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

HEALTH ALLIANCE INTERNATIONAL

Abstract Here the abstract will come

Contents

List of Figures

List of Tables

1 Introduction

If a unique literary form had to be chosen to write or talk about Health Information Systems (HIS), the Complaint would probably be the best pick. Be it complaints on the burden of work involved in collecting, managing and analyzing data in health systems [BIBLIO], or laments on the inexistence of good quality data in most developing countries health systems [BIBLIO], HIS are usually described as non performing burdens of health systems that can only be improved [HMN citation]. This frustration is only matched by the expectations placed in HIS and their widely recognized importance, some authors calling HIS "the foundation of public health"[1].

This proposal defines a research project aiming at exploring an approach to HMIS that expands the way HIS are usually considered. Combining a critical approach to HMIS with some technical research, we want to deconstuct approaches that are too often made about the need for a heavily normative approach to HMIS, and offer specific alternative approaches for an amelioration of HMIS' usefulness.

1.1 Information in Health Systems

The use of statistical information for the management of complex organizations has evolved since the beginning of the XIXþcentury. 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 [BIBLIO]. The needs for information and data thus are varied at different levels of health systems

GRAPH PYRAMIDE ?

Information at national level

Information at

Importance d'information pour l'administration de systèmes complexes

Importance de l'information statistique

Importance de l'information dans les systemes organises [edgar morin] Importance historique dans le developmment des systemes de sante

"Les outils statistiques permettent de découvrir ou de créer des êtres sur lesquels prendre appui pour décrire le monde et agir sur lui" - citation sur la decision rationnelle. mais on sait que c'est un construit. Social et technique. Cas spécifique, statistique sanitaire

tension entre valeur de l'etat et de la science se retrouve jouer à plein dans les HMIS ie HMIS sont objets composites. On veut montrer l'ambivalence de SI comme objet technique et politique. L'appproche utilisée pour les HMIS se fait surtout en termes de standardisation, et de précision croissante au niveau de la collecte de données. Approche critique des systèmes d'information. Réflexion se fait à la rencontre entre approche Science Studies et sciences humaines, et les avancées en science de l'information / statistique / Artificial Intelligence. Using multiple approaches, we wish to be able to defend a vision of HMIS that maximizes value of data collected and analyzed, by defending an approach centered on producing value at local level of health systems, and depending on non standardized processes at higher levels of hte health system to produce value and use data.

We will do so pursuing three aims related to how

Aim 1 Evaluate benefits of improving data collection techniques on the care of patients

Aim 2 Define a method for ensuring interoperability of different data systems commonly in HMIS

Aim 3 Analyze data with innovative approaches for monitoring of health services performance

Aim 4 Provide a broad historical and political analytical framework on how HMIS have been formalized and implemented in selected countries to understand how hmis are thought and conceived by health officials

decrire c'est gérer . role de description et de prescription

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

The use of statistical information in the design, the implementation and the evaluation of public policies is of growing interest in different domain of public life. Different trends in the thoughts and traditions of public life have led to this strengthened importance.

The improvement and the diffusion of tools and resources available for the production of public statistics have certainly be a first trend that has led to a better availability and use of data for decision making. In the meantime, the culture surrounding the production and use of numerical data in modern societies has evolved, both in reaction to increased capabilities of measurement, and of evolution of political and management sensibilities in these societies.

The use of data for pubic decision is consubstantial to the apparition and the development of public as a domain of public action. The "invention of population" in the second half of the XVIIIth century was made possible by the reform and development of demographic information in Europe and the development of demographic methods. In later stages, the development of sampling and inferential statistics methods in the XIXth century was also key to the targeting of specific public health interventions.

The use of data for policy making is thus, as we see, 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 Global Health realm, the use of data for the definition of *evidence based* intervention and policies has emerged as a panacea of project design and management. There are nonetheless difficulties in this regard. The global nature of public health means that statistical data available for analysis is by nature scattered and varied in nature, technical characteristics, quality and scope. In the meantime, the exigence of Global Health practitioners is to use and understand varied data sources in a unified global framework. The Global Burden of Disease initiative is a good example of this exigence of a global assessment of a wide variety of data from multiple contexts.

The challenge of using and processing different kinds of data varies with the nature of the data sources. The design and definition of survey data, for example, is governed by methodological and technical constraints that are comparable between settings and implementations. Meanwhile, data from health systems will be influenced by multiple factors, ranging from the administrative traditions in which they develop, to the level of resources involved in the design and building of these data systems, and to the type of activities performed in these systems. Among them, hospital data could arguably be considered the most impacted by these different factors.

