DEPARTMENT OF GLOBAL HEALTH
UNIVERSITY *of* WASHINGTON

PhD in Global Health

Metrics Track

Methods and Issues for
the valorization of HMIS Data

Grégoire Lurton

IHME — Institute for Health Metrics and Evaluation

I-TECH — International Training & Education Center for Health

BLUE SQUARE .ORG

Abstract   Here the abstract will come

DRAFT

Contents

List of Figures

List of Tables

1 Introduction

If a literary form had to be chosen to write or talk about Health Management Information Systems (HMIS), the complaint would probably be everyone's favorite pick. Be it complaints on the burden of work involved in collecting, managing and analyzing data in health systems, or laments on the inexistence of good quality data in most developing countries health systems, HIS are usually described as a non performing burdens of health systems, that can only be improved [HMN citation]. This frustration has multiple causes, and is only matched by the expectations placed in HIS and their widely recognized importance, some authors calling HIS "the foundation of public health"[1]. Collecting and analyzing information on activities and results of health systems and on the populations served is indeed essential to guide strategic decision making and to inform health policies.

The complexity of designing and operating well performing national HMIS comes from the fact that HMIS have to handle a high diversity of data and information. Meanwhile, producing this information requires the contribution of a multitude of actors, and is a huge organizational and methodological challenge. Data used to produce relevant health information may come from administrative records, organizational documents or population surveys, and are produced by a variety of actors and organizations, with differing cultures and approaches.

Finally, HIS have to be able to adapt rapidly to changing epidemiological, organizational or political situations. They have to be able to produce relevant information on emerging health issues, and to adapt to the entry of new actors or to a modification in the mode of management of health systems.

In richer countries, the issue surrounding health information is often one of regulation and standardization. The existence of well performing and well established data sources on populations, and the relative ease of collecting massive amounts of data on individuals pose questions that are mainly related to the protection of privacy, and to the definition of standards for interoperability. The definition of what information should or can be produced is usually a legislative and matter, handled by dedicated entities.

In developing countries, where population data is scarce and data collection can be much more of a challenge, the issues are much different. Health policies depend heavily on the financial, political and technical support of international actors, with differing orientations and priorities. As a consequence, developing countries HIS strengthening programs can be classified into two great families : technology based solutions and institutional reforms. The first family is mainly targeted towards improvement of data collection and relies on solutions relying on the revolution in data collection capabilities. The second family relies on the adequation of health information produces to policy makers needs, and usually promotes a top-down and normative approach to Information Systems.

If each of these approaches has its benefits and some success stories, they appear to miss an important part of what makes information systems work. They put their emphasis to the extremes of the information production chains (cf. Figure 2) and undervalue the middle tiers of these systems. This proposal designs research program oriented towards exploring ways in which this middle tier can be mobilized to improve the value of information produced by HMIS.

This document will start by an introduction of the conceptual framework surrounding the proposed research. We will define Health Management Information Systems, using two classic schematic approaches. We will then define the objectives we will pursue in our research, and will finally describe our different aims and the methods we want to use to complete them.

2 Conceptual framework

We will first look at a representation of HMIS through the goals they are meant to fulfill. Once we will have understood why HMIS are used, we will look at a representation of how they are organized.

2.1 HMIS definitions

2.1.1 Goal approach

A first approach to HMIS is a consideration of the stated goals of these information systems. Figure 1 shows what these main goals are.
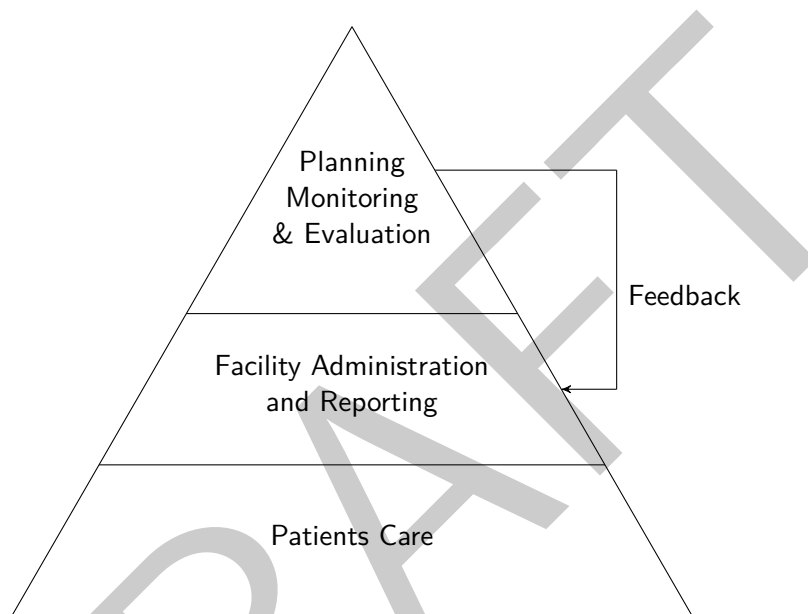

Figure 1: Information needs for HMIS

Patients Care Taking care of patients is the primary goal of a health facility. To do so, it is necessary to collect data on these patient, data that will be transmitted (to other services), stored and reused during further follow-ups.

Facility Administration and Reporting At facility level, HMIS data is used in daily activities to quantify and forecast needs in health inputs, and to create reports for higher levels of the health system.

Planning, Monitoring & Evaluation People in charge of the administration of health systems at local or national also need data to monitor activities in the health system, to evaluate the results of interventions, to report to funders or to plan later interventions.

The pyramidal representation of these needs is used to show that these goals fill data needs at different levels of health systems. The different needs for information can roughly thought as being the needs of different type of actors of the health system. Meanwhile, this understanding is not fully true, as at local level, actors will often hold multiple roles and will thus have to use information in different situations. For example, a physician may also be in charge of managing his health facility, and will thus need to plan activities and report on them.

2.1.2   Functional approach

A first way to approach HMIS is to describe the four principal functions that are necessary to have a HMIS to run. Figure 2 presents a simplified sketch of the principal functions that are to be filled in order for HMIS to produce useful information.
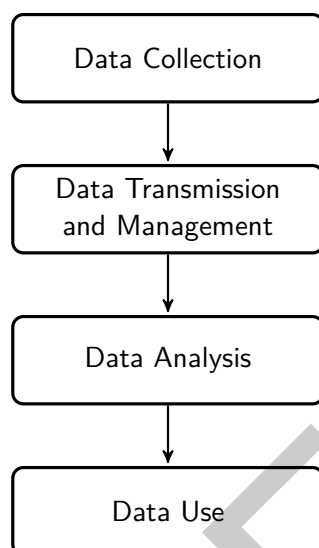

Figure 2: Different functions inside the Health Information Systems

Data Collection Primary data collection is essential to the production of any information system. In the case of HMIS, data collection happens in health facilities, and is made by health professionals. Data collected in health facilities can be individual patient data collected in patients files or cards. It can also be a first level of aggregation of this data, as for indicators that are reported on a regular basis by facilities to higher levels of the health system. This reporting usually happens through standardized reports, that are then transmitted by successive aggregation to the top of the health pyramid.

Data Management Data collected in health facilities has to be stored and archived, to be later accessed and reused. Data management work can encompass managing paper data, or managing computerized data. Individual patient data will be computerized in Electronic Medical Records (EMR) whereas aggregated indicators are stored in data-warehouses, like the DHIS2 software.

Data Analysis Data that is collected and stored in HMIS can then be analyzed. Analysis can be defined as the transformation of data into information. The results of data analysis can be varied, from collection of graphs and maps that can constitute dashboards, to the results of complex models that provide evidences of causality.

Data Usage Information is the end product of the HMIS, and is used by decision makers or health workers to achieve their tasks. For example, a nurse in a health post may need the monthly consumption of a health product to place an order. A District Health Officer may consider the evolution of monthly number of cases of malaria in his district to plan malaria prevention activities. At national level, a worker at the Ministry of Health may use the number of patients tested positive for HIV to design grant applications for the Global Fund.

Even though the schematic representation of this functions is linear, it should be noted that this linearity is not true in practice. Even once data has been analyzed the results of this analysis

has to be archived, and transmitted to the information end users. In some situations, data can be used in its raw form. For example, a physician may use a biological result that he has received in a raw form. Finally, for some data, data collection won't happen inside the health facility. For example, population survey results may be used to plan and target some interventions, but primary data collection will have happened outside of the health system, and the first function to be used in the HMIS will be the data management function.

