integration viability

Similar to SPL form data, the challenge to integration exists in the administrative level variable. Only the district is reported. However, it is possible to extract village names from the "Comments" column. The process is laborious because only the village name is given, and the block and panchayat name has to be filled through search the village databases. In addition, outbreak cases from both rural and urban regions are reported in the same table. They are not differentiated, and the comments only mention a region name without specifying whether it's an urban town or rural village. SBM and IMIS databases are focused on rural regions only. Hence, extra searches are required to determine whether the outbreak is in a rural region before village name extraction can be done.

Overall, compared to integration within IMIS and SBM, much more data processing and manual editing is required to create outbreak datasets at the appropriate geographic granularity.

Analysis on the data utility is carried out below.

- Acceptability

Outbreak incidents are considered crisis situations. If an outbreak is actually observed and confirmed, the response, monitoring and action should be prompt. Within our dataset, the date of outbreak start and the date of reporting are at most 4 days apart. There are a few cases of make-up reports from previous weeks, arriving 1-2 weeks late.

Overall, the data is monitored closely at the district and state surveillance unit, and considering its high priority, the motivation to monitor and control outbreaks should be high.

- Other factors

Considering that occurrence outbreak is the ultimate consequence of other water and sanitation related risk factors and violations, outbreak data is the direct and final indicator of severe violation in the WaSH-Health chain. Hence, analysis of sensitivity, predictive value and timeliness does not apply to this data table.

5.4 Cross-database integration

5.4.1 Database schema

As summarized through section 5.1-5.3, key variables in IMIS, SBM and IDSP are generally connected through the base database schema of administrative levels, as shown in Figure 5-22.

The administrative level at the top is all highlighted in red, while all IMIS data are highlighted in yellow, SMB data blue and IDSP data gray. The three tables of water sources in IMIS are consolidated into one water source table. While SBM mobile data, SBM aggregated physical progress data and IDSP SPL forms data are not included in the final analysis, they are still incorporated in the schema for reference.

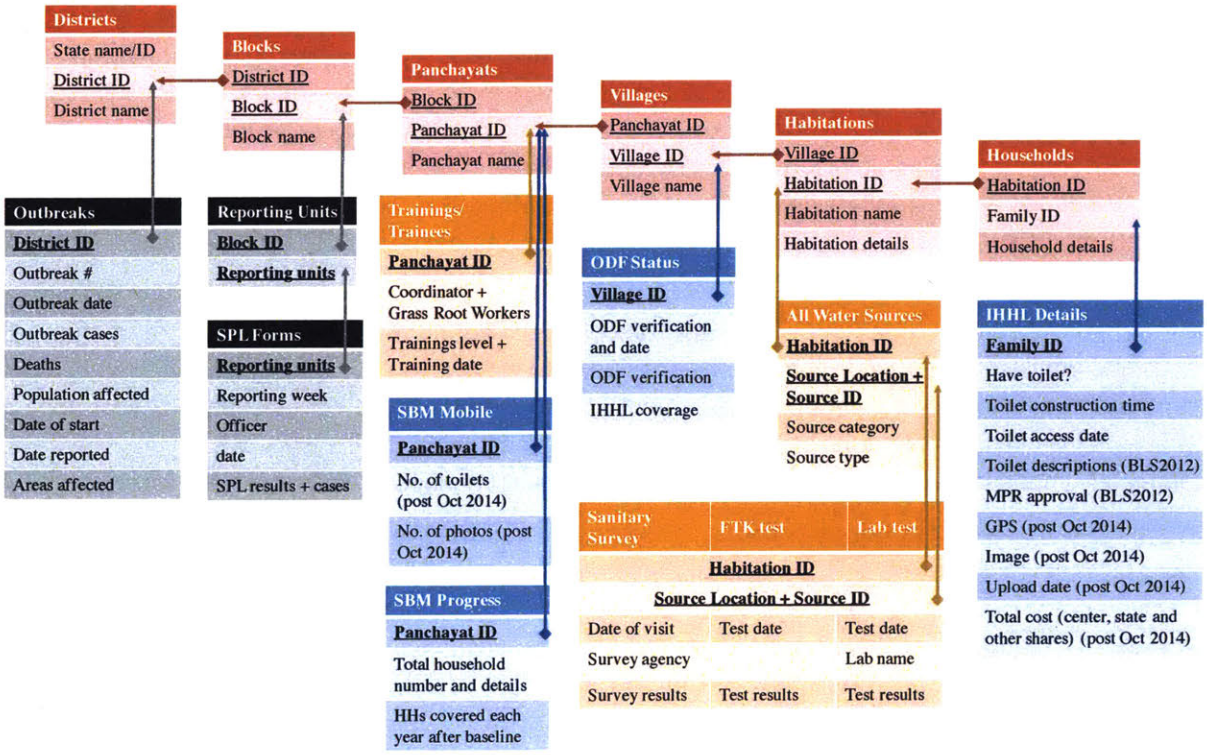


Figure 5-22: Combined schema of all tables of interest within IMIS (blue tables), SBM (yellow tables) and IDSP (gray tables).

Evaluation on the integrated WaSH database and the cost to integration are carried out in the following aspects, as outlined in chapter 3.

5.4.2 Database Structure

We first look at the database structure, focusing on evaluating its simplicity, flexibility and stability.

The relationships in the database are generally simple and straightforward. All data are connected at different levels of the rural administration region, and can be easily aggregated at a specific level for further analysis.

Considering that the administrative levels are the basic structure of this schema, the flexibility of the database depends largely on the adaptability of the administrative units. The current structure is not flexible in such a manner above the household level. At the household level, family ID is available in the SBM database, which makes it possible to track each family despite changes in household details. However, while there seems to be habitation IDs at the backend of the IMIS database, there is not a visible tracking ID for every single habitation. Without a consistent ID, name changes of the habitations across the years would raise challenge over matching the correct dataset across the years.

There have been frequent recent changes in the administrative units. In SBM, as mentioned in Section 5.2, it was possible to track the changes in block names changes only through matching household IDs. For example, there is an increase of 971 habitations in Navsari in the past year, most of which followed the case in Table 5-25 with a large increase in the number of habitations under one village. After matching the population number, it can be concluded that the new habitations are completely separate from the original habitation in 2015-2016 so there is not a need to adjust previous variables under the original habitation. These processes can be completely avoided if there is a clear and consistent way that each administrative level is uniquely identified, so that any revisions to the names would not cause any disruption within the data system. More flexibility can be added if the identifier can accommodate any future consolidation or separation of administrative units. These are all very likely cases in the context of rural India.

The current database does not have the flexibility to adapt to these changes, and manual rearrangements are required. In comparison, the SBM database is slightly more flexibility. Even though some extra matching is required, consistent family IDs can at least still help identify any rearranged administrative units. IMIS does not have such consistent unit IDs in place, which may make it completely impossible to adapt older data to any significant changes in administrative units.

Table 5-25: Change in habitations across 2015-2016 to 2016-2017 in IMIS.

District	Block	Panchayat	Village	Habitation	2015-2016 population	2016-2017 population
Navsari	Jalalpore	Chokhad	Chokhad	Chokhad	1047	1047
Navsari	Jalalpore	Chokhad	Chokhad	Choramala falia	-	112
Navsari	Jalalpore	Chokhad	Chokhad	Ero falia	-	101
Navsari	Jalalpore	Chokhad	Chokhad	Hanuman falia	-	111
Navsari	Jalalpore	Chokhad	Chokhad	Harijanvas falia	-	160
Navsari	Jalalpore	Chokhad	Chokhad	Navi nagari	-	113
Navsari	Jalalpore	Chokhad	Chokhad	Patel falia	-	102
Navsari	Jalalpore	Chokhad	Chokhad	Tekara falia	-	124

The stability of the database depends on consistent availability and accessibility of the data. Across the three databases, IDSP provides data in the timeliest fashion and has the strongest database server

that would allow smooth delivery of information. SBM data are also relatively consistently provided. However, delays in the updates of ODF verification data for some villages question its reliability in providing the most updated sanitation status. Compared to SBM, IMIS has a much larger number of datasets of a much wider variety, and the increased complexity resulted in an unstable database with a large number of inaccessible data tables. The data collection and provision are also less reliable than the SBM database, due to a less clearly defined monitoring target.

In conclusion, while the database structure is relatively simple and straightforward to operate, the lack of a flexible administrative unit unique identifier reduces the overall adaptability of the database. The stability of the database varies by the different data sources. While we are not able to access IDSP data, in theory the data is provided with the highest consistency. SBM data are generally provided reliably. IMIS is the most challenging data source with frequent accessibility failures.

5.4.3 Integration Viability

Although the administration region based integration structure is simple, the cross-database integration process remains challenging. Most IMIS data are at the habitation level, and it has the most comprehensive quantity of habitations, so it is set as the base administrative unit for integration.

To begin with, SBM Household Details data has to be aggregated to the habitation level. 2015-2016 habitation names require updates to the revised names in 2016-2017 through matching family IDs at the household level. However, even after updating and removing the several blocks that have changed their district jurisdiction, many habitation still do not match exactly across SBM and IMIS. Different spellings (e.g. panchayat "Kavath" vs. "Kawath") or additional name segments (panchayat "Sultanpur (Vadadhara)" in SBM vs "Sultanpur" in IMIS) are common occurrences. Over 1000 SBM habitations are unable to be matched directly to IMIS habitations. Hence, fuzzy matching algorithms are used in SQL including Levenshtein distance (the moves it takes for one word to be transformed to another), Soundex (whether two words sound similar) and so on.

In the end, the following table join algorithms are instituted:

- District names exactly the same
- Block names with a Levenshtein distance less than 3
- Panchayat names with a Levenshtein distance less than 3, or one of the panchayat name contains to the panchayat name, or both block names have the maximum sound similarity
- Village and habitation names with a Levenshtein distance less than 3, or both names have the maximum sound similarity.

With this algorithm, around 300 habitations out of the total 4099 habitations in SBM are still unmatched due to other irregular name difference that was hard to account for (e.g. panchayat name "Ukardina muvada" in SBM and "Ukardi na mu" in IMIS, where one word contains another but with an additional space added in between). Many of these required manual matching in the end.

Overall, apart from the frequent insertion or deletion of one or two characters, a few additional patterns can be observed such as the interchangeable use of v/w, "Muvada"/ "Mu"/"Na Muvada" and so on. However, these patterns may occur based on the locality. Time-intensive adaptation and optimization

of the matching model based on different regions is required for cross-database integration. Manual observation and matching is likely still required at the end.

Even after all the possible matching has been conducted, the exact IMIS match for the 10 habitations outlined in Table 5-26 are still not found. The reasons are also outlined in Table 5-26, which are all related to difference in the habitations under the same village. In many of the cases, while there is one habitation that shares the same name as the panchayat and village in one database, this one habitation disappears and a number of additional habitation emerges in the other. This seems to be a common restructure theme of habitations, considering that most habitations are also changed in the similar manner in IMIS from 2015-2016 to 2016-2017. Without further information on the details of these habitation restricting processes, we can only leave the 10 SBM habitations unmatched.

In the end, the rest 4089 habitations can be matched with the corresponding IMIS habitations, which is only around 75% of the total number of habitations (5445) listed in IMIS for 2015-2016. This may be due to the incomplete SBM IHHL entries, or due to additional habitations created under IMIS. If we consider all 972 new habitations instituted in 2016-2017 in IMIS, the discrepancy increases even further.

Table 5-26: SBM habitations with no IMIS match. The detailed list of IMIS-SBM habitation matching is attached in Appendix A.

