tvst

DOI: 10.1167/tvst.6.3.1

Article

Does Testing More Frequently Shorten the Time to Detect Disease Progression?

Johannes Ledolter[1] and Randy H. Kardon[2]

[1] Departments of Management Sciences/Statistics & Actuarial Science, University of Iowa, Iowa City, IA, USA
[2] Department of Ophthalmology and Visual Sciences, University of Iowa Hospital and Clinics and Iowa City VA Medical Center, Iowa City, IA, USA

Correspondence: Johannes Ledolter, Tippie College of Business, S352 Pappajohn Business Building, University of Iowa, Iowa City, IA 52242, USA. e-mail: johannes-ledolter@uiowa.edu

Received: 11 October 2016
Accepted: 17 March 2017
Published: 1 May 2017

Keywords: trend change detection; patient monitoring; sampling frequency; autocorrelation; statistical power

Citation: Ledolter J, Kardon RH. Does testing more frequently shorten the time to detect disease progression? Trans Vis Sci Tech. 2017;6(3):1, doi:10.1167/tvst.6.3.1

Purpose: With the rise of smartphone devices to monitor health status remotely, it is tempting to conclude that sampling more often will provide a more sensitive means of detecting changes in health status earlier over time, when interventions may improve outcomes.

Methods: The answer to this question is derived in the context of a model where observations are generated from a linear-trend model with independent as well as autocorrelated autoregressive-moving average, or ARMA(1,1), errors.

Results: The results imply a cautionary message that an increase in the sampling frequency may not always lead to a faster detection of trend changes. The benefit of rapid successive observations depends on how observations, taken closely together in time, are correlated.

Conclusions: Shortening the observation period by half can be accomplished by increasing the number of independent observations to maintain the same power for detecting change over time. However, a strategy to detect progression of disease sooner by taking numerous closely spaced measurements over a shortened interval is limited by the degree of autocorrelation among adjacent observations. We provide a statistical model of disease progression that allows for autocorrelation among successive measurements, and obtain the power of detecting a linear change of specified magnitude when equal-spaced observations are taken over a given time interval.

Translational Relevance: New emerging technology for home monitoring of visual function will provide a means to monitor sensory status more frequently. The model proposed here takes into account how successive measurements are correlated, which impacts the number of measurements needed to detect a significant change in status.

Introduction

In medical surveillance, patient characteristics are assessed at consecutive clinic visits with the intent to detect medically relevant trend changes among consecutive observations. In the absence of disease, measurements tend to fluctuate around a stable level, but often deteriorate progressively once the disease has set in. Because of cost considerations and burden of travel for patients, assessments for common eye conditions that progress, such as glaucoma, are frequently carried out at 6-month intervals. Other conditions, such as macular degeneration, may be monitored monthly, once intravitreal injections of anti-vascular endothelial growth factor are administered for treatment of the wet variety of macular degeneration. The objective is to detect a medically meaningful trend change within as short of a time period as possible so that adjustments in treatment can be made to prevent further damage.

The development of inexpensive smartphone monitoring systems that can be used by the patient at home raises the question of whether a medically relevant trend change could be detected sooner if observations are taken more frequently, and perhaps over a shorter time period.[1,2] As an illustration, assume that the clinician takes observations of visual function or structure every 6 months wishing to detect a trend change of a certain magnitude within 3 years

(that is, $n = 7$ observations). This is a strategy used by many clinicians, for conditions such as glaucoma that can slowly progress if not treated optimally. The sampling frequency is limited by the resources available to accommodate more follow-up visits per year and the burden placed on patients who are asked to return more frequently for follow-up measurements. Of course, it would seem to be more beneficial to increase the sampling frequency and take observations every 3 months, every month, every week, and so on, if it would help detect a trend earlier or with more certainty, when a treatment intervention may lead to preservation of vision. Increasing the sampling frequency to improve the ability to detect a change over a given sampling period would certainly bring benefits, as long as the observations we collect are statistically independent, with little correlation between measurements at adjacent time intervals.

A second competing strategy could take the same number of observations, but over just the first half of the observation interval (e.g., during the first 18 months of the 3-year interval), hence reducing the time between successive visits by half. Better yet, if measurements could be collected on a portable device at home with an increase in sampling frequency, then could the observation period over which a relevant change could be detected be shortened even further? The question arises whether this second strategy would provide similar power of detecting a meaningful change over a shorter time period. Furthermore, if there is little cost to taking observations, the sampling frequency over this new shorter observation interval could be increased even further; in theory, patients could take daily or evenly hourly measurements on their state of health. Would this be better? If this were the case then one could, by taking numerous observations in rapid succession, detect a relevant change even sooner.

In this note, we discuss the advantages and disadvantages of such an approach, and we address the question whether home-based surveillance methods currently being developed to sample more frequently can lead to detection of medically relevant trend changes over a shorter time period.