The way a program considers and plans each of these functions will define how a specific HMIS will work. We will now present three common approaches to HMIS.

2.1.3   Three HMIS archetypes

Functions of HMIS (cf. section 2.1.2) are not independent of each other. Defining the relative importance of different functions of HMIS in the overall systems can change greatly the way a HMIS functions, and the output it produces. We differentiate three paradigmatic types of HMIS, varying on the respective influence of different functions. Building on the idea that a HMIS is used to provide an image of the activities and performances of a health system, we describe each function as a different way of making an image.

Jigsaw Puzzle HMIS   - A common way to design HMIS in developing countries can be considered as a Jigsaw Puzzle approach. A series of indicators are designed by program managers. These indicators are deemed to be *sensitive* and *specific*, and are supposed to allow managers to track and identify precisely the performance of health systems, and to provide important information on health system's results. The HMIS will then be organized to produce carefully designed indicators at facility level, and to transmit these indicators to higher levels for aggregation.

In these types of system, a lot of importance given to data collection functions, as the quality of this primary data collection is key to the rest of the work in the system. Data management in these system is often limited to aggregating some data and transmitting it to different actors in the health systems. Data analysis is usually mainly descriptive and is limited to presentation of time series values or mapping of indicators along administrative boundaries.

These systems are similar to jigsaw puzzles, made of very specific pieces, to compose a predetermined picture. When they are well designed, these systems can provide very useful information on health systems. Meanwhile, they are very vulnerable to any variation in primary data collection. As for jigsaw puzzles having a piece missing will jeopardize the possibility to get the whole picture right.

Pixel HMIS   - Another way to conceive HMIS is built on the collection and use of a multitude of individual data collected through Electronic Medical Records (EMR). Once the data is collected, program managers can query different indicators on different levels of aggregation, that can be extracted from different EMRs. In the best situations, interoperability of multiple EMRs present in a country allow for a central analysis of the data [10].

These systems allow a great variety of analysis, with a great variety of approaches. Analysis can be led varying geographic and time focus, or changing definitions of computed quantities. It also allows longitudinal analysis that are more difficult to perform with other approaches.

This approach thus involves a great investment in primary data collection and management, and allows elaborate data analysis. Meanwhile, it requires a technological investment and maturity that is seldom achieved in rich countries, and thus is very rare in developing countries.

Tangram HMIS   - Between the two extremes that are puzzle and pixel HMIS is a third, less prevalent approach to HMIS. This approach will be compared to the tangram game, in which simple forms are used and reused to draw different pictures. In this approach, the emphasis is put on the management functions of HMIS. Simple data elements are collected and stored, and are used and combined in different ways depending on the analysis that is done.

A key component of this model is thus the ability to store and reuse data, thus putting an emphasis on the middle tiers of data management. The use of data warehouses for computerized data is thus a characteristic of this approach. Meanwhile, it also requires an emphasis on data analysis in order to provide relevant information to end users.

2.2   HMIS strengthening strategies

Depending on the HMIS model that is used, programs will implement different type of HMIS strengthening approaches. Programs who privilege a jigsaw puzzle approach to HMIS will tend to focus on standardizing procedures and methods for data collection and data analysis. Meanwhile, programs who privilege a pixel approach to HMIS will tend to favour solutions geared towards the implementation of new and performant data collection tools.

The institutional approach operates under the assumption that all functions of information systems should be geared towards and submitted to the end information users. This approach tends to be extremely normative as any activity in the information system has to be oriented towards one main predefined goal. In doing so, this approach undervalues the benefits of both the integration of external data, and the positive externalities data collection and analysis may have on multiple users.

The technological approach relies on the assumption that collecting data and making it available is a sufficient enabler for all other functions of information systems to operate. In this sense, a direct link is made between an information need and a data gap. This approach comes at a cost, and provides only limited benefits if it is not supported by improved data analysis. These solutions tend to provide highly specialized and siloed data collection systems.

We argue these two approaches focus on the most expensive ways to strengthening HMIS (data collection and systemic reforms), and are emphasizing the design of systems and tools that are specific to precisely defined data needs, thus limiting the possibility to implement secondary data usage and the positive externalities of their interventions. The archetype of these pitfalls are the well known parallel and siloed data systems present in many developing countries health systems.

Meanwhile, some of the most significant successes in the strengthening of health information systems in developing countries have been reached precisely through the strengthening of their middle tier. The District Health Information System (DHIS2) project has become a pervasive system to store and organize data collected in developing countries health systems. The DHIS2 approach to health information is based on the organization and storage of multiple data types and sources in a generic data warehousing approach. Its versatility and its ability to adapt to different contexts and data has made it increasingly used in multiple context, thus arguably improving the storage and the availability of health data in low resources countries. Other approaches geared towards the promotion of interoperability of different dimensions of data systems, such as the Open Health Information Exchange framework are also gaining traction.

If these approach have provided efficient solutions to organize and access HMIS data, there is still a lack of solutions to analyze and use HMIS data. Indeed, the high dimensionality of HMIS data and its average low quality make it essentially hard to analyze using standard methods available in developing countries health systems. We will now describe the research project developed to explore ways to analyze this data.

2.3   Approach and research questions

This project aims at exploring methods to improve the analysis and use of HMIS data by providinwg innovative approaches to this data. To do this, we will use both a technical approach using innovative methods from the data science field, and a critical approach of health information systems as social objects.

The generic question we ask is : How can data currently routinely used data sources be used to provide actionable information for decision makers ?

More precisely, we will ask three main data analysis questions :

- How can metadata collected in an EMR be used in EMR data analysis ?

- How can different non standardized HMIS sources be mapped and jointly analyzed ?

- How can multiple data source be integrated to HMIS and analyzed to provide information at local level ?

- How do decision makers in health systems consider HMIS generated data, and how does it influence the way HMIS are engineered ?

Our method is taken in a decision theoretic framework. We do not aim at providing substantial knowledge with the results of our analysis, but rather we aim at providing information that will emopower decision makers to take decisions.

Cadre bayésien

We operate in decision theoretic framework, our end objective being to inform decision making for the

Approche Probabilistique des données

Control Charts

Each of these questions explores a different aspect of how standard HMIS data analysis can be expanded to produce useful information with HMIS data. Metadata like the times of creation and savings of EMR forms are indeed seldom used. Meanwhile they provide useful information on how data is collected, and on the working patterns inside health facilities. Our second question explore a different problem, which is often taken as a question of interoperability. Indeed the multiplicity of programs and actors working in many health systems generates a multiplicity of indicators used, that are often related but not identical. There is a need for simple and effective methods to map and conjointly analyze data from different HMIS systems. Finally, HMIS should be useful for local level analysis and decision making. Meanwhile, HMIS data is seldom useful on its own and has to be integrated in larger analysis frameworks to produce interesting information. This integrating can be difficult at national level, but it is even more complicated at local levels, has mapping precisely different data becomes more and more of a challenge at small scale.

We also aim at providing a critical evaluation of the way HMIS are thought of in developing countries societies. Authors like Alain Desrosières and Ted Porter have shown how statistics and computation have come from and generated different cultures of public action in modern societies. The ambivalence of numbers as descriptors or norms has an influence on how information systems are thought of as top-down normative systems instead of knowledge systems. We aim at interrogating how this perception has its roots in long term local historical trends as well as in the tradition and methods of quantitative public health.

To answer these question, we will conduct four distinct research projects.

Aim 1 Evaluation of the benefits of improved data collection for HIV patient care in Kenya.

Aim 2 Test of multiple semantic approaches to interoperability in Bénin.

Aim 3 Definition and test of a local malaria elimination metric in Namibia.

Aim 4 Analyze of the theory and practices of Health Information Systems for national decision makers in Mali.

These aims have been designed to provide insights on the problematic posed for each HMIS goal and function. We will now describe each of these aims in more details.

