ppseq: An R Package for Sequential Predictive Probability Monitoring

by Emily C. Zabor, Brian P. Hobbs, Alexander M. Kaizer, and Michael J. Kane

Abstract Clinical trials in oncology are making increasing use of larger phase 1 and phase 2 sample sizes with the addition of baskets for different disease sub-types or multiple dosing groups, expansion cohorts to further study safety and obtain preliminary efficacy information, or randomization to further define efficacy or compare across doses, among other design elements. With the enlarged sample sizes come increasing ethical concerns regarding patients who enroll in clinical trials, and the need for rigorous statistical designs to ensure that trials can stop early for futility while maintaining traditional control of type I error and power. The R package **ppseq** provides a framework for designing early phase clinical trials using sequential monitoring based on the Bayesian predictive probability. Trial designs can be compared using interactive plots and optimized based on efficiency or accuracy.

Introduction

In the context of cytotoxic treatments, phase 1 trials in oncology traditionally have the primary aim of identifying the maximum tolerated dose (MTD), defined as the highest dose that still maintains a certain pre-specified rate of toxicity, most commonly 30%, in a dose-escalation phase. Designs for dose-escalation trials include the rule-based 3+3 design and the model-based continual reassessment method, among others. But with increasing study focused on non-cytotoxic treatments, such as immunotherapies, the MTD either may not exist or may not be relevant, and toxicities may either develop much later or even be chronic, making these treatments difficult to study with traditional dose-escalation designs. As a result, it is becoming increasingly common to include dose-expansion cohorts, in which additional patients are enrolled in phase 1 after the dose-escalation phase is complete. In this setup, the dose-escalation phase is considered phase 1a and used to assess the initial safety of multiple doses, then the dose-expansion phase is considered phase 1b and can have a variety of aims including to further refine the safety of one or more doses, to assess preliminary efficacy, to explore the treatment in various disease-specific subtypes that all share a common biomarker that the treatment is targeting, or to further characterize the pharmacokinetics and/or pharmacodynamics. The use of dose-expansion cohorts increased from 12% in 2006 to 38% in 2011 (Manji et al., 2013) and trials with dose-expansion cohorts led to higher response rates and more frequent success in phase 2 trials (Bugano et al., 2017).

But dose-expansion cohorts are not always planned in advance, and therefore may be subject to on-the-fly decision making that can lead to large sample sizes and poor statistical properties. For example, the KEYNOTE-001 trial of pembrolizumab, initially designed as a 3+3 dose-escalation trial, included multiple protocol amendments and ultimately enrolled a total of 495 non-small-cell lung cancer patients across six disease-specific expansion cohorts (Garon et al., 2015). In a basket trial of atezolizumab, an anti-PD-L1 treatment in patients with a variety of cancers and both with and without PD-L1 expression, an expansion cohort in metastatic urothelial bladder cancer ultimately enrolled 95 patients, despite the fact that no expansion cohort in this disease subtype was originally planned in the trial protocol but was rather added later in a protocol amendment in which the sample size was increased from what was initially planned (Petrylak et al., 2018; Powles et al., 2014).

Bayesian sequential predictive probability monitoring provides a natural framework for early phase oncology trial design, allowing flexibility to stop early for futility or safety concerns while also maintaining traditional levels of power and control of type I error, and has been proposed as an approach to interim monitoring in clinical trials previously [Dmitrienko and Wang (2006); Lee2008; Saville2014]. However, in order to be useful to investigators designing such trials, software must be made available that streamlines design. This paper introduces the **ppseq** package for the R software language (R Core Team, 2020), which provides functions to design one-sample and two-sample early phase clinical trials using sequential predictive probability monitoring. Interactive plots produced using **ggplot2** and **plotly** assist investigators in comparing designs based on different thresholds for decision making, and we suggest optimization criteria for selecting the ideal design.