Materials, Methods, and Results

Trend Model with Independent Observations

In order to shed light on such questions, one needs a statistical model. We let time vary over the unit-time interval [0, 1] and let n equally spaced measurements be generated from the model

$$y_i = \alpha + \beta t_i + \varepsilon_i = T_i + \varepsilon_i \quad \text{for } i = 1, 2, \ldots, n \qquad (1)$$

with $t_i = (i - 1)/(n - 1)$. Equation 1 describes a stationary linear-trend model. The trend component $T_t = \alpha + \beta t$ expresses the time progression of the measurement and ε_t is a measurement error with mean 0 and variance σ^2. In this section we assume that the errors ε_t are independent. This is reasonable provided there is no instrument carry-over from one measurement to the other and the linear trend is indeed deterministic. In the following section, we allow for autocorrelations among the errors. Autocorrelation in the errors can arise from instrument carry-over, but also because the progression of many anthropometric signals is not purely deterministic, but also affected by stochastic perturbations.

Our main interest is in detecting a change in the trend progression β. What is the power of detecting a specified slope change with just $n = 7$ observations equally spaced on the unit-time interval? How does the power change if we take more than seven observations over the same unit-time interval? And what are the consequences of reducing the observation period in half (or to any other fraction of the unit-time interval) and taking seven (or more) observations over the reduced time interval? We start with the obvious: By restricting attention to only the first half of the time interval, we do not obtain observations from the second half of the interval. We lose the opportunity to learn whether there are changes to the trend during the second half of the interval. Trends are typically not stable, and an approach that looks at only part of the time interval certainly limits one's ability to check for changing trends. Assuming that changes in the trend are constant over time is certainly a very strong assumption.

However, for the purpose of this paper, we assume that the slope is constant across the unit-time interval. We investigate the effects of reducing the observation interval, but increasing the sampling frequency. We derive an expression for the statistical power of detecting a change in the slope, from baseline value β_{Base} to the new value $\beta_{\text{Base}} + \beta_*$, for known significance level α (usually, 0.05) and error standard deviation σ.

We write the regression model in Equation 1 in its mean-corrected form, $Y_i = \bar{Y} + \beta(t_i - \bar{t}) + \varepsilon_i$ with $t_i = (i - 1)/(n - 1)$ for $i = 1,2,\ldots n$, and $\bar{t} = 1/2$. We consider the test of $H_0 : \beta = \beta_{\text{Base}}$ against $H_1 : \beta = \beta_{\text{Base}} + \beta_*$,

with $\beta_* > 0$. The standard error of the least squares estimate $\hat{\beta}$ is given by

$$\sigma_{\hat{\beta}} = \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(t_i - \bar{t})^2}} = \frac{\sigma}{\sqrt{n(n+1)/12(n-1)}};$$

see Abraham and Ledolter[3] for the standard error; the last expression follows from straightforward algebra. The critical limit for the hypothesis test is $CL = \beta_{\text{Base}} - z_\alpha \sigma_{\hat{\beta}}$, where z_α is the percentile of the standard normal distribution; for significance level $\alpha = 0.05$, $z_\alpha = -1.645$. The power of the test is given by

$$\begin{aligned}\text{Power} &= P\left[\hat{\beta} \geq CL | \beta = \beta_{\text{Base}} + \beta_*\right] \\ &= P\left[Z \geq \frac{\beta_{\text{Base}} - z_\alpha \sigma_{\hat{\beta}} - (\beta_{\text{Base}} + \beta_*)}{\sigma_{\hat{\beta}}}\right] \\ &= P\left[Z \geq -z_\alpha - \frac{\beta_*}{\sigma_{\hat{\beta}}}\right] = \Phi\left(z_\alpha + \frac{\beta_*}{\sigma_{\hat{\beta}}}\right) \\ &= \Phi\left(z_\alpha + \frac{\beta_*}{\sigma}\sqrt{\frac{n(n+1)}{12(n-1)}}\right), \qquad (2)\end{aligned}$$

where $\Phi(.)$ is the cumulative distribution function of the standard normal distribution.

Reducing the observation interval and taking n equally spaced observations on the interval $[0, P < 1]$ changes the power to

$$\text{Power} = \Phi\left(z_\alpha + P\frac{\beta_*}{\sigma}\sqrt{\frac{n(n+1)}{12(n-1)}}\right) \qquad (3)$$

The ratio β_*/σ is a critical parameter; the power decreases if smaller changes need to be detected. For $\alpha = 0.05$, $P = 1$ (using the full time interval, such as 3 years), $\beta_*/\sigma = 2.5$, and $n = 7$, the power is 0.712. Given measurement variability $\sigma = 0.4$, seven equally spaced observations over the full 3-year time interval allow us to detect an increase of one unit; detecting a smaller change, such as a half unit change over three years, with just seven observations is almost impossible (power 0.294). If we took $n = 15$ observations over the full time period, the power of detecting an increase of one unit is larger (0.910), and we are fairly certain to detect a change of that magnitude. This is expected, as more observations are always better than fewer.