1.2 Importance of Health Management Information Systems

GRAPH FONCTIONS

Difficulté supplémentaire vient du fait qu'on parle de systèmes qui doivent s'inscrire dans la durée et qui doivent permettre à la fois une connaissance de la santé de la population et permettre le monitoring des activités des HRH (dual usage) mais also qu'ils impliquent le travail

Among different data sources Health Management Information Systems (HMIS) are specific in so far as they are designed and thought, from the very beginning, to fulfill multiple purposes. If a survey is implemented, the only objective of its data collection tools is to collect data fitted to the sole purpose of completing the survey's objectives. In the meantime, HMIS typically rely on personals and resources whose primary goal is not the collection of data, but have other functions in the health system. This is also a difference with data stemming from sources specialized systems, like a survey implementations, for which every resource involved is aiming at producing quality data.

This non specificity and non specialization of HMIS is key to understand both its importance and its challenges. The overarching importance of Health Management Information Systems (HMIS) in modern health systems[1] is as well recognized as the inability of most developing countries to implement well-performing HMIS. HMIS are important to provide information on the BIBLIO.

The low performance of HMIS comes from multiple origins. BIBLIO.

We contend that the difficulty of designing and implementing HMIS that deliver usable information for decision makers comes from technical difficulties, as well as from detrimental technical and organizational choices. HMIS are primarily considered functions of a health systems, and not as primarily statistical systems. As a result, in many situations, the logics and choices that are made in health systems are guided by administrative culture and the organization of health care in specific contexts, and are not primarily designed to perform as data systems. In this regard, HMIS are mainly aimed at very specific goals, limiting interoperability of systems and building systems that are not flexbile and are not designed to give information outside of a very limited, preset framework.

M&E = version la plus degradee

what is and what is not part of HMIS The reasons for this weakness are as varied as HMIS are complex objects. Many functions are involved in developing and implementing an HMIS. There are different views that can be adopted to describe a HMIS. Some authors privilegiate the demand side of HMIS, by describing HMIS trough their end users and why they need information. Some authors will privilegy goals of a specific HMIS users. Some authors will concentrate on describing what should be included as being part of an HMIS. Finally some other actors will prefer to address different functions that are exerted inside of HMIS.

Even if the understanding of what should be considered part of a HMIS may vary depending on authors, every source regarding HMIS usually refers to the importance of using data for decision making at every level of health systems.

In the meantime, there is a role for HMIS in organizing work in health systems more widely, as the way data is collected has an impact on how work is organized inside health systems.

What are the main parts of HMIS ?SUBSYSTEMS : cf document liberia / openHIE : tous les differents composants d'un systeme d'information

Different uses ?

HMIS mainly considered as a normative object in society. We contend this is not the only proper approach to it, and problematization can add benefits and create space for new solutions. -> confusion between creating norm number and norm tools

Thinking about and working around HMIS requires different levels of thinking.

organizational : what makes an HMIS work, and what functions should technological : right level of technology. adapt for best usage by different users (Illich) analytical : what techniques of analysis to use with very specific data political : how this data should be used to inform decisions ?

The objective of this proposal is to shape our angle of analysis in each of these approaches, and to identify research directions that will help furthering the understanding of HMIS in developing countries and offer solutions for currently important issues.

2 Conceptual framework - issues in HMIS designs

2.1 Functional approach

A first way to approach HMIS is to describe the principal functions that are necessary to have a HMIS to run. Figure 1 presents a simplified sketch of the principal functions that are to be filled in any HMIS.

Four main functions can be found in HMIS.

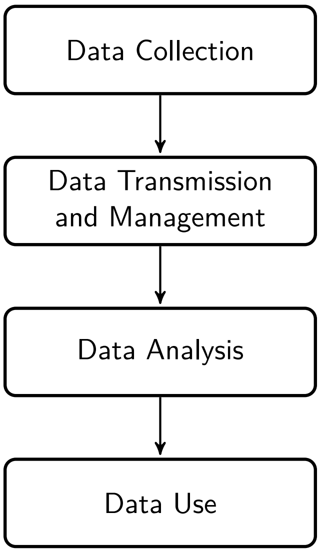

Figure 1: Different functions inside the Health Information Systems

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 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 happen through standardized reports, and 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, of the type of the DHIS2 software.

Data Analysis Data that is collected and stored in HMIS can then be analyzed. The type of analysis that is doable with EMR data will be different from the type of analysis that is possible to make with indicator type data.

Data Usage What kind of decisions ? Memoire Cheickna.

2.2 Goal approach

Another approach to HMIS is a consideration of the stated goals of these information systems. Figure 2 shows what these goals are. The pyramidal representation of these needs is used to show that these goals fill data needs at different levels of health systems.

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.

