Sample sizes for estimating the sensitivity of a monitoring system that generates repeated binary outcomes with autocorrelation

Albert E. Parker, James W. Arbogast

Sample sizes for estimating the sensitivity of a monitoring system that generates repeated binary outcomes with autocorrelation

Albert E. Parker[1] [2], James W. Arbogast[3] [4]

Abstract

Sample size formulas are provided to determine how many events and how many patient care units are needed to estimate the sensitivity of a monitoring system. The monitoring systems we consider generate time series binary data that are autocorrelated and clustered by patient care units. Our application of interest is an automated hand hygiene monitoring system that assesses whether healthcare workers perform hand hygiene when they should. We apply an autoregressive order 1 mixed effects logistic regression model to determine sample sizes that allow the sensitivity of the monitoring system to be estimated at a specified confidence level and margin of error. This model overcomes a major limitation of simpler approaches that fail to provide confidence intervals with the specified levels of confidence when the sensitivity of the monitoring system is above 90%.

Keywords

sensitivity, sample size calculation, autoregressive, time series, binary, mixed effects logistic regression, hand hygiene, automated hand hygiene monitoring, hand hygiene compliance, electronic compliance monitoring

[1] Center for Biofilm Engineering, Montana State University, Bozeman, MT 59717
[2] Department of Mathematical Sciences, Montana State University, Bozeman, MT 59717
[3] GOJO Industries Inc., Akron, OH
[4] JW Arbogast Advanced Science Consulting, LLC

1. Introduction

Validating the sensitivity and specificity of monitoring systems in healthcare facilities is crucial for maintaining patient and healthcare worker safety. Whereas a diagnostic test produces a single binary outcome (e.g., healthy or diseased), a monitoring system produces a time series of binary outcomes that may be serially correlated. Furthermore, monitoring data from healthcare facilities typically are clustered by patient care units. Hence, approaches for estimating sensitivity and specificity of diagnostic tests that assume independence of observations are not appropriate for estimating the sensitivity and specificity of monitoring systems. This paper provides explicit formulas that recommend how many outcomes to observe when validating a monitoring system with the goal of estimating the sensitivity with a high level of precision and statistical confidence. Although this paper focuses on sensitivity, it is straightforward to use the same techniques presented here to also estimate specificity. If the goal is to estimate the sensitivity of a monitoring system in a single patient care unit (i.e., one cluster), the sample size formula is based on an autoregressive order 1 (AR(1)) logistic regression that models the serial correlation of the outcomes over time. If the goal is to estimate the sensitivity of a monitoring system across an entire healthcare facility based on only a random sample of units, a sample size formula is provided for determining the number of units and the number of outcomes at each unit. This second formula is based on an AR(1) logistic regression model with a random effect for patient care unit. Results are compared to the more simplistic "*z*-test of proportions" approach that fails to generate CIs with the correct confidence levels when the sensitivity of the monitoring system is above 90% even when the outcomes are independent.

2. Motivating Example

Our application of interest is an automated hand hygiene monitoring system (AHHMS). Achieving high hand hygiene compliance among healthcare workers (HCW) is crucial for improving patient safety[1,2]. The evidence is clear that HCWs do not perform hand hygiene as often as they should, with compliance rates reported to be less than 50% in many healthcare facilities.[3-5] One way to monitor compliance is an AHHMS. An AHHMS monitors two different types of events: sensors in doorways indicate an entry into or an exit from a patient's room; while sensors in dispensers near the room indicate a dispense of soap or alcohol based handrub (ABHR). Hence, the AHHMS generates bivariate binary data: 1's when there is an exit into or out of a patient's room, 1's when there's a soap or ABHR dispense near the room, and 0's otherwise. In practice, the sensitivity of the doorway sensors is validated separately from the dispenser sensors. We recommend that the approach in this paper be applied to estimate the sensitivity of each sensor type separately. A HCW is assumed to satisfactorily perform hand hygiene if a soap or alcohol dispense accompanies the entry into the patient's room and the exit from the patient's room. The Leapfrog Group, a nonprofit patient safety organization that collects and reports healthcare facility performance to the public and assigns a letter grade to facilities based on their record of patient safety[6], publicly grades hospitals based on different performance criteria including whether hand hygiene compliance is monitored and whether the monitoring system is working properly, i.e., has a high level of sensitivity. In this context, sensitivity is the percentage of times that the AHHMS correctly captures an event (room entry, room exit; or a hand hygiene dispense) as corroborated by a human observer. Recent literature[7] suggests that a validation study of a monitoring system should have two phases: a planned path phase and a behavioral phase, see Figure 1.

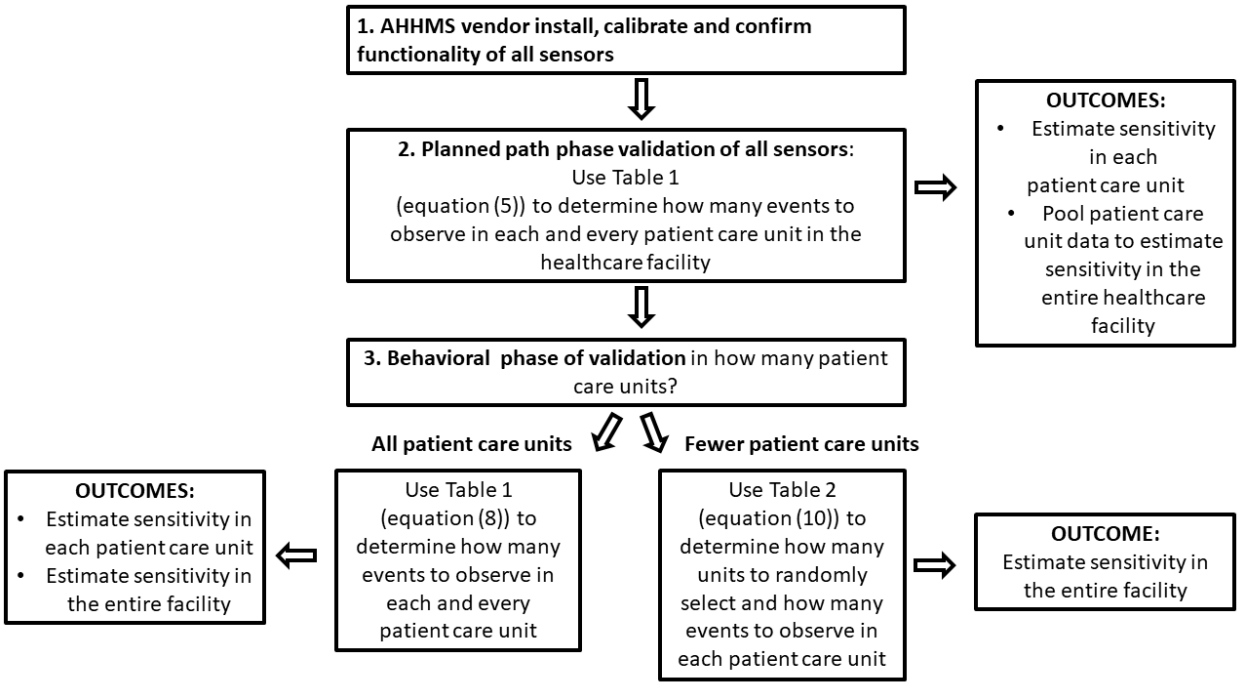

Figure 1. How to validate an AHHMS? How many patient care units to include? How many events in each patient care unit?

During the planned path phase, AHHMS sensors are purposefully activated by the observer. Because the observer controls the timing and physical locations of the activations, we assume that the events during the planned path phase are independent. Hence we provide a sample size calculation for estimating sensitivity of the AHHMS during the planned path phase that assumes the events are independent using logistic regression. We do not recommend the simplistic "*z*-test of proportions" approach that fails to generate CIs with the correct confidence levels when the sensitivity of the AHHMS is above 90%. During the behavioral phase, AHHMS sensors are activated by HCWs during their normal workflow. Because preliminary investigation suggests that hand hygiene events may be serially correlated over time, we calculate sample sizes to estimate sensitivity of the AHHMS during the behavioral phase using

an AR(1) logistic regression model that accounts for this serial correlation. In both the planned path and behavioral phases, the sensitivity of the doorway sensors is validated separately from the dispenser sensors. The doorway and dispenser sensors of an AHHMS each generate a time series of binary data for each patient care unit where it is installed. Indeed, it is common to compare hand hygiene compliance across different patient care units. Some healthcare facilities may have the resources to complete the behavioral phase in every patient care unit. However, this approach may be too resource intensive for other facilities. In this case we propose completing the behavioral phase with a few randomly chosen units and then estimating the sensitivity of the AHHMS across the entire facility by analyzing the data with an AR(1) logistic regression model with a random effect for patient care unit.

