Published in final edited form as:

Stat Med. 2014 July 20; 33(16): 2746–2757. doi:10.1002/sim.6145.

Extending cluster Lot Quality Assurance Sampling designs for surveillance programs

Lauren Hund^{a,*} and Marcello Pagano^b

^aDepartment of Family and Community Medicine, University of New Mexico 2400 Tucker Avenue Northeast Albuquerque, NM 87131

^bDepartment of Biostatistics, Harvard School of Public Health 655 Huntington Avenue, Boston, Ma, 02115

Abstract

Lot quality assurance sampling (LQAS) has a long history of applications in industrial quality control. LQAS is frequently used for rapid surveillance in global health settings, with areas classified as poor or acceptable performance based on the binary classification of an indicator. Historically, LQAS surveys have relied on simple random samples from the population; however, implementing two-stage cluster designs for surveillance sampling is often more cost-effective than simple random sampling. By applying survey sampling results to the binary classification procedure, we develop a simple and flexible non-parametric procedure to incorporate clustering effects into the LQAS sample design to appropriately inflate the sample size, accommodating finite numbers of clusters in the population when relevant. We use this framework to then discuss principled selection of survey design parameters in longitudinal surveillance programs. We apply this framework to design surveys to detect rises in malnutrition prevalence in nutrition surveillance programs in Kenya and South Sudan, accounting for clustering within villages. By combining historical information with data from previous surveys, we design surveys to detect spikes in the childhood malnutrition rate.

Keywords

Acceptance sampling; longitudinal surveillance; LQAS; nutrition; survey design

1 Introduction

Lot Quality Assurance Sampling (LQAS), also referred to as sampling for attributes or acceptance sampling, has a long history of applications in industrial quality control [1]. In its simplest form, LQAS is a binary classification tool for identifying poorly performing groups, with performance measured using a binary indicator. In the past several decades, LQAS has become a popular tool in global health care surveys, including applications in quality control for health records and community health workers; assessing HIV/AIDS control programs; and monitoring maternal and child health [2]. While LQAS surveys are

*Corresponding author, lhund@salud.unm.edu.

now a widely accepted tool in global health, these surveys have been criticized for lack of transparent interpretation and poor classification properties [3, 4, 5, 6]. Simple and transparent methods for designing and interpreting LQAS surveys are useful for program managers conducting surveillance, who often do not have access to a statistician for survey design.

Two-stage cluster sampling [7] for estimating the coverage of a health indicator is a commonly used survey design in public health surveillance and monitoring and evaluation (e.g. [8, 9, 10, 11]). These two-stage cluster surveys are typically designed to estimate a population parameter with certain amount of precision. In practice, the choice between using LQAS or a conventional two-stage cluster survey depends on the goals of the surveillance program. LQAS surveys are designed for binary classification of coverage, rather than estimation of coverage. Because the goal is classification, rather than estimation, LQAS surveys typically require sampling fewer individuals and are less expensive to implement than two-stage cluster surveys. However, LQAS surveys have historically relied on simple random sampling from the target population. When the target population is spread across multiple distant villages, using cluster sampling in LQAS designs is more cost-effective than using simple random sampling [12, 13].

Recently, several methods have been proposed for preserving classification risks in LQAS designs when clustering is present [12, 14, 15, 16, 17], with applications in nutrition surveillance [12, 13, 18], childhood vaccinations [14, 15, 16], and quality assurance among community health workers [17]. The methods in [14, 15, 16, 17] rely on parametric modelling assumptions and subsequently only apply in certain situations (namely, when the number of clusters in the population is large and the model is correct). Alternative approaches do not adjust for clustering, instead relying on small within-cluster sample sizes and low clustering, which then result in minimal variance inflation [9, 12]; relying on low clustering may not be accurate [10] and relying on small within-sample cluster sizes *a priori* may not be cost-effective [4].

Uniting cluster LQAS designs with standard survey sampling theory alleviates the need for context-specific designs and parametric assumptions; and allows for the straightforward application of survey sampling results for LQAS designs. In this paper, we develop this united approach for LQAS survey sampling and apply the framework to the design of surveillance programs for monitoring acute malnutrition. These programs inform decisions regarding when and how to allocate resources in emergency settings. However, no consensus has been reached regarding how to design LQAS surveys in the nutrition setting [4, 5, 12, 19].

One challenging design question for nutrition LQAS surveys (and other LQAS surveys as well) is how to select the LQAS survey design parameters to balance precision and feasibility [3, 4, 5, 6, 19, 20, 21]. We address the how to incorporate uncertainty in the underlying health status of the population into the survey design; existing LQAS surveys do not allow for such uncertainty. This method of design parameter selection will generally be relevant during the development of a longitudinal surveillance programs. When longitudinal surveillance programs are relatively new, we may have limited historical information about

the coverage of the health indicator in the population. Baseline data is available after the first round of surveillance. As more rounds of surveillance are conducted, more data becomes available. We begin to learn more about the underlying health of the target population, and can account for varying levels of uncertainty about coverage over time in a longitudinal program.

