Toward Rigorous Idiographic Research in Prevention Science: Comparison Between Three Analytic Strategies for Testing Preventive Intervention in Very Small Samples

Ty A. Ridenour · Thomas Z. Pineo ·
Mildred M. Maldonado Molina ·
Kristen Hassmiller Lich

Abstract Psychosocial prevention research lacks evidence from intensive within-person lines of research to understand idiographic processes related to development and response to intervention. Such data could be used to fill gaps in the literature and expand the study design options for prevention researchers, including lower-cost yet rigorous studies (e.g., for program evaluations), pilot studies, designs to test programs for low prevalence outcomes, selective/indicated/adaptive intervention research, and understanding of differential response to programs. This study compared three competing analytic strategies designed for this type of research: autoregressive moving average, mixed model trajectory analysis, and P-technique. Illustrative time series data were from a pilot study of an intervention for nursing home residents with diabetes ($N=4$) designed to improve control of blood glucose. A within-person, intermittent baseline design was used. Intervention effects were detected using each strategy for the aggregated sample and for individual patients. The P-technique model most closely replicated observed glucose levels. ARIMA and P-technique models were most similar in terms of estimated intervention effects and modeled glucose levels. However, ARIMA and P-technique also were more sensitive to missing data, outliers and number of observations. Statistical testing suggested that results generalize both to other persons as well as to idiographic, longitudinal processes. This study demonstrated the potential contributions of idiographic research in prevention science as well as the need for simulation studies to delineate the research circumstances when each analytic approach is optimal for deriving the correct parameter estimates.

Keywords ARIMA · Diabetes · Blood glucose · Indicated prevention · Mixed model trajectory analysis · P-technique · Selective prevention · Sliding scale

T. A. Ridenour (✉)
Department of Pharmaceutical Sciences,
Center for Education and Drug Abuse Research,
University of Pittsburgh,
3520 Forbes Avenue, 2nd Floor, Room 226,
Pittsburgh, PA 15213, USA
e-mail: tar27@pitt.edu

T. Z. Pineo
University of Pittsburgh Medical Center,
Pittsburgh, PA, USA

M. M. Maldonado Molina
University of Florida,
Gainesville, FL, USA

K. Hassmiller Lich
The University of North Carolina at Chapel Hill,
Chapel Hill, NC, USA

Prevention science focused on psychosocial outcomes has relied almost exclusively on large sample randomized designs, consistent with its traditional emphasis on universal intervention (Biglan et al. 2000; Ennett et al. 2003). These methods have been used well to advance prevention in certain ways such as estimating the efficacy of a prevention program. Further diversification of prevention methodologies could facilitate much needed statistical, theoretical, and applied innovations (Biglan et al. 1996; Kazdin and Blase 2011; Molenaar 2004). This paper demonstrates three analytic strategies with promise for rigorous prevention research that uses within-person, small sample or case study methods. Each strategy can be used in tandem with innovative designs recently introduced to prevention science such

as adaptive designs or Systems Dynamics modeling. These small sample designs offer low-cost alternatives to large sample studies, filling a critical need in prevention science during economic downturns and in settings with tight budgets (e.g., schools). To underscore a distinguishing emphasis of these designs compared to traditional large-sample clinical trials, the term "impact" herein refers to the response to an intervention observed in an individual or homogeneous group (in contrast to efficacy, the average population outcome from an intervention).

Need for Small Sample, Intensive Intraindividual Designs

Recent studies illustrate the range of applications, potential, and need for small sample study designs in prevention. Microtrials hold promise for elucidating genetic, and complex non-genetic, mechanisms within prevention research (Howe et al. 2010). Medium-sized or larger intervention effects can be detected using small samples (e.g., $N<15$) with as few as three time points and conservative statistical tests (Ridenour et al. 2009). Policy and media studies have demonstrated prevention program impact using low-cost, single case studies (e.g., one state or community) (Biglan et al. 1996; Maldonado Molina and Wagenaar 2010). Some specific needs for small sample, intensive within-person designs in prevention science include rigorous designs for testing selective/indicated/tailored intervention, pilot studies, identification of "active ingredients" and mechanisms of prevention programs, differential intervention response, and testing efficacy of programs to avert low prevalence pathologies (e.g., suicide attempts).

Failing to assimilate intensive within-person research impedes prevention science in fundamental ways. To illustrate, psychological characteristics have been nearly always analyzed as if they are invariant, fixed-effect traits, until acted upon by intervention. One example is using a single wave of baseline data to test whether pre-intervention characteristics predict intervention outcomes. This approach assumes that within-person variability in risk factors is uninformative when in fact patterns in pre-intervention variability of characteristics are probably equally (and maybe more so) predictive of response to an intervention (Hintze and Marotte 2010). The most critical effect of an intervention may be to alter the within-person variability of an outcome rather than the average level. A key barrier to testing pre-intervention variability and trends as predictors of efficacy is the prohibitive cost to collect multiple waves of baseline data in large samples. Yet, personal and environmental risk factors are rarely, if ever, static including personality traits and psychiatric disorders (Boker et al. 2009; Borkenau and Ostendorf 1998; Molenaar 2004).

The Ergodicity Assumption

Nearly all extant psychosocial prevention research assumes ergodicity (between-person variation generalizes to longitudinal within-person variability), which is rarely tenable in any social or health science (Molenaar 2004). Thus, the current prevention science literature that focuses on program efficacy (consisting almost entirely of large sample studies with few waves of data per person), may generalize well to the corresponding populations (demarcated by sample age, gender, region, etc.) but lacks generalization to longitudinal processes that occur within individuals who receive an intervention (Boker et al. 2009). Small sample designs permit delineation of the size and mechanisms of a prevention program impact on specific types of persons and well-specified variation between subgroups. In lieu of intensive within-person designs to elucidate developmental and intervention mechanisms, theory regarding prevention program efficacy must currently rely largely on conjecture rather than evidence-based grounding.

Early within-subject experimental designs that did not assume ergodicity were used in psychology and education (Hoyle 1999). These lines of investigation lost prominence partly due to the decline of the subdisciplines in which they were used (e.g., behaviorism), but also because of statistical shortcomings such as failing to account for serial dependency. Since then, in fields mostly outside of psychology and education, analytic methods have been developed and widely used to overcome statistical shortcomings of conducting rigorous case or small sample studies using time series data (Boker et al. 2009; Hoyle 1999; Molenaar 2004). Herein, three such analytic methods were demonstrated for small sample prevention experiments: time series analysis (specifically ARIMA), mixed model trajectory analysis and P-technique.

Foci of the Present Study

Each analytic strategy was evaluated in terms of statistical power to detect intervention effects, relative strengths and limitations and potential uses to address prevention-oriented research questions. The strategies differ in how intervention effects are estimated, method to account for serial dependency, and modeling of non-intervention variables (e.g., personal characteristics such as age or gender).

A secondary goal was to illustrate the potential utility of these techniques for applied uses and expansion of the scope of prevention science. Consistent with this special issue theme, the outcome consisted of blood glucose level (mg/dL), a well-documented biological mediator of medical complications from diabetes. Conducting this study using a large-scale, randomized clinical trial would be logistically

formidable, costly and time consuming. Moreover, the intervention was designed for a well-specified group of persons, an ideal scenario for a small-scale pilot study.

