Title
Does accounting for seizure frequency variability increase clinical trial power?

Authors
Goldenholz, Daniel M
Goldenholz, Shira R
Moss, Robert
et al.

Publication Date
2017-11-01

DOI
10.1016/j.eplepsyres.2017.07.013

Peer reviewed

HHS Public Access

Author manuscript

Epilepsy Res. Author manuscript; available in PMC 2018 November 01.

Published in final edited form as:

Epilepsy Res. 2017 November ; 137: 145–151. doi:10.1016/j.eplepsyres.2017.07.013.

Does accounting for seizure frequency variability increase clinical trial power?

Daniel M Goldenholz, MD, PhD[1,10], **Shira R. Goldenholz, MD, MPH**[10], **Robert Moss**[2], **Jacqueline French, MD**[3], **Daniel Lowenstein, MD, PhD**[4], **Ruben Kuzniecky, MD**[3], **Sheryl Haut, MD**[5], **Sabrina Cristofaro**[3], **Kamil Detyniecki, MD**[6], **John Hixson, MD**[4], **Philippa Karoly, MS**[7], **Mark Cook, MBSS, MD**[7], **Alex Strashny, PhD**[8], **William H Theodore, MD**[1], and **Carl Pieper, DrPH**[9]

[1]Clinical Epilepsy Section, NINDS, NIH

[2]SeizureTracker LLC

[3]New York University

[4]UCSF

[5]Montefiore Medical Center/Albert Einstein College of Medicine

[6]Yale University

[7]University of Melbourne

[8]Centers for Disease Control

[9]Duke University Medical Center, Dept. of Biostatistics and Bioinformatics

[10]Beth Israel Deaconess Medical Center

Abstract

Objective—Seizure frequency variability is associated with placebo responses in randomized controlled trials (RCT). Increased variability can result in drug misclassification and, hence, decreased statistical power. We investigated a new method that directly incorporated variability into RCT analysis, Z_V.

Methods—Two models were assessed: the traditional 50%-responder rate (RR50), and the variability-corrected score, Z_V. Each predicted seizure frequency upper and lower limits using prior seizures. Accuracy was defined as percentage of time-intervals when the observed seizure frequencies were within the predicted limits. First, we tested the Z_V method on three datasets (SeizureTracker: n=3016, Human Epilepsy Project: n=107, and NeuroVista: n=15). An additional

Corresponding author: Daniel Goldenholz, National Institutes of Health, NINDS, Clinical Epilepsy Section, CNP, DIR, 10 Center Drive, 10-CRC, Room 5S-207, MSC 1408, Bethesda MD 20892-0001.

CONFLICTS OF INTEREST

None of the authors have any disclosures.

independent SeizureTracker validation dataset was used to generate a set of 200 simulated trials each for 5 different sample sizes (total N=100 to 500 by 100), assuming 20% dropout and 30% drug efficacy. “Power” was determined as the percentage of trials successfully distinguishing placebo from drug ($p<0.05$).

Results—Prediction accuracy across datasets was, Z_V: 91–100%, RR50: 42–80%. Simulated RCT Z_V analysis achieved >90% power at N=100 per arm while RR50 required N=200 per arm.

Significance—Z_V may increase the statistical power of an RCT relative to the traditional RR50.

Keywords

Epilepsy; clinical trials; placebo effect; prediction; seizure frequency; natural variability

1. Introduction

There is a need for new epilepsy drugs, given the 35% prevalence of drug-resistant epilepsy[1,2]. However, drug development remains challenging due to high expense and frequent trial failure. Trials suffered from rising placebo response rates over the past several decades[3], typically 4–27%[4] but recently up to 40%[5]. This can translate into unsuccessful trials[6], increased sample size, and increased development costs[7]. Seizure frequency variability at the patient level, typically unreported, may explain a significant portion of placebo responses, because natural frequency fluctuations are sufficiently large to produce a “response” even without treatment[8]. Uncertainty about variability may hamper randomized clinical trial (RCT) interpretation. With current methods, variability represents “noise” obscuring the drug efficacy “signal”. With lower noise, trials are expected to cost less and have fewer failures.

The RR50 (the percentage of patients with 50% seizure reduction in each trial arm), is the preferred outcome measure of the European Medicines Agency (EMA)[9]. The U.S. Food and Drug Administration prefers median-%-change (MPC). Trials typically require co-primary RR50 and MPC endpoints. RR50 is less statistically efficient than MPC[10], and typically used in power calculations for patient enrollment. However, based on recent evidence, the RR50 likely overestimates clinically relevant measures[8]. Simulations based on 1767 patient seizure diaries show that many RCT 50%-responders may subsequently become non-responders due to large natural variability. Consequently, models incorporating expected variability may improve epilepsy RCT interpretability, generalizability, and efficiency. Obviously, such models would only be of use if adopted by regulatory agencies.

Standard clinical practice includes implicit judgments about natural variability as well. Physicians are expected to make medication changes based on whether seizure rates have exceeded some arbitrary upper bound. If a drug adjustment results in rate decreases below an arbitrary lower boundary, the adjustment is considered beneficial. For patients with years of seizure-freedom, variability computations are irrelevant. But if seizure-freedom is short-lived, measured over a short duration, or if the patient is not seizure-free, no formal clinical tools exist to calculate expected bounds on seizure rates.

Clinicians and trialists would benefit from a robust method for predicting natural seizure frequency variability. This study represents the first attempt to account for the impact of variability on seizure frequency measurements, using a multi-modal data-driven approach.

2. Materials and Methods

2.1. Overview

This work presents a novel method for assessing RCTs called Z_V (Methods 2.2). Z_V and RR50 were compared in their ability to predict seizure frequencies several months into the future (Figure 1, Methods 2.3). In three datasets, each patient diary was divided into 6-month intervals to mimic typical RCT duration[11]. In each interval, early seizure rates were used to predict later rates using RR50 and Z_V.

To assess Z_V utility in an RCT (Figure 2), we generated a set of simulated clinical trials based on realistic seizure data (Methods 2.4). Five sets of 200 trials each included 100, 200, 300, 400 or 500 patients. Statistical power was computed for each series, and each calculation method (RR50, MPC and Z_V), to determine the minimum number of patients needed for the trial to achieve 90% power for each method.

