Pune Epidemiological Database Project
PEP DATA PROJECT

Version 1.0
Date: 23-Feb-2021

A Pune Knowledge Cluster Initiative

Summary

<table>
<tr><td>Title</td><td colspan="2">Pune Epidemiological Database Project (PEP-DATA): A Pune Knowledge Cluster (PKC) Initiative</td></tr>
<tr><td>Vision</td><td colspan="2">Establish an integrated epidemiological database in Pune district that would facilitate advanced analysis and prediction of outbreaks. The database that will be scalable, secure, open source and interoperable will help strengthen the health systems through responsive and evidence-based decisions on all public health policies.</td></tr>
<tr><td rowspan="4">Approach</td><td>Objectives</td><td>Timeline</td></tr>
<tr><td>1. Assessment of the landscape of data sources in public health systems
Understand and assess the complex data ecosystem of various public health data sources and reporting mechanisms. The formative assessments will be carried out in public sector health facilities.</td><td>12 months</td></tr>
<tr><td>2. Design and build an epidemiologic database module for acute febrile illnesses.
Identify minimal variables that need to be collected for reportable diseases and for identifying outbreaks but also meet the global data collection standards. Based on these, PEP-data database will be designed that can be uniformly implemented for different health facility levels.</td><td>24 months</td></tr>
<tr><td>3. Develop and implement analytical algorithm for identification of disease outbreaks.
Collaborate with IT partners to develop web-based links that will show the data for public viewing in graphical manner as well as in tables. analytical algorithms, using machine learning where necessary, will be used to integrate to automate the data visualization.</td><td>24 months (in parallel with objective 2)</td></tr>
<tr><td>Partners (not exhaustive)</td><td colspan="2">Government Departments
• MH/PMC/PCMC Health Departments
• Vital registration
• Pune Smart City Initiative
National Health Programs
• IDSP
• NTEP
• NACO
• NVBDCP
• Other
IT Analytics
• TCS
Persistent System

Academic Partners
• IISER Pune
• IUCAA
• NCL
• JHU (India)
• AFMC
Hospitals/Research Institutes
• ICMR/NARI/NIV
• DBT NCCS
• IRSHA
• Naidu Hospital, Sassoon</td></tr>
</table>

1. Introduction:

Overview. The COVID-19 pandemic highlighted the importance of a robust public health surveillance system that can facilitate timely intervention and aid in designing public health policies.[1] Prior to the pandemic, systematic data collection of health-related data at the national level has been infrequent and was restricted to a few vertical programs.[2] For example, HIV and TB programs collect individual patient level data but these data are not publicly accessible. Similarly, infectious disease surveillance program (IDSP) compiles aggregate data on 33 reportable illnesses. But these datasets are not designed to identify new local or national level, small or large disease outbreaks that need immediate local management policies and resource allocation. In addition, these datasets are not curated to understand the epidemiology of the diseases such as risk factors, disease progression and geographic variation, comorbidities and so on. Therefore, there is an urgent need for establishing publicly accessible curated epidemiologic databases to rapidly identify new disease or syndrome outbreaks which forms the basis for timely action from public health authorities.[3] Further, these databases can also lend itself to understanding the epidemiology of relevant public health problems.

Pune Municipal Corporation and Pune Knowledge Cluster Collaboration. During the early stages of the pandemic, Indian Government and Indian Council of Medical Research constituted guidelines for centralized collection of demographic and clinical data such as total case count, deaths, hospital occupancy as well as guidelines for clinical management. During the beginning of the pandemic, Pune city, managed by Pune Municipal Corporation (PMC), successfully initiated, and maintained Government mandated data compilation from clinic and hospital-level data at the city level. Around the same time, the Pune Knowledge Cluster (hereafter referred as PKC) was set up and funded by the cluster initiative of the office of the Principal Scientific Advisor to the Government of India. The PKC brings together Pune-based academic institutes, R&D labs, industry and government organisations to identify gaps in health, environment and social structures and policies and to facilitate and develop the standard procedures to address these gaps. The PMC and PKC entered into a collaborative agreement in April 2020 to develop and implement local policies based on the analysis of the COVID-19 patient-level data. The major activities of this collaboration included curation of the data, analysis of data at sub-region level called Prabhags, modelling of the data to project the pandemic curve, constitution of a project to assess the prevalence of positive serology among the population and coordination for resource procurements and allocation (**Figure 1**).