Next, let us assume that we want to make a decision within the first half of the observation period of 3 years, and do so with the same number of observations ($n = 7$), which now are equally spaced over the first 1.5 years. Using $P = 0.5$, Equation 3 results in power 0.294. As expected, this power is smaller than the one we get when spreading the seven observations over the whole 3 years, as it is equivalent to the power of detecting half of the change, $(0.5)\beta_*$, over the full time interval (Equation 3). Of course, one can increase this unacceptable low power by adding more observations and sampling more often. For example, with $n = 15$ observations equally spaced over the first 1.5 years, the power is 0.440. Figure 1 shows that we need 36 observations to achieve the same power (0.712) that we obtain with seven observations equally spaced over 3 years.

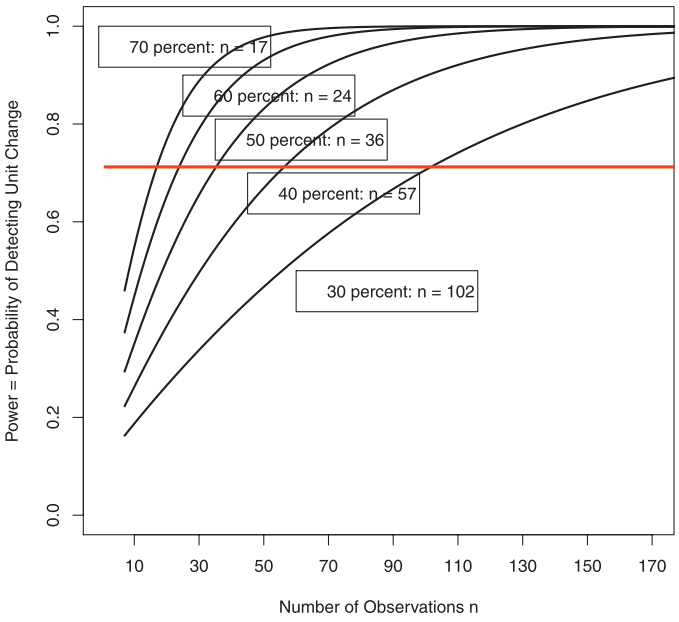

Figure 1. Independent observations. Reduction of the testing period to 70%, 60%, 50% (18 months), 40%, and 30% (~12 months) of the original 36-month (3-year) period. In order to obtain the same power that is achieved with seven observations over the full 36-month time interval (0.712, as indicated by the *red line*), 17 observations are needed if the sampling period is reduced to 70% of the full 36 months, 24 observations are needed if the sampling period is reduced to 60% of the full 36 months, 36 observations are needed if the sampling takes place over only the first 18 months (reduction of the sampling period to 50% of the full 36 months), and 102 observations are needed if sampling is restricted to the first 12 months (reduction the sampling period to 30% of the full 36 months).

What about reducing the interval even further, to just 3/10 of the original 3 years ($P = 0.3$), but increasing the number of observations over this short time period even more? The results in Figure 1 show that we need 102 independent observations to attain the same power (0.712) that we obtain with $n = 7$ observations spaced evenly over 3 years. It takes more independent observations to compensate for the

Table. Results of the Simulation Study: False Rejection Probabilities and Power of the Two Strategies.

	Strategy 1		Strategy 2	
	False Rejection	Power	False Rejection	Power
$k = 2$	0.053	0.811	0.142	0.839
$k = 1$	0.049	0.296	0.140	0.388

Single test (Strategy 1) vs. repeated tests on successive observations (Strategy 2).

reduced observation interval, but the same power can always be achieved by increasing the sample frequency.

Increasing the sampling frequency certainly strengthens the reliability of a trend estimate, but also increases the danger of spurious observations. Identification and removal of outlier observations from a trend estimate will minimize the risk of responding to measurements when no response is needed. In current in-clinic evaluations of visual function that are sampled every 6 months, the clinician looks at every single observation, and with more and more observations there is a tendency to make treatment changes on the basis of the unusual recent results even when no changes are warranted. Often interventions are made when an observation exceeds its 2- or 3-sigma bounds. While the chance of being outside the 2-sigma limits is 5% for a single observation, the chance that we observe one of say 20 (independent) observations outside the 2- sigma limits is quite high (38%, using the binomial distribution).

