Supplementary Material for Contextual Constrained Learning for Dose-Finding Clinical Trials

A PRELIMINARIES

Before providing the proof of the theorems, we introduce some regularity assumptions on the dose-toxicity model as follows.

Assumption 1. *There exist* $C_{1,s,k} > 0$, $1 < \gamma_{1,s,k}$, $C_{2,s,k}$, *and* $0 < \gamma_{2,s,k} \leq 1$ *such that* $|p_{s,k}(a) - p_{s,k}(a')| \geq C_{1,s,k}|a - a'|^{\gamma_{1,s,k}}$ *and* $|p_{s,k}(a) - p_{s,k}(a')| \leq C_{2,s,k}|a - a'|^{\gamma_{2,s,k}}$ *for all* $s \in \mathcal{S}$, $k \in \mathcal{K}$, *and* a, $a' \in \mathcal{A}$.

We then immediately have the following proposition.

Proposition 1. *For* $p_{s,k}(a)$, $\forall s \in \mathcal{S}, \forall k \in \mathcal{K}$ *satisfying Assumption 1,*

1. $p_{s,k}(a)$ *is invertible;*

2. For each $s \in \mathcal{S}$, $k \in \mathcal{K}$, *and* d, $d' \in \mathcal{P}$, *we have* $|p^{-1}_{s,k}(d) - p^{-1}_{s,k}(d')| \leq \bar{C}_{1,s,k}|d - d'|^{\bar{\gamma}_{1,s,k}}$, *where* $\bar{\gamma_{1,s,k}} = \frac{1}{\gamma_{1,s,k}}$ *and* $\bar{C}_{1,s,k} = C_{1,s,k}^{-\frac{1}{\gamma_{1,s,k}}}$.

For notational simplicity, we denote $C_{1,s} = \min_{k\in\mathcal{K}} C_{1,s,k}$, $C_{2,s} = \max_{k\in\mathcal{K}} C_{2,s,k}$, $\gamma_{1,s} = \max_{k\in\mathcal{K}} \gamma_{1,s,k}$, $\gamma_{2,s} = \min_{k\in\mathcal{K}} \gamma_{2,s,k}$, $\bar{\gamma}_{1,s} = \gamma_{1,s}^{-1}$, and $\bar{C}_{1,s} = C_{1,s}^{-\bar{\gamma}_{1,s}}$.

B PROOF OF THEOREM 1

From the Hoeffding bound, the following upper bound of the probability is given:

$$\mathbb{P}[\hat{a}(t) - a_s^* > \alpha_s(t) + \epsilon] \leq \exp(-2N_s(t)(\alpha_s(t) + \epsilon)^2).$$

In addition, the difference between the MTD threshold and the expected toxicity is also bounded as

$$\begin{aligned} p_{s,I(t)}(a_s^*) - \zeta &\leq p_{s,I(t)}(a_s^*) - \zeta + \zeta - p_{s,I(t)}(a_s^* - \alpha_s(t)) \\ &\leq C_{2,s}|a_s^* - \hat{a}_s(t) + \alpha_s(t)|^{\gamma_{2,s}}. \end{aligned}$$

By rearranging the terms, we have

$$\begin{aligned} \mathbb{P}\left[\frac{\sum_{t=1}^{N_s(T)} p_{n,I(N_s^{-1}(t))}(a_s^*)}{N_s(T)} - \zeta < C_{n,2}\epsilon^{\gamma_{n,2}}\right] &\geq 1 - \exp(-2N_s(T)(\alpha_s(N_s(T)) + \epsilon)^2) \\ &\geq 1 - \delta_s. \end{aligned}$$

C PROOF OF THEOREM 2

C.1 Case 1: $k_s^* \neq 0$

We first bound the probability that the recommended dose error for subgroup s occurs with C3T-Budget if $k_s^* \neq 0$. The event that the recommended dose error occurs satisfies the following:

$$\left\{\hat{k}_s^* \neq k_s^*\right\} \subseteq \left\{p_{s,k_s^*}(\hat{a}_s(T)) > \zeta\right\} \cup \left\{\bar{q}_{s,k_s^*}(T) < \theta\right\} \cup \left\{\max_{k\in\mathcal{K}_s} \bar{q}_{s,k}(T) \neq k_s^*\right\}.$$

Thus, the probability of the recommended dose error for subgroup s can be bounded as

$$\mathbb{P}\left[\hat{k}_s^* \neq k_s^*\right] \leq \mathbb{P}\left[p_{s,k_s^*}(\hat{a}_s(T)) > \zeta\right] + \mathbb{P}\left[\bar{q}_{s,k_s^*}(T) < \theta\right] + \mathbb{P}\left[\max_{k\in\mathcal{K}_s} \bar{q}_{s,k}(T) \neq k_s^*\right].$$

Then, we can bound the probability by obtaining the bound for each term.