Diabetes in the Elderly

Diabetes is generally characterized by either a deficiency of (Type I) or an insensitivity to (Type II) insulin, a hormone which facilitates transfer of glucose from blood plasma into cells. Organ damage (visual loss, stroke, heart attack, kidney disease, nerve damage, etc.) results both from sustained high blood glucose levels and because cells lack glucose. Acute spikes in glucose levels also can cause alterations in cognition including loss of consciousness. Normal fasting blood glucose ranges from 70 to 125 mg/dL (serving the role of a universal screener). Two random non-fasting sugars over 200 mg/dL is diagnostic for diabetes. Ketoacidosis (accumulation of acid or ketones in the blood stream consequent to poorly controlled Type I diabetes) occurs at a blood glucose of 240 mg/dL or greater which, if untreated, can lead to diabetic coma or even death.

As with psychosocial prevention programs, alternative interventions exist to control blood glucose by targeting different risk factors or mechanisms. Compared to psychosocial prevention, greater delineation of matching treatment to types of persons with diabetes has occurred (largely due to small sample studies). In the elderly, many treatments for diabetes have iatrogenic side effects (e.g., metformin may cause diarrhea, exenatide and liraglutide may cause vomiting) or become ineffective over time (other noninsulin medications). Although a carbohydrate consistent diet keeps meal size constant (while allowing for variability in glucose levels), nursing home residents are rarely able to adhere long-term to this inflexible regimen. The "sliding scale" is a commonly used insulin regimen consisting of a basal insulin dosage supplemented with bolus doses that are often adjusted every 2 weeks, based on the patient's average glucose level. Because the sliding scale fails to account for a patient's meal size (Pickup and Keen 2002), it has been criticized for its lack of prospective dosing. A popular alternative is the insulin pump which is convenient, portable, and provides prospective insulin dosing based on meal size and glucose reading. Nursing home residents are often treated using a sliding scale due to a variety of factors including practitioner preference and habit, ease of sliding scale use, and staff or practitioner discomfort with insulin pumping technology. Use of the insulin pump with nursing home residents also is limited by patient factors (inability to purchase the expensive equipment, confusion leading to misuse of equipment, risk of infusion site dislodgement, unpredictable eating patterns) or lack of insurance coverage.

Tested herein was a novel "manual pancreas" intervention, consisting of nurses manually (a) reading glucose levels, (b) using algorithms to compute needed adjustments to insulin dosing (e.g., Bolderman 2002), and (c) delivering bolus doses of insulin. Manual pancreas was expected to reduce average glucose levels and frequency of glucose spikes compared to care as usual (sliding scale). Accordingly, the first hypothesis was that clinically important differences between interventions for the sample ($N=4$) and for each individual patient ($n=1$) would be statistically detected.

Each of the three analytic strategies offers an advantage for the purpose of modeling how glucose levels vary over time and may (not) respond to an intervention. Glucose levels follow a circadian rhythm and are affected by many other factors with food intake considered the most influential. ARIMA is most adept for modeling the cyclic patterns in glucose (e.g., circadian rhythm, nutritional content of breakfasts may vary less than suppers from day-to-day). P-technique offers the advantages of statistical control of circadian rhythm and other forms of serial dependency, convenient tests for whether intervention changes variance in an outcome, and the option of modeling change in outcome as a latent variable. Accordingly, the second hypothesis was based on the expectation that ARIMA is best able to handle the high intrapersonal variability in glucose patterns. Specifically, modeling of observed glucose levels was hypothesized to be most accurate for ARIMA, less accurate for P-technique and least accurate for trajectory analysis.

For the purpose of testing the manual pancreas intervention (expected to consist of a mean difference between baseline and intervention phases), cyclical variation in glucose is largely a nuisance factor to be parsed out of the analysis. Mixed model trajectory analysis can control for a range of patterns of serial dependency while offering statistical power for detecting change over time (phase differences, slopes, etc.). The third study hypothesis was that estimates of intervention effect size would differ little among the techniques. This hypothesis implies that for simple efficacy studies (e.g., mean differences between study phases that are conducted within persons), mixed model trajectory analysis sufficiently epitomizes data, even for outcomes with large and seasonal intrapersonal variability.

Methods

Sample

Consistent with university IRB approval, participants were from a single nursing home in a medium-sized northeastern town. The inclusion criterion was having Type I diabetes or Type II diabetes severe enough to require both basal and

bolus dosages of insulin. Data were from medical charts. Study data consisted of gender, age and blood glucose levels at 7:30am, 11:30am, 4:30pm and 8:30pm. Blood glucose circadian rhythms typically peak between noon and 6pm, but vary considerably between persons.

Procedures and Intervention

Data were collected between August 1st, 2005 and April 30th, 2007. Each participant experienced a baseline phase of sliding scale followed by a manual pancreas phase. An intermittent baseline design was used to greatly reduce potential for spurious effects (Biglan et al. 1996; Cook and Campbell 1979) and attain important strengths above and beyond conventional pre/post quasi-experimental evaluations. Strengths include additional assurance that conditions preceding the intervention were not atypical (e.g., reducing the potential for regression-to-the-mean effects), increased control over contemporaneous unanticipated events during study implementation, and an opportunity to examine the functional form of intervention effects. Study timeline and patient intervention phases appear in Fig. 1.

Nursing home staff was trained to compute algorithms to determine bolus insulin doses based on the insulin sensitivity factor and carbohydrate to insulin ratio (available from the 2nd author). Training lasted 1 h, detailed the protocol for computing insulin dosage and was videotaped for subsequent staff review. The physician trainer and nurse manager were available daily for consultation, direct patient care and random checking of insulin dosing calculations.

Analyses

Data analyses included autoregressive moving average (ARIMA) analysis (with Proc ARIMA, SAS 9.1), mixed model trajectory analysis (MMTA) (with Proc Mixed, SAS 9.1), and P-technique (with AMOS 17.0). Software packages were chosen because they are easily accessed by prevention researchers. Each method has an established literature, but has

rarely been used for small sample prevention efficacy research. Following are succinct descriptions of each method as they were used herein with references to thorough overviews. For simplicity, statistical notation assumes the use of standardized variables (means = 0; SDs = 1).

Exploring the functional form of a time series helps detect such trends as step functions (sudden mean shifts such as might be expected from implementing an intervention) vs. linear slopes vs. exponential changes as well as longer-term asymptotic effects. Overall, a time-series design permits evaluation of (1) long-term manifold, even hundreds of, repeated measures; (2) a broad range of resolutions of data (e.g., annual to multiple times daily); (3) comparisons between study groups; and (4) within-person intervention phases. The competing analytic techniques were compared statistically in terms of (a) accuracy of predicted glucose levels in terms of correlation with observed glucose levels, (b) consistency among techniques for estimating intervention effect size, and (c) consistency in predicted glucose levels among techniques.