The following simulation study illustrates this issue in the regression context when we test whether a meaningful change has occurred over a 3-year period. For our simulations, we assume that the mean of the observations changes linearly from baseline 0 to the value $k\sigma$ after a period of 3 years, for both $k = 1$ and $k = 2$. Independent normal observations with standard deviation σ are generated every 6 months. Two different estimation and testing strategies are compared. Strategy 1 uses (only) the seven observations that are available at the end of the 3-year period to test whether the slope of the linear regression through the origin has increased (using significance level 0.05). Strategy 2 starts the testing after the first four observations have been collected (i.e., after having observed the measurement at 18 months) and repeats the test with each successive observation, for a total of four tests. It concludes a change in the slope when one or more of these tests reject the no-change hypothesis.

We conduct 10,000 simulations for the null hypothesis when no change has occurred, with the proportion of rejections representing the error of a false rejection of a true null. We do the same for the alternative, with the proportion of rejections representing the power of the detection procedure. The results are given in the Table. Without adjustments for multiple testing, repeated tests carried out on frequent readily available measurements will increase the false rejection (in our simulation from 5% to 14%), leading clinicians to make treatment changes when no changes are warranted. Adjustments for multiple testing are essential as more and more data can be sampled frequently and become more readily available in a home testing scenario.

Materials, Methods, and Results

Linear Trend Model with Autocorrelated Observations

The trend model in Equation 1 assumes a deterministic linear trend and independent measurement errors. Independent measurement errors are reasonable as instrument carry-over from one measurement error to the next is unlikely. However, the progression of the anthropometric signal $T_t = \alpha + \beta t$ is often not purely deterministic but also affected by stochastic perturbations r_t that lead to persistent slow-moving deviations from the deterministic linear trend, analogous to a slow moving wave. Persistence implies that a signal at time t above the trend line tends to be followed by signals that are above the trend line as well. In other words, signals tend to stay above (or below) the trend line for several periods in a row. Such persistence can be modeled with a first-order autoregressive model, $r_t = (1/(1 - \varphi B))\ \xi_t = \xi_t + \varphi\xi_{t-1} + \varphi^2\xi_{t-2} + \ldots$. Here, B is the backshift operator, φ is the autoregressive parameter (which, for statistical stationarity, has to be between -1 and 1), and ξ_t are independent mean zero random variables with variance σ_ξ^2. The first-order autoregressive model for r_t implies autocorrelations $Cor(r_t, r_{t-k}) = \varphi^k$ and variance $\sigma_r^2 = \frac{\sigma_\xi^2}{1-\varphi^2}$. Persistence is achieved when the autoregressive parameter φ is positive and close to 1. The autoregressive model becomes the (nonstationary) random walk when $\varphi = 1$. A random walk can

take very long persistent excursions from the deterministic trend line. For a detailed discussion of time series models (including the backshift operator notation, stationarity and nonstationarity, and autoregressive and moving average models) we refer the reader to Abraham and Ledolter[4] and Box et al.[5]

Incorporating anthropometric persistence into the trend model leads to the following more realistic model of change, $Y_t = \alpha + \beta t + r_t + \varepsilon_t$. Subtracting the deterministic linear trend from the measurements, leads to trend deviations:

$$\tilde{Y}_t = Y_t - (\alpha + \beta t) = r_t + \varepsilon_t = \frac{1}{(1 - \varphi B)}\xi_t + \varepsilon_t \quad (4)$$

The model for the trend deviations can be written as $(1 - \varphi B)\tilde{Y}_t = \xi_t + (1 - \varphi B)\varepsilon_t$, and is known as the autoregressive-moving average, or ARMA(1,1), model: there is just one lagged autoregressive term and the autocorrelations of the moving average component on the right-hand side of the model are zero after lag 1. It is straightforward to show that the standard deviation and the autocorrelations of the deviations from the linear trend model $\tilde{Y}_t = Y_t - (\alpha + \beta t)$ are $\sigma_{\tilde{Y}} = \sigma = \sqrt{\sigma_r^2 + \sigma_\varepsilon^2}$, $\rho_1 = \varphi\frac{1}{1+(1-\varphi^2)\left(\sigma_\varepsilon^2/\sigma_\xi^2\right)} = \varphi\frac{1}{1+\left(\sigma_\varepsilon^2/\sigma_r^2\right)}$, and $\rho_k = (\rho_1)\ \varphi^{k-1}$ for $k \geq 1$. For $\sigma_\varepsilon^2 = 0$ (when there is no measurement error), the ARIMA(1,1) model simplifies to the first-order autoregressive model with variance σ_r^2 and autocorrelations $\rho_k = \varphi^{k-1}$.

Persistence is modeled through the autoregressive parameter, and let us assume $\varphi = 0.8$. The ratio $\sigma_\varepsilon^2/\sigma_r^2$ compares the variance of the independent measurement errors with the variance of the persistent stochastic trend movements. We assume variance ratio $\sigma_\varepsilon^2/\sigma_r^2 = 3$ as the stochastic trend component should not deviate too much from the deterministic linear trend and most of the variability should come from the measurement noise. With these choices of parameters the autocorrelations of $\tilde{Y}_t = Y_t - (\alpha + \beta t)$ are $\rho_1 = 0.8/(1+3) = 0.2$ and $\rho_k = (0.2)(0.8)^{k-1}$ for $k \geq 1$. While the lag 1 autocorrelation is moderate in size ($\rho_1 = 0.2$), there is a persistent slow decay in the autocorrelations from lag 1 onward.