2.2. The variability-corrected Z_V method

The Z_V model assumed seizure frequency variability during both experimental and baseline periods remained unchanged. Typically, "seizure frequency" refers to a 28-day seizure count; here, we focus on 14-day seizure counts. For mathematical simplicity, an individual's seizure frequencies were assumed to follow a Gaussian distribution. Each patient's seizure count for each 2-week interval of time was represented by $C_{i,j}$, the count of the i^{th} interval in the j^{th} patient. Because we chose a 2-month baseline (Figure 1), 4 intervals of 2-weeks were considered. The model calculated an estimated mean ($\widehat{\mu_{\text{ðÌŽě}}}$), and standard deviation ($\widehat{\sigma_{\text{ðÌŽě}}}$) of the set of $C_{i,j}$'s during the baseline (Equations 1,2), with the 4 values of $C_{i,j}$ (M=4):

$$\widehat{\sigma_{\text{ðÌŽě}}} = \sqrt{\frac{1}{M-1}\sum_{i=1}^{M}(C_{i,j} - \widehat{\mu_{\text{ðÌŽě}}})^2} \quad \text{(Eq. 1)}$$

$$\widehat{\mu_{\text{ðÌŽě}}} = \frac{1}{M}\sum_{i=1}^{M} C_{i,j} \quad \text{(Eq. 2)}$$

The Z_V model assumes under the null hypothesis (i.e. no treatment effect) that the patient-specific experimental period standard deviation and mean ($\widehat{\widehat{\sigma_{\text{ðÌŽě}}}}$ and $\widehat{\widehat{\mu_{\text{ðÌŽě}}}}$) are equal to those of the baseline period. Equations 3 and 4 predicts the seizure frequency distribution during the 6 experimental phase intervals for null treatments:

$$\widehat{\widehat{\sigma_{\text{ðÌŽě}}}} = \widehat{\sigma_{\text{ðÌŽě}}} \quad \text{(Eq. 3)}$$

$$\widehat{\widehat{\mu_{\text{ðİŽě}}}}=\widehat{\mu_{\text{ðİŽě}}} \quad \text{(Eq. 4)}$$

During the 3-month experimental period, the parameter $Z_V(i, j)$ was computed for the i[th] 2-week interval seizure count ($C_{i,j}$) of the j[th] patient, in the form of a negative Z score (see Figure 2A):

$$Z_V(i,j)=-\left[\frac{C_{i,j}-\widehat{\mu_{\text{ðİŽě}}}}{\widehat{\sigma_{\text{ðİŽě}}}}\right] \quad where\, i \in [1 \ldots 6] \quad \text{(Eq. 5)}$$

Note that $Z_V(i,j)$ has a negative sign in front by convention, to ensure positive values when seizure frequency decreases (a central goal in epilepsy clinical trials). For instance, if a patient had a rate of 10±3 seizures/interval (i.e. $\widehat{\mu_{\text{ðİŽě}}}=10, \widehat{\sigma_{\text{ðİŽě}}}=3$) during baseline and $C_{1,j}$ was 1 during the first 2-week interval of the experimental phase then $Z_V(1,j)$ would be 3. In this context, 3 indicates that the seizure rate in 1[st] 2-week interval showed a *decrease* 3 standard deviations below the baseline rate. As a second example, if that patient's next 2-week interval with 13 seizures, the $Z_V(2,j)$ would be −1, representing an *increase* in expected rate by 1 standard deviation above baseline.

To obtain an overall estimate of trial success (Figure 2B), Z_V was compared between the two arms of a trial using mixed effects model with an order-1 autocorrelation (AR1) structure. Each patient can contribute up to 6 Z_V observations during the experimental phase, depending on whether and when dropout occurs. The mixed effects model controls for random effects of individual patients, and for (potential) correlation structure between repeated measures. It adjusts weights from patient data based on the amount of information present allowing for dropout. The AR1 structure was selected because there may be some degree of "memory" in longitudinal seizure counts[12].

In the unusual case where $\widehat{\sigma_{\text{ðİŽě}}}=0$, Z_V is undefined. This case occurs when all baseline $C_{i,j}$ values are identical. To prevent Z_V being undefined, the following equation is applied:

$$\text{if } \widehat{\sigma_{\text{ðİŽě}}}=0: \quad \widehat{\sigma_{\text{ðİŽě}}}=\sqrt{\frac{1}{M-1}} \quad \text{(Eq. 6)}$$

In the case of four baseline values (M=4), we force $\widehat{\sigma_{\text{ðİŽě}}}=1/\sqrt{3}$.

2.3. Prediction testing

2.3.1. Data—Data came from three patient diary databases (Table 1). Each dataset was managed in de-identified format, consistent with NIH Office of Human Subject Research Protections, Protocol #12301. For each dataset, data were redacted into diary format. Patients were not required to have fixed, unchanging medication regimens; some changed their medications often. Due to incomplete information on medication changing as well as medication compliance, these factors were not formally addressed.

The first study (NeuroVista) utilized subdural electrodes which were chronically implanted in 15 patients as part of a seizure warning system[13]. Despite low enrollment, the NeuroVista study represents the longest completely characterized seizure dataset available. All patients were adults with confirmed focal epilepsy. The data consisted of several types of seizures: type 1, which were clinical seizures (reported or confirmed to be clinical by audio review) that had electrographic correlation; type 2, unconfirmed clinical (unreported) seizures with electrographic pattern identical to type 1; and type 3, subclinical, non-reported seizures with electrographic patterns that differed from types 1 and 2. Patients maintained implants for 7–24 months (median 12). All electrographic seizure patterns were confirmed by visual inspection from a qualified epileptologist.

A second dataset was obtained from the Human Epilepsy Project (HEP)[14], an ongoing multi-center study based on a highly screened set of adult patients with focal epilepsy. Patients were enrolled early in their diagnosis, and had comprehensive data recording, including self-reported data quality measures. Data included all 263 patients enrolled between July 2012 to March 2016. This second dataset represents one of the most reliable patient-reported seizure databases available, because of the extensive physician oversight and independent review of data. Diary data for each patient tracked between 1–46 months (median 16).

A third dataset was obtained from SeizureTracker.com[15], a free online, mobile service, representing one of the world's largest patient-managed seizure diary databases. The database includes adults and children with focal or generalized epilepsy. The SeizureTracker database consisted of a data export of all consecutive data entered from the project start in December 2007 through October 2015, comprising 12,946 patients and 1,060,680 seizures. A second export of SeizureTracker (October 2015 through May 2016) was obtained for a validation stage (see section 2.5 below) adding 149,356 new seizures from 1835 patients (846 of whom were new patients).

2.3.2. Preprocessing—Some preprocessing was required to ensure data interpretability. In all three datasets, we required each patient to have at least six months of diary data and at least six seizures recorded to be included for further analysis (see Table 1). A minimum duration was required because simulations were standardized to 6-month blocks.