Autoregressive Moving Average (ARIMA) Analysis Glucose levels have seasonal variation due to circadian rhythm and weekly diet patterns. A strength of ARIMA is adeptness at testing and accounting for all aspects of cyclic patterns and stochastic autocorrelation structure (serial dependency) of a time series (Velicer and Colby 1997). By first accounting for serial dependency, a benchmark is provided for testing an intervention. The general multiplicative seasonal model was considered for each patient's time series. For an outcome variable, Y_t, a seasonal ARIMA model with structure $(0,1,1)(0,1,1)_4$ is:

$$\left(1 - B^4\right)Y_t = \alpha + \omega I_t + \left(1 - \Theta B^4\right)u_t \qquad (1)$$

where Y_t is an outcome (i.e., glucose level) from the first, $t=1$, to last observation; α is a constant; ω represents the effect of intervention; I_t is a time-varying step function that equals 0 for baseline observations and 1 during intervention; Θ is the first order seasonal moving average parameter; u_t is random error,
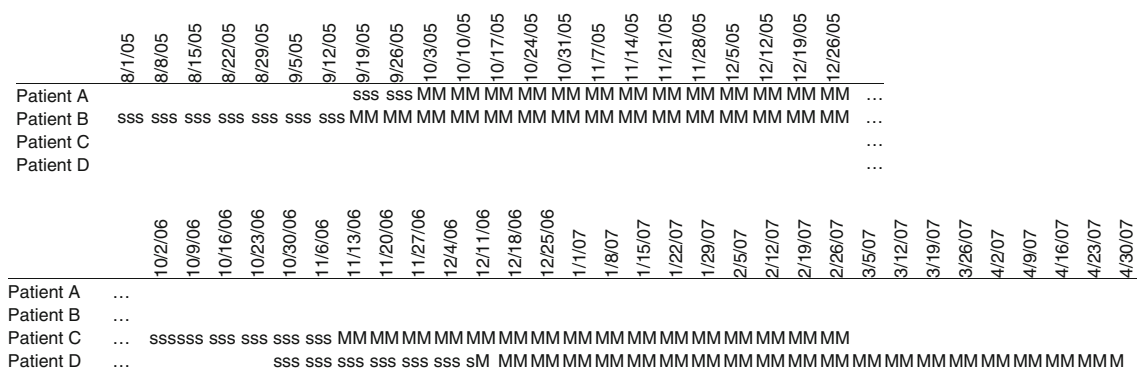

Fig. 1 Timeline of data collection. Note: s = baseline phase using sliding scale. M = intervention phase using manual pancreas

and B is the backshift operator such that $B^4(Y_t)$ equals Y_{t-4}. A backshift operator transforms an observation of a time series to the previous observation to account for seasonality. An ARIMA model is appropriate only after the series is stationary (has a constant mean, variance, autocorrelation) which typically requires hundreds of observations.

Mixed Model Trajectory Analysis (MMTA) Whereas ARIMA detects well the error covariance structure(s) of time series, MMTA offers (a) statistical and logistical parsimony, (b) modeling flexibility, (c) statistical power, and (d) familiarity to clinical researchers (Hedeker and Gibbons 2006). For within-person analyses, individuals' outcomes occurring over time are analyzed at level 1 which are clustered in persons at level 2 (where individual differences are tested). Ridenour et al. (2009) demonstrated using MMTA for small sample randomized clinical trials having few waves, including for cross-over and intermittent baseline designs. Within-person MMTA can be represented using a single regression equation:

$$Y_{it} = \beta_0 + u_{0i} + \beta_1 + u_{1i} + \beta_2 \text{Intx}_{it} + \beta_3 (\text{Intx*Time})_{it} + e_{it} \tag{2}$$

where Y_{it} is an outcome for individual i at time t; the intercept for individual i (outcome for the observation at which time is centered, typically the first of the study) is a function of the average sample intercept (β_0, a constant) plus the individual's deviation from this average (u_{0i}); change in the outcome over time is modeled as a function of the sample average trend (β_1) plus the individual's deviation from that trend (u_{1i}); baseline vs. intervention phases were modeled as differences between means ($\beta_2 \text{Intx}_{it}$) and trends (change over time or $\beta_3(\text{Intx*Time})_{it}$); and e_{it} denotes random error. When modeling a trend for a single person, β_1 model the average for that person over time, u_{0i} is dropped from the model (i.e., equals zero), and u_{1i} models the deviation of specific observations from the person's trend. Intervention phases were dummy coded as 0 (sliding scale) and 1 (manual pancreas) and tested as fixed effects. Maximum likelihood estimation was used to test fit of the overall model, specific predictors and error covariance structures (using likelihood-ratio χ^2, Akakie's Information Criterion, Bayesian Information Criterion fit statistics) with restricted maximum likelihood used to acquire parameter estimates.

When the error covariance structure of a time series is misspecified in MMTA, the analysis likely generates biased estimates of parameter variance and random effects and possibly biased estimates of fixed effects (Kwok et al. 2007; Sivo et al. 2005). Several error covariance structures were tested (autoregressive, heterogeneous autoregressive, autoregressive moving average, and factor analytic, each with a lag 1 per the proc mixed options)

to account for the greatest amount of variance due to serial dependence. To reduce potential for Type I error, the Kenward-Roger adjusted F-test for small samples was used (Littell et al. 2006).

P-technique P-technique models latent-level within-person longitudinal processes based on multivariate time series data (Molenaar and Nesselroade 2009). For instance, P-technique studies (and meta-analysis thereof) identified core processes of client and therapist contributions to psychotherapy (Russell et al. 2007). P-technique uses an equivalent covariance structure as factor analysis, thereby providing an analytic paradigm which is familiar to many prevention researchers for longitudinal, within-person, small sample research. Although P-technique has evolved in the form of dynamic factor models, the two methods perform equivalently for modeling simple latent within-person processes (Molenaar and Nesselroade 2009). A general P-technique model for y_t, similar to traditional factor analysis, is:

$$\mathbf{y}_t = \Lambda \eta_t + \varepsilon_t \tag{3}$$

where y_t denotes a person's value at time t over the course of a (p-variate) time series; Λ represents a vector of factor loadings; η comprises a corresponding vector of latent factors; and ε represents random (measurement) error.

Glucose levels at different times of day were analytically handled as indicators of a single latent construct (daily glucose level). Multiple glucose measures were needed per day because of circadian rhythms. An important assumption of P-technique which held for the present data, but often does not in time series data, is that the time series is weakly stationary (in effect, that within an individual, the mean and variance of a time series are constant and finite).

Integrated Method A novel approach tested herein was an attempt to capitalize on the strengths of ARIMA and MMTA to potentially surmount each of their limitations. ARIMA is more adept for elucidating error covariance structure than MMTA. However, ARIMA is more sensitive to missing data and outliers than MMTA and offers less power. Thus, MMTA was conducted a second time, using the error covariance structure detected in ARIMA; results of the first and second MMTA were then compared.