COVID-19 database utilization. The COVID-19 patients level data compiled at the city level supported several analyses as the datapoints includes demographic variables such as age, sex, address, date of testing, positivity and outcomes (deaths). Using the data from early stages of the pandemic in Pune, the PKC academicians accurately forecasted the incidence of COVID-19 over the next several months. This enabled procurement of SARS-CoV-2 testing, ventilators and establishment of additional test facilities and flu clinics, and COVID-cate centers. Furthermore, COVID-19 hotspot pockets within the city with high population density were identified and those areas were contained to limit the transmission. Moreover, these data were curated to assess the epidemiology and impact of lockdowns in incidence of COVID-19 cases. Notably, a Pune COVID-19 web application was developed and implemented that used analytical methods to derive the incident COVID-19 cases over time, case doubling time, and trajectory at the city levels and sub-region levels called prabhags, which can be visualized publicly in real-time.[4] Although

these data were used to forecast the trajectory of the cases, the dataset shared by the PMC posed a few challenges. First, it included minimum variables, focused to disseminating summary reports to administration and press. Second, there were several missing data points such as date of testing, correct addresses and date of hospitalization and clinical outcomes which made it difficult to assess the burden of disease. Despite these limitations, several lessons were learned on how optimally collected data can help manage public health crisis brought on by the highly transmissible COVID-19.