3 EMR and individual health[1]

A first aim will be to understand how data collection itself impacts quality of care. As we postulate that data collection is not a neutral activity, we want to look into how primary data collected in HIV care setting can impact the outcome of care and organizational capabilities of HIV services. The case we will explore for this project is provided through a project implemented by ITECH in Kenya.
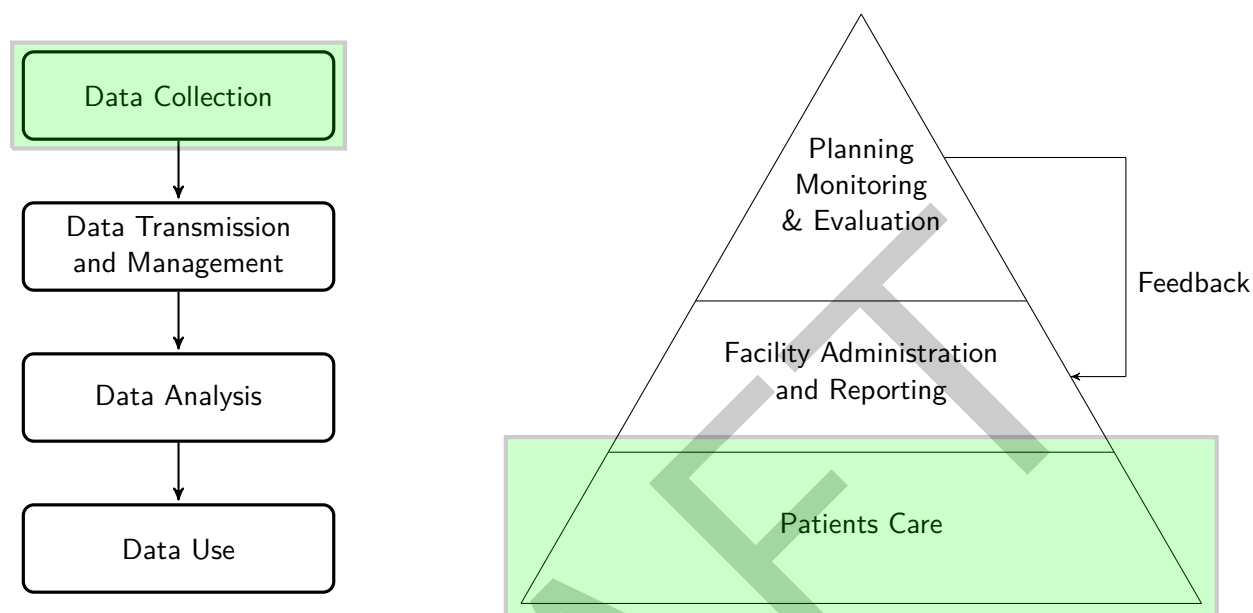

Figure 3: Objective one definition

3.1 Setting

In Kenya, I-TECH has implemented an EMR for HIV care, called KenyaEMR, in 341 facilities. The evaluation of this program is currently being carried out. One objective of this evaluation is to assess the effectiveness of KenyaEMR implementation. This effectiveness will be evaluated on two dimensions:

1. Improvement of reporting quality in facilities after KenyaEMR implementation

2. Improvement of quality of care metrics after KenyaEMR implementation

3.2 Data

Kenya's legal framework for protection of confidentiality of personal health information prohibit transfer of individual patient-level data from any health care facility, even if the data is de-identified. For this reason, the data we will use for this evaluation will be indicators of quality of care, aggregated monthly at facility level, and used for Continuous Quality Improvement (CQI) (see section 3.5). These indicators will be aggregated on site in Kenya and transmitted for data analysis.

To monitor the maturity of implementation of KenyaEMR (see section 3.3), we will measure the delay in data entry using metadata stored with KenyaEMR forms, with time stamps for form creation. We will also trace utilization of reporting features of KenyaEMR by using time stamps linked to the use of reports generation. All this data will be extracted and transmitted in raw form for analysis.

To measure the quality of the reports produced for different periods (see section 3.4), we will consider counts of number of forms entered for a given period, and mean completeness of entered

---

[1] This section is heavily based on KenyaEMR evaluation protocol.

forms. These will be aggregated on site and transmitted for data analysis. We will also use results from Routine Data Quality Assessments (RDQA) that have been conducted in different sites with KenyaEMR implementation. Data for these RDQA are collected in Excel format, and will be used as an external measure of the quality of data entered in KenyaEMR. In the remaining of this document, we will thus use the following terms:

- Patient data refers to the data collected by health workers during patients' visits. They are stored in paper patients' files, or entered in KenyaEMR forms. We will thus refer to paper patient data or to electronic patient data. This data will not be directly used for analysis in this project.

- CQI indicators refers to aggregated indicators used to measure quality of care.

- CQI Report refers to a set of CQI indicators computed for a specific month for a specific facility.

- DHIS Report refers to the MOH 731 and MOH 711 reports. We will differentiate between paper reports for which the data and the computation of indicators have been made without any digitalization of patients' data, and electronic reports for which patients' data has been digitalized. We will be able to use the paper reports as they have been entered in DHIS2 or other data collected by health districts administrations.

- Patient Forms Metadata refers to the metadata generated by KenyaEMR when patient forms are entered. The metadata used should mainly be timestamps related to time of data entry.

- Reporting Metadata refers to timestamps generated by KenyaEMR when different types of report are generated.

- RDQA Data refers to raw data collected during RDQA exercises.

3.3 Implementation maturity

We distinguish three different periods in the implementation of the EMR. Each of these periods is characterized by different ways the data is collected, entered, analyzed or used. For each of these periods, we will also have access to different types of data. We will describe the characteristics of each of these periods, and present a strategy to categorize the available data in each of these periods, using DHIS and CQI reports and metadata.

3.3.1 Paper Based

In the first period, no patient data entry is made in the facility. Patient data is collected in paper files, and reports are computed manually using these files. In the meantime, health workers can only use paper data to follow their patients.

The data we will be able to access from this period is:

- The patient data that will have been retrospectively entered in KenyaEMR

- The paper reports that will have been entered in DHIS2 or other reports available from health districts administrations

3.3.2 Retrospective Data Entry

In a second period, data entry has been implemented in the facility. The backlog of paper patient data has to be retrospectively entered, and current patient data is also entered in KenyaEMR after a delay. During this period, health workers will still refer to paper data to follow their patients. This is the Routine Data Entry phase (RDE).

The data we will be able to access from this period is:

- The patient data entered in KenyaEMR

- The metadata for patient data entered in KenyaEMR

- The CQI and DHIS reports computed from this data

- The reporting data metadata

- Evaluations of data quality from RDQA

3.3.3   Point of Care

In a third period, the patient data is entered either by the health worker or by a specialized data clerk based on patient data collected on paper by the health worker, in quasi real time with the medical consultation. We call this phase the Point of Care (POC) phase.
    The data we will be able to access from this period is:

- The patient data retrospectively entered in KenyaEMR

- The metadata for patient data entered in KenyaEMR

- The CQI and DHIS reports computed from this data

- The reporting data metadata

- Evaluations of data quality from RDQA

3.3.4   Transition periods

There may be some overlaps between different periods. For example, the limit between the paper collection and routine data entry may not be clear cut, as some facilities may have tried to start entering more recent patient forms during the RDE period, to be on top of the work quickly. Similarly, some facilities may have been at the same time doing retrospective data entry for some forms, and POC data entry for some others, depending on the organization of care.
    To take this into account, we will need to consider overlapping periods for different aspects of the data.

- Data quality: the process to collect and enumerate patient data is identical in paper based period and RDE period. Meanwhile, in the POC period, data is possibly directly entered in KenyaEMR, without using a paper form. Also, rapid data entry may allow to go back to the HW to complete missing data, or to correct unclear information.

- Report computation quality: Once the data entered in KenyaEMR, the reports can be computed automatically. Thus, the quality of computation of reports will be identical in POC and RDE, but will likely differ from the Paper Based period (see Section 3.4 for more details on Reporting Quality).

- Quality of care: in the paper based period as in the RDE period, HW can only access patient data through paper files. They thus can't use automated reminders, or summary information offered by KenyaEMR. Meanwhile, starting in the RDE period, some reports can be edited through Kenya EMR that would allow health worker to better track late and defaulting patients, and thus would allow them to pass reminders calls, or plan lab tests. We would thus anticipate to see a slightly improved quality of care for RDE period compared to Paper Based period, and to see an additional improvement for POC period compared to RDE period.