Results

The sample of four European-American nursing home patients ranged in age from 72 to 85 years and included three males. Patients A, C and D had Type 1 diabetes; B had poorly controlled Type II diabetes. Consistent with the population

with diabetes, patients' glucose was above the normal range on average. Patient A's $\overline{X} = 153.7$ ($SD=55.7$); sliding scale $\overline{X} = 192.6$ ($SD=72.8$, $n=37$) and manual pancreas $\overline{X} = 150.0$ ($SD=52.4$, $n=338$). Patient B's $\overline{X} = 190.2$ ($SD=93.5$); sliding scale $\overline{X} = 262.0$ ($SD=109.5$, $n=196$) and manual pancreas $\overline{X} = 155.2$ ($SD=58.7$, $n=403$). Patient C's $\overline{X} = 221.3$ ($SD=103.5$); sliding scale $\overline{X} = 253.7$ ($SD=126.0$, $n=157$) and manual pancreas $\overline{X} = 209.0$ ($SD=90.7$, $n=413$). Patient D's $\overline{X} = 183.7$ ($SD=81.1$); sliding scale $\overline{X} = 232.1$ ($SD=87.6$, $n=151$) and manual pancreas $\overline{X} = 170.5$ ($SD=74.04$, $n=554$). As the large SDs suggest, extreme glucose spikes were observed in each patient with maximum glucose of 355 for A and near 550 for patients B, C and D. Few missing observations occurred: 0.7 % for Patient A, 1.5 % for B, 2.8 % for C and 2.7 % for D.

Figure 2 illustrates the time series data as well as certain challenges of analyzing them. Visual inspection of Panel a reveals neither when the intervention was initiated nor its impact for Patient C at 11:30am. Yet, manual pancreas was associated with lower blood glucose on average ($p<.001$) (Table 1). For comparison, Patient D's data at 8:30pm (Panel b) and the superimposed trend line demonstrates an effect size that resembles the effect hidden in Panel a. Neither linear nor nonlinear trends occurred in the data aside from a phase shift associated with intervention, so results focus on mean differences between study phases.

ARIMA Autoregression is present for each patient (unconditional AIC for patient A=4,917; B=7,849; C=7,483; and D=8,889). A model with lag of 1 (autocorrelation with the preceding glucose level) improves overall fit to the data (AIC for patient A=4,935; B=7,390; C=7,241; and D=8,397), but fails to account for all serial dependence of any patient. Additional modeling of autocorrelation for time of day, via lag 4 autocorrelation (AIC for A=4,949; B=7,069; C=7,261; and D=8,501) and moving average of 4 (AIC for patient A=4,459; B=6,472, C=6,719, and D=7,902),

Fig. 2 Exemplar patient time series data for two patients. Panel **a**: Time series of blood glucose levels at 11:30am for Patient C. Panel **b**: Time series of blood glucose levels at 8:30pm for Patient D with the trend from results of mixed model trajectory analysis of intervention impact appearing in *bold*
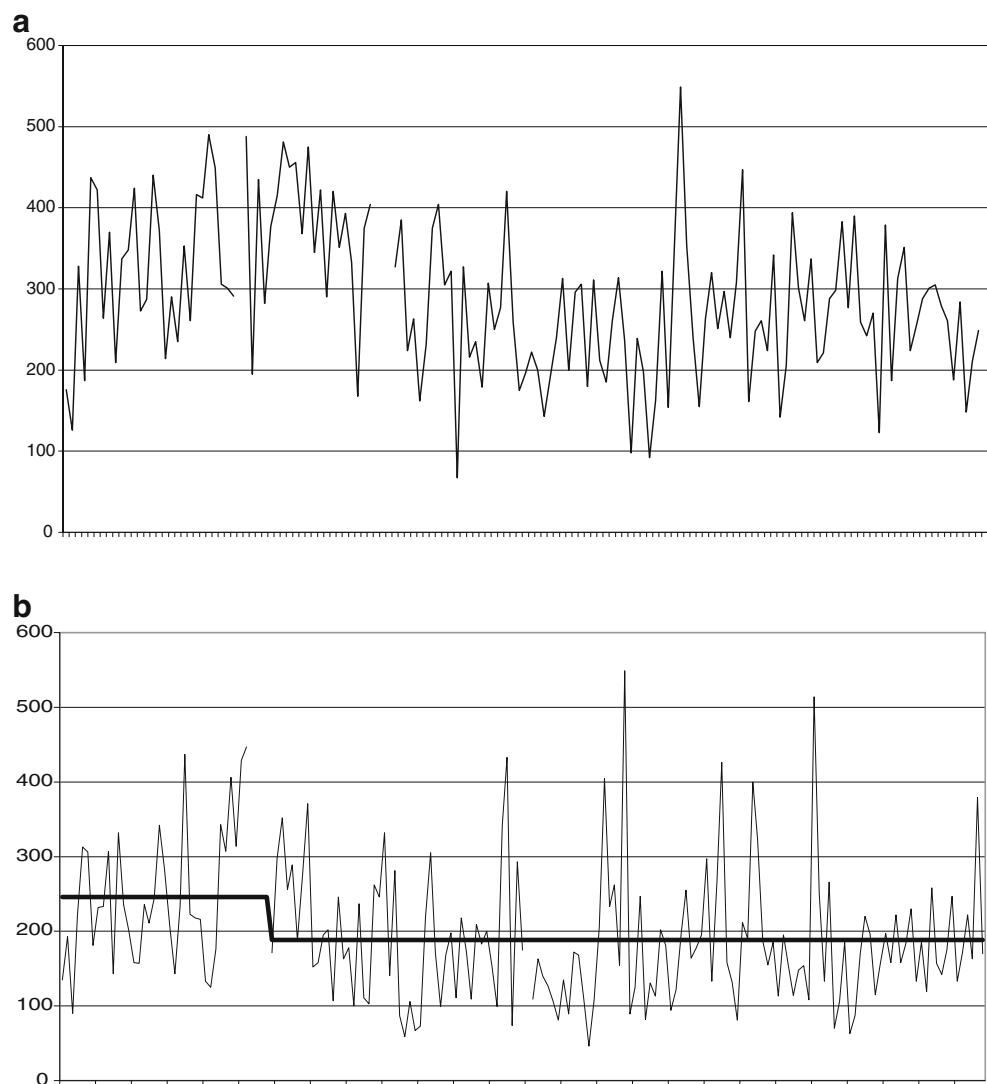
Table 1 Change in blood glucose associated with manual pancreas (compared to sliding scale)

	Aggregated times[A]	7:30am	11:30am	4:30pm	8:30pm
MMTA					
Entire Sample	−49.4[a] (9.2)	−35.9[b] (9.8)	−43.3[a]* (194.2)	−59.4[b] (9.7)	−59.1[a]* (277.9)
Patient A	−40.9[b] (10.7)	0.2[b]* (11.1)	1.8[a]* (24.4)	−50.4[b] (20.2)	−104.2[b] (19.4)
Patient B	−107.9[b] (11.8)	−32.2[b] (8.8)	−117.3[a] (23.0)	−156.3[b] (19.3)	−122.2[b] (17.0)
Patient C	−22.6[b]* (15.3)	11.5[b]* (27.5)	−66.6[a] (26.8)	−35.5[b]* (25.4)	3.0[b]* (27.7)
Patient D	−24.6[b] (10.1)	−112.1[b] (16.0)	26.3[a]* (17.6)	43.5[b] (17.7)	−57.3[b] (24.3)
P-technique					
Entire Sample	−64.9 (6.8)	−32.4 (7.7)	−89.3 (7.6)	−98.8 (6.5)	−83.1 (6.2)

A = intervention effect aggregated over all times of the day for the sample or specific patient. a = heterogeneous autoregression, lag 2, error covariance structure. b = factor analytic, lag 2, error covariance structure. *Change in glucose was NS ($p>.01$). Parenthetical values are standard errors. Change attributable to time (slope) and time-intervention interaction were statistically nonsignificant in all MMTA

accounts for remaining autocorrelation of residuals ($p>.05$) for all patients, except D. A strict significance criterion of $p<.001$ was used to evaluate all models for effects of outliers; no outliers are included in the results for ARIMA.