3. Sample size calculations

The goal of validating a monitoring system is to estimate the true sensitivity (π) with a point estimator (s) and a one-sided lower confidence limit (LCL),

$$LCL = s - m \tag{1}$$

where m is the margin of error that depends on the confidence level C (see, e.g., Newcombe 1998, McCracken & Looney 2017).[8,9] That is, at a confidence level C, the goal is to conclude that the true sensitivity of the system is larger than the LCL. We will calculate the number of events to observe at a single patient care unit in section 3.1, and then consider how many events to observe at multiple units in section 3.2.

3.1 Number of events at one unit

We will show how to calculate the LCL for the sensitivity of a monitoring system at a single patient care unit and then provide formulas that predict how many events to monitor during validation to attain desired levels of statistical confidence and precision. When the events are independent, we recommend equation (5) below as a sample size calculator for determining the number of events to monitor to validate a monitoring system. When the events are serially correlated according to an AR(1) process, we recommend (8). First, however, a more common approach is reviewed that has limitations which are overcome by use of equations (5) and (8).

Let $\{y_t\}_{t=0}^{N-1}$ be a time series of binary outcomes from a monitoring system from a single patient care unit. In our motivating example, the events we want to detect from doorway sensors are entries and exits from a patient's room; the events we want to detect from dispenser sensors are hand hygiene events; the sensitivity of both kinds of sensors will be established separately. The outcome $y_t = 1$ indicates that the monitoring system detected an event at time t. We will use $x=1$ to indicate if an event truly occurred, as supervised by a human observer who is tasked with validating the system; otherwise, $x=0$. The proportion

$$\pi = p(y_t = 1 | x = 1)$$

is the true sensitivity of the monitoring system. Put another way, we will focus on a subset of the time series that corresponds to when $n_{\text{events}} \leq N$ true events occur, $\{y_t | x=1\}$, to estimate sensitivity.

3.1.1 Independent events: a common simple approach

The simplest approach for calculating the LCL uses a sample proportion $s = \bar{y}|(x = 1)$ to estimate sensitivity. This assumes that there are a sufficient number of events n_{events} so that the distribution for s is approximately normal, in which case the margin of error at a $C\times100\%$ confidence level is

$$(2) \qquad m = z_C\sqrt{\frac{s(1-s)}{n_{events}}}$$

(see, e.g., Newcombe 1998, Obuchowski 1998, Hajian-Tilaki 2014, McCracken & Looney 2017).[8-11] In equation (2), n_{events} is the number of events occurring in a single patient care unit. To assess sensitivity of the monitoring system in a single unit at 95% confidence, one might solve equation (2) for the sample size to get the standard formula from a first semester statistics class

$$(3) \qquad n_{\text{events}} = \frac{z_C^2 s(1-s)}{m^2}$$

where:

- z_C is the multiplier depending on the confidence level C. At 95% confidence, $z_{0.95} = 1.645$
- m is the desired margin of error
- s is some estimator of the true sensitivity π before data collection, so it is typically calculated from a pilot study or from literature. After binary data from n_{events} events

are collected, a new estimator s of the sensitivity is calculated by simple averaging which is then plugged into equations (1) and (2) to generate a CI for the true sensitivity π.

It is well-known (see, e.g., Newcombe 1998, p858; McCracken & Looney 2017)[8,9] that when estimating percentages close to 100%, as we do here when estimating sensitivities, equations (2) and (3) provide CIs that have less than $C\times100\%$ confidence unless the sample sizes are very large. For example, Limper et al., (2016)[7] use an equation similar to (3) to propose monitoring $n_{\text{events}} = 123$ events to validate a monitoring system in a patient care unit when estimating sensitivities of 90% or higher at a 95% confidence level. Simulations using equation (2) show that the actual confidence level of the resulting CIs constructed from equations (1) and (2) is 93.7% when the true sensitivity is $\pi = 90\%$, 85.4% when $\pi = 95\%$, and 69.5% when π=99%. This degradation of the confidence level is because the normal approximation becomes increasingly worse as π approaches 100%, see Figure 2.

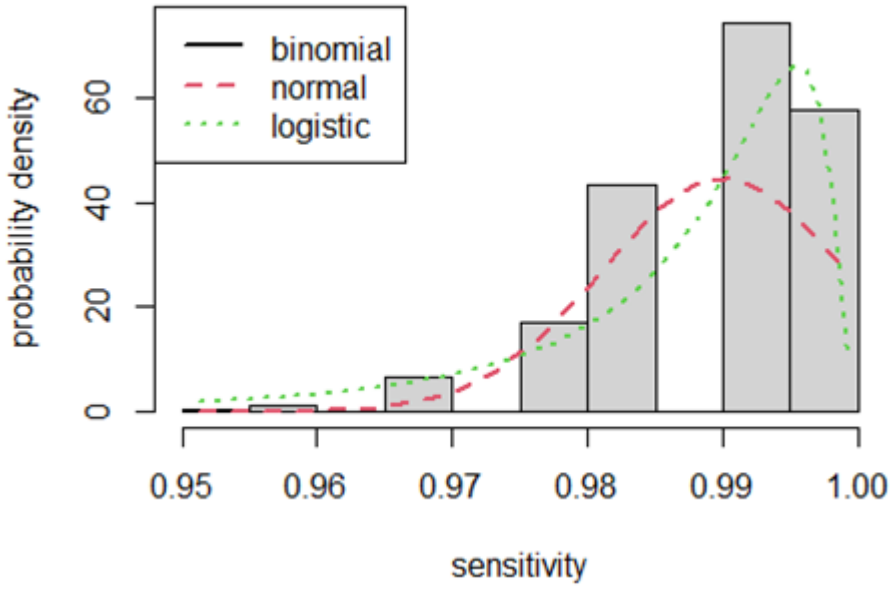

Figure 2. Three different distributions of the sensitivity estimator *s* obtained from observing n = 123 events when the true sensitivity is π = 99%. When *s* is the sample proportion as assumed by (2) and (3), the bars indicate the "exact" binomial distribution while the red dashed curve indicates the ill fit of the associated normal approximation. The green dotted curve indicates the logistic regression distribution of the sample proportion *s* assumed by equations (4) and (5) that better approximates the binomial distribution when the true sensitivity is close to 100%.

3.1.2 Independent events: a more flexible approach

There are other approaches for calculating the LCL that do not assume that the estimator *s* for sensitivity is approximately normal (see, e.g., Newcombe 1998).[8] Here we focus on constructing a margin of error to be used in equation (1) to construct a LCL with better realized confidence levels. When the events are independent, we model the data $\{y_i|x=1\}$ with logistic regression with model equation

$$\text{log-odds} = \ln\left(\frac{\pi}{1-\pi}\right) = \mu$$

(see, e.g., Neter et al. 1996, p. 574)[12] where μ is the log-odds of the monitoring system after an event really occurs (when $x = 1$). The parameter of interest is $\pi = \frac{\exp(\mu)}{1+\exp(\mu)}$. When the events are independent (as is assumed during the planned path phase in our motivating example) then the autocorrelation $\text{Cor}(y_t, y_{t+k}) = 0$ for any $k > 0$. Under the assumption that the estimator $\hat{\mu}$ is normally distributed, then logistic regression gives a Cx100% one-sided LCL for the log-odds for sensitivity μ as

$$\text{LCL}_{\text{log-odds}} = \ln\left(\frac{s}{1-s}\right) - z_C\sqrt{\frac{1}{n_{\text{events}}s(1-s)}}$$

where the Cramer-Rao estimator $\widehat{Var(\hat{\mu})} = \frac{1}{n_{\text{events}}s(1-s)}$ is used and s is the sample proportion from n_{events} independent binary data. From this, a LCL for the sensitivity π as in equation (1) can be calculated using the standard transform from logistic regression,

$$\text{LCL} = s - m = \frac{\exp(\text{LCL}_{\text{log-odds}})}{1 + \exp(\text{LCL}_{\text{log-odds}})}$$

which gives the following formula for the margin of error

$$(4) \qquad m = s - \frac{\exp\left(-z_C\sqrt{\frac{1}{n_{\text{events}}s(1-s)}}\right)}{1/s - 1 + \exp\left(-z_C\sqrt{\frac{1}{n_{\text{events}}s(1-s)}}\right)}.$$

Just as equation (2) was used to derive sample sizes when s is normal, we used equation (4) to derive a formula for the number of events to observe from a monitoring system when $\hat{\mu}$ is normal,

(5) $$n_{\text{events}} = \frac{z_C^2}{s(1-s)\left(\ln\left(\frac{s(1-s+m)}{(s-m)(1-s)}\right)\right)^2}.$$

(see section A.1 in the Appendix for a derivation). Because (5) is calculated before data collection, *s* in equation (5) (as in (3), (8) and (10)) is typically calculated from a pilot study or from literature. After binary data from n_{events} events are collected, a new estimator *s* of the sensitivity is calculated by logistic regression which is then plugged into equations (1) and (4) (or (7) or (9) below)) to generate a CI for the true sensitivity π.

