Supplementary Material "Estimating quantile treatment effect on the original scale of the outcome variable: a case study of common cold treatments"

Harri Hemilä and Matti Pirinen
University of Helsinki

2025-08-26

Supplement to a manuscript submitted to Trials https://trialsjournal.biomedcentral.com/

Contents

1. Interpretation of BQTE
   1.1 When does BQTE estimate an individual-level treatment effect?
   1.2 Bounds on the individual-level treatment effect
   1.3 Variance of the individual-level treatment effect

2. Validating the `bqte` function
   2.1 Generating continuous example data
   2.2 The range of valid estimates
   2.3 The number of quantile points used in estimation
   2.4 Magnitude of estimation error and coverage of confidence intervals in discrete data
   2.5 Comparison between BQTE and QTE
   2.6 Variance of BQTE and QTE

3. Source of data of the included examples

4. Influence of missing data in the Carrageenan data sets

1. Interpretation of BQTE

Back-transformed quantile treatment effect (BQTE) is a function that describes the difference between the distributions of outcome values of two populations. We will call the two populations that we compare as the **treatment** population and the **control** population, motivated by the need to compare how a particular treatment affects the outcome value in a randomized controlled trial (RCT). In an RCT, the study sample is randomly split into the treatment group (that represents the treated population) and the control group (that represents the control population). The treatment of interest is applied to the treatment group and placebo is applied to the control group. The possible difference in the outcome distributions between the treatment group and the control group is interpreted as the population-level treatment effect. We will use the BQTE function to quantify this difference between the two populations.

Definition of BQTE. For any random variable X, let $F_X(x) = \mathrm{Prob}(X \leq x)$ be its cumulative distribution function, and $F_X^{-1}(p) = \inf\{x \,|\, F_X(x) \geq p\}$ the corresponding quantile function defined for all quantile levels $p \in (0, 1)$. Let C be an outcome variable of the control population and T an outcome variable of the treated population. When both populations have **continuous** distributions, we define,

$$\mathrm{BQTE}(x) = {F_T}^{-1}\left(F_C(x)\right) - x.$$

Here is an intuitive summary how to find BQTE(x) for continuous outcome variables:

1. Find which proportion of the control population has outcome values $\leq x$. Call this proportion $F_C(x)$.
2. Find what is the corresponding outcome value $y(x)$ in the treated population for which the proportion of the treated population having outcome values $\leq y(x)$ is the same $F_C(x)$.
3. Define $\mathrm{BQTE}(x) = y(x) - x$.

In case the outcome variable is **discrete**, we define BQTE by averaging in the treated population over the quantile levels corresponding to the control population's value x. Thus, let $I(x) = \{p \,|\, F_C^{-1}(p) = x\} \subseteq [0, 1]$ be the interval of the quantile levels corresponding to x in the control population, and define

$$\mathrm{BQTE}(x) = \mathrm{E}_{p \sim I(x)}\left({F_T}^{-1}(p)\right) - x = \frac{1}{|I(x)|}\int {F_T}^{-1}(p)\, \mathbb{I}\left(p \in I(x)\right)\, \mathrm{d}p - x.$$

Above, the expectation is taken over the random variable p having a Uniform distribution on interval $I(x)$, $|I(x)|$ is the length of the interval, and $\mathbb{I}(\cdot)$ is the indicator function.

Here is an intuitive summary how to find BQTE(x) for discrete outcome variables:

1. Find the interval of quantile levels $I(x) = (p_1(x), p_2(x)]$ for which the control population has outcome value x.
2. Find what is the corresponding average outcome value $y(x)$ in the treated population corresponding to the same interval of quantile levels $I(x)$.
3. Define $\mathrm{BQTE}(x) = y(x) - x$.

Thus, BQTE describes the difference between the outcome value of the treated population and the control population when this comparison is done between treated and control individuals whose outcome values rank similarly among their respective populations. As the value of BQTE may vary across outcome values, we present BQTE as a function of outcome value x in the control group.

The difference in the means of the populations, called the **average treatment effect** (ATE), is the most common approach to assess the effect of the treatment at the population level. However, the BQTE function measures the difference between the two populations in a more detailed way than the single value of ATE. For example, if we observe that BQTE is not constant across the outcome values, then we can conclude that the individual-level treatment effect (see below for definition) cannot be constant across individuals. Or, if the relative BQTE (defined as $\mathrm{BQTE}(x)/x$) is not constant, then we know that the individual-level treatment effect cannot be proportional to the untreated outcome value by a single constant. Neither of these pieces of information can be inferred from the value of the ATE alone. Next, we consider what additional information we can extract from the BQTE function.

1.1. When does BQTE estimate an individual-level treatment effect?

We define two instances of the concept of **treatment effect**. A **population-level treatment effect** is the difference between the outcome distributions of the treatment and control populations and it can be estimated from RCT data by different approaches such as by ATE or BQTE which contain different levels of information about the population-level treatment effect. Instead, an **individual-level treatment effect** is a theoretical construct that cannot, in general, be estimated from an RCT without some additional assumptions. To formally define the individual-level treatment effect, we follow the framework of the potential outcomes by Rubin (1974; Journal of Educational Psychology, 66(5), 688–701). Suppose that we have not yet carried out the experiment. For any individual i, we can imagine two potential outcome variables:

- X_i would be the outcome variable if individual i belonged to the control population,
- Y_i would be the outcome variable if individual i belonged to the treated population.

Before the experiment, we consider both X_i and Y_i as random variables and their difference, $\Delta_i = Y_i - X_i$, defines a random variable that represents the theoretical treatment effect for individual i. Once we do the experiment, we will observe a value for either X_i (if i was assigned to the control group) or Y_i (if i was assigned to the treatment group) but never for both X_i and Y_i simultaneously. Thus, the experiment does not provide a direct way to estimate the treatment effect Δ_i of individual i. Note that from now on, we will refer by the random variables X and Y to the untreated and treated potential outcomes of the *same* individual even when we do not specify that same individual by using a common subscript as above. In particular, the random variables X and Y are not independent because they refer to the same individual.

To conceptually connect BQTE, that quantifies the population-level treatment effect, to the individual-level treatment effects, we will consider a theoretical concept of order-preserving treatments.

Definition. We say that a **treatment preserves order** (or that the **treatment is order-preserving**) if the untreated outcome values of a set of individuals have, on average, the same relative order as the treated outcome values of the same set of individuals would have had, had these individuals been treated. Formally, we define that the treatment preserves order when, for every outcome value x, $F_T\left(\mathrm{E}(Y \mid X = x)\right) = F_C(x)$.

For example, consider a trial where the outcome is the duration of the common cold, the treatment group is treated with zinc acetate lozenges and the control group is given placebo. The assumption that the treatment preserves order means that the durations of the colds of the control individuals would have had approximately the same relative order if those individuals had been treated with zinc acetate lozenges rather than given placebo.