Predictive probability monitoring

Consider the setting of a binary outcome, such as tumor response as measured by the RECIST criteria in the setting of a study in patients with solid tumors, where each patient, denoted by i, enrolled in the trial either has a response such that $x_i = 1$ or does not have a response such that $x_i = 0$. Then $X = \sum_{i=1}^{n} x_i$ represents the number of responses out of n currently observed patients up to a maximum

of N total enrolled patients. Let p represent the probability of response, where p_0 represents the null response rate under no treatment or the standard of care treatment and p_1 represents the alternative response rate under the experimental treatment. Most dose-expansion studies with an efficacy aim will wish to test the null hypothesis $H_0 : p \leq p_0$ versus the alternative hypothesis $H_1 : p \geq p_1$.

The Bayesian paradigm of statistics is founded on Bayes rule, which is a mathematical theory that specifies how to combine the prior distributions that define prior beliefs about parameters, such as the response rate p, with the observed data, such as the total number of responses X, yielding a posterior distribution. Here the prior distribution of the response rate $\pi(p)$ has a beta distribution $Beta(a_0, b_0)$ and our data X have a binomial distribution $bin(n, p)$. Combining the likelihood function for the observed data $L_x(p) \propto p^x(1-p)^{n-x}$ with the prior, we obtain the posterior distribution of the response rate, which follows the beta distribution $p|x \sim Beta(a_0 + x, b_0 + n - x)$. Using the posterior probability, which represents the probability of success based only on the data accrued so far, we would declare a treatment efficacious if $\Pr(p > p_0|X) > \theta$, where θ represents a pre-specified posterior decision threshold. The posterior predictive distribution of the number of future responses X^* in the remaining $n^* = N - n$ future patients follows a beta-binomial distribution $Beta - binomial(n^*, a_0 + x, b_0 + n - x)$. Then the posterior predictive probability (PPP), is calculated as $PPP = \sum_{x^*=0}^{n^*} \Pr(X^* = x^*|x) \times I(\Pr(p > p_0|X, X^* = x^*) > \theta)$. The posterior predictive probability represents the probability that the treatment will be declared efficacious at the end of the trial when full enrollment is reached. We would stop the trial early for futility if the posterior predictive probability dropped below a pre-specified threshold θ^*, i.e. $PPP < \theta^*$. Predictive probability thresholds closer to 0 lead to less frequent stopping for futility whereas thresholds near 1 lead to frequent stopping unless there is almost certain probability of success. Predictive probability provides an intuitive interim monitoring strategy for clinical trials that tells the investigator what the chances are of declaring the treatment efficacious at the end of the trial if we were to continue enrolling to the maximum planned sample size, based on the data observed in the trial to date.

Package overview

The **ppseq** package facilitates the design of clinical trials utilizing sequential predictive probability monitoring. The main challenge in designing such a trial is joint calibration of the posterior probability and posterior predictive probability thresholds to be used in the trial in order to maintain the desired level of power and type I error. The `calibrate_thresholds()` function will evaluate a grid of posterior and predictive thresholds provided by the user as vector inputs through the argument `pp_threshold` for posterior thresholds and the argument `ppp_threshold` for predictive thresholds. Other required arguments include the null response rate `p_null`, the alternative response rate `p_alt`, a vector of sample sizes at which interim analyses are to be performed `n` as well as the maximum total sample size `N`, the direction of the alternative hypothesis `direction`, a vector of the two hyperparameters of the prior beta distribution `prior`, and the number of posterior samples `S` and the number of simulated trial datasets `nsim`. The additional argument `delta` can specify the clinically meaningful difference between groups in the case of a two-sample trial design. The function returns a list, the first element of which is a `tibble` containing the posterior threshold, predictive threshold, the mean sample size under the null and the alternative, the proportion of positive trials under the null and alternative, and the proportion of trials stopped early under the null and alternative. The proportion of positive trials under the null represents the type I error and the proportion of positive trials under the alternative represents the power. The `print()` option will print the results summary for each combination of thresholds, filtered by an acceptable range of type I error and minimum power, if desired.

Optimization

After obtaining results for the all combinations of evaluated posterior and predictive thresholds, the next step is to select the ideal design from among the various options. The **ppseq** package introduces two optimization criteria to assist users in making a selection. The first, called the "optimal accuracy" design, identifies the design that minimizes the Euclidean distance to the top left point on a plot of the type I error by the power. The second, called the "optimal efficiency" design, identifies the design that minimizes the Euclidean distance to the top left point on a plot of the average sample size under the null by the average sample size under the alternative, subject to constraints on the type I error and power. The `optimize_design()` function will return a list that contains the details of each of the two optimal designs.

