Test Power for Drug Abuse Surveillance

Jarad Niemi[1], Meredith Smith[2], and David Banks[1]

[1] Duke University, Durham, NC 27708 USA
[2] Purdue Pharma L.P., Stamford, CT 06901 USA
(now at Abbot Laboratories, Abbot Park, IL 60064 USA)

Abstract. Syndromic surveillance can be used to assess change in drug abuse rates and to find regions in which abuse is most common. This paper compares the power of three syndromic surveillance procedures (a paired-sample test, a process control chart, and a conditional autoregressive model) for detecting change in opioid drug abuse patterns, using data from two reporting systems (the OTP and PCC datasets). We find that the conditional autoregressive model provides good power and geographic information and that the OTP data carry the strongest signal.

1 Introduction

The substantial rise in nonmedical use and abuse of prescription opioid analgesics over the past decade offers an important opportunity for comparing syndromic surveillance methods [4]. Prescription drug abuse has similarities to infectious disease due to the inherent geographical effects [3], multiple reporting systems which vary in coverage and data quality, and it is a major concern in public health management. Prescription opioid analgesic abuse cost the U.S. an estimated $8.6 billion in 2001 due to increased health care, workplace, and criminal justice costs [1].

This paper contrasts three different strategies for syndromic surveillance:

- A paired-difference two-sample test, which looks for differences in abuse rates over time at each reporting site.
- A sequential process control procedure, using the CUSUM chart, similar to that used by the CDC [9].
- A conditional autoregressive (CAR) model which incorporates covariates as well as a model for geographic dependence.

These methods are compared with respect to their power in detecting simulated signal using historical abuse data and in their ability to detect hot spots of this abuse.

The data sets used in this study operate under the auspices of the Researched Abuse, Diversion, and Addiction-Related Surveillance (RADARS®) system:

- OTP. The Opioid Treatment Programs study collects quarterly questionnaires from abusers enrolled in Methadone Maintenance Treatment Programs (MMTPs) and thus captures a key population of sophisticated abusers.

D. Zeng et al. (Eds.): BioSecure 2008, LNCS 5354, pp. 131–142, 2008.
© Springer-Verlag Berlin Heidelberg 2008

– PCC. The Poison Control Center network records information on help calls resulting from intentional drug exposures; not all poison control centers participate, but its coverage is about 70% of the U.S. by population.

Besides these databases we also considered: the National Survey on Drug Use and Health (NSDUH), an annual federal survey; Monitoring the Future (MTF), a nationally representative cohort study of self-reported drug use by 8th, 10th, and 12th grade students, who are then followed in biennial surveys until age 29; and the Drug Abuse Warning Network (DAWN), which provides an annual cross-sectional sample of Emergency Department visits related to nonmedical use of drugs. Although NSDUH and MTF survey 70,000 and 50,000 respondents respectively, the percentage of these individuals abusing opioids is small. Therefore their effective sample sizes for detecting change in a syndromic surveillance program would be inadequate. The effective sample size in DAWN is larger, but DAWN studies only a few major metropolitan areas in the country and therefore spatial information is limited.

We focused on one specific medication, the opioid analgesic OxyContin® (oxycodone HCI, controlled-release) Tablets, since it has been the target of abuse over several years [2]. We focus on change detection for a one-sided alternative which specifies that the drug abuse rate has decreased over time. This approach is simpler than two-sided alternatives, reflects federal interest in measuring the effectiveness of drug prevention programs, and our results extend directly to the symmetric hypothesis that drug abuse has increased.

All power studies were performed by simulation. For each combination of database and analysis, we examined power as a function of simulated levels of abuse reduction. The simulations were performed by bootstrapping [5] from the original data sets, after adjustment to achieve specified reduction levels.

Our goals are to determine the tradeoffs among the three analyses, in terms of power, geographic localization, and operational requirements (computing time, statistical complexity). We also want to determine the tradeoffs among the two databases used in this study. Section 2 describes the methodology; Section 3 presents the results; and Section 4 summarizes the comparisons.

