1. Introduction

In 1967, war broke out between Biafra, a breakaway area of Nigeria, and the rest of the country. The ensuing civil war resulted in over a million deaths and a profound display of the world's inability to provide humanitarian aid. No standard approaches existed for measuring the nutritional status of children, oral rehydration therapy for treating diarrhoea was not yet available, and medical services were woefully inadequate and often inappropriate. After working in Biafra, the former CDC Director Bill Foege opined that humanitarian assistance was the 'disaster within the disaster'.

In 1994, under the eyes of the world's TV cameras, approximately 800,000 Rwandans fled into the border town of Goma, Zaire, over a three-day period. Within a month, approximately 7% of the population had died, primarily from cholera, dysentery and simple exhaustion with dehydration. Efforts to deliver water to these refugees were slow and ineffectual, and in the initial weeks clinical treatment practices were appalling. Two critiques of the crisis came to similar conclusions: the relief community could not marshal the organisation and resources to focus on the primary needs of water provision and appropriate medical treatment (Goma Epidemiology Group, 1995; ODI, 1996).

Following this crisis, efforts were undertaken to increase the efficacy of humanitarian aid. Primary among these has been the Sphere Project, whose mandate is to 'improve the quality of assistance provided to people affected by disasters, and to enhance the accountability of the humanitarian system in disaster response' (Sphere Project, 2004). The approach of the Sphere Project has been to lay out a set of minimum standards and guiding principles of humanitarian assistance in the hope that, if the quality of services is kept above some specific level, the likelihood of positive programme effects will increase. The ICRC has run a three-week intensive Health Emergencies in Large Populations (H.E.L.P.) course in several locations around the world each year. Columbia University, in partnership with the International Rescue Committee and World Education Inc., has run a series of similar

two-week courses. Likewise, the Sphere Project, UNICEF, the University of Wisconsin, and the group RedR, have all established courses which have each trained hundreds of field practitioners in various aspects of humanitarian assistance. Most of these efforts were designed to increase the sophistication or technical skills of humanitarian aid workers in the hopes of improving the level of relief programmes.

A separate set of efforts has resulted in attempts to standardise the way that programmes are evaluated. Most prominent among these is the SMART Project, which is funded by the Canadian aid agency CIDA and USAID. CIDA emergency programming has based funding of projects on the cost-effectiveness estimates of received proposals. Within the Food and Health-Related Programmes section of Humanitarian Assistance within CIDA, officers attempt to only fund projects which will result in a death averted for every $300 US spent. Starting in 2004, the USAID Office of Foreign Disaster Assistance (OFDA) will require all emergency health programmes to produce improvements in nutritional status or reductions in the affected population's mortality rate. These efforts by CIDA and OFDA require aid agencies to be able to estimate the nutritional status and mortality rates in those populations served. The SMART Project is an attempt to develop a standardised approach and a guiding template for measuring the nutritional status and crude mortality rate within a population.

All of the aforementioned efforts since Goma have shared a set of common assumptions. They all assume that humanitarian aid plays a positive role in the health of those served. They all assume that a set of skills and techniques exists which, if properly employed, would reduce the suffering of those affected by emergency conditions. To some extent, all of these efforts are, or can be, used to advocate for the existing humanitarian relief system.

This report attempts to develop a framework for assessing the influence of humanitarian aid on the health of those targeted as the beneficiaries of this aid. This

opens up the possibility that aid could have a negative impact on the health of the beneficiaries, and attempts to view aid as something that happens to people affected by natural disasters and war. It is hoped that, by taking an advocacy-free approach, this paper will facilitate relief workers' ability to serve their clients.

Box 1: Some definitions

Health impact: The health impact of a project is defined as the change in the primary health objective measure that occurred because of a project's implementation. Because variations in health measures occur over time, a change in the primary health objective measure concurrent with the project's implementation cannot be assumed to be the result of the project without additional supportive evidence.

Health consequences: The health consequences of a project are the sum of all of the health-related influences, intended and unintended, of that project.

Indicator: An indicator is a qualitative or a quantitative measure designated to reflect some process associated with a desired outcome. An indicator may be a measure of an outcome of interest, or it may be a measure of some activity which is believed to be linked to the outcome.

Process indicator: An indicator that measures a level of activity, knowledge, or material action, but that is not itself a measure of health status.

Rate: A rate is considered in this paper to be a number of events occurring per unit population per unit time.

Point estimate: A point estimate is the calculated rate or number from the data available, without regard for the precision of the estimate. Usually, point estimates are presented with a corresponding range of uncertainty.

2. A theoretical model for measuring project impact

Four general considerations are proposed for evaluating the impact of humanitarian health-promoting projects: the match of the societal level of the project and the level

of the observed impact; the strength of causation between the intervention and the change in health status; the validity of a baseline of comparison; and the validity of the indicator employed. Each of these considerations is described below, followed by a schematic model for evaluating the health impact of a humanitarian aid project.

2.1 Project level

In general, programmes designed to influence the health of a population in need focus on one of three societal levels: the individual or patient, the household, or the entire community.

Examples of programmes focused on the individual include: virtually all curative programmes, most counselling programmes, immunisations, targeted supplementary and therapeutic feeding, literacy programmes, and health education programmes designed to change the trainees' attitudes and behaviour. Most medical efforts are focused on individuals.

Examples of programmes focused on the household include: latrine construction and water provision, some fly and mosquito control efforts, health education designed to influence household dynamics or caretaking practices for children, most income generation and food self-sufficiency programmes, and housing and shelter programmes.

Examples of programmes that benefit communities include many public health initiatives, such as: educational activities aimed at preventing individuals from harming others (promoting latrine use, condom promotion among those who are HIV-positive, food vendor hygiene), activities that occur in the collective areas of the community, such as controlling mosquito breeding sites and drainage, outbreak prevention measures and health information systems.

While some programmes, such as immunisations and HIV or STD treatment and prevention, might be providing individual and community benefits, all specific measurable objectives should match the societal level where the desired outcome is expected. For an aid programme to claim to have produced a benefit, either the levels of programme implementation and evaluation need to be the same, or some clear logic needs to be presented to explain why the level of benefit is different from the level of the intervention.