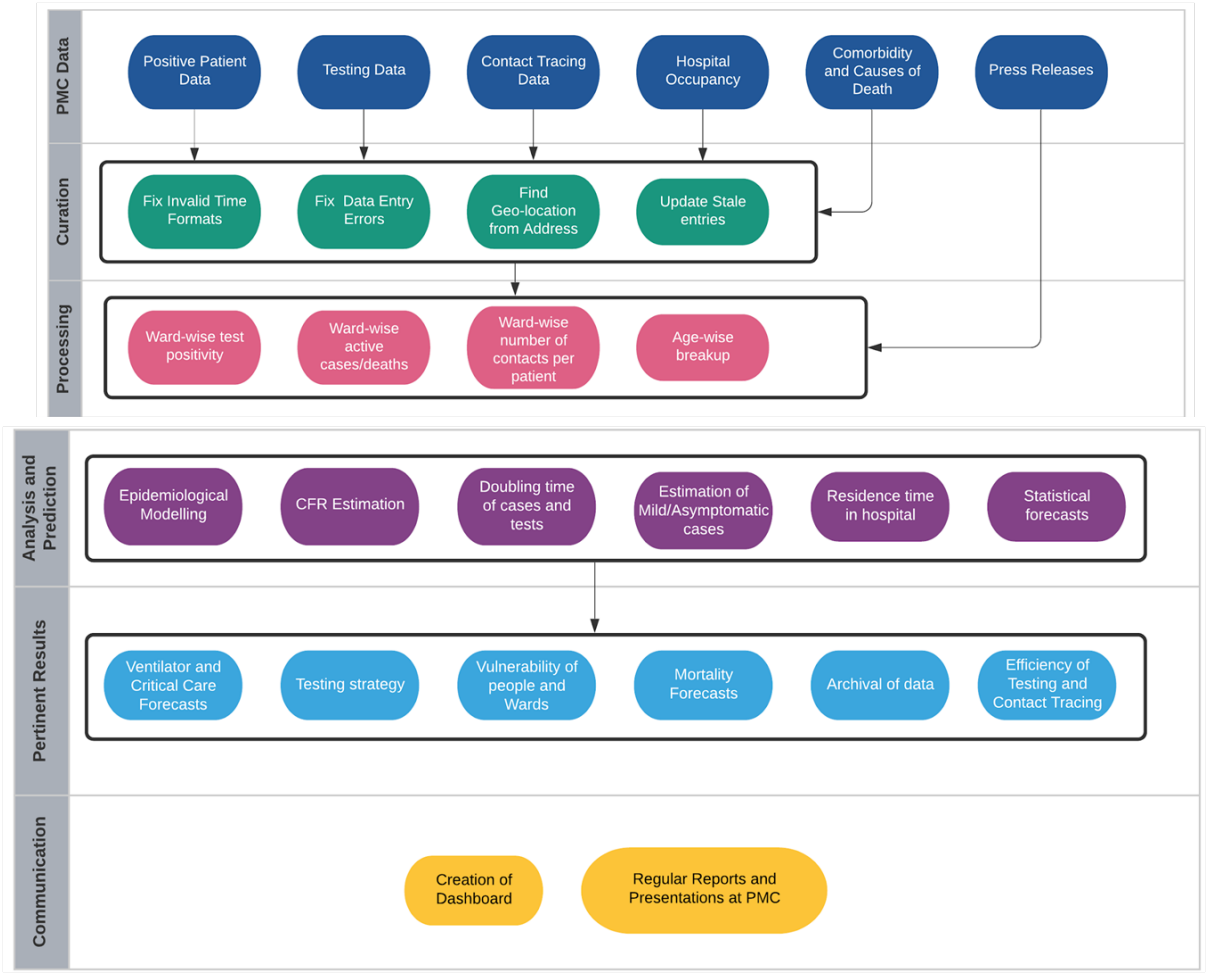

Figure 1: COVID-19 activities undertaken by PKC in collaboration with Pune Municipal Corporation.

Historically, infectious diseases epidemics occurred more frequently at regional levels than at international levels as was seen with COVID-19. Therefore, regional level epidemiological databases are a critical need. A few disease specific databases for cancers and diabetes exist in Pune city that have more granular data points and have been used to inform public health interventions. However, these databases are not designed to identify outbreaks. Our positive experience with PMC-PKC collaboration for COVID-19 pandemic has built a successful template for public-private collaboration, development and management of publicly accessible epidemiologic databases that yields advanced analytics to predict epidemic progression. Here, we propose establishment of a robust public health surveillance and epidemiological database entitled “Pune Epidemiological Database Project (PEP-DATA)” that is feasible, acceptable, scalable and sustainable across

different levels of health facilities (primary health centers (PHCs), taluka hospitals and tertiary centers) and that will establish a standardized format for data collection to facilitate recommendations for public health policies to local officials and help forecast future outbreaks.

2. Innovation.

The PKC envisages a centrally placed database that will contribute to the following major innovations with the establishment of the PEP-DATA project. As shown in **Figure 2**, the vision for the project is to strengthen health systems through responsive and evidence-based decisions supported by a scalable, secure, open source and interoperable database that is quality controlled and becomes amenable to advanced analytics. Initially, this project will focus on developing a module for acute febrile illnesses and when the template has been successfully developed and implemented, this module can be adopted for multiple diseases-communicable (diarrheal and respiratory illnesses) and non-communicable (diabetes, heart diseases).

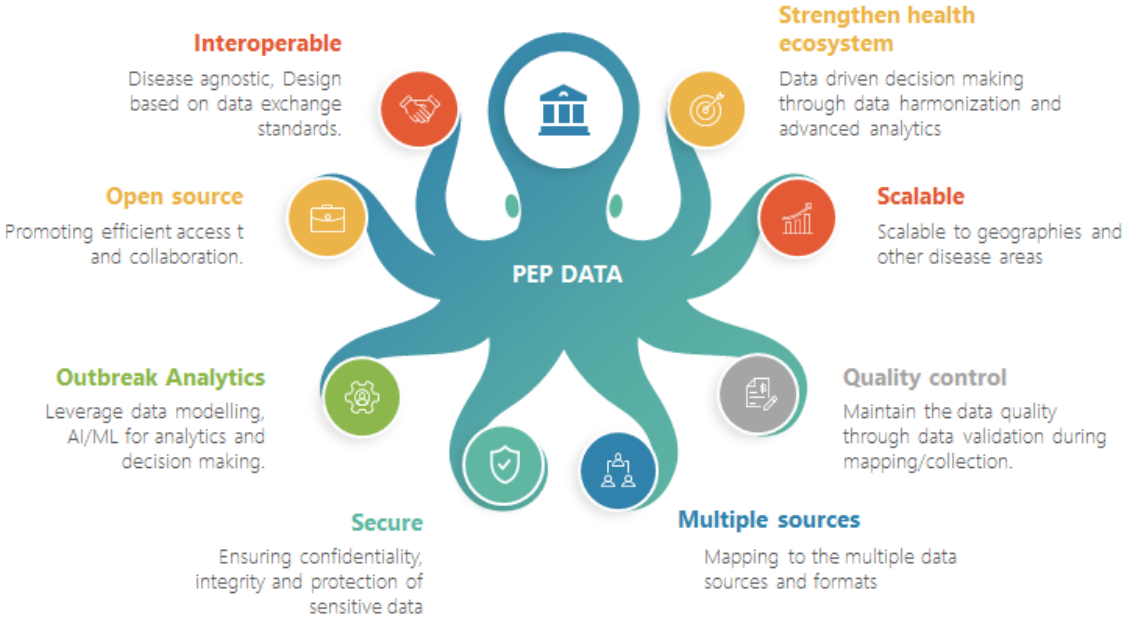

Figure 2. Vision for the PEP-DATA project

A.Digitization of data capture, processing, storage, analysis of the epidemiological data. As has been done with the COVID-19 database to some extent, we propose to capture multiple segments of data such as background, case data, surveillance data, medical data, laboratory data as shown in **Figure 3**. These data will be collected electronically, will be encrypted,and undergo routine data quality and validation checks to maintain high quality complete and granular data. The data collection will ensure maintenance of confidentiality and privacy of individual patients as per the Digital Health guidelines.