To assess whether equation (5) deals with the pitfalls encountered by equation (3), consider again the example from Limper et al. (2016)[7], who used an equation similar to (3) to propose monitoring $n_{\text{events}} = 123$ events to validate a monitoring system at a single patient care unit when estimating sensitivities of 90% or higher at a 95% confidence level. Simulations of $n_{events} = 123$ binary data were used to construct nominal 95% CIs via equations (1) and (4), which showed that the actual confidence level was 96.1% when the true sensitivity was $\pi = 90\%$, 98.9% when $\pi = 95\%$, and 99.9% when π=99%. These results show that these CIs are overly conservative. The reason that the CIs from equation (4) have higher confidence levels than those attained by equation (2) is that the distribution assumed for the sensitivity *s* more closely aligns with the actual distribution when the sensitivities get close to 100%, see Figure 2. We do not investigate other techniques that are more closely aligned with the nominal confidence levels because (4) and (5) lend themselves to two important extensions: time series modeling and clustering by patient care unit.

3.1.3 Serially correlated events

For serially correlated time series, we model the full time series $\{y_t\}_{t=0}^{N-1}$ with the logistic regression model equation

(6) $$\text{log-odds} = \ln\left(\frac{\pi_x}{1-\pi_x}\right) = \mu_x$$

where μ_1 is the log-odds of the monitoring system after an event really occurs (when $x = 1$) and μ_0 is the log-odds of the monitoring system when no events occur (when $x = 0$). The parameter of interest is the true sensitivity $\pi \coloneqq \pi_1 = \frac{\exp(\mu_1)}{1+\exp(\mu_1)}$. The events are dependent, with dependence defined by the AR(1) autocorrelation function (see section A.2 in the Appendix). When modeling serially correlated time series with an AR(1) process, there is a simple variance inflation factor

$$v = \frac{1+r}{1-r}$$

to incorporate into the formula for the margin of error in equation (4) where r is the "one step correlation" over time. The resulting margin of error for sensitivity to use in equation (1) is

(7) $$m = s - \frac{\exp\left(-z_C\sqrt{\frac{v}{n_{\text{events}}s(1-s)}}\right)}{1/s-1+\exp\left(-z_C\sqrt{\frac{v}{n_{\text{events}}s(1-s)}}\right)}$$

(see section A.2 in the Appendix for a proof) where s is the estimator of π from an AR(1) logistic regression applied to n_{events} binary data. This easy way to account for correlation is one reason we chose to model the sensitivity using logistic regression. When the events are independent, then the variance inflation factor is $v = 1$ and the margin of error in equation (7) simplifies to

equation (4). When there is a serial correlation among events over time, however, then $v > 1$, which increases the margin of error. Solving equation (7) for the sample size shows that

$$(8) \qquad n_{\text{events}} = \frac{v \times z_C^2}{s(1-s)\left(\ln\left(\frac{s(1-s+m)}{(s-\text{m})(1-s)}\right)\right)^2}$$

is the number of events to monitor during system validation. In equation (8), s is typically calculated from a pilot study or from literature. The difference between equations (5) and (8) is multiplication by the inflation factor v. If there is a moderate level of serial correlation, $r = 0.5$, then the inflation factor is $v = 3$. Plugging this into equation (8) predicts a three-fold increase in sample size for serially correlated events versus when the events are independent!

3.2 Number of events at multiple units

The equations presented earlier for a single patient care unit can be updated to provide the number of events and number of randomly selected patient care units to observe in a large healthcare facility with the goal of estimating the sensitivity (π) of the monitoring system across the entire facility (i.e., pooled over all patient care units) at a high level of confidence.

Consider the binary AR(1) time series $\{y_{t,i}\}_{t=0}^{N-1}$ generated by the monitoring system in unit i (sensors indicate an event or not). The logistic regression model for multi-unit data includes a random effect (u_i) for patient care unit,

$$\text{log-odds} = \ln\left(\frac{\pi_{x,i}}{1-\pi_{x,i}}\right) = \mu_x + u_i,$$

where μ_1 is the log-odds that the monitoring system detects a real event (x = 1) pooled across all units in the facility, μ_0 is the log-odds of the monitoring system when no real events occur (x = 0), $u_i \sim N(0, \sigma^2_{\text{unit}})$ and $\mu_x + u_i$ is the log-odds for the i^{th} unit (see, e.g., Moerbeek et al. 2001, Abebe et al. 2015).[13,14] The parameter $\pi_i \coloneqq \pi_{1,i} = p(y_{t,i} = 1|x = 1)$ is the true sensitivity for patient care unit i. The variance of the log-odds among units σ^2_{unit} can be estimated by V_{unit} which is found by fitting an AR(1) logistic mixed effects model to the binary time series from multiple units at the same healthcare facility. The parameter of interest is $\pi = \frac{\exp(\mu_1)}{1+\exp(\mu_1)}$, which is the sensitivity pooled over all patient care units (i.e., when u_i=0 for all i). To build a confidence interval for π, we focus on the subset $\{y_{t,i} | x=1\}$ of n_{events} events from each of the n_{units} units. The variance of the estimator of the overall log-odds is

$$\text{Var}(\hat{\mu}_1) = \frac{v}{n_{\text{events}} n_{\text{units}} \pi(1-\pi)} + \frac{\sigma^2_{\text{units}}}{n_{\text{units}}}$$

(see section 3.1.3). When σ^2_{units} must be estimated from data by V_{units} then this equation provides a variance estimator that is biased downward (Moerbeek et al, 2001, Van Breukelen & Candel 2015).[13,15] Van Breukelen & Candel (2015)[15] suggest using

$$\widehat{Var(\hat{\mu}_1)} = \frac{5}{4}\left(\frac{v}{n_{\text{events}} n_{\text{units}} s(1-s)} + \frac{V_{\text{units}}}{n_{\text{units}}}\right)$$

to overcome the bias. This adds the unit-to-unit variance (V_{units}) to the margin of error in equation (7) when estimating the sensitivity π at a $C\times100\%$ confidence level,

$$(9) \qquad m = s - \left(\frac{\exp\left(-\frac{\sqrt{5}z_C}{2}\sqrt{\frac{v}{n_{\text{events}} n_{\text{units}} s(1-s)} + \frac{V_{unit}}{n_{\text{units}}}}\right)}{1/s - 1 + \exp\left(-\frac{\sqrt{5}z_C}{2}\sqrt{\frac{v}{n_{\text{events}} n_{\text{units}} s(1-s)} + \frac{V_{unit}}{n_{\text{units}}}}\right)} \right)$$

where s is the estimator of π from an AR(1) mixed effects logistic regression applied to n_{events} binary data from each of n_{units} units (see section A.3 in the Appendix). This straightforward way to account for the variance among units in the same facility is one reason we chose to model the sensitivity using logistic regression. From equation (9) one can set the number of events that need to be observed at each unit (n_{events}) and the number of units to be randomly selected from the facility (n_{units}) to calculate the LCL for the sensitivity across the facility using equations (1) and (9).

From equation (9) one can fix the number of units (n_{units}) to be randomly selected for validation, then find the number of events (n_{events}) that need to be observed at each of these units as in the following equation,

$$(10) \qquad n_{\text{events}} = \frac{5 \times v \times z_C^2}{s(1-s)\left[4n_{\text{units}}\left(\ln\left(\frac{s(1-s+m)}{(s-\text{m})(1-s)}\right)\right)^2 - 5z_C^2 V_{\text{units}}\right]}.$$

In equation (10), s is typically calculated from a pilot study or from literature. Equation (10) is the number of events to monitor at each of n_{units} randomly selected patient care units to validate a monitoring system when the events are either independent ($v = 1$) or serially correlated ($v > 1$). Notice that equation (10) suggests valid sample sizes, $n_{\text{events}} > 0$, as long as the unit-to-unit variance σ^2_{units} (and its estimate V_{units}) is not too large. This is because when the unit-to-unit variance is large, the study must include more units (n_{units}) to decrease the margin of error; it is not possible to decrease the margin of error simply by increasing the number of events (n_{events}) at just a few units.

4. Illustrative Examples

In the context of the motivating example in section 2, sensitivity is the percentage of times that the AHHMS correctly captures an event. There are two kinds of sensitivity to assess for an AHHMS: sensitivity of the door sensors where an event is either an entry into or an exit from a patient's room; and sensitivity of the dispenser sensors where an event is a dispense from either a soap or ABHR dispenser.