2.2 Criteria of causation

Over time, a great deal of debate has arisen over what epidemiological evidence constitutes proof that some exposure or input produced an effect versus what evidence simply implied an association. This distinction can be more than academic as was seen over three decades of debate regarding the effects of smoking on health. Sir Austin Bradford Hill put forward criteria for attempting to ascribe causation between an exposure and a health outcome (www.dsru.org/publications/DRSU_143.html, accessed 14/1/2004). These criteria are so widely utilised that some introductory textbooks simply refer to them as the epidemiological criteria of causation (Mausner and Kramer, 1985). While Hill's main motive was to attribute the causation of a disease to exposure of a chemical or biological agent, the logic of these criteria also apply to assessing the positive effects of favourable exposure, such as health programmes.

Hill said that all of the following conditions can contribute to the argument that an exposure induces a health consequence:

1. The greater the strength of the association, the more likely it is that it is causative.
2. There is a dose-response relationship between the exposure and the health outcome.

3. Exposure consistently induces the health consequence in different settings at different times.
4. The exposure occurs before the health outcome.
5. There is a biologically plausible explanation for the exposure resulting in the health outcome.
6. There are no more plausible explanations for the health outcome.
7. Experimental results add particular weight to the evidence.

For virtually all cause and effect health relationships, some of these criteria will not apply. For example, not everyone exposed to a pathogen becomes ill and most people who smoke never develop lung cancer. Nonetheless, having several of these criteria met greatly strengthens the argument that some intervention influenced the health status of a programme. For a programme to be shown to have an impact, criteria 4, 5 and 6 should always be met. Of particular concern to programme evaluation is the issue of biological plausibility and the amount of service provided. Programmes need to be evaluated with particular regard to the likelihood that the level of inputs provided could plausibly result in the outcome reported. That is, the number of clinic visits, or the amount of food provided per child, need to be sufficient to induce the health effects observed.

Few experimental trials with randomised allocations of services have ever been conducted in emergency or refugee situations (Roberts et al., 2001; Tomashek et al., 2001). For medical and public practices that have been well studied in non-emergency situations (e.g. virtually all standard treatment protocols), there is little reason to question the logic of the intervention, and evaluations tend to focus on the relative success of implementation. For programmes that are designed specifically for the unique circumstances, the best evaluations tend to be observational in nature, that is comparing the haves with the have-nots. Some researchers have argued that observational studies are susceptible to bias, and can be relatively misleading (Loannidis et al., 2001).

These criteria of causation can be applied to populations and programmes as readily as Hill applied them to specific disease agents. For interventions with a vast literature documenting the attributable benefits (e.g. measles vaccination or Vitamin A supplements), the need to show 'proof' that the intervention produced a health benefit may be small, but for many other emergency interventions (e.g. HIV prevention through educational efforts or health benefits from shelter) there may be little or no evidence that such programmes produce any health benefits, making the importance of documenting any benefits great. Most humanitarian programmatic efforts fall somewhere in between, employing types of programmes that have produced documented benefits in other settings, have failed in some settings, and may or may not be producing benefits in the setting at hand. Illustrative examples of how Hill's criteria of causation can be applied to evaluate programmes will be presented in the next section.

2.3 Validity of the baseline of evaluation

It is exceedingly difficult to show that a humanitarian programme has had an impact without knowing the rate at which something was occurring before the intervention began and after the intervention was implemented. A comparable population that lacks the specific intervention of interest can sometimes be used as a proxy to determine what happens in the absence of the intervention. Likewise, when people are arriving in a new location or are returning home, it is often impossible to determine the baseline before their arrival. In those cases, established norms can be applied as an assumed baseline or as a threshold above or below which the indicator should not fall. Programmes lacking in a baseline rate or a comparison group, which keep mortality low or keep water and food provision high, can rightly claim success, but they can never quantify the impact of their efforts.

Baseline information should ideally be collected in the same way as the final evaluation data will be collected. Programmes should be evaluated on the level of

the intervention (individual, household or population); therefore, baseline data should be collected on the same level. This rarely occurs without rigorous programme planning. More commonly, programmes establish some baseline in the early stages of the project and the same process is repeated at the end of the project. If the project has a lag period between the start and the corresponding health consequences, and if the baseline information is not important for designing the project, this approach can be effective.

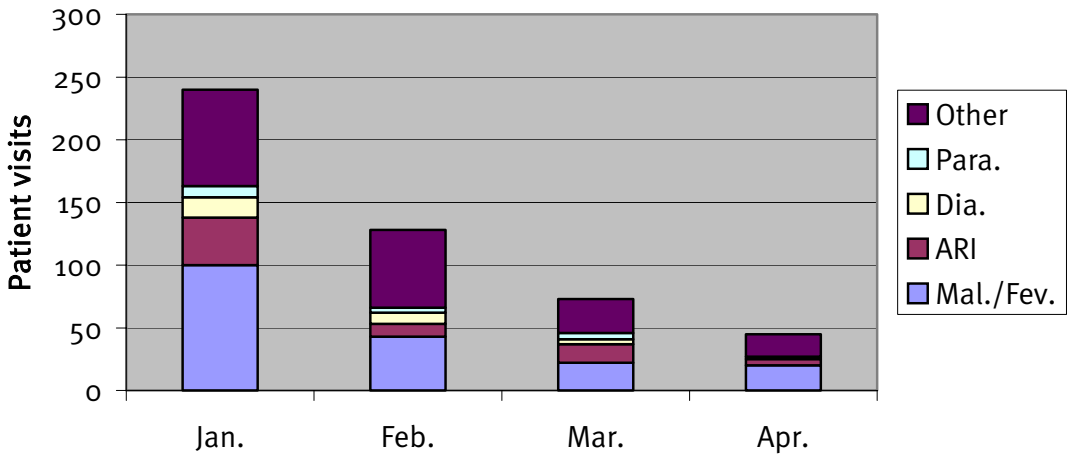
2.4 Validity of the indicator employed

Often, the indicator employed by an organisation is the actual objective of interest. For example, general comprehensive health programmes designed to reduce the mortality rate often use the crude mortality (CMR) or mortality among children less than five years of age (‹5MR) as the outcome of interest. This is optimal and, in the absence of some other temporal trend, a reduction in the indicator implies a favourable programme impact. Sometimes organisations will not know how to measure mortality in the population or find it too difficult to do so, and they will evaluate a programme designed to reduce mortality by monitoring a proxy, such as the number of patients treated. An increase in numbers of patients treated has no direct correlation with mortality. For mortality to be reduced while all other factors remained constant, the fraction of people at risk of death that go to the clinic would have to go up, or the efficacy of treatment once they reach the clinic would have to increase, or both.