Based on this periodization, we will want to test three main hypothesis:

1. Observed data quality is similar in paper and RDE period, and better in POC period.

2. Computation quality is bad in paper-based period but then improved in RDE and POC periods.

3. Quality of care is worst in paper-based period, improves in RDE and is best in POC period.

3.3.5   Methods for periodization

For each facility included in this analysis, we will have to define when they enter or exit each of these periods. To do this, we will use programmatic data collected by I-TECH staff to monitor the implementation of KenyaEMR, and time stamps associated with forms entered in KenyaEMR, and Building on the characteristics of the different periods, we will categorize the different dimensions of the data collection and use separately:

1. Data Quality: The passage between stage 1 and stage 2 of data collection will be tracked looking at the delay of data entry of forms. Looking at the distribution of this delay, and using I-TECH monitoring data for confirmation, we will define a threshold to define stage 2 data entry. We will also use comparison of data completeness between different periods.

2. Report Computation: The passage between stage 1 and stage 2 of report computation will be tracked looking at the source of the reports available for the facility. Existence of reports from DHIS2 or similar source that were not produced using KenyaEMR computation will lead to the categorization of the stage of report computation as stage 1. Reports computed with KenyaEMR will lead to a categorization of the period as a stage 2 for report computation. The categorization will be validated with data from I-TECH monitoring, and by a comparison of results reported in DHIS2 and results computed for the same month from KenyaEMR.

3. Data usage: The passage between stage 1 and stage 2 for data usage will be used considering metadata from different reports, and delay of data entry. A different threshold as the one used for data quality will be used to categorize a facility as stage 1 or 2 for quality of care.

Using available data to explore this different dimensions, we will be able to categorize each facilities' reports into its corresponding period. As we anticipate some exceptions due to unclear transition periods, we will design a continuous index of maturity of implementation of KenyaEMR, to be included in latter stages of the analysis. Depending on the results of the exploratory work, we will use a continuous index or a discrete periodization of the intervention.

3.4   Reporting quality

To estimate the impact of KenyaEMR on the quality of reporting, we will compare aggregated monthly reports on HIV activities in facilities produced before and after implementation of KenyaEMR. Evolution of reporting quality involves two evolutions: amelioration of primary data quality, and amelioration of report computation quality.

We will measure data quality by looking at specific metrics:

- Proportion of data fields used to compute reports that have contain valid data

- Mean monthly number of visits by active patients

We will also use RDQA data to evaluate the quality of the data. Using RDQA results as training results, we will explore systematic classification of data quality based on reports indicators and patient forms metadata distribution.

We will then measure the evolution of data quality between RDE and POC data in KenyaEMR and we will perform simple comparisons to evaluate changes in data quality when entering data directly in computerized form.

Also, we expect computation quality to have multiple measurable impacts:

- Greater coherence of indicators involving longitudinal data analysis,

- Greater coherence of indicators involving multiple data sources

- Greater coherence of indicators evolution in time, as computerized computation will be exactly the same in time

- Greater coherence of indicators between facilities, as computerized computation will be exactly the same in all facilities.

We will compare reports generated for the same facilities and same months, in Period 1 and Period 2, and we will perform simple comparisons to evaluate changes when using standardized computation methods.

Based on these two dimensions of reporting quality, we will finally design an index of reporting quality that will be used in subsequent analysis. Quality of reporting will then be modelled using, using facility characteristics as covariate, and the index of maturity of implementation. The coefficient associated to maturity of implementation will be considered as the measure of the impact of KenyaEMR on reporting quality (see section Quality of Care and patient health outcome4 for presentation of the modeling strategy).

3.5   Quality of Care and patient health outcome

Using existing aggregate-level longitudinal data from KenyaEMR sites, we will retrospectively compare quality of care and patient health outcome indicators during each period of the EMR transition. The specific quality of care and patient health outcome indicators to be examined will be determined in collaboration with CDC and the MOH, based on commonly-used indicators within Kenya and globally. A list of these indicators can be found in Annex C.

To model the association between using KenyaEMR and the level of each quality of care and patient health outcome indicators, we will use Generalized Estimating Equations (GEE) that will allow us to take into account the temporal correlation of observations. Covariates that we will introduce in this model include:

- Facility type

- KenyaEMR implementation maturity index

- Reporting quality index

- Number of patients followed for HIV in the facility

- Number of HW involved in HIV care in the facility

- Time trend

The coefficient estimated for KenyaEMR implementation maturity index in this model will be considered as the measure of the impact of KenyaEMR on the quality of care and the health outputs of HIV patients. The index will be introduced in continuous form or in dichotomized form. Alternative proxy of KenyaEMR implementation will also be tested such as period of implementation as defined for quality of care in table 1.

3.6   Timeline

Figure 4 presents a timeline for the realization of this objective. Even though the data collection process could be a sort of blackbox, we expect this paper to be finished by February 2017.
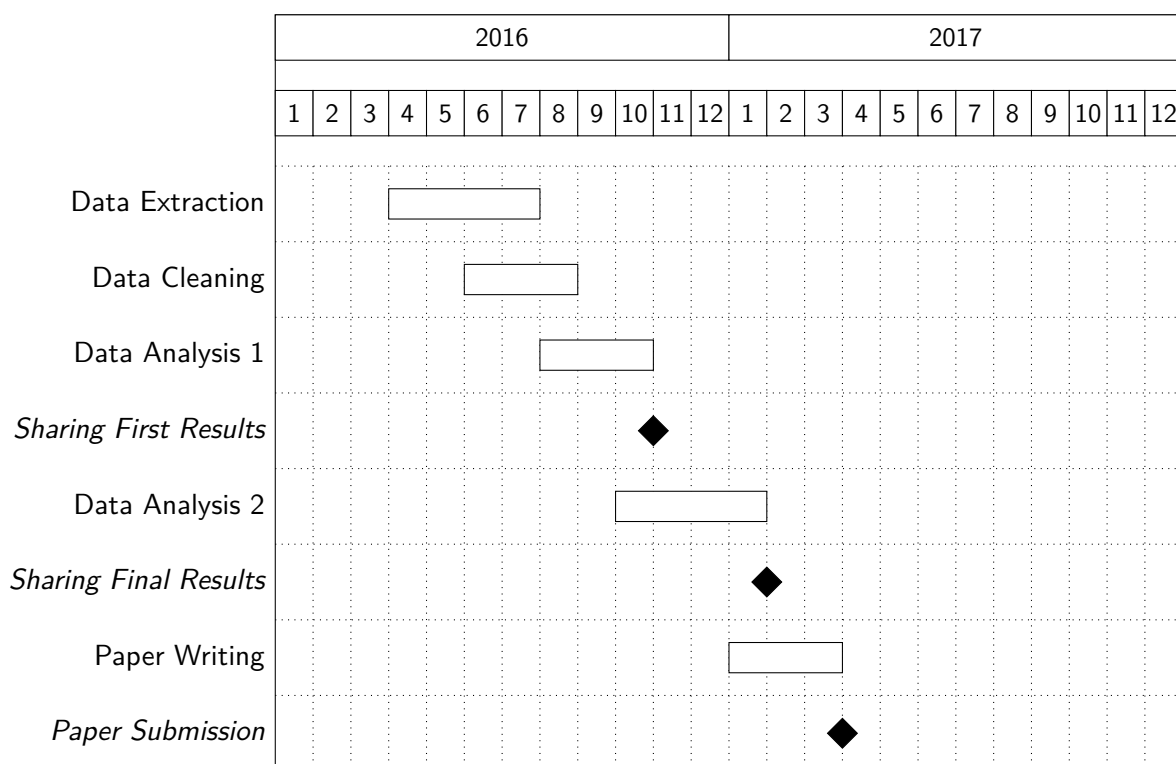
Figure 4: Gantt Chart for Paper 1

4 Semantic approaches to HMIS interoperability for external data validation

The second paper of this dissertation regards the management of data collected in hospitals in developing countries. The lack of standardization of indicators computed in different places or by different HMIS makes the use of HMIS data a complicated task. This issue is often handled as an issue of interoperability between systems.

4.1 Introduction to Interoperability work

There are multiple reasons why interoperability between HMIS datasets is a great asset :

Comparison Being able to compare the results of different health systems is essential to be able to benchmark these results, or to make different systems benefit from each others' experiences.

Validation / Completion When multiple systems are operating in the same place, one can wish to compare results from different system in order to validate the data, or to fill missing data from another system.