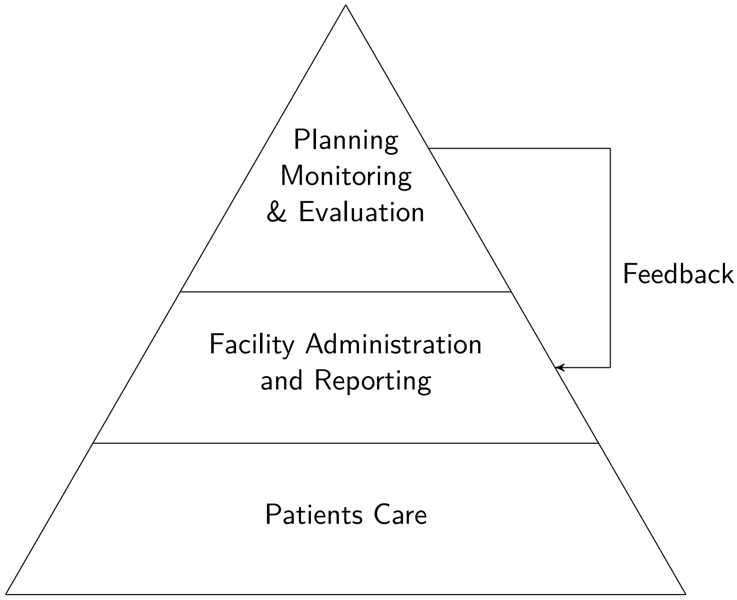

Figure 2: Objectives of HMIS

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.

2.3 Problematization and problem analysis

Using the framework we presented in sections 2.1 and 2.2, we identify important issues in the way HMIS are designed, implemented and used in developing countries. Our analysis is looking at the adequacy of HMIS methods and practices with the aims of these systems. We posit there is an overemphasis that is put on data collection in a lot of systems, which jeopardizes the way health information systems perform.

The reasons for this overemphasis can be traced to the intellectual frameworks that has surrounded the design and development of statistical systems in developing countries, and the evolution of what collecting data for organizations means.

2.3.1 The administrative legacy

There has been a long term evolution since the early XIXth century as of how data should be produced and used in health systems. As

reductio ad M&E

Question de la statistique coloniale. Les différents niveaux de la statistique administrative, importance de la justification et du contrôle dans l'utilisation faites des données administratives.

There is a primary problem in the use of HMIS data. Alain Desrosières has shown the richness and complication of the production and use of statistics in modern societies. Desrosières shows how two traditions have been cohabiting in the early ages of the production of social statistics[2].

> 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 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 developing countries' 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 other dimensions, how can civil society help for HMIS design and evaluation

2.3.2 Three HMIS strategies

Functions of HMIS (cf. section 2.1) 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 [3].

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, Most of interventions to improve HMIS are geared toward improving *data quality* or its availability, all characteristics that concern the data collection function. Meanwhile, facility reporting

Problème est l'équilibrage des différentes fonctions. Mauvaise articulation. Comment rééquilibrer.

Statistique publique

Non adptation de l'administration et des systèmes utilisés, comme montré dans le cas de l'algérie

Data collection should be the same for all the goals. Highly specialized data collection. Low specialization of other functions. This shows a bizarre profile. Data users everywhere. Very little data specialists.

2.3.3 Data collection as most important function

Our approach to HMIS relies on two premises. First, in resource limited health care systems, data collection tasks often rely on non specialized personnel, that has to cary out other tasks, such as taking care of patients or providing support services for the care of patients. HMIS tasks should be as simple as not to be detrimental to patients care.

Secondly, HMIS should be conceived as data systems at list as much health system functions. This means adapting a framework around HMIS that perceives HMIS mainly as a system that shadows other health system functions to collect data, and think of HMIS mainly as systems which primary goal to collect and analyse data, and should thus be optimized in this regard.

Holding these two premises in mind leads us to think and desing on methods that will be aimed at imrpoving the higher functions of HMIS, such as data management and data analysis, without putting an unwelcome data collection burden on actors of health systems. Our theory of change for HMIS is thus based on a strengthening of data management, data analysis and data usage in HMIS.

data collection improvements are strongly limited by current constraints whereas there is much more room for improvement in other functions.

Schema

Comment faire pour améliorer l'utilisation des données dans les systèmes de santé. Quels sont les aspects importants ?

Theory of change for HMIS ?

2.4 Objectives

Our main objective is to explore methods and approaches that can be used to improve the comprehension and the operations of HMIS in developing countries. Looking at different angles of HMIS, we aim at understanding how each of the different functions of HMIS can be leverage to improve the outcomes of HMIS operations.

The dissertation will be built in four parts, each part being a look at a specific function of HMIS. For each of these functions, we aim at providing a better understanding of the challenges met, or to offer solutions to improve outcomes of HMIS.

3 EMR and individual health

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.

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

HEALTH SECTOR AIDS RESPONSE GROUP (ARG/MOHS)
MONTHLY PMTCT DATA COLLECTION/SUMMARY FORM

Name of PMTCT Site:..

