HHS Public Access

Author manuscript

Clin Trials. Author manuscript; available in PMC 2020 June 01.

Published in final edited form as:

Clin Trials. 2019 June ; 16(3): 273–282. doi:10.1177/1740774519833679.

Improving pragmatic clinical trial design by using real-world data

Susan M Shortreed[1,2], **Carolyn M Rutter**[3], **Andrea J Cook**[1,2], and **Greg E Simon**[4]

[1]Biostatistics Unit, Kaiser Permanente Washington Health Research Institute, USA

[2]Department of Biostatistics, University of Washington, USA

[3]RAND Corporation, USA

[4]Kaiser Permanente Washington Health Research Institute, USA

Abstract

Background: Pragmatic clinical trials often use automated data sources such as electronic health records, claims, or registries to identify eligible individuals and collect outcome information. A specific advantage that this automated data collection often yields is having data on potential participants when design decisions are being made. We outline how this data can be used to inform trial design.

Methods: Our work is motivated by a pragmatic clinical trial evaluating the impact of suicide-prevention outreach interventions on fatal and non-fatal suicide attempts in the 18 months after randomization. We illustrate our recommended approaches for designing pragmatic clinical trials using historical data from the health systems participating in this study. Specifically, we illustrate how electronic health record data can be used to inform the selection of trial eligibility requirements, to estimate the distribution of participant characteristics over the course of the trial, and to conduct power and sample size calculations.

Results: Data from 122,873 people with patient health questionnaire (PHQ) recorded in their electronic health records between July 1, 2010 and March 31, 2012, were used to show that the suicide attempt rate in the 18 months following completion of the questionnaire varies by response to item nine of the PHQ. We estimated that the proportion of individuals with a prior recorded elevated PHQ (i.e. history of suicidal ideation) would decrease from approximately 50% at the beginning of a trial to about 5% 50 weeks later. Using electronic health record data, we conducted simulations to estimate the power to detect a 25% reduction in suicide attempts. Simulation-based power calculations estimated that randomizing 8,000 participants per randomization arm would allow 90% power to detect a 25% reduction in the suicide attempt rate in the intervention arm compared to usual care at an alpha rate of 0.05.

Conclusions: Historical data can be used to inform the design of pragmatic clinical trials; a strength of trials that use automated data collection for randomizing participants and assessing outcomes. In particular, realistic sample size calculations can be conducted using real-world data from the health systems in which the trial will be conducted. Data-informed trial design should yield more realistic estimates of statistical power and maximize efficiency of trial recruitment.

Corresponding Author: Susan M Shortreed, 1730 Minor Ave, Ste 1600, Seattle, WA 98101, Phone: 206-287-2088, susan.m.shortreed@kp.org.

Keywords

Electronic medical records; study design; power calculations; sample size calculations; randomized trial design; mental health; suicide prevention; pragmatic clinical trials

Introduction

Pragmatic randomized trials have been proposed to address policy and healthcare questions in real-world settings. Leveraging technology advancements, pragmatic trials often use automated data sources such as often electronic health records, claims, or registries, to identify eligible individuals (or individuals within clusters for cluster-randomized trials) and assess outcomes without ever contacting participants.[1–4] This makes enrolling a large, broad population feasible. Expenses associated with recruitment and data collection restrict how many individuals can feasibly be enrolled in a randomized clinical trial (RCT). Small sample sizes limit power to detect treatment effects; thus, investigators often restrict study eligibility criteria to target populations expected to have large treatment benefits. This approach does not allow assessment of treatment effects across diverse populations nor treatment effectiveness if delivered in a real healthcare setting. Additionally, high participant burden associated with eligibility screening and data collection is one reason potential participants may refuse participation and it may lead to missing information among participants. Addressing patient burden can increase generalizability of study results and decrease bias associated with missing data. Reducing patient burden (e.g. to assess eligibility or outcomes) is a key component that can set pragmatic trials apart from traditional RCTs designed to assess efficacy.[1] Pragmatic RCTs are becoming more viable as the number of research networks increases (e.g. Federal Drug Administration's Sentinel Initiative,[5] Health Care Systems Research Network,[6–8] National Institutes of Health's Mental Health Research Network,[9] Patient Centered Outcomes Research Network,[10] NorthWest EHealth,[11] NHS National Institute for Health Research).[12]

Many advantages of using clinical data to facilitate pragmatic trial implementation, including automated assessment and standardized recording of clinical assessments, have been described elsewhere along with other design considerations.[13] Here we focus on a specific advantage of pragmatic RCTs conducted within health systems: the availability of clinical data on potential participants prior to trial design. We propose approaches that use real-world data to inform trial design; specifically, for determining eligibility criteria, assessing how patient characteristics may change over the course of trial enrollment, and estimating necessary sample size. Our work is motivated by a pragmatic RCT designed to evaluate the impact of suicide-prevention outreach interventions on suicide attempts funded by the NIH Common Fund Health Care Systems Research Collaboratory and National Institute for Mental Health. This suicide prevention trial is used as an illustrative example; these approaches are applicable to all trials embedded in healthcare systems. We provide relevant background on this suicide prevention trial, before describing and illustrating our approaches to trial design.

Methods

Suicide prevention outreach trial – background and available data sources

Suicide is the tenth leading cause of death in the United States. Effective suicide prevention interventions exist but are often resource intensive and effectiveness of implementing these interventions at a health system level is unknown. The suicide prevention outreach trial (SPOT) was designed to compare the effectiveness of two low-intensity outreach interventions to usual care on suicide attempt rate, fatal and non-fatal, in the 18 months post randomization. One intervention focused on maintaining engagement in mental health care; the second focused on improving self-management of suicidal ideation and behavior using Dialectical Behavior Therapy skills.[14] Both interventions were delivered primarily online using the health systems' electronic patient portal – allowing large-scale delivery at low cost. While SPOT is a three-armed RCT, we consider two-armed trials – one active intervention compared to usual care. The extension to three or more group comparisons is straightforward.