The SeizureTracker data required additional preprocessing to reduce noise, as there was no physician curating the original database. Repeated patient profiles were removed. Seizures reported to occur after the export date were excluded. Seizures reported with identical start times were removed except for the first one, under the assumption that these represented erroneous repeat entries. Seizures erroneously reported to occur prior to patients' date of birth were excluded. Patients with unreported or impossible ages were excluded due to difficulty in verifying seizure dates.

2.3.3. Prediction Testing—To test the accuracy of prediction of seizure frequency, we simulated a series of 6-month clinical trials of using the three datasets. Diary data was segmented into as many 6-month trial periods as available, comprising 2 months of baseline, a skipped titration month, and a 3-month experimental phase. The number of seizures in

each two-week block within each trial period was represented by $C_{i,j}$ (the i^{th} 2-week seizure count, for the j^{th} patient). Because seizures are very rapid events typically lasting less than two minutes[16], truncation of events at the edges of 2-week segments was considered unnecessary.

Two approaches for seizure frequency predictions were tested on the individual patient level (Figure 1), the RR50 and the Z_V methods. For the Z_V method, the 95%-confidence limits of expected experimental $C_{i,j}$ rates were computed:

$$C_{i,j} \in \left[\widehat{\mu_{\text{ðİŽě}}} - 2\widehat{\sigma_{\text{ðİŽě}}}, \widehat{\mu_{\text{ðİŽě}}} + 2\widehat{\sigma_{\text{ðİŽě}}}\right] \quad \text{(Eq. 7)}$$

The RR50 model has been required by the EMA for traditional epilepsy RCTs, and therefore has been employed for many years. It makes no assumptions about the distribution (unlike the Gaussian assumption of the Z_V model). Rather, it only specifies the lower limit of the 2-week counts during the experimental phase from the j^{th} patient ($C_{i,j}$) as follows:

$$C_{i,j} \in [0.5\,\widehat{\mu_{\text{ðİŽě}}}, \infty) \quad \text{(Eq. 8)}$$

Predictions from Models Z_V and RR50 were tested for each available trial period, of each available patient, within each dataset. The duration of an individual's diary was defined as the time including the first through the last reported seizure entry. This was done to avoid the association with "diary fatigue", wherein a patient stops recording entries despite ongoing events.

Each 2-week seizure count of the experimental period ($C_{i,j}$) was compared to the predicted range (i.e. Eq.7 and 8). The number of $C_{i,j}$ values that were within the predicted range was tallied.

An example: suppose the j^{th} patient had a baseline rate of $\widehat{\mu_{\text{ðİŽě}}} = 10$, and $\widehat{\sigma_{\text{ðİŽě}}} = 3$. During the experimental phase, the third 2-week count $C_{3,j} = 4$. Under the RR50 model (Eq. 8), the predicted limits would be [5, ∞), so $C_{3,j}$ would be noted as a 'failure' of the prediction. Under the Z_V model (Eq. 7), the same $C_{3,j}$ would be within predicted limits of [4,16], thus Z_V would annotate this $C_{3,j}$ response as a 'success'. Similar to this example, these binary outcomes (0=failure, 1=success) were collected for all such 2-week counts from the experimental phase of all available diary data, across all 3 datasets.

2.4. Validation testing of Z_V

To demonstrate the utility of Z_V, we simulated a set of trials from the second data export of SeizureTracker (Methods 2.3.1). SeizureTracker was selected due to the large size of the dataset, and because of the relative consistency of the predictions (Methods 2.3.3) across datasets. The trial parameters were similar to a typical RCT[17]: baseline-8 weeks, titration-4 weeks, and experimental phase-12 weeks. Analogous to a recent trial[18], we required a minimum of three seizures per month and no 21-day seizure-free period during baseline, and at least one seizure after the entire trial duration. Qualified patients (Table 1) were selected

randomly (with replacement) from the new export for these simulated trials. The start time of the data was selected from a uniform distribution of possible start times, allowing for numerous possible virtual patients from each patient diary. A typical[19] drug “strength” of 30% was used, representing the probability that drug treatment would prevent any given seizure[8]. For example, if a patient had 10 seizures recorded during the experimental phase, each seizure would have a 30% chance of being removed, with an expected total of 7 seizures after treatment simulation. Some patients would have more and some would have fewer (because of the probabilistic method of modeling the drug effect). The number of possible drug-exposed virtual patients became extremely large, because even if two identical patient diaries were used, random number generators were applied to the experimental phase representing the effect of the drug, thereby making unique diaries each time. Placebo was modeled as unchanged seizure diaries, because natural variability has been shown to produce realistic responder rates[8]. The number of patients for a single trial was fixed initially at 100, with random allocation to placebo or drug using a 1:1 ratio. A probability of dropout for any given patient was set to 20%. The timing of the patient dropout was simulated independently for each patient by randomly choosing an integer between 1 and 5 for the number of 2-week periods dropped at the end of the experimental phase. For example, if the 12th patient were randomly selected to have experienced dropout, a random number such as 4 might have been chosen, meaning that he would have only completed 2 of the 6 experimental periods (each period is 2 weeks). Trial success was indicated by the ability of the test statistic to distinguish drug from placebo ($p<0.05$).

Traditional trial success[10] based on RR50 was compared to the Z_V method (Equation 5, Figure 2). Comparisons between placebo and drug arms were calculated for Z_V with a mixed effects model and for RR50 with Fisher Exact test. For RR50, the percent-change values were computed using intention-to-treat analysis, using last observation carried forward in the case of early dropout (as is commonly practiced in epilepsy trials). Thus, with each simulated trial, the performance of the two methods was compared.

The fraction of 200 simulated trials that were successful estimated the method power. For instance, if 120 out of 200 trials achieved statistical significance with the RR50 method, then power would be 60%. If that same set of 200 trials was analyzed by the Z_V method resulted in 180 significant trials, then Z_V power would be calculated as 90%. Iterating 200 times obtained stable power estimates. The entire procedure for simulation was repeated again for number of patients set to 200, 300, 400 and 500 as well. In this way, the power of each of the three methods was compared at each of five different trial sizes, using a total of $5 \times 200 = 1000$ virtual trials, and $200 \times (100+200+300+400+500) = 300{,}000$ virtual patients.