Month:.. Year:..

1. ANC	15 - 24	25 - 34	35 - 44	45 - 49	TOTAL
Total expected pregnant women for the month					
Women seen for the first ANC/New ANC attendees					
Women received pre - test counselling for HIV					
Women tested for HIV					
Women HIV positive					
Women HIV 1					
Women HIV 2					
Women HIV 1 & 2					
Women received post test counselling and test result					
Women counselled for infant feeding					
Male partners tested for HIV					
Male partners with a positive test result					
Discordant partners or couples					
HIV positive women assessed for ART eligibility by CD4 counting					
HIV positive women assessed for ART eligibility by clinical staging					
HIV positive women received ARV treatment for own health					
HIV positive women referred for FP from ANC					
HIV positive women offered FP during ANC/Post natal					
HIV positive women received ZDV at ANC					
2. Maternity / Delivery					
Expected Delivery in the health facility among HIV + women					
HIV positive women who delivered in the health facility (Normal)					
HIV positive women who delivered in the health facility (C/S)					
HIV positive women received NVP during labor					
HIV positive women received ZDV and 3TC during labor					
Women lost to contact					
Reported death of mother					
3. Postnatal					
HIV positive women on ZDV and 3TC after delivery (Tail)					
Women received complete ARVs for PMTCT					
Total number of expected HIV exposed children					
Number of HIV exposed children registered during the course of the					
Infants who received NVP within 72 hours					
Number of HIV exposed infants received Ziduvudine (ZDV) and Lamividine (3TC) for 7 days (tail treatment					

Submitted by:.. Received by:.. Verified by:..

Remarks:..

Figure 3: Sierra Leone PMTCT registry (2012)

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 4). These indicators will be aggregated on site in Kenya and transmitted for data analysis.

To monitor the maturity of implementation of KenyaEMR (see section 2), 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), we will consider counts of number of forms entered for a given period, and mean completeness of entered

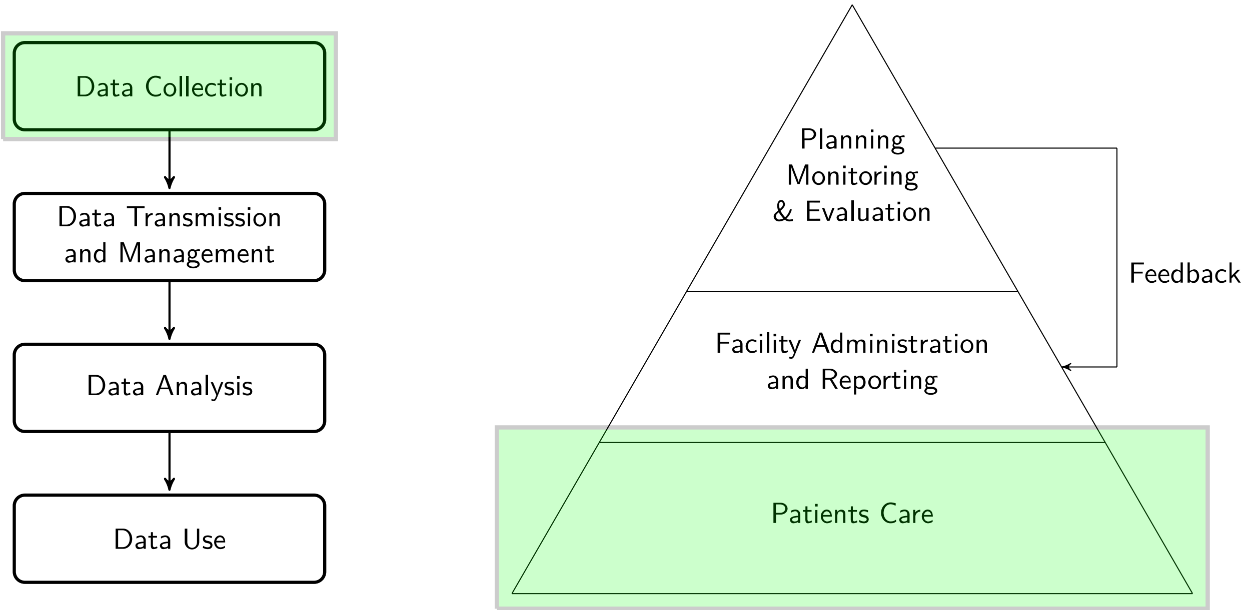

Figure 4: Objective one definition

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 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.

Table 1 presents a summary of the different periods described. 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 XXX 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 DDDD.