While, based on RCT data, we cannot directly confirm that a treatment preserves order, we can make the following remarks.

- First, the assumption that the treatment preserves order seems more plausible the more homogeneous the individuals in the study population are, because within a homogeneous population there are less differences between the individuals that could affect how the treatment applies to each individual. Thus, for example, the assumption of order-preserving treatment seems more plausible in an RCT that studies females between 40-45 years of age than in another RCT that studies both males and females between 18-80 years of age.

- Second, in case that some covariate information were available on the individuals included in the RCT (such as sex, age, medication, disease history etc.), then we would expect that under an order-preserving treatment the way how the covariates predict the outcome values would be similar in the treated population as it is in the control population. For example, if male individuals tend to have longer disease duration than female individuals in the control group, then the same should be true in the treatment group if the treatment preserves order. If, on the other hand, we observe that the covariates such as sex do not predict the disease duration consistently in both the control and the treated populations, then the assumption of order-preserving treatment may not be appropriate. Note that the treatment could preserve order in a subpopulation (such as in males only) even if it did not preserve order in a larger population (such as in both sexes together).

The connection between the order-preserving treatments and the individual-level treatment effects (for continuous outcome variables) can be formulated as follows: **BQTE can be interpreted as the expected value of the individual-level treatment effect if and only if the treatment preserves order.**

To justify the claim, consider a treatment that preserves order for a continuous outcome, that is, $F_T\left(\mathrm{E}(Y \mid X = x)\right) = F_C(x)$ for all x. By applying ${F_T}^{-1}$ on both sides of the equation we get

$$\mathrm{E}(Y \mid X = x) = {F_T}^{-1}\left(F_C(x)\right)$$

and since by definition $\mathrm{BQTE}(x) = {F_T}^{-1}(F_C(x)) - x$ it follows that

$$\mathrm{BQTE}(x) = \mathrm{E}(Y \mid X = x) - x = \mathrm{E}(Y - X \mid X = x).$$

Thus, BQTE is the conditional expectation of the individual-level treatment effect and we can interpret BQTE as an estimator of the individual-level treatment effect.

On the other hand, suppose that the treatment does not preserve order. Then there exists some outcome value x in the control population whose quantile level $F_C(x)$ does not remain constant under the treatment: $F_T\left(\mathrm{E}(Y \mid X = x)\right) \neq F_C(x)$. By applying ${F_T}^{-1}$ on both sides of this inequality, and following the steps above, we see that

$$\mathrm{BQTE}(x) \neq \mathrm{E}(Y \mid X = x) - x = \mathrm{E}(Y - X \mid X = x).$$

Thus, if the treatment does not preserve order, $\mathrm{BQTE}(x)$ does not match with the conditional expectation of the individual-level treatment effect.

We have seen that the BQTE can be interpreted as the expected value of the individual-level treatment effect if and only if the treatment preserves order. Unfortunately, there is no general way to verify whether a given treatment preserves order based on RCT data. However, as we show below, we can extract useful information about the individual-level treatment effects from the BQTE function even without the assumption about the preservation of order.

1.2 Bounds on the individual-level treatment effect

In this section, we compute upper and lower bounds on the individual-level treatment effect and further extend these bounds to the ATE in the tails of the distribution.

Consider a treatment that does not necessarily preserve order. Let us denote the range of the possible outcome value of the treated population by (L, U), where L and U refer, respectively, to the lower and upper bounds of the outcome value in the treated population. We use the following estimators for these bounds:

$$\begin{aligned} \widehat{L} &= \min\{\text{all outcome values of the treatment group}\}, \\ \widehat{U} &= \max\{\text{all outcome values of the treatment group}\}, \end{aligned}$$

that is, we use the observed minimum and maximum of the outcome values of the treatment group as the estimates of L and U, respectively. This leads to an estimated lower bound of $\widehat{L} - x$ and an estimated upper bound of $\widehat{U} - x$ for the treatment effect of an untreated individual with an outcome value $X = x$. The logic is that since we expect that (almost) all outcome values in the treatment population are within the interval $\left[\widehat{L}, \widehat{U}\right]$, we expect that the treatment effect at $X = x$ is (almost always) bounded within $\left[\widehat{L} - x, \widehat{U} - x\right]$. The accuracy of these bounds increases with the sample size of the treatment group and decreases with increasing kurtosis (fatness of the tails) of the outcome distribution of the treated population.

We can extend the idea of the bounds to the ATE of the upper or lower tail as follows. For the outcome value x, define the **lower tail average treatment effect (LTATE)** as

$$\mathrm{LTATE}(x) = \mathrm{E}\left(Y \mid X \leq x\right) - \mathrm{E}\left(X \mid X \leq x\right).$$

and the **upper tail average treatment effect (UTATE)** as

$$\mathrm{UTATE}(x) = \mathrm{E}\left(Y \mid X \geq x\right) - \mathrm{E}\left(X \mid X \geq x\right).$$

While we cannot directly estimate these quantities for a general treatment, we can get a lower bound for LTATE by **lower tail back-transformed quantile treatment effect (LTBQTE)** and an upper bound for UTATE by **upper tail back-transformed quantile treatment effect (UTBQTE)** that are defined next.

Let $I_L(x) = [0, p_L(x))$ be the maximal half-open interval of quantile levels for which $X \leq x$ and $I_U(x) = (p_U(x), 1]$ be the maximal half-open interval of quantile levels for which $X \geq x$, that is, $p_L(x) = F_C(x)$ and $p_U(x) = 1 - \mathrm{Prob}(X \geq x)$. We define,

$$\begin{aligned}\mathrm{LTBQTE}(x) &= \mathrm{E}_{p\sim I_L(x)}\left({F_T}^{-1}(p)\right) - \mathrm{E}\left(X \,|\, X \leq x\right),\\ \mathrm{UTBQTE}(x) &= \mathrm{E}_{p\sim I_U(x)}\left({F_T}^{-1}(p)\right) - \mathrm{E}\left(X \,|\, X \geq x\right),\end{aligned}$$

where the expectations of the type $\mathrm{E}_{p\sim I}$ are taken with respect to p having a uniform distribution over the interval I.

For the case of continuous outcome distributions, we can also write these quantities as

$$\begin{aligned}\mathrm{LTBQTE}(x) &= \mathrm{E}\left(Y \,|\, Y \leq {F_T}^{-1}(F_C(x))\right) - \mathrm{E}\left(X \,|\, X \leq x\right),\\ \mathrm{UTBQTE}(x) &= \mathrm{E}\left(Y \,|\, Y \geq {F_T}^{-1}(F_C(x))\right) - \mathrm{E}\left(X \,|\, X \geq x\right).\end{aligned}$$

Thus, LTBQTE (UTBQTE) is the difference in the expected values of the treatment and control populations when the two expectations are computed for the same proportion of the population from the lower (upper) tail. Note that LTBQTE(x) approaches ATE as x gets large while UTBQTE(x) approaches ATE as x gets small.