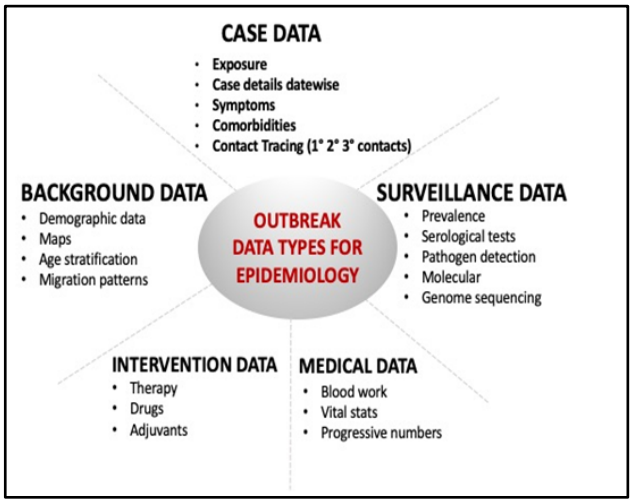

Figure 3. Data elements proposed to be captured in PEP-DATA.

B. Advanced integrated analytics for outbreaks. Outbreak analytics represent tools and methods used to collect, curate, visualize, analyze, model, interpret and report on outbreak data. These outputs are central to the surveillance pillar of any outbreak response and aid in forecasting and decision analysis. The major outputs and deliverables of the outbreak analytics and a COVID-19 illustration are shown in **Figure 4a and b.**

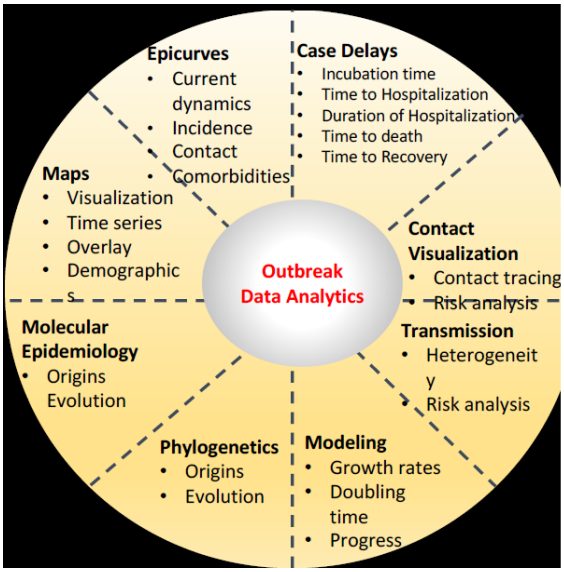

Figure 4a. A schema of potential outputs and deliverables from integrated analytics of outbreaks Figure 4b. An example of COVID-19 visualization of real-time data

C. Modelling, forecasting and decision analysis. Epidemiologic databases make it possible to develop mathematical, geospatial modelling that helps identify migration of population, and understand transmission dynamics using network approaches (**Figure 5**). Importantly, multi-model consensus of refined individual models can be derived to make projections of impacts and risks of each management strategy that can be used for deliberations. These modelling exercises will provide data to implement adaptive management strategies as the outbreak progresses.