A reduction in blood glucose of 52.2 mg/dL (se=22.6) associated with manual pancreas is estimated via ARIMA for the sample in aggregate ($t=2.31$, $p=.021$). In N-of-1 analyses, for patients A and C, manual pancreas does not statistically reduce overall glucose levels vs. sliding scale ($t=−1.47$, $p=.15$ for A; $t=−0.72$, $p=.48$ for C), although the effects are in the hypothesized direction: −25.2 for A and −47.9 for C. Patient B's glucose is significantly reduced ($t=−2.11$, $p=.04$) by 62.5 units. Patient D's results are confounded by (a) missing data immediately preceding implementation of manual pancreas (Fig. 2, Panel b) due to patient hospitalization and (b) outlier levels of high blood glucose during the few days preceding hospitalization. Omitting patient D's missing data points permits the model to fit patient D's data; under this condition, glucose is reduced by 61.5 ($t=−5.42$, $p<.001$). ARIMA models of outcomes for specific times of day are too unstable to provide estimates of intervention impact because of needing more observations and the missing data.

MMTA Using MMTA, two error covariance structures best fit the data, depending on which subset of data was analyzed (see Table 1; fit statistics are not presented due to space restrictions but are available from the 1st author). Analyses were repeated for each time of day to: control for circadian rhythm variation (cf. ARIMA lag 4 autocorrelation), clarify between-patient differences and elucidate mealtime-specific intervention impact. The factor analysis lag 1 covariance structure best fits glucose time series for 7:30am, 4:30pm and 8:30pm; heterogeneous autoregression lag 1 (which does not require variances to be approximately equivalent) best fits the 11:30am time series.

Table 1 summarizes MMTA results. In aggregate, manual pancreas is associated with reduced blood glucose (by 49.4 mg/dL, se=9.2; $p<.001$). This effect is statistically significant only at 7:30am and 4:30pm (due to large standard errors at other times), but increases throughout the day. Patient-specific analyses suggested a reduction in overall blood glucose for patients A, B and D but not C (also partly due to a larger standard error). Within-patient results differ appreciably among them, including the times of day at which manual pancreas is associated with decreased blood glucose. No significant interaction between study phase and time occurs.

As noted earlier, analyses were repeated using the error covariance structure identified in ARIMA analyses. To replicate the ARIMA lag 4 structure for entire patient time series (rather than time-specific series), a Toeplitz lag 5 error structure is required in Proc Mixed. Results for these replicated analyses differed little from those reported earlier.

P-technique Figure 3 presents the P-technique model with lag 1 autocorrelation at the latent level and lag 4 autocorrelation for specific times of day. Fit to data is best by freeing parameters to allow for between-patient and between-
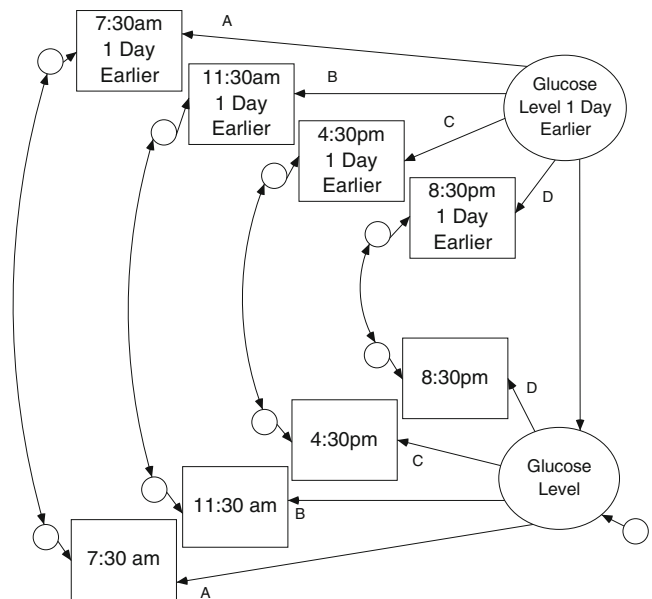
Fig. 3 P-technique factor model for daily glucose level with Lag 1

intervention differences (Table 2). The aggregate estimate of intervention efficacy from P-technique is a reduced blood glucose level of −64.3 mg/d (se=6.8; $p<.001$). Patient-specific analysis does not generate stable estimates of intervention effect because the large number of model parameters (and subsequent low ratio of number of parameters to observations). Time-specific analysis replicates the aggregate result that lower blood glucose is associated with manual pancreas (Table 1); standard errors for these analyses consist of a pooled standard error based on the corresponding parameters from sliding scale and manual pancreas estimates. The reduction in blood glucose associated with intervention is least for 7:30am and greatest at 4:30pm.

"Outliers" as Outcomes The term "outlier" appears in quotation marks because glucose outliers are nuisance values only in terms of statistical modeling. To the health of patients, they are of utmost importance, as blood glucose spikes suggest poor diabetes control and reductions in their frequency is a clinical goal. Statistical "outliers" also are frequently targets of psychosocial prevention programs (e.g., aggression, suicide attempts).

Thus, for rigorous small sample prevention research, change in frequency of such outliers can be central to testing a program's impact on an individual as well as the program's efficacy (and requiring their omission to analyze data may be a critical limitation). Using McNemar's χ^2, the following changes in frequency of occurrences of glucose >240 are associated with sliding scale vs. manual pancreas, respectively. (Phase marginal values for 2×2, cross-tabulation tables appear earlier in Results section "n"s.) In aggregate, a reduction in spike frequency of nearly 2/3 is associated with manual pancreas (43.8 % vs. 14.2 %, $p<.001$). This result also is observed per patient (all χ^2 tests $p<.001$): A had 15.0 % vs. 4.6 %; B had 54.6 % vs. 7.7 %; C had 48.4 % vs. 30.3 %; D had 32.5 % vs. 13.5 %. To estimate the effect of omitting outliers in MMTA and P-technique (to which ARIMA is sensitive), a sensitivity analyses was conducted comparing results with all data to results when outliers were omitted. Virtually identical results were found, thus only results for the full dataset are reported for each strategy.

Comparisons Among Techniques Regarding hypothesis two, observed glucose levels were predicted by ARIMA with $r=.57$, by MMTA with $r=.43$ and by P-technique with $r=.65$ (all $p<.01$, using Pearson r). Additionally, the r between predicted glucose values of ARIMA and MMTA was .53, between ARIMA and P-technique was .66 and between MMTA and P-technique was .56. For hypothesis three, Cohen's d was 1.28 for ARIMA, 0.84 for MMTA and 1.34 for P-technique (due to larger standard deviations for MMTA).