In this paper, we propose a simple and flexible non-parametric procedure to incorporate clustering effects into the LQAS sample design to appropriately inflate the sample size, accommodating finite numbers of clusters in the population when relevant. We use this framework to then develop principled selection of survey design parameters in longitudinal surveillance programs, accounting for uncertainty about the health status of the population. We provide an R package for implementing the proposed methods, to facilitate the use of the methods in practice. In Section 2, we review LQAS surveys for binary classification of a health indicator, and introduce the surveillance problem of monitoring the prevalence of malnutrition in children. In Section 3, we propose a simple adjustment to sample size calculations for LQAS surveys to incorporate intraclass correlation. In Section 4, we introduce a framework for incorporating uncertainty in target coverage rates into LQAS longitudinal surveillance programs. Throughout the paper, we illustrate our methods using LQAS survey data collected from various sites in Kenya and South Sudan over 3–4 month intervals. The proposed methods provide principled guidelines for designing an LQAS-type classification procedure to detect changes in malnutrition prevalence over time using cluster sampling, while preserving the simplicity of LQAS methods.

2 Lot quality assurance sampling

LQAS is a binary classification procedure based on binomial probabilities; the design is defined by four design parameters p_l , p_u , α , and β [22]. Define X as the number of successes out of n independent Bernoulli trials. We model $X \sim \text{Binomial}(n, p)$, where p is the probability of success. Classification of a lot as acceptable or unacceptable is based on a classification rule d . In public health applications, lots are referred to as supervision areas (SAs), which are the regions from which the sample of size n is collected. For instance, in nutrition surveillance programs, field workers collect measurements on n children in an SA, and count X , the number of children with malnutrition. Malnutrition prevalence in a region would be classified as high if $X > d$ (the ‘lot’ is rejected); and acceptable if $X \leq d$ (the ‘lot’ is accepted).

The goal is to select a sample size n and decision rule d such the risk of misclassifying an SA as acceptable or unacceptable is small. The following two equations control the risk profile of the classification procedure and consequently determine the survey design:

$$\begin{aligned} P(X > d | p = p_l) &\leq \beta \\ P(X \leq d | p = p_u) &\leq \alpha \end{aligned} \quad (1)$$

We want to minimize the risk of classifying as ‘high p ’ when the true probability of success p is not high, *i.e.* when $p < p_l$; and minimize the risk of classifying as ‘low p ’ when p is not low, *i.e.* $p > p_u$. To define allowable risk, we specify α , the probability of classifying as

acceptable when $p = p_u$; and β , the probability of classifying as unacceptable when $p = p_l$. Supervision areas with true prevalence between p_l and p_u are said to lie within the ‘grey region;’ classification risks are not restricted within the grey region. The width of the grey region ($p_u - p_l$) and choice of α and β influence the precision of the classification procedure.

2.1 Data Example: Nutrition Surveillance using LQAS

The binary indicator global acute malnutrition (GAM) is a measure of child malnutrition, defined as a weight-to-height Z-score (WHZ) less than -2 and/or bipedal edema; alternatively, GAM is defined as a middle-upper arm circumference (MUAC) less than 125 mm and/or presence of an edema. We define malnutrition using the WHZ-score definition of GAM throughout this paper, following the literature for LQAS nutrition surveys. The World Health Organization (WHO) classifies malnutrition prevalence as critical if the prevalence of GAM in a population is greater than 15% [23]. Malnutrition prevalence can be assessed using sample surveys in children age 6–59 months [18]. Due to the infeasibility of implementing simple random sampling in emergency settings, children are often sampled within villages (using cluster sampling).

Two-stage cluster sampling surveys (30 clusters of 30 children) have historically been used to assess malnutrition prevalence [24, 25]. To decrease costs, Deitchler *et al.* [18] propose using LQAS with cluster sampling to classify malnutrition status as acceptable or unacceptable [13, 18]. The authors select $\alpha = 0.1$, $\beta = 0.2$ as risks and select 3 couplets for p_l and p_u : 5–10%, 10–15%, and 15–20% [18]. Based on these design features, the authors choose 33×6 ($n = 198$) and 67×3 ($n = 201$) cluster sampling designs as guide designs for monitoring malnutrition prevalence, with respective decision rules 13, 23, and 33 corresponding to the three couplets [18]. (Note that these designs have classification risks that are near $\alpha = 0.1$ and $\beta = 0.2$, but often do not strictly meet these cut-offs).

Existing nutrition cluster LQAS designs do not explicitly adjust for within-cluster (intracluster) correlation, but require sampling a large number of clusters to minimize the impact of this correlation [12]. The cost-effectiveness and feasibility of survey designs that require visiting over 60 different clusters is questionable [4]; the target population may not even contain 60 different clusters (villages). Clustering effects may be low enough in the nutrition setting that assuming simple random sampling is valid [13]; other studies suggest clustering of nutrition status exists at the household-or village-level [10, 26].

We obtained data from 24 different 33×6 recurring LQAS nutrition surveys collected by Action Against Hunger. Trained enumerators conducted surveys at one site in South Sudan and seven sites in Kenya over approximately 3–4 month intervals during 2008 and 2009, with three rounds of surveillance per site. Surveys were conducted using the procedures described in [13]. We performed all analyses using R version 2.15.2 [27]. For each of the 24 surveys, we classify malnutrition status as critical or acceptable using the three different 33×6 LQAS decision rules. Most of the surveys include more than $33 \times 6 = 198$ children (most had ~ 220 children in the survey); we calculate the frequency of critical classifications from a sample of 198 using the hypergeometric distribution. These findings represent the frequency that the LQAS survey would have produced a critical classification if the survey protocol had been followed exactly and 198 children were sampled.