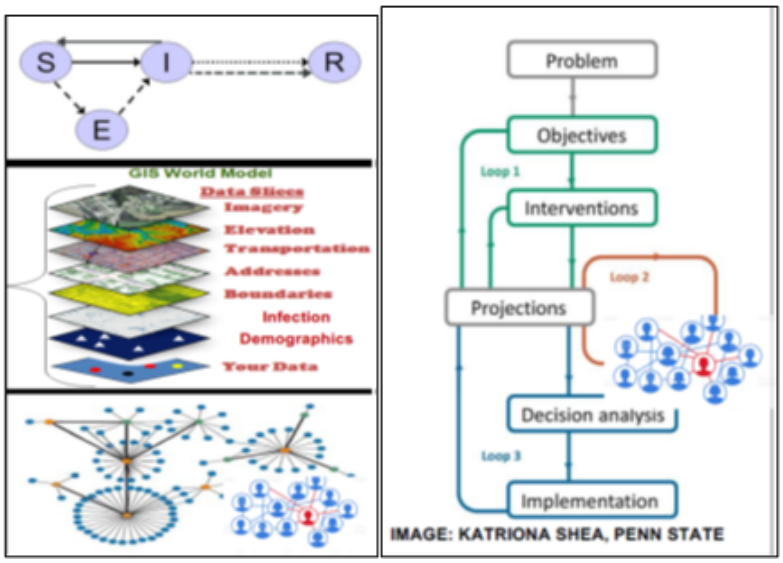

Figure 5. Illustration of mathematical, geospatial and decision models

D. Web-application based real-time data dissemination. For COVID-19 database, the PKC has created a web-based easy-to-use informative dashboard, as shown in **Figure 4b**. Using this template, a web application will be created that will have in-built machine learning algorithms to visualize the data for communicable and non-communicable illnesses.

3. Goals and Objectives:

The overarching goal of this proposal is to establish an integrated epidemiological database in Pune district that would facilitate analysis and prediction of outbreaks and is scalable across Maharashtra State and beyond. Data may be generated at different sources and in different formats but still could be used to create a database by defining a minimum data standards and elements to facilitate advanced analytics and modelling for the purposes mentioned above. To achieve this goal, we propose the following specific objectives.

Objective 1. Assess the landscape of data sources in public health systems using surveys and formative assessments.

Objective 2. Design and build epidemiologic database module that include minimum variables that can be uniformly collected across different health levels for febrile illnesses that can be adapted for various communicable and non-communicable diseases.

Objective 3. Develop and implement tools to create publicly accessible web-based, real-time visualization of data for identification of disease outbreaks.

4. Approach: To achieve the above goals and objectives, we have assembled a collaborative team of experts including but not limited to epidemiologists, statisticians, data scientists, modelers, and information technologists with their institutional support. The organizational chart of this project is shown in Figure 6. The PKC principal investigators, **Professor L. Shashidhara and Professor Ajit Kembavi and senior advisor Dr Anita Kane**, will facilitate the project and obtain necessary approval and funding required for this project. The PMC, Pune district authorities, **Drs. Vidya Mave, Nikhil Gupte** will be EPI-DATA project leads and will supervise development of the proposal, data abstraction forms, design of the database and implementation. **Dr Prasad Bogam** will provide technical

inputs for the data forms and data variables to be collected; **Dr Nishi Suryavanshi** will lead the implementation of the project; **Dr Mandar Paradkar** will assist in reviewing the data points for the pediatric population. **Dr. Rupa Mishra** will be the PKC liaison and will help supervise staff on the ground and will be involved in day-to-day management of the project personnel. The collaborating teams and institutions include several Pune-based academic institutions and industry partners as shown in **Figure 6.**