Bound of First Term: The probability in the first term can be transformed and bounded as

$$\begin{aligned}
\mathbb{P}\left[\hat{a}_s(N_s(T)) < p^{-1}_{s,k^*_s}(\zeta)\right] &\leq \mathbb{P}\left[|a^*_s - \hat{a}_s(N_s(T))| > \Gamma_{U_s}\right] \\
&\leq \sum_{k\in\mathcal{K}} \mathbb{P}\left[\left|\hat{p}_{s,k}(N_s(T)) - p_{s,k}(a^*_s)\right| > \left(\frac{\Gamma_{U_s} N_s(T)}{N_{s,k}(T)\bar{C}_{1,s}K}\right)^{\gamma_{1,s}}\right] \\
&\leq \sum_{k\in\mathcal{K}} 2\exp\left(-2N_{s,k}(T)\left(\frac{\Gamma_{U_s} N_s(T)}{N_{s,k}(T)\bar{C}_{1,s}K}\right)^{\gamma_{1,s}}\right) \qquad (1) \\
&\leq 2K\exp\left(-2\left(\frac{\Gamma_{U_s}}{\bar{C}_{1,s}K}\right)^{\gamma_{1,s}} N_s(T)\right), \qquad (2)
\end{aligned}$$

where $\Gamma_{U_s} = |a^*_s - p^{-1}_{s,k}(\zeta)|$. The inequality in (1) follows from the Hoeffding's inequality and the inequality in (2) follows from the regularity assumption $\gamma_{1,s} > 1$.

Bound of Second Term: From the Chernoff-Hoeffding's inequality, we have

$$\begin{aligned}
\mathbb{P}\left[\bar{q}_{s,k^*_s}(T) < \theta\right] &= \mathbb{P}\left[\bar{q}_{s,k^*_s}(T) < q_{s,k^*_s} - (q_{s,k^*_s} - \theta)\right] \\
&\leq \exp\left(-2N_{s,k^*_s}(T)\Delta^2_{s,\theta}\right) \\
&\leq \exp\left(-\frac{8}{25}\frac{cN_s(T)\Delta^2_{s,\theta}}{\Delta^2_{s^*}}\right), \qquad (3)
\end{aligned}$$

where $\Delta_{s^*,\theta} = |q_{s,k^*_s} - \theta|$ and

$$\Delta_{s^*} = \begin{cases} 9(\min_{k\in\mathcal{K}, k\neq k^*_s} q_{s,k^*_s} - q_{s,k}), & \text{if } k^*_s = \max_{k\in\mathcal{K}}\{q_{s,k}\}, \\ \max_{k\in\mathcal{K}}\{q_{s,k}\} - q_{s,k^*_s}, & \text{otherwise.} \end{cases}$$

The inequality in (3) follows from the lemmas in (Audibert et al., 2010).

Bound of Third Term: The bound of the third term can be obtained by following the proof of Theorem 1 in (Audibert et al., 2010). The third term is bounded as

$$\mathbb{P}\left[\max_{k\in\mathcal{K}_s} \bar{q}_{s,k} \neq k^*_s\right] \leq 2N_s(T)K\exp\left(-\frac{2cN_s(T)}{25}\right). \qquad (4)$$

Total Bound: From the bounds in (2), (3), and (4), we can bound the probability of the recommended dose error for subgroup s as

$$\mathbb{P}\left[\hat{k}^*_s \neq k^*_s\right] \leq \exp\left(-M_{(1a)}N_s(T)\right) + 2K\left(\exp\left(-M_{(1b)}N_s(T)\right) + N_s(T)\exp\left(-M_{(1c)}N_s(T)\right)\right),$$

where $M_{(1a)} = \frac{8}{25}\frac{c\Delta^2_{s,\theta}}{\Delta^2_{s^*}}$, $M_{(1b)} = 2\left(\frac{\Gamma_{U_s}}{\bar{C}_{1,s}K}\right)^{\gamma_{1,s}}$, and $M_{(1c)} = \frac{2c}{25}$.

C.2 Case 2: $k^*_s = 0$

For the case with $k^*_s = 0$, we can bound the probability of the recommended dose error for subgroup s similar to Case 1. We have

$$\left\{\hat{k}^*_s \neq 0\right\} \subseteq \bigcup_{k\in\mathcal{K}} \left\{p_{s,k}(\hat{a}_s(T)) \leq \zeta\right\} \cup \bigcup_{k\in\mathcal{K}} \left\{\bar{q}_{s,k}(T) \geq \theta\right\}$$

and

$$\mathbb{P}\left[\hat{k}^*_s \neq 0\right] \leq \sum_{k\in\mathcal{K}} \mathbb{P}\left[p_{s,k}(\hat{a}_s(T)) \leq \zeta\right] + \sum_{k\in\mathcal{K}} \mathbb{P}\left[\bar{q}_{s,k}(T) \geq \theta\right].$$