Table 1 contains summary statistics and classification decisions for the 24 surveys by location. Using the 5–10% couplet, critical classification occurs in approximately 22 surveys; using the 10–15% couplet, in approximately 21 surveys; and using the 15–20% couplet, in approximately 16 surveys. Trends in malnutrition prevalence by location across time are shown in Web Figure 1 in the Web-based Supporting Materials. GAM prevalence has seasonal and regional variations, but is chronically high in most regions, with the highest prevalence in the Mandera region.

3 Incorporating clustering

When individuals within clusters are similar, cluster sampling inflates the classification risks of the LQAS survey, with the amount of inflation depending on the sample size per cluster and on the amount of clustering [7]. Following current practice in LQAS designs, we assume that the within-cluster sample size is constant, prioritizing simplicity of implementation. To use Equation 1 for classifications, we must sample using a self-weighting design, assuming either that clusters are of approximately equal sizes or that sampling occurs via probability proportional to size sampling.

We propose a simple algorithm for determining the survey sample size and decision rule in the presence of clustering. Define K as the number of clusters in the population; k as the number of clusters sampled; M as the population size within each cluster (assumed equal across clusters); and m as the sample size per cluster. The intraclass correlation coefficient ρ quantifies the magnitude of the between cluster variability in prevalence relative to the within cluster variability. The intraclass correlation is $\rho = 1 - (M/\{M - 1\}) * (SSW/SSTO)$, where SSW is the within-cluster sum of squares and $SSTO$ is the total sum of squares [7].

Following [28], we exploit the relationship between the intraclass correlation (ρ) and the design effect to adjust the sample size. The design effect (D_{eff}) is the ratio of the variance of X (the number of children with malnutrition) under the cluster sampling survey design to the variance of X under simple random sampling, fixing the sample size n [7]. The design effect equals $D_{eff} = 1 + (m - 1)\rho$ when the following assumptions hold: the number of individuals sampled per cluster, m , is constant; the population size within each cluster, M , is large and equal across clusters; and the number of clusters in the population, K , is large [29].

Removing the assumption that K is large, the design effect is $D_{eff} = 1 + (fm - 1)\rho$, where $f = (1 - k/K)$ is the first stage finite population correction (see Appendix 1 in the Web-based Supporting Materials for a derivation). More general forms of the design effect can be derived by further relaxing the assumptions, such as varying population size by cluster; we address the situation in which K is small relative to the number of sampled clusters k , as this scenario is likely common in LQAS surveys.

The effective sample size (ESS) is the sample size required under simple random sampling to achieve the same variance as in the cluster survey design, and $ESS = n/D_{eff}$ [28]. To calculate the survey sample size, we first determine the necessary sample size under simple random sampling, and then back-calculate to determine the sample size for a cluster survey. We iterate through choices of m and k until we find a decision rule that meets the

classification risks α and β for a given choice of p_l and p_u . The algorithm proceeds as follows:

1. For a given m and k , calculate the effective sample size, $n^* = n/D_{eff}$, where $n = mk$. Round n^* to the nearest integer.
2. Search for a decision rule d^* using the standard binomial LQAS model with sample size n^* , specifying α , β , p_l , and p_u .
3. To obtain the final decision rule, calculate d^*D_{eff} and round this quantity to the nearest integer, to obtain the decision rule d .
4. The final sample size is n , consisting of k clusters of size m , and the decision rule is d .

We have written an R package `lqasdesign`, which contains a set of functions for designing cluster LQAS surveys using this algorithm. The package is available for download at https://dl.dropboxusercontent.com/u/4659680/lqasdesign_1.0.zip. Details about these functions are available in the help files and in the package vignette, accessible by typing `vignette("manual", package="lqasdesign")` in R.

For a given k , we are not guaranteed that a sample size and decision rule exist that meet the risk thresholds, α and β . To find the minimum number of clusters that must be sampled, we consider the properties of the effective sample size as m (the within-cluster sample size) increases. As m gets large, the effective sample size converges to $n_{max}^* = k / (\{1 - k/K\}\rho)$ (see Appendix 1 in the Web-based Supporting Materials for derivation). We must select a large enough k (namely, $k = n_{max}^* / [\rho(1 + n_{max}^*/K)]$) such that a decision rule exists for n_{max}^* that meets the design specifications under simple random sampling using the binomial model.

Conceptually, we consider X to be an overdispersed binomial random variable, where the overdispersion results from aggregating heterogeneous clusters; and $\text{var}(X) = D_{eff}^* np(1 - p)$. This method for sample size calculation extends the existing cluster LQAS methods (e.g. [14, 15, 16, 17]) by allowing for a finite number of clusters in the population and not requiring specification of a parametric distribution for X (e.g. in [17], X is modelled as a convolution of beta-binomial random variables). These existing methods also rely on the assumptions of constant m and large M . As in [17], our method requires an *a priori* estimate of the intraclass correlation ρ . In an ongoing longitudinal surveillance program, we can update estimates of ρ over time.

LQAS is intended as a rapid and feasible surveillance tool, and minimizing cost (time) is pivotal. In survey sampling, the optimal cluster size k for a fixed cost is calculated by minimizing the variance of an estimator subject to the cost constraint. In LQAS surveys, the optimal cluster size k is found by minimizing cost subject to a variance constraint (namely, the effective sample size). A common cost constraint considered in survey sampling is a linear cost function, $C = c_1k + c_2km$, where C is the total cost, c_1 is the cost of adding a cluster, and c_2 is the cost of adding an additional individual. In Appendix 1 of the Web-based Supporting Materials, we illustrate that the standard result for the optimal within-

cluster sample size for a linear cost function [7] holds for LQAS cluster surveys, namely that the optimal within cluster sample size is either $m = \sqrt{c_1(1 - \rho)/(c_2\rho)}$ or $m = C/(c_1 + c_2K)$ (sampling all clusters) if $k \leq K$.