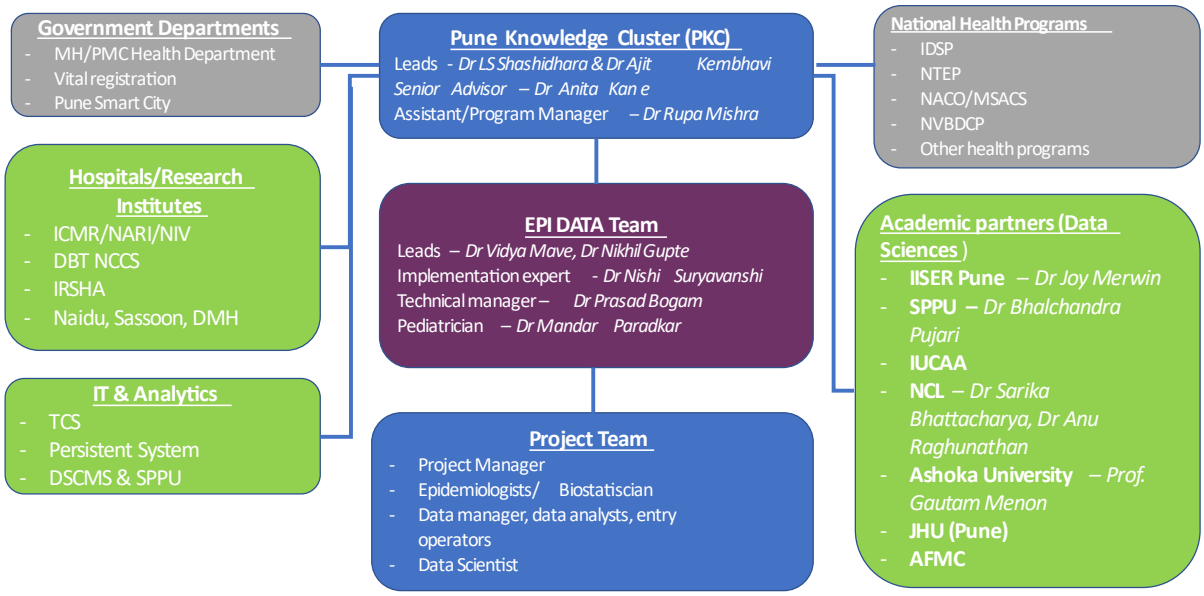

Figure 6. Organizational chart of PEP-DATA project

To accomplish each of three objectives, we propose three Phases.

Objective 1, Phase 1. Assessment of the landscape of data sources in public health systems.

Rationale.

In public and private sector health facilities, the data collection points (reception, doctors and exit) may significantly vary. Furthermore, the tools used to capture the health data may be paper based or on e-platforms. Moreover, data storage for programs like IDSP may be in different formats. Importantly, even if these data are collected routinely, data quality including missing data information may not have been systematically assessed. Therefore, formative assessment that describes the landscape of data sources, data quality and mode of data capture, are critical. To address this, we propose the following methodology.

Methodology (Figure 7).

Settings: To get a basic understanding of the current data sources and government mandated aggregate data capturing, we propose formative assessments of public sector health facilities. These include PHCs, CHC, taluka civil hospitals, corporation hospitals, specialized and tertiary care hospitals hospitals (such as Naidu Hospital Programs), and community-based workers (ANMs, ASHA, Anganwadi). Approximately 5% of 600 PHCs,

2-3 civil and corporate hospitals and all tertiary care facilities will be targeted for this exercise.

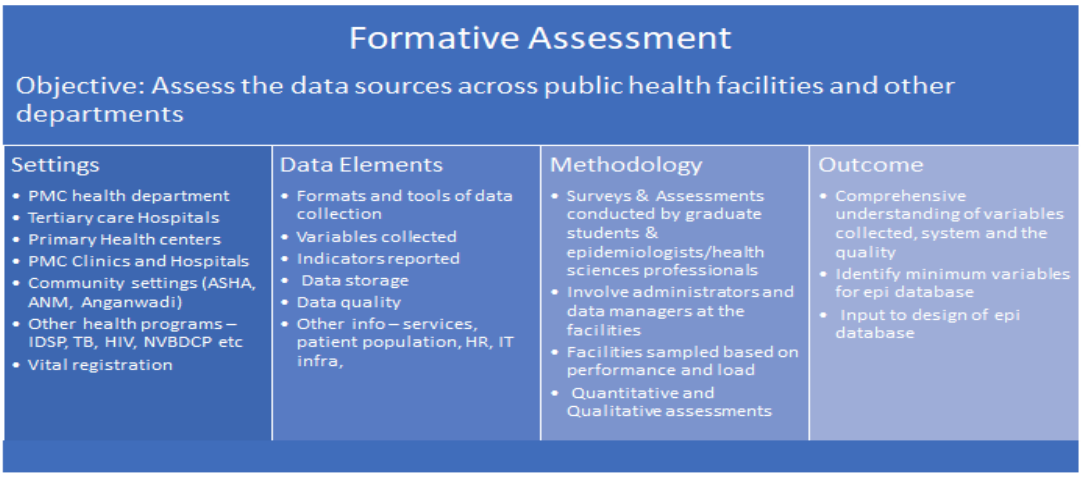

Figure 7. Phase 1: Formative assessment schema

Elements of formative assessments: Information on number of outpatients, inpatients admissions, patient flow, mode of the data collection, data collection templates and data archiving will be obtained. In addition, retrospective data for 3 months will be abstracted for all sources of data collected at each health facility. Furthermore, data flow and storage and reporting of laboratory and radiography data will be recorded. These data will be collected by data collectors who have a minimum of master's degree in health sciences and will be recruited by the PKC.

Assessment tools: We will be developing formats for a systematic data collection. The data collection formats of survey will be configured in a real time data capturing tool for easier and faster data collation and curation process.

Mapping of the data: We propose to analyze the collected data to make the following themes.