Similar to Case 1, we can bound the probability as

$$\mathbb{P}\left[\hat{k}^*_s \neq 0\right] \leq K\exp\left(-M_{(2a)}N_s(T)\right) + 2K^2\exp\left(-M_{(2b)}N_s(T)\right),$$

where $M_{(1a)} = \frac{8}{25}\frac{c\Delta^2_{s^*\theta}}{\underline{\Delta}^2_{s^*}}$, $M_{(1b)} = 2\left(\frac{\bar{\Gamma}_{U_s}}{\bar{C}_{1,s}K}\right)^{\gamma_{1,s}}$, $\bar{\Delta}_{s^*,\theta} = \max_{k\in\mathcal{K}}|\theta - q_{s,k}|$, $\underline{\Delta}_{s^*} = \max_{k\in\mathcal{K}}\{q_{s,k}\} - \min_{k\in\mathcal{K}}\{q_{s,k}\}$, and $\bar{\Gamma}_{U_s} = \max_{k\in\mathcal{K}}\Gamma_{U_s}$.

C.3 Recommended Dose Error Bound

Finally, we have the theorem with $M_{R1} = \max\{1 + 2K + N_s(T), K + 2K^2\}$ and $M_{R2} = \min\{M_{(1a)}, M_{(1b)}, M_{(1c)}, M_{(2a)}, M_{(2b)}\}$.

D Worst-Case Reget Bound For Total Efficacy of C3T-Budget-E

To evaluate C3T-Budget-E, we compare its performance to that of an algorithm with the complete knowledge of $q_{s,k}$'s and $p_{s,k}$'s called an oracle algorithm. We denote the expected total cumulative efficacy achieved by the oracle algorithm by $E^*(T,B)$. Then, the regret of C3T-Budget-E is defined as

$$R(T,B) = E^*(T,B) - E(T,B), \tag{5}$$

where $E(T,B)$ is the expected total cumulative efficacy. Then, we provide the efficacy regret bound of C3T-Budget.

Theorem 1. *Given a fixed $\rho \in (0,1)$, the worst-case regret of C3T-Budget-E is bounded as*

$$R(T,B) \le T\bar{\delta}\bar{\Delta} + q_1^*\sqrt{\rho(1-\rho)T} + M_E \log T + O(1),$$

where M_E is a non-negative constant (provided in our supplementary material).

Proof. For the regret bound of C3T-Budget-E, we first define the optimal value of the LP problem that can be obtained by solving the LP problem with q_s^*'s (see (3) in our main paper) as

$$v(\rho) = \sum_{s=1}^{\tilde{s}(\rho)} \pi_s d_s^* + \psi_{\tilde{s}(\rho)+1}(\rho)\pi_{\tilde{s}(\rho)+1} d^*_{\tilde{s}(\rho)+1}.$$

This optimal value $v(\rho)$ can be considered the maximum expected reward in a single round with average budget ρ. Thus, using $v(\rho)$, we can bound the total expected cumulative efficacy of the oracle $E^*(T,B)$ as the following lemma.

Lemma 1. *(Wu et al., 2015) If the time-horizon and budget are given by T and B, respectively, then we have $\hat{E}(T,B) = Tv(\rho) \ge E^*(T,B)$.*

Then, the upper bound of the expected cumulative efficacy of the oracle $E^*(T,B)$ as follows.

$$\begin{aligned} R(T,B) &= E^*(T,B) - E(T,B) \\ &\le \hat{E}(T,B) - E(T,B) \\ &= Tv(\rho) - \sum_{s\in\mathcal{S}}\sum_{k\in\mathcal{K}} q_{s,k}\mathbb{E}[N_{s,k}(T)]. \end{aligned}$$

From the regret using $\hat{E}(T,B)$, we can partition the regret according to the source of regret as follows:

$$\begin{aligned} R(T,B) &= Tv(\rho) - \sum_{s\in\mathcal{S}}\sum_{k\in\mathcal{K}} q_{s,k}\mathbb{E}[N_{s,k}(T)] \\ &= \sum_{s\in\mathcal{S}}\sum_{k\in\mathcal{K}} \Delta^{(s)}_{s,k}\mathbb{E}[N_{s,k}(T)] + Tv(\rho) - \sum_{s\in\mathcal{S}} q_s^*\mathbb{E}[N_s(T)] \\ &= \underbrace{\sum_{s\in\mathcal{S}}\sum_{k\in\mathcal{K}} \Delta^{(s)}_{s,k}\mathbb{E}[N_{s,k}(T)]}_{=R^{(1)}(T,B)} + \underbrace{\sum_{t=1}^{T}\mathbb{E}\left[v(\rho) - \sum_{s\in\mathcal{S}}\hat{\psi}_s(\rho(t))\pi_s q_s^*\right]}_{=R^{(2)}(T,B)}. \end{aligned}$$

Recall that $\Delta^{(s')}_{s,k}$ is the difference between the optimal expected efficacy of subgroup s' and the expected efficacy of subgroup s with dose k, $\Delta^{(s')}_{s,k} = q_{s'} - q_{s,k}$. The decomposed regret $R^{(1)}$ represents the regret due to taking suboptimal doses and the other decomposed regret $R^{(2)}$ represents the regret due to ordering errors in subgroups. It is worth noting that in $R^{(2)}$, it is supposed that the optimal doses are chosen. Finally, the regret of C3T-Budget-E is bounded as

$$R(T,B) \le R^{(1)}(T,B) + R^{(2)}(T,B).$$

Then, we can bound the regret of C3T-Budget by obtaining the bound of each regret.

