REPORT ON PROMISE PRIMARY STATISTICAL ANALYSIS

Bruce Levin, Ph.D.

07/09/2021

The primary analysis consists of estimating the intervention effect using fully conditional maximum likelihood estimation together with an exact, conditional *P*-value and an exact test-based 95% confidence interval. Some supportive sensitivity analyses were also prespecified in the published protocol paper[1], namely, Wald, score, and generalized log-likelihood ratio methods, which we report below. In addition, we present a series of descriptive statistics for (a) the within-pair *site effects*, i.e., the log odds ratios (LORs) on TVS (yes vs. no) comparing "early" switchers (sites switching to the new CCP program in Period 1) vs. "late" switchers (sites switching to the new CCP program in Period 2) in each matched pair of sites, adjusted for any intervention effect; (b) the within-pairs *period effects*, i.e., the LORs on TVS (yes vs. no) comparing Period 1 vs. Period 0 and LORs comparing Period 2 vs. Period 0, adjusted for any intervention effect; and (c) the baseline log odds on TVS for late switchers in Period 0. We also present a table of *smoothed TVS rates*, i.e., fitted expected proportions of TVS for each site in each period estimated by the underlying analysis model.

Below we refer to a given matched pair of sites as a "block" and index them by $b=1,\ldots,8$. The data may be arrayed as an $8 \times 2 \times 2 \times 3$ table of frequencies for blocks by sites within blocks by TVS outcomes by period. It will be convenient to introduce the following notation. Within each block, the variable *Site* takes value 1 for the site randomized to early switching in Period 1 or value 2 for the other site randomized to late switching in Period 2. The period variable *Per* takes the values 0, 1, and 2. The data appear in Table 1 below.

[1] Irvine MK, Levin B, Robertson MM, et al. PROMISE (Program Refinements to Optimize Model Impact and Scalability based on Evidence): a cluster-randomised, stepped-wedge trial assessing effectiveness of the revised versus original Ryan White Part A HIV Care Coordination Programme for patients with barriers to treatment in the USA. *BMJ Open* 2020;10:e034624. doi:10.1136/bmjopen-2019-034624.

Table 1
TVS outcomes in PROMISE

Block	*Site*	*Outcome*	*Period 0*	*Period 1*	*Period 2*	*Total*
1	1	*TVS*	7	26	12	45
1	1	*No TVS*	10	18	10	38
		Total	17	44	22	83

Block	*Site*	*Outcome*	*Period 0*	*Period 1*	*Period 2*	*Total*
1	2	*TVS*	5	1	13	19
1	2	*No TVS*	4	3	6	13
		Total	9	4	19	32

Block	*Site*	*Outcome*	*Period 0*	*Period 1*	*Period 2*	*Total*
2	1	*TVS*	5	25	18	48
2	1	*No TVS*	12	20	14	46
		Total	17	45	32	94

Block	*Site*	*Outcome*	*Period 0*	*Period 1*	*Period 2*	*Total*
2	2	*TVS*	18	11	33	62
2	2	*No TVS*	6	5	11	22
		Total	24	16	44	84

Block	*Site*	*Outcome*	*Period 0*	*Period 1*	*Period 2*	*Total*
3	1	*TVS*	5	7	3	15
3	1	*No TVS*	4	3	2	9
		Total	9	10	5	24

Block	*Site*	*Outcome*	*Period 0*	*Period 1*	*Period 2*	*Total*
3	2	*TVS*	9	10	17	36
3	2	*No TVS*	1	4	8	13
		Total	10	14	25	49

Block	*Site*	*Outcome*	*Period 0*	*Period 1*	*Period 2*	*Total*
4	1	*TVS*	1	10	2	13
4	1	*No TVS*	2	11	0	13
		Total	3	21	2	26

Block	*Site*	*Outcome*	*Period 0*	*Period 1*	*Period 2*	*Total*
4	2	*TVS*	3	2	2	7
4	2	*No TVS*	2	6	8	16
		Total	5	8	10	23

Blocks (*Site* 1; *Site* 0):

1 = (Services for the Underserved, Inc.; Argus Community, Inc.)
2 = (Research Foundation of State University of New York; HHC Kings County Hospital Center)
3 = (Wyckoff Heights Medical Center; Sunset Park Family Health Center Network of Lutheran Medical Center)
4 = (The Institute for Family Health; Community Health Project, Inc.)