Figure 1 below shows attendance at Matoumbou clinic in the Republic of Congo for early 2001. The clinic was virtually closed in late 2000 due to a lack of drug stocks, and during late December 2000 the ICRC delivered drugs to the clinic. The figure shows that attendance went from zero to over 200 patients per month, and then decreased steadily as the drug supply again ran out. This demonstrates how clinic attendance can be influenced by factors barely related to the health of the population. Many organisations use process indicators such as drug doses supplied,

clinics supported or staff trained to justify general health programmes designed to reduce mortality. These indicators are even further removed from the desired health effects.

Figure 1: Monthly attendance by diagnosis following December drug distribution, Matoumbou Clinic, Kinkala Dist., Congo



For an indicator to show that a health benefit is being induced by a programme, several factors need to coalesce: a) The indicator either needs to be a health outcome or be repeatedly and consistently shown to be associated with a health outcome (i.e. such as measles immunisations being associated with reduced risk of acquiring measles); b) the indicator needs to be from valid data collected on the same societal level as the expected health effects; c) the level of the indicator needs to improve and supportive data (such as a comparison group or complementary process indicators) needs to show that the benefit was most likely from the programme.

Few humanitarian programmes utilise indicators that fulfil criteria a) above. Rarely are humanitarian organisations encouraged or expected to fulfil criteria c). Table 1

characterises some commonly used indicators with regard to their strength of association to health and the ease with which they can be monitored.

Table 1: Characteristics of indicators commonly used to justify health programmes

Certainty of significance	Indicator	General ease of acquiring data to show health effects
Highest	Crude Mortality, <5 mortality Case fatality rate	Difficult in rural/diffuse settings, easier in camps
High	Nutritional status of children Disease rates Immunisation status of children	Easy on the clinic data level, difficult but more valid with population surveys
	Patient-specific mental health evaluations Safety of blood supply	Logistically easy, requires skill on part of evaluator
Moderate	Food-basket evaluations Water and sanitation availability	Easy in camps, more difficult in more diffuse populations
	Reduction in MMR through RH services Improved patient outcomes via referrals Impregnated bednets distributed	Very difficult to measure even though benefits are likely to be occurring
	Comprehensive, timely health info. system Good coordination between sectors Knowledge and attitudes about services available Population practices	Nearly impossible. These are difficult to measure, and all require a series of events to induce a health benefit
Low	People given seeds and tools, shelter or other materials Drainage, fly control activities or tasks Number of clinic visits Distance to facilities, health workers/capita	Easy to measure. Links to health are likely to be mediated via many steps.
	Trainings conducted, numbers trained Change in knowledge w/o documented change in behaviour Messages/curricula developed	Easy to measure. May produce no effects on health

Of note in this table is an apparent tendency for the indicators with the weakest links to health outcomes to be the most easily measured. They also are employed more often than the health outcomes listed at the top of Table 1.

It should be noted that indicators are utilised for many reasons: to monitor implementation of a programme, to determine when aspects of the programme are off-track, to monitor for graft or theft, and only sometimes to document the impact of a programme. For example, in the Sphere 2004 guidelines, there are 34 indicators related to shelter, but none have health data to support that this specific task will produce a health effect (op. cit.). Thus, the indicators put forward by the Sphere project are not designed to show project impact. They are social and technical prompts suggested by sector specialists with the expectation that, if all of these facets are remembered and employed, it is likely that the optimal project impact will occur.

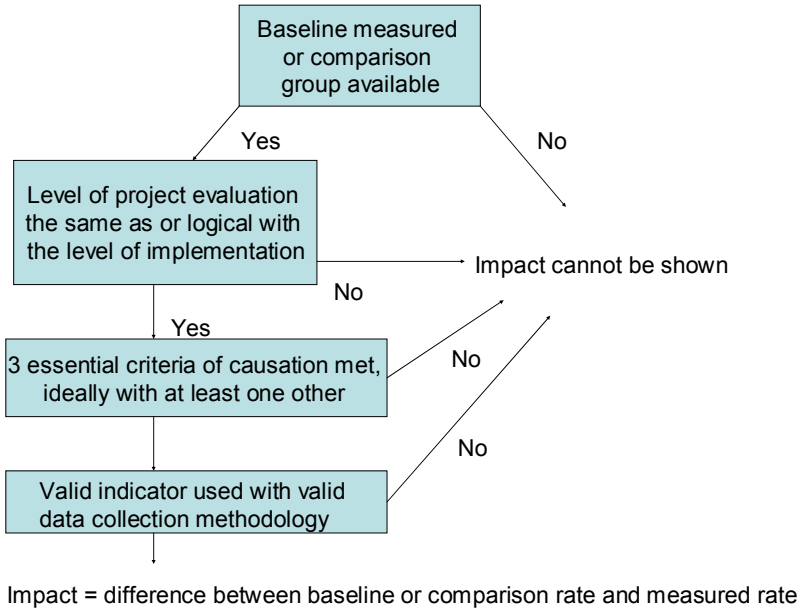
2.5 A theoretical model for evaluating project impact

Figure 2 employs the concepts presented above as a method to evaluate if a project was likely to have produced an impact on health status and to form a point estimate of the project impact.