1. Data variables collection and patient flow similarities and variations at different health levels by the facility (eg. PHC, hospitals) for routine care
2. Mode of data collection and storage
3. Templates used for various national level and local level reporting to authorities.
4. Quality of data (e.g., proportion of missing data) based on the retrospective data collected at different health facilities.

Data mapping exercise will then identify the common data points that are collected across all facilities for reporting and routine management purposes. This exercise will form the basis for planning of the next objectives.

Timeline. We propose 12- months for this objective. The first 4-6 months will be used for the data collection of this objective and the next 6 months will be used to analyze the data for mapping. During pilot, a simple data collection tool will be deployed that captures the mode of data collection and the health facility contacts using the existing PKC staff.

Project team for this objective. Under the supervision of project leads, **Dr. Rupa Mishra** will assist as the PKC liaison and will help supervise staff on the ground. **TBA program manager** will be responsible for day-to-day supervision of staff and responsible for quality assessment of all personnel. **TBA data collectors** (X) will be responsible for collecting all requisite data at health facilities. **TBA data scientists** along with project leads will compile and analyze the data supervised by project leads.

Expected outcomes. At the end of this phase, we hope to have a basic understanding of the status of the reportable diseases, data capture modes, data storage and common variables recorded across different public sector health facilities. Furthermore, we will identify the scope of the PEP-DATA project by the end of this phase.

Objective 2, Phase 2: Design and build epidemiologic database module for acute febrile illnesses.

Rationale:

Acute febrile illness syndrome generally signals the development of outbreaks when unique patterns are identified. Therefore, developing the database module with this syndrome would be ideal and may form the template for other significant public health relevant diseases. In addition, several infectious diseases already are reportable and they need to be captured in certain universal templates. Based on Objective 1, we propose to identify minimal variables that meet the global data collection standards based on which EPI-data database will be designed that can be uniformly implemented for different health facility levels.

Methodology

Development of a list of variables based on global standards. Global and National data standards for health/clinical records and interchange such as EHR standards, HL7 or CDISC etc exist across healthcare or clinical research domain.[5-8] We will adapt these standards and identify the list of variables that need to be captured and/or mapped across different health systems based on the reportable diseases. Furthermore, the data capture must include a minimum set of demographic variables such as age, sex, and address to help identify the geographic location and vulnerable population for any future outbreak assessments.

Stakeholder meetings. A stakeholder meeting will be held after the development of the list of variables. Discussion points may include mode of feasible data capture, feasibility of capturing the list, training needed for uniform data capture based on the global data capture standards, data migration to a central repository.

Data capture and integration. Based on the above information collected, the data capture may still continue with the existing systems or a new data capture mode using an app may be developed. Different health systems will be given these two options. If data capture continues with the existing system, data abstraction for the list of variables mentioned above will be done using a digital template. To ensure sustainability, assessment of human resources available for data capture for the basic variables for acute febrile illnesses and reportable diseases will be performed. These data from different sources will then be merged using a digital platform to develop a central repository of the database.

Training for uniform data capture. Regardless of data capture modes, the training of staff at different health levels would be needed. As the scope of this project is very large even for

acute febrile illnesses, only one PHC, and one taluka or corporation hospital will be targeted for training and implementation of the data capture method/s.

Design of the database. The IT sector partners of the PKC will be engaged for this activity. These partners have generously agreed to provide IT support for data collection capture tools as well as designing the database. The basic rules of the database will be followed during the design of the database that include, but will not be limited to, routine assessment of quality, ability to query the data, and logical checks.

Timeline: This Phase may take 1-2 years.

Project team for Phase 2. Depending on the scope of the project which will be determined at the end of Phase 1, the project team personnel may vary. The project leads will closely work with PKC PIs to determine the personnel needed for this Phase.

Expected outcome. At the end of this Phase, we expect to have the database implemented in a few health centers in Pune district. Further, we will have developed a febrile illness database that can be analyzed for the burden of acute febrile illnesses and its outcomes. Based on the experience of implementation and ease of data integration and analysis, the expansion of the EPI-DATA across all facilities will be determined.

Objective 3, Phase 3: Develop and implement analytical tools, including machine learning, for identification of disease outbreaks.

Rationale.

Large data from the EPI-DATA project will require sophisticated technologies to understand epidemiology and track disease burden to understand, estimate, and predict disease burdens and their relationships to risk factors. In this project, we will collaborate with data scientists from our IT partners to develop such algorithms to create publicly accessible web-based, real-time visualization of data for identification of disease outbreaks.