D.1 Bound of $R^{(1)}$

We first bound the first part of the regret, $R^{(1)}$. In C3T-Budget-E, the set of candidate recommended doses is constructed in each round and the dose is chosen among the doses in the set. Thus, taking the suboptimal doses can occur due to not only the inaccurate estimation of the efficacy but also the inaccurate estimation of the toxicity. To reflect this, we decompose the regret $R^{(1)}$ into two parts according to whether the optimal dose is included in the set of the candidate doses or not as follows.

$$R^{(1)}(T,B) = \underbrace{\sum_{t=1}^{T}\sum_{s\in\mathcal{S}} \mathbb{I}\{H(t)=s\}\mathbb{P}[k_s^* \notin \mathcal{K}_s(t)]\bar{\Delta}_s}_{=R^{(1a)}(T,B)} + \underbrace{\sum_{t=1}^{T}\sum_{s\in\mathcal{S}} \mathbb{I}\{H(t)=s\}\mathbb{P}[k_s^* \in \mathcal{K}_s(t)]R_{s,2}(t)}_{=R^{(1b)}(T,B)},$$

where $\bar{\Delta}_s = \max_{k\in\mathcal{K}} \Delta_{s,k}^{(s)}$.

Bound of $R^{(1a)}(T,B)$: We bound the regret $R^{(1a)}$. Since the event $\{k_s^* \notin \mathcal{K}_s(t)\}$ can be bounded by $\{p_{s,k_s^*}(\hat{a}_s(t) - \alpha_s(t)) > \zeta\} \cup \{\hat{q}_{s,k_s^*}(t) < \theta\}$, we can bound the regret $R^{(1a)}$ as

$$\begin{aligned}
R^{(1a)}(T,B) &\le \sum_{t=1}^{T}\sum_{s\in\mathcal{S}} \mathbb{I}\{H(t)=s\}\left(\mathbb{P}[p_{s,k_s^*}(\hat{a}_s(t)-\alpha_s(t)) > \zeta] + \mathbb{P}[\hat{q}_{s,k_s^*}(t) < \theta]\right)\bar{\Delta}_s \\
&= \underbrace{\sum_{t=1}^{T}\sum_{s\in\mathcal{S}} \mathbb{I}\{H(t)=s\}\mathbb{P}[p_{s,k_s^*}(\hat{a}_s(t)-\alpha_s(t)) > \zeta]\bar{\Delta}_s}_{=R^{(1a-1)}(T,B)} + \underbrace{\sum_{t=1}^{T}\sum_{s\in\mathcal{S}} \mathbb{I}\{H(t)=s\}\mathbb{P}[\hat{q}_{s,k_s^*}(t) < \theta]\bar{\Delta}_s}_{=R^{(1a-2)}(T,B)}.
\end{aligned}$$

We first bound $R^{(1a-1)}(T,B)$. In the following lemma, we show that the safe dose for each subgroup (i.e., the toxicities of the dose levels are below the MTD threshold) are included in the set of the candidate doses with high probability.

Lemma 2. *For each subgroup s, $\mathbb{P}[p_{s,k}(\hat{a}_s(t) + \alpha_s(t)) > \zeta] \le \delta_s$, for any $p_{s,k}(a_s^*) \le \zeta$.*

Proof. We have

$$\begin{aligned}
\mathbb{P}[\hat{a}_s(t) + \alpha_s(t) < a_s^*] &= \mathbb{P}[a_s^* - \hat{a}_s(t) > \alpha_s(t)] \\
&\le \sum_{k\in\mathcal{K}} \mathbb{P}\left[|\hat{p}_{s,k}(t) - p_{s,k}(a_s^*)| > \left(\frac{\alpha_s(t)N_s(t)}{N_{s,k}(t)\bar{C}_{s,1}K}\right)^{\gamma_{s,1}}\right] \\
&\le \sum_{k\in\mathcal{K}} 2\exp\left(-2N_{s,k}(t)\left(\frac{\alpha_s(t)N_s(t)}{N_{s,k}(t)\bar{C}_{s,1}K}\right)^{2\gamma_{s,1}}\right) \qquad (6) \\
&\le 2K\exp\left(-\left(\frac{\alpha_s(t)}{\bar{C}_{s,1}K}\right)^{2\gamma_{s,1}} N_s(t)\right) = \delta_s
\end{aligned}$$