Visualizations

To assist users in comparing the results of the various design options, a `plot()` option is also available, that allows creation of static plots using the **ggplot2** package (Wickham, 2016) or interactive plots using the **plotly** package (Sievert, 2020). Two plots are produced, one plotting type I error by power and indicating the optimal accuracy design, and one plotting the average sample size under the null by the average sample size under the alternative and indicating the optimal efficiency design. The motivation for including an interactive graphics option was the utility of the additional information available when hovering over each point. Instead of simply eyeballing where points fall along the axes, users can see the specific type I error, power, average sample size under the null, average sample size under the alternative, the posterior and predictive thresholds associated with the design, as well as the distance to the upper left point on the plot.

Decision rules

Finally, to ease the implementation of clinical trials designed with sequential predictive probability monitoring, once a design has been selected, a table of decision rules can be produced using the `calc_decision_rules()` function. The function takes a vector of sample sizes at which interim analyses are to be performed `n` as well as the maximum total sample size `N`, the null value to compare to in the one-sample case `p0`, the posterior threshold of the selected design `theta`, the predictive threshold of the selected design `ppp`, the direction of the alternative hypothesis `direction`, a vector of the two hyperparameters of the prior beta distribution `prior`, and the number of posterior samples `S`. The additional argument `delta` can specify the clinically meaningful difference between groups in the case of a two-sample trial design. The function results in a `tibble` with the sample size at each interim analysis and the decision point r. The trial would stop at a given look if the number of observed responses is $\leq r$ otherwise the trial would continue enrolling if the number of observed responses is $> r$. At the end of the trial when the maximum planned sample size is reached, the treatment would be considered promising if the number of observed responses is $> r$.

Case study

To demonstrate the functionality of the **ppseq** package, we focus on a re-design of a dose-expansion cohort for the study of atezolizumab in metastatic urothelial carcinoma patients (mUC) using sequential predictive probability monitoring. Atezolizumab is an anti-PD-L1 treatment that was originally tested in the phase 1 setting in a basket trial across a variety of cancer sites harboring PD-L1 mutations. The atezolizumab expansion study in mUC had the primary aim of further evaluating safety, pharmacodynamics and pharmacokinetics and therefore was not designed to meet any specific criteria for type I error or power. An expansion cohort in mUC was not part of the original protocol design, but was rather added later through a protocol amendment. The expansion cohort in mUC ultimately enrolled a total of 95 participants (Powles et al., 2014). Other expansion cohorts that were included in the original protocol, including in renal-cell carcinoma, non-small-cell lung cancer, and melanoma, were planned to have a sample size of 40. These pre-planned expansion cohorts were designed with a single interim analysis for futility that would stop the trial if 0 responses were seen in the first 14 patients enrolled. According to the trial protocol, this futility rule is associated with at most a 4.4% chance of observing no responses in 14 patients if the true response rate is 20% or higher. The protocol also states the widths of the 90% confidence intervals for a sample size of 40 if the observed response rate is 30%. There was no stated decision rule for efficacy since efficacy was not an explicit aim of the expansion cohorts. In the re-design we assume a null, or unacceptable, response rate of 0.1 and an alternative, or acceptable, response rate of 0.2. We plan a study with up to a total of 95 participants. In our sequential predictive probability design we will check for futility after every 5 patients are enrolled. We consider posterior thresholds of 0, 0.7, 0.74, 0.78, 0.82, 0.86, 0.9, 0.92, 0.93, 0.94, 0.95, 0.96, 0.97, 0.98, 0.99, 0.999, 0.9999, 0.99999, and 1, and predictive thresholds of 0.05, 0.1, 0.15, and 0.2.