We have provided motivation why the ARMA(1,1) is a useful error model for trend regressions. There is also evidence in the literature[6,7] that errors in regressions of anthropometric time series data on deterministic functions of age follow ARMA(1,1) models. Carrico et al.[7] show that, in a regression of young-adult blood pressure on linear and quadratic functions of age, body mass index, and height, ARMA(1,1) errors are preferable to AR(1) and errors with compound symmetry.

Our new model:

$$Y_i = \bar{Y} + \beta(t_i - \bar{t}) + \varepsilon_i \quad \text{with } t_i = (i-1)/(n-1) \text{ for } i = 1, 2, ..., n \quad \text{and } \bar{t} = 1/2 \quad (5)$$

assumes that the errors ε follow an ARMA(1,1) model, implying an $n \times n$ error covariance matrix V with elements $v_{ij} = \sigma^2$ for $i = j$ and $v_{ij} = \sigma^2\rho_1\varphi^{|i-j|-1}$ for $i \neq j$. The generalized least squares (GLS) estimator of β in the model in Equation 5 is given by $\hat{\beta}_{GLS} = (X^TV^{-1}X)^{-1}X^TV^{-1}(Y - \bar{Y})$. Here V is the $n \times n$ covariance matrix specified above, $X = (t_1 - \bar{t}, t_2 - \bar{t}, \ldots, t_n - \bar{t})^T$ is the $n \times 1$ column vector of times, and $Y - \bar{Y} = (Y_1 - \bar{Y}, Y_2 - \bar{Y}, \ldots, Y_n - \bar{Y})^T$ is the $n \times 1$ column vector of mean-corrected observations. The superscript ^{T} denotes the transpose. The GLS estimator is the most efficient estimator among all linear unbiased estimators, with the smallest sample variance $\sigma^2_{\hat{\beta}_{GLS}} = (X^TV^{-1}X)^{-1}$.[3] Substituting this standard error into Equations 2 and 3 leads to the power

$$\text{Power} = \Phi\left(z_\alpha + P\frac{\beta_*}{\sigma}\sigma_{\hat{\beta}_{GLS}}\right) \quad (6)$$

We return to our example with $z_{0.05} = -1.645$, $\beta_* = 1$, $\sigma = 0.4$ and an observation interval that is reduced from the original 3 years ($P < 1$), but now assume that the error is characterized by the ARMA(1,1) model with weekly autoregressive coefficient $\varphi_W = 0.8$ and variance ratio $\sigma_\varepsilon^2/\sigma_r^2 = 3$. The n observations on the reduced unit-time interval [0, $P < 1$] are spaced $156P/(n-1)$ weeks apart. Hence, the autoregressive coefficient between successive observations is $(\varphi_W)^{156P/(n-1)}$. This value, $\sigma = 0.4$ and the variance ratio $\sigma_\varepsilon^2/\sigma_r^2 = 3$ are used for the calculation of the covariance matrix V of the n observations equally spaced over the interval [0, $P < 1$].

With the original 3-year observation interval ($P = 1$) and $n = 7$ observations, the power calculated from Equation 6 with $\varphi_W = 0.8$ is still 0.712, the same power we obtain when there is independence. This is because observations are 26 weeks apart ($156P/(n-1) = 156(1)/6 = 26$), $\varphi_W^{26} \approx 0$, and V is a diagonal matrix with zero autocorrelations. The power is affected only for much larger weekly autoregressive coefficient very close to 1.

Figure 2 shows results for the ARMA(1,1) model with weekly autoregressive coefficient $\varphi_W = 0.8, \sigma = 0.4$ and variance ratio $\sigma_\varepsilon^2/\sigma_r^2 = 3$. In order to obtain