Other non-linear cost functions may also be reasonable. For instance, suppose we know that approximately 20 individuals can be sampled per day. Additionally, suppose the cost of adding an additional cluster is a function of the number of clusters visited per day. As a simple example, we could consider a cost function where $C = c_1k + c_2km$ if $20 \bmod m = 0$ (where mod is the modulo operation) and $C = (a + c_1)k + c_2km$ otherwise, where a is a cost inflation factor for having to visit more than one cluster per day. Aside from the linear cost function setting, optimal cluster sizes will generally not have a closed form and can be selected by iterating through total cost for different values of m and k under a fixed effective sample size. In the R package, we have implemented an option for the user to specify a cost function (rather than specifying m or k).

3.1 Simulation Study

Asymptotic results for the design effect adjustment are proved in [28]. For this adjustment to perform well in populations with a finite number of clusters, we require that $P(X < d) \approx P(X^* < d^*)$, where X^* is a binomial random variable with sample size n^* and probability p . X and X^* have the same mean and variance by definition. Due to the discrete nature of the binomial distribution and the rounding in steps (1) and (3) of the algorithm, the distributions of X and X^* are not exactly equal. A limitation to the clustering adjustment is the potential for this approximation to inflate the classification risks, α and β . We verify through simulation that the approximation $P(X < d) \approx P(X^* < d^*)$ is reasonable in smaller sample sizes.

To assess the performance of the design effect correction with a finite number of clusters (small K), we evaluate the properties of the design with $p_l = 0.10$, $p_u = 0.15$, $\alpha = 0.1$, and $\beta = 0.2$ by simulation. To construct the simulated populations, we generate $K = 20$ cluster level prevalences from a Beta distribution, with correlation coefficient $\rho = 0.05$ and mean prevalences equal to p_l and p_u . We then scale the 20 prevalence estimates so that their means are exactly equal to p_l or p_u , and the intraclass correlation is exactly 0.05. We also generate samples from a beta-binomial distribution with means p_l and p_u and $\rho = 0.05$, to compare the finite cluster sample size to the sample size when $K = \infty$. More details of the simulation are shown in Appendix 2 in the Web-based Supporting Materials.

Results of the simulation are shown in Figures 1 and 2. Assuming the number of clusters in the population is infinite will result in a much larger sample size than necessary, when the true number of clusters in the population is small (Figure 1). In the simulated data, the estimated classification risks (calculated analytically from the sample size algorithm) are close to the empirically calculated risks determined through simulation (Figure 2). Differences between the estimated and empirical risks are driven by rounding error and the inexactitude of the above approximation, resulting in an increase in one of α or β and a decrease in the other.

3.2 Data Application - Impact of clustering on the survey design

Using the 33×6 LQAS data from Kenya and South Sudan, we assess the impact of clustering on the survey design. Code for reproducing these analyses using the R package are provided in Web Appendix 4. We assume that there is a large (infinite) number of clusters at each survey site. This assumption is most likely violated, but we do not have information about how the clusters were enumerated. For each survey, we estimate the intraclass correlation ρ using maximum likelihood estimation, assuming the data are generated from a beta-binomial distribution. Ridout *et al.* [30] summarize estimators of ρ for binary data. If the data do not follow a beta-binomial distribution, our estimate of ρ may be poor.

The average intraclass correlation over all of the surveys is $\hat{\rho} = 0.037$, and ranges from 0.00 to 0.13 across the surveys. The intraclass correlation estimate $\hat{\rho} = 0.037$ is somewhat low, but non-negligible. For instance, in a 33×6 survey using the 15–20% couplet, if $\rho = 0$, the α and β risks are 0.14 and 0.22, respectively; if $\rho = 0.037$, these risks are now 0.16 and 0.24. These risks will be higher for the surveys with $\rho > 0.037$. Therefore, this 33×6 survey design does not meet the pre-specified α - and β - levels of 0.1 and 0.2.

We no longer need to stay within the confines of the 67×3 or 33×6 designs to ensure that clustering does not substantially inflate our classification risks; sampling very few children per cluster may not be cost-effective, depending on the distance between clusters. We can apply the design effect sample size algorithm to determine the number of clusters to sample to meet these risk levels. We again consider the 15–20% couplet with $\alpha = .1$ and $\beta = .2$. Fixing $m = 6$, when $\rho = 0$, we would need to sample 44 clusters to meet the predetermined risk levels. With $\rho = 0.037$, we would need to sample 57 clusters. For $\rho = 0.1$, we would need to sample 66 clusters. If the number of clusters in the population was finite, say $K = 50$, then we would sample 46 clusters. Failing to account for the finite number of clusters in the population would result in a design with $k > K$.

We now consider choosing the optimal number of clusters for a fixed cost function. First, we assume a linear cost function, where the cost of adding an additional cluster is 5 times as expensive as adding an individual. When $\rho = 0.037$, the optimal number of clusters is $k = 31$, sampling $m = 12$ children per cluster. In remote settings, the cost of the survey will decrease if travel expenses and additional logistics are substantially reduced. For instance, if a field team can collect data on $m = 20$ children per day, it may be more cost-efficient to sample 20 children per cluster to minimize travel costs. We now assume the same linear cost function, except if $m \bmod 20 = 0$, the cost of adding an additional cluster is now 7 times as expensive as adding an individual (rather than 5 times). In this case, the optimal cluster size is $k = 23$, sampling $m = 20$ children per cluster.