SBM Habitations with no IMIS match					Mismatch Comments	IMIS Comments				
District	**Block**	**Panchayat**	**Village**	**Habitation**		**District**	**Block**	**Panchayat**	**Village**	**Habitation**
Navsari	Chikhali	Achhavani	Achhavani	**Bezzari faliya**	There is only "BEZZARI FALIA" in IMIS and matched to another SBM habitaiton already					
Navsari	Chikhali	Achhavani	Achhavani	**Dadri faliya**	There is only "DADARI FALIA" in IMIS and matched to another SBM habitaiton already					
Navsari	Gandevi	Bigri	Bigri	**Bigri**	The left are all unmatched habitations from Bigri village in IMIS, but no single Bigri habitation.	Navsari	Gandevi	Bigri	Bigri	**Halpativas falia** **Nishal falia** **Ramvadi (harijan) falia** **Suitalvadi falia**
Navsari	Gandevi	Ponsri	Ponsri	**Ponsri**	Unmatched IMIS villages from Ponsari to the left, same as above	Navsari	Gandevi	Ponsari	Ponsari	**Bhuliya falia** **Maskara falia**
Surat	Kamrej	Dharutha	Dharutha	**Bhathiji faliya** **Iay faliya**	Only one Dharutha habitation to the left in the IMIS village and already matched to another SBM habitation	Surat	Kamrej	Dharutha	Dharutha	**Dharutha**
Surat	Mahuva	Vanskui	Vanskui	**Gamtal faliya** **Naher faliya**	There are 4 other IMIS habitations in Vanskui and cannot be matched to the 2 names in SBM	Surat	Mahuva	Vanskui	Vanskui	**Gocher falia** **Madhudurlabh falia** **Pana falia** **Pramukhshri falia**
Surat	Umarpada	Ghanawad	Ghanawad	**Rajanvadi**	Only one unmatched IMIS habitation in Ghanawad village, does not match with SBM name	Surat	Umarpada	Dhanavad	Ghanawad	**Kantanvadi**
Surat	Umarpada	Rudhi gavan	Rudhi gavan	**Rundh gavan**	One habitation under Rudhi Gavan village in IMIS (to the left), already matched to another SBM habitation.	Surat	Umarpada	Rundh gavan	Rudhi gavan	**Rundhgavan**

The final cross database integration results are shown in Table 5-27 and Figure 5-23. The biological lab test results, sanitary survey results and trainee numbers are significantly low in coverage compared to the other variables. SBM ODF data are not available in snapshot format. To simulate the snapshot, we selected villages that declared or verified ODF before the end of 2015-2016, and labelled all the rest of the villages undeclared. Similarly, outbreak records are only reported across 12 villages with all the rest concluded as no outbreaks. Hence, ODF and outbreak are considered covering 100% administrative regions after such preprocessing, so their entries are not compared in the integration results table.

The mismatch identified in the table generally resulted from the following:

- Habitation name discrepancy across SBM/IMIS
- Wrong source ID
- Supply scheme sources require both scheme ID and source ID for unique identification, but only source ID are available in the testing records, resulting in duplicate matches.

Other mismatches that can be corrected through administrative region name adjustments are not included in the mismatch counts.

Table 5-27: Cross-database integration results

Base Unit Counts	**Habitation (5445)**		**Panchayat (1413)**		**Source (~80000)**	
Matching Variable	Coverage	Mismatch	Coverage	Mismatch	Coverage	Mismatch
Lab - biological tests	929 (17%)	0			3453 (4%)	0
Lab - all tests	5242 (96%)	0			16740 (20%)	0
FTK – biological test	4167 (77%)	0			11914 (15%)	12
Sanitary Surveys	755 (14%)	0			965 (1%)	1
Training			1412 (>99%)	0		
Trainees			305 (22%)	0		
SBM IHHL details	3857 (71%)	10				

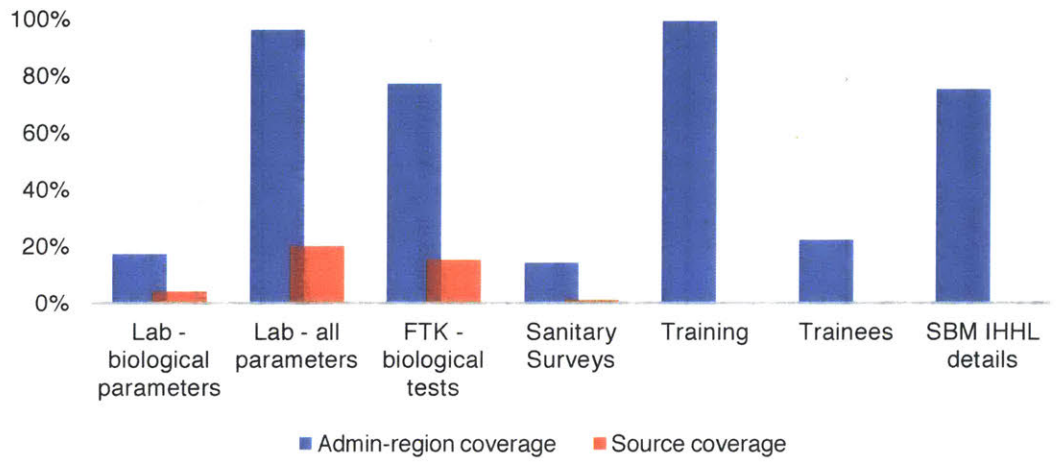


Figure 5-23: Data coverage across administrative levels (habitation/panchayat) and water sources

5.4.4 Uniformity

As suggested by the evaluation on table joins within each of the database, inconsistent data entries are frequent within and across the databases.

Figure 5-22 shows that critical nodal points of data table linking are the administration units and the source ID. Common inconsistencies in administration units result from the following scenarios:

- General difference in spelling and wording of the same name
- Inconsistent habitation entries within the same village
- Difference in block jurisdiction of panchayats
- Difference in district jurisdiction of blocks
- Difference in panchayat names.

Common inconsistencies in water source identification generally result from the following scenarios:

- Difference in source ID number due to digits being cutoff
- Source ID is not a unique identifier for Water Supply Scheme sources, but is used as one in other tables.

A flexible ID system for the administrative units that can adapt to frequent changes, as described in Section 5.4.2, would be critical in decreasing the barrier to integration. In addition, all tables should have a primary key, and all future link between data tables should be implemented through a foreign key setup that links two table through a unique primary key. The current duplicate records in data tables and discrepancies in the one-to-many links are likely results of missing or problematic primary key and foreign key setups.

Apart from inconsistencies in the data table linking process, there are also inconsistencies in some of the same variables across the different databases. For example, household numbers are reported across SBM and IMIS databases, but the scatter plot in Figure 5-24 shows quite a few discrepancies in household numbers of the same habitation. A higher IMIS household number can be explained because of incomplete IHHL entries in SBM. A higher SBM household number, on the other hand, suggest either that households are missing in IMIS, or that SBM habitation units are defined differently than IMIS. Considering that SBM have over 1000 habitations missing compared to IMIS records, there is also the possibility that the missing IMIS habitations may be considered a segment of a nearby habitation in SBM definition, resulting in some SBM habitations having more households. This can largely affect cross-sector data matching and analysis.

Further investigation and local interviews are needed to uncover the root of such data inconsistencies.

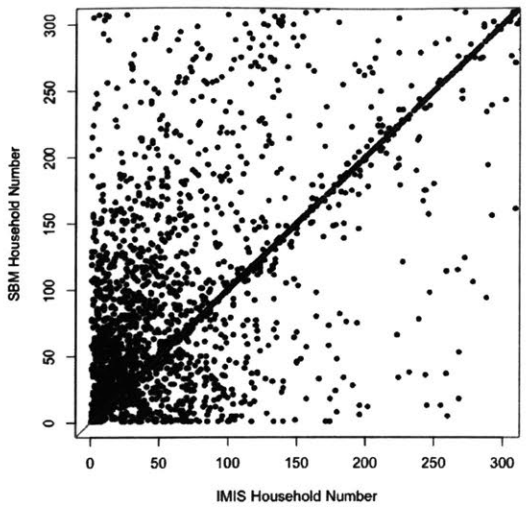


Figure 5-24: Scatter plot of the SBM vs. IMIS household number for each habitation. The dots are expected to fall along y=x, the red diagonal line.

5.4.5 Completeness

A comprehensive visualization of all key integrated variables across all 5356 habitations (excluding habitations where IMIS water quality or sanitary results are unavailable) are shown in Figure 5-25. All results are aggregated at the habitation level. Water quality test, FTK test and sanitary survey results are all binary, where contamination found in any source, or any source with medium-high risk would result in 1. Binary variables are selected because any contamination in any single source may result in widespread disease outbreak. It is challenging to account for the relationships between difference sources within the same habitations – larger contamination sample numbers might not necessary mean a larger population at risk. Hence, to avoid the uncertain complexity, we are only using binary variables for the aggregated results. Training and trainee entries are the numbers at the panchayat level, assigned to each habitation. IHHL coverage percentage is an aggregation of toilet ownership at the household level. ODF categories 0, 1, 2 are defined as "not declared," "declared but not verified" and "verified," same as in Section 5.2. Outbreak are also reported in binary records where habitations that belong to the 12 affected villages are labelled 1.

The data are ordered by district, block, panchayat, village and habitation name accordingly, and the first column is the index from 1 to 5356. Neighboring entries are likely entries from the same administrative region.

All missing data are shaded in grey. As we can observe, apart from the challenge in the integration process, there is also a serious challenge of data availability in the integrated database. A large number of habitations do not have any biological lab test records. While most habitations have chemical lab test records such as pH or TDS, very few violations are detected in comparison to the biological contamination detected through FTK, so they are not a good proxy for biological contamination. Sanitary surveys and trainee numbers are also sparse. Other variables such as training numbers, IHHL coverage and FTK biological contamination can be integrated more effectively with a manageable percentage of missing entries.

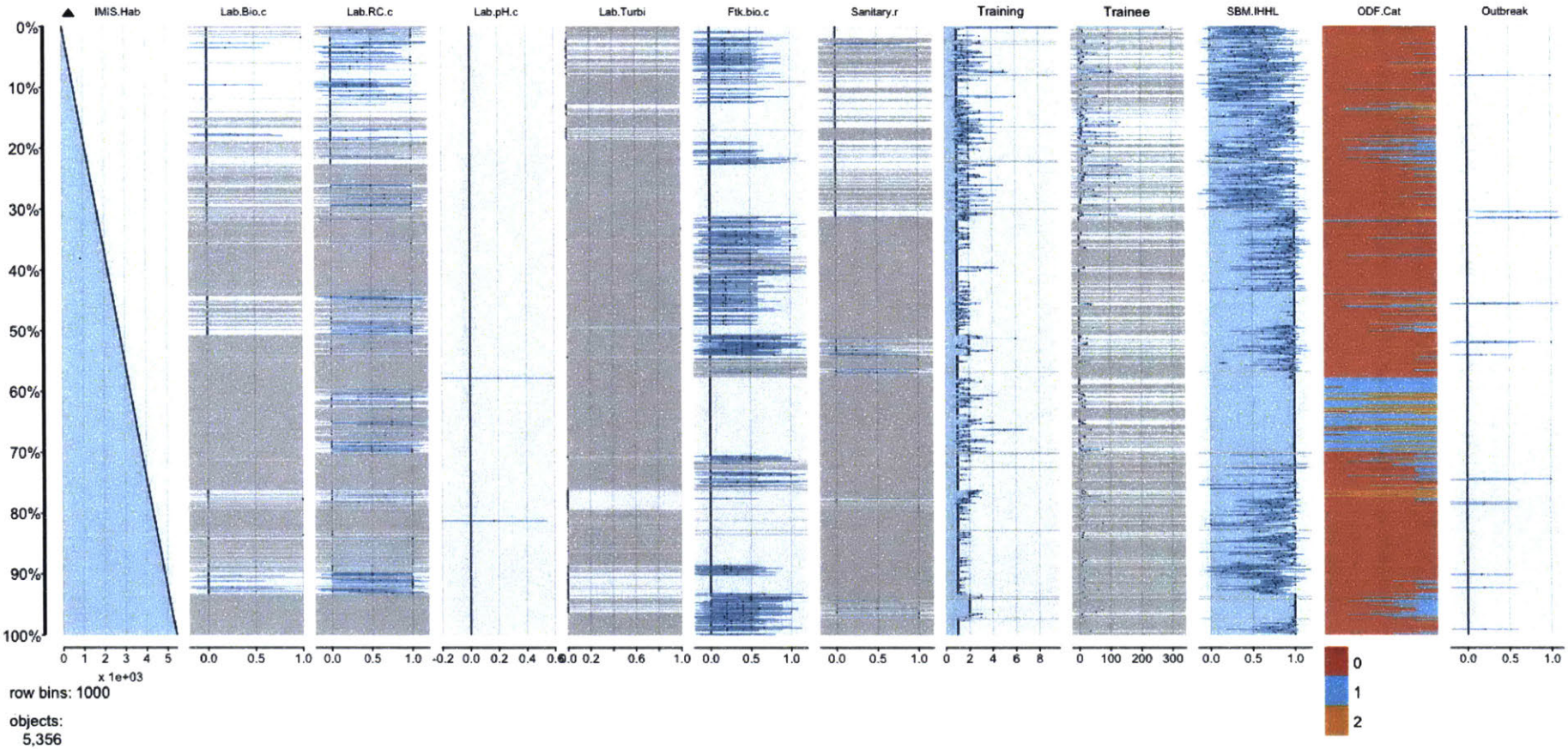


Figure 5-25: Table plot of key outbreak-related variables across IMIS, SBM and IDSP. The lab result columns are (from right to left): any biological contamination (E. coli or coliform) found, any residual chlorine violation found, any pH contamination found, any turbidity violation found. The next columns are: any biological contamination by FTK, average sanitary survey risk level, training numbers, trainee numbers, SBM IHHL percentage, ODF category and outbreak records.