Co-analysis Finally, pooling results from different systems provides higher power for analysis that can be made on different subjects.

It should be noted that the conditions for interoperability can be seen in different ways for each of these uses. If comparison of results necessitates that measured indicators have quasi identical definitions and methods, it is less the case for validation and completion, where a set of indicators can be used as a proxy to check the coherence or impute values of another data set. Finally, co-analysis may or may not require an exact mapping of indicators from different systems, depending of the subject matter.
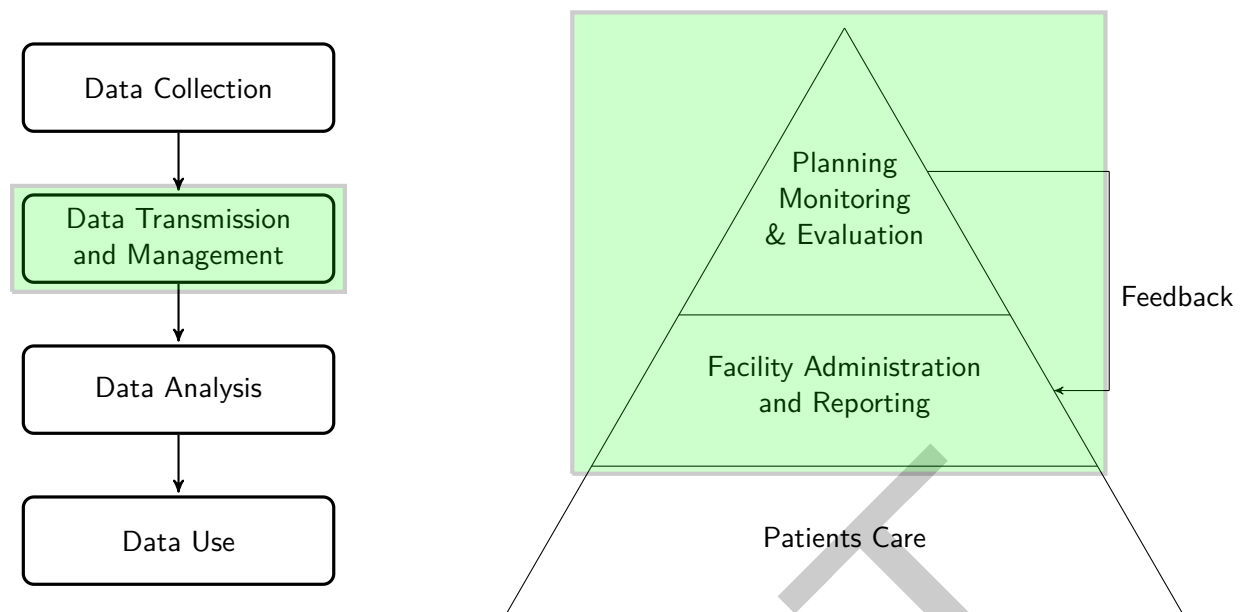
Figure 5: Objective two definition

There are multiple levels at which interoperability of data systems has to be enforced [2]. At the Syntactic-Technical level, protocols have to be designed and implemented, to ensure that different data-systems can communicate. At the Organizational level, processes have to be implemented to allow the exchange and to define the condition of usage and aggregation of different data systems. Finally, at the Semantic level, qualitative meaning and understanding of the nature of the data being exchanged and compared has to be enforced.

I am interested in considering the semantic level of interoperability. Indeed, I will make the hypothesis that perfect standardization of HMIS indicators across contexts and platforms is an elusive goal, and may not even be desirable, as it is a factor of rigidity for HMIS, which should be able to evolve rapidly. I am thus be interested in defining methods to ensure ex-post semantic interoperability between different HMIS systems, in order to map indicators and with an objective of data validation and missing data imputation.

4.2   Research questions

This project is in the framework of a larger project, defining an interoperability framework between different systems. As part of this project, a tagging system has been defined that allows mapping of indicators from different standard hospital indicators sets.

Meanwhile, this tagging approach, if it is effective, has an important entry cost for users. Tagging a set of multiple dozens of indicators can indeed been an harrowing task in a field where quick fixes are the gold standard. We are thus working on two ways to improve this uptake. One is the definition of automatized learning methods that could help users in the tagging task. Another approach is to provide users with sufficiently strong incentives that the tagging work will be worth going through. As actors working towards interoperability of systems are likely to be middle tier actors in the information system, we think empowering them to use the benefits of interoperability for validation of data, correction of faulty data and completion of missing data would be a strong incentive.

I will thus explore methods to validate, correct and complete data sets from routine HMIS, and measure the benefits of indicators matching between different dataset to improve validation, correction and completion. This will be considered looking at different intermediary questions :

1. What is a good metric to assert data quality and completeness of a given HMIS data set ?

2. What is the performance of different approaches to HMIS data validation, correction and completion ?

3. What is the benefit of using mapped datasets on this last metric ?

Our main aim will be to gauge the credibility of a data value or a dataset. Additionally, we want to provide some evidence as of what the source of error or missingness in a dataset is. There are three main situations we want to be able to differentiate between :

simple outlier are situations when an isolated data value in a dataset is wrong, for one facility once. These are the situations that are the most commonly recognized as outliers.

outlying report are situations when all values of one report appear to be off. This may be due to an update in the tools or methods for some indicators, or to the training of a new Health Worker in the facility who does not fully comply with usual ways to compute indicators. To identify this type of issue, it is necessary to compare

outlying facility are situations when one facility is consistently reporting numbers that are different from surrounding facilities. This may happen when structural conditions in one facility are leading to discrepancies in data collection, or in data computation.

Our general approach will be to attach to each data value a probabilistic value for its credibility. The combination of these credibility measures for a dataset will give an overall estimation of the credibility of this dataset, and the type of error it may suffer from. In a third and last step, we will explore methods for imputation of data values with low credibility or for methods with low credibility.

4.3   Data Validation / Method

We will test two approaches to do so :

Error prediction Using the validation dataset from OpenRBF, we will try and predict wrong data values using a simple predictive approach and bagging different Machine Learning Classification methods. Result of this approach will be a probabilistic assessment of data quality for each indicator value.

Variable imputation Using all available data, we will impute all data value and get a posterior estimated distribution of the value.

4.3.1   Error Prediction

For a each indicator value, we will model the probability that the value is right. We will use a logistic model specified as :

Using *Forward model selection* and training models on data on verified values (cf. 4.6), we will find the best models to predict data value errors. We will then use this model out of sample to compute the probability a value is true on each value.

4.3.2   Variable Imputation

For a given indicator X_{i*} measured at time $t*$ in facility $f*$, we will compute an imputed distribution based on all other information available in our data. A general representation of this approach could be written as:

$$\widetilde{X}_{i*,t*,f*} \sim \mathrm{D}\left(f\left(X_{i,t,f}\right)_{(i,t,f)\neq(i*,t*,f*)}\right)$$

where D is an unspecified distribution derived from a function of all available data excluding the data point $\widetilde{X}_{i*,t*,f*}$.

Using verified data as our validation set, we will define a threshold for credibility of data that we will later use as a decision rule for considering data as regular or outlying (see 4.4).

We will test different approaches for this model.

4.4   Credibility Pattern Screening

The validity probability of each data value will be pooled at report, facility and district levels. An analysis of the distribution of credibility at each level of aggregation will be made to characterize the type of data error pattern in one of four categories : No error, Single outliers, Outlying report or Outlying facility.

No error will be situations in which no data values will be over the fixed threshold of acceptable credibility.

Single Outliers will be situations in which less than 10% of data values in a single report for a facility will be under the fixed threshold of acceptable credibility.

Outlying report will be situations in which more than 10% of data values in a single report for a given facility will be under the fixed threshold of acceptable credibility.

Outlying facilities will be situations in which more than 10% of data values will be under the fixed threshold of acceptable credibility in more than two consecutive reports for the same facility.

4.5   Performance Metric

Validation of these methods will be made using cross-validation. We will first evaluate the data validation framework independently of the indicators mapping framework. We will then test the performance of the combination of different indicator mapping methods and data validation methods.

4.6   Data

This work will be conducted in collaboration with the Belgian startup Bluesquare. Bluesquare has developed a data system for management and validation of Results Based Financing indicators, called OpenRBF. This solution has been implemented in Bénin in XX facilities in YY départements. Indicators are collected on a monthly basis in OpenRBF, and a data validation system is in place, to check the accuracy of reported indicators.