Clinical data available for designing SPOT included the nine-item Patient Health Questionnaire (PHQ), a measure of depressive symptom severity. PHQ item nine asks how often in the last two weeks an individual has "Thought that you would be better off dead, or of hurting yourself in some way?" Response options are: not at all, some of the days, most of the days, and nearly every day and are coded: 0 (not at all), 1, 2, 3 (nearly every day). It has been shown PHQ item nine is associated with suicide attempt risk in the following 2 years.[15,16] Individuals who respond with 0 to item nine are at the lowest risk. Individuals who respond with 3 are at the highest risk; the number of patients who respond with 3 is relatively small.

In addition to longitudinal PHQ data, information is available on suicide attempts from multiple sources. For care delivered within the health system, suicide attempt information is available in the electronic health record. For care delivered at an external health system this information is captured through claims because all participating sites are integrated health systems, providing both care and insurance coverage. Fatal suicide attempts are ascertained using state mortality records, which are available for all individuals residing in the state at the time of their death. When an individual disenrolls from their health plan, we cannot be certain a suicide attempt would be recorded in our data base, therefore they are censored. Each SPOT health system updates member enrollment files regularly, making censoring individuals straightforward to implement.

The SPOT study identified eligible individuals using PHQ item nine responses and used a modified Zelen randomization design.[17] In this design, all patients who meet eligibility criteria are randomized to an offer of an intervention or continued usual care. This design requires a waiver of consent for randomization and a modified consent process, in which individuals are approached by study interventionists and provide consent for participating in the intervention. Individuals randomized to usual care are not contacted, and individuals who do not consent to the intervention do not receive it and are no longer contacted (they remain assigned to their randomized intervention arm for analyses). This randomization procedure, i.e. through a waiver of consent, should only be conducted when outcome data is

collected through automated systems, thus available for all participants, including those randomized to usual care and who decline intervention participation. If it was necessary to contact participants for outcomes assessment, this randomization scheme would lead to biased outcome ascertainment. The primary comparison in SPOT is an intent-to-treat comparison of all individuals randomized to usual care to all individuals randomized to an offer of an intervention, regardless of how much of the intervention (if any) they were exposed. The SPOT study design and all procedures were reviewed and approved by Institutional Review Boards of all health systems involved.

Leveraging real-word data to design pragmatic clinical trials

Electronic health records data, and other automated data sources, provide an opportunity to use historical information to emulate hypothetical clinical trials; we illustrate this approach here. We used data from HealthPartners, Kaiser Permanente Colorado, and Kaiser Permanente Washington (formerly Group Health Cooperative) members who had a PHQ response recorded in their electronic health record between January 1, 2007 and September 30, 2013. For inclusion in hypothetical trials, we considered PHQ item nine responses recorded between January 1, 2010, when routine PHQ use was in effect at all health systems, and March 31, 2012 to ensure 18 months of historical follow-up.

To emulate a hypothetical trial, we followed each person in our database forward from January 1, 2010 until they met eligibility criteria, then assessed outcomes in the 18 months following. An individual could only enter a hypothetical trial once. If they met eligibility criteria multiple times after January 1, 2010, we assessed outcomes in the 18 months following the earliest date. If an individual never met eligibility criteria they were never “enrolled”.

Identifying trial populations and estimating event rates.—Determining study eligibility criteria can be a difficult task. Investigators must identify a population who could benefit from the intervention, but the eligible pool needs to be large enough to conduct the study. We estimated how many individuals might be eligible for hypothetical trials with different eligibility criteria, and estimated 18-month suicide attempt risk in those individuals. We considered four eligibility criteria based on PHQ item nine response, patients who:

1. Reported thoughts of self-harm nearly every day (response of 3);
2. Reported thoughts of self-harm most of the days or nearly every day (response of 2 or 3);
3. Endorsed suicidal ideation by responding positively (response of 1, 2 or 3);
4. Completed item nine during a clinical encounter regardless of their response (response of 0, 1, 2, or 3).

It is important to note patients may be assessed with the PHQ at several visits. An individual may have a PHQ item nine response of a 1 at one visit, and at a later visit a 3 could be recorded. Using historical information, we can evaluate not only who might be enrolled, but also *when they would be enrolled*. For example, the patient described above would be enrolled in all trials, but at different times depending on the criteria used. This timing may

be important, for statistical power and health care delivery; suicide attempt risk may not be the same after these different times.[15, 16, 18] Determining when, within the symptom course, an individual should be enrolled is important for trials of interventions for non-acute conditions in which symptom severity fluctuates overtime.

We estimated the number of patients who met each of the four eligibility criteria, the suicide attempt rate in the 18 months following the first eligible visit using Kaplan-Meier curves,[19] and cumulative incidence rates. To assess potential site heterogeneity, we produced estimates for each site separately and combined across sites. The SPOT study randomized patients with a recorded PHQ item nine response of 2 or 3; thus, we use this eligibility criterion (i.e. criterion 2) throughout the remainder of the paper.

Examining how populations change over time.—We estimated how enrollment of specific subgroups might change over the course of trial enrollment. This is especially important if heterogeneity of treatment effects or differential event rates within subgroups are expected. For example, it is useful to estimate the proportion of individuals with previously recorded versus no prior recorded suicidal ideation. At the beginning of a trial, a large proportion of those randomized will have been living with depression and/or suicidal ideation for several months or years. As the trial continues, weekly enrollment rates will decline and the "prevalent" pool (i.e. individuals with prior recorded suicidal ideation) will become exhausted. In later months, eligible individuals are mostly comprised of patients without prior recorded suicidal ideation. This enrollment pattern will be observed regardless of how many sites are involved, but for a fixed sample size, more participating sites will lead to a larger proportion of prevalent patients in the final trial population. All RCTs will encounter declining enrollment rates and changing patient characteristics over the course of enrollment. In trials of interventions for non-acute conditions with evolving symptom severity, it is particularly important to evaluate how the proportion of prevalent individuals enrolled changes over time. We estimated changes in PHQ item nine response at randomization, PHQ item nine response prior to randomization, age, and race/ethnicity for each site separately and all sites combined. These characteristics are known to be related to suicide attempt risk.[15]

Examining censorship patterns.—A particularly novel consideration associated with pragmatic RCTs that assess outcome information using automated data sources is what it means for a participant to be "lost-to-follow-up." Recall, in the SPOT study, outcomes are observed (i.e. suicide attempts) for individuals who are enrolled in participating health systems. Once an individual disenrolls, they are censored as we no longer observe their outcome. We examined censoring patterns (i.e. health system disenrollment patterns) in the 18 months following hypothetical randomization and if PHQ item nine response at randomization, prior PHQ item nine response, and age were associated with different censoring patterns using stratified Kaplan-Meier curves, for each site separately and all sites together.