In addition to missing data entries across the habitations, it is also useful to analyze the missing variables along the outbreak disease pathway to better understand the capability and limitations of the current database in the chain of outbreak control.

Here we revisit the DPSEEA framework with more specificities. The DPSEEA framework for waterborne diseases is defined in Figure 5-26. There are other diseases associated with water, such as water-washed diseases where diseases occur due to lack of water to ensure basic handwashing or food washing processes, or water-based diseases caused by contacting contaminated water. Compared to the contamination of public drinking water sources that the entire community share, water-washed or water-based diseases are less likely to develop into community-scale outbreaks. While some of the outbreaks in this study, such as cholera, Hepatitis E and Jaundice are not included as a health effect in Figure 5-26, the pathways should be similar. The response for these outbreaks all included water quality testing, chlorination and sanitary education, suggesting the similar causation of pathogens concentration in drinking water.

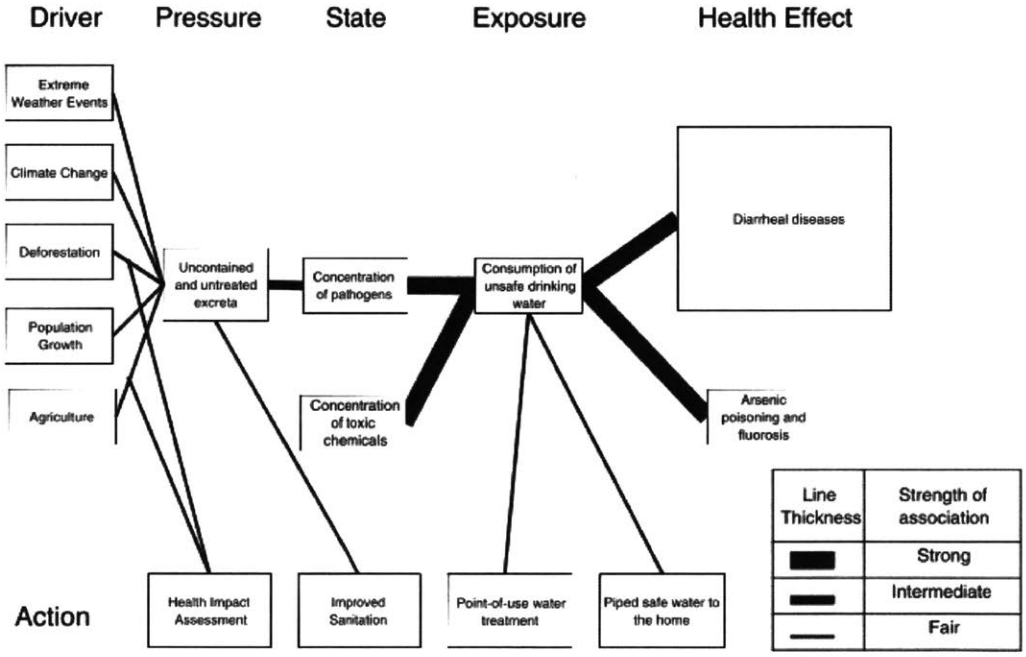


Figure 5-26: DPSEEA model for "water-borne diseases", as defined by the Bradley Classification. (Gentry-Shields and Bartram, 2014).

Based on the framework, the existing data from IMIS, SBM and IDSP are connected via the DPSEEA framework as shown in Figure 5-27. Analysis on key missing variables along the chain is carried out below.

- Driver

Population pressure and general environmental pressures are not accounted for in this study. Factors such as climate change, deforestation and so on are not available within the databases. While population data is available at the habitation levels and growth rates are possible to account for, it might be more useful in a time series analysis. In addition, for the year 2015-2016, it is clear that SBM

initiative is the main driver for changes in open excreta. Hence, population growth is not included in the model.

- Pressure

Open excreta can be inferred from ODF status data, but it is not directly documented. ODF status is also only available for locations with 100% latrine construction and free from open excreta. The level of open excreta for the majority of habitations that have not declared ODF is not available. However, the "Action" variable of IHHL physical progress directly impacts the likelihood of open excreta. In theory, the IHHL progress variable can only impact the DPSEEA chain through the changing levels of open excreta. Hence, it can be considered a proxy for open excreta in locations where ODF has not been declared. The limitations of this proxy should be noted, considering that installation of physical structures do not immediately lead to behavioral change.

Other sanitary risks are documented in the sanitary survey form. However, considering the low quantity of surveys and the overall dubious quality of the records, the reliability of sanitary survey results as a pressure factor is low.

- State

Biological water quality contamination status by lab and FTK is an effective reflection of pathogen concentration in drinking water. Considering the low completeness of biological water quality lab tests, results from the two sources may need to be combined to create a more comprehensive record.

- Exposure

Information on exposure is not available in this analysis. Data on the population served by each source is available in IMIS only at each source level, making them too challenging to access within the timeline of this study. The length of time that the contamination has been present and actions after contamination found are not documented in the datasets, making it hard to estimate exact risk exposure.

- Health Effect

The direct effect along this DPSEEA chain is waterborne disease. Outbreaks are only a manifestation of diseases clusters. While outbreaks can be considered a rough proxy of serious large-scale health effects, there are also many health effects from unsafe exposure that are not elevated to the level of outbreaks. Using outbreak occurrence as the outcome effect in the casual chain loses a lot of information, especially considering that outbreaks occurred only in 12 out of the 1670 villages.

- Action

While SBM implementation is hard to directly account for, IHHL progress is an effective reflection of the implementation process. While water access is not in the original DPSEEA framework, it is generally required for a functional IHHL and thus included in association with IHHL construction. Increased local training is also part of the SBM initiative that can improve sanitary status.

On the other hand, training can also impact the casual chain through increasing water quality risk awareness, which can induce better water drinking habits and likely decrease consumption of unsafe water. As explained in Section 5.1.2, however, it is not possible to differentiate between the sanitation

training and water quality awareness training, because the training types entries are not consistent and cannot be categorized (as indicated in the evaluation Table 5-15).

Piped water coverage and chlorination treatment are also actions that may directly impact contamination exposure. Water source type data are available for all water sources. However, considering the doubts with the total source number, piped water percentage calculated from all sources may not be reliable. It is not included in the model. While residual chlorine data are help in approximating the chlorination treatment coverage, very few residual chlorine data are available to effectively contribute to the analysis.

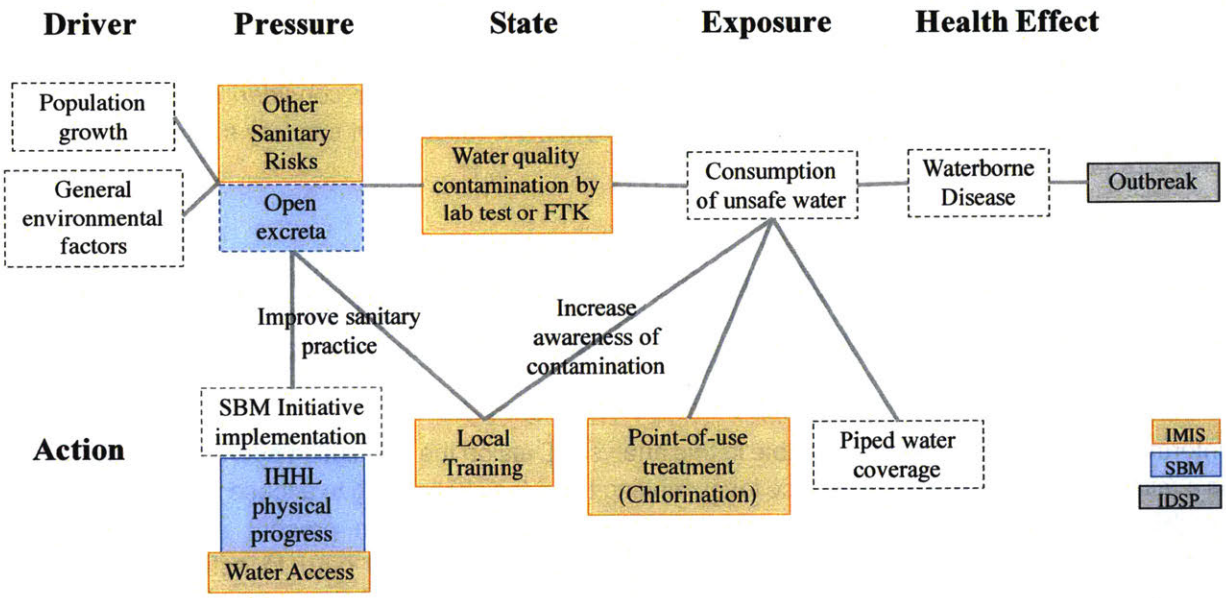


Figure 5-27: Adaptation of the DPSEAA framework for waterborne diseases in the context of India's 3 databases. Variables inside dashed lines are missing from the integrated database.

5.4.6 Major Limitations and Costs to Integration

In conclusion, the following major limitations exist within the integrated WaSH database. The required solutions are also the cost barriers to an effective implementation of a WaSH-integrated database.

Barriers to data table joining

The inconsistent habitation names and source unique identifiers result in challenges during the data table integration process, especially for cross-database joins.

Solutions:

- An improved identification system should be setup for administrative units that allows for highly flexibility and adaptability to administrative region reconfigurations.

- Administrative region coverage should be assigned to IDSP reporting units, or estimated by the geographic location of the reporting unit.
- A better understanding of the major habitation restructuring process is required to evaluate how the process may affect the corresponding datasets.
- A better unique identifier should be setup for water supply schemes.
- Primary keys and foreign keys should be effectively instituted throughout the database to prevent any future inconsistencies in the table joining process.

Credibility concerns of Water Source Inventory

There is a larger number of water sources, but only less than 20% have been covered through water testing. Concerns have been raised over the exact number of functional water sources, and interviews review that sometimes the different tests for the same source can be entered as multiple sources. The lack of a credible source inventory prevents from evaluating piped supply coverage and water quality testing coverage, both of which are likely to impact waterborne disease occurrence. The water supply scheme inventory is also inaccessible.

Solution:

- A review of all current water sources is required where defunct sources should be removed from the inventory and sources with similar information can be consolidated.
- Water source tables should be categorized in alignment with the five source lists (as indicated by the alphanumeric source ID): water supply scheme source, delivery point source, public source, private source, surface water source.
- Past snapshots of the water source inventory at the end of each fiscal year should be available.
- The "source location" entry should be clearly defined so that duplicate sources can be easier to identify.
- Overall performance improvement of IMIS is needed to ensure consistent access to water supply scheme inventory.