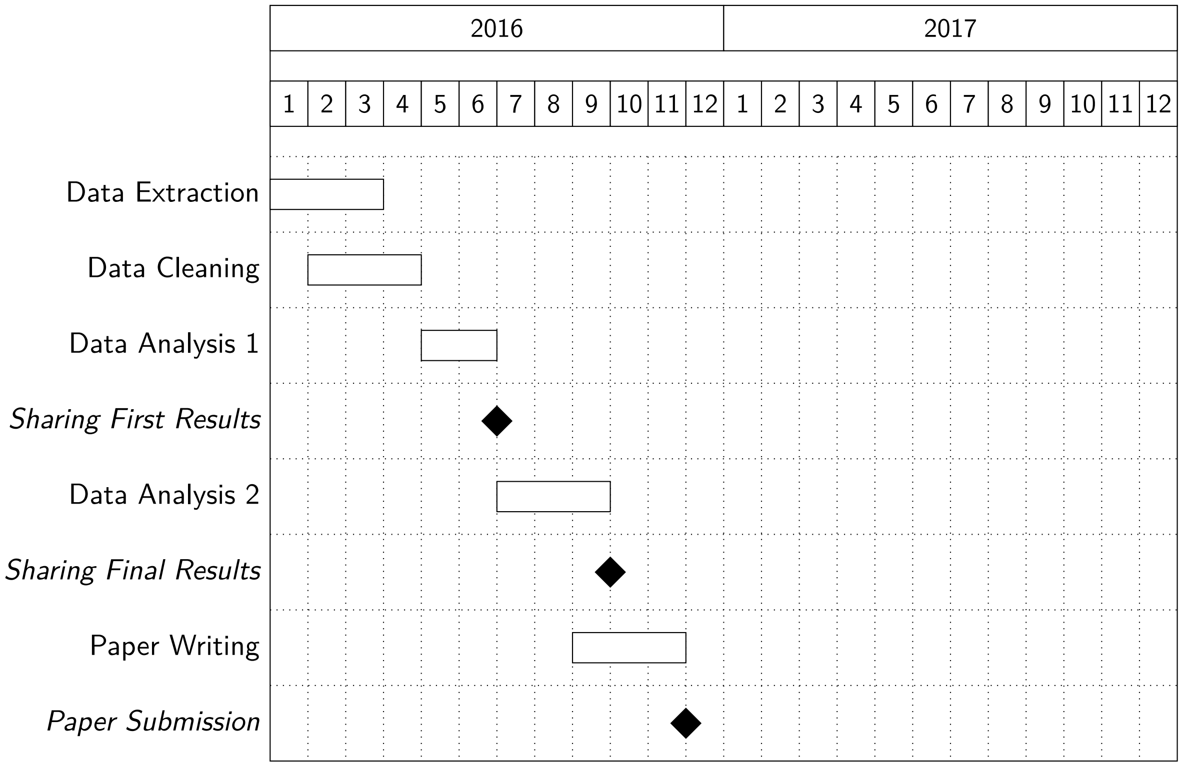

Figure 5: Gantt Chart for Paper 1

4 Data Management

The second approach of this dissertation regards the management of data collected in hospitals in developing countries. Many systems have been developed to store this data and use it in different situations. Meanwhile, some problems are frequently found, that prevent statisticians and public health scientists to use this data. Issues regarding data completeness and data quality are of primary importance. An issue regarding these questions is the lack of validation frameworks for data systems. Data coming from one system is seldom confronted, or even less validated with data from a different system. As a result, monitoring and definition of data quality is usually mainly a procedural concept for data producing entities, and a critical and often restrictive aspect of data analysis.

We aim at offering a set of methods that allow the use of different existing data sources in order to enforce better data quality and data completeness. This entails two main aspect :