2 Methodology

The methods developed in this paper follow this protocol: 1) a generative model is assumed for the data, 2) simulations with artificial signal are generated from this model, and 3) multiple surveillance techniques are applied to the resulting data. In the OTP data, the generative model is a CAR model including covariates. In the PCC data, the generative model is a log-linear model. In both data sets, the surveillance techniques used include a two-sample test and a process control chart. We also use the CAR and the log-linear model for surveillance in the OTP and PCC data sets, respectively, but the models include a term to detect the difference in abuse between time points.

2.1 Statistical Tests

Many statistical approaches to syndromic surveillance could be considered: repeated measures MANOVA, longitudinal analysis, time series analysis, various regression models, process control charts, two-sample tests, and so forth. In picking the three methods used in this paper, we sought transparency as well as adequate statistical power.

Two-Sample Tests. These tests look for a change between two time points. We use the classic one-sided test for a difference in binomial proportions. The test statistic is:

$$z = (\hat{p}_1 - \hat{p}_2)/\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

where $\hat{p}_1$ is the observed proportion of abusers in the previous quarter and $\hat{p}_2$ is the observed proportion in the current quarter. This test statistic is referred to a standard normal table.

The two-sample test can be improved when the same sites report each quarter. If there are k such sites, one can perform the two-sample test separately at each, and pool the resulting P-values according to Fisher's rule [6]. Let p_i, for $i = 1, \ldots, k$, be the P-value for site i; then

$$\chi^2_{2k} = -2\sum_{i=1}^{k} \ln p_i$$

which can be referred to a chi-squared table. Thus many slightly significant reductions can be pooled to give stronger evidence of a reduction. Repeated use requires adjustment for multiple testing. To give an indication of geographical variability in the reductions, one can map the P-values by region.

Process Control Charts. Control charts check whether a succession of observations has drifted away from a baseline value. A CUSUM chart plots the sum of the differences between the previous quarters' proportions and a baseline proportion. As described in [8], when this sum falls below a lower control line, the result is statistically significant. The CUSUM procedure is more complex than the two-sample test or the Shewhart control chart, but still fairly simple.

Control charts assume that the baseline is fixed and known. This is reasonable in manufacturing, but in syndromic surveillance, we do not know the baseline abuse rate; we can only estimate this, with uncertainty, from historical data. The assumption of no trend in this historical data is critical.

CAR Models. The previous tests make no use of covariate information. If covariates are influential, then a regression model should have more power and greater ability to notice specific regions with unusual behavior.

We use the Generalized Linear Model (GLM) in conjunction with the Conditional Autoregressive (CAR) model. Our GLM uses the logit function (log odds) to linearize the dependence of a proportion upon covariates [7].

CAR models include spatial or temporal dependence through a neighborhood structure, so that reporting units that are near each other have correlated data. This is reasonable in syndromic surveillance. In our application, there are known hot spots of opioid drug abuse in Appalachia and Maine. For this analysis, we aggregated the geographic information to the state level. The methodology could be extended to other spatial resolutions, but we found this aggregation to be effective for our purposes

Inference is done through Markov chain Monte Carlo [10], but in a problem of this scale there are computational challenges. Sometimes it took a full day of computing to provide a single point in the power curve.

2.2 Simulation Procedures

All of the power curve figures in Section 3 were produced by simulation. These simulations were produced using either the model in (1) or (2) where the mean was multiplied by the appropriate fraction to produce, on average, a linear decline in abuse over three years. The parameters used in these simulations were drawn from the posterior distributions for the parameters using only the pre-intervention data. We focus on power for two significance levels: $\alpha = .05$ and $\alpha = .05^2 = .0025$ that bracket loose and stringent levels for Type I error. Each plotted point is based upon 200 simulations with a specific, simulated decrease in abuse from the previous historical record in each of the databases.