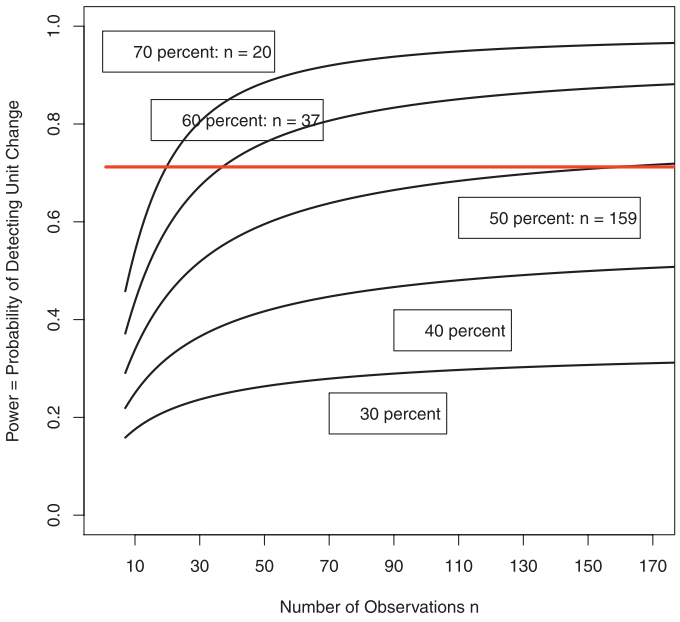

Figure 2. ARMA(1,1) correlation between subsequent errors with weekly autoregressive coefficient $\varphi_w = 0.8$ and variance ratio $\sigma_\varepsilon^2/\sigma_r^2 = 3$. Reduction of the testing period to 70%, 60%, 50% (18 months), 40%, and 30% (~12 months) of the original 36-months (3-year) period. In order to obtain the same power that is achieved with seven observations over the full 36-month time interval (0.712, as indicated by the *red line* in Fig. 1), 20 observations are needed if the sampling period is reduced to 70% of the full 36 months, 37 observations are needed if sampling is reduced to 60% of the full 36 months, 159 observations are needed if sampling is reduced to 50% of the full 36 months. Many more observations are needed for larger reductions of the original observation period.

the same power that is achieved with seven observations over the full 36-month time interval (0.712), 20 observations are now needed if the sampling period is reduced to 70% of the full 36 months. This is larger than the 17 observations that are needed in the independent case. Thirty-seven observations are needed if sampling is reduced to 60% of the full 36 months, which is larger than the 24 observations in the independent case. For an observation window that is cut in half we find that it takes more (correlated) observations to compensate for the shortened observation interval. Now 159 observations are needed to attain the same power, instead of the 36 independent observations.

When reducing the time interval to 40% (or even less) of the original time interval, the increase in the number of correlated observations that are needed becomes very large and cannot compensate for the shortened observation interval, as adjacent observations over the reduced interval are now so close together that their autocorrelations approach 1. This implies that there is no benefit to taking such extra observations. For an observation interval reduced to 30%, we limit ourselves to $(156)(0.3) = 46.8$ weeks. With $n = 100$, for illustration, adjacent observations are 0.46 weeks apart and $(\varphi_W)^{0.46} = (0.8)^{0.46} = 0.90$. Off-diagonal elements in the covariance matrix V are large, which indicates that there is little benefit to collecting observations that are so close together in time.

A Comment on the Choice of Parameters in Figure 2

Persistence parameter $\varphi = 0.8$ and variance ratio $\sigma_\varepsilon^2/\sigma_r^2 = 3$ imply a moderate initial autocorrelation $\rho_1 = \varphi/4 = 0.20$ with subsequent slow decay. For smaller variance ratios, when trend changes dominate independent measurement errors, our calculations show that increases in the number of required samples are even larger than the ones reported in Figure 2. Larger variance ratios decrease the autocorrelations. For $\sigma_\varepsilon^2/\sigma_r^2 = 9$, the autocorrelations start their exponential decay from $\rho_1 = \varphi/10 = 0.08$, and the increases in the number of required samples become smaller. We need 18 samples when the sampling period is reduced to 70% of the initial interval (20 samples are required when $\sigma_\varepsilon^2/\sigma_r^2 = 4$, and 17 samples are required when observations are independent), 28 samples when the sampling period is reduced to 60% (37 samples when $\sigma_\varepsilon^2/\sigma_r^2 = 4$ and 24 samples when independent), and 49 samples when the sampling period is reduced to 50% (159 samples when $\sigma_\varepsilon^2/\sigma_r^2 = 4$ and 36 samples when independent). The increase depends on the parameters in the ARMA(1,1) model. If prior data are available, we recommend to calculate maximum likelihood estimates of all parameters in the linear-trend model with ARMA(1,1) errors. The estimates allow us to obtain (1) the resulting increase in the required sample size, and (2) the correct standard error of the trend estimate β. This is important, as standard errors derived under independence are incorrect if errors are autocorrelated. For positive autocorrelation standard errors assuming independence are too small, which leads to spurious significance.[8,9]

Discussion

The paper by Crabb and Garway-Heath[10] is related to this discussion. They investigate, through simulations, whether it is better to collect more observations at the beginning and at the end of the observation period ("wait and see" approach) than to

space the observations evenly throughout the observation window. Their finding that the power of detecting a change is increased with the "wait and see" strategy can be predicted from theory without any simulations, as the standard error of a slope estimate $\sigma_{\hat{\beta}}$ becomes smaller when the settings of the regression predictor are located at the boundary of the experimental region; see Materials, Methods, and Results. Our paper derives results theoretically and also allows for correlation among adjacent observations.