Several explanatory comments need to be provided regarding this schematic model. First, saying that an impact cannot be shown is not the same as saying the project had no impact. It is likely that many projects produce health benefits even though they can never be shown to do so. Secondly, intentionally withholding services believed to be beneficial from a population in order to be able to document a benefit among those served may be unethical and beckons ethical review by an appropriate authority. Third, the criteria for causation may not apply for measures on the individual level with an unwavering medical link to an outcome (e.g. many complicated deliveries will result in foetal death without intervention, removing HIV+ blood from the blood supply will avert HIV transmission, stopping a severe

haemorrhage will be likely to avert a death). Fourth, if done properly, an evaluation should hold potential for showing a negative impact of the intervention. Next, the further up the 'certainty of significance' listed in Table 1 an indicator belongs, the more certain is the validity of the impact finding. Finally, the schematic in Figure 2 is not specific in time, except that the benefit must follow the programme's implementation. True project impact occurs over an infinite period, however meaningless that concept may be. The longer the period over which the impact evaluation is conducted, the better able it will be to assess the overall eventual impact.

Figure 2: Schematic for the evaluation of project impact



The two boxes below contain summaries of two projects evaluated using the schematic in Figure 2.

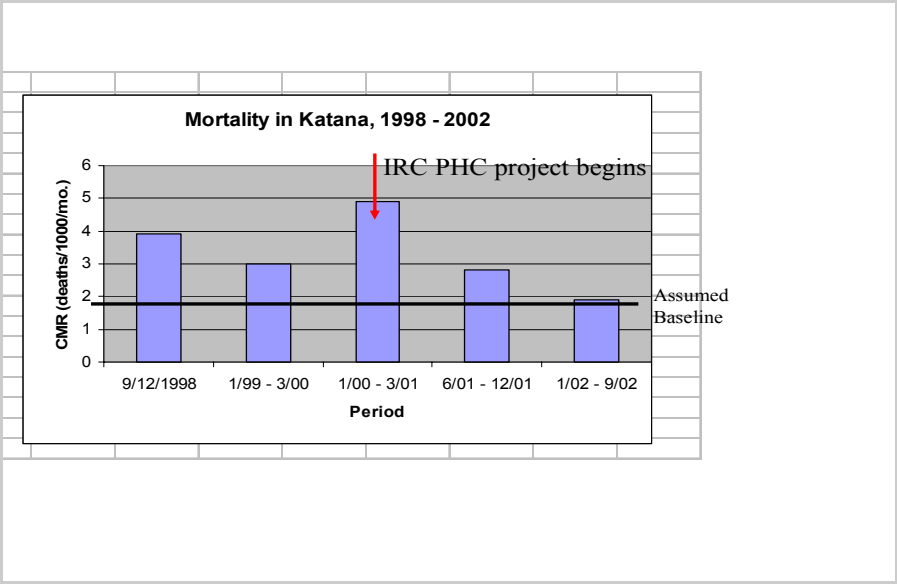
Box 2: International Rescue Committee health programme in Katana, DRC

Starting in December 2000, the International Rescue Committee (IRC) began a general health programme to support government services in Katana Health Zone, DRC. IRC conducted population-based mortality surveys in this area with 345,000 mostly rural residents. The programme consisted of the provision of drugs, supplies, training and medical oversight in the clinics, a water provision and hygiene education programme in villages with the highest rates of cholera in 2000, a measles immunisation and vitamin A provision campaign, and support to the local health committees, which included the donation of vouchers for the most indigent community members. The figure below shows the CMR over the period covered by five surveys conducted between 1999 and 2002. IRC claims to have reduced the excess CMR by 60% (from 4.9 to 2.8 deaths per 1,000/ month where the baseline is assumed to be 1.5) during the period from six to 12 months after implementation; and by 70% (from 2.8 to 1.9 deaths per 1,000 per month) over the period from 12 to 24 months after implementation. IRC reports that the total cost per death averted (including overhead and management salaries) was $227 the first year, and $132 per death averted over the two years of funding.

In support of these results in the figure below being a consequence of the health programme, IRC reported that:

- Attendance at the clinic rose by 147% between 1999 (~7,400 visits per month) and 2001 (~18,300 visits per month average).
- 70% of treatments were for malaria and diarrhoea, the main reported causes of death in the 1999 and 2000 surveys, and decreased as a cause of death in 2001 and 2002.
- CMR in the five eastern provinces of DRC was estimated by IRC to have increased slightly in 2001 compared to 2000. (continued)

- A survey in November of 2001 found that 60% of residents that had experienced fever in the preceding two weeks had sought treatment at a clinic.



Employing the schematic above: there is a baseline, the project goals, most of the implementation, and the evaluation were on the population level, the benefit occurred after implementation, the findings are biologically plausible (although one visit per resident per year seems low), alternative explanations for the reductions cannot be ruled out given the variance over time and the dramatic changes in violent conflict although IRC reports that the violence did not dramatically subside until 2002, the magnitude of the reduction and the fact that IRC's two other areas of health programmes had similar reductions (but somewhat less dramatic) implies significance and repeatability. Finally, the fact that the CMR was measured by an apparently valid survey method implies that IRC probably did reduce mortality in Katana.

Box 3: IRC laboratory activities in Kisangani, DRC

In the city of Kisangani, also in DRC, IRC undertook an effort to support laboratory activities in hospitals, specifically with regard to the testing of blood supplies for HIV which had not been done for the preceding three years. At the start of the programme, 7% of blood donors were HIV+ and 200 transfusions were occurring per month, almost exclusively among children experiencing extreme anaemia induced by malaria. At the end of the programme two years later, 17% of blood donors were HIV+ and 120 transfusions per month were occurring. IRC assumed that all transfusions of HIV+ blood could infect an individual, that all blood would be used (as it was always solicited for a specific patient) and that 93% of recipients at the start of the project and 83% at the end of the project were HIV-, and capable of being infected. This is conservative since children tend to be less HIV+ than blood donors. IRC reported a total programme cost of $476 per case of HIV averted at the start, and $327 at the end.