The process control chart simulations assume that, after the last historical quarter, the abuse rate in subsequent quarters drops linearly over three years to a new level that was 5%, 10%, 15% or 20% lower. Extensive pre-simulation runs were made to estimate the lower control line values for these charts.

3 Results of the Power Analyses

The following subsections describe the power curves. There is a short review of each dataset and special analytic issues that they pose.

3.1 OTP

The OTP data derive from questionnaires administered to abusers enrolled in selected Methadone Maintenance Treatment Programs (MMTPs). It captures information on opioid abuse in the past 30 days, the primary drug, and geographic/demographic information. The data are quarterly in 2005.

OTP Two-Sample Test. Figure 1 shows the estimated power of the two-sample test using OTP data with Fisher's test for change at MMTP clinics that appear in both time periods. This "blocking" of an MMTP with itself automatically controls for many biases and reduces the variance in comparisons.

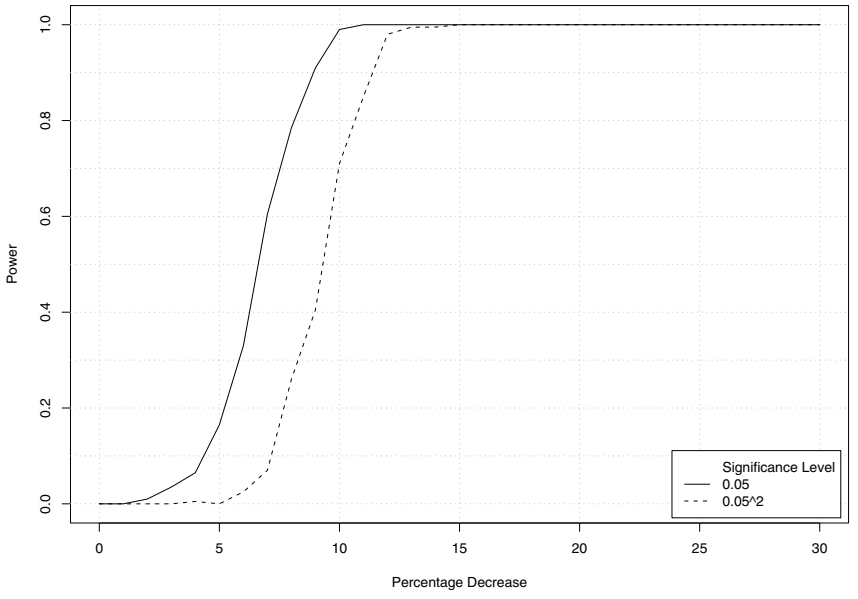

Fig. 1. Power curves for a two-sample OTP test. The most recent four quarters are combined to give the pre-sample abuse level.

OTP Control Chart. For control charts, one cannot plot power for all possible values of reduction (percentage decrease). So Figure 2 plots the probability of rejection for four different reduction levels: 20%, 15%, 10% and 5%, reading the curves from left to right.

Interpreting Figure 2 requires some care. Note that the two-sample tests use a year's worth of data, whereas the quarterly data has necessarily smaller sample size. Looking at the power in the fourth quarter gives a basis for comparison, but recall that control charts do not adjust for multiple testing.

OTP CAR Model. OTP data capture age, gender, race, and location. This enables use of a CAR model that incorporates spatial correlation structure.

The CAR model used in this power study is:

$$\begin{aligned} Y_{ik}(t) &\sim \text{Bernoulli}(p_{ik}(t)) \\ \text{logit}(p_{ik}(t)) &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + b_k \end{aligned} \tag{1}$$

where $Y_{ik}(t)$ is the outcome for individual i living in state k (i.e., it is 1 if the individual has used an oxycodone product in the past 30 days and 0 otherwise). The covariates describe gender, race, and age, respectively, where race has been dichotomized to white or non-white and age has been broken down into 17 age categories. Flat priors were used for all coefficient parameters. The spatial random effect for U.S. state are assumed to have a CAR prior distribution

$$\pi(b_1, \ldots, b_K) \propto \exp\left(-\frac{\tau}{2}\sum_{i \neq j} w_{ij}(b_i - b_j)^2\right)$$

where τ is the precision parameter, given a Gamma(1,1) prior, and w_{ij} is obtained from the matrix of binary weights that indicate whether two states share a border. We examined the use of a quadratic term in age, but it was not significant and thus excluded from this model.