Limited access to details at the source/scheme level

Important details including population served, contamination records and functionality of each source are not readily accessible in data table formats.

Solutions:

- Ideally, details available at the source level should be included in the water source table as separate columns.
- Otherwise, improved multi-level scraping is required to obtain data source by source and aggregate all details into a comprehensive data table.

Low quantity of biological water quality lab results

Habitation coverage of biological water quality lab tests are only 17%. Risks of contamination cannot be effectively detected through such low coverage level.

Solutions:

- Routine biological water quality monitoring should be enforced at all water labs.
- Water quality FTK results can be used to supplement lab results, but data quality needs to be ensured.
- Chemical water quality results can be used as an indicator for potential biological contamination. If increased monitoring of biological contaminants is challenging, monitoring should be prioritized for sources that found FTK, pH or Turbidity contamination.

Low quantity and quality of sanitary survey results

Sanitary survey results cover less than 14% of the habitations. Water quality concerns, rather than the actual facility and environmental risks, are concluded in the sanitary survey dataset. Inconsistent reporting formats raise quality concerns over the dataset.

Solutions:

- Regular sanitary inspection should be enforced at all WASMO local offices.
- More standardized sanitary survey result reporting structure should be in place.
- Instead of reporting the safety or chlorination requirement of the water source, details on the actual risk factors of the water facility and its surrounding environment should be reported.

Interdependency of the concurrent data records

When different data are collected at the same time, such as sanitary survey data and FTK data, it is unclear whether the results may have influenced each other. Sanitary survey conclusions on chlorination requirements are frequently in sync with positive FTK results. Data analysis results can be significantly impacted depending on whether the monitoring data are collected independently.

Solutions:

- Further interviews on the exact data collection process are required.
- Independent data collection routines for each of the datasets should be set in place.
- For datasets that are related to each other, a tracking note should be available in the database. For example, a water sample tested positive through FTK may be sent on to be further tested in a local lab. A connector variable should be available to identify the two water quality records as interdependent.

Lack of sanitary behavioral and hygiene data

While the BLS-2012 collected data on toilet functionality and usage, the data columns are not available for the 2015-2016 dataset. Hygiene data are also not available. While IHHL and ODF can be proxies for open excreta risks, the scope is limited, especially considering delays in behavioral change after physical progress have been frequently reported.

Solutions:

- Continuous monitoring requirements should be set up for existing latrines.
- Toilet usage and other sanitation and hygiene practice data should be collected to the extent possible.
- The ODF+ status should be instituted to encourage long-term behavioral monitoring and ODF sustainability.

Outbreak as a proxy for health effects

Outbreaks are too low in occurrence to act as an effective proxy for SPL form data. They cannot comprehensively reflect the general health effects of contaminated water sources. They are also likely to go unreported, especially in rural regions.

Solutions:

- SPL form data is required to create a useful WaSH-health integrated database. Better and more timely access to the data would be necessary.

-

6 DECISION SUPPORT EFFECTIVENESS

Chapter 6 focuses on analyzing the effect of the integrated WaSH-health framework based on the DPSEEA casual chain. The chain is segmented into two components where factors associated with biological water quality contamination are analyzed in Section 6.1 and factors associated with outbreak occurrence are analyzed in Section 6.2. Summary on framework effectiveness and limitations to a more comprehensive evaluation are detailed in Section 6.3.

DECISION SUPPORT EFFECTIVENESS

After the thorough analysis of the integrated database and the cost barriers to implementation, this chapter focus on the validity of the DPSEEA model, and how effectively it can assist outbreak management. To begin with, the strength of association among the different segments of the DPSEEA framework is analyzed to understand the validity as well as the limitations of the current framework.

While the DPSEEA framework have multiple chains of association, due to the lack of certain variables and the exclusion of variables with large numbers of missing entries, only the following two associations can be analyzed:

Action – (Pressure) – State: analyzing factors affecting likelihood of biological water contamination

State + Action – (Exposure) – Health Effect: analyzing factors affecting likelihood of outbreaks

The statistical analysis is detailed in the following sections.

6.1 Water contamination outcome

6.1.1 Analysis

We first analyze the Action-(Pressure)-State association, with biological water quality contamination as the outcome. This section of the DPSEEA framework can be shown in Figure 6-1.

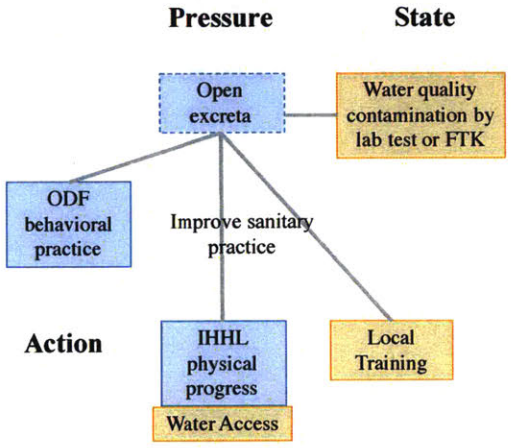


Figure 6-1: Action – Pressure – State segment of the DPSEEA framework for waterborne diseases

The goal is to understand how multiple actions affect open defecation status and in turn impact water quality contamination status. While ODF can be a rough indication of open excreta level as we outlined in Section 5.4.5, it is important to note that the ODF variable in the integrated database cannot signify any open defecation related information before 100% IHHL achievement. ODF declaration and verification, rather than a director indicator of "open excreta", is more analogous to an extension "action"

component from 100% IHHL achievement. The two variables are interdependent and separately represents actions before and after physical latrine construction completion.

Consequently, the problem above is separated into two questions of interest:

- whether latrine construction progress, water access and local trainings are significantly associated with ultimate decrease in the water quality contamination state (among habitations where 100% IHHL have not been achieved)
- whether ODF status, water access and local trainings are significantly associated with decrease in water quality contamination (among habitations where 100% IHHL have been achieved)

Key variables include in the model are listed in Table 6-1. Background variables are other general characteristics of the population that are likely to affect the biological contamination outcome. They are not considered within the DPSEEA framework and their results are not key to this study, but including these variables are expected to produce more rigorous and significant results for the other key variables of interest.

Table 6-1: Key variables in the Action-Pressure-State model

Category	Variable	Data type	Definition
Outcome	Biological contamination	Binary (0 or 1)	Whether biological contamination has been found in any of the sources in the habitation, either through lab tests or FTK tests.
Key independent variables	IHHL coverage	Numeric (0-1)	Percentage latrine coverage within a habitation
	ODF status	Categorical	3 categories: not declared ODF, declared but not verified, verified ODF.
	Water access	Numeric (0-250)	Liters per capita per day (LPCD) of water access
	Training number	Integer (0-9)	Total training sessions participated at the gram panchayat level
Background variables	BPL %	Numeric (0-1)	Percentage of households below poverty line
	SC %	Numeric (0-1)	Percentage of scheduled caste population
	ST %	Numeric (0-1)	Percentage of scheduled tribe population

With a binary outcome variable, this becomes a logistic regression model. Model selection is conducted through an exhaustive best subset selection process using AIC criteria. Compared to BIC, AIC takes into account complexities within the model system, which aligns with reality of the WaSH-health integrated model. Considering that water access can affect latrine functions, the interaction between IHHL coverage and water access is also considered in the model.

6.1.2 Results

Incomplete IHHL Coverage Case

Collinearity among independent variables are first considered through the correlation matrix plot in Figure 6-2. There are no alarming correlations, so we proceed with model selection.

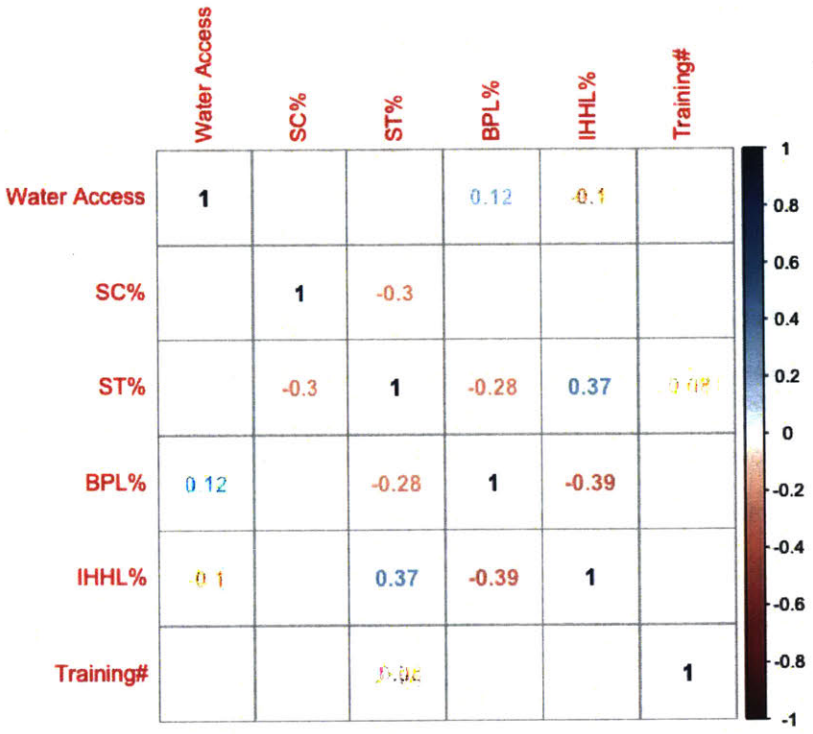


Figure 6-2: Correlation matrix among all independent variables

For habitations with less than 100% IHHL coverage (1610 entries), the following independent variables are selected by the AIC criteria in the final logistic regression model. Their coefficients and significant level are summarized in

Table 6-2.

Table 6-2: Summary table for logistic regression results (IHHL coverage < 100%).

Significance codes (as in R): 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**

Variable	Coefficient estimate	Standard error	Significance level
Water Access	-0.007982	0.006756	0.2374
IHHL coverage	-0.447176	0.286571	0.1187
Training number	0.172535	0.080773	0.0327 *
ST (%)	1.571377	0.229182	7.06E-12 ***

As expected, results suggest that increasing IHHL coverage results are correlated with slightly lower odds of water contamination within a habitation. The effect, however, is not drastic. With each additional 10% in IHHL coverage, the odds of finding biological contamination in the habitation only decreases to approximately 96% of the original level. The correlation is also not significant.

Increasing training are correlated with increasing odds of contamination, which is slightly counterintuitive. However, in Figure 6-1, there is a potential neglected link between the water quality "state" and the local training "action" – low water quality may increase the amount of mandated training for the local panchayat. In addition, panchayats that received more training may conduct more rigorous FTK testing, which can increase the likelihood of contamination discovery. These associated factors may have countered training's the positive effect on water quality, resulting in the negative correlation between training number and water safety.

Water access also is also correlated with a decreased odds of water contamination, as expected. However, the coefficient is small and the correlation is not significant.

Considering that the correlation between the key variables and the water quality outcome is at most weakly significant, all coefficients should be interpreted with caution.

On the other hand, the background variable of scheduled tribe population percentage is the only highly significant independent variable. An increase in ST population by 10% would increase the odds of contamination by 17%. Considering the historically disadvantaged status of ST population, its correlation with increased water contamination is expected.

Complete IHHL Coverage Case

For habitations with 100% IHHL coverage (2225 entries), the following independent variables are selected by the AIC criteria in the final logistic regression model. Their coefficients and significant level are summarized in Table 6-3.

Table 6-3: Summary table for logistic regression results (IHHL coverage = 100%).

Variable	Coefficient estimate	Standard error	Significance level
Water Access	-0.009814	0.005495	0.0741 .
ODF category 1: Declared but not verified	-1.270748	0.192661	4.23E-11 ***
ODF category 2: ODF verified	-1.88168	0.356234	1.28E-07 ***
SC %	9.466329	2.197612	1.65E-05 ***
ST %	1.32967	0.268463	7.31E-07 ***
BPL %	-0.94517	0.305303	0.00196 **

To begin with, the distribution of habitations among the 3 difference ODF categories are shown in Figure 6-3. After restricting to full IHHL coverage, the discrepancy among the different categories are more reasonable compared to Figure 5-16.