A few examples are provided in this section.

4.1 Example #1: sample size for the planned path phase

If a healthcare facility wants to estimate sensitivity during the planned path phase of validation in a single patient care unit (see Figure 1) with a margin of error m = 10% at a C=95% confidence level, and they expect the sensitivity to be π=90%, then equation (5) shows that:

$$n_{\text{events}} = \frac{1.645^2}{0.9(1-0.9)\left(\ln\left(\frac{0.9(1-0.9+0.1)}{(0.9-0.1)(1-0.9)}\right)\right)^2} = 45.7.$$

This suggests that 46 events (either entries and exits, or dispenses) should be monitored at the patient care unit to estimate monitoring system sensitivity during the planned path phase of validation (compare with Table 1). Assuming that n = 46 events are observed in this one unit and the monitoring system detects 42 of these events, then sensitivity is estimated to be s = 42/46 = 91%. To build a CI for the true sensitivity π as in equation (1), first plug into equation (4) to get the margin of error,

$$m = 0.91 - \frac{\exp\left(-1.645\sqrt{\frac{1}{46\times0.91\times0.09}}\right)}{1/0.91-1+\exp\left(-1.645\sqrt{\frac{1}{46\times0.91\times0.09}}\right)} = 0.098.$$

Table 1. Number of observed events to validate a monitoring system in a single patient care unit for different confidence levels, margins of error, and expected sensitivities using equations (5) and (8). The number of serially correlated observations was calculated assuming a moderate level of correlation over time ($r = 0.5$).

Confidence (*C*)	**Margin of error (*m*)**	**Estimated sensitivity (*s*)**	**Number of independent observations**	**Number of serially correlated observations**
80%	10%	95%	11	31
		90%	12	36
	5%	95%	28	81
		90%	37	111
85%	10%	95%	16	47
		90%	19	55
	5%	95%	41	122
		90%	56	168
90%	10%	95%	24	71
		90%	28	84
	5%	95%	62	186
		90%	86	256
95%	10%	95%	39	117
		90%	46	138
	5%	95%	103	307
		90%	141	422
99%	10%	95%	78	234
		90%	92	275
	5%	95%	205	613
		90%	281	843

This is close to the specification above for a margin of error of m = 10%. This means that the 95% LCL for the true sensitivity is 91% - 9.8% ≈ 81%. So, given the data in this example with n = 46 observations in a single patient care unit, we would be 95% confident that the true sensitivity π is larger than 81%. Simulations of $n_{events} = 46$ binary data were used to construct nominal 95% CIs for the true sensitivity π via equations (1) and (4), which showed that the attained confidence level was 99.9%, much higher than 95% even for this relatively small sample size when the true sensitivity is π = 90%. To estimate the sensitivity across the entire facility, we recommend analyzing the data across all units with a logistic regression model with a random effect for unit. For balanced independent data (when there are n_{events} collected at each of the n_{units} in the facility), the analysis will provide a LCL for the sensitivity across the entire facility as provided in equations (1) and (9) with v=1.

A healthcare facility should not use the conventional approach outlined in equations (2) and (3) to design a validation study. In this example the margin of error from equation (2) is

$$m = 1.645\sqrt{\frac{0.91(1-0.91)}{46}} = 0.069.$$

Plugging this into equation (1) gives LCL = 84% for the true sensitivity. Simulation studies show that the actual confidence level of the associated CI is 92% which is below the nominal value of 95%.

4.2 Example #2: sample size for the behavioral phase

In this example, we give the number of events to monitor to estimate the sensitivity of the AHHMS in a single unit in the healthcare facility during the behavioral phase of validation (see Figure 1). During the behavioral phase of the study, it is expected that the observed events will come in bursts over time. That is, there is a high correlation that a period of activity (or inactivity) by the monitoring system will be followed by another period of activity (or inactivity). Similar to Example #1, we would like a margin of error $m = 10\%$ at a $C = 95\%$ confidence level and expect the sensitivity to be $\pi = 90\%$ during the behavioral phase of monitoring system validation. Modelling the time series of events with an AR(1) time series model with a serial correlation of $r = 0.5$, equation (8) yields

$$n_{\text{events}} = \frac{3 \times 1.645^2}{0.9(1-0.9)\left(\ln\left(\frac{0.9(1-0.9+0.1)}{(0.9-0.1)(1-0.9)}\right)\right)^2} = 137.1$$

which suggests that 138 events (either entries and exits, or dispenses) should be monitored to estimate sensitivity during the behavioral phase of monitoring system validation at a single patient care unit (compare with Table 1). Assume that $n = 138$ events are observed in one unit and the monitoring system detects 132 of these events. Then sensitivity in the behavioral phase of this study at this one unit is estimated to be $s = 132/138 = 96\%$. To build a CI for the true sensitivity at this one unit as in equation (1), plug into equation (7) to get the margin of error

$$m = 0.96 - \frac{\exp\left(-1.645\sqrt{\frac{0.5}{138\times0.96\times0.04}}\right)}{1/0.95-1+\exp\left(-1.645\sqrt{\frac{1}{138\times0.96\times0.04}}\right)} = 0.086$$

which is close to the specification above for a margin of error of m = 10%. This means that the 95% LCL for the true sensitivity is 87%. So, given the data in this example with n = 138 observations in a single patient care unit, we would be 95% confident that the true sensitivity π is larger than 87%.

To estimate the sensitivity across the entire facility, we recommend analyzing the data across all units with a logistic regression model with a random effect for unit and an AR(1) serial correlation over time. For balanced data (when there are n_{events} collected at each of the n_{units} in the facility), the logistic regression will provide a LCL for the sensitivity across the entire facility as provided in equations (1) and (9) with the variance inflation factor v determined by the one step correlation r estimated from the AR(1) process.

4.3 Example #3: sample size for the behavioral phase at multiple units

One way to estimate the sensitivity across a large healthcare facility with many patient care units is to pool together the data from all of the individual units, where data is collected from each unit as in Example #2. In this example, we consider the scenario where it may be too resource intensive to perform the behavioral phase of validation at each and every unit in the facility. To estimate the sensitivity across the entire facility, in this example we consider randomly selecting and performing validation in n_{units} = 4 units. As in the previous examples, we would like a margin of error m = 10% at a C = 95% confidence level and expect the sensitivity to be π = 90% during the behavioral phase of monitoring system validation. As in Example #2, the time series of events will be modelled with an AR(1) process with a serial correlation r = 0.5. Because we

are randomly sampling from multiple units, an estimate of the variance across units (σ^2_{units}) is required. If there are no pilot data, then we will estimate the unit-to-unit variance as some proportion (*ICC*) of the total variance,

$$V_{\text{units}} = ICC\left(\frac{v}{s(1-s)} + V_{\text{units}}\right) \Leftrightarrow V_{\text{units}} = \frac{ICC}{1-ICC} \times \frac{v}{s(1-s)}.$$

Importantly, *ICC* is the intra-class correlation of the events that occur in the same patient care unit. Estimates of *ICC* can be obtained from a review of 31 multi-cluster studies that found that 50% of the time, $ICC \leq 0.01$; and 90% of the time, $ICC \leq 0.055$ (Table 3 in Adams et al. 2004).[16] Using the median ICC=0.01 suggests that, for this example, $V_{\text{units}} = \frac{0.01}{0.99} \times \frac{3}{0.9 \times 0.1} = 0.33$. Plugging into equation (10) yields

$$n_{\text{events}} = \frac{5 \times 3 \times 1.645^2}{0.9(1-0.9)\left[4 \times 4 \times \left(\ln\left(\frac{0.9(1-0.9+0.1)}{(0.9-0.1)(1-0.9)}\right)\right)^2 - 5 \times 1.645^2 \times 0.33\right]} = 75.6$$

which suggests that 76 events (either entries and exits, or dispenses) should be monitored in each of the n_{units} = 4 patient care units to estimate sensitivity at 95% confidence across the entire facility during the behavioral phase of monitoring system validation (compare with Table 2).

Table 2. Number of randomly chosen units and number of serially correlated events to validate a monitoring system across an entire healthcare facility for different confidence levels and expected sensitivities using equation (10). The margin of error was set at $m = 10\%$ with a moderate level of correlation over time ($r = 0.5$) and low to moderate unit-to-unit variability (i.e., with intra-class correlation ICC=0.01 and 0.05 respectively as described in Example #3).