Table 1 shows the posterior credible regions (Bayesian confidence intervals) on the coefficients for gender, race, and age. The estimated value for the gender coefficient is .46 and its credible region excludes 0, so men are more likely to abuse opioid drugs. Similarly, the coefficient of 1.15 on race means that whites are more likely to abuse opioids. The age coefficient is negative, so the elderly are less likely to abuse.

The location terms b_k are random effects, and correlated. Since data are aggregated by state, the correlation structure is coarse: two states directly interact if they are contiguous; otherwise, states are conditionally independent given their neighbors. (Alaska is the only reporting state that had no neighbors.) The model for the correlation is multivariate Gaussian with unknown but common correlation for states that share boundaries.

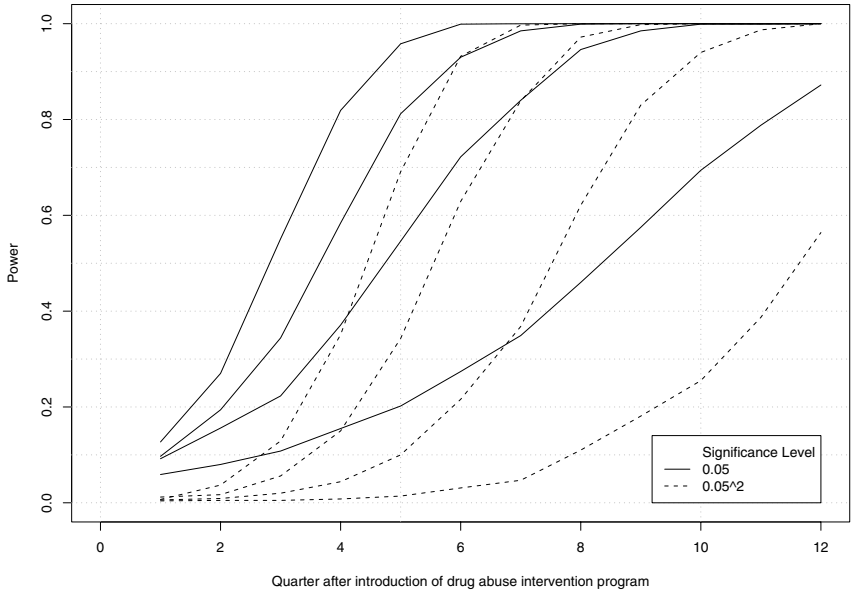

Fig. 2. Power curves for CUSUM testing with OTP data. The solid (broken) lines correspond to .05 (.0025) level tests. The lines, reading from the bottom up, correspond to 5%, 10%, 15% and 20% reductions in abuse rate, pro-rated over three years.

Table 1. Percentile points of the posterior distributions on the fixed-effect terms in the CAR model

	2.5%	50%	97.5%
Intercept	-2.2	-1.7	-1.2
Gender	0.31	0.46	0.58
Race	0.89	1.15	1.47
Age	-0.051	-0.043	-0.034
CAR precision	0.19	0.36	0.64