The inequality in (6) follows from the Hoeffding's inequality. □

From this lemma, the probability that the event $\{p_{s,k_s^*}(\hat{a}_s(t) - \alpha_s(t)) > \zeta\}$ occurs is bounded by δ_s since the set of the candidate doses for subgroup s is constructed by $\{k \in \mathcal{K} : p_{s,k}(\hat{a}_s(t) + \alpha_s(t)) \le \zeta\}$ in C3T-Budget-E. Then, the regret $R^{(1a-1)}$ can be simply bounded as

$$\begin{aligned}
R^{(1a-1)}(T,B) &\le \sum_{s\in\mathcal{S}} N_s(T)\delta_s\bar{\Delta}_s \\
&\le T\bar{\delta}\bar{\Delta},
\end{aligned}$$

where $\bar{\delta} = \max_{s\in\mathcal{S}} \delta_s$ and $\bar{\Delta} = \max_{s\in\mathcal{S}} \bar{\Delta}_s$.

We bound the regret $R^{(1a-2)}(T,B)$. For the minimum efficacy threshold, we have the following lemma.

Lemma 3. *Let For each subgroup s,* $\mathbb{P}[\hat{q}_{s,k_s^*}(t) < \theta] \le N_s(t)^{-2c}$.

Proof. We have

$$\begin{aligned}\mathbb{P}[\hat{q}_{s,k_s^*}(t) < \theta] &\le \mathbb{P}\left[\bar{q}_{s,k_s^*}(t) < q_{s,k_s^*} - \sqrt{\frac{c\log N_s(t)}{N_{s,k_s^*}(t)}}\right] \\ &\le N_s(t)^{-2c}\end{aligned}$$

The first inequality follows from the fact $q_{s,k_s^*} \ge \theta$ and the second inequality follows from the Chernoff-Hoeffding inequality. □

Then, for $c \ge \frac{1}{2}$, the regret $R^{(1a-2)}(T,B)$ is bounded as

$$\begin{aligned}\sum_{t=1}^{T}\sum_{s\in\mathcal{S}} \mathbb{I}\{H(t)=s\}\mathbb{P}[\hat{q}_{s,k_s^*}(t) < \theta]\bar{\Delta}_s &\le \sum_{t=1}^{T}\sum_{s\in\mathcal{S}} \mathbb{I}\{H(t)=s\}N_s(t)^{-2c}\bar{\Delta}_s \\ &\le S\sum_{t=1}^{T} t^{-2c}\bar{\Delta} \\ &\le S\log(T)\bar{\Delta} \qquad (7)\end{aligned}$$

In summary, we have the bound of $R^{(1a)}(T,B)$ as following.

$$R^{(1a)}(T,B) \le (T\bar{\delta} + S\log T)\bar{\Delta}$$

Bound of $R^{(1b)}(T,B)$:We now bound the regret $R^{(1b)}$. In this case, the optimal dose is included in the set of the candidate doses. Then, the error occurs when the suboptimal doses are chosen in the set due to the inaccurate parameter estimation $\hat{a}_s$ and the inaccurate efficacy estimation. To bound this, in the following lemma, we show that the unsafe dose for each subgroup (i.e., the toxicities of the dose levels exceed the MTD threshold) are excluded from the set of the candidate doses with high probability.

Lemma 4. *Let* $\Gamma_s = \min_{k\in\mathcal{K}} |a_s^* - p_{s,k}^{-1}(\zeta)|$ *and*

$$\alpha_s(t) = \bar{C}_{s,1}K\left(\frac{\log\frac{2K}{\delta_s}}{2N_s(t)}\right)^{\frac{\bar{\gamma}_{s,1}}{2}}.$$

For each subgroup s, if $N_s(t) \ge t_s^{(1b)} = \frac{1}{2}\left(\frac{\bar{C}_{s,1}K}{\Gamma_s-\epsilon}\right)^{2\gamma_{s,1}}\log\frac{2K}{\delta_s}$, *then we have* $\mathbb{P}[p_{s,k}(\hat{a}_s(t)+\alpha_s(t)) \le \zeta] \le \exp(-2N_s(t)\epsilon^2)$, *for any* $p_{s,k}(a_s^*) > \zeta$.

Proof. From the Hoeffding's inequality, we have

$$\alpha_s(t) \le p_{s,k}^{-1}(\zeta) - a_s^* - \epsilon = \Delta_{s,k} - \epsilon,$$

where $\Delta_{s,k} = |a_s^* - p_{s,k}^{-1}(\zeta)|$. With the definition of $\alpha_s(t)$, we can conclude the lemma. □