1. A framework for interoperability of different data systems
2. A framework for validation of data

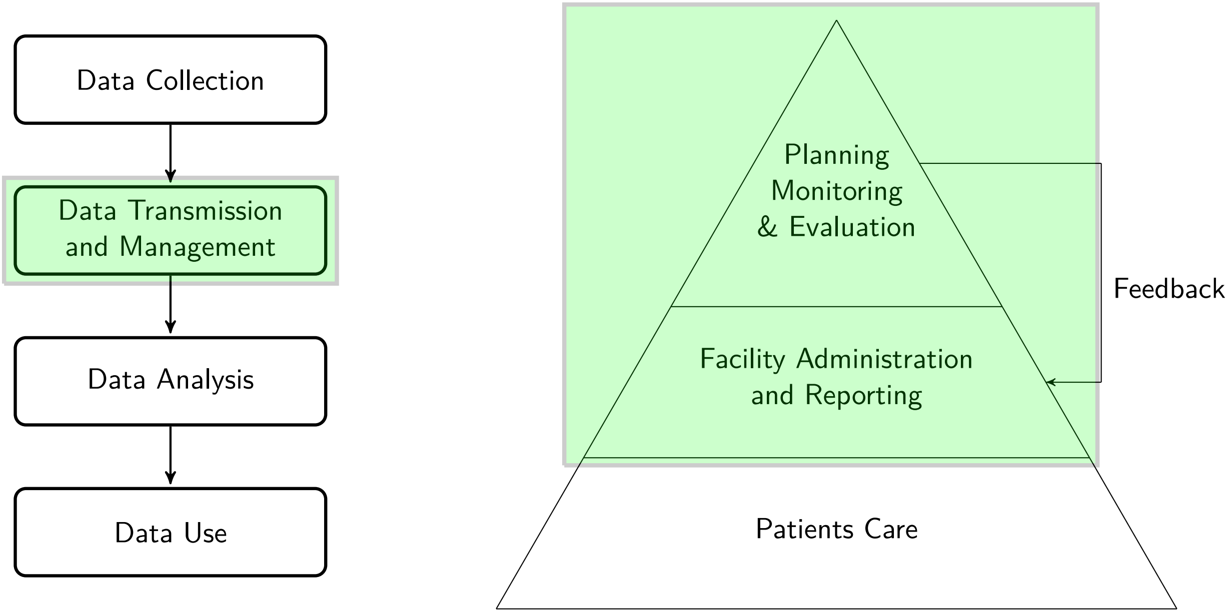

Figure 6: Objective two definition

4.1 Interoperability framework

POINT SUR INTEROPERABILITY BRAA / SAHAY / EXPLICATION TRAVAIL DEJA FAIT

Our aim will be to define and test these two frameworks, building on some work currently happening around interoperability frameworks. In this project, conducted with health information startup Bluesquare, we aim at defining a tagging based approach to interoperability. With this approach, each health service indicator in a given facility reporting framework is defined based on three dimensions : health issue, population target and type of service provided. A first level of testing has already implemented in Bénin, matching DHIS2 indicators with OpenRBF indicators. Using this simple method, it was possible to match correctly XX% of GG indicators from the OpenRBF framework with a subset of YY% indicators from the DHIS2 framework.

One of the biggest benefits of this approach is its ability to define a notion of *distance* between indicators. Indicators that related by one dimension will be farther apart from each other than indicators that are related by two dimensions. Also, for dimensions that are not defined on a discrete basis (for example age boundaries for a population), incomplete correspondence can be quantified.

Building on this flexibility, it is possible to automatically map datasets between each other, and to offer possibilities of joint analysis between data sets. A first level of analysis is the validation of data, informed by comparison of different data sets.

Used for data validation, imputation / completion.

4.2 Research questions

1. Define and test an algorithmic approach to assert data credibility from a given data set
2. Define and test an approach to compare and complete data missingness using imputation and external data

1. What is the performance of different approaches

2. What is a good metric to assert data quality and completeness of a given data set
3. What is a good metric to assert comparability of two indicators or data sets.

4.3 Validation framework

Our main aim will be to gauge the credibility of a data value or of a dataset. We can anticipate different situation we would like to investigate :

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 considered outliers. This may be the most straightforward situation.

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.

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.

We want to attach to each value a probabilistic value for the credibility of a data value, or of a data set. The combination and comparison of the credibility of multiple data points or reports at multiple levels in turn gives a fine grained picture of data credibility, and orients actions to be taken.

The data validation work will be conducted using data from OpenRBF Bénin, with data values entered and verified being recorded.

Our approach will thus be divided in multiple stages, as we want to be able to assess : local coherence , period coherence , neighboring coherence.

Estimation Multiple models

Description Dashboards

Decision Decision

DEFINITIONS ON DATA VALUES AND OTHER RELATED CONCEPTS

Using the validation dataset from OpenRBF, we will try and predict error using a simple predictive approach bagging different Machine Learning Classification methods. Result of this approach will be a probabilistic assessment of data quality for each indicator value. This, in turn, will be turned in a facility level dashboard, tracking estimated quality of reporting in time. A district level dashboard could also be created, to track average data quality in different districts.

PRECISER MODELES ET METHODES

In predicting data validity, we will test the introduction of weighting of different predictors, based on their estimated distance from a given indicator. This will be made for each indicator, internally and externally, and testing different definitions of inter-indicator distance.

Finally, we will use similar methods as previously to impute true values of missing or false indicators. Validation of these will be made using standard validation techniques (k-Fold, etc...)

4.4 timeline

Figure XXX 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 DDDD.