For these quantities, the following relations apply:

$$\mathrm{LTATE}(x) \geq \mathrm{LTBQTE}(x) \quad \text{and} \quad \mathrm{UTATE}(x) \leq \mathrm{UTBQTE}(x).$$

To verify the argument, let us consider the first inequality in the continuous setting. We will show that $\mathrm{LTATE}(x) \geq \mathrm{LTBQTE}(x)$, which equals to showing that

$$\mathrm{E}\left(Y \,|\, X \leq x\right) \geq \mathrm{E}\left(Y \,|\, Y \leq {F_T}^{-1}(F_C(x))\right).$$

Both quantities in the above inequality can be presented as integrals over random variable Y, the left-hand side over the event $A = \{X \leq x\}$ and the right-hand side over the event $B = \{Y \leq {F_T}^{-1}(F_C(x))\}$. Let μ be the underlying probability measure of the two-dimensional random variable (X, Y). By definition of the marginal CDFs F_C of X and F_T of Y, both events have the same probability $\mu(A) = \mu(B) = F_C(x)$. Intuitively, among all events with the same probability, the event B is the one with the smallest expected value of Y because B represents exactly the extreme left tail of the distribution of Y. To formally prove this, let's consider the sample space Ω of the experiment. Then, $A = \{\omega \in \Omega \,|\, X(\omega) \leq x\}$ and $B = \{\omega \in \Omega \,|\, Y(\omega) \leq {F_T}^{-1}(F_C(x))\}$. In particular, for all $\omega \in A \backslash B$, $Y(\omega) > {F_T}^{-1}(F_C(x))$ and for all $\omega \in B \backslash A$, $Y(\omega) \leq {F_T}^{-1}(F_C(x))$. Additionally, $\mu(A \backslash B) = \mu(B \backslash A)$ because $\mu(A) = \mu(B)$. Hence,

$$\begin{aligned}\mathrm{E}\left(Y \,|\, X \leq x\right) \cdot F_C(x) &= \int_A y \,\mathrm{d}\mu(y)\\ &= \int_{A\cap B} y \,\mathrm{d}\mu(y) + \int_{A\backslash B} y \,\mathrm{d}\mu(y)\\ &\geq \int_{A\cap B} y \,\mathrm{d}\mu(y) + \mu(A\backslash B) \cdot {F_T}^{-1}(F_C(x))\\ &= \int_{A\cap B} y \,\mathrm{d}\mu(y) + \mu(B\backslash A) \cdot {F_T}^{-1}(F_C(x))\\ &\geq \int_{A\cap B} y \,\mathrm{d}\mu(y) + \int_{B\backslash A} y \,\mathrm{d}\mu(y)\\ &= \int_B y \,\mathrm{d}\mu(y)\\ &= \mathrm{E}\left(Y \,|\, Y \leq {F_T}^{-1}(F_C(x))\right) \cdot F_C(x).\end{aligned}$$

As a conclusion, we have shown that LTBQTE cannot overestimate the ATE in the lower tail.

The condition $\text{UTATE}(x) \leq \text{UTBQTE}(x)$ can be verified similarly.

Estimating LTBQTE(x) and UTBQTE(x). Simple estimators for LTBQTE(x) and UTBQTE(x) result by replacing the expectations by the numerical averages of the observations in the tails. Assume that we have N_C values from the control population and N_T values from the treated population, in either the original or the bootstrapped sample. As with the estimator $\widehat{\text{BQTE}}$, we first find K empirical quantile points corresponding to the quantile levels of $1/(K+1), \ldots, K/(K+1)$ from both groups, and these quantile points are denoted here by x_i and y_i for the control and treatment groups, respectively. To compute an estimate $\widehat{\text{LTBQTE}}(x)$ at a value x observed in the control sample, we first find the largest quantile index ℓ for which $x_\ell = x$, and estimate

$$\widehat{\text{LTBQTE}}(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} y_i - \frac{1}{\ell} \sum_{i=1}^{\ell} x_i.$$

Similarly, to estimate $\widehat{\text{UTBQTE}}(x)$, we find the smallest quantile index u for which $x_u = x$, and estimate

$$\widehat{\text{UTBQTE}}(x) = \frac{1}{K-u+1} \sum_{i=u}^{K} y_i - \frac{1}{K-u+1} \sum_{i=u}^{K} x_i.$$

Finally, for values not observed in the control sample, the estimate is interpolated by linear interpolation or by smoothing with splines.

1.3. Variance of the individual-level treatment effect

In this section, we consider how much the individual-level treatment effect varies across the individuals and how that variance depends on the assumption of order-preserving treatment. For that, we write the variance of the individual-level treatment effect $Y-X$ in terms of correlation $r = \text{cor}(Y, X)$ and variance $v_C = \text{var}(X)$ of the control population and $v_T = \text{var}(Y)$ of the treated population as follows

$$\begin{aligned} \text{var}(Y - X) &= v_C + v_T - 2\,\text{cov}(Y, X) \\ &= v_C + v_T - 2\,r\,\sqrt{v_C \cdot v_T}. \end{aligned}$$

Thus, the variance of the treatment effect is the smallest possible when $r = 1$, in which case the treatment preserves order, and is the largest possible when $r = -1$, in which case the treatment completely flips the order around.

We conclude that the order-preserving treatment with a linear relationship between the treated and untreated outcome values is the most conservative treatment in the sense of minimizing the variance of the individual-level treatment effect in the population.

2. Validating the `bqte` function

Let's test `bqte()` function by generating data for the treatment (T) group and the control (C) group.

Continuous case. For a given choice of continuous distributions, we can compute the exact BQTE for each possible outcome value in controls by first using cumulative density function (CDF) in controls to get a correct tail probability in the control distribution and then applying the quantile function (inverse of CDF) in the treatment group to get the corresponding outcome value in that group. Difference between the two outcome values is the exact BQTE.

Discrete case. For discrete distributions, the simulation setting is slightly more complex. There, each possible outcome value in controls corresponds to a probability mass which in turn corresponds to a probability interval in CDF. When this interval is mapped by the inverse of CDF in the treatment group, the result contains possibly multiple outcome values in the treatment group. To get the exact BQTE, we compute the expectation of the outcome value in the treatment group conditional on the event that corresponds to the particular interval of CDF. The difference between this expectation and the outcome value in the controls is then the exact BQTE.

We apply three estimators of BQTE. The first two are versions of the estimator presented in our main text and the third was defined by Doksum (1974).