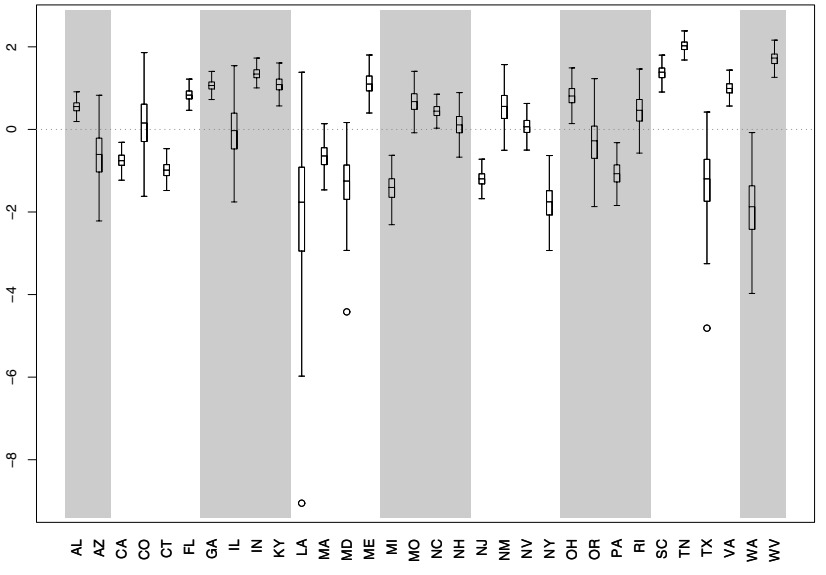

Fig. 3. A display, by state, of the 95% credible regions on the location (or state) effect. If the range of the line for a state straddles zero on the y-axis, then there is more than a 5% chance that that state has no effect on the abuse rate, after accounting for gender, race and age. The central boxes contain the middle 50% of the probability mass for the magnitude of the state effect, and the midline within the box is the point estimate of the magnitude of the state effect.

Figure 3 shows estimates of the state effects in the CAR model. Three states with large positive effects were Tennessee, West Virginia, and Virginia. This accords with previous reports of high opioid abuse rates in Appalachia. Some states, such as California and Connecticut, have lower than expected rates of opioid abuse. The wide interval for Louisiana surely reflects uncertainty in the data due to Hurricane Katrina.

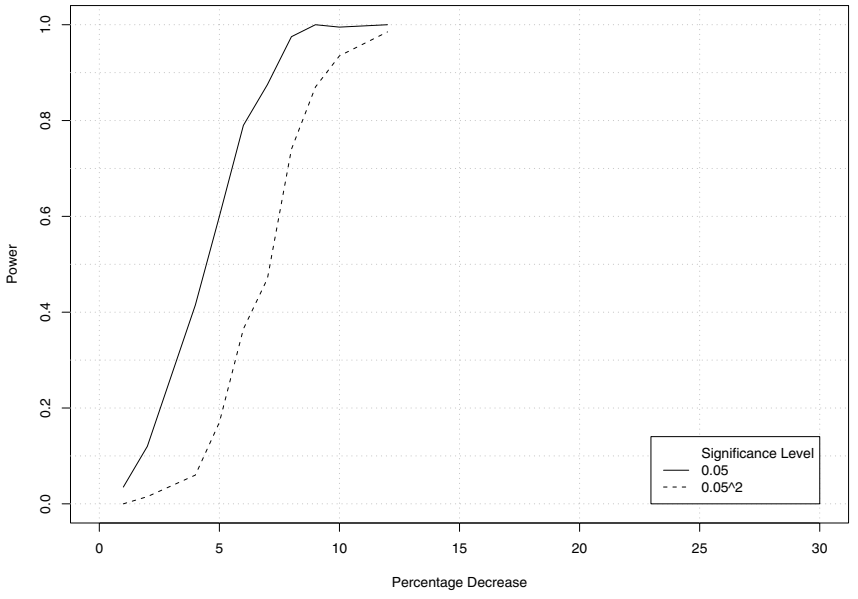

Fig. 4. Power curves for a CAR model test of abuse reduction using OTP data. The solid line is for an alpha level of .05; the dashed line is for an alpha level of .0025.