Using the same virtual trials, except without simulated “drugs”, RR50 and Z_V were recalculated as an assessment of Type 1 error rates. For example, in a trial with 100 patients, all patients were given placebo. The Z_V and RR50 methods compared the first 50 placebo patients to the second 50 placebo patients. Such a trial would be expected to achieve statistical significance at the $p<0.05$ level about 5% of the time. Thus, these recalculated RCTs verify that the methods do not artificially elevate power at the expense of unacceptably high Type 1 error.

The same calculations done with for RR50 and Z_V were repeated one the same data in order to compare MPC and Z_V. The difference between treatment arms using MPC was tested using the Wilcoxon Signed Rank test[10]. Analysis was conducted in R (v3.2.3) and Matlab (2016b).

3. Results

3.1. Prediction testing

In the first SeizureTracker export of the available 12,946 patient profiles with at least 1 seizure recorded, 12,651 were retained after preprocessing requirements were met (see Methods 2.3.2). Of those, 3016 patients were retained after inclusion criteria (at least 6 seizures recorded in at least 6 months) were applied. Of the 263 patients from HEP, 107 met inclusion criteria. All 15 NeuroVista patients were included.

For each dataset, the two models were run sequentially for as many trial periods as were available in each patient's diary. For example, if a patient had 18 months of diary data, then three trial periods were tested (with baseline and experimental phases included in each trial period). Individual predictions were "correct" if the expected seizure rates were obtained. In Z_V, seizure rates ≤2 predicted standard deviations from the predicted mean were "correct", corresponding to the 95% confidence interval expected. The predictions were 41–80% correct with RR50, and 91–100% for Z_V (Figure 3).

3.2. Validation of Z_V

The second data export of SeizureTracker included 1835 patients, of which 403 patients met inclusion criteria. The computation of statistical power in the example simulation is shown in Figure 4. Z_V outperformed the RR50 method. In particular, Z_V achieved power >90% with N=200, while the RR50 method required N=400. The Type I error rate remained at approximately 5% for all values of N.

Figure 5 shows a comparison of the MPC method with Z_V. Again, the Type I error rate remained approximately 5%. The MPC method required N=200 to achieve 90% power, similar to Z_V.

A more detailed analysis of the simulated trials is included in the Appendix (A1).

4. Discussion

Our study found that the expected range of seizure frequencies can be predicted accurately using variability-corrected Z_V. Using three data sources, the prediction from Z_V outperformed the traditional RR50. With a separate data set from SeizureTracker, we also showed that Z_V improved statistical power over the EMA-preferred RR50 (although not the FDA-preferred MPC), even in the presence of low sample size, patient dropout and/or weak drug effects. With regulatory acceptance, this method could lead to less costly clinical trials, and improved clinical care models.

4.1. Advantages of variability prediction

Z_V can be interpreted as “deviation from expected rate”. Given that current methods fail to address natural variation, this may be a valuable clinical quantity.

RR50 assumes that any reduction in seizure rate 50% below baseline represents binary improvement (Figure 1). In contrast, Z_V computes a patient-specific continuous metric of deviation from prediction. Z_V had higher power than RR50 (Figure 4). The Z_V method can also be calculated retrospectively on existing RCTs.

If regulators would allow Z_V to replace RR50, trial cost could be reduced. MPC and Z_V may be complementary, as they quantify different RCT attributes.

It is currently unknown if the rising placebo responder rates[3] are due to changes in trial design, populations, geography, natural variability, or other factors[4]. Nevertheless, accounting for fluctuations in event rates in a disease that clearly shows considerable fluctuations[8] is likely to be beneficial.

In epilepsy RCTs, dropout rates can reach as high as 27–31%[20,21]. Traditionally, the method of last-observation carried forward (LOCF) is used to account for dropout. However, using LOCF may result in overestimated effect sizes[3]. In contrast, Z_V manages dropout using the weights from the mixed effects modeling. Thus, use of Z_V may permit more accurate effect size estimates, and higher statistical power, even in the presence of patient dropout.

Quantitative decision support could remove the ambiguity of “worsening” and ”improvement” in seizure frequency for clinic patients. A calculation analogous to Z_V could be computed via app, website, or an electronic medical record “plugin”. Integrating such metrics into clinical practice would assist providers in decisions about maintaining or adjusting treatments.

4.2. Limitations of Z_V

The Z_V method has some assumptions. First, it assumes that seizure frequency variability is predictable using a Gaussian distribution. Despite reproducibility (Results 3.1), these findings may not generalize across all forms of epilepsy. Second, it assumes that the baseline measurement is sufficient to estimate true seizure frequency variability. The accuracy of that estimate will be a topic for future studies. Third, the methodology assumes that seizure rates have nonzero variance. Based on clinical experience and quantitative evaluation of catamenial epilepsy[22], we anticipate zero-variance to be practically nonexistent. Next, the Z_V framework assumes that placebo response can be accounted for largely by variability alone[8]. If, conversely, placebo responses are dominated by other factors (such as regression-to-the-mean, or psychological influences)[23], the framework may require modification. Finally, it is assumed that a therapy (e.g. drug) decreases the average seizure rate by a certain percentage relative to the baseline. Although this is also built into traditional RCT analyses, it is worthwhile to recognize that alternatives exist, and they may require further exploration.

4.3. Data considerations

The three datasets come from diverse sources. The NeuroVista data was derived from few patients with medication-resistant focal epilepsy; the other two datasets included focal and generalized epilepsy. NeuroVista data is the "gold standard" in terms of reliability of seizure detection, since intracranial electrodes were used to identify each seizure. HEP required that patients enroll early in their disease course, whereas the other two did not. HEP data was patient-reported, however, it was reviewed by multiple physicians, improving reliability. The SeizureTracker dataset included longitudinal data spanning years and more patients than most datasets worldwide. It was the only dataset that didn't include physician oversight. Self-reported data has additional biases[15]. Perhaps the most challenging is "diary fatigue"—if the patient/caregiver loses interest in the diary, no straightforward correction exists. Of note, all epilepsy phase III RCTs use patient self-report as well. SeizureTracker also uniquely included children and generalized forms of epilepsy. Despite these differences, common results emerged, strengthening the possibility that the findings are generalizable. Specifically, regardless of the degree of reporter reliability (which varied across datasets), the variability prediction appeared to provide accurate boundaries for future seizure frequencies. Of note, all government approved anti-seizure medications and devices are currently approved based on the imperfect method of outpatient self-report. Therefore, accurate predictions from self-reported outpatient diary data is of central importance for analysis of clinical trials.

An important consideration, especially to the HEP dataset, is the possibility of medication changes influencing seizure frequencies and variability. HEP was unique; all patients were recently diagnosed with epilepsy, their medications were likely changed more than some other populations. Although this may have influenced the predictions (Figure 3), adjusting for this would have further improved the estimates. Thus, unadjusted values are presented here as a lower bound for the possibility of prediction.