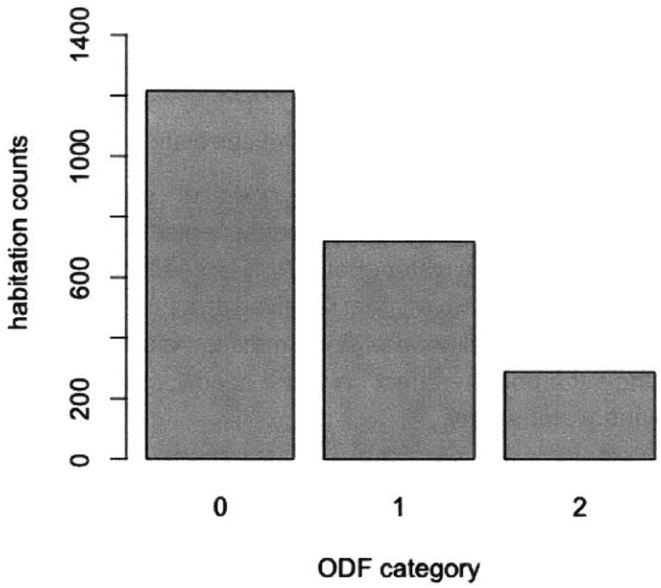


Figure 6-3: Habitation (with 100% IHHL coverage) distribution among the different ODF category

As expected, results suggest that declaration and verification of ODF is strongly associated with decreased likelihood of water quality contamination. Compared to other habitations with 100% IHHL coverage in villages that have not yet declared ODF, habitations in ODF-declared villages have only a 28% odds of finding biological water contamination. This ratio decreases to 15% for habitations in ODF-verified villages.

Increased water access is only weakly associated with decreased likelihood of contamination. The effect is also weaker, where each additional 10 LPCD water access rate brings about a 10% decrease in the odds of contamination. Training numbers are no longer selected for this model.

All background variables are also moderately to highly significant. The association between increasing SC, ST ratios and increasing contamination is expected. However, it is interesting to note that increasing BPL% is associated with decreasing odds of contamination. This may result from the fact that BPL households are often given priority in water and sanitation initiatives, such as a larger subsidy for latrine construction compared to APL households. BPL% is also associated with a lower likelihood of contamination among habitations that have not achieved 100% IHHL, as shown in in

Table 6-2, but the correlation is not significant.

Considering the potential association between BPL% and ODF status, multicollinearity in this model is tested through variance inflation factors with no resulting significance. Interaction terms have been considered but they are excluded in the model selection process.

6.1.3 Summary

As shown in the results above, among actions of IHHL construction (i.e. physical improvements) and ODF status achievement (i.e. additional behavioral improvements), only ODF status are significantly associated with decreasing levels of biological water contamination. ODF practices after latrine progress completion are associated with as much as an 85% reduction in the odds of contamination,

in the case of a verified ODF status. Despite a generally weak correlation, water access level still becomes more relevant after the latrine coverage completion.

6.2 Outbreak outcome

6.2.1 Analysis

The next segment along the DPSEEA framework is the State + Action – (Exposure) – Health Effect association, with outbreak as the ultimate outcome. Relevant variables in this section of the DPSEEA framework can be shown in Figure 6-4.

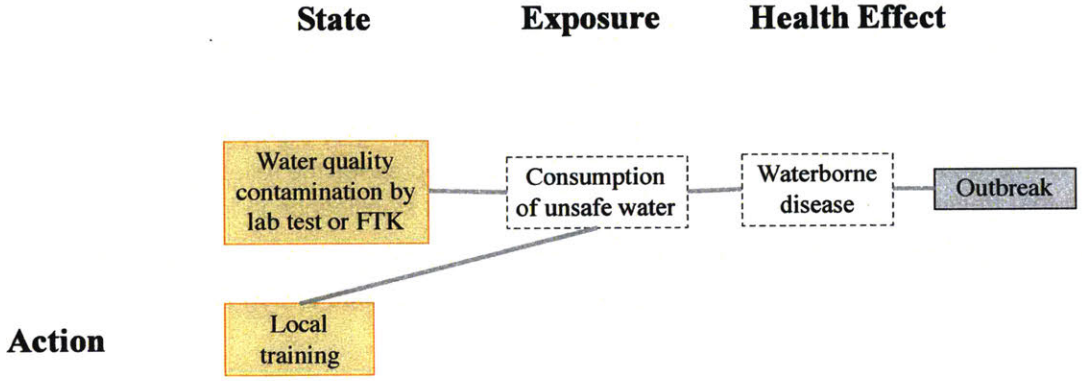


Figure 6-4: Action + State – Exposure – Effect segment of the DPSEEA framework for waterborne diseases

In this chain, the goal is to evaluate the connection between consumption of biologically contaminated water and the ultimate health effects. However, with no exposure information available in the integrated database, we have to consider water quality contamination and water awareness training as an indicator for the likelihood of contaminated water consumption. Similarly, waterborne disease data are not available, and outbreaks are the only alternative indicator of only the most severe disease clusters. The limitations of these proxies are analyzed previously.

Consequently, the goal above is transformed to the analysis of water quality contamination and trainings and their relationship with outbreak occurrence.

Key variables included in the model are listed in Table 6-4. All variables are analyzed at the habitation level. Background variables are similarly identified.

Table 6-4: Key variables in the Action + State – Exposure – Effect model

Category	Variable	Data type	Definition
Outcome	Outbreak occurrence	Binary (0 or 1)	Whether outbreak has occurred in the village that the habitation is located in.
Key independent variables	Biological contamination	Binary (0 or 1)	Whether biological contamination has been found in any of the sources in the habitation, either through lab tests or FTK tests.
	Training number	Integer (0-9)	Total training sessions participated at the gram panchayat level

Background variables	BPL %	Numeric (0-1)	Percentage of households below poverty line
	SC %	Numeric (0-1)	Percentage of scheduled caste population
	ST %	Numeric (0-1)	Percentage of scheduled tribe population

While the number of population affected by the outbreaks is available, without exposure data, the scale of diseases and outbreaks cannot be reflected in water contamination status. Hence, the binary outcome of outbreak occurrence is selected. While biological contamination can also be represented through the ratio of sources or samples with contamination found, there is insufficient information on water quality test coverage to determine whether these ratios are representative of the entire habitation.

It is not possible to separate sanitation training from water quality training, so the same training number variable is used across the models. Since most water quality training is related to FTK usage and education, it may be possible to approximate water quality only training by the frequency of FTK tests per source. However, the lack of an accurate source count also decreases the reliability of such factors. Considering that training number only showed moderate to weak significance in the action-pressure-state chain, it is still included in this model for analysis.

A logistic regression model is run, and model selection is again conducted through an exhaustive best subset selection with AIC criteria.

6.2.2 Results

Model selection by AIC resulted in the following independent variables as shown in Table 6-5. The fit overall was poor and none of the independent variable showed any significance.

Table 6-5: Summary table for logistic regression results

Variable	Coefficient estimate	Standard error	Significance level
Biological contamination	0.58892	0.42256	0.163
Training number	0.20904	0.13332	0.117
BPL %	0.01087	0.5008	0.983

As expected, contamination records have a positive correlation with the odds of outbreak occurrence, where existence of biological contamination is associated with an 80% increase in the odds of outbreak occurrence. However, the result is not significant and the standard error is relatively high.

The case is similarly with the training number variable. Similar to the positive correlation with water contamination, increased training is also associated with increased outbreak occurrence likelihood. This points to further investigation on how training participation is decided. It seems likely that factors along the DPSEEA chain – such as high contamination or high disease rates – are one of the determinants for increased training requirements. In the end, however, the high standard error also renders the correlation relatively insignificant.

6.2.3 Summary

Compared to the Action-Pressure-State model in 6.1, the fit of this model is significantly poorer. One of the most obvious cause is the lack of disease data. Outbreak occurred in only 49 out of a total of 5356 habitations in the dataset. Many issues resulting from poor water quality are not translated into the outbreaks. In addition, outbreak recognition also depends on many extra factors that are not accounted for in the waterborne disease DPSEEA model chain. For example, effectiveness of local health workers and rapid response teams may have larger influence the reporting and detection of outbreaks. None of these factors are accounted for.

The lack of exposure data also affected the specification of the model. With limited information on the population exposed to contamination, it is challenging to estimate the likelihood of large scale disease occurrence.

Overall, the State + Action – (Exposure) – Health Effect associations cannot be effectively concluded due to the lack of critical variables.

6.3 Effectiveness and limitations

Based on the analysis above, the DPSEEA model can be updated to include the extra information on strength of association between the relevant variables (Figure 6-5). The blue dotted lines are not part of the DPSEEA framework. They only represent the significant association discovered through statistical analysis. Based on these significant relationships, we can conclude on significant associations in the actual DPSEEA casual chain framework, as shown by the blue lines from ODF to "open excreta" to water quality contamination. No conclusions can be drawn on the rest of the associations along the chain.

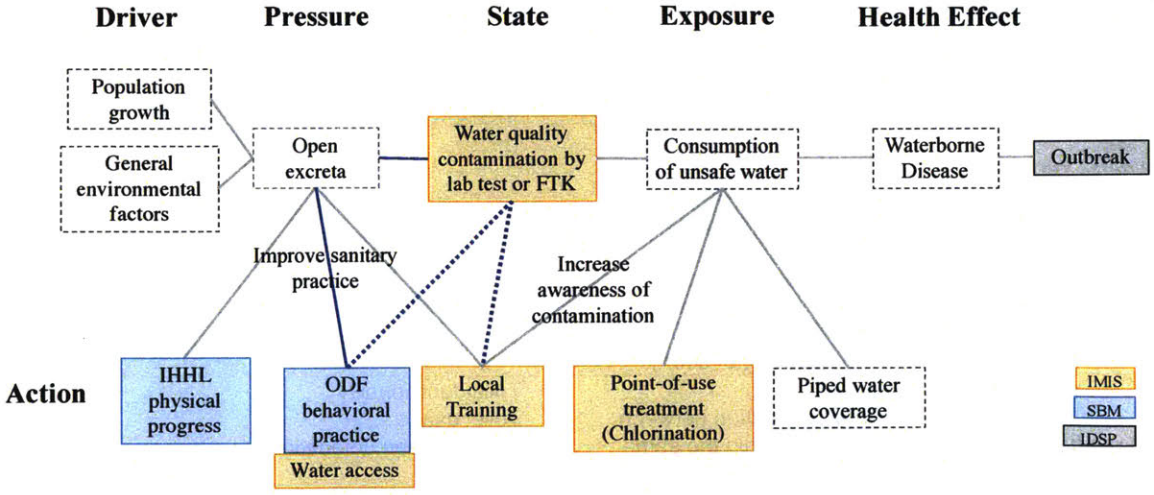


Figure 6-5: DPSEEA framework for waterborne diseases in the India context, with updated strength of association. All grey lines are association that still remain unknown. Blue dotted lines are the significant (but indirect) relationships found through statistical analysis. Blue lines are estimated significant association based on the blue dotted lines.

More specifically, ODF behaviors are associated with decreased likelihood of water quality contamination. Based on the framework, we can hypothesize that this is due to the "action – pressure - state" chain, where positive sanitary actions resulted in less open excreta, eventually leading to less biologically contaminated water sources. Association between training and decreased water quality status is also observed statistically, but currently it is not possible to map this association onto the framework chains, due to insufficient information on training contents and reasons behind training participation.

Overall, increased information on the integrated DPSEEA model can allow for effective risk assessments. For example, with the associations displayed in Figure 6-5, it is possible to estimate the difference in water contamination likelihood between villages that have failed ODF verification and villages that passed. The increased ability to predict likelihood of hazards along the WaSH-Health framework is a significant positive benefit that only such an integrated system can bring forth.

There is a lot more potential to be explored with this integrated system, especially considering the numerous associations along the DPSEEA chain that are not yet concluded. A number of additional actions are required to increase its effectiveness for the purpose of outbreak control.