An important reason against shortening the observation period is that such approach will miss changes in the trend that occur within the interval that is not being monitored. But even if the trend stays the same, a strategy of increased sampling of correlated measurements over a shortened interval may not generate the same amount of information as does the traditional approach of obtaining measurements every 6 months. Because of autocorrelation among adjacent observations, the benefit of taking numerous closely spaced measurements may be exaggerated. For autocorrelated observations, an increase in the sampling frequency may not compensate for a shortened observation interval. However, if the autocorrelation is observed to be low for a certain type of measurement (e.g., the pupil light reflex), there may be significant benefits for increasing the sampling frequency over a shortened observation interval.

For biological measurements, especially those used in clinical medicine, where a trend analysis may be critical to understanding whether a treatment intervention is recommended, it is important to consider how correlated adjacent measurements are to one another in order to design the optimal sampling interval for the problem being monitored. One strategy would be to determine the autocorrelations empirically by taking measurements across subjects at different time intervals, allowing us to check the ARMA(1,1) model assumption in Equation 5. Knowing the autocorrelation value, the desired amount of change that one wants to detect, and the power and level of statistical significance desired will inform potential developers on an optimal sampling strategy that will take advantage of telemedical devices.

Our paper assumes that the variability of all observations is the same and the variance of measurements taken at home will be the same as in a supervised clinic environment. This would, of course, depend on the type of measurement being evaluated. Behavioral tests of vision, such as visual field sensitivity are influenced by the cognitive ability of the subject, distractions that may be present, and the effort being put forth on the day of testing and may not be the same at home compared with a monitored clinical setting. On the other hand, objective measurements of vision, such as the pupil light reflex or evoked potentials from the visual system may be more ideally suited toward at-home testing in an unsupervised, but familiar environment. Better yet, an automated, real-time, built in video-based monitoring of a patient's behavior during home testing may provide a type of behavior supervision that would optimize behavioral tests of visual function. For independent observations, the standard error of an average is obtained by dividing the standard deviation of a single measurement by the square root of the number of observations. Hence, with doubling the standard deviation of individual measurements, the sample size of the less-precise group must be increased by a factor of 4 in order to obtain the same precision and power. Therefore, objective readouts of visual function that have low repeat measurement variability would be prime candidates for at-home frequent testing over time to detect the status of ocular diseases.

Our paper shows that the shrinkage of the observation period by half can be compensated by increasing the number of independent observations so that the powers of the two strategies are the same (Fig. 1). The result is not surprising, as one knows that even the smallest change over a short-time window becomes significant if one increases the number of observations. But the result is highly dependent on how much autocorrelation exists between successive measurements (Fig. 2). It also requires the very strong assumption that the progression of the condition is constant. But in many types of disease, it is unlikely that the progression of disease is constant, and it is more plausible that progression trends are stochastic. Progression changes as time goes on, with some periods when the progression is rather flat, and other periods when conditions change quickly. Hence, it would be a misguided approach in the case where progression may vary significantly from one measurement surveillance period to the next to focus on ever-smaller time windows and to compensate with more frequent measurements over these small windows. Clinicians know about meaningful changes over a period of 3 years, but they know much less about changes that can be expected over brief periods in slowly progressive disorders, such as glaucoma or multiple sclerosis. The promise of personalized medicine with optimal monitoring of

disease being enabled by frequent measurements with a home-based system is within reach with new technological advances of monitoring devices. However, the promise is more likely to become a reality if the patient is monitored indefinitely during the course of their disease to better understand patterns of progressive change and when treatment interventions are warranted for an individual patient.

Even if one could assume that trends are deterministic – which is unlikely – a strategy of taking more observations over shorter time windows has its challenges, even if such observations are independent. Our simulation results in the Table illustrate that appropriate adjustments for multiple testing are needed when analyzing and interpreting abundant successive measurements, to prevent the increased risk of diagnosing a change in the status of the disease being studied when one really does not exist.

We show that a strategy to detect progression sooner by taking closely spaced measurements over a shortened interval is limited by the degree of autocorrelation. We show this result by studying the linear-trend model with ARMA(1,1) errors, but have observed the same finding for first-order autoregressive errors in the linear-trend model as well as in the proportional linear-trend model with constant coefficient of variation. An important conclusion of our study is that one should always check whether errors are indeed independent and incorporate any serial correlation when making inferences about the trend parameters.

Accurate monitoring of progression is essential for patient management. The standard model considered in the literature combines a structural component that postulates a linear time progression and a noise component that specifies uncorrelated errors. Recently, several papers have started to question these assumptions and have proposed more general models, both for the structural time progression and for the noise component.

