Abstract

With the increasing need to mitigate the negative impact of the global food, and financial crises on children, there is compelling need for fresh data on ECD (in child nutrition, health, early childhood education) to be used in planning for the response and decision making. We propose a systematic procedure, accompanied by a supporting software architecture called STAND, to collect and monitor global ECD related data in near-time, analyze that data with human analysts and software tools, and provide policy makers with an experimental environment within which they can learn about and simulate options they need to determine what interventions would provide the highest impact in combating early childhood disease.

I. Introduction

According to a United Nations 2007 report [1], approximately 9.7M young children died each year around the world – approximately 50% of this number die in Africa alone. Many of these deaths are attributed to disease, malnutrition, and other treatable factors. Despite carefully designed aid programs designed to collect data about such diseases, it is clear that collection of data at the grass roots level suffers from many major flaws that prevent the design and implementation of timely policy instruments. For instance:

- Data is painstakingly collected over a period of years and is often several years out of date (for instance, most existing data available in ECD today is from before 2007, more than two years old).
- Data is often collected at a coarse geospatial granularity, causing policy interventions to be implemented across a wide region, rather than being
pinpointed in the very specific areas where the interventions could be maximally helpful.

- Data analysis is carried out by hand by subject matter experts – despite their best efforts, this may cause many important inferences that can be drawn from the data to be missed, leading to potentially valuable options being invisible to policy planners.
- The ability to formulate policies that maximize the expected “return on investment” in terms of ECD outcomes and “experiment” with policies in a virtual laboratory is absent.
- The ability to efficiently and accurately assess the impact of policy interventions is compromised.

In this paper, we propose the STAND process which transforms the best resources available today in ECD research and development, together with novel new data collection methods and business models, to deliver real-time, actionable information to the policy maker, together with computational tools to support the exploration of different policies and their repercussions. STAND has the following attributes:

- The ability to collect data at a fine grained level – often on a clinic by clinic or hospital by hospital basis – providing a detailed level of visibility relatively uncommon today.
- Relatively little delay in data reporting by providing an innovative incentive structure to grass roots data collection efforts.
- Processes to flag “bogus” data.
- Processes to automatically apply data mining methods to understand what types of ECD interventions work, and the conditions/circumstances under which they work.
- Processes to automatically suggest policies to policy makers, together with tools to explore the strengths and weaknesses of those policies and generate a new policy that maximize the probability of achieving the desired ECD objectives.

Processes to understand the relationship between allocation of resources and desired ECD outcomes. More importantly, STAND can be implemented and executed almost completely using existing scientific and technological advances. However, the successful execution of STAND will require significant support and help from the national and local governments of the countries in question – we see STAND implementations varying from country to country, taking into account national and local concerns in those countries and always being executed in collaboration and under the aegis of the local national and regional leadership of the country.

II. The STEP Process
Figure 1 below shows a schematic view of the STAND process and framework. The rest of this section will explain this process.

II.A Data Collection Agents

The key idea in STAND is that data will flow from health care workers or ECD preschool teachers at health clinics, and early childhood centers via mobile telephones to the STAND database which will include a whole suite of tools for data deconfliction. *Any health care worker or ECD teacher* at a clinic or ECD center that we wish to monitor can register with the STAND system, and become part of the STAND ecosystem. After this registration, they will be able to access the STAND database and the STAND social network via either a mobile telephone or via an Internet connection. The job of data collection agents is to collect data and feed it into the STAND database via their mobile phones (or computers where available), as well as to correct erroneous information in the STAND database. An important question here is incentivizing this kind of reporting – we will discuss this in greater detail in Section II.B.

Figure 1. Schematic of the proposed STAND framework

The data collection agents will be charged with reporting “local” health care, ECD and nutrition data. For each health care facility, we want to create a *table*. The rows of the table will correspond to time periods (e.g. probably weeks). The columns will
correspond to different variables that we want to measure for that health care facility or clinic. Examples of variables might include:

- **Disease variables**
  - Number of patients admitted to the facility with malarial symptoms that week
  - Number of patients who died of malaria that week
  - Number of patients admitted to the facility with cholera symptoms that week
  - Number of patients who died of cholera that week
  - Similar numbers for other diseases

- **Patient demographic variables**
  - Number of children under 5 admitted to facility with malarial symptoms that week
  - Number of children under 5 who died of malaria that week
  - Number of children in the family salary bracket S (some salary ranges will be determined) admitted that week
  - Similar variables for other diseases and other child age groups

- **Medication related variables**
  - Number of prescriptions for medication M prescribed that week
  - Number of prescriptions for medication M prescribed to children under 5 that week
  - Similar variables for other medications and drugs

- **Clinic information variables**
  - Number of doctors present that week
  - Number of doctor-hours the doctors were present that week
  - Number of nurses present that week
  - Number of nurse-hours the nurses were present that week
  - Did prescription drug M run out that week? If yes, what was the cause (e.g. bad weather, no funds, etc). Same for different types of prescription drugs and other medical supplies.