4 Designing surveillance tools to detect changes over time

Selecting the survey design parameters p_l and p_u for a longitudinal surveillance program often requires substantial contextual knowledge of the health indicator in the target population [19]. Consider a nutrition surveillance LQAS design. In practice, the rate of malnutrition varies across populations, due to differences in body-type, different definitions of GAM, as well as differing underlying malnutrition status of the populations. Pastoralist

populations tend to be tall and thin. Relatively healthy children in pastoralist population are more likely to be classified as malnourished than those in other populations, when GAM is defined using WHZ scores [31].

A goal of established surveillance programs is detecting changes in coverage of a health indicator that signal a crisis. For example, if we observe a substantial spike in the malnutrition prevalence in a population, we need to quickly intervene. Detecting changes in coverage is inherently tied to statistical process control [32] (as opposed to quality control [33]). In process control, an indicator is tracked over time, and an alarm is sounded when the indicator goes ‘out of control.’ Knowledge of when a process is ‘in-control’ is necessary for understanding when the process is ‘out of control.’ Hence, knowledge of the underlying coverage when the population is ‘in-control’ is essential to understanding when the population becomes ‘out of control’.

Presumably, there exists some target coverage level such that coverage is acceptable below(above) this threshold and unacceptable above(below) this threshold. If the target coverage, denoted p^* , is known, we can use this threshold to inform selection of p_l and p_u . For example, when monitoring malnutrition prevalence, it might be reasonable to specify $p_l = p^*$ and an upper threshold $p_u = p^* + U$, where U is a known meaningful deviation in prevalence from p^* . This target threshold p^* may change over the course of a surveillance program, as programmatic interventions impact the health of the population.

We now assume that data are collected across multiple time-points t , with X_t denoting the number of individuals with the health indicator out of n_t total sampled individuals. When p^* is unknown, ideally we can estimate this value from existing historical data and incorporate uncertainty in p^* into the survey design at time t . For instance, we use an empirical Bayesian framework to estimate the posterior distribution of p^* and incorporate this uncertainty into the survey design by modelling X_t using the beta-binomial distribution. First, we specify a prior distribution for p^* , namely $p^* \sim \text{Beta}(a_0, b_0)$. If program managers have some information about plausible ranges for p^* , this information can be incorporated into the prior by adjusting a_0 and b_0 . If no historical information is available, we recommend selecting $a_0, b_0 < 1$. Poor specification of a_0 and b_0 can result in incorrect classifications when little data is available, as the prior will dominate the data.

We can incorporate any available data from periods when the population was healthy (*i.e.* periods when it was reasonable to assume that coverage was acceptable); discerning whether the population was healthy should be made by subject-matter experts with contextual knowledge. Without this contextual knowledge, it may be best to conduct an estimation survey (rather than using LQAS), to learn more about the population.

For now, we assume existing historical data were collected when the population was relatively healthy and were collected using simple random sampling. We denote X_H as the number of individuals with the health indicator out of a sample of size n_H . Then, the posterior distribution for p^* is $f(\cdot) = \text{Beta}(a_1, b_1)$, where $a_1 = X_H + a_0$, $b_1 = n_H - X_H + b_0$.

Incorporating data from multiple surveys is trivial; for $j = 1; \dots, J$ surveys, we collect X_j^H and n_j^H , and the posterior parameters are $a_1 = \sum_j X_j^H + a_0$, $b_1 = \sum_j (n_j^H - X_j^H) + b_0$.

Rather than specifying p_l and p_u , we specify deviations from p^* , L and U . We choose upper and lower classification thresholds L , U based on the following criteria: if $p > U$, a notable rise in prevalence occurred; if $p < L$, no notable changes occurred. Often, $L = 0$ is a logical choice. Values between L , U are in the ‘grey region,’ and we do not restrict classification risks within the ‘grey region’ in the survey design. We determine how likely it is that p_t was drawn from $f(\cdot)$ shifted by L and $f(\cdot)$ shifted by U (Figure 3). The misclassification risks are constrained by the equations $Pr(X_t \leq d_l | p = U, f(\cdot)) \leq \alpha$ and $Pr(X_t > d_l | p = L, f(\cdot)) \leq \beta$ (details of calculation are provided in Appendix 3 in the Web-based Supporting Materials). We iterate through choices of N_t and d_t to find the design that minimizes N_t and meets the specified classification risk thresholds. The beta-binomial model for X_t accommodates uncertainty in p^* ; larger sample sizes are generally required for classification due to this additional uncertainty. However, the design protects against misspecification of p^* .

Because the clustering adjustment in Section 3 is rooted in standard survey sampling theory, it is easy to extend this design to accommodate cluster sampling. To use a cluster LQAS design, we can directly apply the algorithm in Section 3, but use the beta-binomial model rather than the binomial model in Step (2). Inflating the sample size to account for a cluster sampling design in the historical data is also simple in this setting. First, we convert the historical data X^H, N^H to the ‘effective sample size’ scale, calculating $X^{H*} = X^H / D_{eff}^H$ and $N^{H*} = N^H / D_{eff}^H$, where D_{eff}^H is the design effect for the historical survey data. Then, we calculate the posterior distribution for p^* using X^{H*} and N^{H*} .