4.4. Further validation

The impact of different trial parameters must be explored to delineate the boundaries of utility of this technique (e.g. sensitivity to baseline and test durations, drug efficacy, etc.). Additionally, Z_V will need to be studied using existing drug trial data under two conditions: known effective and ineffective drug/dosage combinations. In this way, Z_V can be tested for "true positives" and "true negatives".

4.5. Extending prediction

A number of design decisions in the Z_V method (Figure 2) should be considered flexible. For instance, the number of baseline and experimental intervals are adjustable. Additionally, the predictions (Eq. 3 and 4) are only one possibility. A few others are: predicting variance based on measured mean, utilizing additional covariates, and non-normal distribution of seizure frequencies[12]. Also, future extensions could explicitly account for regression-to-the-mean and psychological effects. Indeed, Z_V can be readily extended to multiple arm trials, or more advanced designs, including sequential parallel comparison design[24], two-way-enriched design[25], adaptive methods[26,27], and platform trials[28] because the key innovation is the normalization of the experimental phase based on a variability prediction.

Z_V may be generalized to many other areas of medicine, whenever event data is used for outcomes. Any disease with a seemingly random, episodic symptom may benefit. For instance, neurological conditions (e.g. headache, stroke/TIA, multiple sclerosis, narcolepsy, REM sleep behavior disorder), psychiatric conditions (e.g. psychosis, depression, manic episodes, panic attacks) or general medical conditions (e.g. syncope, diabetic hypoglycemia, asthma, congestive heart failure, inflammatory bowel disease) could be analyzed in an analogous fashion. A variability-correction method may reduce costs and exposures to ineffective medications.

4.6. Conclusions

This study represents the first formal attempt to quantify and use natural variability in seizure frequency for RCT analysis. The findings suggest that variability-correction could dramatically improve the power and efficiency of RCTs. In turn, this could improve the safety of patients via decreased exposure to non-therapeutic doses of medications[29]. Indeed, smaller, more efficient trials could lead to much lower drug trial costs, thereby accelerating drug discovery.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Primary data was obtained by the Human Epilepsy Project team, the NeuroVista team, and the SeizureTracker.com team. Use of the data was facilitated by the International Seizure Diary Consortium (https://sites.google.com/site/isdchome/).

FUNDING:

This research was funded in part by the National Institutes of Neurological Disorders and Stroke, Intramural Research Division.

References

1. Brodie MJ, Barry SJE, Bamagous Ga, Norrie JD, Kwan P. Patterns of treatment response in newly diagnosed epilepsy. Neurology. 2012; 78:1548–54. [PubMed: 22573629]
2. Kwan P, Brodie MJ. Early identification of refractory epilepsy. N Engl J Med. 2000; 342:314–319. [PubMed: 10660394]
3. Rheims S, Perucca E, Cucherat M, Ryvlin P. Factors determining response to antiepileptic drugs in randomized controlled trials. A systematic review and meta-analysis. Epilepsia. 2011; 52:219–33. [PubMed: 21269281]
4. Goldenholz DM, Goldenholz SR. Response to placebo in clinical epilepsy trials-Old ideas and new insights. Epilepsy Res. 2016; 122:15–25. [PubMed: 26921852]
5. French JA, Krauss GL, Wechsler RT, et al. Perampanel for tonic-clonic seizures in idiopathic generalized epilepsy A randomized trial. Neurology. 2015; 85:950–7. [PubMed: 26296511]
6. Halford JJ, Ben-Menachem E, Kwan P, et al. A randomized, double-blind, placebo-controlled study of the efficacy, safety, and tolerability of adjunctive carisbamate treatment in patients with partial-onset seizures. Epilepsia. 2011; 52:816–25. [PubMed: 21320109]
7. PhRMA. Profile Biopharmaceutical Research Industry. 2015. http://www.phrma.org/sites/default/files/pdf/2015_phrma_profile.pdf

8. Goldenholz DM, Moss R, Scott J, Auh S, Theodore WH. Confusing placebo effect with natural history in epilepsy: A big data approach. Ann Neurol. 2015; 78:329–36. [PubMed: 26150090]
9. European Medical Agencies. Guideline on clinical investigation of medicinal products in the treatment of epileptic disorders. London: 2010. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/201-0/01/WC500070043.pdf
10. OS, NH. Primary efficacy endpoint in clinical trials of antiepileptic drugs: Change or percentage change. Drug Inf J. 2010; 44:343–50.
11. Perucca E. What clinical trial designs have been used to test antiepileptic drugs and do we need to change them? Epileptic Disord. 2012; 14:124–31. [PubMed: 22977898]
12. Ahn JE, Plan EL, Karlsson MO, Miller R. Modeling longitudinal daily seizure frequency data from pregabalin add-on treatment. J Clin Pharmacol. 2012; 52:880–92. [PubMed: 21646441]
13. Cook MJ, O'Brien TJ, Berkovic SF, et al. Prediction of seizure likelihood with a long-term, implanted seizure advisory system in patients with drug-resistant epilepsy: a first-in-man study. Lancet Neurol. 2013; 12:563–71. [PubMed: 23642342]
14. French, J., Kuzniecky, R., Lowenstein, D. The Human Epilepsy Project. http://www.humanepilepsyproject.org/
15. Fisher RS, Blum DE, DiVentura B, et al. Seizure diaries for clinical research and practice: limitations and future prospects. Epilepsy Behav. 2012; 24:304–10. [PubMed: 22652423]
16. Theodore WH, Porter RJ, Albert P, et al. The secondarily generalized tonic-clonic seizure: a videotape analysis. Neurology. 1994; 44:1403–7. [PubMed: 8058138]
17. Guekht, aB, Korczyn, aD, Bondareva, IB., Gusev, EI. Placebo responses in randomized trials of antiepileptic drugs. Epilepsy Behav. 2010; 17:64–9. [PubMed: 19919904]
18. French, Ja, Krauss, GL., Biton, V., et al. Adjunctive perampanel for refractory partial-onset seizures: Randomized phase III study 304. Neurology. 2012; 79:589–96. [PubMed: 22843280]
19. Hemery C, Ryvlin P, Rheims S. Prevention of generalized tonic-clonic seizures in refractory focal epilepsy: A meta-analysis. Epilepsia. 2014; 55:1789–99. [PubMed: 25182978]
20. French JA, Abou-Khalil BW, Leroy RF, et al. Randomized, double-blind, placebo-controlled trial of ezogabine (retigabine) in partial epilepsy. Neurology. 2011; 76:1555–63. [PubMed: 21451152]
21. Elger CE, Brodie MJ, Anhut H, Lee CM, Barrett JA. Pregabalin add-on treatment in patients with partial seizures: a novel evaluation of flexible-dose and fixed-dose treatment in a double-blind, placebo-controlled study. Epilepsia. 2005; 46:1926–36. [PubMed: 16393158]
22. Herzog AG, Fowler KM, Sperling MR, Massaro JM. Progesterone Trial Study Group. Distribution of seizures across the menstrual cycle in women with epilepsy. Epilepsia. 2015; 56:e58–62. [PubMed: 25823700]
23. Goldenholz DM, Goldenholz SR. Response to placebo in clinical epilepsy trials-Old ideas and new insights. Epilepsy Res. 2016; 122:15–25. [PubMed: 26921852]
24. Fava M, Evins aE, Dorer DJ, Schoenfeld Da. The problem of the placebo response in clinical trials for psychiatric disorders: Culprits, possible remedies, and a novel study design approach. Psychother Psychosom. 2003; 72:115–27. [PubMed: 12707478]
25. Ivanova A, Tamura RN. A two-way enriched clinical trial design: combining advantages of placebo lead-in and randomized withdrawal. Stat Methods Med Res. 2011 Epub ahead of print.
26. Connor JT, Elm JJ, Broglio KR. ESETT and ADAPT-IT Investigators. Bayesian adaptive trials offer advantages in comparative effectiveness trials: an example in status epilepticus. J Clin Epidemiol. 2013; 66:S130–7. [PubMed: 23849147]
27. Bhatt DL, Mehta C. Adaptive Designs for Clinical Trials. N Engl J Med. 2016; 375:65–74. [PubMed: 27406349]
28. Rugo HS, Olopade OI, DeMichele A, et al. Adaptive Randomization of Veliparib Carboplatin Treatment in Breast Cancer. N Engl J Med. 2016; 375:23–34. [PubMed: 27406347]
29. Ryvlin P, Cucherat M, Rheims S. Risk of sudden unexpected death in epilepsy in patients given adjunctive antiepileptic treatment for refractory seizures: A meta-analysis of placebo-controlled randomised trials. Lancet Neurol. 2011; 10:961–8. [PubMed: 21937278]