Confidence	Estimated sensitivity	Number of units	Number of observations at each unit with low unit-to-unit variance	Number of observations at each unit with moderate unit-to-unit variance
80%	95%	2	24	not possible
		3	15	39
		4	11	20
		10	4	5
	90%	2	30	not possible
		3	18	71
		4	13	28
		10	5	6
90%	95%	2	81	not possible
		3	43	not possible
		4	29	not possible
		10	10	17
	90%	2	110	not possible
		3	54	not possible
		4	36	not possible
		10	12	23
95%	95%	2	277	not possible
		3	96	not possible
		4	58	not possible
		10	18	63
	90%	2	639	not possible
		3	136	not possible
		4	76	not possible
		10	21	175

In Example #2, we found that 138 events needed to be observed at each unit. The reason that the required sample size in Example #2 is higher is that the goal in that case was to estimate

the sensitivity at each individual unit within the facility at a high level of precision (i.e., a 10% margin of error) and confidence (95%). In this example, the goal is to estimate the sensitivity across an entire facility based on 4 randomly chosen units, so fewer observations per unit are needed. The downside to this study design is that, with only 76 observations per unit, the sensitivity at each of the randomly selected units is estimated at a lower confidence level (85-90%, see Table 1) when the desired margin of error is 10%. The sensitivity at other units that did not participate in the validation study can still be estimated using a prediction or tolerance interval that will have worse precision and/or confidence than the levels specified in Tables 1 and 2.

Performing the sample size calculation again using a higher level of unit-to-unit variability, the 90th percentile ICC=0.05, suggests $V_{\text{units}} = \frac{0.05}{0.95} \times \frac{3}{0.9 \times 0.1} = 1.75$, from which equation (10) suggests a bogus sample size of $n_{\text{events}} = -34.1$. This highlights that when the unit-to-unit variance σ^2_{units} is large enough, it is not possible to decrease the margin of error simply by increasing the number of events (n_{events}) at each of n_{units} = 4 patient care units. Instead, the validation study must include more than 4 units to decrease the margin of error to the desired level of 10%.

5. Results

Here we apply the sample size calculations developed in section 3 to determine how many events to observe when the goal is to validate the sensitivity of a monitoring system under the two scenarios of interest: in a single patient care unit; and across an entire facility using data only from a few patient care units.

Table 1 shows the required number of observations to validate a monitoring system in a single patient care unit when the data are either independent or serially correlated according to an AR(1) logistic process. As pointed out after equation (8), if there is a moderate level of serial correlation, $r = 0.5$, then a three-fold increase in sample size is predicted for serially correlated events compared to when the events are independent.

Figure 3 compares the suggested sample sizes for independent observations using the logistic regression approach (equation (5)) to the conventional simple approach (equation (3)). As expected from the parabolic expression in (3), the simple approach suggests that sample sizes are symmetric around the peak at $s = 50\%$. As the sensitivity gets closer to 100%, the simple approach suggests smaller and smaller sample sizes. This is an artifact when using the normal approximation for calculating a margin of error for the sensitivity. As shown earlier, the resulting intervals do not really provide the stated confidence levels as the sensitivity gets closer to 100%. The logistic regression approach gives an asymmetric curve of sample sizes that peaks at $s = 60\%$ (regardless of the correlation r, see (5) and (8)), is larger than the simple approach for $s \geq 54\%$ and a markedly increases as the sensitivity gets very close to 100% to maintain a high level of statistical confidence.

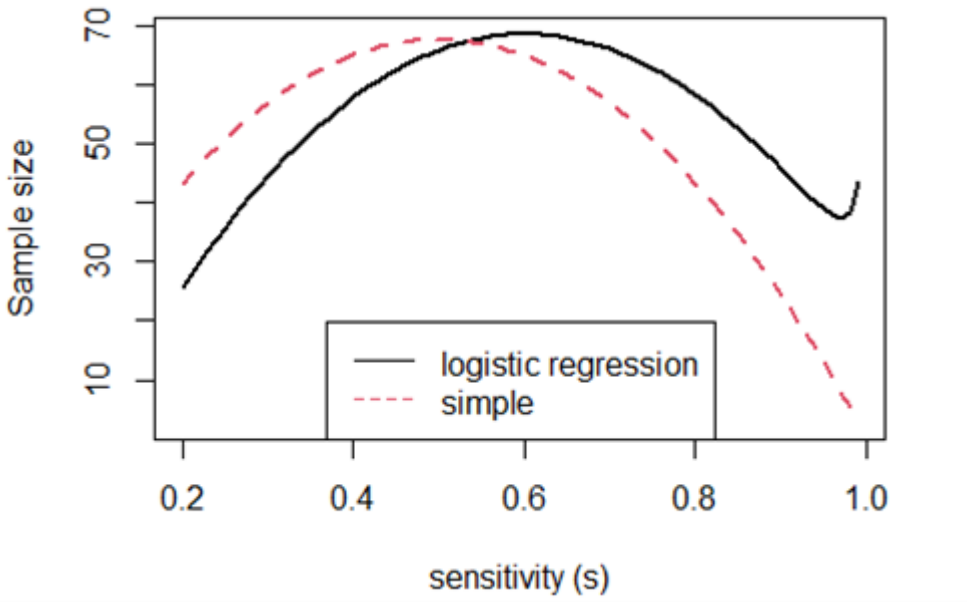

Figure 3. Suggested sample sizes for estimating sensitivity π at a single patient care unit when events are independent using the simple approach in (3) compared to the logistic regression approach in (5) when the estimated sensitivity *s* ranges from 20% to 99%. Results are shown at a *C*=95% confidence level (so z_C = 1.645) and margin of error *m*=10%.

Figure 4 shows how increasing the serial correlation *r* affects the margin of error *m* of a CI for sensitivity when the data are serially correlated according to an AR(1) logistic process. As expected, the margin of error increases as the serial correlation increases.

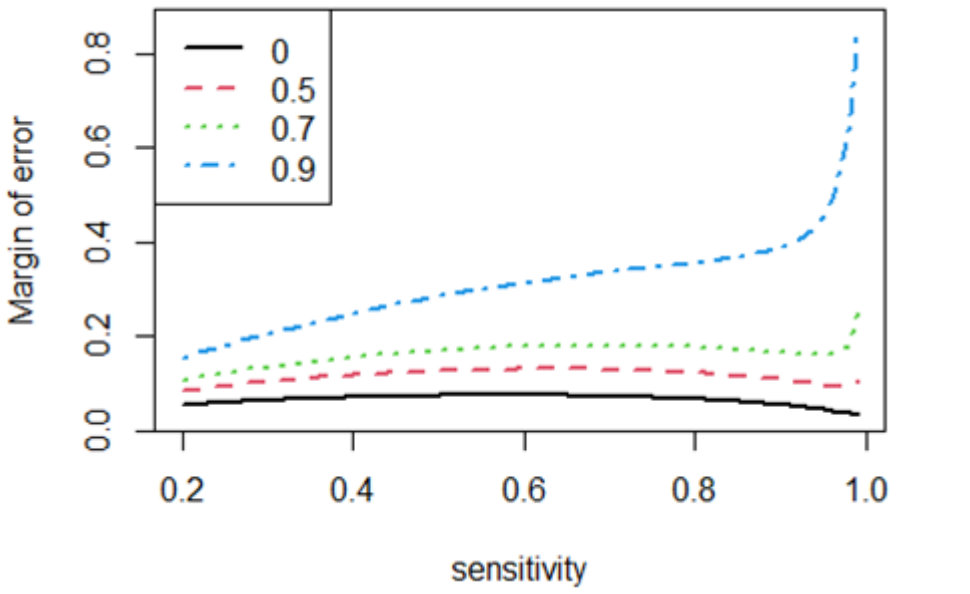

Figure 4. The margin of error (7) for sensitivities from 20% to 99% when events are serially correlated over time at a single patient care unit for different AR(1) correlations (r = 0, 0.5, 0.7, 0.9). Results are shown at a C=95% confidence level (so z_C = 1.645) and n_{events} = 123.

Table 1 and Figures 2-4 focus on estimating the sensitivity at each unit separately with a high level of precision and confidence. Table 2 considers the scenario where one randomly selects n_{units} units to validate which generates an estimate of the sensitivity across the entire healthcare facility. Table 2 suggests that smaller sample sizes be observed at each unit compared to Table 1. This is because the focus of Table 2 is to estimate the sensitivity across the entire healthcare facility with a specified level of precision, however, with a lower level of precision when estimating the sensitivity at any single unit. If one wishes to estimate the sensitivity at each unit with a higher level of precision and confidence, then one should use the sample sizes recommended by Table 1.

When the unit-to-unit variance is large, the study must include a large number of units (n_{units}) to decrease the margin of error. It is not possible to decrease the margin of error simply

by increasing the number of events (n_{events}) at just a few units. This is the scenario when "Not possible" is reported in Table 2 (i.e., (10) yielded a bogus $n_{events} < 0$).