Applying the schematic in Figure 2: the initial HIV rate combined with the knowledge that no blood was tested provide a reasonable baseline for the measure of risk, the assessment was a cumulative set of experiences based on the patient, although which exact patient would have received the HIV+ blood discarded because of testing is unknown (that is, the intervention and evaluation were on the same level), the biological link between being transfused with HIV+ blood and becoming HIV+ is so strong that Hill's criteria of causation are not of concern, and the testing of blood destined for transfusion seems a valid measure of the blood-borne threat. Thus, in this case, with no surveys or data collection beyond the scope of programme activities, it is likely that IRC produced a benefit with this programme.

3. Available techniques and guidelines for measuring impacts of health interventions

A host of guidelines are now available for documenting health problems and health conditions during complex emergencies. WHO distributes a CD-ROM containing almost 200 guidelines and manuals for assisting workers in complex emergencies. Titles cover a range of topics from emergency obstetric care to *Model Guidelines for the International Provision of Controlled Medicines for Emergency Medical Care*. These references are also available on the web at: www.who.int/eha/disasters. Many of these manuals are explicitly on survey, interview and surveillance methodologies. A similar compendium, with fewer manuals but more descriptive documents, has been produced by the Payson Center at Tulane University, and can be found on the web at http://payson.tulane.edu. Many NGOs have produced their own guidelines and manuals. Médecins Sans Frontières (MSF) has perhaps the most complete collection, which includes the book, *Refugee Health: An Approach to Emergency Situations*, which provides guidelines for addressing and collecting information for many of the more commonly encountered health problems seen during emergencies. Thus, there is little need for this document to review or suggest specific guidelines.

Most techniques to document changes in health status can be divided by two different paradigms: quantitative versus qualitative and surveys versus surveillance.

Quantitative measures tend to have an external event or reality that can be detected by an observer. Thus, quantitative methods are often thought of as measures made by an outsider (a doctor, a teacher, a researcher) looking in. Qualitative methods tend to involve the investigation of phenomena which cannot be directly observed (such as trust or fear). Many techniques (such as community-based mapping) involve aspects of both quantitative and qualitative approaches. Most complete assessments also require both kinds of approaches (Sphere Project, 2004).

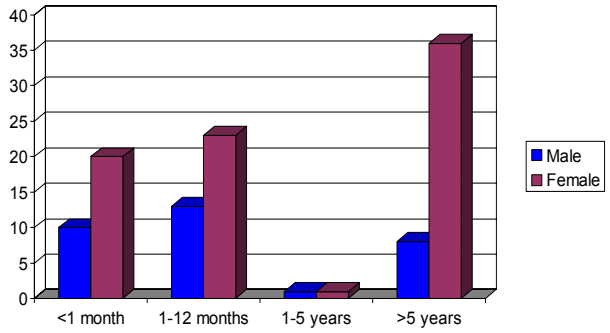
Surveys tend to be a systematic process by which investigators actively go out and collect information. Surveillance is the systematic collection of health information over time for decision-making. Thus, in theory, a process by which nutritional surveys are repeated on a regular basis could be a surveillance system. Like the qualitative/quantitative distinction, the two methods often overlap and are usually needed in consort to provide a timely and insightful picture of the health situation. However, in practice, programmes tend to be evaluated by one means or the other.

With surveillance systems, which are often clinic or hospital-based, information is collected continuously. Thus, there is the potential to identify problems or outbreaks early on, allowing for rapid response and prevention of cases. The CDC has developed a set of guidelines for evaluating surveillance systems which is widely utilised in emergency settings (CDC, 1988). Birth and death registrations, grave counting, monitoring of food baskets or water consumption can all be part of a health surveillance system. Surveillance systems have been used to detect and intervene in infectious disease outbreaks both during the acute phase of crises and over long periods of refugee settlement (Elias, Alexander and Sokly, 1990; Marfin, Moore and Collins 1994).

Several common elements exist among effective surveillance systems. First, there needs to be clear and universally applied case definitions for the health events of concern. Second, the system and the definitions need to be in step with the host government or the contextual and cultural setting in which the crisis exists. Third, surveillance systems should be simple. There is a tendency in many centrally-run Ministries of Health to include too many diseases and demographic categories in surveillance systems. This makes the system burdensome, slow and often ineffectual. On the other hand, information germane to the health crisis needs to be collected, and this can vary from setting to setting. For example, Figure 3 (which describes mortality in a Burmese refugee camp) shows that females, especially those above five years, were dying in greater numbers than boys, even though the

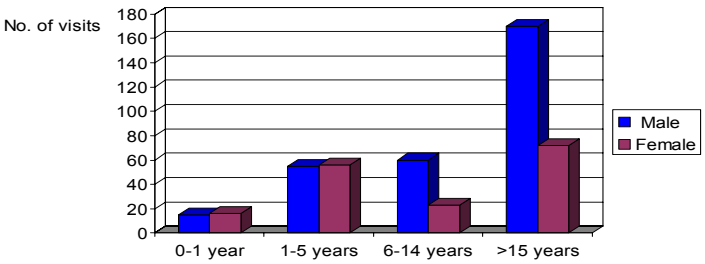
child population was assumed to be equally male and female. Figure 4 shows that males accounted for most of the clinic visits by children. This data led to the conclusion that male children were more readily being taken to clinics when ill, resulting in their lower mortality. This is an example where the breakdown of death and clinic attendance data by gender allowed for the identification and remediation of a social prejudice which resulted in an adverse health effect.

Figure 3: Mortality among Burmese refugees, Ghundum II Camp, Bangladesh, 6 May–26 June 1992



Source: Brent Burkholder UNHCR/CDC

Figure 4: Outpatient visits, Ghundum II Camp, 1–17 July 1992



Source: Brent Burkholder, UNHCR/CDC

Aid agencies often evaluate programmes by establishing a surveillance system at the beginning of a funding cycle, and contrasting the rate of health events at the beginning and the end. This is valid if either: a) all of the events of interest are captured by the surveillance network; or b) the data from within the system is representative of the health conditions of the entire population and remains consistently so over the course of the project. Neither of these conditions is commonly met for clinic-based surveillance systems in rural and urban areas, although both of these conditions are often met in well-defined settings like refugee camps.