Table 1 (continued)
TVS outcomes in PROMISE

Block	*Site*	*Outcome*	*Period 0*	*Period 1*	*Period 2*	*Total*
5	1	*TVS*	4	17	6	27
5	1	*No TVS*	16	39	17	72
		Total	20	56	23	99
Block	*Site*	*Outcome*	*Period 0*	*Period 1*	*Period 2*	*Total*
5	2	*TVS*	10	19	33	62
5	2	*No TVS*	9	10	35	54
		Total	19	29	68	116

Block	*Site*	*Outcome*	*Period 0*	*Period 1*	*Period 2*	*Total*
6	1	*TVS*	1	8	3	12
6	1	*No TVS*	0	14	4	18
		Total	1	22	7	30
Block	*Site*	*Outcome*	*Period 0*	*Period 1*	*Period 2*	*Total*
6	2	*TVS*	9	15	13	37
6	2	*No TVS*	3	3	6	12
		Total	12	18	19	49

Block	*Site*	*Outcome*	*Period 0*	*Period 1*	*Period 2*	*Total*
7	1	*TVS*	3	34	18	55
7	1	*No TVS*	3	28	19	50
		Total	6	62	37	105
Block	*Site*	*Outcome*	*Period 0*	*Period 1*	*Period 2*	*Total*
7	2	*TVS*	3	8	5	16
7	2	*No TVS*	3	3	16	22
		Total	6	11	21	38

Block	*Site*	*Outcome*	*Period 0*	*Period 1*	*Period 2*	*Total*
8	1	*TVS*	7	15	17	39
8	1	*No TVS*	3	11	17	31
		Total	10	26	34	70
Block	*Site*	*Outcome*	*Period 0*	*Period 1*	*Period 2*	*Total*
8	2	*TVS*	6	3	13	22
8	2	*No TVS*	2	1	13	16
		Total	8	4	26	38

Blocks (*Site* 1; *Site* 0):

5 = (St Lukes Roosevelt Hospital; Mount Sinai Medical Center)
6 = (HHC Jacobi Medical Center; HHC Elmhurst Hospital Center)
7 = (Bronx Lebanon Hospital Center; Beth Israel Medical Center)
8 = (Community Health Action of Staten Island and Housing Works, pooled; Asian and Pacific Islander Coalition on HIV AIDS, Inc.)

Now let T denote the *intervention* indicator variable taking the value 1 for the updated CCP program or 0 for the original program. Note that T is a deterministic function of *Site* and *Per* due to the stepped-wedge design, namely, $T = T(Site, Per) = I[Site=1]\cdot I[Per=1] + I[Per=2]$. T takes the value 1 for early switching sites in Period 1 and for both sites in Period 2, else it takes the value 0. The following underlying logit model for TVS outcomes is assumed to hold in the study population.

$$\log\frac{P[TVS = yes \mid block = b, Site, Per]}{P[TVS = no \mid block = b, Site, Per]} = \alpha^{(b)} + \beta T + \gamma^{(b)} I[Site = 1] + \pi_1^{(b)} I[Per = 1] + \pi_2^{(b)} I[Per = 2]$$

The intercept coefficient $\alpha^{(b)}$ is the log-odds on TVS at baseline (Period 0) for "late" switchers in block b; β is the intervention effect, equal to the log odds ratio on TVS vs. no TVS comparing the new CCP vs. the original CCP, adjusted for site and period effects; $\gamma^{(b)}$ is the site effect in block b, equal to the LOR on TVS (yes vs. no) comparing early vs. late switchers, adjusted for intervention and period effects; and $\pi_1^{(b)}$ and $\pi_2^{(b)}$ are the period effects in block b, equal to the LORs on TVS (yes vs. no) comparing Period 1 and Period 2, respectively, to Period 0, adjusted for intervention and site effects.