Let $N_s^{-1}(\tau) = \min_t\{t : N_s(t) = \tau\}$ which represents the round index at which the τ-th patient of subgroup s arrives. Then, we can bound the regret $R^{(1b)}$ as

$$\begin{aligned}R^{(1b)} &\le \sum_{s\in\mathcal{S}}\left[t_s^{(1b)} + (K-U_s)\sum_{t=1}^{N_s(T)}\exp(-2t\epsilon^2) + \sum_{t=t_s+1}^{N_s(T)}\sum_{k:p_{s,k}(a_s^*)\le\zeta}\mathbb{I}\{I(N_s^{-1}(t))=k\}\right] \\ &\le \sum_{s\in\mathcal{S}}\left[t_s^{(1b)} + \frac{K-U_s}{2\epsilon^2} + \sum_{k:p_{s,k}(a_s^*)\le\zeta}\frac{c\log T}{\Delta_{s,k}^{(s)}}\right] \\ &\le \bar{t}^{(1b)} + \frac{K-\underline{U}}{2\epsilon^2} + \sum_{s\in\mathcal{S}}\sum_{k:p_{s,k}(a_s^*)\le\zeta}\frac{c\log T}{\Delta_{s,k}^{(s)}},\end{aligned}$$

where $\bar{t}^{(1b)} = \max_{s\in\mathcal{S}} t_s^{(1b)}$ and $\underline{U} = \min_{s\in\mathcal{S}} U_s$.

Bound of $R^{(1)}(T,B)$: Finally, from the bounds of $R^{(1a)}$ and $R^{(1b)}$, we have the regret bound of $R^{(1)}$ as following:

$$R^{(1)}(T,B) \le (T\bar{\delta} + S\log T)\bar{\Delta} + \sum_{s\in\mathcal{S}}\sum_{k:p_{s,k}(a_s^*)\le\zeta} \frac{c\log T}{\Delta_{s,k}^{(s)}} + O(1). \tag{8}$$

D.2 Bound of $R^{(2)}$

We now bound the second part of the regret, $R^{(2)}$. Recall that the regret R is decomposed into two parts: the regret due to taking suboptimal doses $R^{(1)}$ and the regret due to ordering errors in subgroups $R^{(2)}$. Thus, in here, we do not have to consider the suboptimal doses and consider the ordering errors only.

In Wu et al. (2015), the regret due to the ordering error is analyzed. Compared with the case that is analyzed, we additionally consider the safety constraint. However, we can follow the analysis on the regret due to the ordering error in Wu et al. (2015) for the bound of $R^{(2)}$ since the safety constraint reduces the ordering errors by excluding the unsafe doses which is not the optimal doses. Thus, we can provide the regret bound of $R^{(2)}$ by using the analysis.

Before providing the regret bound, we define some boundary cases according to ρ and η_s's since the bound depends on them. We first define a non-boundary case for a given fixed $\rho \in (0,1)$ as a case in which $\rho \neq \eta_s$ for any $s \in \mathcal{S}$, and define a boundary case for a given fixed $\rho \in (0,1)$ as a case in which $\rho = \eta_s$ for some $s \in \mathcal{S}$. Then, by applying the analysis on our algorithm, we have the regret bounds on the following lemma.

Lemma 5. *(Wu et al., 2015) Given a fixed $\rho \in (0,1)$, the regret $R^{(2)}(T,B)$ is bounded as follows:*

(1) For the non-boundary case,

$$R^{(2)}(T,B) \le [\bar{q}^* + v(\rho)]M_{nb}^{(2)}\log T + O(1)$$

(2) For the boundary case,

$$R^{(2)}(T,B) \le q_1^*\sqrt{\rho(1-\rho)T} + M_b^{(2)}\log T + O(1)$$

where $\bar{q}^* = \sum_{s\in\mathcal{S}}\pi_s q_s^*$,

$$M_{nb}^{(2)} = \sum_{s=1}^{\tilde{s}(\rho)}\sum_{k\in\mathcal{K}} \frac{27}{2g_{\tilde{s}(\rho)+1}^{nb}\left[\Delta_{\tilde{s}(\rho)+1,k}^{(s)}\right]^2} + \sum_{s=\tilde{s}+2}^{S}\sum_{k\in\mathcal{K}} \frac{27}{2g_s^{nb}\left[\Delta_{s,k}^{(\tilde{s}+1)}\right]^2} + 2SK,$$

$$M_b^{(2)} = \sum_{s=1}^{\tilde{s}(\rho)-1}\sum_{k\in\mathcal{K}} \frac{27}{2g_{\tilde{s}(\rho)}^{b}\left[\Delta_{\tilde{s}(\rho),k}^{(s)}\right]^2} + \sum_{s=\tilde{s}+1}^{S}\sum_{k\in\mathcal{K}} \frac{27}{2g_s^{b}\left[\Delta_{s,k}^{(\tilde{s})}\right]^2} + 2SK,$$

$$g_s^{nb} = \min\left\{\pi_s, \frac{1}{2}(\rho - \eta_{\tilde{s}(\rho)}), \frac{1}{2}(\eta_{\tilde{s}(\rho)+1} - \rho)\right\},$$

$$\text{and } g_s^{b} = \min\left\{\pi_s, \frac{1}{2}(\rho - \eta_{\tilde{s}(\rho)-1}), \frac{1}{2}(\eta_{\tilde{s}(\rho)+1} - \rho)\right\}.$$