- **Preschool information variables**
  - Number of preschool children attending preschool
  - Number preschool classes held at the preschool that week

Note that the above list is not exhaustive. It is just a small sample of the set of variables we wish to monitor on a weekly basis at each health clinic participating in this study.

Each data collection agent (health care worker, or preschool teacher) will have a fixed time window within which they may input this information.
The goal of our effort is to ensure that the final table associated with each clinic or preschool is completed every week (for the previous week) and is filled with accurate information.

II.B Incentives

Efforts to collect data on health care today are based on employing relatively small numbers of workers to collect the data, leading to long delays and large holes in the collected data. In addition, such efforts entail enormous cost.

In contrast, we propose an incentive based data collection method that rewards health care workers, or early childhood caregivers or preschool teachers based on a “points” system which we will describe thoroughly in Section II.C below. The “points” that health care workers accumulate can be redeemed for items such as:

- Phone cards with varying numbers of prepaid minutes on them,
- School textbooks
- Medicines
- Food
- Nutritional supplements
- Water purification tablets
- Education CDs or DVDs

This is not intended as an exhaustive list, but rather to show that a variety of products that are consistent with global development efforts can be provided as incentives. Other incentives may be provided on a localized basis, based on what is both attractive to the local population and desirable from a development standpoint. Each item has a certain number of points required to redeem it. Thus, a 30 minute phone card will require fewer points for redemption than a 60 minute phone card.

*It is our belief that with appropriate advertising and word of mouth efforts, the participation of health care workers in STAND will be high.*

II.C STAND Incentive Scoring Algorithm

The key aspect of our framework is the assignment of “points” to health care workers who interact with the STAND system. The assignment of points must:

- Reward those who entered accurate information
- Reward individuals based on how fast they were in entering that information (this also helps ensure the accuracy of the data)
- Reward those who correct inaccurate information
- Penalize those who enter inaccurate information
• Penalize those who corrected accurate information and tried to render it wrong.

The STAND scoring algorithm should obey a law of diminishing returns. The first health care worker to report a (correct) item should get a certain number of points (say 100), while the second to report the same item may get 50, the third may get 25, and so forth. Different pieces of information may have different value – for instance, scores for reporting attendance of health care workers may be lower than for reporting the number of patients admitted with malarial symptoms. After the tenth person to report the information, the points they get may be so small that they are virtually useless. The exact number of points can vary based on the importance of the piece of information.

In the same way, the first person to correct an inaccurate piece of information also gets a certain number of points, while the second person to do so gets less, and so forth. Thus, there is exponential decay in the rewards assigned to people who validate the system. The trick is to ensure that the decay is large enough so a relatively small number of people need to validate a piece of data for us to have confidence in the correctness, without having to pay out too much by way of incentives. Fortunately, there are many such scoring functions that obey these general criteria. While the mathematical details of these scoring functions are complex, we will not go into the details here. The system is therefore a self-policing system where the system users are able to get incentives for correcting erroneous information.

At the same time, however, it is important to have other checks and balances in place so that a group of individuals do not collectively try to defraud the system (though we strongly suspect it will be impossible to root out all fraud).

This is where we can use a class of “trusted users” who have to go through a vetting process to be considered “trusted”. For instance, health care inspectors and trusted NGO workers may play the role of trusted users. When they visit a health care facility, they may be able to provide “ground truth” against which the reports of ordinary health care workers can be verified. In the event of conflicts between reports from trusted users and ordinary users, we will assume that the trusted users’ reports are correct.

In short, the STAND scoring system has several major advantages over existing paradigms:

• There are concrete incentives for workers at health clinics to provide reliable data.
• The total value of incentives provided for a piece of data can be capped, allowing appropriate budgeting. Though incentives are provided for the provision of such data, the financial value of the incentives to the funding agency involved (such as a national Health Ministry, or an international donor such as the World Bank) can be capped by ensuring that the total value of points
awarded is “bounded above” by a fixed dollar value. For instance, if we decide that finding the number of malaria patients admitted to a particular clinic during a given should cost at most $5, then we can allocate $4 for the information and $1 for corrections. Of the $4 for those providing information, we might allocate $2 to the first person who provided the correct information, $1 to the second, $0.50 to the third and so forth. Of the $1 provided for correcting incorrect information, we might provide $0.50 to the first person correctly correcting the information, $0.25 to the second, and so forth.

- **The data is timely and it is a relatively short time before we can be reasonably sure of the accuracy.** This will save substantial time and effort and – we assume – many, many lives.
- **The data will be available to national governments and international donors through a single online database.**
- **The incentives provided to health care workers are compatible with global development goals.** By providing incentives via instruments such as textbooks, medicines, and other worthwhile, developmentally-sensitive means, we supplement existing development initiatives in other areas (e.g. in areas of education).