Determining power and sample size.—We outline two approaches to incorporating real-world data into power and sample size calculations. The first approach uses available data to estimate quantities required for standard sample size or power calculations. We used

PASS version 14[20] to calculate required sample size for 80% and 90% power to detect a 25% decrease in the 18-month suicide attempt rate using estimated event and censorship rates.

The second approach uses a simulation study based on real-world data to estimate power under different scenarios. We used this approach to account for a changing eligible patient population (i.e. individuals with versus without prior recorded suicidal ideation) over time. We varied randomization arm size (intervention and control groups equal to 5000, 6000, 7000, 8000, 9000, and 10000) and number of participating sites (1 to 4 sites). We simulated 10,000 trials for each scenario using R version 3.4.1[21], by repeating steps 2 through 7.

1. We estimated the baseline hazards of, and covariate effects on, time to first suicide attempt using a Cox proportional hazards model[22] including: PHQ item nine response at randomization (2 or 3), site (HealthPartners, Kaiser Permanente Colorado, Kaiser Permanente Washington), age (18 to 29 years, 30 to 64 years, 65 years and older), sex (male, female), race/ethnicity (White non-Hispanic, non-White or Hispanic), and provider type (mental health specialty vs. primary care or other specialty).
2. We used bootstrap resampling to generate simulated trials, randomly selecting patients with replacement from the full hypothetical trial dataset. Assuming each site accrued 50 patients per week, we calculated how many weeks would be needed to enroll the required sample. We assumed the proportion of individuals with previously recorded suicidal ideation (PHQ item 9 response 2 or 3) was 60% in week one and decreased by one percent each week. At week 55 and thereafter, 5% of trial participants were assumed to have previously reported suicidal ideation.
3. Individuals were randomly allocated (50/50) into intervention and control groups.
4. We used baseline hazard and covariate coefficients estimated in Step 1 to calculate each person's daily cumulative hazard based on covariate values and randomly assigned treatment group. The hazard ratio for the intervention group was 0.75 for all simulations. We used an inverse probability lookup approach based on the cumulative hazard for each patient to simulate individual event times.
5. Censoring times were generated using a "hot-deck" imputation approach.[23] For each patient, we identified all patients who had the same covariate values and did not have an event. We randomly sampled a censoring time from among this group for that individual for each simulated trial. All events simulated to occurred after an individual's censoring time were censored in the analysis.
6. For each simulated trial, we estimated two models: an unadjusted model and an adjusted model that controlled for all covariates used in Step 1.
7. Power was estimated by the proportion of simulated trials for which an intervention effect was statistically significant at the 0.05 level.

Results

Between January 1, 2010 and March 31, 2012, there were 296,127 PHQ item nine responses recorded in the electronic health records of 122,873 patients with an average of 3.2 PHQs recorded per person (standard deviation=4.3; minimum=1; median=2.0, maximum=109). In the three years prior (i.e. between January 1, 2007 and December 31, 2009) 233,101 PHQs were recorded on 90,426 patients, which were used to identify individuals with a history of suicidal ideation. Across sites, the number of PHQ responses, members, and average number of PHQ responses per member varied. Health Partners: 30,279 responses from 95,800 members (average of 3.3 per person), Kaiser Permanente Colorado: 28,420 responses from 53,499 members (average 1.95), and Kaiser Permanente Washington: 64,174 responses from 146,828 members (average 2.47). We present results combined across all sites, as enrollment patterns and suicide attempt and censorship rates were consistent across sites.

A trial requiring a PHQ item nine response of 3 between January 1, 2010 and March 31, 2012 would have randomized 5,423 participants. A trial that required response of 2 or 3 (i.e. hypothetical SPOT study; criteria 2) would have randomized 6,970 additional patients, for total of 12,393. Requiring any positive score, would have randomized an additional 18,154 for a total of 30,437 participants. A trial that included anyone who completed a PHQ would have randomized an additional 92,326 for a total of 122,873 participants.

Figure 1 presents Kaplan-Meier curves for time to first suicide attempt in the 18 months following randomization for hypothetical trials with the four eligibility criteria, which represent the survival curves we would expect in each trial's usual care arm. The 18-month cumulative incidence was 0.033 (trial requiring response of 3), 0.026, 0.018, and 0.0078 (PHQ completion). Randomizing individuals who responded with a 3 would result in the highest event rate, while enrolling all individuals who completed the PHQ, regardless of their score, would enroll more participants, but the event rate would be much lower.

Figure 2 presents Kaplan-Meier curves for individuals who would have met multiple trial eligibility criteria; these curves show timing matters. Suicidal ideation, as measured by the PHQ, is a time-varying characteristic; these plots indicate the suicide attempt rate may vary after different item nine responses. The cumulative incidence following a lower response on item nine appears smaller than following a higher response for the same individuals, but it is not as low as the plots show in Figure 1.

Figure 3 shows the changing characteristics of people randomized in the hypothetical SPOT. Figure 3b shows the proportion of individuals who have a prior recorded suicidal ideation is approximately 50% in the first weeks of the trial and decreases over time; about fifty weeks later the proportion is 5%.

Of the 12,178 patients eligible for the hypothetical SPOT study, 36.9% disenrolled from the health system before 18 months of follow-up would have expired. Figure 4 shows Kaplan-Meier curves describing censorship patterns overall and within subgroups (PHQ item nine response at randomization, prior PHQ item nine response, and age).