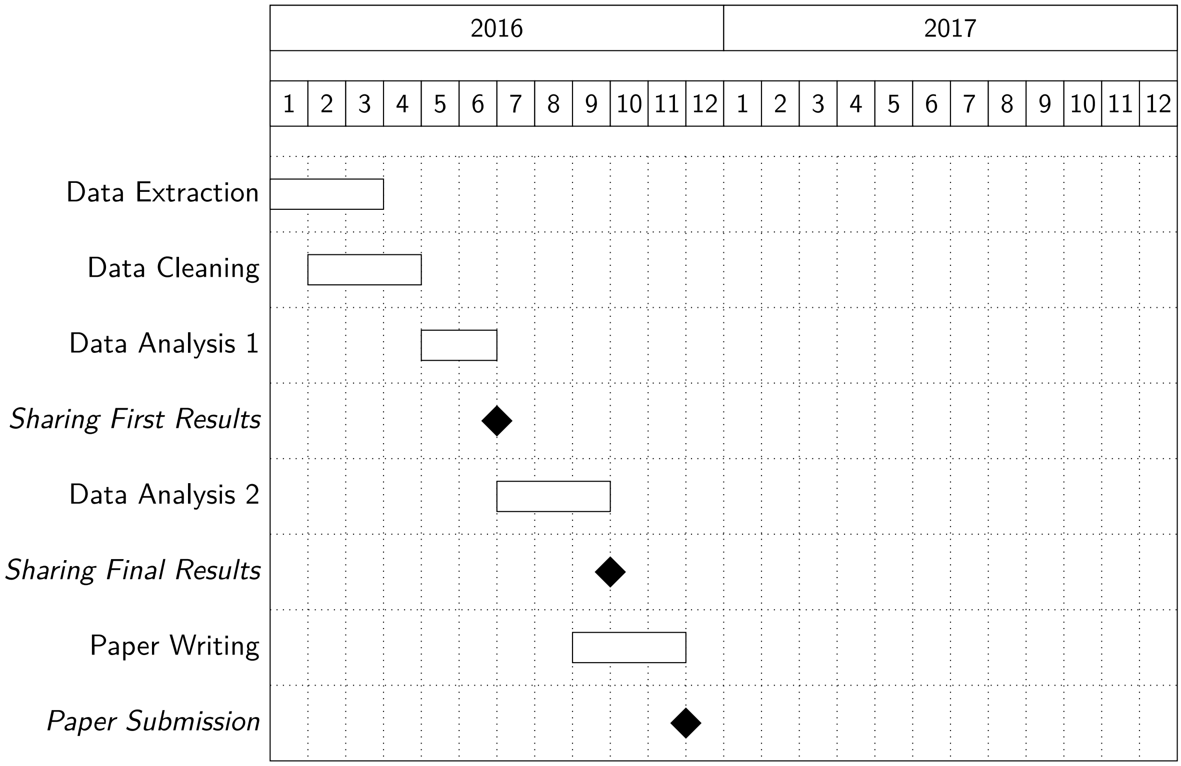

Figure 7: Gantt Chart for Paper 2

5 Data Analysis

HMIS data is most often very highly dimensional. A high number of indicators is collected on a regular basis in a host of facilities. Classic data analysis methods used in Public Health are not always suited to explore this type of data, especially due to the average low quality and completeness of the data. We aim at defining methods that can help make better use of this type of data for policy and decision making purposes. Mainly, we want to orient our work on reducing dimensionality of the HMIS data for analysts and policy makers.

A consequence of this is that policy makers usually concentrate on issues that are not dictated by what they see in their own data but concentrate on issues dictated at higher, natioanl or international levels. This is akin to the classic story of the economist looking for his keys under a lamppost while he lost them somewhere else. Analysis is only directed at variables for which there is external light.

Reducing dimensionality may come in different ways. Most important ones we have easy access to are :

- Reduce interesting indicators
- Reduce organizational unit to concentrate on
- Reduce data values to concentrate on

Building on literature of quality control in the industrial realm, we want to define observed norms and standards and simply identify units that differ from these norms and standards.