First we install and load the **ppseq** package.

```
install.packages("ppseq")

library(ppseq)
```

We use the `calibrate_thresholds()` function to obtain the operating characteristics of designs based on on each combination of posterior and predictive thresholds. Because of the inherent computation intensity in these calculations, this function relies on the **future** (Bengtsson, 2020) and **furrr** (Vaughan and Dancho, 2021) packages to parallelize computations. The user will be responsible for setting up a call to `future::plan()` that is appropriate to their operating environment and simulation

setting. The example in this case study was run on a Unix server with 192 cores, and we wished to use 76 cores to accommodate the 76 distinct designs that result from the 19 posterior by 4 predictive threshold grid. Because the code takes some time to run, the results of the below example code are available as a dataset called `one_sample_cal_tbl` included in the **ppseq** package.

```
library(future)

set.seed(123)

plan(multicore(workers = 76))

one_sample_cal_tbl <-
  calibrate_thresholds(
    p_null = 0.1,
    p_alt = 0.2,
    n = seq(5, 95, 5),
    N = 95,
    pp_threshold = c(0, 0.7, 0.74, 0.78, 0.82, 0.86, 0.9, 0.92, 0.93, 0.94,
                     0.95, 0.96, 0.97, 0.98, 0.99, 0.999, 0.9999, 0.99999, 1),
    ppp_threshold = seq(0.05, 0.2, 0.05),
    direction = "greater",
    delta = NULL,
    prior = c(0.5, 0.5),
    S = 5000,
    nsim = 1000
    )
```

Next we print the results table using the `print()` option, and limited to designs with type I error between 0.01 and 0.2, and a minimum power of 0.7. We find that 35 of the 76 designs meet these criteria for type I error and power.

```
print(
  one_sample_cal_tbl,
  type1_range = c(0.01, 0.2),
  minimum_power = 0.7
  )
```

```
#> # A tibble: 35 x 8
#>    pp_threshold ppp_threshold mean_n1_null prop_pos_null prop_stopped_null
#>           <dbl>         <dbl>        <dbl>         <dbl>              <dbl>
#>  1          0.7           0.1         18.2         0.199              0.543
#>  2          0.7           0.15        16.7         0.182              0.728
#>  3          0.7           0.2         11.6         0.139              0.801
#>  4          0.74          0.1         18.1         0.199              0.544
#>  5          0.74          0.15        16.7         0.182              0.728
#>  6          0.74          0.2         11.6         0.139              0.801
#>  7          0.78          0.1         18.1         0.200              0.544
#>  8          0.78          0.15        16.7         0.182              0.728
#>  9          0.78          0.2         11.6         0.139              0.801
#> 10          0.82          0.1         18.1         0.200              0.544
#> # ... with 25 more rows, and 3 more variables: mean_n1_alt <dbl>,
#> #   prop_pos_alt <dbl>, prop_stopped_alt <dbl>
```

We use the `optimize_designs()` function to obtain the details of the optimal accuracy and optimal efficiency designs, limited to our desired range of type I error and minimum power. We find that the optimal accuracy design is the one with posterior threshold 0.9 and predictive threshold 0.05. It has a type I error of 0.072, power of 0.883, average sample size under the null of 51, and average sample size under the alternative of 91. The optimal efficiency design is the one with posterior threshold of 0.92 and predictive threshold of 0.1. It has a type I error of 0.06, power of 0.796, average sample size under the null of 39, and average sample size under the alternative of 82. For comparison, the original design of the atezolizumab expansion cohort in mUC, with a single look for futility after the first 14 patients, has a type I error of 0.005, power of 0.528, average sample size under the null of 76, and average sample size under the alternative of 92.