In the 18 months following a qualifying PHQ (i.e. the earliest recorded 2 or 3 after January 1, 2010 and before March 21, 2012) 325 suicide attempts were observed. The average time from first eligible PHQ until health plan disenrollment was 14.8 months (SD=5.5 months); approximately 2% disenrolled each month. The estimated 18-month cumulative incidence was 0.033. Assuming a two-sided log-rank test, with a type-1 error rate of 0.05, a trial with 4,798 patients per arm would have 80% power to detect a 25% reduction in the 18-month post-randomization suicide attempt rate in intervention compared to usual care. A sample of size 6,411 per arm would have 90% power.

Results of simulation-based power calculations, varying both the number of participants and number of participating sites, are shown in Table 1. As the number of participants randomized and sites involved varies, the proportion who have prior recorded suicidal ideation also varies. Keeping the sample size constant at 5,000 per arm, a trial conducted at one site would have 12.7% of participants with prior recorded suicidal ideation, while a trial conducted at three sites would have 28.1%. Holding the number of sites constant at three, the proportion of individuals with prior recorded suicidal ideation decreases from 28.1 % to 16.5 % as the sample size increases from 5,000 per arm to 10,000. This simulation-based approach estimates 6,000 per arm should allow at least 80% power to detect a hazard ratio of 0.75 and 8,000 per arm should allow at least 90% power.

Discussion

We have demonstrated how clinical and other automated data sources can be used to inform pragmatic trial design; specifically, to predict the composition of trial participants and to conduct power and samples size calculations. The SPOT study, a large, pragmatic RCT evaluating the impact of outreach interventions on suicide attempts was used as an illustrative example. We presented two approaches for data-informed power calculations. One used event and censorship rates estimated from real-world data and standard software for sample size calculations. The second used simulations to incorporate complexities like evolving patient populations over time. For example, participants randomized early on are more likely to have been struggling with illness for longer (i.e. prevalent cases), while later more participants have newly emerging symptoms (i.e. incident cases). This time-varying pattern in patient enrollment is common in all trials of interventions for non-acute conditions. The proportion of prevalent versus incident cases in the full trial population depends on the number randomized and the number of participating sites. In our example, while the proportion of individuals who had prior recorded suicidal ideation varied as the sample size and number of sites varied, the event rate was constant over these variations. The suicide attempt rate of individuals with and without prior recorded suicidal ideation was similar, thus this variation did not impact sample size estimates. In settings where the event rate is different, properly accounting for this variation will lead to differing sample size estimates.

Addressing the impact of site selection on trial design is a complex and important topic for consideration when conducting multi-sites studies. Site heterogeneity in all design decisions can and should be assessed. In the SPOT study, suicide attempt rates, enrollment patterns, and censorship patterns were consistent across all sites, but this is not often the case.

Assessing site heterogeneity and accounting for its potential impact on study design can have large implications. For example, if there were meaningful differences in outcome rates, it may be necessary to conduct site-stratified analyses and sample size calculations should reflect this stratified analysis.

Large, pragmatic clinical trials often randomized all eligible participants; thus, we expect results will generalize to future patients of participating health systems and other similar health systems. It is important to evaluate the generalizability of study results of trials that do not randomize all individuals or trials for which patients may opt out. Clinical data can be used to evaluate differences between trial participants and non-participants (e.g. those not randomized or who refused to participate).[24] Previous authors have noted obtaining data to evaluate generalizability can be difficult;[25] electronic health record data can be a great resource for this. We have previously demonstrated using clinical data to examine non-response bias in a survey.[26]

Here we compared four eligibility criteria based solely on recorded PHQ item nine response; often more complex criteria are considered, including from multiple data sources, which may not be equally accessible. For example, an additional eligibility criterion in SPOT was recent use of the health system's secure message portal. For trial design, information on secure message use was not able to be linked to recorded PHQ response. Summary information about secure message use across all health system members at each site was used to plan trial implementation. Future work should consider how to best incorporate both individual-level and aggregate data in trial design. Researchers conducting trials embedded in health systems can access historical clinical data as part of prep-to-research activities. This is not always the case; privacy concerns can limit data availability. To share individual-level data researchers must de-identify data to comply with Health Insurance Portability and Accountability Act regulations.[27] If the data contain any direct identifiers or if re-identification risk is deemed unacceptable, data use agreements can be put in place.

Increased use of electronic health record systems in practice has increased the availability of clinical data for research purposes. We have demonstrated the use of this data to improve pragmatic RCT design. This work relies on the assumption that history will be a good predictor of the future. If temporal changes occur within the health systems or the populations they serve, then this assumption may not be valid. An additional limitation of these databases, is they are "living"; codes can be corrected or removed, altering the data from pull to pull.[28] Data from external sources can also be delayed. For example, billing codes can be delayed by late billing and adjudications.

There are many decisions that go into trial design. The more informed by real-world data these decisions are, the more realistic we will be in planning and determining the limitations of trials to detect desired effects. In the SPOT study, use of clinical and administrative data to accurately model recruitment and outcome event rates led to a significant increase in estimates of necessary sample size and time necessary to recruit an adequate sample. While this news may be unwelcome, more informed decisions in the design phase avoid subsequent disappointments regarding actual recruitment rate and precision or statistical power.

Funding

National Institute of Mental Health UH3 MH007755

NIH Common Fund Health Care Systems Research Collaboratory U54 AT007748

Declaration of conflicting interests

Dr. Shortreed has worked on grants awarded to Kaiser Permanente Washington Health Research Institute (KPWHRI) by Bristol Meyers Squibb and by Pfizer. Drs Shortreed and Cook are also co-Investigators on grants awarded to KPWHRI from Syneos Health, who is representing a consortium of pharmaceutical companies carrying out FDA-mandated studies regarding the safety of extended-release opioids.