6. Discussion

This work considers sample sizes for estimating the sensitivity of a monitoring system in healthcare settings, which generates autocorrelated binary time series data that are clustered by patient care unit. In our literature review, we found much work that focuses on assessing the sensitivity of diagnostic tests of disease using a simple z-test of proportions (see, e.g., Buderer 1996, Obuchowski 1998, Hajian-Tilaki 2014).[10,11,17] However, the simple z-test of proportions approach presumes independence of the data and, as has been pointed out by others and confirmed here, it does not maintain nominal confidence levels when the true sensitivity is close to 100%. This paper provides explicit formulas that are based on an AR(1) mixed effects logistic regression that overcomes these limitations. Our motivating application is validation of AHHMS for which assuming independence may be inappropriate during the behavioral phase. Unfortunately, there is recent literature[7] that advocates for assuming independence and applying the *z*-test of proportions for calculating sample sizes for validation studies of AHHMS.

An AHHMS continuously monitors hand hygiene among HCWs. The AHHMS is composed of a network of sensors attached to patient doors, sinks, and wall mounted alcohol-based sanitizer dispensers. Others have advocated for validating AHHMS using a two-phase approach (planned path and behavioral)[7] in every unit. This paper builds upon this previous work by 1) providing concrete guidance regarding how many events and how many patient care units to assess during each phase; 2) using logistic regression to overcome the shortcomings of

the simpler test of proportions approach. This will enable healthcare facilities to have high confidence that their AHHMS is functioning with a high level of sensitivity.

Others have considered study designs based on logistic regression with random effects as we do[13,15], although they do not consider autocorrelation and focus on comparing control and treatment groups whereas our focus is on estimating the log-odds for a single group (i.e., detection of real events as identified by a gold standard – in our motivating example, a human observer). In the presence of random effects (when sensitivity is assessed at multiple patient care units), they found that the estimated variance of the logistic regression coefficients was biased downward. To correct for this Van Breukelen & Candel[15] suggested a multiplicative correction factor that we utilize. Another approach, suggested by Candel & Van Breukelen[18] in the context of linear mixed effects models with Gaussian error, is to add an additional cluster (i.e., in our case, another patient care unit) to maintain high confidence levels. While these approaches explicitly address bias in the unit-to-unit variance estimator (in (9) and (10)), they do not address the imprecision of this estimator when the number of units is small (as in Table 2), which could adversely affect the coverage of the resulting CIs after data are collected and analyzed. Abebe et al[14] considered mixed effects AR(1) logistic regression to determine sample sizes as we do but they consider more general forms of the fixed and random effects that required Bayesian simulations. We use a simpler model in which case we are able to provide explicit equations for computing sample sizes. This is the first time we have seen the standard error update by the variance inflation v in the context of an AR(1) logistic regression. Interestingly, a similar update to the standard error has been shown previously for the sample mean for any stationary AR(1) series by Cryer & Chan (p. 29)[19], which for a sample proportion is $\mathrm{Var}(\bar{y}) \approx \frac{v\pi(1-\pi)}{n_{\mathrm{events}}}$. This could

be used to update the simple margin of error and sample size calculations given in equations (2) and (3) when the sensitivity π is not too close to 100%, although we do not investigate that here.

In this work, data are clustered by unit (as opposed to by room) because in our motivating example HCWs may perform hand hygiene at a dispenser in a hallway that connects multiple rooms in the same unit before entering a room or after exiting the room. There are badge-based AHHMs that can monitor when an individual HCW performs hand hygiene, or enters and exits rooms, or enters and exits the "patient care zone" around the patient's bed.[20,21] To evaluate sensitivity of a badge based AHHMS, the statistical model would include a random effect for patient care unit as we do here, but also include an additional random effect for HCW.

The first step in any time series analysis before fitting an ARMA model is to attempt to model any trend or non-stationarity in the data with linear or polynomial or smoother terms (via general additive models)[22]. Afterwards, monitoring system sensitivity is modeled as a logistic function over time, in which case a facility may be interested in estimating the true sensitivity pooled over all time points (by averaging via integration under the logistic curve), or just during times when sensitivity is predicted to be worst. Fitting such models may alleviate the need to fit an ARMA model at all, including the AR(1) model that we consider here. While we use a simple AR(1) process for determining sample sizes, the final model to use when analyzing the resulting data may require polynomial or smoother terms, some other ARMA model, or no ARMA model at all. Such a determination should be made after scrutinizing the residuals. If it is known ahead of time that the model to be fit to the data will differ from the simple models that

we consider here, alternate approaches for calculating sample sizes are available (see, e.g., Abebe at al 2015)[14].

It is conceivable that when validating a monitoring system, the data suggest 100% sensitivity (although we have not seen this in the context of AHHMS, see Limper et al.[7]). McCracken & Looney[9] consider calculating confidence limits for this case using methods other than logistic regression.

For diagnostic tests of disease, Buderer[17] updated the sample size calculator in equation (3) to account for the prevalence p of the disease in the population, $n_{total} = \frac{z^2 s(1-s)}{pm^2}$. This formula indicates the total number of individuals to test when it is not possible at the outset to differentiate between individuals who have the disease and individuals who do not. Put another way, $n_{total} = \frac{n_{events}}{p}$.[11] Thus, the total number of individuals to test, n_{total}, increases markedly when studying a disease with low prevalence, e.g., with a ten-fold increase if only p=10% of the population being tested has the disease. When estimating the sensitivity of an AHHMS (as described in section 2) that indicates when individuals enter or exit patient rooms or perform hand hygiene, then p = 1, so we consider equations (5), (8) and (10) for determining sample sizes. In other scenarios, where, e.g., one may want to consider hand hygiene just among HCWs, then the equations can easily be updated to account for prevalence p of HCWs among all individuals simply by dividing each equation by the prevalence p.

In some cases, the goal may be to assess sensitivity of a monitoring system at a specified level of power. Equations other than (5), (8) and (10) can be derived from logistic regression to

give sample sizes for a validation study that estimate the sensitivity of the monitoring system at a specified level of power (see section A.4 in the Appendix). Unlike the sample size calculations presented here that require an estimate (s) and a desired margin of error (m) for the sensitivity, sample sizes that incorporate power require that values for the sensitivity under null and alternative hypotheses be specified.

The models that we use here to produce sample sizes (in Table 1 and 2) in order to estimate sensitivity at high confidence and precision can be used directly to produce sample sizes to estimate specificity at high confidence and precision. For example, in (6), μ_0 is the log-odds of the monitoring system when no events occur ($x = 0$). The true specificity is $1 - \pi_0$ $=1-\frac{\exp(\mu_0)}{1+\exp(\mu_0)} = p(y_i=0|x=0)$, which is the probability that the monitoring system correctly does not indicate that an event occurred when there really was no event.

Data Availability Statement

The computer code and results that support the findings of this study are available from the corresponding author upon reasonable request.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: GOJO Industries paid Montana State University a fee for AEP

to develop statistical methodologies and to be lead author of the manuscript. JWA was funded by GOJO Industries via normal budgeting and compensation process while employed by them.

References

1. Boyce JM, Pittet D. Healthcare Infection Control Practices Advisory Committee and the HICPAC/SHEA/APIC/IDSA Hand Hygiene Task Force. Guideline for hand hygiene in health-care settings. *MMWR* 2002;51:1-45.
2. World Health Organization. WHO guidelines for hand hygiene in health care. Geneva, Switzerland: World Health Organization, 2009.
3. Boyce JM, Laughman JA, Ader MH, Wagner PT, Parker AE, Arbogast JW. Impact of an automated hand hygiene monitoring system and additional promotional activities on hand hygiene performance rates and healthcare-associated infections. *Infect Control Hosp Epidemiol* 2019;40(7):741-747.
4. Srejic E. Hand hygiene compliance monitoring provides benefits, challenges. *Infection Control Today* December 6, 2015. Accessed June 8, 2022. https://www.infectioncontroltoday.com/view/hand-hygiene-compliance-monitoring-provides-benefits-challenges.
5. Bredin D, O'Doherty D, Hannigan A, Kingston L. Hand hygiene compliance by direct observation in physicians and nurses: a systematic review and meta-analysis. *J Hosp Infect* 2022;130:20-33.
6. The Leapfrog Group. 2023 Leapfrog Hospital Survey, Section 6D: Hand Hygiene, page 194. Accessed June 6, 2023. https://www.leapfroggroup.org/sites/default/files/Files/2023HospitalSurvey_20230401_v9.0.pdf.
7. Limper HM, Slawsky L, Garcia-Houchins S, Mehta S, Hershow RC, Landon E. Assessment of an aggregate-level hand hygiene monitoring technology for measuring hand hygiene performance among healthcare personnel. *Infect Control Hosp Epidemiol* 2016;38(3):348-352.
8. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Stat Med* 1998;17(8):857-872.
9. McCracken CE, Looney SW. On Finding the Upper Confidence Limit for a Binomial Proportion when Zero Successes are Observed. *J Biom Biostat* 2017;8(2):1000338.
10. Obuchowski NA. Sample size calculations in studies of test accuracy. *Stat Methods Med Res* 1998;7(4):371-392.
11. Hajian-Tilaki K. Sample size estimation in diagnostic test studies of biomedical informatics. *J Biomed Inform* 2014;48:193–204.
12. Neter J, Kutner M, Nachtsheim C, Wasserman W. Applied Linear Statistical Models, 4th edition, Irwin, Boston, 1996.
13. Moerbeek M, Van Breukeelen G, Berger M. Optimal experimental designs for multilevel logistic models. *The Statistician* 2001;50(2):17-30.
14. Abebe H, Tan F, Van Breukelen G, Berger M. Bayesian design for dichotomous repeated measurements with autocorrelation. *Statistical Methods in Medical Research* 2015;24(5): 594-611.
15. van Breukelen GJP, Candel MJJM. Efficient design of cluster randomized and multicentre trials with unknown intraclass correlation. *Statistical Methods in Medical Research* 2015;24(5):540-556