To test for a drop in abuse rate, the CAR model in equation (1) is modified to include a term for time. If the time coefficient is significantly less than zero, this indicates one-sided change (reduction). The magnitude of the effect can be estimated from the coefficient.

The CAR model was fit using Gibbs sampling run through WinBugs with an R interface. The OTP data takes a relatively long time to run (about 10 minutes per simulation run) so the power is only calculated for reductions of 1%, 2%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, and 12%, as shown in Figure 4.

3.2 PCC

The PCC network data consists of calls to regional poison control centers (usually funneled through 911 calls) regarding intentional exposure to drugs. The network covers approximately 70% of the U.S. population. The geographic counts are aggregated at the 3-digit zip code (3DZ) level. The data are highly reliable in terms of the identification of specific drugs, since PCC operators usually obtain the NDC code from the pharmacy label, but demographic data are not consistently captured.

The main explanatory variable in the PCC data is the 3DZ location. We use this to leverage information on the estimated population and the number of unique recipients of dispensed drugs (URDDs) in the region. The fitted model

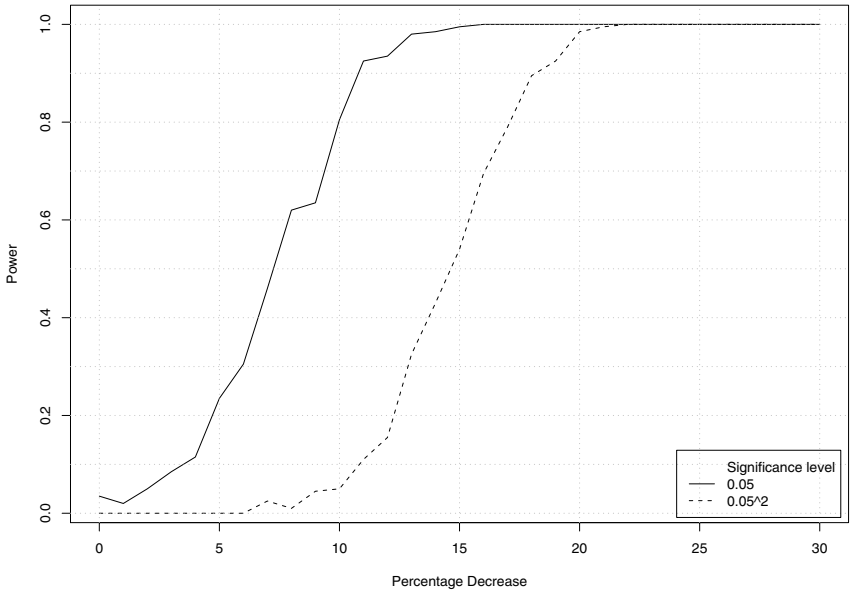

Fig. 5. The solid (broken) line is the probability of rejecting the null hypothesis of no reduction in opioid abuse at the .05 (.0025) level using PCC data with a two-sample test when, in fact, the reduction is as shown on the x-axis