References

1. Thorpe KE, Zwarenstein M, Oxman AD, et al. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. J Clin Epidemiol 2009; 62: 464–475. [PubMed: 19348971]
2. van Staa T, Goldacre B, Gulliford M, et al. Pragmatic randomised trials using routine electronic health records: putting them to the test. BMJ 2012; 344: e55. [PubMed: 22315246]
3. Chalkidou K, Tunis S, Whicher D, et al. The role for pragmatic randomized controlled trials (pRCTs) in comparative effectiveness research. Clin Trials 2012; 9: 436–446. [PubMed: 22752634]
4. Califf RM. Pragmatic clinical trials: emerging challenges and new roles for statisticians. Clin Trials 2016; 13: 471–477. [PubMed: 27378791]
5. Behrman RE, Benner JS, Brown JS, et al. Developing the Sentinel System--a national resource for evidence development. N Engl J Med 2011; 364: 498–499. [PubMed: 21226658]
6. Platt R, Davis R, Finkelstein J, et al. Multicenter epidemiologic and health services research on therapeutics in the HMO Research Network Center for Education and Research on Therapeutics. Pharmacoepidemiol Drug Saf 2001; 10: 373–377. [PubMed: 11802579]
7. Steiner JF, Paolino AR, Thompson EE, et al. Sustaining research networks: the twenty-year experience of the HMO Research Network. EGEMS (Wash DC) 2014; 2: 1067. [PubMed: 25848605]
8. Vogt TM, Elston-Lafata J, Tolsma D, et al. The role of research in integrated healthcare systems: the HMO Research Network. Am J Manag Care 2004; 10: 643–648. [PubMed: 15515997]
9. Rossom RC, Simon GE, Beck A, et al. Facilitating action for suicide prevention by learning health care systems. Psychiatr Serv 2016; 67: 830–832. [PubMed: 27032667]
10. Fleurence RL, Curtis LH, Califf RM, et al. Launching PCORnet, a national patient-centered clinical research network. J Am Med Inform Assoc 2014; 21: 578–582. [PubMed: 24821743]
11. NorthWest EHealth. NorthWest EHealth homepage, https://nweh.co.uk/ (2018, accessed 21 December 2018).
12. National Institute for Health Research. National Institute for Health Research homepage, https://www.nihr.ac.uk/ (2018, accessed 21 December 2018).
13. National Institutes of Health. Rethinking clinical trials homepage. NIH Collaboratory Living Textbook, http://rethinkingclinicaltrials.org/ (2018, accessed 21 December 2018).
14. Simon GE, Beck A, Rossom R, et al. Population-based outreach versus care as usual to prevent suicide attempt: study protocol for a randomized controlled trial. Trials 2016; 17: 452. [PubMed: 27634417]
15. Simon GE, Rutter CM, Peterson D, et al. Does response on the PHQ-9 Depression Questionnaire predict subsequent suicide attempt or suicide death? Psychiatr Serv 2013; 64: 1195–1202. [PubMed: 24036589]
16. Simon GE, Coleman KJ, Rossom RC, et al. Risk of suicide attempt and suicide death following completion of the patient health questionnaire depression module in community practice. J Clin Psychiatry 2016; 77: 221–227. [PubMed: 26930521]
17. Zelen M A new design for randomized clinical trials. N Engl J Med 1979; 300: 1242–1245. [PubMed: 431682]

18. Simon GE, Shortreed SM, Johnson E, et al. Between-visit changes in suicidal ideation and risk of subsequent suicide attempt. Depress Anxiety 2017; 34: 794–800. [PubMed: 28440902]
19. Kaplan EL and Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958; 53: 457–481.
20. NCSS L. PASS 14 Power Analysis and Sample Size Software Kaysville, Utah, USA2015.
21. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2017.
22. Therneau T A package for survival analysis in S. 2.38 ed. 2015.
23. Andridge RR and Little RJA. A review of hot deck imputation for survey non-response. Int Stat Rev 2010; 78: 40–64. [PubMed: 21743766]
24. Westreich D, Edwards JK, Lesko CR, et al. Transportability of trial results using inverse odds of sampling weights. Am J Epidemiol 2017; 186: 1010–1014. [PubMed: 28535275]
25. Stuart EA and Rhodes A. Generalizing treatment effect estimates from sample to population: A case study in the difficulties of finding sufficient data. Eval Rev 2017; 41: 357–388. [PubMed: 27491758]
26. Shortreed SM, Von Korff M, Thielke S, et al. Electronic health records to evaluate and account for non-response bias: a survey of patients using chronic opioid therapy. Obs Stud 2016; 2: 24–38. [PubMed: 28042621]
27. U.S. Department of Health & Human Services. Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (HIPAA) privacy rule, https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf (2012, accessed 21 December 2018).
28. Nelson JC, Cook AJ, Yu O, et al. Challenges in the design and analysis of sequentially monitored postmarket safety surveillance evaluations using electronic observational health care data. Pharmacoepidemiol Drug Saf 2012; 21 Suppl 1: 62–71. [PubMed: 22262594]