1. Indicator wise standardization 2. Spatial smoothing and validation 3. Temporal modeling and validation

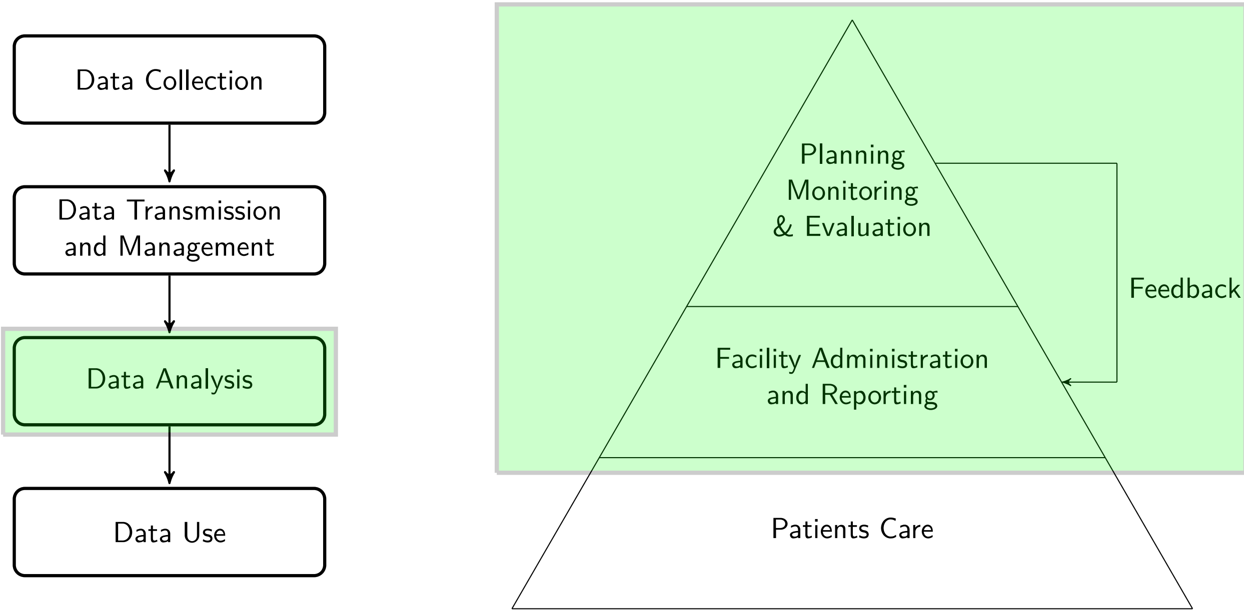

Figure 8: Objective three definition

5.1 Setting

5.2 Data

5.3 Analysis

5.4 Timeline

6 Data Use

Finally, recognizing even an excellent technical framework can not ensure the proper use of statistical information in a health system, we want to reflect on the culture of data use in developing countries health systems. To do so, we want to adopt a post colonial approach, and understand how a vision of statistics resulting from colonial rule and neo-liberal management results in a deficit of data using for health systems related decisions.

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.

Case study : analyse de textes M&E / projets de reforme de systemes hmis, et analyse de la vision des HMIS qu'ils proposent. quelle place pour la societe civile ? inversion des priorites.

6.1 Setting

6.2 Data

6.3 Analysis

6.4 Timeline

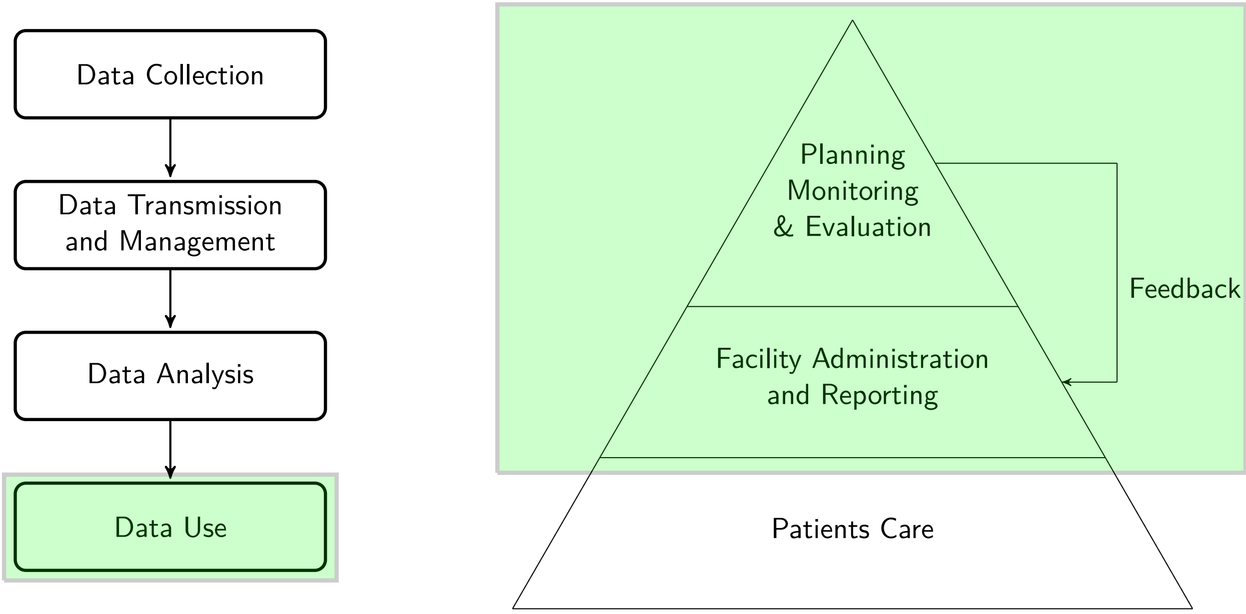

Figure 9: Objective four definition

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] 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.

[3] 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.