- The expected range of seizure frequencies can be predicted several months in advance.
- Using these predictions, a new trial analysis method Z_V was introduced.
- Compared to 50%-responder rates (RR50), Z_V has higher statistical power to distinguish the placebo arm from the therapeutic arm.
- Use of Z_V in trial analysis may allow for design of epilepsy trials with decreased sample size and cost.

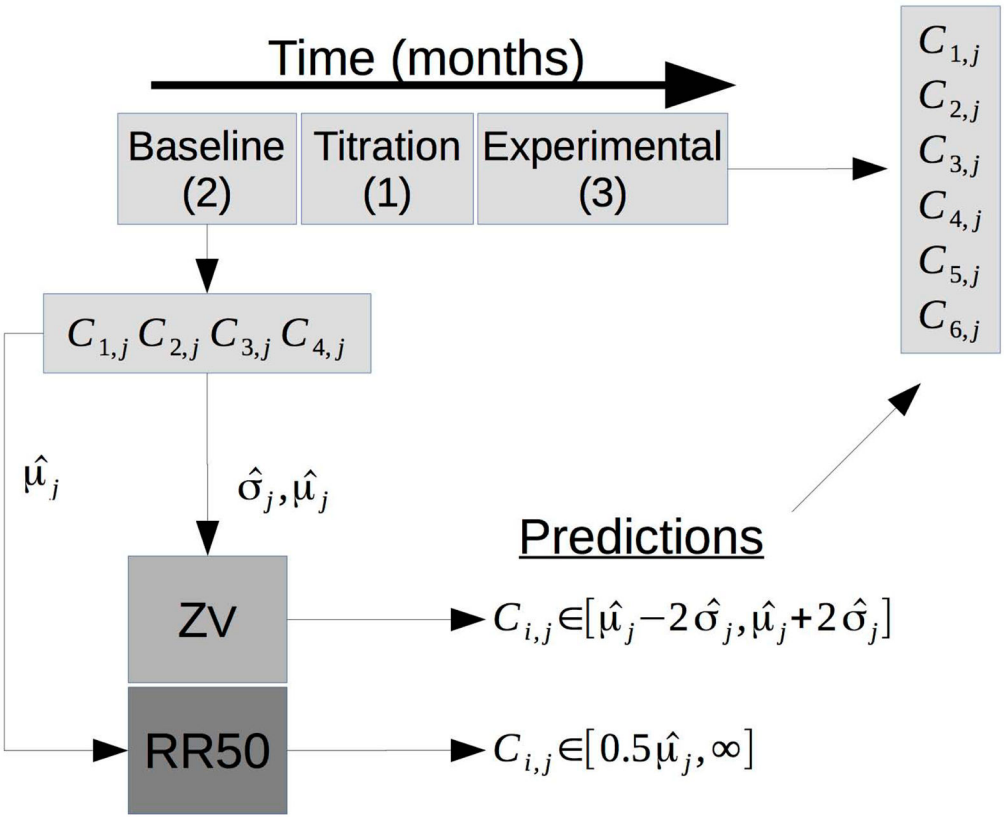

FIGURE 1.

Prediction models. The three phases of a clinical trial are shown: baseline (B), titration (T) and experimental (E). Placebo is given during T and E for those patients assigned to placebo. Drug is titrated up during T, and given at a steady dose during E. The 2 prediction models use the measured $\widehat{\mu_{\text{ðÏŽě}}}$ and $\widehat{\sigma_{\text{ðÏŽě}}}$ (the mean and standard deviation of 2-week seizure counts) from the baseline period, to predict the limits of $C_{i,j}$, the 2-week seizure counts during the experimental phase.