Surveys can take several forms. Pre-treatment and post-treatment evaluations can be a type of survey if they are conducted systematically (such as every patient, or every tenth patient, or ten patients per day). But most surveys are an attempt to actively go out and interview a representative sample of the population. WHO and others have produced manuals specifically to guide health workers to conduct specific kinds of surveys, with nutritional anthropometry and EPI (Expanded Programme on Childhood Immunizations) coverage methodologies being among the most succinctly described (available at www.who.int/eha/disasters).

Most surveys of health outcomes are conducted by following a common set of procedures before heading to the field: identifying what is to be measured, developing the survey questionnaire and data collection methodologies, determining the sample size, defining the sample under study (often called the sampling frame), and selecting the sample. For the most commonly surveyed outcomes, these first three measures are predetermined by academics who have calculated appropriate sample sizes and suggest questions and measurement techniques to employ. In the case of EPI coverage surveys, most teams visit 30 clusters of households with seven children in each cluster. This will typically produce an estimate of the percent of the population that has been vaccinated for one or each specific antigen, with a precision of +/- 10% (Henderson and Sundarasen,

1982). Most nutritional survey methodologies suggest visiting 30 clusters of 30 children. Unfortunately, there is no standard method accepted for measuring the most global measure of health: the mortality rate. There are several reasons for this. First of all, the mortality rate varies widely between populations. Secondly, survey estimates of mortality usually require household reporting of deaths and the willingness of households to report deaths varies widely from culture to culture. Finally, the ease of determining how many people live in each household and how well the reported deaths can be linked to the population at risk of dying also varies widely.

Woodruff reported that NGOs and UN agencies regularly employ one of three methods during retrospective mortality surveys: the past-household census method, the current household census method, and the children-ever-born method (Woodruff, 2002). The past household census method involves asking the household who lived with the family at some point in the past and what has happened to each of these people now. This has the advantage that it allows for people to move in and out of the household and produces a person-time at risk estimate for the household denominator. The present household census method asks who lives in the household now and has anyone died during some period of recall. This method assumes that, on average, the number of people living in the interviewed houses has remained constant over the recall period, which may or may not be true. The current household census method has the advantage of being easy to conduct, requires less interviewee recall, and produces more easily recorded and managed data. No data exist to determine which of these two methods is more accurate. The third method, the children-ever-born method, involves asking mothers how many children they have ever delivered (or over some recall period), and what has happened to those children. This method has several disadvantages. It can only be used to estimate child death rates. Because the recall period is generally longer than employed in the other methods, the results are less likely to reflect the current situation and time. And finally, researchers in Zaire and Liberia have shown that

many, if not most, infant deaths are never reported by this approach (Becker et al., 1993; Taylor et al., 1993).

Most health measures of population-based health conditions require a survey for proper documentation. Unfortunately, most relief workers lack the skills needed to take a valid sample and to analyse the results of a survey. This is why many initiatives to improve the quality of relief programmes have emphasised the importance of training relief workers in survey methodologies.

4. Impact of international activities to increase the efficacy and accountability of humanitarian aid

Following the events in Goma in 1994, there has been a renewed emphasis on improving the effectiveness of humanitarian aid. Several publications documented the limited effects of the relief effort in Goma (Goma Epidemiology Group, 1995), or criticised the efforts outright (Saddique et al., 1995; Van Damme, 1995). A comprehensive review of the humanitarian relief effort during the Rwandan civil war and its aftermath, commissioned by the European Community, likewise pointed out many shortcomings of the relief effort (ODI, 1996).

Griekspoor and Sondorp have reviewed the international initiatives undertaken in recent years to improve the efficacy of humanitarian aid (Griekspoor and Sondorp, 2001). Initiatives they review include:

- The Red Cross Code of Conduct
- The Active Learning Network on Accountability and Performance (ALNAP)
- The Sphere Project
- The Humanitarian Accountability Project
- The Local Capacity for Peace Project.

Among these, Sphere has had the most influence on the day-to-day practices of relief programmes. Most major donors now expect their funded partners to adhere to the minimum guidelines set out by the guide. NGO workers have quickly become conversant with the guidelines and recommendations for their sector, and many managers now review proposals with the Sphere guidelines beside them as a reference. Columbia University has completed a review of the influence of the Sphere Project on the quality of humanitarian relief. While the Sphere were accepted, adhered to and applied widely in the relief community, there is no evidence that the initiative had improved the quality of humanitarian relief (Van Dyke and Waldman, 2004).

USAID and CIDA have funded the Standardized Monitoring and Assessment of Relief and Transition (SMART) initiative. This advocates for 'the use of shared, reliable, standardized benchmark indicators among donors and humanitarian organizations'. (www.smartindicators.org). The project is focused on developing a standard approach to monitoring nutritional status and mortality through surveys. While this initiative is too young to have yet influenced the quality of humanitarian relief worldwide, the project has managed to develop a consensus on an approach for surveying nutritional anthropometry, and is field testing a standardised method for measuring mortality. A file containing a manual and instructions is available for downloading at www.smartindicators.org.

For the past two years, parts of the Humanitarian Assistance Programme at the Canadian aid agency CIDA have attempted to require cost-effectiveness measures for non-food aid emergency programmes. While specific data are not available for public review, the vast majority of non-food aid emergency programmes produce and document some health effects (Bryan Luck, personal communication, 1/22/2004). As seen in the next section, this initiative to document disease rates or mortality appears to induce better reported results than are seen by some other donors.

5. Current practices

While many international agencies and NGOs have extraordinary skills at documenting the impacts of their health programmes, it is believed that there are two major monitoring issues frequently confronting the international community: the use of process indicators in lieu of health outcomes; and a lack of monitoring skills among some humanitarian workers.