16. Adams G, Gulliford MC, Ukoumunne OC, Eldridge S, Chinn S and Campbell MJ. Patterns of intracluster correlation from primary care research to inform study design and analysis. *J Clin Epidemiol* 2004;57:785–794.
17. Buderer NM. Statistical methodology: I. Incorporating the prevalence of disease into the sample size calculation for sensitivity and specificity. *Acad Emerg Med* 1996;3(9):895-900.
18. Candel MJJM, van Breukelen GJP. Sample size adjustments for varying cluster sizes in cluster randomized trials with binary outcomes analyzed with second-order PQL mixed logistic regression. *Statistics in Medicine* 2010;29(14):1488-1501.
19. Cryer, J., Chan, K. Time Series Analysis with Applications in R, 2nd edition, New York, NY: USA: Springer, 2008.
20. Knepper BD, Miller AM, Young HL. Impact of an automated hand hygiene monitoring system combined with a performance improvement intervention on hospital-acquired infections. *Infect Control Hosp Epidemiol* 2020;41(8):931-937.
21. Strauch J, Braun TM, Short H. Use of an automated and hygiene compliance system by emergency room nurses and technicians is associated with decreased employee absenteeism. *Am J Infect Control* 2020;48:575-577.
22. Arbogast JW, Moore LD, DiGiorgio M, Robbins G, Clark TL, Thompson MF, Wagner PT, Boyce JM, Parker AE. The impact of automated hand hygiene monitoring with and without complementary improvement strategies on performance rates. *Infect Control Hosp Epidemiol* 2023;44:638–642.
23. Kac M, Murdock W, Szegö G. On the eigenvalues of certain Hermitian forms. *J. Rational Mech. Anal* 1953;2:767–800.
24. Wackerly D, Mendenhall W, Scheaffer R. Mathematical Statistics with Applications, 7th ed. Mason OH: Brooks/Cole, 2008.

Appendix

A.1 LCL for sensitivity at one unit assuming independence

Starting with

$$\text{LCL} = s - m = \frac{\exp(\text{LCL}_{\text{log-odds}})}{1 + \exp(\text{LCL}_{\text{log-odds}})}$$

from section 3.1.2, the algebraic steps that solve for n_{events} are

$$\text{LCL} = (1 - \text{LCL}) \times \exp(\text{LCL}_{\text{log-odds}})$$

$$\ln(\text{LCL}) = \ln(1 - \text{LCL}) + \ln\left(\frac{s}{1-s}\right) - z_C\sqrt{\frac{1}{n_{\text{events}} s(1-s)}}$$

$$z_C\sqrt{\frac{1}{n_{\text{events}}s(1-s)}} = \ln\left(\frac{s(1-LCL)}{LCL(1-s)}\right).$$

Solving this last equation for n_{events} gives the formula for the sample size in equation (5).

A.2 LCL for sensitivity at one unit using AR(1) process

For serially correlated time series $\{y_t\}_{t=0}^{N-1}$, the logistic regression model equation is given by (6). Assuming that the time series is stationary, dependence is defined by the autocorrelation function $Cor(y_t, y_{t+k}) = \rho(k)$ for any $k > 0$ that can take on any values between -1 and 1. At first we will assume that $x = 1$ for all $N=n_{\text{events}}$ points $\{y_t\}$ in the AR(1) time series, and then consider the time series $\{y_t|x=1\}$ used in practice. That is, at first we will consider the simplified intercept-only AR(1) logistic model

(11) $$\text{log-odds} = \ln\left(\frac{\pi}{1-\pi}\right) = \mu,$$

(model (6) with $x = 1$ for all y_t) then we will consider model (6) that is used in practice.

Following Abebe et al. (p. 597)[14], we apply an extension of the generalized estimating equation approach that gives the estimator

(12) $$\widehat{\text{Var}(\hat{\mu})} = \left[\frac{\partial P}{\partial \mu}^T V^{-1} \frac{\partial P}{\partial \mu}\right]^{-1}$$

where P is a $n_{\text{events}} \times 1$ column vector with t^{th} component $p(y_t)$ and V is the $n_{\text{events}} \times n_{\text{events}}$ covariance matrix of the binary responses,

$$V = \text{Var}(\boldsymbol{y}) = W^{1/2} R W^{1/2}.$$

The matrix W is an $n_{\text{events}} \times n_{\text{events}}$ diagonal matrix $W = \pi(1-\pi)I$, and $R = \text{Cor}(\boldsymbol{y_t})$ is the $n_{\text{events}} \times n_{\text{events}}$ correlation matrix with entries $R_{i.j} = \rho(|i-j|)$, so

$$V = \pi(1-\pi)R.$$

For the model (11), all n_{events} components of $\frac{\partial P}{\partial \mu}$ are equal to $\frac{\partial}{\partial \mu}\left(\frac{\exp(\mu)}{1+\exp(\mu)}\right) = \frac{\exp(\mu)}{(1+\exp(\mu))^2} = \pi(1-\pi)$. Plugging back into (12) gives the estimator

$$\widehat{Var(\hat{\mu})} = \frac{1}{\pi(1-\pi)}[\mathbf{1}^T R^{-1} \mathbf{1}]^{-1}$$

where $\mathbf{1}$ is an $n_{\text{events}} \times 1$ column vector of 1's. This holds for model (11) for any correlation matrix R. For example, when the data are independent, then R=I, and $\mathbf{1}^T R^{-1} \mathbf{1} = n_{\text{events}}$ which yields $\widehat{Var(\hat{\mu})} = \frac{1}{n_{\text{events}}\pi(1-\pi)}$ which is the same variance estimator used in sections 3.1.2 and A.1. For an AR(1) process, $\rho(k) = r^k$, $R_{i.j} = r^{|i-j|}$ and R^{-1} is a tridiagonal matrix with

(13) $\text{diag}(R^{-1}) = \left[\frac{1}{1-r^2}, \frac{1+r^2}{1-r^2}, \ldots, \frac{1+r^2}{1-r^2}, \frac{1}{1-r^2}\right]$

and off-diagonal equal to $\frac{-r}{1-r^2}$ (Kac et al.)[23]. Therefore,

(14) $\mathbf{1}^T R^{-1} \mathbf{1} = \frac{1}{1-r^2}\left(2(1-r) + (1-2r+r^2)(n_{\text{events}} - 2)\right)$

which shows that

$$\text{(15)} \quad \widehat{Var(\hat{\mu})} = \frac{1}{\pi(1-\pi)}\left(n_{\text{events}}\frac{1-r}{1+r} + \frac{2r}{1+r}\right)^{-1},$$

so when n_{events} is large,

$$\text{(16)} \quad \widehat{Var(\hat{\mu})} \approx \frac{1}{n_{\text{events}}\pi(1-\pi)}\frac{1+r}{1-r}.$$