Discussion

Results were generally consistent with the hypotheses. Each analytic method led to similar overall conclusions. First, manual pancreas is associated with healthier glucose levels than sliding scale. Second, large variability occurs between patients in terms of average glucose levels, daily patterns, and response to intervention. Third, in spite of meaningful impacts attributable to manual pancreas, greater glucose control is needed for each patient.

Comparison Among Techniques

Nevertheless, for this particular study each method offered a unique benefit. ARIMA best detected serial dependency (e.g., thus reducing the denominator for estimation of Cohen's d). MMTA estimated the intervention effect with the least difficulty and for the smallest segments of data (i.e., per time of day for each patient). The P-technique model (modeling error covariance as detected by ARIMA) replicated the time series data most accurately. If the study purpose had been to determine the reduction in glucose level (as opposed to Cohen's d), either overall or for specific patients' mealtimes, MMTA would have best served the study. Otherwise, P-technique or ARIMA would have been the most advantageous technique.

Estimates of effect size differed between the analytic techniques, a result that was not surprising given their differing approaches to modeling phases of the time series.

Table 2 Test of differences between interventions and participants on latent processes of blood glucose

Parameters freed	χ^2	df	AIC	LR χ^2, df		
				vs 1	vs 2	vs 3
1. None (fully constrained)	43,866.4	327	43,916.4			
2. Between phases	43,246.7	310	43,330.7	619.7*, 17		
3. Between patients	42,832.2	276	42,984.2	1,034.2*, 51	414.5*, 34	
4. Between phases & patients	41,944.3	207	42,234.3	1,922.1*,120	1,302.4*, 103	767.5*, 56

df = degrees of freedom. AIC = Akakie's Information Criterion. LR χ^2 = likelihood ratio χ^2. *$p<.001$. "Parameters" consists of factor loadings, intercepts and variances.

The most conservative efficacy estimate was from MMTA, which was more similar to the estimate of ARIMA than P-technique. Given the range (11 to 549 mg/dL) and high variance of observations, the three estimates of efficacy were similar. Not surprisingly, estimates of intervention outcome become less similar at lower levels of data analysis, such as for individuals or specific times. Compared to single cases, it appears that samples consisting of a few persons greatly improve accuracy (and generalizability) of effect size estimates in intensive within-person studies. Although evidence is lacking to determine which technique provides the most accurate estimates, this study clearly demonstrated the need for simulation data with known correct results to learn which technique is optimal under specific circumstances.

The three methods handled serial dependence differently. Within MMTA, covariance within-persons is handled as nuisance variability, controlled for with an error covariance structure. ARIMA not only conveniently tests for and explicitly models within-person autocorrelation and cyclical patterns (seasonality), it can capitalize on such trends to forecast outcomes occurring subsequent to the study data (Chatfield 2004). In contrast, software programs in which P-technique could be conducted require the statistician to explicitly model an error covariance structure (as opposed to providing code to add the structure to an analysis).

Finally, the three techniques offer differing strengths which may be leveraged for answering different prevention-related research questions. Illustrated herein were the strengths of MMTA for testing intervention impact, specifically statistical power, robustness to missing data and outliers, as well as an elegant modeling approach that requires fewer parameters and observations than ARIMA or P-technique. MMTA can facilitate comparisons between nomothetic studies (generalizations across people) and idiographic studies (within-person processes) because the exact same techniques and estimates can be derived from both studies.

A strength of ARIMA that was illustrated was detection of error covariance structure. For instance, this strength permitted analytic control of within-person variability, leading to a larger estimate of intervention efficacy (in the form of Cohen's d). A feature of ARIMA that was not illustrated herein is forecasting outcomes (widely used in business). Potentially, an ARIMA model could warn of high probability for a clinical event (e.g., suicide attempt) which could then be averted with intervention.

Most of the strengths which P-technique and dynamic factor models offer for prevention research were not highlighted herein. One strength that was demonstrated was that P-technique most closely modeled the observed levels of glucose. P-technique also allows within-person developmental processes (including response to combinations of interventions) to be investigated at the latent level. To illustrate how useful this feature is, consider that ratings of a child's antisocial behavior are well-known to be inconsistent between the child, parent and teacher and there is no agreed upon method for culling data from each rater (de los Reyes and Kazdin 2005). These methods offer a data-driven method for aggregating ratings from different sources. Additionally, interactions over time between dyads could be modeled.

Study Limitations

Comparisons between the analytic strategies require replication. This study provided working hypotheses that should be tested using simulation studies with known correct results. The scenario investigated herein consisted of a simple mean comparison between interventions. While consistent with emphases in prevention science, how well these study results generalize to more complex scenarios (e.g., modeling error covariance, testing interactions) needs clarification. Data analysis packages can be improved upon by adding features specifically for the small sample strategies considered herein or which can be programmed to specifically handle the features of a particular time series (e.g., a unique error covariance structure). Finally, this study illustrated an important limitation of each technique; they could not statistically test the reduced frequency of outlier clinical events (glucose spikes) associated with the intervention.

Implications and Future Steps

One needed future step is to determine how the strengths of each technique might be used in complementary ways to analyze intensive within-individual analyses. An example alluded to earlier applies ARIMA to identify the error covariance structure and then including that structure in MMTA or P-technique. Potentially, each method is optimal for analyzing different developmental processes within persons or research conditions. Exactly when each technique provides the optimal analytic strategy could be determined in simulation studies in light of varying sample sizes, numbers of observations, effect sizes and serial dependence. Also needed is rigorous simulation work to determine sample sizes needed to generalize parameter estimates to populations of varying homogeneity (e.g., elderly in nursing homes vs. all retired persons).

Regarding the study design, opportunities abound for using small sample designs in the context of prevention. At first, the task of collecting manifold data points for each person may appear onerous. However, this task frequently already occurs in settings where participants receive intervention (Velicer and Colby 1997). Teachers, especially special education teachers, collect outcomes data on students and recently do so more frequently to record student response to intervention (Burns et al. 2010). Any locale in which people come into regular contact with health or education professionals (e.g., hospitals, schools, residential

or inpatient units, institutional leisure activities) provides an opportunity for intensive within-person data collection. Logistical obstacles notwithstanding, the intervention goals of professionals in these settings generally align with the goals of prevention science, providing foundations for collaborating with them.

As illustrated herein, coupling the interrupted time-series design with these analytic techniques provided statistical power to detect effects in very small samples and generates estimates of intervention impact. Such designs offer high external validity to future observations of the patients who are studied as well as other patients who resemble the study sample. Indeed, they could be used by clinicians to monitor a patient's progress and inform clinical decisions (Kazdin and Blase 2011). The between-person variability shown herein illustrates a need for simulation research to determine adequate sample sizes for generalizing results to other persons. To date, generalizability has been addressed adequately in intensive within-person research using replication studies. Having statistically grounded guidelines to determine sample size could improve this aspect of a-priori study design.