4.1 Data application - Examining changes over time

Using the preliminary nutrition surveillance data, we cannot accurately estimate p^* from the data. Instead, we use baseline data to design the subsequent survey to detect a rise in prevalence from baseline, assuming prevalence at baseline was not high. In practice, the program manager should examine the data from the baseline survey and use contextual knowledge to determine if the survey results suggest malnutrition prevalence was high or low at baseline. If the data from the baseline survey are not consistent with the contextual knowledge and prevalence appears high at baseline, then it is best to assume that prevalence was high at baseline and proceed with the monitoring and evaluation program accordingly.

To design the survey, we first estimate p from the baseline survey and use this estimate to calculate the design effect for the next survey. We select $L = 0$, $U = 0.1$, $\alpha = 0.1$, and $\beta = 0.2$; that is, we aim to detect 10 percentage point increases in prevalence at least 90% of the time. We will accept detecting an increase in prevalence when none has occurred at most 20% of the time. Calculations for the sample size and decision rule were conducted in the `lqasdesign` R package, and we provide a sample calculation in Appendix 4 in the Web-based Supporting Materials.

The nutrition data contain approximately 220 children per survey. As in Section 2.1, for each survey design, we randomly sample from the observed data and calculate the probability that we would declare a change in prevalence using the hypergeometric distribution. For some of the surveys, we have not collected a large enough sample to reach a conclusion as to whether or not the prevalence changed over time. The resulting survey designs are presented in Table 2. In the Garissa Riverine and Mandera Pastoral populations, we detect a rise in prevalence from baseline. Using the highest 15–20% couplet, we would classify the Mandera Pastoral population as high prevalence at all three time points. Using the design to detect changes, we obtain more information - not only is the prevalence high in this region, prevalence has increased from the baseline survey.

5 Discussion

In this article, we propose a framework for designing cluster LQAS surveys; by using a survey sampling framework, we improve the accuracy and flexibility of the existing cluster LQAS designs. Using a simple design effect adjustment, the pre-specified classification risks are preserved when cluster sampling is used. Further, if the number of clusters in the population is finite, we adjust the sample size downward. The sample size algorithm is inexact, relying on an approximation for finite sample sizes, and we caution against using the design effect adjustments in surveys with very small sample sizes (< 30) or with very rare events. In these cases, if the number of clusters is large, an exact method (such as [17]) is preferable.

The proposed design framework will add flexibility to LQAS designs, without compromising the simplicity of the design procedure and decision making process. We would encourage program managers to investigate the data beyond simple binary classifications, if such an analysis is feasible, examining spatial and temporal trends to help guide interventions. To adjust for clustering, program managers must provide an estimate of the intraclass correlation coefficient. If surveys repeat over time, tracking the intraclass correlation coefficient over time is feasible, assuming the sampling units and target population are fixed over time. Further, program managers should keep track of previous survey results for estimating the target coverage, when relevant.

By using historical information and data from previous surveys, we detected spikes in malnutrition when the target coverage is unknown. Specifically, we were able to detect rises in malnutrition prevalence in nutrition surveillance programs in Kenya and South Sudan, where the underlying rates of malnutrition are known to vary. Knowing whether the prevalence is changing over time can provide additional valuable information, supplementing current LQAS designs that only address static classifications of prevalences as critical or acceptable. The populations surveyed include a large fraction of pastoralist populations. Consequently, GAM prevalences, and perhaps intraclass correlations, were higher than those that will be observed in many settings. Monitoring populations with lower prevalences requires reevaluation of the survey parameters which will likely impact the survey sample sizes. The proposed tools are flexible enough to adapt to different settings. Further, malnutrition prevalence estimates differ substantially when MUAC, rather than WHZ score, is used to construct the GAM indicator. For instance, the median GAM

prevalence using MUAC scores in our survey populations is 7.4%, compared to 20.3% using WHZ scores. In future work, incorporating information from both MUAC and WHZ scores to improve classification is of interest.

Using this framework, we propose a cohesive surveillance tool for monitoring the prevalence of acute malnutrition over time. We anticipate that this surveillance tool will be more generally useful in surveillance programs operating over time. For example, another potential area of application is routine immunization surveillance, in which LQAS surveys are conducted across time using cluster-sampling to track changes in immunization coverage [14].

While our approach provides a flexible framework for clustering adjustments, the approach has several limitations. First, the design requires an estimate of the intraclass correlation at the time of the survey. This correlation may be hard to estimate in relatively new surveillance programs; misspecification of this parameter will impact the sample size for the survey (affecting both cost and precision). Another limitation to our approach is the reliance on a large-sample asymptotic results for the clustering adjustment. While the simulation study provides evidence that the clustering adjustment performs well in finite samples, this issue should be explored further in future work.

In conclusion, we find that using standard survey sampling results to inflate LQAS sample sizes obviates the need for parametric modeling assumptions and creates a flexible and feasible design framework for field-use. Cluster LQAS designs can be used to detect rises in the prevalence of malnutrition. Future work should explore implementing this design framework in the field and assessing real-world implementation, as in [13].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The data sets used for the analysis were collected by Action Against Hunger; and were provided by David Doledec. The authors would like to thank Megan Deitchler for her comments on the manuscript. NDSEG Graduate Fellowship and National Institutes of Health grant (T32 AI007358) to L.H.; R01 AI097015 to M.P.