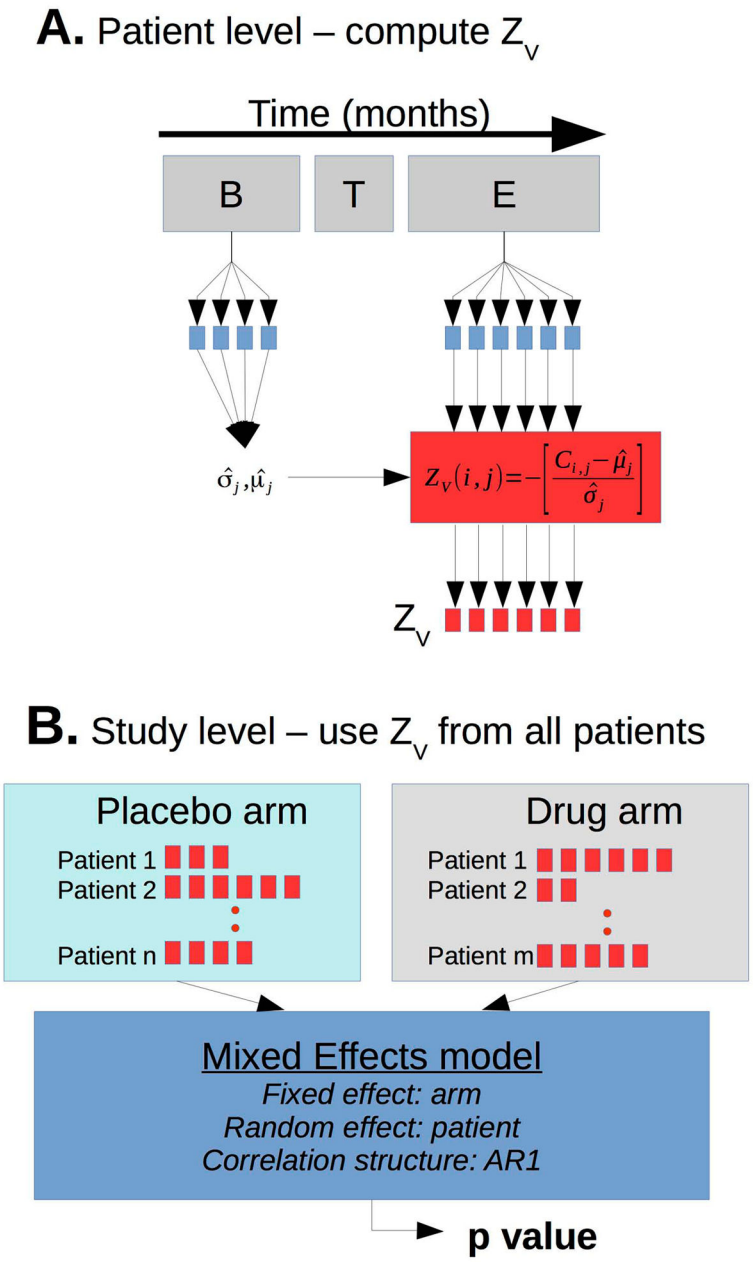

FIGURE 2.

The Z_V analysis method. A. Z_V is calculated for a single patient. A typical trial constructed with baseline (B), titration (T) and experimental (E) phases is shown. The baseline is divided 4 segments in this image, however this number is flexible. Those segments are used to calculate measured $\mu_{Baseline}$ and $\sigma_{Baseline}$, the mean and standard deviation of the seizure counts from each segment. These are then used to compute normalized Z_V from the similarly divided segments of E. Note that in this image 6 segments are represented, though this number is flexible. B. At the study level, all patients contribute a set of Z_V values, however if a patient drops out early then they may contribute less than a full set. Dropout is represented when not all 6 red squares are present for each patient. All completed Z_V values

from each arm are treated with equal weight and are compared with a mixed effects model to obtain a final p value.

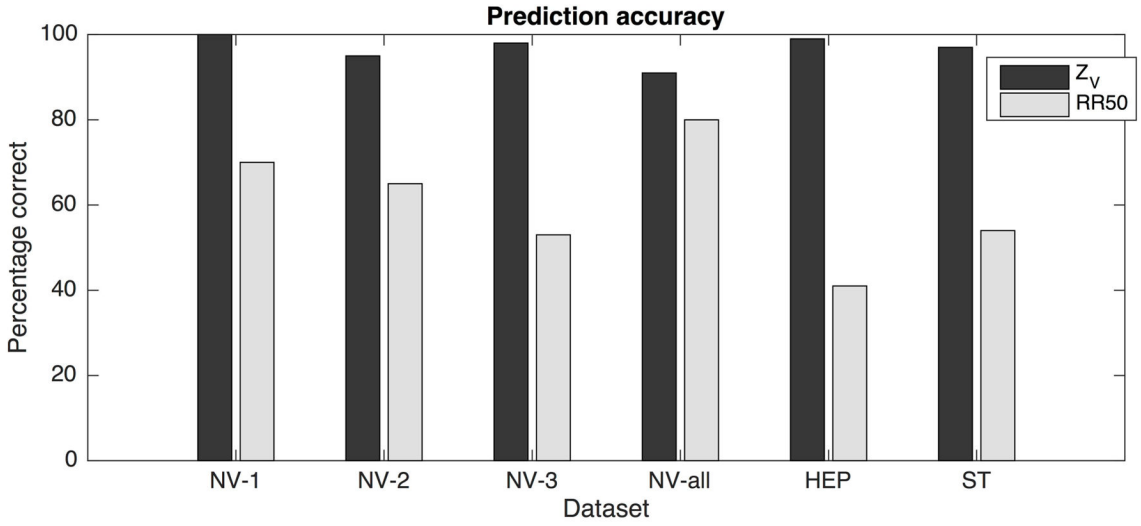

FIGURE 3.

Prediction of variability. For each combination of dataset and model, trial-sized periods of data within each patient were predicted to have certain variability. These predictions had an overall accuracy, which is plotted in each of the graphs. Z_V outperforms RR50 across datasets. NV-1 = NeuroVista clinically reported seizures. NV-2 = Neurovistal clinically equivalent seizures. NV-3 = NeuroVista electrographic seizures. NV-all = NeuroVista subtypes 1 through 3 combined. HEP = Human Epilepsy Project. ST = SeizureTracker data through October 2015.