D.3 Regret Bound of $R(T,B)$

Form Lemma 5, we can see that the boundary case has a worse bound $O(\sqrt{T}\log T)$ than the non-boundary case $O(\log T)$. Hence, with (8) and Lemma 5, we have the worst-case regret bound of C3T-Budget-E in the theorem with

$$M = \sum_{s\in\mathcal{S}}\sum_{k:p_{s,k}(a_s^*)\le\zeta} \frac{c}{\Delta_{s,k}^{(s)}} + \sum_{s=1}^{\tilde{s}(\rho)-1}\sum_{k\in\mathcal{K}} \frac{27}{2g_{\tilde{s}(\rho)}^{b}\left[\Delta_{\tilde{s}(\rho),k}^{(s)}\right]^2} + \sum_{s=\tilde{s}+1}^{S}\sum_{k\in\mathcal{K}} \frac{27}{2g_s^{b}\left[\Delta_{s,k}^{(\tilde{s})}\right]^2} + 2S(K+\bar{\Delta}),$$

$$\text{and } g_s^{b} = \min\left\{\pi_s, \frac{1}{2}(\rho - \eta_{\tilde{s}(\rho)-1}), \frac{1}{2}(\eta_{\tilde{s}(\rho)+1} - \rho)\right\}.$$

□

E DESCRIPTION OF C3T-Budget-E

For C3T-Budget-E, we consider the following formulation:

$$\begin{aligned} \text{maximize} \quad & E_\Pi(T,B) \\ \text{subject to} \quad & \mathbb{P}\left[S_{\Pi,s}(T,B) \le \zeta\right] \ge 1-\delta_s, \ \forall s \in \mathcal{S} \\ & \textstyle\sum_{t=1}^{T} Z_t \le B. \end{aligned}$$

where we have simply substituted the objective function $D_\Pi(T,B)$ in the limited-budget C3T problem (See (2) in our main paper) with $E_\Pi(T,B)$. With this formulation, the agent tries to achieve high efficacies (rather than low dose recommendation error), which results in focusing on subgroups with high efficacies. We now provide a detailed description of C3T-Budget-E to solve the above problem.

Algorithm 1 C3T-Budget-E

Input: Time-horizon T, budget B, and subgroup arrival distributions π_s's
Initialize: $\tau = T$, $b = B$, $t = 1$
while $t \le T$ **do**
  $\hat{a}_s(t) \leftarrow \frac{\sum_{k=1}^{K} \hat{a}_{s,k}(t-1)N_{s,k}(t-1)}{N_s(t-1)}, \forall s \in \mathcal{S}$
  $\mathcal{K}_s(t) = \{k \in \mathcal{K} : p_{s,k}(\hat{a}_s(t) + \alpha_s(t)) \le \zeta\}, \forall s \in \mathcal{S}$
  if $b > 0$ **then**
    if $N_{H(t)}(t) \le K$ **then**
      Sample each dose once $I(t) = N_{H(t)}(t)$
    else
      $k^*_s(t) \leftarrow \text{argmax}_{k \in \mathcal{K}_s} \hat{q}_{s,k}(t), \forall s \in \mathcal{S}$
      $\hat{q}^*_s(t) \leftarrow \max_{k \in \mathcal{K}_s} \hat{q}_{s,k}(t), \forall s \in \mathcal{S}$
      Obtain $\hat{\boldsymbol{\psi}}(b/\tau)$'s by solving the LP problem (See (3) in our main paper) with ordered $\hat{q}^*_s(t)$'s
      Allocate dose $I(t) = \begin{cases} k^*_{H(t)}(t), & \text{with probability } \hat{\psi}_{H(t)}(b/\tau), \\ 0, & \text{otherwise.} \end{cases}$
    end if
  end if
  Observe the efficacy X_t and toxicity Y_t
  Update τ, b, $N_s(t)$, $N_{s,k}(t)$, $\bar{q}_{s,k}(t)$, $\bar{p}_{s,k}(t)$
  $\hat{a}_{s,k}(t) \leftarrow \text{argmin}_a |p_{s,I(t)}(a) - \bar{p}_{s,I(t)}(t)|, \forall s \in \mathcal{S}, \forall k \in \mathcal{K}$
  $t \leftarrow t + 1$
end while

F DESCRIPTION OF BASELINE ALGORITHMS

To evaluate the performance of C3T-Budget and C3T-Budget-R, we implement the following baseline algorithms:

- **Contextual UCB (C-UCB)** (Auer et al., 2002; Varatharajah et al., 2018): C-UCB is an extended version of a traditional UCB algorithm in Auer et al. (2002) for a contextual bandits. We implement it by running S instances of the tradition UCB algorithm for each subgroup as introduced in Zhou (2015). In the algorithm, the safety and budget constraints are not considered. The algorithm updates the empirical expected efficacy of each dose for each subgroup and its confidence bound. It also updates the empirical toxicities, but they are used only for the recommendation. In each round, the algorithm chooses the dose having the highest UCB index for the subgroup arrived in the round. At the end of trial, it recommends a dose for each subgroup as: $\bar{d}_s = \text{argmax}_{k:\hat{p}_{s,k} \le \zeta, \bar{q}_{s,k}(t) \ge \theta} \bar{q}_{s,k}$, where $\hat{p}_{s,k}$ and $\hat{q}_{s,k}$ are the empirical expected toxicity and efficacy of dose k for subgroup s.