Note that the logit model contains no *Per* × *Site* interactions terms. Because a site's intervention T is a function of period and site as noted above, it follows the model also contains no *Intervention* × *Per* or *Intervention* × *Site* interactions.

As explained in Irvine et al. (2020), the nuisance parameters $\alpha^{(b)}$, $\gamma^{(b)}$, $\pi_1^{(b)}$, and $\pi_2^{(b)}$ can be eliminated from the analysis by fully conditioning on the following sufficient statistics (in addition to the total number of clients within each site in each period): (i) the total number of TVS outcomes across periods within each site; and (ii) the total numbers of TVS outcomes within periods in each block. The resulting conditional likelihood function depends only on the single intervention effect parameter β through the single remaining sufficient statistic S, which is the total number of TVS outcomes summed across early switching clinics in Period 1. The primary analysis is based on the conditional distribution of sufficient statistic S.

Results.

The value of the sufficient statistic is $S=142$ TVS outcomes, yielding a conditional maximum likelihood estimate $\hat{\beta}_c$ of the intervention effect (log odds ratio on TVS vs. no TVS comparing new CCP vs. original CCP) of $\hat{\beta}_c = -0.1302$, corresponding to an odds ratio (OR) of $\exp(\hat{\beta}_c) = 0.8779$. Thus, the new program provided slightly *lower* TVS rates than did the original program. The effect is not statistically significant: the exact, two-tailed *P*-value (by the point-probability method[2]) is 0.7445. The corresponding exact, test-based 95% confidence interval for the intervention LOR is (–0.7966, 0.5584) and that for the intervention OR $\exp(\beta)$ is (0.4509, 1.748).

In the protocol paper, we stated we would report results from sensitivity analyses based on the some large sample (normal theory) methods. For the Wald test: the standard error of $\hat{\beta}_c$ is 0.3277, yielding a Wald test *Z*-score of –0.3974 ($P=0.6911$) and an approximate 95% c.i. of (–0.7726, 0.5121) for the LOR and (0.4618, 1.669 for the odds ratio. For the score test: the conditional score statistic (observed sufficient statistic *S* minus its null expected value $E_0[S]$) is 142–143.2173 = –1.2173 with standard error 3.0629, yielding a score-test *Z*-score of –0.3974 ($P=0.6910$). The score-test-based approximate 95% confidence interval for the intervention LOR is (–0.7701, 0.5098) and that for the intervention OR is (0.4630, 1.664). For the generalized log-likelihood ratio statistics: the chi-squared statistic is 0.1544 on 1 df ($P=0.6944$). The likelihood ratio test-based approximate 95% confidence interval for the intervention LOR is (–0.7916, 0.5218) and that for the intervention OR is (0.4531, 1.685). These large-sample approximate results are close to the exact conditional results and do not alter the primary analysis inference.

Table 2 below provides *constrained* unconditional maximum likelihood estimates of the nuisance parameters from the underlying logit model where the intervention LOR is constrained at its conditional maximum likelihood value $\hat{\beta}_c = -0.1302$. A random-effects meta-analysis of the results from the eight blocks provides a summary estimate of the average random effect from the population

[2] See Section 2.7 of J.L. Fleiss, B. Levin, and M.C. Paik (2003), *Statistical Methods for Rates and Proportions, 3rd Ed.* (New York: John Wiley & Sons).

reflected in the trial sites. The method of Dersimonian and Laird was used to estimate the variance components.[3,4]

Columns 6-9 indicate that the period effects are not significantly heterogeneous. There is a modest, non-significant effect of Period 1 (vs. Period 0) (LOR=0.2860 with s.e. 0.2028, P=0.1585) and a small effect of Period 2 (vs. Period 0) (LOR=0.0822 with s.e. 0.2001, P=0.6812).