5.1 Overuse of process indicators

To gain some insight into the present practices of humanitarian aid agencies, all final reports submitted in 2003 of health-related programmes funded by the US Department of State, Bureau of Population, Migration, and Refugees (BPRM) were reviewed. Three interim reports (which did not document health benefits) were not included in the review. Proposals that contained objectives of health-related activities (e.g. shelter provision, food transport) but that did not specifically say they would influence health status were excluded. The remaining 15 final reports were evaluated against five criteria:

- Was there a health-related objective?
- Was the baseline rate measured or a comparison group identified?
- Was the health-related outcome measured and reported?
- Was the social level of the evaluation appropriate given the intervention?
- Were there any major issues supporting or raising concerns about the reported outcome data?

Six of 15 reports did not attempt to measure or report any health-related rates or status. Proposals corresponding to five of these six reports only contained process indicators as the objectives, and thus the lack of documented health benefits was assured before the projects began. An additional three of the 15 reports contained health data-based objectives but did not present any health-status data, instead

reporting process indicators such as the numbers of clinics supported, consultations given, or tons of food distributed.

In the final analysis, only four of 15 final reports could show a benefit as outlined in Figure 2, and three others were likely to have produced a population-based benefit, although this was not documented. The results of this analysis confirm the general conclusion reached at the July 2002 SMART Monitoring and Evaluation Workshop, that while NGOs and agencies often want to monitor health outcomes, they usually monitor process indicators (www.smartindicators.org/events/HIU_workshopreport.htm). Problems with process indicators seen in the BPRM review include: the cited activity may be related to the health outcome but the significance of this effort depends on the activities being done well and in sufficient numbers (e.g. Eritrea and Sierra Leone, wanted to reduce mortality and reported numbers of clinic-based activities) or the health-related objective is only distantly related to the health outcome (e.g. Uganda, wanted to induce 'food self-sufficiency' but reported tons of food distributed). In some cases, the link between the process indicator and the outcome was simply implausible (e.g. Balkans, wanted to reduce dependency on aid of chronically 'Extremely Vulnerable Individuals' and reported doing this for some by distributing school books).

The programme with perhaps the most difficult to measure outcomes – a mental health programme in Guinea in 2002 – had the most rigorous documentation, which included pre-intervention and post-intervention patient evaluations and the use of non-patient controls. Representatives for the other three programmes which documented impacts felt that very little of the project budget (perhaps less than 2%) was spent on documenting the impacts.

The practice of using process indicators poses several threats to the quality of a health programme. As seen in Figure 1, process indicators such as clinic attendance rates may not be related to health conditions. As seen in Table 2, process indicators

sometimes simply have little relation to health status. Finally, issues of quality and proper timing can influence the impact on health of most services as measured by process indicators. Where possible, actual health outcomes are the preferred indicators of most agencies. Several agencies (ACF, SCF, MSF, WHO, UNICEF) have developed standard nutritional survey or mortality survey manuals in an attempt to make health status measures a cornerstone of programme activities. The ability of staff in the field to reliably measure these outcomes has been questioned.

5.2 Lack of monitoring skills

Over 20 NGOs provided general health services in the eastern DRC in 2000 and 2001 with funding from either OFDA or ECHO. According to OFDA, only two of those agencies could show health benefits associated with their programmes (Miriam Lutz, OFDA, Personal communication, 29/12002). This seemed plausible at the time given the violent and chaotic circumstances within which the NGOs operated. The short funding cycles and volatile nature of emergencies often prohibit a systematic and rigorous evaluation of either the impact or the monitoring of multiple agencies in the same setting.

Reviewers from CDC evaluated the monitoring of projects and the measurement of nutritional status and mortality in Somalia from the period 1991–93 (Boss, Toole and Yip, 1994). They developed a set of criteria for evaluating different kinds of information (surveillance and surveys) and systematically reviewed available reports. They found that the range of methodologies employed and outcomes measured were so variable, and of such poor quality, that they prevented widespread comparisons and much of the data was simply not credible.

Spiegel at al. from the Centers for Disease Control and Prevention (CDC) reviewed 125 nutritional surveys conducted in Ethiopia in 1999 and 2000 during a time of famine, but relative peace and stability (Spiegel et al., 2002). The surveys were carried out by 14 organisations with a wide range of survey expertise. Only 67 of the 126 surveys

attempted a sample that represented the population served. Only nine of those 67 surveys assigned clusters to the population in a manner that was proportional to the sub-units of the population, and only six of those possessed the minimum number of clusters (30) and children (900) suggested by most nutritional manuals. Most survey reports did not describe what sampling methods were employed, and few presented confidence intervals around the results. Sixteen reports were 'rapid assessments', without any attempt to take a representative sample. These unstructured surveys measured an average global malnutrition of 32% and severe malnutrition to be 5%. This contrasted with the 67 surveys that attempted to be population-based, which found 12% global and 1% severe malnutrition. Spiegel concluded that NGO workers were unprepared to conduct quantitative assessments of this kind, and that most of the surveys were of such poor quality as to be unhelpful in making sound relief policy decisions.

The measurement of anthropometry is relatively standardised compared to many other health outcomes, such as mortality and mental health. For example, a mortality survey in Kabare Health Zone in the Eastern DRC in 1999 was conducted simultaneously with an EPI coverage survey and mistakenly only included households with a child under five years of age. The resulting estimate of crude mortality (1.9/1,000/mo.) was far lower than a later repeat survey (2.7/1,000/mo.), and in fact, the initial survey missed most of the excess mortality. (Roberts, IRC, unpublished data, 1999). For some project objectives, such as the prevention of HIV transmission, there is not even a widely agreed outcome to be measured. The difficulty of assessing outcomes such as mortality is a principal reason for the use of process indicators in place of health outcomes.

The Sphere Project is an attempt to circumvent the fact that many programme activities cannot be reliably documented to improve the health of those affected by crises. Many indicators suggested by Sphere are not quantifiable, or are not health outcomes. Thus, without improved staff or a significant change in attitude among

donors, it is likely that humanitarian agencies will continue to rely heavily on process indicators and not be expected to prove that programmes influenced the health of the targeted beneficiaries.