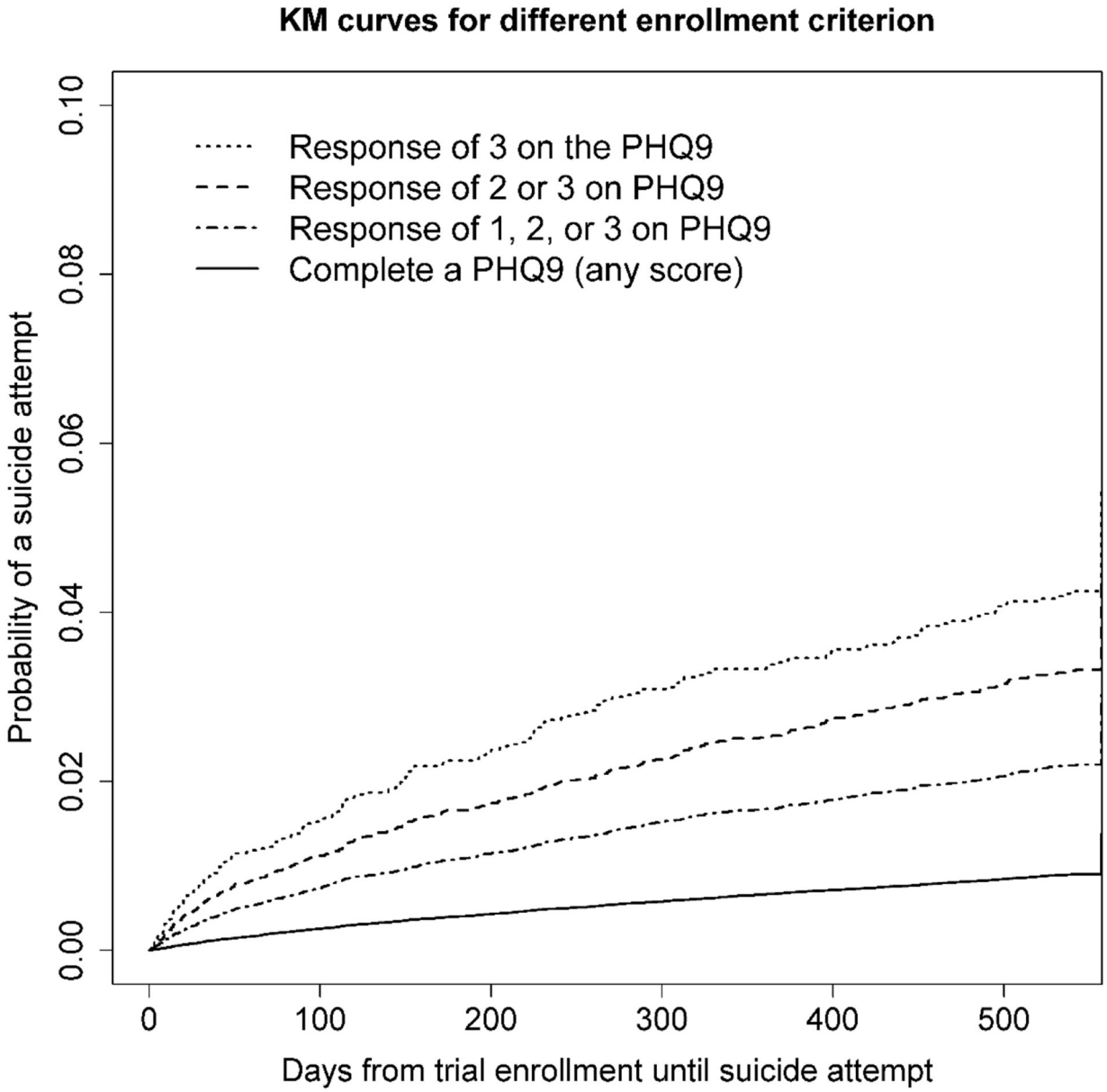

Figure 1.

Kaplan-Meier curves for days until first suicide attempt following response to item nine on the patient health questionnaire (PHQ) for four different trial eligibility criterions. Note, the scale on the y-axis (survival probabilities) has been limited to more readily make comparison between survival curves.

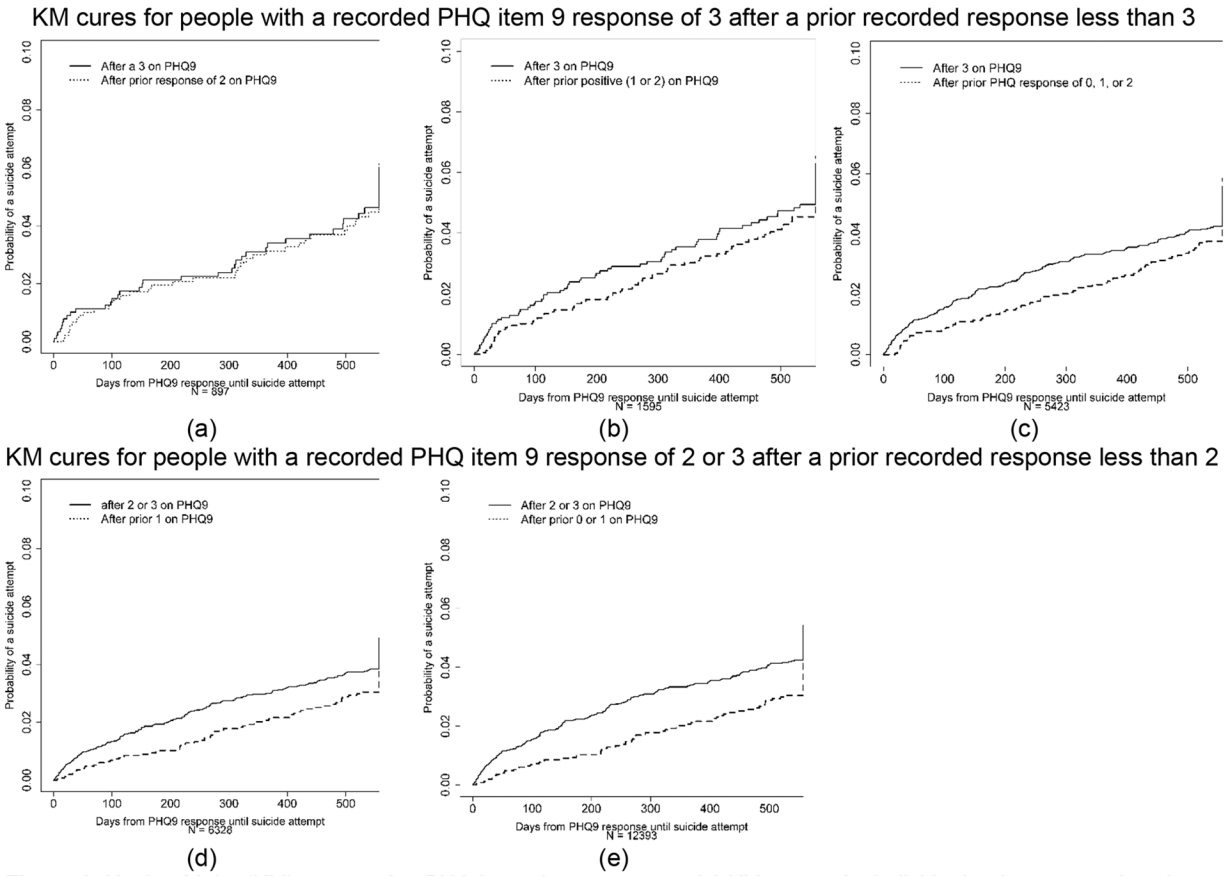

Figure 2.