- **Contextual KL-UCB (C-KL-UCB)** (Garivier and Cappé, 2011; Varatharajah et al., 2018): C-KL-UCB is an extended version of a KL-UCB algorithm in Garivier and Cappé (2011) for a contextual bandits. Similar to C-UCB, we implement it by using S instances of KL-UCB. The algorithm is same with C-UCB except for

using the KL-UCB index instead of the UCB index, which is given by

$$\hat{q}_{s,k}(t) = \sup\{q \geq \bar{q}_{s,k}(t) : N_{s,k}(t-1)I(\bar{q}_{s,k}), q) \leq \log N_s(t) + \log\log N_s(t)\},$$

where $I(p,q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$ is the Kullback-Leibler divergence.

- **Contextual independent Thompson sampling (C-Indep-TS)** (Aziz et al., 2019): C-Indep-TS is an extended version of an Indep-TS algorithm in Aziz et al. (2019); Thompson (1933) for a contextual bandits. Similar to other extended algorithms, we implement it by using S instances of Indep-TS. In Indep-TS, a Bayesian approach is used to estimate the efficacy and toxicity as follows:

$$\hat{q}_{s,k}(t) \sim \text{Beta}(\alpha^q_{s,k}(t), \beta^q_{s,k}(t)) \text{ and } \hat{p}_{s,k}(t) \sim \text{Beta}(\alpha^p_{s,k}(t), \beta^p_{s,k}(t)),$$

where $\alpha^q_{s,k}(t) = X_{s,k}(t) + 1$, $\beta^q_{s,k}(t) = N_{s,k}(t) - X_{s,k}(t) + 1$, $X_{s,k}(t) = \sum_{\tau=1}^{t} \mathbb{I}\{H(\tau) = s, I(\tau) = k\}X(\tau)$, $\alpha^p_{s,k}(t) = Y_{s,k}(t) + 1$, $\beta^p_{s,k}(t) = N_{s,k}(t) - Y_{s,k}(t) + 1$, and $Y_{s,k}(t) = \sum_{\tau=1}^{t} \mathbb{I}\{H(\tau) = s, I(\tau) = k\}Y(\tau)$. In each round t, the efficacy and toxicity for subgroup $H(t)$ is realized based on the posterior distribution as above equation, and then, the dose k that has the maximum realized efficacy $\hat{q}_{H(t),k}$ is chosen. At the end of the trial, it recommends a dose for each subgroup as: $\bar{d}_s = \text{argmax}_{k:\hat{p}_{s,k}(t)\leq\zeta, \hat{q}_{s,k}(t)\geq\theta} \hat{q}_{s,k}(t)$.

- **Contextual 3+3 (C-3+3)** (Storer, 1989): C-3+3 is an extended version of a 3+3 clinical trial design in Storer (1989) for a contextual model. Similar to other extended algorithms, we implement it by using S instances of 3+3 design. In 3+3 design, for each subgroup, the lowest dose is treated to 3 patients. Then, it observes the toxicity of the patients. If the agent observes the toxicity from none of patients, then the next dose is treated to another 3 patients. If the agent observes only one toxicity, the same dose is treated to additional 3 patients. If the agent still observes only one toxicity among the 6 patients, then the next dose is treated to another 3 patients. Otherwise, the trial is stopped and the dose treated before stopping is recommended. If the instance of 3+3 design for a subgroup is stopped once, then the patients in the subgroup are skipped.

References

Audibert, J.-Y., Bubeck, S., and Munos, R. (2010). Best arm identification in multi-armed bandits. *COLT 2010*, page 41.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256.

Aziz, M., Kaufmann, E., and Riviere, M.-K. (2019). On multi-armed bandit designs for phase I clinical trials. *arXiv preprint arXiv:1903.07082*.

Garivier, A. and Cappé, O. (2011). The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376.

Storer, B. E. (1989). Design and analysis of phase I clinical trials. *Biometrics*, 45(3):925–937.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294.

Varatharajah, Y., Berry, B., Koyejo, S., and Iyer, R. (2018). A contextual-bandit-based approach for informed decision-making in clinical trials. *arXiv preprint arXiv:1809.00258*.

Wu, H., Srikant, R., Liu, X., and Jiang, C. (2015). Algorithms with logarithmic or sublinear regret for constrained contextual bandits. In *Advances in Neural Information Processing Systems*, pages 433–441.

Zhou, L. (2015). A survey on contextual multi-armed bandits. *arXiv preprint arXiv:1508.03326*.