6. Conclusions

Despite efforts to improve the quality, accountability and performance of humanitarian aid in the health sector and more broadly, knowledge of the impact of humanitarian aid programmes remains limited. Current humanitarian practice suggests that the health impact of programmes is often assumed, but rarely demonstrated. This can be partly explained by the difficult, often volatile working environment of humanitarian agencies. However, field epidemiology provides potentially useful tools for analysing the impact of aid programmes. These tools are seldom used for that purpose and consequently, humanitarian efforts rest on a limited evidence-basis. There are a number of elements that must be considered for more systematic analysis of the impact of humanitarian health programmes:

- Agencies tend to use performance or process indicators as proxy for impact, without the necessary evidence that the intervention is robustly linked with a health outcome. Monitoring the level of health services does not necessarily provide reliable information of the impact on populations. Further research on the links between particular interventions and health outcomes is required to build up this evidence-base.

- Efforts to document project impact should be woven into monitoring and surveillance activities, not only to reduce costs, but also as a tool to improve programme quality. The absence of systematic monitoring and surveillance in the humanitarian sector is a serious obstacle to assessing the impact of humanitarian aid. The assessment of impact should not be considered as a separate activity that takes place at the end of a project, but should ideally be

included throughout the project cycle, starting with the formulation of objectives.

- For health impacts to be more widely documented, there must be increased collaboration between donors and relief workers, which not only provides sufficient incentives and increased training, but also stimulates a culture that documents programme failures as well as successes as a learning opportunity. For example, new initiatives such as SMART are a potentially useful platform for analysing the global impact of humanitarian aid. However, some agencies fear that these mechanisms reinforce control over humanitarian agencies, instead of increasing the quality and performance of humanitarian aid.

- Finally, it is important that analysis of impact is not reduced to a narrow set of technical questions, regardless of the wider context in which aid is delivered. Whereas a poorly designed programme is unlikely to yield good results, a well-performing programme does not guarantee impact if the wider political context is detrimental to its implementation. Here, the notions of humanitarian space and the respect of basic humanitarian principles are probably as important as the use of appropriate tools and techniques for ensuring the greater possible impact.

References

Becker, S. R. & et al 1993, 'Infant and Child Mortality in Two Counties of Liberia: Results of a Survey in 1988 and trends since 1984', *International Journal of Epidemiology*, vol. 22, no. Supplement 1, p. S56-S63.

Boss, L. P., Toole, M., & Yip, R. 1994, 'Assessments of mortality, morbidity, and nutritional status in Somalia during the 1991-1992 famine. Recommendations for standardization of methods', *Journal of the American Medical Association*, vol. 272, no. 5, pp. 371-376.

CDC 1988, 'Guidelines for Evaluating Surveillance Systems', *Morbidity and Mortality Weekly Report*, vol. 37, no. S-5.

Elias, C. J., Alexander, B. H., & Sokly, T. 1990, 'Infectious disease control in a long-term refugee camp: the role of epi. surveillance and investigation', *American Journal of Public Health*, vol. 80, no. 7, pp. 824-828.

Goma Epidemiology Group 1995, 'Public health impact of Rwandan refugee crisis: what happened in Goma, Zaire, in July 1994', *The Lancet*, vol. 345, pp. 339-344.

Griekspoor, A. & Sondorp, E. 2001, 'Enhancing the Quality of Humanitarian Assistance: Taking Stock and Future Initiatives', *Prehosp Disast Med*.

Henderson, R. H. & Sundarasen, A. 1982, 'Cluster sampling to assess immunization coverage: a review of experience with a simplified sampling method', *Bulletin of the World Health Organisation*, vol. 60, no. 2, pp. 253-260.

Loannidis, J. & et al 2001, 'Any casualties in the clash of randomized and observational evidence? No recent comparisons have studied selected questions, but we do need more data.', *British Medical Journal*, vol. 322, pp. 879-880.

Marfin, A. A., Moore, J., & Collins, C. 1994, 'Infectious disease surveillance during emergency relief to Bhutanese refugees in Nepal', *Journal of the American Medical Association* no. 272, pp. 377-381.

Mausner, J. S. & Kramer, S. 1985, *Epidemiology: An Introductory Text*, J.B Saunders and Co. edn, Philadelphia.

ODI 1996, *The Joint Evaluation of Emergency Assistance to Rwanda: Study III Principal Findings and Recommendations*, ODI, London, 16.

Roberts, L., Chartier, Y., Chartier, O., & et al 2001, 'Keeping clean water clean in a Malawi refugee camp: a randomized intervention trial', *Bulletin of the World Health Organisation*, vol. 79, pp. 280-287.

Saddique, A., Samal, A., Isham, M., & et al 1995, 'Why treatment centers failed to prevent cholera deaths among Rwandan refugees in Goma, Zaire.', *The Lancet*, vol. 345, pp. 359-361.

Spiegel, P., Salama, P., Mahoney, S., & et al. Methodology Case-Study: Ethiopia. Presentation at SMART Monitoring and Evaluation Workshop, Washington DC. July 24, 2002.

Taylor, W. & et al 1993, 'Mortality and Use of Health Services Surveys in Rural Zaire', *International Journal of Epidemiology*, vol. 22, no. Supplement 1, p. S15-S19.

The Sphere Project 2004, *Humanitarian Charter and Minimum Standards in Disaster Response*, 2nd edition edn, Geneva.

Tomashek, KM., Woodruff, B. A., Gotway, CA., & et al 2001, 'Randomized intervention study comparing several regimens for the treatment of moderate anemia among refugee children in Kigoma Region, Tanzania', *American Journal of Tropical Medicine and Hygiene*, vol. 64, no. 3-4, pp. 164-171.

Van Damme, W. 1995, 'Do refugees belong in camps? Experiences from Goma and Guinea', *The Lancet*, vol. 346, pp. 360-362.

Van Dyke, M. & Waldman, R. 2004, *The Sphere Project Evaluation Report*.

Woodruff, B. A. Review of Survey Methodology presented at SMART Workshop, July 23, 2002. 2002.