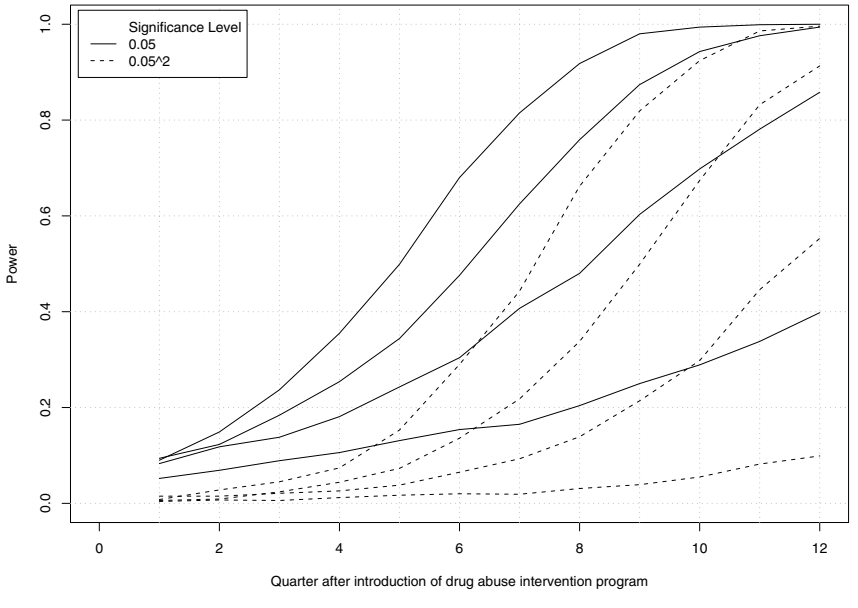

Fig. 6. The power curve for a CUSUM test with PCC data. The solid (broken) lines correspond to .05 (.0025) level tests. The lines, reading from the bottom up, correspond to 5%, 10%, 15% and 20% reductions in abuse rate, pro-rated over three years.

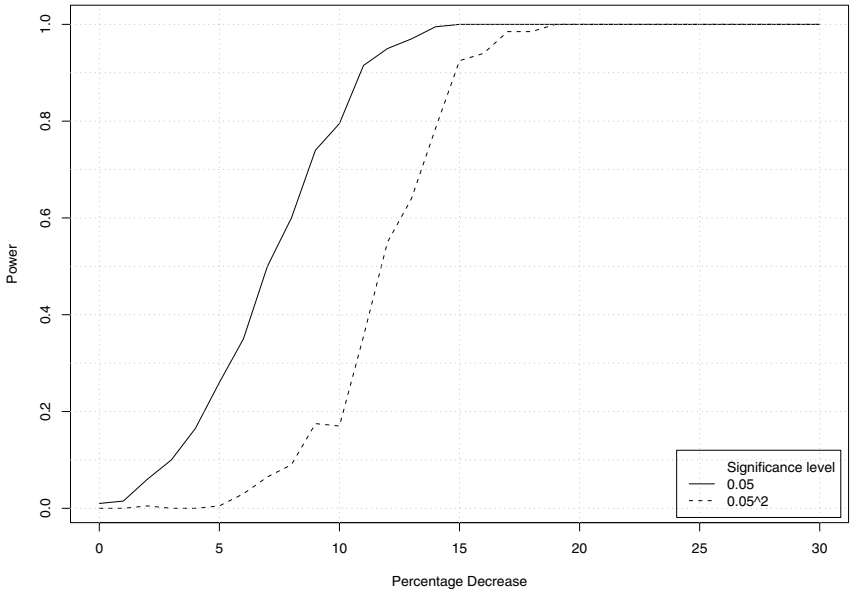

Fig. 7. The solid (broken) line is the probability of rejecting the null hypothesis of no reduction in opioid abuse at the .05 (.0025) level using PCC data and the GLM test when, in fact, the reduction is as shown on the x-axis

assumes the observed count Y_i of specific opioid PCC calls in the ith 3DZ has Poisson distribution with parameter λ_i given by:

$$\ln(\lambda_i) = \alpha \ln(\text{URDD}_i) + \beta \ln(\text{Pop}_i) \tag{2}$$

The population size and URDD provide rough "denominators" for the group at risk for opioid abuse.

PCC Two-Sample Test. After fitting, we used equation (2) to simulate observations for future quarters by drawing counts from Poisson distributions with decremented values for the estimated λ_i such that, on average, the number of simulated abuses decreased to a specified amount linearly over three years. The test statistic for deciding whether there has been a change is:

$$t_{n-1} = \bar{\Delta} / \sqrt{s_\Delta^2/(n-1)}$$

where n is the number of 3DZs and $\bar{\Delta}$ is the average regional difference in the count at the historical baseline of surveillance and the count afterwards. The s_Δ^2 is the sample variance of the observed differences.