Increased data completeness

The majority of inconclusive evidence is due to incomplete datasets. Most of the datasets of interest are available, but they are either in accessible or have too many missing entries to be included in the study, as shown in Table 6-6.

While there is a much higher cost to obtain datasets that are not available, it is viable to find alternative channels to access data or enforce better collection of data. Hence, data in the first two rows of Table 6-6 are likely to become integration-ready in the near future. These crucial datasets can contribute significantly to the improved effectiveness of a WaSH-integrated data system.

Table 6-6: Incomplete datasets excluded from the current integrated database

Incomplete datasets	IMIS	SBM	IDSP
Missing variables / Questionable quality	Sanitary survey results Lab results (biological) Trainee numbers Water sources inventory		
Challenging to access	Population served by source/scheme Contamination status of source/scheme Functionality of source	SBM mobile data	SPL form disease surveillance results
Not available	Interconnection between sources within the same habitation	Toilet functionality Toilet usage General open defecation status Hygiene practices	

Understanding confounding variables

With the assistance of the DPSEEA framework and statistical analysis, it becomes possible to draw conclusions on relationships between the variables. However, there are possible confounding variables that we are leaving out of the framework, which may cause us to draw the wrong conclusions on associations and casual relationships.

For example, if we take a closer look at some of the critical variables included in the models, as shown in Figure 6-6, there is a clear segment of ODF category data at the bottom half of the table showing exceedingly good performance. Referring back to the original data table, we can see that this segment is corresponding to Vansda block in Navsari district. According to the table plot, 0 biological contamination is almost shown consistently across the same block segment. While this may be indication that good water quality is associated with ODF practices, this may also be a result of a very efficient block-level government. In addition, considering that this block is almost 100% tribal, it may have been covered in priority or pioneering programs, resulting in the high level of achievement.

If we leave out Vansda block completely, results from the regression model in Section 6.1.2 (Complete IHHL coverage case), especially coefficients and significance level of the ODF categorical variable, would be affected to a large extent. As shown in the updated logistic regression summary in Table 6-7, the ODF verification status is no longer significant. While the ODF declaration status (without verification) is still significant, it is now correlated with an increased likelihood of water quality contamination.

Data from Vansda block drastically affect the model. Without further understanding of the exact cause of its anomalous performance, inclusion of the data may produce a highly misleading model.

Table 6-7: Updated summary table of the logistic regression for IHHL > 100%, excluding Vansda block

Variable	Coefficient estimate		Standard error		Significance level	
	Updated result	Original result	Updated result	Original result	Updated result	Original result
Water Access	-0.020948	-0.009814	0.006308	0.005495	0.000898***	0.0741 .
ODF category 1: Declared but not verified	0.630939	-1.270748	0.21228	0.192661	0.002957**	4.23E-11 ***
ODF category 2: ODF verified	-0.451632	-1.88168	0.374511	0.356234	0.227847	1.28E-07 ***
SC %	4.178232	9.466329	2.307356	2.197612	0.070167.	1.65E-05 ***
ST %	2.333931	1.32967	0.287335	0.268463	4.56E-16***	7.31E-07 ***
BPL %	-0.715709	-0.94517	0.300972	0.305303	0.017407*	0.00196 **

Overall, strict governmental regulation enforcement, or any other special treatment of a specific region, can be a confounding variable that affects performance across WaSH sectors, resulting in a significant correlation that may not actually imply direct connection or causation. Accounting for these types of confounding variables and irregular data clusters are essential to the ultimate effectiveness of the integrated system. Considering that many governmental practices vary significantly by locality, difference across administrative units may need to be taken into consideration to produce a reliable model.

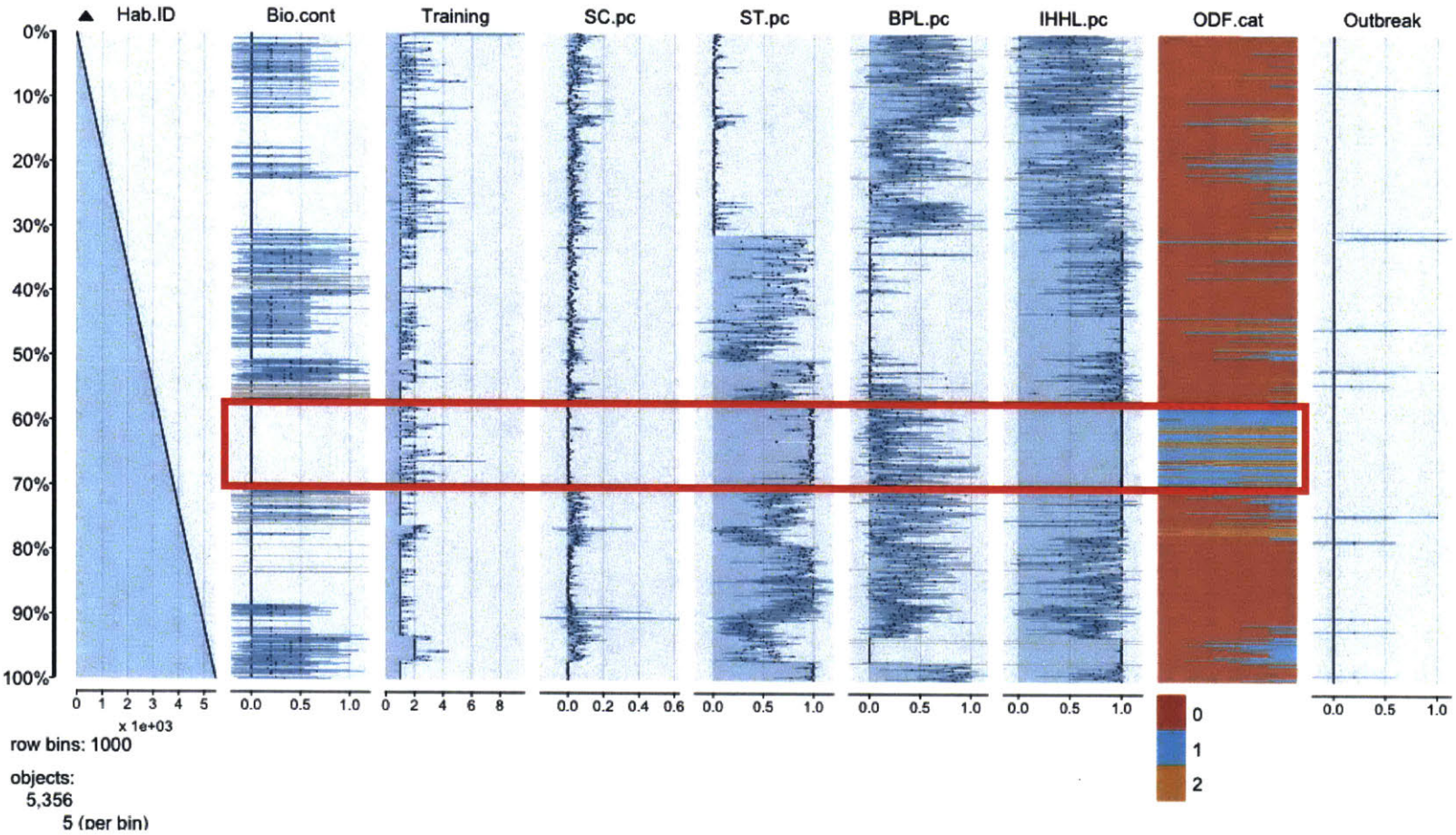


Figure 6-6: Table plot of critical variables included in the DPSEEA models. The variables are again ordered by district, block, panchayat, village and habitation name accordingly. Variables from left to right include: habitation index, biological contamination (binary), training number, SC%, ST%, BPL%, IHHL%, ODF category and outbreak occurrence (binary).

Mixed methods

While qualitative study has been conducted through interviews at agencies across the WaSH-health sectors to gain an understanding of interagency barriers, similar studies are not yet carried out to qualitatively explore the association between the different DPSEEA segments. For example, to understand the negative relationship between water safety and increase levels of training, field surveys are needed to investigate the requirements, content and purpose of all the training sessions.

In addition, ambiguities within the data can only be better explored through a qualitative process. For example, the exact status of villages that declared ODF but have not been confirmed after over 1 year is unclear. They may have failed the first round of inspection and still does not meet ODF standards a year later. We are setting them in the same category as villages that just declared ODF and have yet to have gone through the verification process. With a better understanding of the implications behind these status, the variables may be much better configured to support a more rigorous outbreak control model.

Risk Prediction

As mentioned in the literature review, the final goal of the risk assessment systems such as the integrated DPEESA model is to effectively assess and predict contamination and outbreak risks so that the risks can be effectively managed. The effectiveness of such a model depends on the strength of association between factors along the DPSEEA chain, but more critically, it also depends on how correctly the model can predict risks.

While it is possible to segment the current dataset into training and test data to observe the accuracy of the system in predicting outcomes, there are still too many data quality and completeness issues for the prediction results to be meaningful. Thus, machine learning methods are not implemented at the current stage. However, despite challenging barriers to data integration, basic statistical analysis still shows promising results. Significant relationships are found. Potential confounding variables are identified through visualizing the data distribution.

In conclusion, compared to general correlation studies and risks assessments, this integrated DPSEEA framework as adapted in the India context holds much greater potential. Strength of association is already found through preliminary analysis. It is possible to fill many of the data gaps through obtaining access to existing data and working with the government to improve data collection and validation processes. A more robust and effective model is highly feasible.

7 CONCLUSION

Chapter 7 concludes on barriers identified throughout the study, and proposes pathways forward. Barriers to interagency collaboration and possible solutions are outlined in Section 7.1. Barriers including data inconsistency, data quality and overall integrated database design issues are outlined in Section 7.2 and 7.3. Barriers to a comprehensive cost-effectiveness analysis of the integrated framework are described in Section 7.4. Section 7.5 points to the future pathways toward implementing the final integrated water, sanitation and health framework in outbreak control processes for rural India

CONCLUSION

The WaSH-health integrated approached to disease and outbreak management based on the DPSEEA framework shows significant promise in the WaSH-health monitoring landscape in Gujarat, India. Health effect of any positive interventions in the WaSH system can be evaluated along the casual chain. Negative risks within the integrated system can also be evaluated along the chain to estimate their impacts on disease and outbreak occurrence. Effective utilization of this approach may generate useful evidence for decision support in the WaSH-health governance system, and improve the improve the control of waterborne diseases and outbreaks.

Preliminary steps to create the integrated model already demonstrated encouraging results where a significant association among the "action", "pressure" and "state" factors can be concluded. Through analyzing all the associations along the chain, the ideal decision support system can be created to address the root causes leading to waterborne disease outbreak. On the other hand, many barriers to the implementation of such an integrated approach have also been identified throughout the study. They include agency collaboration barriers, data integration barriers and model construction barriers, all of which can increase the challenge of setting up this integrated system and developing it to its full potential. Pathways towards overcoming the implementation gaps are consequently analyzed in this study.

A conclusive summary of major barriers and likely solutions are laid out in the following sections.

7.1 Interagency collaboration

Based on the discussion in Chapter 4, key barriers in interagency collaboration result from a lack of administrative incentive due to the narrowly defined procedure and action focused targets in each of the respective sector. These sector-based targets give rise to three very independently operating agencies with limited direct channels of cooperation.

Possible pathways to overcome these barriers concluded through interview analysis, and the concluding results are summarized in Table 7-1.

Table 7-1: Pathways towards interagency collaboration.

Barriers	Possible solutions
Lack of administrative motivation for routine collaborative practices	- ODF districts that are now aiming for ODF+ status has the strongest incentive for an integrated system and may be considered piloting locations. - State and district surveillance units as well as SBM consultants have a moderately high incentive and may be considered alliance in the effort. - For continued incentive, it would important to encourage enforcement of ODF+ or other cross-sector performance-based requirements, which are broader than the current narrow sector-based targets and can capture indicators across water, sanitation and health.