In their analysis of longitudinal perimetry data from glaucoma patients, Pathak et al.[11] propose a structural model for the progression that includes the exponential of time and an autoregressive noise component that allows for the temporal correlation among adjacent observations. In their analysis of longitudinal cardiac imaging data, George et al.[12] consider several models for the temporal and the spatial correlations that can be expected across time and across different image locations. Lawton et al.[13] consider a longitudinal model for disease progression of multiple sclerosis patients. They argue quite convincingly that the structural regression component should not merely include linear time trends, but also fractional polynomials of time t, such as $\log(t)$ and $\text{sqrt}(t)$. In addition, their models include parameters for the autocorrelation among adjacent observations. Taketani et al.[14] study how to best predict for a given glaucoma patient his/her response at a future visit. Their model includes nonlinear components for the time progression (quadratic, exponential, and logistic terms of time), and they consider alternatives to the standard least squares estimation by considering robust statistical estimation methods. The study by Chan et al.[15] makes a convincing argument that longitudinal studies (in their application, the movement of a subject's arm over time) must generalize the noise component to allow for possible temporal correlation. Allowing for autocorrelation helps avoid a common mistake of adopting spurious results regarding the structural progression component of the model.[3]

Our paper reflects these new developments in modeling longitudinal data as our model allows for temporal correlation among the observations. We recognize the importance of studying how time trends change over time.

Acknowledgments

The authors thank the two referees and the associate editor for helpful comments.

Supported in part from the VA Rehabilitation Research and Development Center Grant C9251-C, C1786-R, 2I01 RX000889-05A2, 1IO1 RX002101, Chronic Effects of Neurotrauma Consortium CENC0056P (VA Rehabilitation Research and Development and DOD). Also, grants from NEI 1R01EY023279-01, R009040554, and the Department of Defense, CDMRP W81XWH-10-1-0736, W81XWH-11-1-0561, W81XWH-16-1-0071, and W81XWH-16-1-0211.

Disclosure: **J. Ledolter**, None; **R.H. Kardon**, None

References

1. Chew EY, Clemons TE, Bressler SB, et al. Randomized trial of a home monitoring system for early detection of choroidal neovasculariza-

tion. Home Monitoring of the Eye (HOME) study. *Ophthalmology*. 2014;121:535–544.
2. Winther C, Frisen L. Self-testing of vision in age-related macula degeneration: a longitudinal pilot study using a smartphone-based rarebit test. *J Ophthalmol*. 2015;2015: http://dx.doi.org/10.1155/2015/285463.
3. Abraham B, Ledolter J. Introduction to Regression Modeling. Belmont: Thompson Higher Education; 2006.
4. Abraham B, Ledolter J. Statistical Methods for Forecasting. New York: Wiley; 1983.
5. Box GEP, Jenkins GM, Reinsel GC. Time Series Analysis, Forecasting and Control. 3rd ed. New York: Prentice Hall; 1994.
6. Beckett LA, Rosner B, Roche AF, Guo S. Serial changes in blood pressure from adolescence into adulthood. *Am J Epidemiol*. 1992;135:1166–1177.
7. Carrico RJ, Sun SS, Sima AP, Rosner, B. The predictive value of childhood blood pressure values for adult elevated blood pressure. *Open J Pediatr*. 2013;3:116–126.
8. Crabb DP, Garway-Heath DF. Intervals between visual field tests when monitoring the glaucomatous patient: wait-and-see approach. *Invest Ophthalmol Vis Sci*. 2012;53:2770–2776.
9. Box GEP, Newbold P. Some comments on a paper by Coen, Gomme and Kendall. *J R Stat Soc*. 1971;A134:229–240.
10. Granger CWJ, Newbold P. Spurious regressions in econometrics. *J Econom*. 1974;2:111–120.
11. Pathak M, Demirel S, Gardiner SK. Nonlinear, multilevel mixed-effects approach for modeling longitudinal standard automated perimetry data in glaucoma. *Invest Ophthalmol Vis Sci*. 2013;54:5505–5513.
12. George B, Denney T, Gupta H, Dell'Italia L, Aban I. Applying a spatiotemporal model for longitudinal cardiac imaging data. *Ann Appl Stat*. 2016;10:527–548.
13. Lawton M, Tilling K, Robertson N, et al. A longitudinal model for disease progression was developed and applied to multiple sclerosis. *J Clin Epidemiol*. 2015;68:1355–1365.
14. Taketani Y, Murata H, Fujino Y, Mayama C, Asaoka R. How many visual fields are required to precisely predict future test results in glaucoma patients when using different trend analyses? *Invest Ophthalmol Vis Sci*. 2015;56:4076–4082.
15. Chan MF, Giddings DR, Chandler CS, Craggs C, Plant RD, Day MC. An experimentally confirmed statistical model on arm movement. *Hum Mov Sci*. 2004;22:631–648.