References

1. Dodge H, Romig H. A method of sampling inspection. The Bell System Technical Journal. 1929; 8(398):613–631.
2. Robertson S, Valadez J. Global review of health care surveys using lot quality assurance sampling (LQAS), 1984–2004. Social Science & Medicine. 2006; 63(6):1648–1660. [PubMed: 16764978]
3. Sandiford P. Lot quality assurance sampling for monitoring immunization programmes: cost-efficient or quick and dirty? Health policy and planning. 1993; 8(3):217–223. [PubMed: 10128376]
4. Bilukha O. Old and new cluster designs in emergency field surveys: in search of a one-fits-all solution. Emerging Themes in Epidemiology. 2008; 5(7)
5. Bilukha O, Blanton C. Interpreting results of cluster surveys in emergency settings: is the LQAS test the best option. Emerging Themes in Epidemiology. 2008; 5(25)
6. Rhoda D, Fernandez S, Fitch D, Lemeshow S. LQAS: user beware. International Journal of Epidemiology. 2010; 39(1):60–68. [PubMed: 20139433]

7. Lohr, SL. Sampling: design and analysis. Boston: Cengage Learning; 2010.
8. Henderson RH, Sundaresan T. Cluster sampling to assess immunization coverage: a review of experience with a simplified sampling method. *Bulletin of the World Health Organization*. 1982; 60(2):253–260. [PubMed: 6980735]
9. Stroh, G.; Birmingham, M. Protocol for assessing neonatal tetanus mortality in the community using a combination of cluster and lot quality assurance sampling. Geneva: Department of Vaccines and Biologicals, World Health Organization; 2002. URL: <http://apps.who.int/iris/handle/10665/67193> [accessed 2/7/14]
10. Fenn B, Morris S, Frost C. Do childhood growth indicators in developing countries cluster? Implications for intervention strategies. *Public health nutrition*. 2004; 7(07):829–834. [PubMed: 15482606]
11. Aliaga, A.; Ren, R. Optimal sample sizes for two-stage cluster sampling in Demographic and Health Surveys: DHS working paper. Calverton: 2006. URL: <http://www.measuredhs.com/pubs/pdf/WP30/WP30.pdf> [accessed 2/7/14]
12. Olives C, Pagano M, Deitchler M, Hedt B, Egge K, Valadez J. Cluster designs to assess the prevalence of acute malnutrition by lot quality assurance sampling: a validation study by computer simulation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 2009; 172(2): 495–510.
13. Deitchler M, Deconinck H, Bergeron G. Precision, time, and cost: a comparison of three sampling designs in an emergency setting. *Emerging Themes in Epidemiology*. 2008; 5(6)
14. Pezzoli L, Pineda S, Halkyer P, Crespo G, Andrews N, Ronveaux O. Cluster-sample surveys and lot quality assurance sampling to evaluate yellow fever immunisation coverage following a national campaign, Bolivia, 2007. *Tropical Medicine & International Health*. 2009; 14(3):355–361. [PubMed: 19187516]
15. Pezzoli L, Andrews N, Ronveaux O. Clustered lot quality assurance sampling to assess immunisation coverage: increasing rapidity and maintaining precision. *Tropical Medicine & International Health*. 2010; 15(5):540–546. [PubMed: 20214765]
16. Greenland K, Rondy M, Chevez A, Sadozai N, Gasasira A, Abanida E, Pate M, Ronveaux O, Okayasu H, Pedalino B, et al. Clustered lot quality assurance sampling: a pragmatic tool for timely assessment of vaccination coverage. *Tropical Medicine & International Health*. 2011; 16(7):863–868. [PubMed: 21481106]
17. Hedt-Gauthier B, Mitsunaga T, Hund L, Olives C, Pagano M. An approach for incorporating clustering effects into the design of lot quality assurance sampling. *Emerging Themes in Epidemiology*. 2013; 10(11)
18. Deitchler M, Valadez J, Egge K, Fernandez S, Hennigan M. A field test of three LQAS designs to assess the prevalence of acute malnutrition. *International journal of epidemiology*. 2007; 36(4): 858–864. [PubMed: 17517808]
19. Olives C, Pagano M. Choosing a design to fit the situation: how to improve specificity and positive predictive values using bayesian lot quality assurance sampling. *International Journal of Epidemiology*. 2013; 43(1):346–355. [PubMed: 23378151]
20. Olives C, Pagano M. Bayes-LQAS: classifying the prevalence of global acute malnutrition. *Emerging Themes in Epidemiology*. 2010; 7(3)
21. Hund L. New tools for evaluating LQAS survey designs. *Emerging themes in epidemiology*. 2014 Accepted.
22. Pagano M, Valadez J. Commentary: Understanding practical lot quality assurance sampling. *International journal of epidemiology*. 2010; 39(1):69–71. [PubMed: 20139434]
23. World Health Organization. The management of nutrition in major emergencies. Geneva: World Health Organization; 2000.
24. Binkin N, Sullivan K, Staehling N, Nieburg P. Rapid nutrition surveys: How many clusters are enough? *Disasters*. 1992; 16(2):97–103. [PubMed: 20958740]
25. Spiegel P. Who should be undertaking population-based surveys in humanitarian emergencies? *Emerging Themes in Epidemiology*. 2007; 4(12)
26. Katz J. Sample-size implications for population-based cluster surveys of nutritional status. *The American journal of clinical nutrition*. 1995; 61(1):155–160. [PubMed: 7825528]