1. **Bagging estimator** that equals to the average of the BQTE estimates across the bootstrap samples. This is our default estimator and it is shown in Figures unless otherwise stated in the Figure caption. The motivation for bootstrap averaging ("bagging") is that it can reduce variance of the estimator in some contexts, see Fig. S4.
2. **Estimator without bagging**, where we apply our BQTE estimator to the original data and not to the bootstrapped samples as in bagging. (In `bqte()`, this estimator can be chosen by setting `bagging = FALSE`.)
3. **Doksum estimator** defined as $\widehat{\Delta}(x) = y_{i(x)} - x$ where index $i(x) = \left\lfloor n\,\widehat{F}(x) \right\rfloor + 1$ with $\widehat{F}$ being the empirical cumulative distribution function of the control group, n the sample size of the treatment group and y_i the value in the treatment group having rank i.

2.1 Generating continuous example data

To study the continuous case, let's assume

$$T \sim \text{Weibull}(1.2, 2) \text{ and } C \sim \text{Weibull}(3, 5).$$

Figure S1 demonstrates this setting and shows the estimates of treatment effect from `bqte()` function.

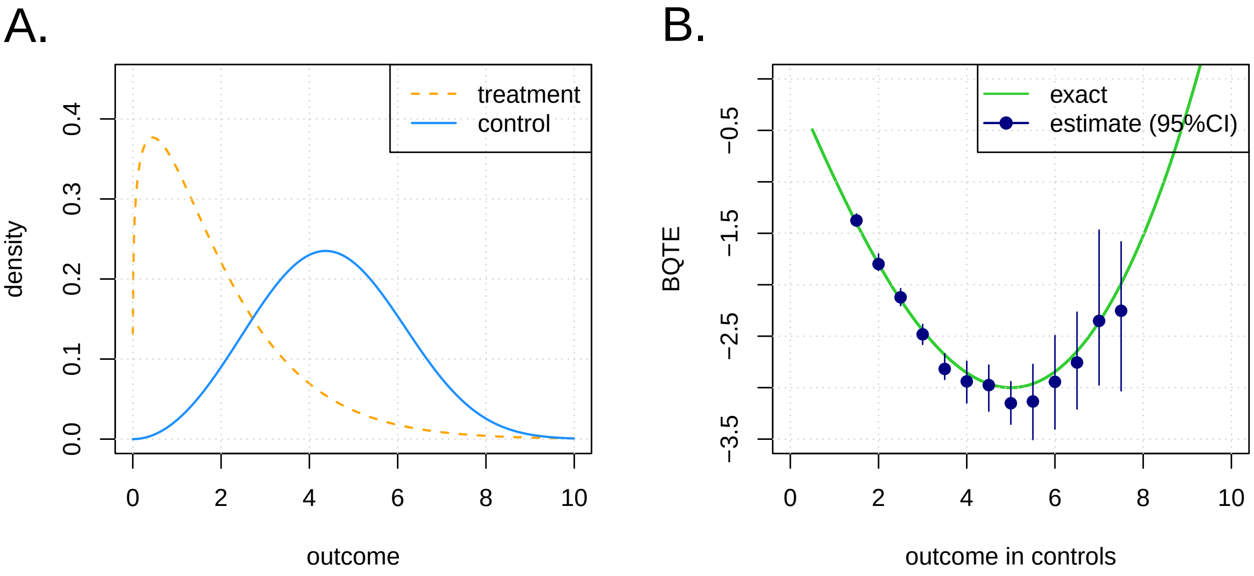

Figure S1. Panel A shows the theoretical distributions in the two groups: Weibull$(1.2, 2)$ in the treatment group and Weibull$(3, 5)$ in the control group. Panel B shows the exact back-transformed quantile treatment effect (BQTE) as a function of the outcome value in controls (green line), together with estimates and their 95% CIs when `bqte()` was applied on a simulated data set with $N = 500$ individuals in each group.

2.2. The range of valid estimates

By default, `bqte()` estimates BQTE in the interval between the empirical quantile levels of $\frac{10}{N}$ and $\frac{N-10}{N}$, where N is the minimum sample size of control and treatment groups. For example, in Fig. S1B, $N = 500$ and hence the range of estimation is between the 2nd and the 98th percentile of the observed data, which correspond approximately to the range from 1.4 to 7.9.

To confirm that the default range for estimation leads to well calibrated confidence intervals (CIs), for values of $N = 50$ and $N = 500$, we generate $R = 100$ data sets as above from the two Weibull distributions and apply `bqte()` function to estimate BQTE at the outcome values that correspond to the following 5 quantile levels in controls: $\frac{5}{N}, \frac{10}{N}, \frac{1}{2}, \frac{N-10}{N}, \frac{N-5}{N}$. For each N and each quantile level, we then compute the empirical coverage of the 95%CIs and the empirical mean of the BQTE estimates over the R data sets. Results are show in Fig S2.

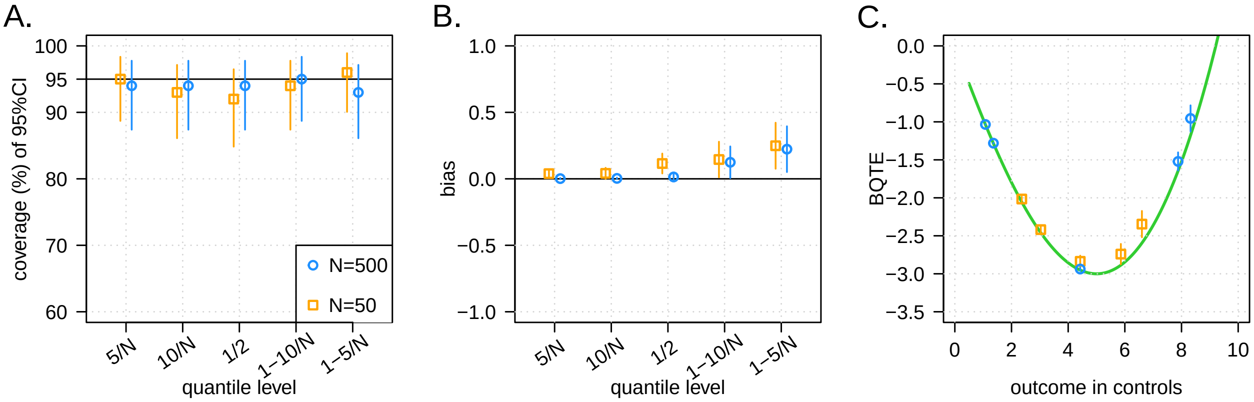

Figure S2. Observed coverage of 95% confidence intervals (panel A), observed bias (panel B) and the mean of the estimates over the 100 replicates compared to the exact values shown by the green line (panel C). The value of N is shown in the legend of panel A. The intervals around each point show 95% confidence intervals based on the sample size of 100 replicates.