Table 2
Meta-analysis results from constrained maximum likelihood estimates of $\alpha^{(b)}$ (intercept log odds on TVS for later-switching sites in Period 0), $\gamma^{(b)}$ (site effects), and $\pi_1^{(b)}$ and $\pi_2^{(b)}$ (period effects).

Block	$\alpha^{(b)}$	s.e.($\alpha^{(b)}$)	$\gamma^{(b)}$	s.e.($\gamma^{(b)}$)	$\pi_1^{(b)}$	s.e.($\pi_1^{(b)}$)	$\pi_2^{(b)}$	s.e.($\pi_2^{(b)}$)
1	–0.0315	0.4946	–0.1880	0.4596	0.5752	0.5053	0.7099	0.5100
2	0.6778	0.3572	–1.0146	0.3425	0.5637	0.4387	0.6338	0.4110
3	1.3401	0.6100	–0.6092	0.5673	–0.1175	0.6963	–0.4079	0.6851
4	–0.3948	0.7729	1.0688	0.7038	–0.6536	0.8756	–0.3651	0.9684
5	0.0040	0.3746	–1.2462	0.3242	0.5763	0.4245	0.1197	0.4183
6	1.3511	0.6756	–1.4784	0.5480	–0.1059	0.8071	–0.3251	0.8021
7	–0.1573	0.6131	0.3146	0.4022	0.3047	0.6410	–0.3355	0.6416
8	1.0304	0.5813	–0.1331	0.4321	–0.3958	0.6584	–0.8248	0.5864
χ^2 for homogeneity on 7 dfs (P-value)	9.851 (0.1972)		20.757 (0.0041)		3.992 (0.7807)		7.122 (0.4163)	
Estimated variance of true effects	0.1092		0.3791		0		0.005694	
Estimated average of true effects	0.4552		–0.4774		0.2860		0.0822	
s.e.(average)	0.2202		0.2729		0.2028		0.2001	
Z-score	2.0668		–1.750		1.410		0.4108	
P-value	0.0387*		0.0802		0.1586		0.6812	

Columns 4 and 5 indicate that the block-specific site effects exhibit significant heterogeneity (chi-squared for homogeneity = 20.757 on 7 df, P=0.004). The estimated summary average of the random site effects was –0.4774 corresponding to an odds ratio on TVS (comparing early- vs. late-switching sites, adjusted for period and intervention effects) of 0.6204, which is not significantly different from zero (P=0.0802). There was thus only a suggestion that early-switching sites had lower TVS rates than late-switching sites, apart from period and intervention effects.

[3] R. Dersimonian and N. Laird (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials* 7:177-176.

[4] See Section 10.9.2 of *Statistical Methods for Rates and Proportions*, supra.

Columns 2 and 3 indicate that the base log odds on TVS for late-switching sites in Period 0 exhibit modest, non-significant heterogeneity (chi-squared for heterogeneity = 9.851 on 7 df, $P=0.1972$). The summary estimate of the average log odds on TVS is 0.4552, corresponding to an overall TVS rate of 0.6119.

Table 3 below presents the fitted rates of TVS from the constrained logit model. Note that in each block, the odds ratios comparing early- vs. late-switching sites in Period 0 are equal to those in Period 2, and are also equal to the odds ratios in Period 1 reduced by the intervention odds ratio of 0.8961, consistent with the underlying logit model.