In the meantime Bénin has been implementing and using DHIS2 nationwide since AAAA. There is considerable interest in mapping indicators from the two systems, and using the two systems as validation and completion solutions.

4.7   Timeline

Figure 6 will present a timeline for the realization of this objective.
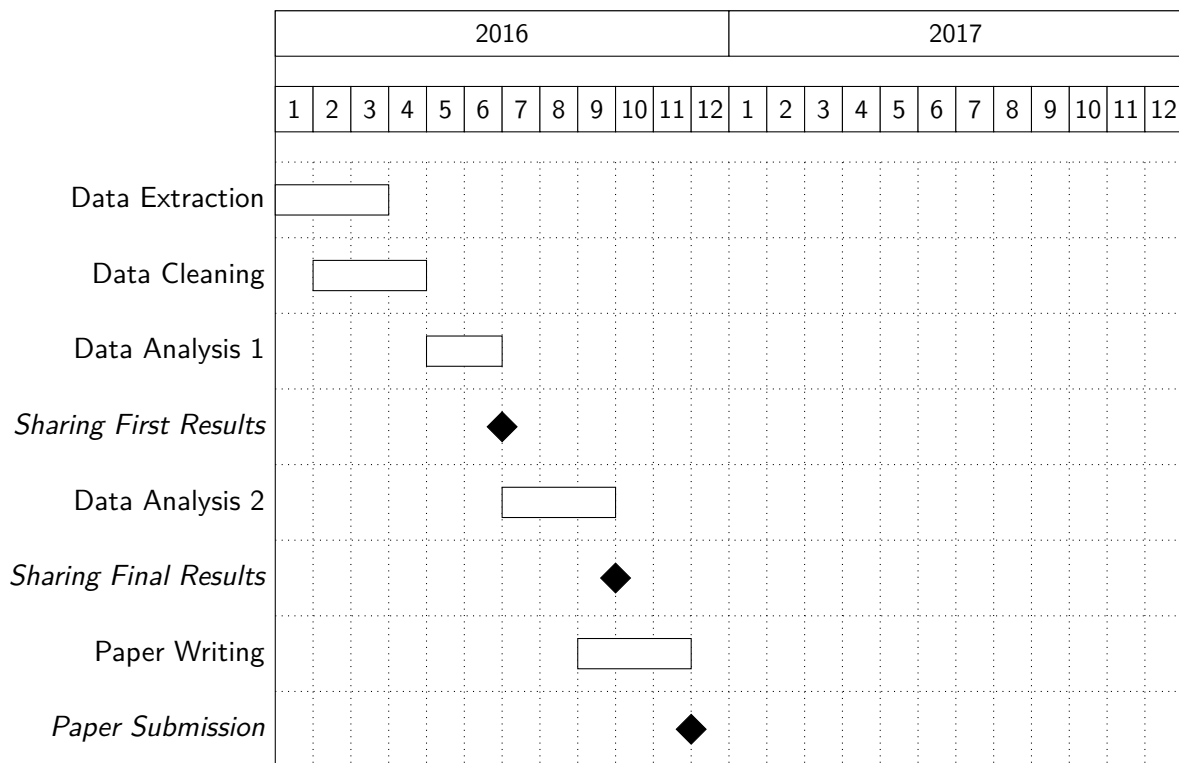
Figure 6: Gantt Chart for Paper 2

5 Innovative analytic approach - data integration for malaria elimination[2]

Our third aim will be to explore an innovative approach to data integration, to provide an elimination metric for malaria elimination.

5.1 Context

Malaria elimination requires programs that are able to monitor and analyze big and complex amounts of data in order to make effective decisions to fight local transmission hotspots. In this situation, reducing the geographic scope to monitor and concentrating attention on areas where malaria transmission is still happening or is likely to happen in the future is essential[6]. By tracking and investigating any new case of malaria, programs are able to focus their action on the most at-risk zones and communities. Surveillance is thus a key element of malaria elimination programs[7] , [8]. Passive and active case detection systems are used to identify and characterize malaria transmission hotspots and orient preventive measures. One of the challenges of malaria elimination is in identifying cases that are increasingly scattered and rare and thus more likely to be missed[4]. Our project aims to develop a novel surveillance analysis tool to assist decision makers to strategize and prioritize their efforts towards elimination, by measuring the degree of confidence they can have that no malaria transmission happened in a given zone.

Elimination is indeed only defined negatively as the absence of detected incidence cases. The closest we can measure is a probability that no transmission is happening. Malaria elimination specialists' work should thus be oriented towards maximizing the likelihood that malaria has been eliminated (i.e. minimizing the probability that a local malaria transmission will happen in the future), and reducing their uncertainty surrounding this probability to a minimum. In order to do so, their portfolio of methods includes: implementing prevention strategies in order to lower the risk

---

[2]This Section is currently the adaptation of a grant proposal submuitted for this aim
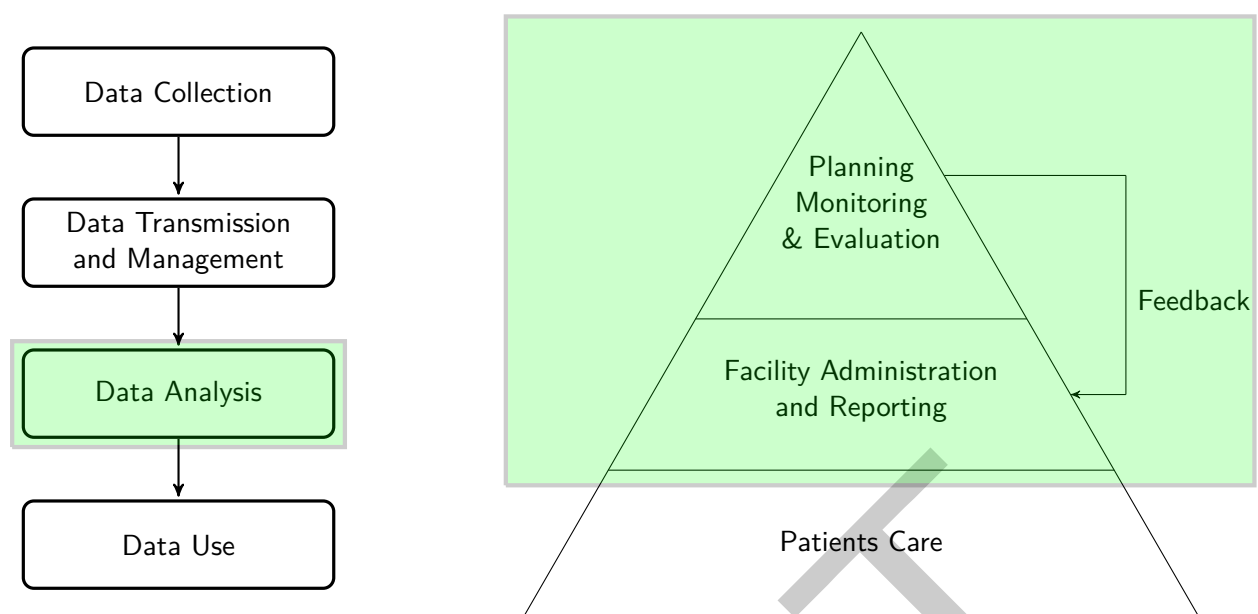
Figure 7: Objective three definition

of transmission; improving the health sector's supply of malaria diagnostics and treatment services in order to improve passive case detection, and thus diminish the uncertainty around untested cases and lower the risk of further transmission; launching and targeting active case detection campaigns to confirm the absence of incident cases in the population; and planning other data collection activities (ecological measures, population practices...) in order to improve the uncertainty of background models of malaria transmission risk. For example, if no malaria transmission is recorded in a community, an entomological survey may be an essential piece of information to maximize the likelihood of elimination, and may be more efficient than implementing active case detection or strengthening facility based diagnostic, but it may be hard to compare the informational benefits of each of these strategies. Conversely, in a zone with low elimination likelihood, a public health professional will be able to understand whether this low likelihood is due to a high observed incidence, necessitating improved prevention, or to low completeness of health service records, or by long overdue contextual data collection. The measure will thus allow malaria elimination strategists to target their resources on the most efficient action.