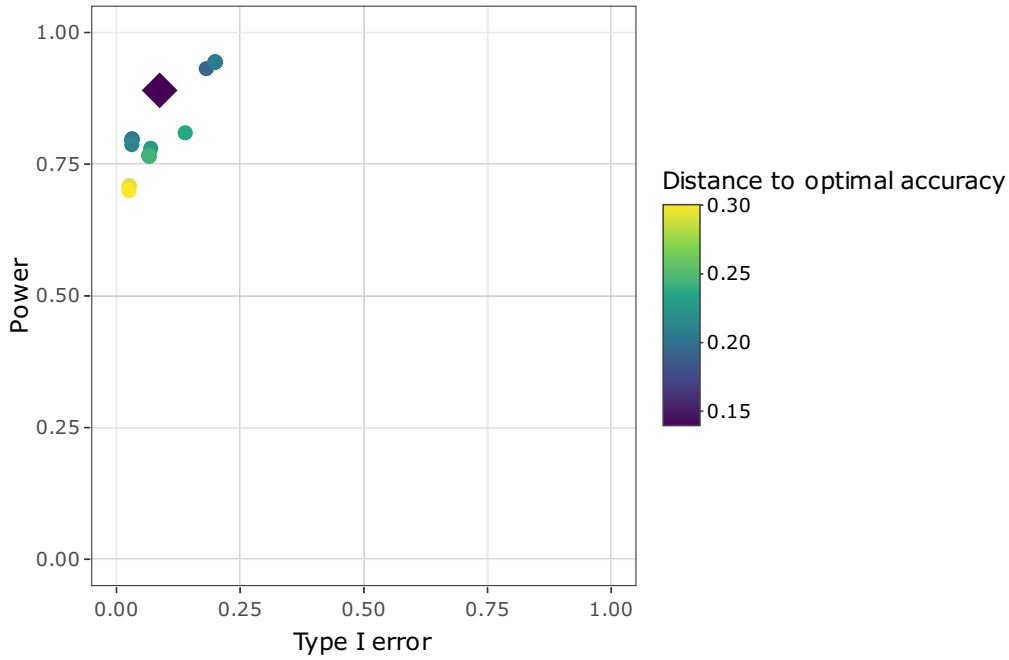
```
optimize_design(
  one_sample_cal_tbl,
```

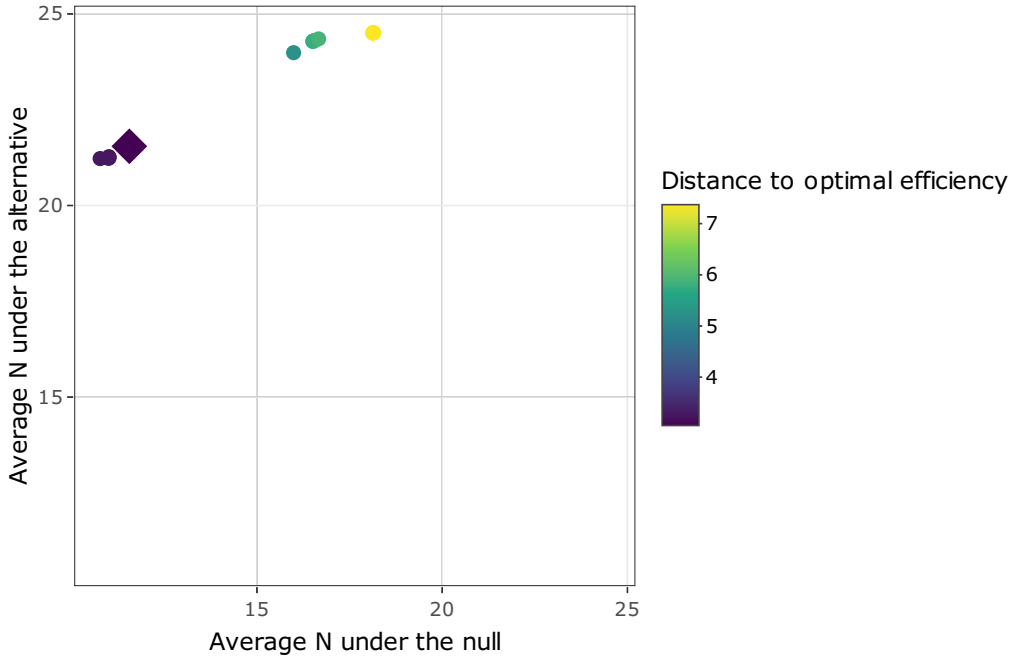
```
  type1_range = c(0.05, 0.1),
  minimum_power = 0.7
  )
```

```
#> $`Optimal accuracy design:`
#> # A tibble: 1 x 6
#>   pp_threshold ppp_threshold `Type I error` Power `Average N under the null`
#>          <dbl>         <dbl>          <dbl> <dbl>                      <dbl>
#> 1         0.93           0.1         0.0872 0.890                       16.7
#>   `Average N under the alternative`
#>                               <dbl>
#> 1                              24.3
#>
#> $`Optimal efficiency design:`
#> # A tibble: 1 x 6
#>   pp_threshold ppp_threshold `Type I error` Power `Average N under the null`
#>          <dbl>         <dbl>          <dbl> <dbl>                      <dbl>
#> 1         0.93          0.15         0.0692 0.780                       11.6
#>   `Average N under the alternative`
#>                               <dbl>
#> 1                              21.5
```