The analytic techniques used herein can be coupled with adaptive designs such as Sequential Multiple Assignment Randomized Trial (Murphy et al. 2007). By not requiring large samples, such complex designs would be neither cost prohibitive nor logistically burdensome (by offsetting fewer persons with greater number of observations to attain power), two barriers that have impeded wide use of these designs (Murphy et al. 2007). Using adaptive designs in applied settings (schools, residential units, probation visits, inpatient hospitals) where data could be (and frequently are) collected at least daily (e.g., attendance, behavior problems, monitoring response to treatment) permits evaluation of intervention phases in terms of trends as well as the traditional mean efficacy.
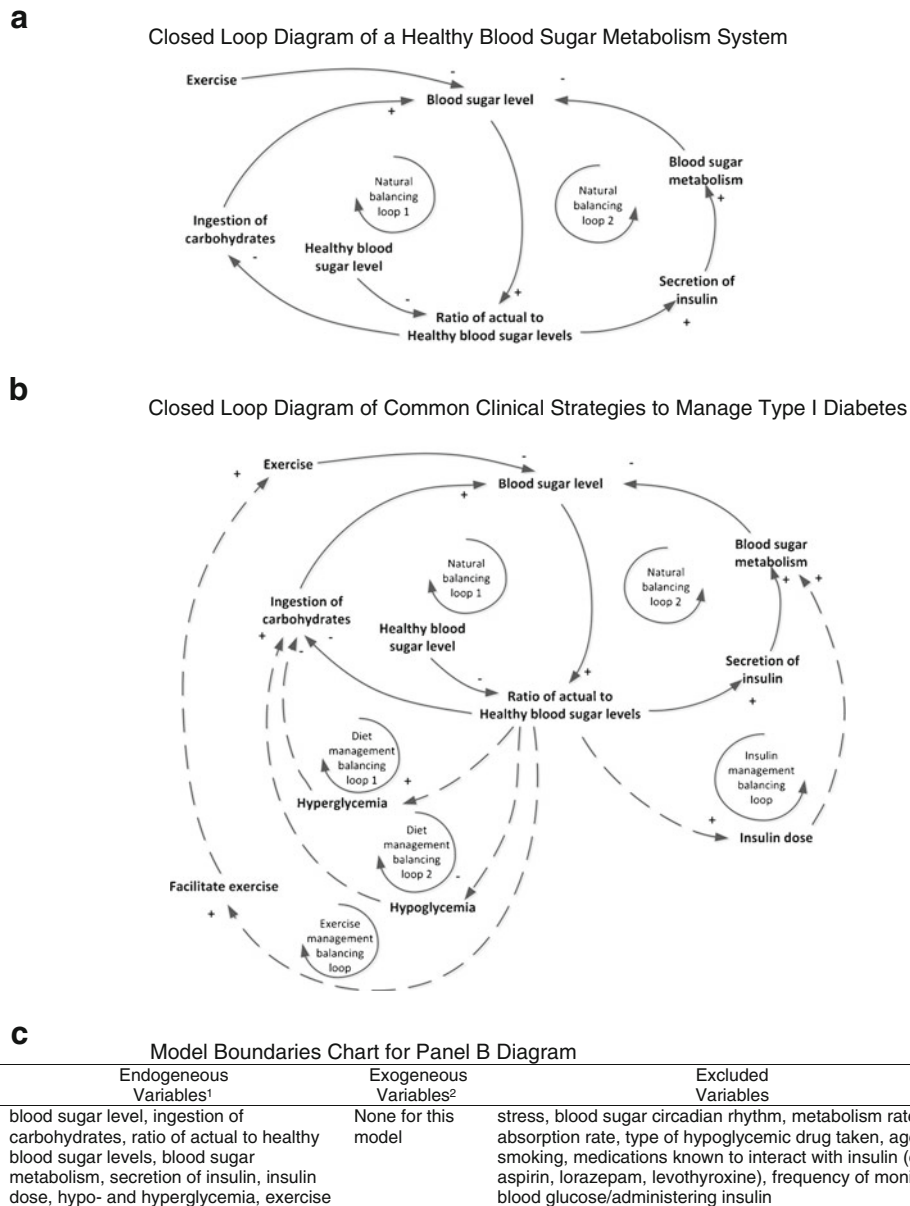
Research methods shown herein could also be useful in System Dynamics studies to address complex and dynamic prevention science problems (Hassmiller Lich et al. 2012). In System Dynamics projects, a model or "dynamic hypothesis" is built to represent the key components of a system in which a documented and dynamic problem is occurring, so that the effects of individual and combined complexities (e.g., mediating factors, delays, interactions, non-linearities) can be understood within the context of the entire system. Alternate intervention scenarios can be developed and simulated, and their behavior could be compared virtually. To illustrate, Panel A in Fig. 4 presents a causal loop diagram depicting basic dynamics of diet, exercise and insulin within the glucose metabolism system largely inspired by a diagram developed by Gaynor (1998, p. 121). Panel b presents an expanded diagram depicting the three common clinical strategies for managing Type I diabetes (using two mechanisms of diet,

exercise and insulin) (Lehmann et al. 2011). Arrows between variables should be interpreted as "an increase/decrease in the first variable causes an increase/decrease in the second variable, all other things being equal." A "+" on an arrow indicates that the two variables move in the same direction (that is, an increase in the first variable leads to an increase in the second variable while a decrease in the first variable leads to a decrease in the second variable, all other things being equal; a "-" notes that the variables move in the opposite direction. Dashed arrows in Panel B illustrate potential interventions, adding additional balancing loops that will bring system behavior (here, blood sugar) under control over time.

The causal loop diagram provides a medium to conceptualize the actions and interactions of various agents (e.g., risk factors) upon an outcome, visually organize the leverage points of intervention, and quantify the relationships among the agents. Virtual simulations based on the quantified model can then identify which elements of intervention lead to meaningful change in the outcome. Small sample within-person studies could provide the data for quantifying parameters and testing the validity of such systems models (a challenge, as each model represents a single testable hypothesis, and data collection to validate the model can require substantial effort). When the model focuses on forces affecting outcomes at the individual level (either person, organization, or community), between-person comparisons can be made to identify important moderating factors between individual entities.

Panel c of Fig. 4 demonstrates a model boundary chart, an important step in creating causal loop diagrams. Prevention programs attempt to alter the putatively most critical factors that contribute to a pathological outcome rather than all of the factors that potentially affect the outcome. The boundary chart makes explicit which variables are included vs. excluded from the model (and intervention), thereby compelling justification for each inclusion/exclusion decision. The chart can inform revision of programs. To illustrate, Zhou et al., Receding horizon control of Type I diabetes based on a data-driven linear time-varying state-space model (unpublished) recently showed that glucometer readings and insulin administration occurring every 30 min. (with feedforward-feedback controls to adjust dosing algorithms specifically per patient) could result in glucose levels that remain between 70 and 175 mg/dL. Their report suggests that by adding the previously excluded variable "frequency of monitoring blood glucose/administering insulin" (Fig. 4, Panel c) to the model/intervention strategy, individuals' blood sugar levels could be appreciably improved.