We will develop a robust method for analyzing malaria surveillance data, using a continuous metric of malaria elimination. We will define, measure and evaluate this innovative metric, combining the evidence of elimination from available data with a measure of data quality and completeness. This will provide an analysis tool to plan and measure local efforts for malaria elimination.

This project will focus on data from Namibia. Namibia has set malaria elimination as one of its priorities[11], and has fully integrated surveillance and tactical planning as its approach for the elimination[3],[9]. Moreover, the National Malaria Strategic Plan comes to an end in 2016, and if successful, our project could contribute to the design of the next strategic plan. Finally, Namibia is currently undergoing an outbreak of malaria the analysis of which will offer a good validation framework to our project. Our team has relationships with the MoH in Namibia and we are confident we will able to access most of the data we need for this project.

5.2 Data

We will first collate a data warehouse of all relevant data available in Namibia. This will include ecological, epidemiologic, HMIS and case detection data; socio-behavioral studies; and wide scale demographic studies. We will also access and store results from the Malaria Atlas Project. All data will be geolocalized. For data sources with no precise geolocalization, we will implement an

approximate location approach. In order to assess data quality at local level, we will also keep a running list of existing data sources that we could not access. We will assess data quality of each data source, combining detailed examination, algorithmic approaches and expert advice.

5.3   Likelihood Metric Modelization

Our first modelling step will be to create a measure of data quality. We will define a relative measure of available information quality for a certain time in a certain zone. This measure will be a composite of data availability and data quality of available sources, with maximum being the best observed score for any time period. We will then fit a simple model of the probability of malaria elimination, as informed by currently available data. The probability of elimination will be measured at a certain time in a certain zone as the probability of occurrence of locally transmitted cases in later periods. Covariates will include the available modelled results of malaria risk, and other localized epidemiologic and contextual data collected in the previous stage. Multiple approaches to integrating impact of data quality will be tested in this model.

Success of the modelling stage will be evaluated using the ability of the likelihood to give a meaningful measure of malaria elimination using case detections as validation in zones where case detection is known to be most complete. Our general approach will be validated if we find a significantly lower likelihood of elimination in zones with new outbreaks than in zones with no new outbreaks.

We will test the sensitivity of the model to the different health officials' decisions defined earlier, regarding additional data generation activities or public health intervention, by simulating how each decision could impact the elimination likelihood. If sufficient cost data can be accessed for each intervention, we will also cost each intervention. For different scenarios, we will consider our approach to be useful if we can recommend a clear best decision.

Our model could be expanded to be usable by malaria programs. We will develop a software that will generalize our data aggregation approach and will allow users to query elimination likelihood and expected results of interventions at precise locations. This surveillance analysis tool will be piloted at different levels of the Namibian health system and will be made available to other interested programs. This approach could be generalized to other geographies and public health issues (surveillance, intervention coverage).

5.4   timeline

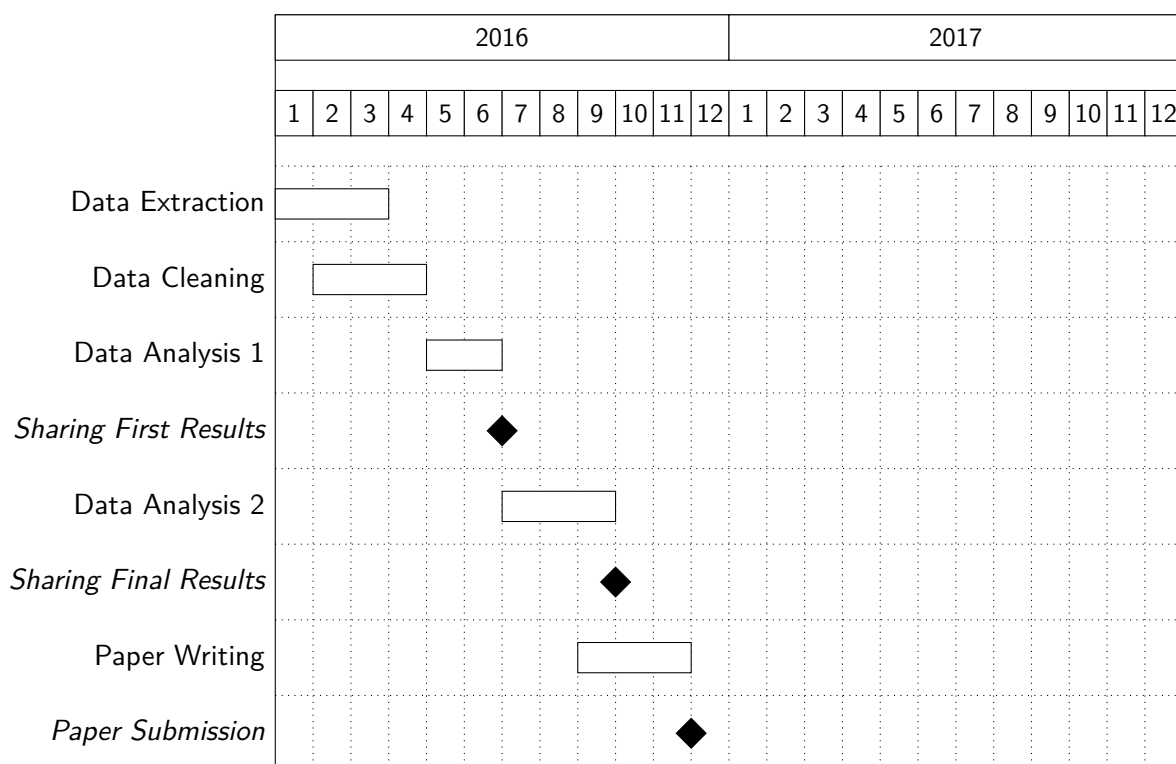Figure 8 will present a timeline for the realization of this objective.

Figure 8: Gantt Chart for Paper 3

6 Data Use

construction d'un espace politique d'équivalence et de codage

"Les outils statistiques permettent de découvrir ou de créer des êtres sur lesquels prendre appui pour décrire le monde et agir sur lui"

6.1 Information in Health Systems

The use of statistical information for the management of complex organizations has evolved since the beginning of the XIXth century. Since the invention of population by XVIIIth century demographers[DESROSIERES], and the integration of numbers in the political and administrative language in the second XIXth century [PORTER], multiple types of information have been used for the orientation of public policies and the administration of public services. Meanwhile, the rise of epidemiology and the critalization of a body of knowledge around the institutions in charge of the defense of Public Health helped creating a specific Public Health oriented spin on quantitative information for health systems.

The use of data for policy making is a combination of data sources, statistical methods, and political or social norms, that will define the conditions of utilisation of statistical evidence for policy making. Finding the proper data source, being able to analyze it and incorporating the results of this analysis in a political process is essential to the proper use and utilization of information systems. In this regard, Alain Desrosières has shown how two traditions have been cohabiting in the early ages of the production of social statistics[5].