In Fig. S2A we see that, independent of sample size N, the 95%CIs are well calibrated for the quantile levels tested. Fig. S2B shows that there is more bias at the right hand tail and the bias does not depend much on the sample size in this example. (Note also that, for any fixed outcome value in controls, the bias would disappear as sample size grows; this is demonstrated by the estimates at the median point in panel B where the bias disappears as N grows from 50 to 500.) From Fig. S2C we observe how the estimates perform with respect to the theoretical values given by the green curve. Based on the results, our default option is to restrict the estimation of BQTE and QTE between quantile levels of $\frac{10}{N}$ and $\frac{N-10}{N}$ but we note that in some other distributions the bias may also be larger and, to confirm reliable results, one should study by simulation the types of distributions that one is working with.

Figure S2 shows only the bagging estimator but we also evaluated the other two estimators on the same simulation setting. The root mean square error (RMSE) of the three estimators at the quantile levels of $10/N$, $1/2$ and $1-10/N$ were,
for the sample size of 50:

```
##   quantile       at RMSE_bagging RMSE_no_bagging RMSE_doksum
## 1     10/N 3.032714    0.2168484       0.2333651   0.2444095
## 2      1/2 4.424985    0.3971765       0.4010331   0.4191274
## 3   1-10/N 5.859512    0.6947341       0.6618659   0.7072629
```

and for the sample size of 500:

```
##   quantile       at RMSE_bagging RMSE_no_bagging RMSE_doksum
## 1     10/N 1.361779   0.03028491      0.03151704  0.03387012
## 2      1/2 4.424985   0.11280326      0.11268059  0.11370297
## 3   1-10/N 7.878384   0.61818896      0.62016916  0.66894113
```

2.3. The number of quantile points used in estimation

The `bqte()` function takes as input the number K of quantile points, corresponding to the probability values of $1/(K+1), 2/(K+1), \ldots, K/(K+1)$, that are used in the estimation of BQTEs. Between these points, BQTE is estimated by linear interpolation. By default, we set $K = N_C$, where N_C is the size of the control group. The value N_C is a natural upper bound for K because when $K = N_C$, there is already a separate estimate corresponding to each control observation. But could some smaller value for $K < N_C$ be a better choice? Let's use a similar simulation procedure as in Fig. S2 to assess how smaller values of $K < N_C$ compare to our default choice $K = N_C$.

For this simulation, the sample size in both groups ($N = N_C = N_T$) takes values $N = 50$ or $N = 500$, and for each value of N, we will run 100 simulations for each value of $K = \frac{N}{4}, \frac{N}{2}, \frac{2N}{3}, N$. The results are in Fig. S3.

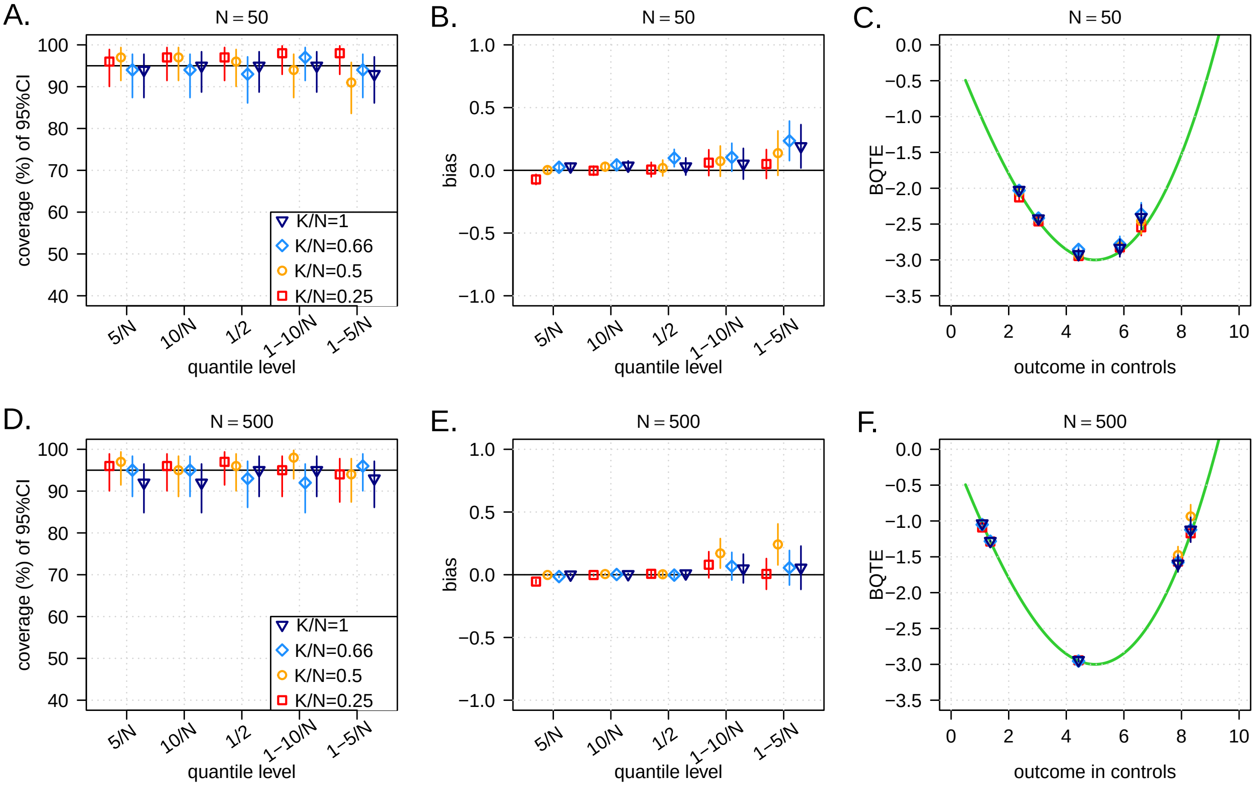

Figure S3. Observed coverage of 95% confidence intervals (panels A, D), observed bias (panels B, E) and the mean of the estimates over the 100 replicates compared to the exact values shown by the green line (panels C, F). The ratio K/N is shown in the legend of panels A and D, and N is shown in the title of each panel. The intervals around each point show 95% confidence intervals based on the sample size of 100 replicates.

In Figure S3, within the recommended quantile level range $10/N$ and $1 - 10/N$, the choice $K = N = N_C$ seems a suitable default value to be used in the `bqte()` function given its small bias throughout these simulations.

2.4. Magnitude of estimation error and coverage of confidence intervals in discrete data

We used the Mossad data (1996) from the main paper to generate realistic data about the duration of cold in a randomized trial when the duration is measured in days. We focus on the recommended interval between quantile levels $10/N$ and $1 - 10/N$ that in these data corresponds to interval (4,15) days. For each integer value in this interval, we computed the exact BQTE values for the observed distribution of the Mossad data using function `exact_bqte_discrete()`. Then we generated 100 simulations from the empirical distribution from Mossad data with sample sizes equaling to the Mossad data (Treatment group = 49, Control group = 50). We estimated the BQTE using four estimators: our `bqte()` function, first with linear interpolation either with boostrap averaging (bagging) or without it (no bagging) and then with the spline interpolation with bagging. The fourth method was the estimator of Doksum. Figure S4 shows the root mean squared error of the four estimators as well as the coverage of the 95% confidence interval (CI) by our bootstrap method with linear interpolation. (Note that the CI is the same with or without bagging.)