Methodology.

Application of machine learning algorithms for data analysis and visualization. Based on the data scientist's recommendation, various machine learning techniques- supervised, unsupervised and reinforcement learning, will be applied to understand the febrile illness data.

Development of web-based real-time data visualization. Similar to COVID-19 data visualization shown in **Figure 3b**, we will collaborate with IT partners to develop web-based tools that will show the data for public viewing in graphical manner as well as in tables.

Timeline: This Phase will occur in parallel to Phase 2.

Project team. The project leads and program managers will work with data scientists and IT partners for this Phase.

Expected outcomes. At the end of this Phase, the EPI-DATA project would have completed the entire database development cycle for febrile illnesses. At this time, the database module is ready to be adapted to other diseases.

References.

1. Sundararaman T. Health systems preparedness for COVID-19 pandemic. Indian J Public Health. 2020 Jun;64(Supplement):S91-S93. doi: 10.4103/ijph.IJPH_507_20. PMID: 32496232.
2. NIti Ayog (2020). VISION 2035 PUBLIC HEALTH SURVEILLANCE IN INDIA. https://niti.gov.in/sites/default/files/2020-12/PHS_13_dec_web.pdf
3. Kraemer, M. U. G., Scarpino, S. V., Marivate, V., Gutierrez, B., Xu, B., Lee, G., Hawkins, J. B., Rivers, C., Pigott, D. M., Katz, R., & Brownstein, J. S. Data curation during a pandemic and lessons learned from COVID-19.
4. COVID-19 in Pune | DSCMS, PKC. Data Analysis and Forecasts of Covid-19 in Pune. Retrieved from http://cms.unipune.ac.in/%7Ebspujari/Covid19/Pune2/
5. Meta Data and Data Standards for Health Domain. https://main.mohfw.gov.in/sites/default/files/Part-I%20Overview%20Report%20Health%20MDDS.pdf
6. Electronic Health Record (EHR) Standards for India – 2016. https://main.mohfw.gov.in/sites/default/files/17739294021483341357.pdf
7. Health Level 7 (HL7). https://www.hl7.org/
8. Clinical Data Interchange Standards Consortium (CDISC). https://www.cdisc.org/standards

Acronyms

- IDSP – Integrated Disease Surveillance Program
- NTEP – National Tuberculosis Elimination Program
- NACO – National AIDS Control Organization
- NVBDCP – National Vector Borne Disease Control Program
- IISER Pune – Indian Institute of Science Education and Research
- IUCAA – The Inter-University Center for Astronomy and Astrophysics
- NCL- National Chemical Laboratory
- JHU (India) – Johns Hopkins University India
- AFMC – Armed Forces Medical Colleges
- ICMR/NARI/NIV – Indian Council for Medical Research / National AIDS Research Institute / National Institute of Virology
- DBT NCCS – Department of Biotechnology National Center for Cell Sciences
- IRSHA – Interactive Research School for Health Affairs
- PHC – Primary Health Center
- CHC – Community Health Center
- ANM – Auxiliary Nurse Midwife
- ASHA – Accredited Social Heath Activist
- EHR – Electronic Health Records
- HL7 – Health Level 7 standards
- CDISC – Clinical Data Interchange Standards Consortium

Proposed budget for Phase 1: Assessment of landscape of data sources in public health systems.

The total budget for first phase is Rs. 21,480,000. It includes personnel, travel, computers, communication, trainings, consultancy and miscellaneous.

Personnel/Travel			
Roles	**No.s**	**Duration (in months)**	**Total expense/Year (in Rs.)**
Program Manager	1	12	₹1,800,000
Program Associates	10	12	₹8,400,000
Data Analyst	2	12	₹1,200,000
Data scientist	1	12	₹3,000,000
Travel allowance (4000 per month)	10	4	₹160,000

Computers/Communications/Apps			
Items	No.s	Rate	Total (in Rs)
Tablet	10	35000	₹350,000
Laptops	4	90000	₹360,000
Desktop	1	100000	₹100,000
Printer	1	50000	₹50,000
Monthly communication cost for 14 personnel (1000/person)	12	14000	₹168,000
App Development for data collection			₹1,000,000

Trainings/Workshops			
Activities	No.s	Rate	Total expense
Consultation workshops	2	150000	₹300,000
Training	3	80000	₹240,000

JHU India group consultancy – ₹ 2,640,000
Miscellaneous cost (10%) – ₹1,712,800