To compare these optimal designs with all other designs, we can use the `plot()` function with the `plotly = TRUE` option to obtain interactive visualizations.

```
plot(
  one_sample_cal_tbl,
  type1_range = c(0.01, 0.2),
  minimum_power = 0.7,
  plotly = TRUE
  )
```


1.00
0.75
0.50
0.25
0.00
Power
0.00
0.25
0.50
0.75
1.00
Type I error
Distance to optimal accuracy
0.30
0.25
0.20
0.15



In this case we may choose to use the optimal efficiency design, which has the desirable trait of a very small average sample size under the null of just 39 patients, while still maintaining reasonable type I error of 0.06 and power of 0.796. This design would allow us to stop early if the treatment were inefficacious, thus preserving valuable financial resources for use in studying more promising treatments and preventing our human subjects from continuing an ineffective treatment. Finally, we generate the decision table associated with the selected design for use in making decision at each interim analysis during the conduct of the trial. Because of the computational time involved, the results of the below example code are available as a dataset called `one_sample_decision_tbl` included in the **ppseq** package. In the results table, we see that at the first interim futility look after just 5 patients, we would not stop the trial. After the first 10 patients we would stop the trial if there were 0 responses, and so on. At the end of the trial when all 95 patients have accrued, we would declare the treatment promising of further study if there were >=14 responses.

```
set.seed(123)

one_sample_decision_tbl <-
  calc_decision_rules(
    n = seq(5, 95, 5),
    N = 95,
    theta = 0.92,
    ppp = 0.1,
    p0 = 0.1,
    direction = "greater",
    delta = NULL,
```

```
    prior = c(0.5, 0.5),
    S = 5000
    )

one_sample_decision_tbl

#> # A tibble: 19 x 2
#>        n     r
#>    <dbl> <int>
#>  1     5    NA
#>  2    10     0
#>  3    15     0
#>  4    20     1
#>  5    25     1
#>  6    30     2
#>  7    35     2
#>  8    40     3
#>  9    45     4
#> 10    50     4
#> 11    55     5
#> 12    60     6
#> 13    65     7
#> 14    70     8
#> 15    75     8
#> 16    80     9
#> 17    85    10
#> 18    90    11
#> 19    95    13
```

Summary

With the focus of early stage clinical trial research in oncology shifting away from the study of cytotoxic treatments and toward immunotherapies and other non-cytotoxic treatments, new approaches to clinical trial design are needed that move beyond the traditional search for the maximum tolerated dose. Bayesian sequential predictive probability monitoring provides a natural and flexible way to expand the number of patients studied in phase 1 or to design phase 2 trials that allow for efficient early stopping for futility while maintaining control of type I error and power. The **ppseq** package implements functionality to consider a range of posterior and predictive thresholds for a given study design and identify the optimal design based on accuracy (i.e. type I error and power) or efficiency (i.e. average sample sizes under the null and alternative). Interactive visualization options are provided to ease comparison of the resulting design options. Once an ideal design is selected, a table of decision rules can be obtained to make trial conduct simple and straightforward.