> The first tradition is administrative, and is based on political science and the law, on the German Staatenkunde, from the time of Conring and Achenwall. It is more taxonomic than metrological: it is designed to classify facts systematically rather than measure them, which is the essence of the other tradition, the "English" tradition. The latter, inspired
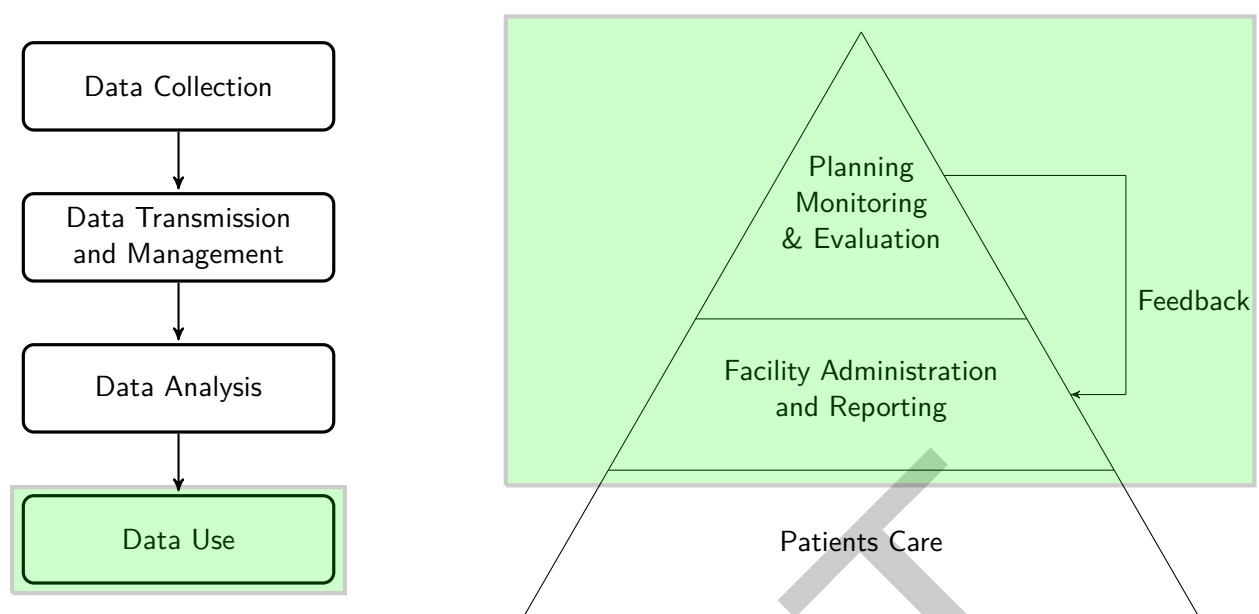
Figure 9: Objective four definition

more by the natural sciences and by progress made in measurement and probability theories, is a distant relation of the English political arithmetic of Graunt and Petty.

Desrosières later shows how these two traditions have bee reconciled in the modern figure of the statistician, at the same time administrator and scientist. It is useful to keep considering this tension when thinking about maturing statistical systems like HMIS. Being able to distinguish between situations when actors of HMIS are acting as administrators, and when the position is that of a metrician is essential to understand HMIS issues and offer informed solutions.

This distinction is essential at many levels. The whole debate around the level of uncertainty that is bearable around a measurement is not only important for statisticians. Choosing a given approach will have an impact on how primary data will be collected, how it will be analyzed, and how it will be used. In many usages of HMIS, complete enumeration is deemed necessary, but this can be discussed. What is the level of confidence one can bear around the estimation of a stock of drugs ?

In sub-Saharan Africa, this tension is reinforced by a political tradition that has been structured around the colonial state. The structures and political traditions coming from this specific have complex relationships with the notions of uncertainty and control. Moreover, these structure are reenacting the colonial culture of exogeneous power structure, through that international actors take in the governance of African country.

This last paper will aim at understanding how some program managers in Mali the data available in HMIS, and how it impacts the way they think, design and implement HMIS programs. We will interrogate the notions of uncertainty, sampling, control and norms for this managers, and their appreciation and use of numerical evidence.

UNFINISHED

Reflections on social conditions of HMIS data usage / politics of administrative statistics.

Data is not produced to create knowledge, but to implement disciplinary monitoring. Thinking mainly in terms of indicators.

Figure 10 will present a timeline for the realization of this objective.
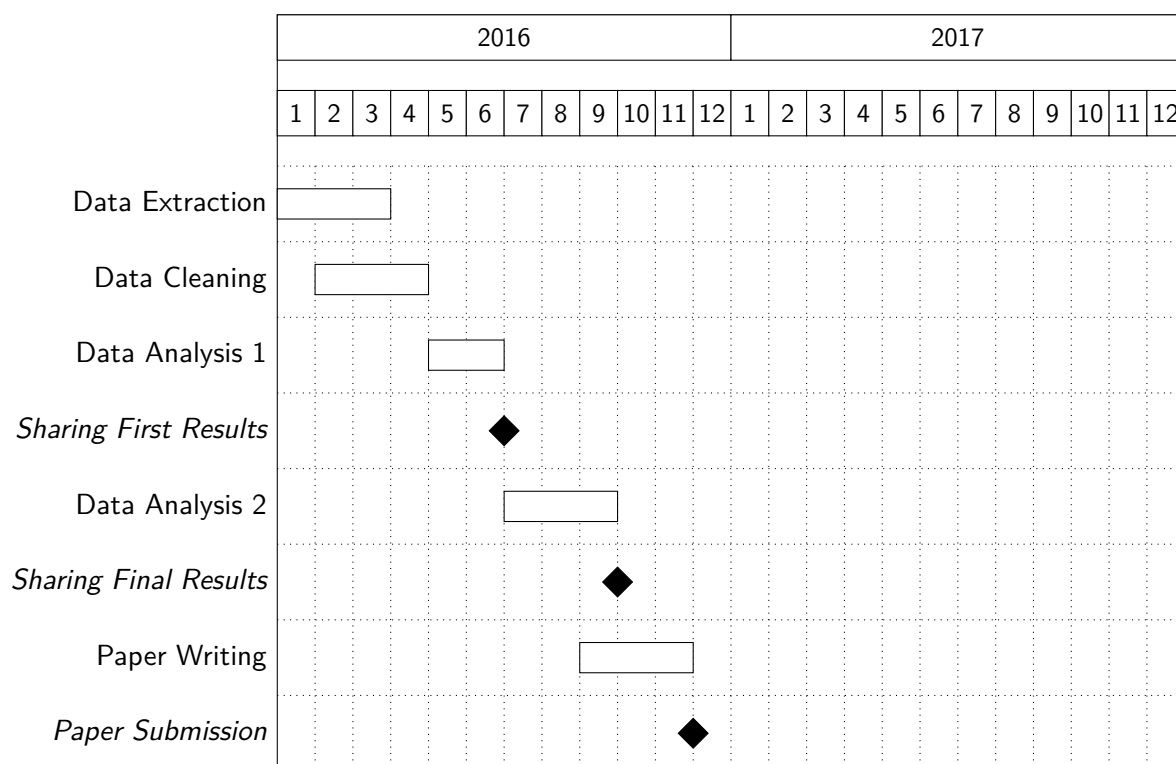
Figure 10: Gantt Chart for Paper 4

References

[1] Abou-Zahr, C., & Boerma, T. (2005). Health information systems: the foundations of public health. *Bulletin of the World Health Organization*, *83*, 578–583.

[2] Braa, J., & Sahay, S. (2012). *Integrated Health Information Architecture : Power to the users*.

[3] Chanda, E., Ameneshewa, B., Angula, H., Iitula, I., Uusiku, P., Trune, D., Islam, Q., & Govere, J. (2015). Strengthening tactical planning and operational frameworks for vector control: the roadmap for malaria elimination in namibia. *Malaria Journal*, *14*, 302.

[4] Cotter, C., Sturrock, H., Hsiang, M., Liu, J., Phillips, A., Hwang, J., Smith Gueye, C., Fullman, N., Gosling, R., & Feachem, R. (2013). The changing epidemiology of malaria elimination: new strategies for new challenges. *The Lancet*, *382*(9895), 900–911.

[5] Desrosières, A. (1997). The administrator and the scientist: How the statistical profession has changed. *Statistical Journal of the UN Economic Commission for Europe*, *14*, 31–50.

[6] Feachem, R., Phillips, A., Hwang, J., Cotter, C., Wielgosz, B., Brian, G., Sabot, O., Rodriguez, M., Abeyasinghe, R., Ghebreyesus, T., & Snow, R. (2010). Shrinking the malaria map: progress and prospects. *The Lancet*, *376*(9752), 1566–1578.

[7] Group, G. H. (2014). *Surveillance Systems to Facilitate Malaria Elimination*.

[8] Organization, W. H. (2012). *Disease Surveillance for Malaria Elimination, an operational manual*.

[9] Programme, N. V.-B. D. C. (2010). *Namibia Malaria Strategic Plan 2010-2016*.

[10] Pugliese, P., Cuzin, L., Cabié, A., Poizot-Martin, I., Allavena, C., Duvivier, C., El Guedj, M., De La Tribonnière, X., Valantin, M., & Dellamonica, P. (2009). A large french prospective cohort of hiv-infected patients: the nadis cohort. *HIV medicine*, *10*(8), 504–511.

[11] Smith Gueye, C., Gerigk, M., Newby, G., Lourenco, C., Uusiku, P., & Liu, J. (2014). Namibia's path toward malaria elimination: a case study of malaria strategies and costs along the northern border. *BMC Public Health*, *14*, 1190.