Lack of existing direct connection channels among the water, sanitation and health monitoring agencies	- Both disease surveillance units and WASMO are collecting water quality data at a local level, and this existing connection can be considered as a launching point to for collaboration expansion - WASMO and NIC are nodal agencies that all three sectors work directly with, and their role in the collaboration development process should be effectively utilized.

7.2 Data inconsistency

Based on the data characteristics and data utility analysis across the WaSH-health sectors in Chapter 5, key barriers to a consistent data system result from frequent changes to administrative units, concerning quality of data collected by grassroots workers or local representatives, and varying data collection process across sectors resulting in data misalignment.

Possible pathways to overcome data inconsistencies are concluded in Table 7-2.

Table 7-2: Pathways to overcome data inconsistency

Barriers	Possible solutions
Many updates and variation across the administrative units resulting in high integration challenge across database and across different years	- A more flexible cross-sector rural administration unit ID system should be instituted that allows high adaptability to the constant administrative region reconfigurations.
Concerns over the quality of FTK and sanitary surveys	- Data tables should have details on personnel conducting the tests and supervisors validating the results - FTK result table should contain information on the parameters tested and the test kit utilized. - Sanitary survey table require more standard definition of each entry and should contain details on the actual risks factors identified.
Lack of a reliable annual inventory of total functional sources	- A comprehensive review of all existing sources should be conducted to remove sources that are likely non-functional or duplicates of each other. - Investigation and possible consolidation of all supply scheme sources with the same source ID can will allow source ID to become the unique identifier of sources across the different source tables. - A better and consistent definition of source location entries may help with future consolidation of repeat sources. - Past snapshots of the water source inventory at the end of each fiscal year should be available. - All five lists of sources (private, public, supply scheme, delivery point, surface water) may be set up in parallel within the same dataset for convenience.
Contradictory results across databases	- All key discrepancies should be noted through cross-database comparisons of the same variable. - Further exploration of variations in the monitoring and collection process of these variables can identify the root cause of the discrepancy, as well as the necessary steps to resolve these discrepancies.

	- Uniform data collection protocol for the same set of variables should be instituted in the long run across the sectors.

7.3 Database design

Based on the database integration analysis in Chapter 5, key barriers to a high-functioning database result from lack of uniformity in the names of administrative unit – the key linking variable across the databases, ineffective primary key and foreign key setups within each database, and the lack of stability in database access (especially for IMIS).

Possible pathways towards a more reliable and well-performing integrated database system are concluded in Table 7-3.

Table 7-3: Pathways towards a high-performing integrated database

Barriers	Possible solutions
Frequent duplicate entries within a single data table	- Primary keys should be set up within each table to enforce a unique identifier of each entry within the table
Inconsistent entries across different data tables sharing the same set of variables	- Foreign keys should be set up so that any linking variable in a secondary data table must match the primary key of the base data table that it's linked to.
High variations in administrative names which cross-database links are dependent on	- Implementation of location and language-based advanced data matching algorithms that minimizes the need for manual name matching. - Considering that administrative units is the basis for all cross-database integrations, the implementation of a cross-sector administrative unit ID is essential. - In the case of IDSP, identifying the administrative unit coverage of each reporting unit is essential to integrate the disease surveillance results at the most desired granularity.
Concern over stability of data access	- Considering that IDSP has the most advanced data operating network among the three databases, it might lead the effort to setup a stably integrated data system that overcomes the webpage loading and data accessibility issues frequently encountered in IMIS and SBM.

7.4 Model creation

Based on the discussion in Chapter 6, while an integrated DPEESA model for outbreak control can be formulated, its effectiveness and full risk prediction capability is yet to fully demonstrated. Multiple barriers and constraints prevented the analysis on model effectiveness, including missing or incomplete datasets in key components along the casual chain of the model, the existence of possible confounding variables and the lack of sufficient qualitative information to further corroboration quantitative conclusions on the model.

Possible pathways to overcome these barriers are concluded in Table 7-4.

Table 7-4: Pathway towards demonstrating an effective WaSH-health model for ultimate outbreak control purposes

Barriers	Possible solutions
Challenge to identify confounding variables	- For all results related to water sources, a tracking note should be available if test results are related to each other, such as repeat tests, follow-up tests or concurrent tests that took each other's results into account of the other result. Unless otherwise noted, all testing records should ideally be independent from each other. - Differences in policy enforcement across administrative units should be noted and taken into account in data analysis. - Other
Datasets, especially biological water quality tests, with large numbers of missing entries	- A more consistent data collection coverage and data collection frequency should be set up across the key variables of interest. - For biological water quality parameters that may be more challenging to collect, chemical water quality parameters can be used as a flag indicator for the necessity of biological testing to reduce the overall required number of biological tests (only when absolutely necessary). - A better understanding of interconnection between local water sources can help extrapolate missing water quality data, which may also reduce the total required number of biological water tests (only when absolutely necessary).
Datasets that are missing from the integrated framework	- For datasets that are challenging to access – especially data on water contamination exposure and disease effects, it is necessary to find alternative channels to obtain data. - For datasets that are not yet available in the WaSH-health sectors, it may be possible to find similar data in other governmental sectors or public organizations. - For highly essential data, such as latrine functionality and hygiene practices, possible data collection routines should be considered.
Challenge in confirming the casual association along the DPSEEA chain	- Qualitative field investigations are necessary to supplement the quantitative analysis on associations between different components in the DPSEEA framework.

After following along these pathways to fill in the gaps of the model, it will become possible to demonstrate the capacity of the model through a comprehensive review on its cost-effectiveness. Ultimately, with these gaps closed, the same pathways can potentially go on to lead to the successful adoption of an integrated water, sanitation and health approach for outbreak control and management.

7.5 Future work

With numerous gaps to overcome before creation of an integrated system, future work will focus on bridging these gaps so that the cost-effectiveness of the system can eventually be demonstrated.

Specifically, to complete the WaSH-health integrated model, the following components of future studies are crucial:

- Gaining access to essential data that were challenging to access during the course of this study, such as SPL form data on waterborne disease and symptom occurrence, water scheme functionality and population coverage details, piped water coverage and so on
- Incorporating local policy enforcement as a potential confounding variable, and utilizing better statistical analysis and visualization methods that can help uncover other potential confounding variables
- Incorporating environmental drivers in the DPSEEA framework
- Expanding the scope of study to more districts across more years.

In addition, mixed methods studies are required to better understand ambiguity in the database which prevented effective analysis and model construction. These studies likely require interviews and collaborative investigations with local India government. Some of the key components of this aspect in future studies include:

- Understanding the reality behind the drastically shifting administrative units and their impacts on current datasets
- Understanding the discrepancy in the number of administrative units across the different databases
- Understanding the contents of local training, how they impact the DPSEEA chain and whether separate impacts can be accounted for differently.
- Exploring data discrepancies among the databases, such as the variation in total household number within the same habitation
- Review and consolidation of the existing water sources to create an updated water source inventory, so that water test and sanitary coverage rates become reliable indicators and can be effectively utilized in the integrated model
- Examination of potential interconnections of water sources within a habitation (e.g. if water scheme sources water quality is reflective of delivery point water quality, and if local groundwater wells are connected to the same groundwater channel) and evaluate the possibility of utilizing spatial extrapolation methods to estimate water quality parameters.

With improved data and improved database integration, a final cost effective analysis on the decision support power of the resulting model is necessary. The effects of any interventions along the DPSEEA chain on health and outbreak can be studied, and the validity of risk assessment and outcome predictions by the DPSEEA framework can be evaluated.

Lastly, apart from statistical models, the integrated framework can also utilize spatial analysis for outbreak risk assessments and geographical risk hotspot identification. The integrated framework may also be useful other components in outbreak control such as detection and outbreak investigation. All these prospects show great promise and should be explored in the future. A district-level pilot in collaboration with local agencies may be the logical next step to evaluate the promise of the abovementioned measures.

8 BIBLIOGRAPHY

Andersson, Y. and Bohan, P. (2001) 'Disease surveillance and waterborne outbreaks', in WHO (ed.) *Water Quality: Guidelines, Standards and Health*. London, UK: IWA Publishing, pp. 115–133.

Babin, S. M., Burkom, H. S., Mnatsakanyan, Z. R., Ramac-Thomas, L. C., Thompson, M. W., Wojcik, R. a, Lewis, S. H. and Yund, C. (2008) 'Drinking Water Security and Public Health Disease Outbreak Surveillance', *Johns Hopkins APL Technical Digest*, 27(4), pp. 403–411.

Bain, R. E., Gundry, S. W., Wright, J. A., Yang, H., Pedley, S. and Bartram, J. K. (2012) 'Accounting for water quality in monitoring access to safe drinking-water as part of the Millennium Development Goals: lessons from five countries', *Bull World Health Organ*, 90(3), p. 228–235A. doi: 10.2471/blt.11.094284.

Bartram, J., Brocklehurst, C., Fisher, M. B., Luyendijk, R., Hossain, R., Wardlaw, T. and Gordon, B. (2014) 'Global monitoring of water supply and sanitation: history, methods and future challenges.', *International journal of environmental research and public health*, 11(8), pp. 8137–65. doi: 10.3390/ijerph110808137.

Bartram, J. and Cairncross, S. (2010) 'Hygiene, sanitation, and water: Forgotten foundations of health', *PLoS Medicine*, 7(11), pp. 1–9. doi: 10.1371/journal.pmed.1000367.

Buehler, J. W., Hopkins, R. S., Overhage, M., Sosin, D. M. and Tong, V. (2004) 'Framework for Evaluating Public Health Surveillance Systems for Early Detection of Outbreaks', *Morbidity and Morality Weekly Report*, 53(RR05), pp. 1–11.

Bureau of Indian Standards (2012) *Indian Standard: Drinking water - Specification (Second Revision)*. New Delhi.

Burkom, H. S., Ramac-Thomas, L., Babin, S., Holtry, R., Mnatsakanyan, Z. and Yund, C. (2011) 'An integrated approach for fusion of environmental and human health data for disease surveillance', *Statistics in Medicine*, 30(5), pp. 470–479. doi: 10.1002/sim.3976.

Cairncross, S., Bartram, J., Cumming, O. and Brocklehurst, C. (2010) 'Hygiene, Sanitation, and Water: What Needs to Be Done?', *PLoS Medicine*, 7(11), p. e1000365. doi: 10.1371/journal.pmed.1000365.

Cassady, J. D., Higgins, C., Mainzer, H. M., Seys, S. a, Sarisky, J., Callahan, M. and Musgrave, K. J. (2006) 'Beyond compliance: environmental health problem solving, interagency collaboration, and risk assessment to prevent waterborne disease outbreaks.', *Journal of epidemiology and community health*, 60(8), pp. 672–674. doi: 10.1136/jech.2005.040394.

CDC (2001) 'Updated Guidelines for Evaluating Public Health Surveillance Systems Recommendations from the Guidelines Working Group', *Morbidity and Morality Weekly Report*, 50(RR13), pp. 1–35.

Centers for Disease Control and Prevention (2013) 'Surveillance for Waterborne Disease Outbreaks Associated with Drinking Water and Other Nonrecreational Water — United States, 2009–2010', *MMWR Morb Mortal Wkly Rep*, 62(35), pp. 714–20. Available at: http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6235a6.htm.

CMS India (2011) *Assessment Study of Impact and Sustainability of Nirmal Gram Puraskar*.

Commissionerate of Rural Development (2016) *Swachh Bharat Mission (Gramin) Annual Implementation Plan 2016-17*.

Craun, G. F., Frost, F. J., Calderon, R. L., Hilborn, E. D., Fox, K. R., Reasoner, D. J., Poole, C. L., Rexing, D. J., Hubbs, S. a and Dufour, a P. (2001) 'Improving waterborne disease outbreak investigations', *International journal of environmental health research*, 11(3), pp. 229–243. doi: 10.1080/09603120120070847.

Department of Health & Family Welfare (2014) *Assessment of Water, Sanitation and Hygiene services and practices at Functional Delivery Points of 8 High Priority Districts, Gujarat (Period.* Gandhinagar, Gujarat.