Table 3
Fitted probabilities of TVS
(adjusted for site and period effects, assuming an intervention OR of 0.8779)

Block	Site	Period 0	Period 1	Period 2
1	1	0.4453	0.5561	0.5891
1	2	0.4921	0.6327	0.6337
2	1	0.4166	0.5242	0.5416
2	2	0.6633	0.7758	0.7652
3	1	0.6750	0.6185	0.5480
3	2	0.7925	0.7725	0.6904
4	1	0.6624	0.4726	0.5445
4	2	0.4026	0.2595	0.2911
5	1	0.2241	0.3108	0.2222
5	2	0.5010	0.6411	0.4984
6	1	0.4682	0.4102	0.3583
6	2	0.7943	0.7765	0.7101
7	1	0.5392	0.5822	0.4235
7	2	0.4608	0.5368	0.3491
8	1	0.7104	0.5918	0.4856
8	2	0.7370	0.6535	0.5188

Checking for heterogeneity of intervention effects across blocks.

In general, when a small overall summary effect arises upon combining evidence across several blocks of data, it is important to check whether the effects are homogeneous across blocks and small in most blocks, or, on the contrary, whether the small summary effect results from large, heterogeneous effects in opposite directions which happen to cancel out. The latter case would constitute an important qualitative interaction and would render the overall finding of a null effect misleading.

We checked for that possibility by estimating the fully conditional intervention log odds ratio in each of the eight blocks and using a 7 df chi-squared test of homogeneity.[5] Table 4 presents the results. Though the estimated intervention effects vary from block to block, the individual standard errors are large and the chi-squared statistic was close to its degrees of freedom ($X^2_{\text{homog}} = 8.9972$ on 7 df, $P=0.2529$). We conclude that there is no significant heterogeneity in true intervention effects and that it is reasonable to summarize the overall intervention effect as reported above ($\hat{\beta}_c = -0.1302$, s.e. $= 0.3277$).[6]

Table 4
Block-specific fully conditional maximum likelihood estimates of the effect of intervention

Block	Site	Log Odds Ratio	s.e.
1	1	1.9466	1.2626
1	2		
2	1	0.6669	0.7271
2	2		
3	1	0.9018	1.1233
3	2		
4	1	0.0514	1.3314
4	2		
5	1	−0.3090	0.6430
5	2		
6	1	−1.3370	1.0895
6	2		
7	1	−1.5767	0.8745
7	2		
8	1	−0.6793	1.2698
8	2		

[5] See Section 10.1 of *Statistical Methods for Rates and Proportions*, supra.

[6] We note that the log odds ratios are positive in the first four blocks and negative in the final four blocks, but there is no apparent substantive reason for this pattern.

Assessing the key assumption of no period-by-site interaction.

All of the preceding analyses were prepared under the key assumption of no period-by-site interaction within blocks in the underlying logistic regression model. The stepped-wedge design allows for a partial check on that assumption, as follows. Let $\lambda_j^{(b)}$ denote the log odds ratio on *TVS* (yes vs. no) comparing the earlier-switching Site 1 vs. the later-switching Site 2 for block b in period j ($b=1,\ldots,8$ and $j=0$, 1, and 2). The underlying model implies that $\lambda_0^{(b)} = \lambda_2^{(b)} = \gamma^{(b)}$ and $\lambda_1^{(b)} = \gamma^{(b)} + \beta$. If, on the other hand, the underlying model were to require additional non-zero interaction terms, say,

$$\ldots + \delta_1^{(b)} I[Site=1] I[Per=1] + \delta_2^{(b)} I[Site=1] I[Per=2],$$

then we would have $\lambda_0^{(b)} = \gamma^{(b)}$, $\lambda_1^{(b)} = \gamma^{(b)} + \delta_1^{(b)} + \beta$, and $\lambda_2^{(b)} = \gamma^{(b)} + \delta_2^{(b)}$. Thus, by considering the differences $\lambda_2^{(b)} - \lambda_0^{(b)}$ in TVS-by-site log odds ratios between Period 2 and Period 0, we may test the hypothesis of homogeneous *Per 2 × Site* interactions ($\delta_2^{(1)} = \cdots = \delta_2^{(8)} = \delta_2$, say) and, if found homogeneous, we may test whether the assumed constant interaction δ_2 differs significantly from zero.