27. R Core Team. R: A Language and Environment for Statistical Computing. Austria: R Foundation for Statistical Computing, Vienna; 2012. URL <http://www.R-project.org/>, ISBN 3-900051-07-0
28. Rao J, Scott A. A simple method for the analysis of clustered binary data. *Biometrics*. 1992; 48(2): 577–585. [PubMed: 1637980]
29. Kerry S, Martin Bland J. Unequal cluster sizes for trials in English and Welsh general practice: implications for sample size calculations. *Statistics in medicine*. 2001; 20(3):377–390. [PubMed: 11180308]
30. Ridout M, Demétrio C, Firth D. Estimating intraclass correlation for binary data. *Biometrics*. 1999; 55(1):137–148. [PubMed: 11318148]
31. Myatt M, Duffield A, Seal A, Pasteur F. The effect of body shape on weight-for-height and mid-upper arm circumference based case definitions of acute malnutrition in Ethiopian children. *Annals of Human Biology*. 2009; 36(1):5–20. [PubMed: 19085192]
32. Colosimo, B.; Del Castillo, E. Bayesian process monitoring, control and optimization. Boca Raton: Chapman & Hall/CRC; 2007.
33. Sower V, Motwani J, Savoie M. Are acceptance sampling and SPC complementary or incompatible? *Quality Progress*. 1993; 26 85–85.

Figure 1.

Figure 2.

Figure 3.

Table 1

Summary statistics by location. The columns contain the survey date; the total number of children sampled (n); the estimated GAM prevalence, with a corresponding exact binomial 95% confidence interval; and the frequency of binary classification of GAM prevalence as unacceptable/critical using the 33×6 LQAS survey designs.

Location	Survey Date	n	GAM Prevalence (CI)	LQAS critical classification		
				5–10%	10–15%	15–20%
Garissa Pastoral	12/08	219	.183 (.130, .240)	1.00	1.00	0.94
	3/09	211	.156 (.110, .213)	1.00	1.00	0.00
	6/09	238	.126 (.087, .175)	1.00	0.78	0.00
Garissa Riverine	12/08	224	.125 (.085, .176)	1.00	0.79	0.00
	3/09	227	.207 (.156, .266)	1.00	1.00	1.00
	6/09	222	.171 (.124, .227)	1.00	1.00	0.61
Garissa Urban	12/08	230	.170 (.123, .224)	1.00	1.00	0.53
	3/09	234	.115 (.077, .163)	1.00	0.37	0.00
	6/09	227	.154 (.110, .208)	1.00	1.00	0.04
Mandera Pastoral	12/08	198	.227 (.171, .292)	1.00	1.00	1.00
	3/09	253	.308 (.252, .369)	1.00	1.00	1.00
	6/09	226	.301 (.242, .365)	1.00	1.00	1.00
Mandera Riverine	12/08	194	.366 (.298, .438)	-	-	-
	3/09	201	.383 (.316, .454)	1.00	1.00	1.00
	6/09	263	.350 (.292, .411)	1.00	1.00	1.00
Mandera Urban	12/08	215	.256 (.199, .320)	1.00	1.00	1.00
	3/09	210	.229 (.174, .291)	1.00	1.00	1.00
	6/09	253	.158 (.115, .209)	1.00	1.00	0.18
Mather Slum	2/09	232	.207 (.157, .265)	1.00	1.00	1.00
	4/09	234	.043 (.021, .077)	0.00	0.00	0.00
	7/09	235	.060 (.033, .098)	0.08	0.00	0.00
Sudan Urban	3/08	198	.258 (.198, .324)	1.00	1.00	1.00
	6/08	198	.197 (.144, .259)	1.00	1.00	1.00
	11/08	637	.231 (.199, .265)	1.00	1.00	0.99

Table 2

Examining changes in prevalence over time in the surveyed populations. For each surveillance round, rows 1 and 2 contain estimated prevalence and intraclass correlation; rows 3–6 contain the survey design sample size, decision rule, and probability of declaring a change in prevalence from existing survey data when clustering is ignored; and rows 7–9 present the same information, incorporating clustering variance ination into the design.

Sur.	Garissa			Mandera			Mathere		Sudan	
	Past.	River	Urban	Past.	River	Urban	Slum	Urban	Urban	Urban
1 st	$\hat{p}_1$	.183	.125	.170	.227	.366	.256	.207	.258	.258
	$\hat{\rho}_1$	.051	.062	.042	.005	.047	.000	.051	.032	.032
2 nd	$\hat{p}_2$	.156	.207	.115	.308	.383	.229	.043	.197	.197
	$\hat{\rho}_2$	.038	.067	.048	.000	.000	.132	.029	.051	.051
$\rho_2 = 0$	N_2	116	84	109	156	238	158	129	174	174
	d_2	25	13	22	41	96	46	31	51	51
$\rho_2 = 0$	$P(X_2)$	.002	.908	.000	.967	-	.000	.000	.000	.000
	d_2	36	20	31	44	174	49	46	67	67
3 rd	$\hat{p}_3$	.126	.171	.154	.301	.350	.158	.060	.231	.231
	$\hat{\rho}_3$	.006	.057	.000	.097	.054	.052	.000	.006	.006
$\rho_3 = 0$	N_3	101	84	82	156	238	148	46	130	130
	d_3	19	13	12	41	96	39	3	30	30
$\rho_3 = 0$	$P(X_3)$	.004	.623	.517	.955	.000	.000	.284	.449	.449
	d_3	34	19	17	44	174	110	4	48	48
3 rd	$P(X_3)$	.000	.646	.645	.976	-	-	.270	.385	.385