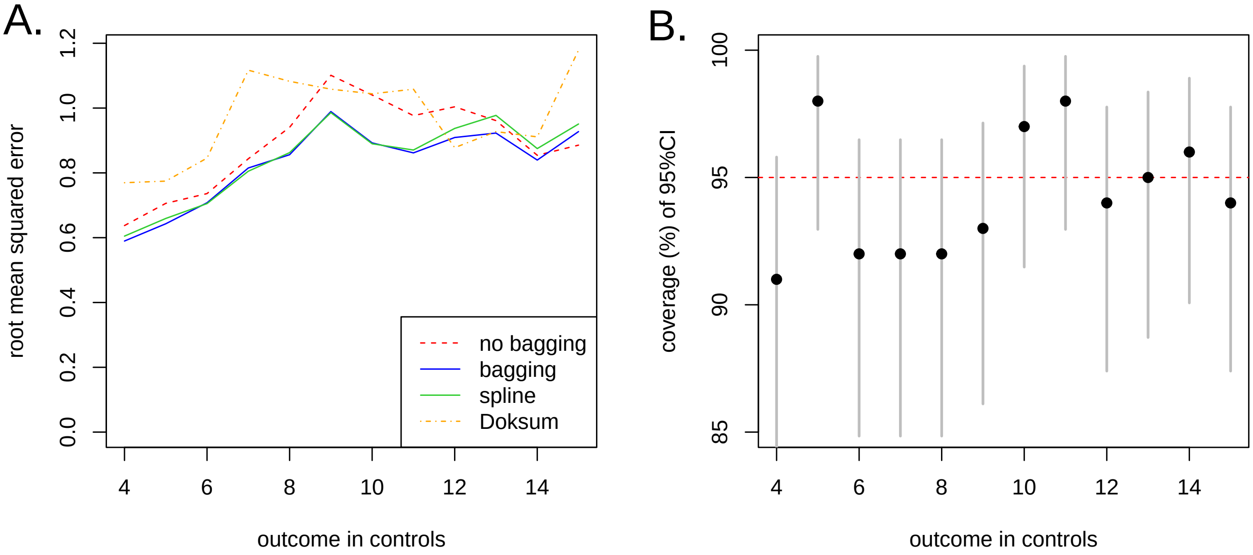

Figure S4. Panel A shows root mean squared error (RMSE) for four estimators (linear interpolation with bagging in blue, and without bagging in red, spline interpolation in green and Doksum estimator in orange) at each of the integer-valued cold duration between 4 and 15 days. The RMSE is estimated over 100 simulated data sets that follow the observed distributions of the Mossad data (1996) from the main text. Panel B shows the empirical coverage of the 95% confidence intervals for the linear interpolation method estimated with bootstrap, at the same grid of values as in panel A. The intervals around the point estimates are 95% binomial confidence intervals based on 100 samples and the red line denotes the target coverage of 95%.

From Figure S4 we see that the RMSE seems the smallest for the bagging estimator with linear interpolation. Based on the result of Figure S4A, we have used the bagging estimator as our default estimator in the discrete data sets that we have analyzed in the main text.

Panel B shows that the 95% confidence intervals do not show evidence of falling short of the target coverage of 95% as the confidence intervals of the coverage estimates in panel B either overlap with the target value or remain above the target value in every case.

2.5 Comparison between BQTE and QTE

We demonstrate the differences between a standard quantile treatment effect (QTE) and our back-transformed quantile treatment effect (BQTE) using the Mossad data from the main text (See section 3 of this supplement for direct link.). The data sets record duration of colds, measured in days, from a randomized, double-blind and placebo controlled zinc gluconate trial, with sample size $N_C = 50$ and $N_T = 49$.

Figure S5 shows the empirical distributions of the treatment group (Panel A) and control group (Panel B). Let's consider the estimate of BQTE at 8 days in the control group and the estimate of QTE at the quantile level of 0.50, which corresponds to 8 days in the control group.

To estimate BQTE(8), we find the quantile levels that correspond to the value 8 in the control distribution, here (0.48,0.58], and identify the corresponding part of the treatment group, which here is distributed between day 5 (with a weight 0.71) and day 6 (weight of 0.29), as highlighted in yellow color in Figure S5A. BQTE(8) is estimated as the difference between the expected values of the highlighted parts of the treatment and control groups. In this example, the numerical estimate is $\widehat{\text{BQTE}}(8) \approx (0.71 \cdot 5 + 0.29 \cdot 6) - 8 = 5.29 - 8 = -2.71$.

To estimate QTE(0.50), we find the level 50% quantile (i.e. the median) from the two groups, which here are 5 days in the treatment group and 8 days in the control group, as marked by red arrows in Figure S5 A and B. Then, $\widehat{\text{QTE}}(0.5) = 5.0 - 8.0 = -3.0$.