### II.D The STAND Database

All of the data (including conflicting reports, inappropriate attempts to correct data) provided by users will be stored in the STAND database. Thus, every single thing ever entered by a user will be in the STAND database even if the data is corrected later.

Our STAND database will include tools to present a *consistent and temporally accurate view* of the unclean data in STAND to the policy expert or analyst.

For example, even if there is considerable error in the STAND database because of lots of corrections, corrections to corrections, discrepancies in entries, etc., it is wise to maintain this record of corrections for future reference. However, policy analysts expect “clean” data to be provided to them. The way this is usually done in databases is to carve out a “cleansed” view of the database that resolves the inconsistencies in the data.

The STAND database will also contain information about user scores so that they can be used to check whether the users are eligible for incentives. We have worked in the past on methods to formulate policies about missing and inconsistent data [2,3,4]. We will use such inconsistency and incomplete information management tools so that users and policy analysts can also express policies to handle such data that are consistent with their own knowledge of the situation on the ground in a particular part of Africa. The figure below shows a screenshot of a system for managing inconsistent data (albeit in a
different domain) that can be adapted to store, track, and resolve inconsistencies in clinical data.

Figure 2. Screenshot of University of Maryland SIMS system to store, track, and manage inconsistent data using user and system specified policies

In addition to the input provided by users, the STAND database will have a set of other data in it as well. For instance, it will contain geospatial information about the location of clinics. It will include information on other geospatial aspects of the region in question such as:

- Location of settlements
- Location of dams, rivers, and lakes,
- Location of schools, universities and educational facilities
- Location of train stations

Figure 3 below shows a screenshot of a previous system called GIDSTAR for tracking diarrheal diseases in Kenya – the system was developed at the University of Maryland. This system includes the ability to visualize (very limited) clinical information (usually out of date), but also track location resources such as dams, rivers, lakes, etc. Built on top of GoogleMaps, GIDSTAR provides a first example of the kind of capabilities that STAND plans to revolutionize.
II.E The STAND Social Network

In addition to the incentives offered to data collectors by STAND, we also plan to include a social networking component that will allow STAND to bring together health care workers, or early childcare givers with similar concerns.

For instance, health care workers in Kenya dealing with diarrheal diseases may be able to learn from their compatriots in other parts of the world who might tell them how they filtered drinking water through items of clothing so as to reduce the level of dangerous organisms in the water [5].

Simple interventions such as these may play a significant role in reducing the occurrence of diseases from around the world.

The STAND social network we are designing will allow users in different countries or different provinces of the same country to share such experiences with one another so that they can leverage the lessons learned from each other.

II.F STAND Policy and Investment Tools
STAND will contain a component through which policy analysts and decision makers can reason about the data and try and learn – from the collected data about different clinics and ECD centers – when certain health care interventions or preschools worked and when they did not.

The goal of an ECD policy analyst is to decide how to maximize the probability that a given ECD related goal will be achieved. What set of actions can the ECD policy analyst recommend so that a given goal (say of reducing malaria deaths by 5%, or increasing access to preschools) can be achieved. In addition, the policy analyst has certain resources that may be available to achieve these goals.

The question is whether a computing system can help the policy maker in this task. We see that the STAND database system will have at least 50-100 variables. It is almost impossible for a single analyst to analyze a table with such a large set of columns and be confident that he or she did not miss any valuable inferences they could have drawn.

We are working on this problem, leveraging a suite of data mining tools [6,7] and algorithms previously developed at the University of Maryland to find a set of actions, given some resource constraints, that maximize the probability that a goal is achieved.

Such tools, when incorporated into the final implementation of STAND, will greatly assist policymakers make better informed policy decisions.

III. Conclusions

With the active support of the World Bank, the University of Maryland is developing a mechanism to implement STAND in collaboration with partners in certain African countries. Initial discussions have focused on a state-wide pilot effort in a Nigerian state, though this is still in the process of negotiation.

Many of the tools and technologies are already in place – however, adaptations are needed to support the specific requirements of the data we plan to collect. We are developing specific and localized incentivization schemes to incentivize local people on their specific needs.

STAND promises to be a valuable tool, when fully implemented and deployed, for real-time data collection, analysis, and policy formulation not only in early childhood related diseases, but also in other age groups, and in related nutritional efforts. In order to succeed, we need valuable buy-in and partnerships with local government, clinics, health care personnel, as well as local industries who might be able to contribute some incentives (e.g. phone cards, textbooks) for free. Major challenges therefore remain to be overcome.
Nonetheless, to be successful, STAND will need significant help from national governments to overcome legitimate local and national concerns. The set of incentives offered to health care workers will need to be sensitive to local concerns and may vary from country to country. The values of the incentives provided may also vary from country to country. Implementation in countries which have widely varying health care systems and dramatically different mobile phone regimes can also be a challenge. Such challenges will require significant local government assistance in order for concrete success to be achieved.

REFERENCES