We stated above that the stepped-wedge design allows only a partial check on the key assumption because the *Per 1 × Site* interactions are aliased with the intervention effect β, such that only the sums $\delta_1^{(b)} + \beta$ are estimable. We are therefore limited to contemplating some reasonable values for the *Per 1 × Site* interactions and assessing the impact of those on the intervention effect.

For the test of homogeneity in $\delta_2^{(b)}$, we first consider only Period 0 and 2 data, because doing so requires the fewest assumptions and allows us to avoid the aliasing issue in Period 1. Using exact log odds ratio regression[7] to analyze the eight pairs of fourfold tables extracted from Table 1 corresponding to TVS (yes or no) cross-classified by Site (1 or 2) in Periods 0 and 2, we find the conditional maximum likelihood estimates of $\delta_2^{(b)}$ to be –0.019, 1.082, 1.529, 0.478, 1.090, and 0.237 with standard errors 1.081, 0.731, 2.455, 0.789, 1.586, and 1.347 for blocks $b=1$, 2, 3, 5, 7, and 8, respectively. (Blocks 4 and 6 had singular mle's, $\hat{\lambda}_2^{(4)} = \hat{\lambda}_0^{(6)} = \infty$, and so were uninformative

[7] See, e.g., Breslow, N., Regression analysis of the log odds ratio: A method for retrospective studies. Biometrics 32, 409-416 (1976). Conditional logistic regression may be used equivalently; see Section 12.4.2 at equation (11.64) or Section 14.3.2 of *Statistical Methods for Rates and Proportions*, supra.

concerning the corresponding interactions.) The chi-squared test of homogeneity is 1.268 on 5 dfs (P=0.938) and we conclude it is reasonable to assume a common *Per 2 × Site* interaction δ_2. Under that assumption, all eight blocks of data become informative and the conditional maximum likelihood estimate of the common interaction is $\hat{\delta}_2 = 0.7215$ with s.e. 0.4153 (P=0.0823).

Next, consider estimating $\delta_1 + \beta$ and δ_2 simultaneously from all three periods under the assumption of a common value of $\delta_1^{(1)} = \cdots = \delta_1^{(8)} = \delta_1$ using conditional logistic regression for the model $\lambda_0^{(b)} = \gamma^{(b)}$, $\lambda_1^{(b)} = \gamma^{(b)} + \delta_1 + \beta$, and $\lambda_2^{(b)} = \gamma^{(b)} + \delta_2$. Here the conditional mle of δ_2 is somewhat larger than when using only two periods of data and the estimate is of nominal significance: $\hat{\delta}_2 = 0.825$ with s.e. 0.4115 (P=0.0449). The maximum likelihood estimate of $\delta_1 + \beta$ is 0.420 with s.e. 0.4352. If we allow for a non-zero value of δ_2 and assume δ_1 takes the same value of 0.825 as estimated for δ_2, then the estimated intervention effect would be a bit worse than the primary analysis result of –0.130, namely, 0.420 – 0.825 = –0.405, with the same standard error of 0.4352 (because only $\delta_1 + \beta$ is estimable). If, on the other hand, we were to assume a true value of $\delta_1 = 0$, then the estimated intervention effect would become positive, namely, +0.420, again with the same standard error of 0.4352. Other assumptions, e.g., a growth model in which the *Per 1 × Site* interaction is half of δ_2, would lead to an estimated intervention effect close to zero.

The bottom line is that all of the above estimates lie within the 95% confidence interval for the intervention effect in the primary analysis under the key assumption, and because none of the above estimates reaches statistical significance at the 5% level, it is reasonable to conclude that the null findings of the primary analysis are not substantively affected by modest departures from the key assumption.