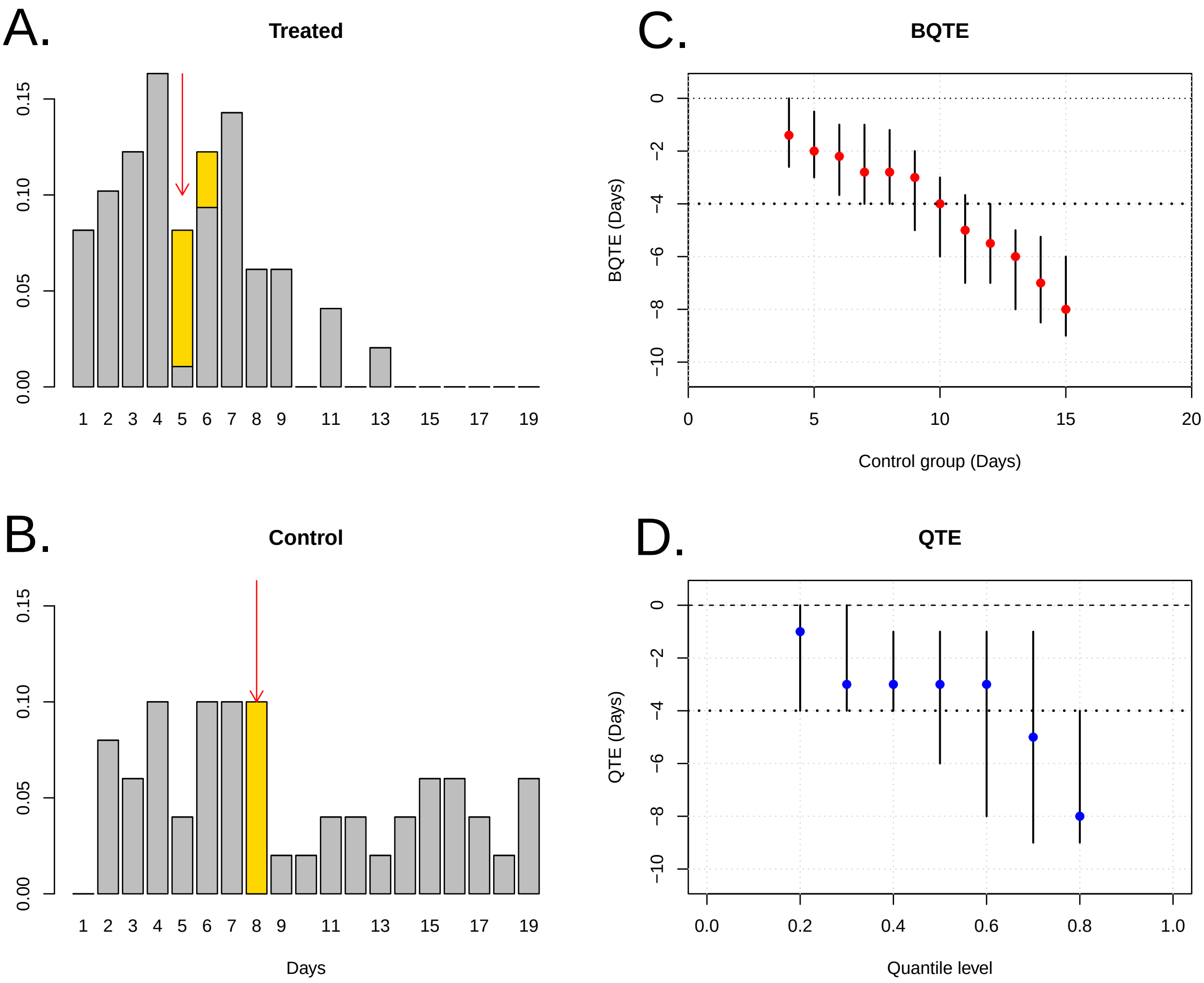

Figure S5. The empirical distributions of the treatment group (Panel A, $N_T = 49$) and control group (Panel B, $N_C = 50$) of the Mossad data. The medians are denoted by red arrows. The probability mass corresponding to the control outcome value 8 is highlighted in yellow color. Panel C shows the BQTE estimates for days between 4 and 15 using the `bqte()` function. Panel D shows the QTE estimates between quantile levels 20% and 80% using the `qte()` function. The 95% confidence intervals are estimated by the bootstrap. In panels C and D, we have not used bagging to make the point estimate match exactly to the values explained in the text above, i.e, $\widehat{\text{BQTE}}(8) = -2.71$ and $\widehat{\text{QTE}}(0.5) = -3.0$. (Consequently, these point estimates differ slightly from the main text Figure 2.)

We can make the following remarks:

- The BQTE function is represented entirely in terms of the outcome value, which, in an applied context, is often simpler to interpret than a quantile level. For example, above the BQTE function tells that for the control value of 8 days, the comparison point in the treatment group is 5.29 days (= 8 - 2.71), whereas the QTE function tells only that, at the median, the difference between the two groups is -3 days.
- Here, the confidence intervals are wider for QTE than for BQTE. This is typical at those quantiles where the treated population is more heavily concentrated than the control population. In the opposite

situation, the relationship reverses. (See variance formulas below.)

- For discrete data, the BQTE value is defined as the difference between the expected values corresponding to the certain range of quantile levels, and this averaging can further reduce variance compared to an estimator that did not average over a quantile range.

2.6 Variance of BQTE and QTE

Let's consider a continous outcome distribution for which we estimate QTE at a fixed quantile level p and BQTE at the corresponding fixed outcome value $x = F_C^{-1}(p)$. Additionally, denote $y = F_T^{-1}(p)$. Thus, $\text{BQTE}(x) = \text{QTE}(p) = y - x$.

The asymptotic distribution of an empirical quantile $\widehat{F^{-1}}(p)$ is known to be $\mathcal{N}\left(F^{-1}(p), \frac{p(1-p)}{N(f(F^{-1}(p)))^2}\right)$, where F is the cumulative distribution function, f the density function and N the sample size. When the control group has size N_C and the treatment group has size N_T,

$$\begin{aligned}
\text{Var}\left(\widehat{\text{QTE}}(p)\right) &= \text{Var}\left(\widehat{F_T^{-1}}(p) - \widehat{F_C^{-1}}(p)\right) = \text{Var}\left(\widehat{F_T^{-1}}(p)\right) + \text{Var}\left(\widehat{F_C^{-1}}(p)\right) \\
&\approx \frac{p(1-p)}{N_T\,(f_T(y))^2} + \frac{p(1-p)}{N_C\,(f_C(x))^2}, \text{ for large } N_C \text{ and } N_T.
\end{aligned}$$

As asymptotic distribution of $\widehat{F_C}(x)$ is $\mathcal{N}\left(p, \frac{p(1-p)}{N_C}\right)$, by using the delta method, we can derive that

$$\text{Var}\left(F_T^{-1}\left(\widehat{F_C}(x)\right)\right) \approx \frac{p(1-p)}{N_C\,(f_T(F_T^{-1}(p)))^2} = \frac{p(1-p)}{N_C\,(f_T(y))^2}.$$

By using the law of total variance,

$$\begin{aligned}
\text{Var}\left(\widehat{\text{BQTE}}(x)\right) &= \text{Var}\left(\widehat{F_T^{-1}}\left(\widehat{F_C}(x)\right) - x\right) = \text{Var}\left(\widehat{F_T^{-1}}\left(\widehat{F_C}(x)\right)\right) \\
&= \text{Var}\left(\text{E}\left(\widehat{F_T^{-1}}\left(\widehat{F_C}(x)\right) \mid \widehat{F_C}(x)\right)\right) + \text{E}\left(\text{Var}\left(\widehat{F_T^{-1}}\left(\widehat{F_C}(x)\right) \mid \widehat{F_C}(x)\right)\right) \\
&\approx \text{Var}\left(F_T^{-1}\left(\widehat{F_C}(x)\right)\right) + \text{E}\left(\frac{\widehat{F_C}(x)\left(1-\widehat{F_C}(x)\right)}{N_T\left(f_T\left(F_T^{-1}\left(\widehat{F_C}(x)\right)\right)\right)^2}\right) \\
&\approx \frac{p(1-p)}{N_C\left(f_T\left(F_T^{-1}(p)\right)\right)^2} + \frac{p(1-p)}{N_T\left(f_T\left(F_T^{-1}(p)\right)\right)^2} \\
&= \frac{p(1-p)}{\left(f_T\,(y)\right)^2}\left(\frac{1}{N_C} + \frac{1}{N_T}\right),
\end{aligned}$$

where, in the approximation on the second last line, the first term was derived above and the second term results by replacing $\widehat{F_C}(x)$ by its expected value p.