Now we will focus on a subset $\{y_t|x=1\}=y_{t_1},\ldots,y_{t_{n_{\text{events}}}}$ that corresponds to when $n_{\text{events}}< N$ true events occur with model equation (6). The $n_{\text{events}}\times n_{\text{events}}$ correlation matrix S for the n_{events} $\{y_t|x=1\}$ has components $S_{ij} = r^{|t_i-t_j|}=r^{|\Delta t_j|}$. As in equation (12),

$$\text{(17)} \quad \widehat{Var(\hat{\mu}_1)} = \left[\frac{\partial P}{\partial \mu}^T (\pi(1-\pi)S)^{-1}\frac{\partial P}{\partial \mu}\right]^{-1} = \frac{1}{\pi(1-\pi)}[\mathbf{1}^T S^{-1}\mathbf{1}]^{-1}$$

where S^{-1} is a tridiagonal matrix with

$$\text{(18)} \quad \text{diag}(S^{-1})=\left[\frac{1}{1-r^{2\Delta t_1}}, 1+\frac{r^{2\Delta t_1}}{1-r^{2\Delta t_1}}+\frac{r^{2\Delta t_2}}{1-r^{2\Delta t_2}},\ldots,\ 1+\frac{r^{2\Delta t_{n_{\text{events}}-1}}}{1-r^{2\Delta t_{n_{\text{events}}-1}}}+\frac{r^{2\Delta t_{n_{\text{events}}}}}{1-r^{2\Delta t_{n_{\text{events}}}}}, \frac{1}{1-r^{2\Delta t_{n_{\text{events}}}}}\right]$$

and off-diagonal $\left[-\frac{r^{\Delta t_1}}{1-r^{2\Delta t_1}}, \ldots, -\frac{r^{\Delta t_{n_{\text{events}}}}}{1-r^{2\Delta t_{n_{\text{events}}}}}\right]$. When $\Delta t_j = 1$ for all j, then S^{-1} is the standard AR(1) inverse correlation matrix given in equation (13). As a small example, consider a time series of length N=12, $y_0, \ldots, y_{11}$, and assume that $\{y_t|x=1\} = \{y_0, y_2, y_5, y_{11}\}$ so n_{events}=4,

$$S = \begin{bmatrix} 1 & r^2 & r^5 & r^{11} \\ r^2 & 1 & r^{5-2} & r^{11-2} \\ r^5 & r^{5-2} & 1 & r^{11-5} \\ r^{11} & r^{11-2} & r^{11-5} & 1 \end{bmatrix}$$

and

$$S^{-1} = \begin{bmatrix} \frac{1}{1-r^{2\cdot 2}} & -\frac{r^2}{1-r^{2\cdot 2}} & 0 & 0 \\ -\frac{r^2}{1-r^{2\cdot 2}} & 1+\frac{r^{2\cdot 2}}{1-r^{2\cdot 2}}+\frac{r^{2\cdot 3}}{1-r^{2\cdot 3}} & -\frac{r^3}{1-r^{2\cdot 3}} & 0 \\ 0 & -\frac{r^3}{1-r^{2\cdot 3}} & 1+\frac{r^{2\cdot 3}}{1-r^{2\cdot 3}}+\frac{r^{2\cdot 6}}{1-r^{2\cdot 6}} & -\frac{r^6}{1-r^{2\cdot 6}} \\ 0 & 0 & -\frac{r^6}{1-r^{2\cdot 6}} & \frac{1}{1-r^{2\cdot 6}} \end{bmatrix}.$$

Plugging in S^{-1} from (18) into (17) shows that

$$\mathbf{1}^T S^{-1} \mathbf{1} = 1 + \sum_{i=1}^{n_{\text{events}}} \frac{1-r^{\Delta t_i}}{1+r^{\Delta t_i}}$$

which yields

(19) $$\widehat{\text{Var}(\hat{\mu}_1)} = \frac{1}{\pi(1-\pi)}\left[1 + \sum_{i=1}^{n_{\text{events}}} \frac{1-r^{\Delta t_i}}{1+r^{\Delta t_i}}\right]^{-1}$$

and also the approximation $\widehat{\text{Var}(\hat{\mu}_1)} \approx \frac{1}{\pi(1-\pi)}\left[\sum_{i=1}^{n_{\text{events}}} \frac{1-r^{\Delta t_i}}{1+r^{\Delta t_i}}\right]^{-1}$ when n_{events} is large.

Combining this with (14) and (15), we have shown that

$$\frac{1}{n_{\text{events}}} \times \frac{1+r}{1-r} > \mathbf{1}^T R^{-1} \mathbf{1} > \left[\sum_{i=1}^{n_{\text{events}}} \frac{1-r^{\Delta t_i}}{1+r^{\Delta t_i}}\right]^{-1} > \mathbf{1}^T S^{-1} \mathbf{1}.$$

Because Δt_i are not known prior to conducting a study, we use the following to (mildly) overestimate the variance when planning sample sizes

(20) $$\widehat{\text{Var}(\hat{\mu}_1)} \approx \frac{1}{n_{\text{events}} \pi(1-\pi)} \frac{1+r}{1-r}.$$

In summary, we have shown that, for an AR(1) time series of binary outcomes modelled by either (6) or (11), the variances of the estimators $\hat{\mu}_1$ and $\hat{\mu}$ are (19) and (15) respectively, which are bounded above by (20) and (16) respectively, with (16) giving a good approximation to (15) when n_{events} is large. In either case, the LCL for the true log odds is

$$\text{LCL}_{\text{log-odds}} = \ln\left(\frac{s}{1-s}\right) - z_C \sqrt{\frac{v}{n_{\text{events}} s(1-s)}}.$$

The same algebraic steps shown in section A.1 can be applied to this equation to derive equations (7) and (8).

A.3 LCL for sensitivity across multiple units at the same healthcare facility

The algebra in section A.1 can be applied to derive equations (9) and (10).

A.4 Sample sizes that attain a specified level of power

Here we provide sample size formulas, alternatives to equations (5), (8) and (10), that allow one to calculate the number of events to observe to validate a monitoring system at a specified level of power 100×(1 – β)%. The null and alternative hypotheses to be tested with an upper one-sided test are:

$$H_0: \mu = \mu_0 = \ln\left(\frac{\pi_0}{1-\pi_0}\right) \Leftrightarrow H_0: \pi = \pi_0 \text{ (the true sensitivity is } \pi_0)$$

$$H_a: \mu = \mu_a = \ln\left(\frac{\pi_a}{1-\pi_a}\right) \Leftrightarrow H_a: \pi = \pi_a \text{ (the true sensitivity is } \pi_a > \pi_0)$$

Following Hajian-Tilaki (equation (6.8))[11], Obuchowski (equation (T1))[10] and Wackerly et al. (p. 509)[24], under the assumption of normality of $\hat{\mu}$, then the number of events to observe is

$$(21) \quad n_{events} = \left(\frac{z_C\sqrt{n}SE(\hat{\mu}|\mu=\mu_0) + z_{1-\beta}\sqrt{n}SE(\hat{\mu}|\mu=\mu_a)}{\mu_a - \mu_0}\right)^2.$$

When the events are independent from a single unit (see section 3.1.2), equation (21) becomes

$$n_{\text{events}} = \left(\frac{\frac{z_C}{\sqrt{\pi_0(1-\pi_0)}} + \frac{z_{1-\beta}}{\sqrt{\pi_a(1-\pi_a)}}}{\ln\left(\frac{\pi_a(1-\pi_0)}{\pi_0(1-\pi_a)}\right)}\right)^2$$

which is the number of independent events to observe from a monitoring system to assess sensitivity at 100×*C*% confidence and 100×(1 – β)% power (cf. equation (5)).

When the events are serially correlated according to an AR(1) process from a single unit (see section 3.1.3), equation (21) becomes

$$n_{\text{events}} = v\left(\frac{\frac{z_C}{\sqrt{\pi_0(1-\pi_0)}} + \frac{z_{1-\beta}}{\sqrt{\pi_a(1-\pi_a)}}}{\ln\left(\frac{\pi_a(1-\pi_0)}{\pi_0(1-\pi_a)}\right)}\right)^2$$

which is the number of AR(1) events to observe from a monitoring system to assess sensitivity at 100×*C*% confidence level and power 100×(1 – β)% power (cf. equation (8)).

When the events are serially correlated according to an AR(1) process from each of n_{units} units (see section 3.2), equation (21) becomes

$$n_{\text{events}} = \left(\frac{\frac{z_C\sqrt{v}}{\sqrt{\pi_0(1-\pi_0)}} + \frac{z_{1-\beta}\sqrt{v}}{\sqrt{\pi_a(1-\pi_a)}} + (z_C + z_{1-\beta})\sqrt{V_{\text{units}}}}{\frac{4\sqrt{n_{\text{units}}}}{5}\ln\left(\frac{\pi_a(1-\pi_0)}{\pi_0(1-\pi_a)}\right)}\right)^2$$

which is the number of AR(1) events to observe from a monitoring system at each unit to assess sensitivity at 100×*C*% confidence level and power 100×(1 – β)% power across the entire healthcare facility (cf. equation (10)).