To summarize, the impact of the manual pancreas intervention was pilot tested using an intermittent baseline design, time series data and analytic strategies. Not only was manual pancreas associated with improved glucose control compared to care-as-usual, the improvement was found for each patient and especially with regard to dangerous spikes

a

Closed Loop Diagram of a Healthy Blood Sugar Metabolism System

b

Closed Loop Diagram of Common Clinical Strategies to Manage Type I Diabetes

c

Model Boundaries Chart for Panel B Diagram

Endogenous Variables[1]	Exogenous Variables[2]	Excluded Variables
blood sugar level, ingestion of carbohydrates, ratio of actual to healthy blood sugar levels, blood sugar metabolism, secretion of insulin, insulin dose, hypo- and hyperglycemia, exercise	None for this model	stress, blood sugar circadian rhythm, metabolism rate, absorption rate, type of hypoglycemic drug taken, age, smoking, medications known to interact with insulin (e.g., aspirin, lorazepam, levothyroxine), frequency of monitoring blood glucose/administering insulin

Fig. 4 Panel a Closed Loop diagram of a healthy blood sugar metabolism system. Panel b closed loop diagram of common clinical strategies to manage type I diabetes. Panel c model boundaries chart for panel b diagram. Panel A presents a causal loop diagram illustrating the putative causal sequences that keep the blood sugar metabolism system balanced, heavily inspired by a similar diagram developed by Gaynor (1998). An *arrow* connecting two variables indicates that a change in the first causes a change in the second, all other things being equal. A "+" on the arrowhead indicates the variables move in the same direction, while a "−" indicates they move in opposite directions. Loops are formed when causal sequences circle back on themselves; balancing loops move a system into equilibrium. Panel B presents an expanded causal loop diagram of common clinical strategies to manage type 1 diabetes. *Dashed lines* suggest potential interventions

in blood glucose. The study had high external validity and clinical feasibility as it was conducted in a community nursing home by the nursing home staff. Equally important, innovative strategies to conduct small sample, intensive within-person clinical research were compared and demonstrated for prevention in applied settings. By coupling intermittent baseline designs with techniques to analyze within-individual time series data, highly applied research strategies can be conducted in clinical settings to address research questions that traditional nomothetic methodologies cannot undertake. Also demonstrated was the need for several lines of simulation research to delineate which analytic techniques to use under varying research circumstances.

References

Biglan, A., Ary, D. V., & Wagenaar, A. C. (2000) The value of interrupted time-series experiments for community intervention research. *Prevention Science, 1*, 31–49.

Biglan, A., Ary, D., Koehn, V., & Levings, D. (1996). Mobilizing positive reinforcement in communities to reduce youth access to tobacco. *American Journal of Community Psychology, 24*, 625–638.

Boker, S. M., Molenaar, P. C. M., & Nesselroade, J. R. (2009). Issues in intraindividual variability: Individual differences in equilibria and dynamics over multiple time series. *Psychology and Aging, 24*, 858–862.

Bolderman, K. M. (2002). *Putting your patients on the pump* (pp. 39–52). Alexandria, VA: American Diabetes Association, Inc.

Borkenau, P., & Ostendorf, F. (1998). The Big Five as states: How useful is the five-factor model to describe intraindividual variations over time? *Journal of Research in Personality, 32*, 202–221.

Burns, M. K., Christ, T. J., Boice, C. H., & Szadokierski, I. (2010). Special education in an RTI model: Addressing unique learning needs. In T. A. Glover & S. Vaughn (Eds.), *The promise of response to intervention* (pp. 267–285). New York: Guilford.

Chatfield, C. (2004). *The analysis of time series: an introduction* (6th ed.). Boca Raton, FL: Chapman and Hall/CRC Press.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally

De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin, 131*, 483–509.

Ennett, S. T., Ringwalt, C. L., Thorne, J., Rohrbach, L. A., Vincus, A., Simons-Rudolph, A., & Jones, S. (2003). A comparison of current practice in school-based substance use prevention programs with meta-analysis findings. *Prevention Science, 4*, 1–14.

Gaynor, A. K. (1998). *Analyzing problems in schools and school systems: a theoretical approach*. Mahwah, N.J., L.: Erlbaum Associates.

Hassmiller Lich, K., Ginexi, L., Osgood, N., & Mabry, P. (2012). A call to address complexity in prevention science research. doi:10.1007/s11121-012-0285-2.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. Hoboken: Wiley.

Hintze, J. M., & Marotte, A. M. (2010). Student assessment and data-based decision making. In T. A. Glover & S. Vaughn (Eds.), *The promise of response to intervention* (pp. 57–77). New York: Guilford.

Howe, G. W., Beach, S. R. H., & Brody, G. H. (2010). Microtrial methods for translation gene- environment dynamics into preventive interventions. *Prevention Science, 11*, 343–354.

Hoyle, R. H. (1999). *Statistical strategies for small sample research*. Thousand Oaks, CA: Sage.

Kazdin, A. E., & Blase, S. L. (2011). Rebooting psychotherapy research and practice to reduce the burden of mental illness. *Perspective on Psychological Science, 6*, 21–37.

Kwok, O., West, S. G., & Green, S. B. (2007). The impact of mis-specifying the within-subject covariance structure in multiwave longitudinal multilevel models: A Monte Carlo study. *Multivariate Behavioral Research, 42*, 557–592.

Lehmann, E. D., Tarín, C., Bondia, J., Teufel, E., & Deutsch, T. (2011). Development of AIDA v4.3b diabetes simulator technical upgrade to support incorporation of lispro, as part and glargine insulin analogues. *Journal of Electrical and Computer Engineering*, 2011, 1–17. http://www.hindawi.com/journals/jece/2011/427196/.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS for mixed models* (2nd ed.). Cary, NC: SAS Press.

Maldonado Molina, M. M., & Wagenaar, A. C. (2010). Effects of alcohol taxes on alcohol-related mortality in Florida: Time-series analyses from 1969 to 2004. *Alcoholism Clininical and Experimental Research, 34*, 1915–1921.

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Brining the person back into scientific psychology, this time forever. *Measurement, 2*, 201–218.

Molenaar, P. C. M., & Nesselroade, J. R. (2009). The recoverability of P-technique factor analysis. *Multivariate Behavioral Research, 44*, 130–141.

Murphy, S. A., Lynch, K. G., Oslin, D., McKay, J. R., & Tenhave, T. (2007) Developing adaptive treatment strategies in substance abuse research. *Drug and Alcohol Dependence, 88*(Suppl 2), S24–S39.

Pickup, J., & Keen, H. (2002). Continuous subcutaneous insulin infusion at 25 years: Evidence base for the expanding use of insulin pump therapy in type 1 diabetes. *Diabetes Care, 25*, 593–598.

Ridenour, T. A., Hall, D. L., & Bost, J. E. (2009). A small sample randomized clinical trial methodology using N-of-1 designs and mixed model analysis. *American Journal of Drug and Alcohol Abuse, 35*, 260–266.

Russell, R. L., Jones, M. E., & Miller, S. A. (2007). Core process components in psychotherapy: A synthetic review of P-technique studies. *Psychotherapy Research, 17*, 273–291.

Sivo, S., Fan, X., & Witta, W. (2005). The biasing effects of unmodeled ARMA time series processes on latent growth curve model estimates. *Structural Equation Modeling, 12*, 215–231.

Velicer, W. F., & Colby, S. M. (1997). Time series analysis for prevention and treatment research. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 211–249). Washington, DC.: American Psychological Association.