Directorate of Economics and Statistics (2012) *Statistical Outline Gujarat State 2012.* Gandhinagar, Gujarat.

Eastern Research Group Inc. (1992) *Inventory of Exposure-related Data Systems Sponsored by Federal Agencies.* Lexington,.

Escamilla, V., Wagner, B., Yunus, M., Streatfield, P., van Geen, A. and Emch, M. (2011) 'Effect of deep tube well use on childhood diarrhoea in Bangladesh', *Bulletin of the World Health Organization*, 89(7), pp. 521–527. doi: 10.2471/BLT.10.085530.

Fewtrell, L. and Colford, J. M. (2004) 'Water, Sanitation and Hygiene: Interventions and Diarrhoea: A Systematic Review and Meta-analysis', *The International Bank for Reconstruction and Development / The World Bank* , (July).

Gelting, R., Sarisky, J., Selman, C., Otto, C., Higgins, C., Bohan, P. O., Buchanan, S. B. and Meehan, P. J. (2005) 'Use of a systems-based approach to an environmental health assessment for a waterborne disease outbreak investigation at a snowmobile lodge in Wyoming', *International Journal of Hygiene and Environmental Health*, 208(1–2), pp. 67–73. doi: 10.1016/j.ijheh.2005.01.009.

Gentry-Shields, J. and Bartram, J. (2014) 'Human health and the water environment: Using the DPSEEA framework to identify the driving forces of disease', *Science of the Total Environment.* Elsevier B.V., 468–469, pp. 306–314. doi: 10.1016/j.scitotenv.2013.08.052.

George, J., An, W., Joshi, D., Zhang, D., Yang, M. and Suriyanarayanan, S. (2015) 'Quantitative Microbial Risk Assessment to Estimate the Health Risk in Urban Drinking Water Systems of Mysore, Karnataka, India', *Water Quality, Exposure and Health.* Springer Netherlands, 7(3), pp. 331–338. doi: 10.1007/s12403-014-0152-4.

Gerstman, B. (2003) 'Outbreak investigation', in *Epidemiology Kept Simple: An Introduction to Traditional and Modern Epidemiology.* 2nd edn, pp. 351–364. Available at: http://ocw.jhsph.edu/courses/fundepi/PDFs/Lecture2.pdf.

Godfrey, S., Labhasetwar, P., Wate, S. and Pimpalkar, S. (2011) 'How safe are the global water coverage figures? Case study from Madhya Pradesh, India', *Environmental Monitoring and Assessment*, 176(1–4), pp. 561–574. doi: 10.1007/s10661-010-1604-3.

Hlaing, Z. N., Mongkolchati, A. and Rattanapan, C. (2016) 'A Household Level Analysis of Water Sanitation Associated with Gastrointestinal Disease in an Urban Slum Setting of South Okkalapa Township, Myanmar', *EnvironmentalAsia*, 2(9), pp. 91–100. doi: 10.14456/ea.2016.12.

Hunter, P., Andersson, Y., Von Bonsdorff, C., Chalmers, R., Cifuentes, E., Deere, D., Endo, T., Kadar, M., Krogh, T., Newport, L., Prescott, A. and Robertson, W. (2003) 'Surveillance and investigation of contamination incidents and waterborne outbreaks', in *Assessing microbial safety of drinking water: Improving approaches and methods.* Cornwall, UK: IWA Publishing, pp. 205–36. doi: 10.1787/9789264099470-en.

IDSP (2015) *Disease Surveillance Under IDSP: Manual for Health Workers.*

IDSP (no date a) *Integrated Disease Surveillance Project.* Available at: http://idsp.nic.in/index4.php?lang=1&level=0&linkid=313&lid=1592 (Accessed: 1 December 2016).

IDSP (no date b) *MODULE A: Data Reporting and Analysis.*

IDSP (no date c) *Training Manual on Data Management.*

IMIS (no date) *Integrated Management Information System (IMIS).* Available at: http://indiawater.gov.in/IMISReports/MenuItems/AboutSite.aspx (Accessed: 20 June 2012).

Jalava, K., Rintala, H., Ollgren, J., Maunula, L., Gomez-Alvarez, V., Revez, J., Palander, M., Antikainen, J., Kauppinen, A., Räsänen, P., Siponen, S., Nyholm, O., Kyyhkynen, A., Hakkarainen, S., Merentie, J., Pärnänen, M., Loginov, R., Ryu, H., Kuusi, M., Siitonen, A., Miettinen, I., Santo Domingo, J. W., Hänninen, M. L. and Pitkänen, T. (2014) 'Novel microbiological and spatial statistical methods to improve strength of epidemiological evidence in a community-wide waterborne outbreak', *PLoS ONE*, 9(8). doi: 10.1371/journal.pone.0104713.

Jalba, D. I., Cromar, N. J., Pollard, S. J. T., Charrois, J. W., Bradshaw, R. and Hrudey, S. E. (2010) 'Safe drinking water: Critical components of effective inter-agency relationships', *Environment International*. Elsevier Ltd, 36(1), pp. 51–59. doi: 10.1016/j.envint.2009.09.007.

Katakwar, M. (2016) 'Narmada river water: Pollution and its impact on the human health', *International Journal of Chemical Studies*, 4(2), pp. 66–70.

Khan, R., Phillips, D., Fernando, D., Fowles, J. and Lea, R. (2007) 'Environmental health indicators in New Zealand: Drinking water - A case study', *Ecohealth*, 4(1), pp. 63–71. doi: DOI 10.1007/s10393-007-0089-1.

Kumar, A., Goel, M., Jain, R. and Khanna, P. (2014) 'Tracking the implementation to identify gaps in integrated disease surveillance program in a block of district Jhajjar (Haryana)', *Journal of Family Medicine and Primary Care*, 3(3), p. 213. doi: 10.4103/2249-4863.141612.

Liu, S., McGree, J., Hayes, J. F. and Goonetilleke, A. (2016) 'Spatial response surface modelling in the presence of data paucity for the evaluation of potential human health risk due to the contamination of potable water resources', *Science of the Total Environment*. Elsevier B.V., 566–567, pp. 1368–1378. doi: 10.1016/j.scitotenv.2016.05.200.

MDWS (2013) *Uniform Water Quality Monitoring Protocol*. New Delhi, India.

Medema, G. J., Payment, P., Dufour, A., Robertson, W., Waite, M., Hunter, P., Kirby, R. and Andersson, Y. (2003) 'Safe Drinking Water: An Ongoing Challenge', in OECD and WHO (eds) *Assessing Microbial Safety of Drinking Water: Improving Approaches and Methods*. Cornwall, UK: IWA Publishing, pp. 11–46. doi: 10.1016/S0048-9697(04)00275-X.

Ministry of Drinking Water and Sanitation (2011) *Strategic Plan – 2011- 2022: Department of Drinking Water and Sanitation – Rural Drinking Water*.

Ministry of Drinking Water and Sanitation (2013a) *National Rural Drinking Water Programme Guidelines*.

Ministry of Drinking Water and Sanitation (2013b) *Uniform Water Quality Monitoring Protocol*. New Delhi, India.

Ministry of Drinking Water and Sanitation (2016) 'Annual Report 2015-16', (April).

Ministry of Drinking Water and Sanitation Government of India (2014) *Guidelines for Swachh Bharat Mission (Gramin)*. Available at: http://www.oecs.org/about-the-oecs/mission-a-objectives.

Mohanty, F. C. (1997) *Environmental health risk analysis of drinking water and lead in Hyderabad city, India*. Harvard University, Cambridge, MA.

National Academy Press (1983) *Risk Assessment in the Federal Government: Managing the Process*. Washington, D.C.

National Homeland Security Research Center (2012) *Electronic Surveillance System for the Early Notification of Community-Based Epidemics (ESSENCE) Water Security Module*.

Niemeijer, D. and de Groot, R. S. (2008) 'A conceptual framework for selecting environmental indicator sets', *Ecological Indicators*, 8(1), pp. 14–25. doi: 10.1016/j.ecolind.2006.11.012.

Niskar, A. S. (2007) 'Environmental data assessment for use in public health surveillance.', *Journal of environmental health*, 70(4), pp. 43–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18044253.

Novellino, M. (2015) *Analysis of Slipback of Rural Water Supply Systems in India using FIETS Framework and IMIS database – Gujarat Case Study by.*

Parsai, A. and Rokade, V. (2016) 'Study of national rural drinking water programme implementation during last five years (from FY 2010-2011 to 2014-2015) in state of Madhya Pradesh, India', *Journal of Water, Sanitation and Hygiene for Development*, 6(1), pp. 170–183. doi: 10.2166/washdev.2016.155.

Rajiv Gandhi National Drinking Water Mission (2006) *Guidelines for National Rural Drinking Water Quality Monitoring and Surveillance Programme Department of Drinking Water Supply Ministry of Rural Development Government of India.* New Delhi.

Rheingans, R., Dreibelbis, R. and Freeman, M. C. (2006) 'Beyond the Millennium Development Goals: public health challenges in water and sanitation.', *Global public health*, 1(1), pp. 31–48. doi: 10.1080/17441690500443139.

Rizak, S. and Hrudey, S. E. (2007) 'Achieving safe drinking water — risk management based on experience and reality', *Environmental Reviews*, 15, pp. 169–174. doi: 10.1139/A07-005.

Schuster, C. J., Ellis, A. G., Robertson, W. J., Charron, D. F., Culleton, N., Donovan, C., Fraser, D., Fyfe, M., Gignac, M., Hutchinson, S., Isaac-renton, J., Moores, B., Keefe, K. O., Stanley, R., Marshall, B. J. and Medeiros, D. T. (2001) 'Infectious Disease Outbreaks Related to Drinking Water in Canada, 1974-2001', *Canadian Journal Of Public Health*, 96(4).

Schwemlein, S., Cronk, R. and Bartram, J. (2016) 'Indicators for Monitoring Water, Sanitation, and Hygiene: A Systematic Review of Indicator Selection Methods', *International Journal of Environmental Research and Public Health*, 13(3), p. 333. doi: 10.3390/ijerph13030333.

Sexton, K., Selevan, S. G., Wagener, D. K. and Lybarger, J. A. (1992) 'Estimating human exposures to environmental pollutants: availability and utility of existing databases', *Archives of Environmental Health*, 47(February 2015), pp. 398–407. doi: 10.1080/00039896.1992.9938381.

Strosnider, H., Zhou, Y., Balluz, L. and Qualters, J. (2014) 'Engaging academia to advance the science and practice of environmental public health tracking', *Environmental Research*, 134, pp. 474–481. doi: 10.1016/j.envres.2014.04.039.

Taylor, D. L., Kahawita, T. M., Cairncross, S. and Ensink, J. H. J. (2015) 'The impact of water, sanitation and hygiene interventions to control cholera: A systematic review', *PLoS ONE*, 10(8), pp. 1–20. doi: 10.1371/journal.pone.0135676.

Tillett, A. H. E., Louvois, J. De and Wall, P. G. (1998) 'Surveillance of Outbreaks of Waterborne Infectious Disease: Categorizing Levels of Evidence', *Epidemiology and Infection*, 120(1), pp. 37–42.

Wescoat, J. L., Fletcher, S. and Novellino, M. (2016) 'National rural drinking water monitoring: progress and challenges with India's IMIS database', *Water Policy*, pp. 1–18. doi: 10.2166/wp.2016.158.

Wolff, C., Eng, M. S., Vaidyanathan, A. and Enve, M. S. (2008) 'Drinking Water and Environmental Public Health Tracking : Assessing Statewide and Local Water Quality Data Sources', pp. 1–9.

Yoder, J., Roberts, V., Craun, G. F., Hill, V., Hicks, L. a, Alexander, N. T., Radke, V., Calderon, R. L., Hlavsa, M. C., Beach, M. J. and Roy, S. L. (2008) 'Surveillance for waterborne disease and outbreaks associated with drinking water and water not intended for drinking- United States, 2005-2006', *Morbidity and mortality weekly report*, 57(9), pp. 39–62. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21937977.