Hence, in large samples,

$$\text{Var}\left(\widehat{\text{QTE}}(p)\right) - \text{Var}\left(\widehat{\text{BQTE}}(x)\right) \approx \frac{p(1-p)}{N_C}\left(\frac{1}{\left(f_C\,(x)\right)^2} - \frac{1}{\left(f_T\,(y)\right)^2}\right).$$

Thus, asymptotically, $\text{Var}\left(\widehat{\text{QTE}}(p)\right) > \text{Var}\left(\widehat{\text{BQTE}}(x)\right)$ when $f_T\,(y) > f_C\,(x)$, which happens when the treatment values have a higher concentration near the quantile value y than the control values have at the corresponding quantile value x. On the other hand, when $f_C\,(x) > f_T\,(y)$, then the estimator of BQTE is expected to have a larger variance than the estimator of QTE.

To confirm the calculations, Table S1 below reports a comparison between the above formulas and empirical variances of the two estimators.

```
##     bqte_var_thr bqte_var_emp qte_var_thr qte_var_emp
## 0.2   0.01717384   0.01729326  0.05703340  0.05708496
## 0.5   0.04708114   0.04826083  0.06097619  0.06061872
## 0.8   0.14221891   0.14700535  0.10631669  0.10494609
```

Table S1. Simulated data for $N_C = 100$ controls from Weibull$(3, 5)$ and $N_T = 200$ treated individuals from Weibull$(1.2, 2)$. (Densities of these distributions are shown in Figure S1.) For quantile levels 0.2, 0.5 and 0.8, theoretical approximations to the variances Var $\left(\widehat{\mathrm{BQTE}}(F_T^{-1}(F_C(p)))\right)$ and Var $\left(\widehat{\mathrm{QTE}}(p)\right)$ were computed using the formulas above and are shown in columns `bqte_var_thr` and `qte_var_thr`, respectively. Then, 100,000 datasets were generated and BQTE and QTE were estimated at each of the three quantile levels. The empirical variances of the estimates are in columns `bqte_var_emp` and `qte_var_emp`, respectively.

From Table S1 we see that the theoretical variance approximations capture the variance of the estimators of BQTE and QTE reasonably accurately. In this example, variance of $\widehat{\mathrm{BQTE}}$ is smaller than variance of $\widehat{\mathrm{QTE}}$ at the quantile levels 0.2 and 0.5 whereas the opposite is true at the quantile level 0.8.

3. Source of data of the included examples

Mossad study. The data are from pages 2-4 of the Additional file 2 of https://doi.org/10.1186/s12874-017-0356-y.

Zinc acetate studies. The data are from page 8 of the Supplementary file 2 of https://doi.org/10.1093/ofid/ofx059

Carrageenan studies. The data are from page 3 of the Supplementary file of https://doi.org/10.1002/prp2.810

The data files used in this study are available at GitHub: https://github.com/mjpirinen/bqte/

4. Influence of missing data in the Carrageenan data sets

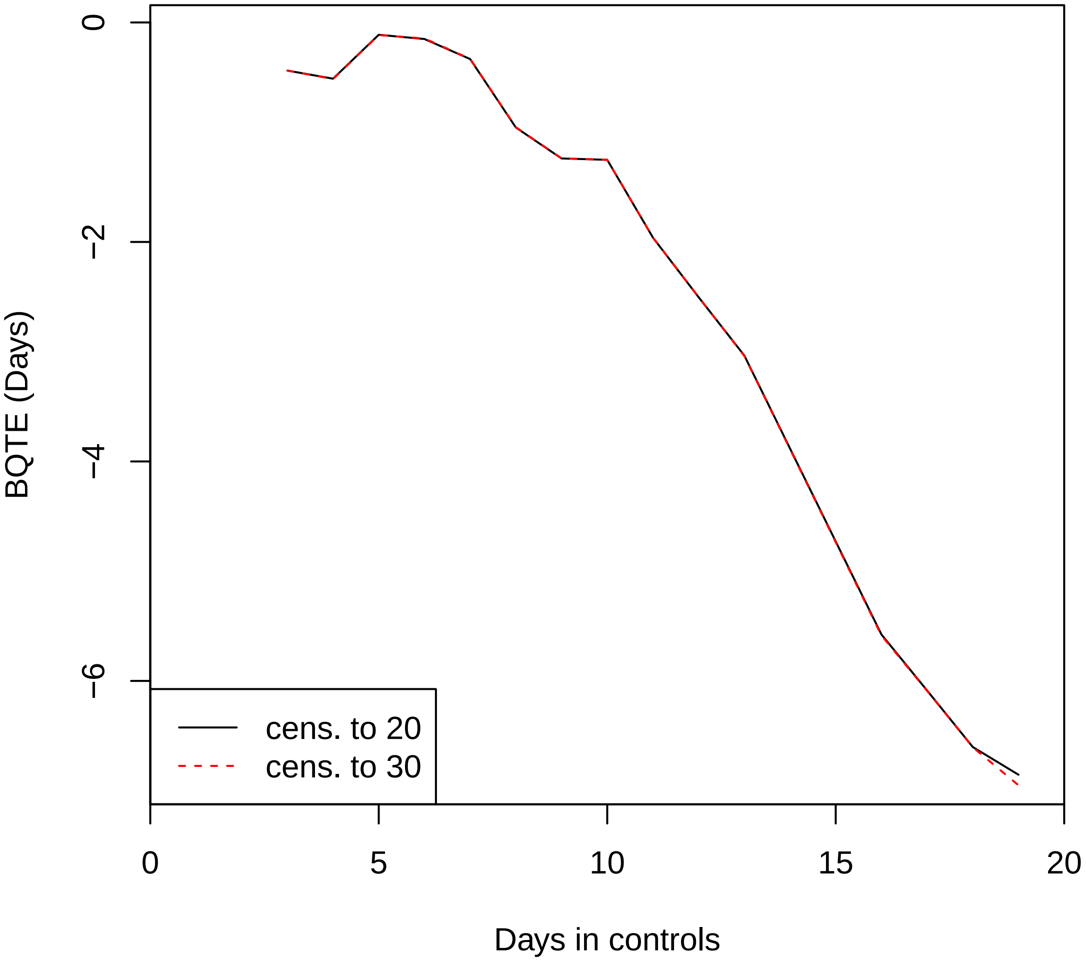

Figure S6. Effect of imputation of the censored observations in the Carrageenan data set. The continuous black line shows the BQTE function when the censored observations were imputed using the value of 20 days. The dashed red line shows the BQTE function when the censored observations were imputed using the value of 30 days. In this calculation, we used B = 20,000 bootstrap samples to ensure stable estimates. There is only a very small difference between the curves at the right-hand side, indicating that the curves are robust to the type of imputation applied in this case.