Kaplan-Meier curves after PHQ item nine responses. **(a)** Kaplan-Meier curves for individuals who report a 3 on item nine of the PHQ, who previously reported a 2. The Kaplan-Meier curve for the 18 months following the reported 3 is shown with a solid line, while the Kaplan-Meier curve for the 18 months following the reported 2 is shown with a dotted line. **(b)** Kaplan-Meier curves for individuals who report a 3 on item nine of the PHQ and who previously reported a 1 or 2. The Kaplan-Meier curve for the 18 months following the reported 3 is shown with a solid line, while the Kaplan-Meier curve for the 18 months following the reported 1 or 2 is shown with a dotted line. **(c)** Kaplan-Meier curves for individuals who report a 3 on item nine of the PHQ and who previously reported a 0, 1, or 2. The Kaplan-Meier curve for the 18 months following the reported 3 is shown with a solid line, while the Kaplan-Meier curve for the 18 months following the reported 0, 1, or 2 is shown with a dotted line. **(d)** Kaplan-Meier curves for individuals who report a 2 or a 3 on item nine of the PHQ and who previously reported a 1. The Kaplan-Meier curve for the 18 months following the reported 2 or 3 is shown with a solid line, while the Kaplan-Meier curve for the 18 months following the reported 1 is shown with a dotted line. **(e)** Kaplan-Meier curves for individuals who report a 2 or a 3 on item nine of the PHQ and who previously reported a 0 or 1. The Kaplan-Meier curve for the 18 months following the reported 2 or 3 is shown with a solid line, while the Kaplan-Meier curve for the 18 months following the reported 0 or 1 is shown with a dotted line.

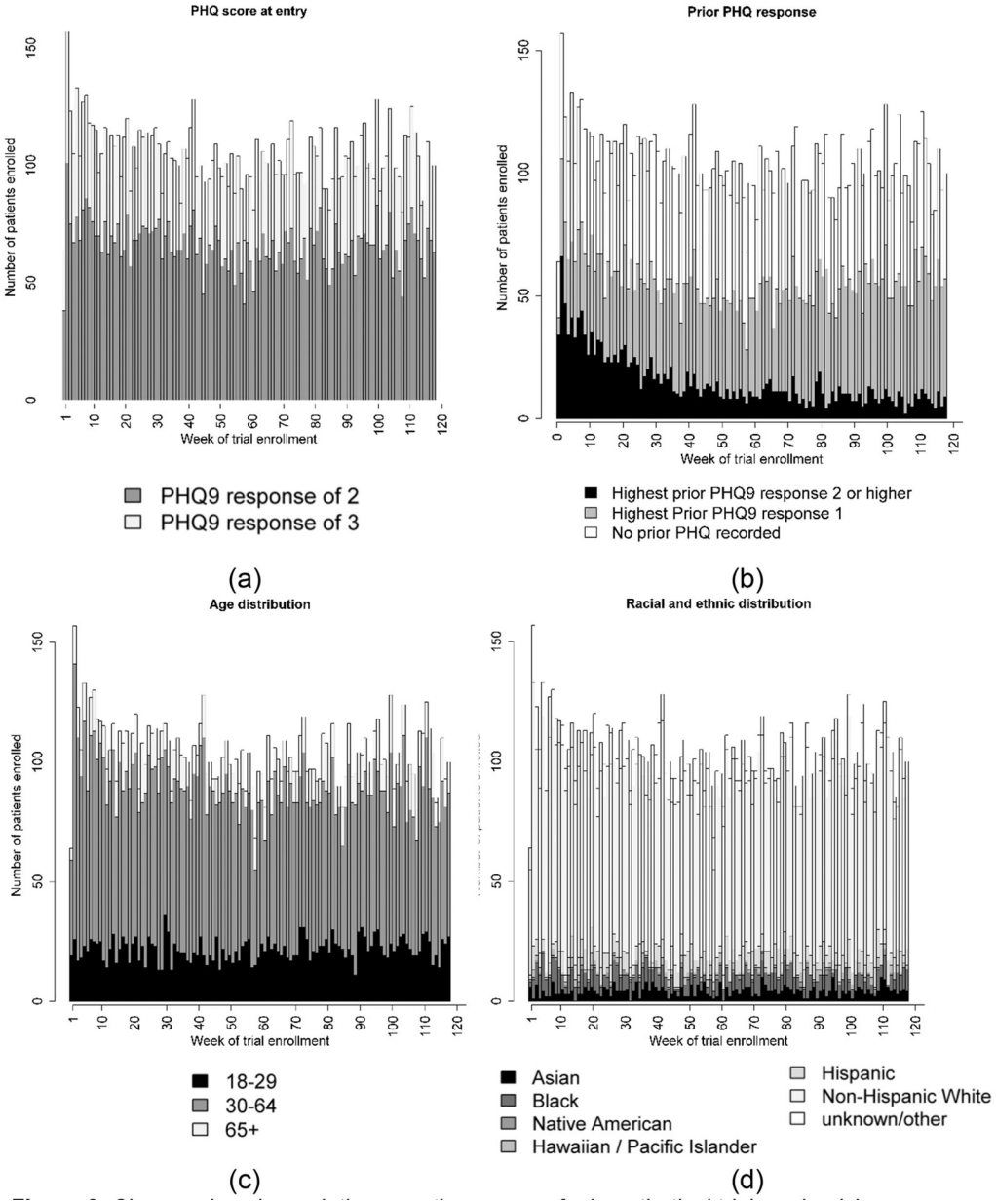

Figure 3.

Changes in subpopulation over the course of a hypothetical trial randomizing individuals who have a PHQ item nine response of a 2 or 3 recorded in their electronic health record between January 1 2010 and March 31 2012. **(a)** The proportion of individuals who are eligible for the trial each week because they responded with a 2 versus with a 3 on item nine of the PHQ. **(b)** The proportion of individuals each week that have prior recorded PHQ item nine response of a 2 or a 3. **(c)** The weekly age distribution. **(d)** The weekly race/ethnicity distribution.