In calculating power, we used only the 571 3DZs that reported in all quarters of 2005. Figure 5 shows the result.

PCC Control Chart. Using the simulation procedure described previously and thus taking advantage of information on URDD and population size, we generated data for the CUSUM test in Figure 6. The mean of the initial year was the baseline against which deviations should be discovered.

PCC Regression Model. We fit a GLM model using URDD and population size as covariates, with the addition of a term for time (as we did with the CAR model for OTP). The effort needed to map the 3DZs to a state-level adjacency matrix prevented a full CAR analysis, although this could be done. Figure 7 gives power curves for tests at the .05 and .0025 levels.

4 Conclusions

Disease surveillance seeks to identify a sudden increase in the prevalence of an illness against a background of relatively low rates. But drug abuse usually has a higher background prevalence which does not increase drastically. So drug abuse surveillance focuses on finding decreases in abuse that document successful intervention investments. This paper has compared methods and datasets that support that objective.

For OTP data, on a three-year horizon, the CUSUM is more powerful than the CAR test, which is more powerful than the two-sample test. But the CUSUM does not adjust for multiple testing, so its apparent power is misleading. Also, the three-year time frame gives it a larger effective sample size than the CAR or two-sample tests. For the PCC data, the same conclusions and caveats apply. The regression test is better than the two-sample test, but both lose to the CUSUM, which enjoys unfair advantages. (The CUSUM makes 12 tests, one for each quarter, all at a .05 level. So the overall probability of Type I error is actually $1 - (1 - .05)^{12} = .46$.)

In comparing the two data sets, OTP has a larger effective sample size than the PCC, so procedures that use OTP will generally be more powerful. Additionally, in this analysis, the regional information was more accessible and could be meaningfully interpreted for the OTP data.

People who contact a Poison Control Center are probably less sophisticated abusers than OTP clients. This may make them of greater (or less) public health interest. Data quality is also a issue. PCC data usually include the NDC code, but OTP data are self-reports from addicts based on recall. Other datasets (NSDUH, MTF, and DAWN) have similar data quality concerns.

References

1. Birnbaum, H.G., White, A.G., Reynolds, J.L., Greenberg, P., Mingliang, Z., Vallow, S., Schein, J., Katz, N.P.: Estimated costs of prescription opioid analgesic abuse in the United States in 2001: A societal perspective. Clinical Journal of Pain 22, 667–676 (2006)

2. Carise, D., Dugosh, K., McLellan, A.T., Camilleri, A., Woody, G., Lynch, K.G.: Prescription OxyContin abuse among patients entering addiction treatment. American Journal of Psychiatry 164, 1750–1756 (2007)
3. Cicero, T.J., Inciardi, J.A., Munoz, A.: Trends in the abuse of OxyContin and other opioid analgesics in the United States: 2002-2004. Journal of Pain 6, 662–672 (2005)
4. Compton, W.M., Volkow, N.D.: Major increases in opioid analgesic abuse in the United States: Concerns and strategies. Drug and Alcohol Dependency 81, 103–107 (2006)
5. Efron, B.: Bootstrap methods: Another look at the jackknife. Annals of Statistics 7, 1–26 (1979)
6. Fisher, R.: Combining independent tests of significance. The American Statistician 2, 30 (1948)
7. McCullagh, P., Nelder, J.: Generalized Linear Models. Chapman and Hall, London (1989)
8. Montgomery, D.: Introduction to Statistical Quality Control. John Wiley & Sons, Inc., New York (2001)
9. Rolka, H., Burkom, H., Cooper, G.F., Kulldorff, M., Madigan, D., Wong, W.-K.: Issues in applied statistics for public health bioterrorism surveillance using multiple data streams: Research needs. Statistics in Medicine 26, 1834–1856 (2007)
10. Waller, L.A., Carlin, B.P., Xia, H., Gelfand, A.: Hierarchical spatio-temporal mapping of disease rates. Journal of the American Statistical Association 92, 607–617 (1997)