Bibliography

H. Bengtsson. A unifying framework for parallel and distributed processing in r using futures, aug 2020. URL https://arxiv.org/abs/2008.00553. [p3]

D. D. G. Bugano, K. Hess, D. L. F. Jardim, A. Zer, F. Meric-Bernstam, L. L. Siu, A. R. A. Razak, and D. S. Hong. Use of expansion cohorts in phase i trials and probability of success in phase ii for 381 anticancer drugs. *Clin Cancer Res*, 23(15):4020–4026, 2017. ISSN 1078-0432 (Print) 1078-0432. doi: 10.1158/1078-0432.Ccr-16-2354. [p1]

A. Dmitrienko and M. D. Wang. Bayesian predictive approach to interim monitoring in clinical trials. *Stat Med*, 25(13):2178–95, 2006. ISSN 0277-6715 (Print) 0277-6715. doi: 10.1002/sim.2204. [p1]

E. B. Garon, N. A. Rizvi, R. Hui, N. Leighl, A. S. Balmanoukian, J. P. Eder, A. Patnaik, C. Aggarwal, M. Gubens, L. Horn, E. Carcereny, M. J. Ahn, E. Felip, J. S. Lee, M. D. Hellmann, O. Hamid, J. W. Goldman, J. C. Soria, M. Dolled-Filhart, R. Z. Rutledge, J. Zhang, J. K. Lunceford, R. Rangwala, G. M. Lubiniecki, C. Roach, K. Emancipator, and L. Gandhi. Pembrolizumab for the treatment of non-small-cell lung cancer. *N Engl J Med*, 372(21):2018–28, 2015. ISSN 0028-4793. doi: 10.1056/NEJMoa1501824. [p1]

A. Manji, I. Brana, E. Amir, G. Tomlinson, I. F. Tannock, P. L. Bedard, A. Oza, L. L. Siu, and A. R. Razak. Evolution of clinical trial design in early drug development: systematic review of expansion cohort use in single-agent phase i cancer trials. *J Clin Oncol*, 31(33):4260–7, 2013. ISSN 0732-183x. doi: 10.1200/jco.2012.47.4957. [p1]

D. P. Petrylak, T. Powles, J. Bellmunt, F. Braiteh, Y. Loriot, R. Morales-Barrera, H. A. Burris, J. W. Kim, B. Ding, C. Kaiser, M. Fassò, C. O'Hear, and N. J. Vogelzang. Atezolizumab (mpdl3280a) monotherapy for patients with metastatic urothelial cancer: Long-term outcomes from a phase 1 study. *JAMA Oncol*, 4(4):537–544, 2018. ISSN 2374-2437 (Print) 2374-2437. doi: 10.1001/jamaoncol.2017.5440. [p1]

T. Powles, J. P. Eder, G. D. Fine, F. S. Braiteh, Y. Loriot, C. Cruz, J. Bellmunt, H. A. Burris, D. P. Petrylak, S. L. Teng, X. Shen, Z. Boyd, P. S. Hegde, D. S. Chen, and N. J. Vogelzang. Mpdl3280a (anti-pd-l1) treatment leads to clinical activity in metastatic bladder cancer. *Nature*, 515(7528):558–62, 2014. ISSN 0028-0836. doi: 10.1038/nature13904. [p1, 3]

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL http://www.R-project.org/. ISBN 3-900051-07-0. [p1]

C. Sievert. *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC, 2020. ISBN 9781138331457. URL https://plotly-r.com. [p3]

D. Vaughan and M. Dancho. *furrr: Apply Mapping Functions in Parallel using Futures*, 2021. URL https://CRAN.R-project.org/package=furrr. R package version 0.2.2. [p3]

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org. [p3]

Emily C. Zabor
Department of Quantitative Health Sciences & Taussig Cancer Institute, Cleveland Clinic
9500 Euclid Ave. CA-60
Cleveland, OH 44195 USA
http://www.emilyzabor.com/
ORCiD: 0000-0002-1402-4498
zabore2@ccf.org

Brian P. Hobbs
Dell Medical School, The University of Texas at Austin
true
Austin, TX 78712
ORCiD: 0000-0003-2189-5846
brian.hobbs@austin.utexas.edu

Alexander M. Kaizer
Department of Biostatistics & Informatics, University of Colorado-Anschutz Medical Campus,
13001 E. 17th Place Campus Box B119
Aurora, CO 80045
ORCiD: 0000-0003-2334-5514
alex.kaizer@cuanschutz.edu

Michael J. Kane
Department of Biostatistics, Yale University
22 Mill Pond Drive
Guilford, CT 06437
ORCiD: 0000-0003-1899-6662
michael.kane@yale.edu