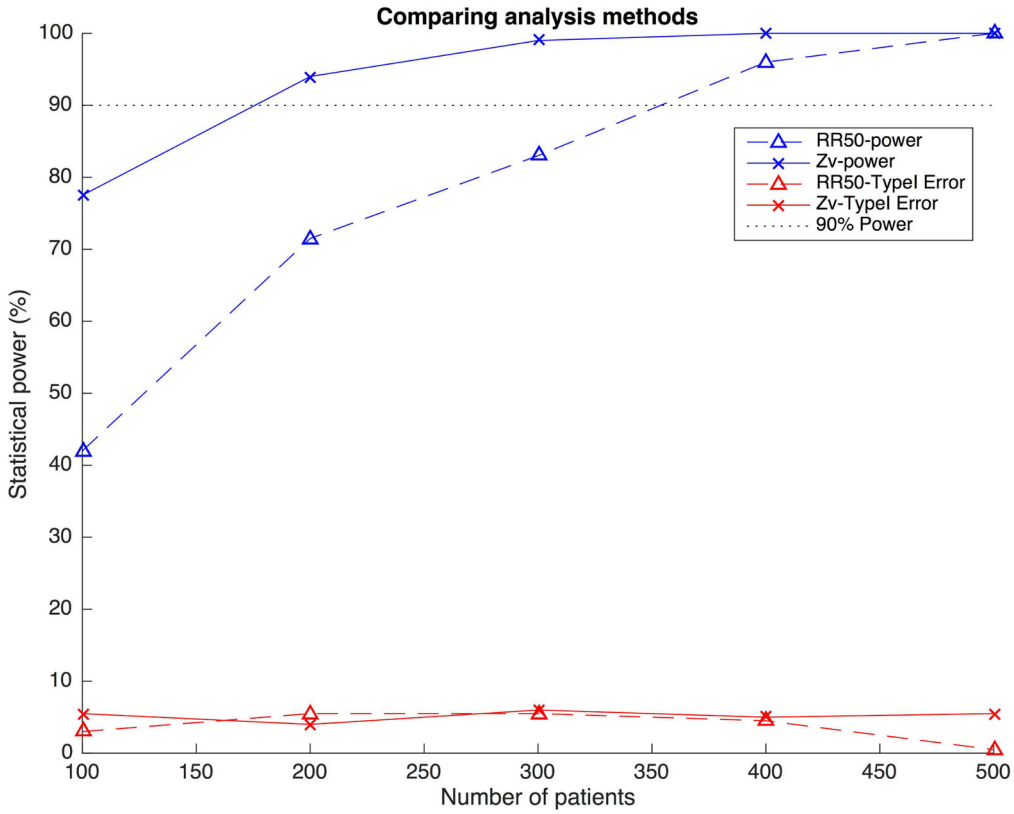

FIGURE 4.

Statistical power and type I error of analysis methods. Using the Seizuretracker dataset a simulation of 200 clinical trials at each trial size (100,200,300,400 and 500) shows that natural variability correction Z_V results in higher statistical power over a traditional RCT method: 50%-responder rate (RR50) to measure the difference between drug and placebo. At each trial size, a set of 200 clinical trials were simulated, thus the entire figure summarizes 1000 simulated trials. Shown here are both the statistical power (upper traces) and the type I error rates (lower traces) of the two methods. Type I errors were calculated by not introducing drug lowering to the same sets of trials as used for the power calculations, thus comparing placebo to placebo.

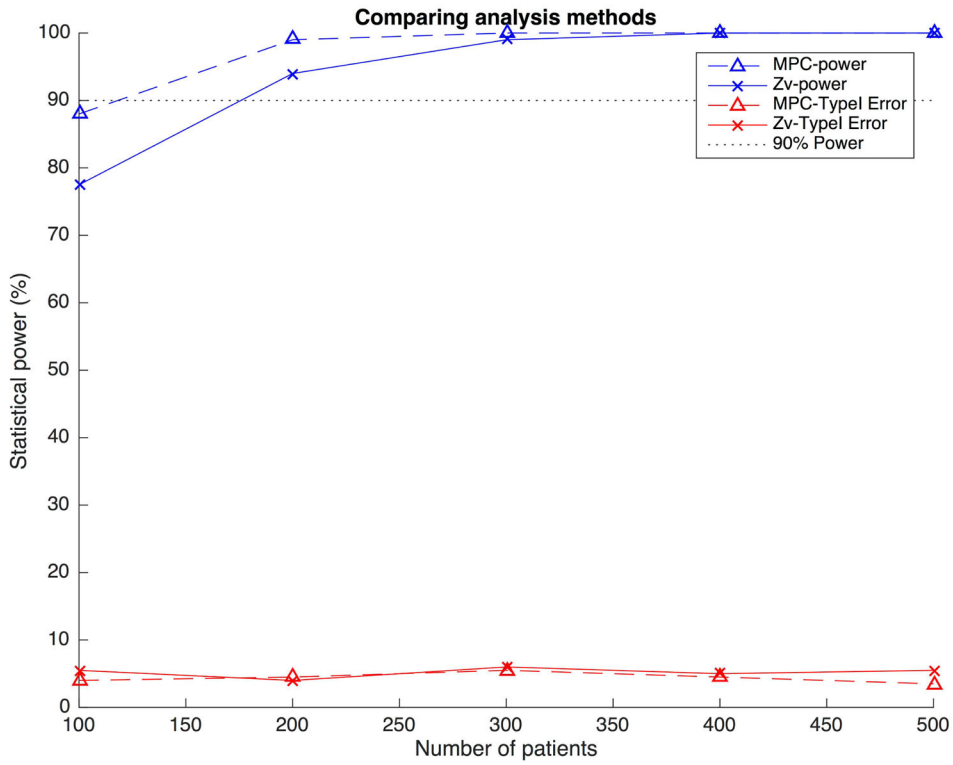

FIGURE 5.

Statistical power and type I error of analysis methods. The same Seizuretracker simulations as Figure 4 are analyzed using natural variability correction Z_V and median percentage change (MPC) to measure the difference between drug and placebo. At each trial size, a set of 200 clinical trials were simulated, thus the entire figure summarizes 1000 simulated trials. Shown here are both the statistical power (upper traces) and the type I error rates (lower traces) of the two methods. Type I errors were calculated by not introducing drug lowering to the same sets of trials as used for the power calculations, thus comparing placebo to placebo.

TABLE 1

Data Sets. Shown here are the 3 datasets used for testing Model V and Model F (NeuroVista, HEP, and SeizureTracker), as well as the additional dataset (denoted with *) from SeizureTracker used in the validation simulation with Z_V. In the additional SeizureTracker dataset, further exclusion criteria were used to obtain the final set of patient data (see Section 2.5).

	N	N (after exclusions)	Study duration in months	Diary durations after exclusion critiera	Ages	Epilepsy
NeuroVista	15	15	7–24 (12)	7–24 (12)	adults	Focal
Human Epilepsy Project	263	107	1–46 (16)	8–42 (22)	adults	Focal
SeizureTracker.com	12946	3016	0–596 (1)	6–596 (20)	adults + children	Focal and generalized
SeizureTracker.com(*)	1835	403	0–8 (3)	6–8 (8)	adults + children	Focal and generalized