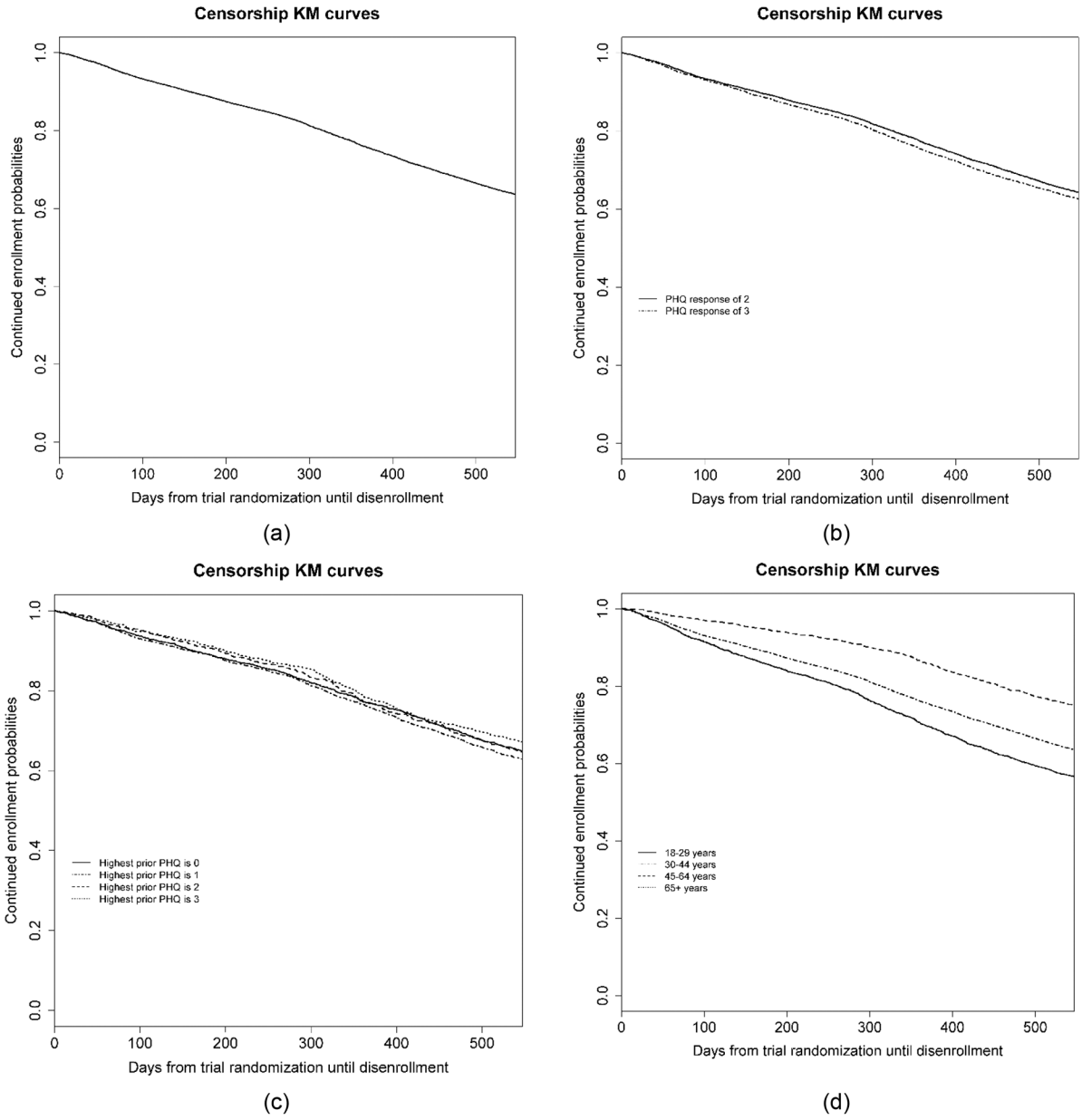

Figure 4.

Censorship patterns of individuals enrolled in a hypothetical trial randomizing patients after a response of a 2 or a 3 on the patient health questionnaire (PHQ). **(a)** Kaplan-Meier curve until disenrollment from the health plan (i.e. censorship). **(b)** Kaplan-Meier curve for censorship by response to item nine at randomization. **(c)** Kaplan-Meier curve for censorship by prior response on item nine of the PHQ. **(d)** Kaplan-Meier curve for censorship by age group.

Table 1.

Results of simulations using clinical and administrative data to estimate the sample size needed to detect a 25% reduction in the suicide attempt rate in the 18 months following randomization.

Number per arm	Number of sites	Percent with prior PHQ item 9 response of 2 or 3	Event rate, usual care arm	Event rate, intervention arm	Power, unadjusted	Power, adjusted for covariates
5000	1	12.7	0.048	0.036	0.76	0.76
6000	1	11.4	0.048	0.036	0.83	0.83
7000	1	10.5	0.048	0.036	0.88	0.88
8000	1	9.8	0.048	0.036	0.91	0.91
9000	1	9.3	0.048	0.036	0.95	0.95
10000	1	8.9	0.048	0.036	0.96	0.96
5000	2	20.4	0.048	0.036	0.76	0.76
6000	2	17.8	0.048	0.036	0.83	0.83
7000	2	16.0	0.048	0.036	0.88	0.88
8000	2	14.6	0.048	0.036	0.93	0.92
9000	2	13.6	0.048	0.036	0.95	0.95
10000	2	12.7	0.048	0.036	0.96	0.97
5000	3	28.1	0.048	0.036	0.76	0.76
6000	3	24.5	0.048	0.036	0.83	0.83
7000	3	21.4	0.048	0.036	0.89	0.88
8000	3	19.4	0.048	0.036	0.93	0.93
9000	3	17.8	0.048	0.036	0.94	0.94
10000	3	16.5	0.048	0.036	0.97	0.97