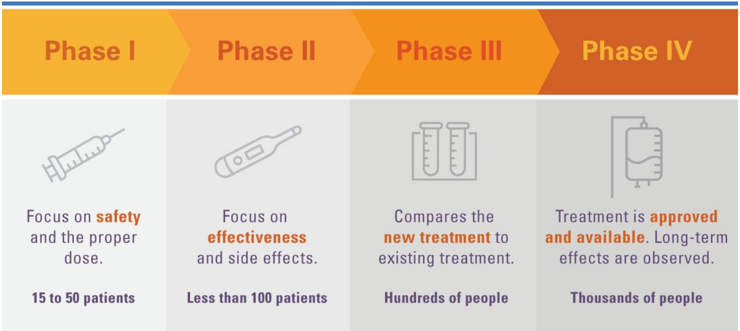
Bandit Algorithms for Early-Stage Clinical Trials

Emilie Kaufmann,

based on a joint work with
Maryam Aziz (Spotify) and Marie-Karelle Riviere (Sanofi)

Journées MAS, Rouen, August 2022

The four phases of clinical trials



Phase I	Phase II	Phase III	Phase IV
Focus on **safety** and the proper dose.	Focus on **effectiveness** and side effects.	Compares the **new treatment** to existing treatment.	Treatment is **approved and available**. Long-term effects are observed.
15 to 50 patients	**Less than 100 patients**	**Hundreds of people**	**Thousands of people**

source: MD Anderson Cancer Center

This talk: phase I, phase I/II

A stochastic model for dose-finding

Early stage trials are often about finding the right dose (or combination of doses) of a given treatment.

	Dose 1	Dose 2	$\cdots$	Dose K
toxicity probability	p_1	p_2	$\cdots$	p_K
efficacy probability	eff_1	eff_2	$\cdots$	eff_K

After selecting a dose $D_t \in \{1, \dots, K\}$ ("arm") for patient t,

- observe whether un-desired side effects occur: $X_t \sim \mathcal{B}(p_{D_t})$

$$\mathbb{P}(X_t = 1 | D_t = d) = p_d \quad \mathbb{P}(X_t = 0 | D_t = d) = 1 - p_d$$

- observe whether the treatment is efficient: $Y_t \sim \mathcal{B}(\mathrm{eff}_{D_t})$ (in phase I/II designs)

Question: what is a good arm?

A stochastic model for dose-finding

Early stage trials are often about finding the right dose (or combination of doses) of a given treatment.

	Dose 1	Dose 2	$\cdots$	Dose K
toxicity probability	p_1	p_2	$\cdots$	p_K
efficacy probability	eff_1	eff_2	$\cdots$	eff_K

After selecting a dose $D_t \in \{1, \dots, K\}$ ("arm") for patient t,

- observe whether un-desired side effects occur: $X_t \sim \mathcal{B}(p_{D_t})$

$$\mathbb{P}(X_t = 1|D_t = d) = p_d \qquad \mathbb{P}(X_t = 0|D_t = d) = 1 - p_d$$

- ~~observe whether the treatment is efficient: $Y_t \sim \mathcal{B}(\text{eff}_{D_t})$ (in phase I/II designs)~~

Question: what is a good arm?

A non-standard bandit problem

Maximum Tolerated Dose (MTD)

Given a specified threshold θ, the MTD is the dose whose probability of toxicity is closest to θ:

$$k^\star = \operatorname*{arg\,min}_{k \in [K]} |\theta - p_k|$$

Two possible goals with this alternative notion of optimal arm :

- identify the MTD as quickly as possible
  ($\simeq$ best arm identification)
- treat as many patients as possible with the MTD
  ($\simeq$ regret minimization)

Ideally both, but they are known to be conflicting objectives.

Outline

1. Solving (unstructured) MTD identification

2. Exploiting monotonicity assumptions

3. Beyond MTD identification

Outline

1. Solving (unstructured) MTD identification
2. Exploiting monotonicity assumptions
3. Beyond MTD identification

Sequential Halving for MTD Identification

```
Input: total number of patients $T$ (fixed-budget)
number of doses $K$
Initialization: $S_0 = \{1, \dots, K\}$;
For $r = 0$ to $\lceil \log_2(K) \rceil - 1$, do
    sample each arm $i \in S_r$ for $t_r = \left\lfloor \frac{T}{|S_r|\lceil \log_2(K)\rceil} \right\rfloor$ times;
    let $\hat{p}_i^r$ be the empirical toxicity of dose $i$;
    let $S_{r+1}$ be the set of $\lceil |S_r|/2 \rceil$ arms with smallest $\hat{d}_i^r := |\theta - \hat{p}_i^r|$
Return $\hat{k}_T$ the unique arm in $S_{\lceil \log_2(K)\rceil}$
```

Upper bound on the error probability [Aziz et al., 2021]

$$\mathbb{P}\left(\hat{k}_T \neq k^*\right) \leq 9 \log_2 K \cdot \exp\left(-\frac{T}{8H(\boldsymbol{p}) \log_2 K}\right),$$

where $H(\boldsymbol{p}) := \sum_{k=1}^{K} \frac{1}{\Delta_k^2}$ with $\Delta_k = |\theta - p_k| - |\theta - p_{k^\star}|$.

Sequential Halving for MTD Identification

```
Input: total number of patients $T$ (fixed-budget)
number of doses $K$
Initialization: $S_0 = \{1, \dots, K\}$;
For $r = 0$ to $\lceil \log_2(K) \rceil - 1$, do
    sample each arm $i \in S_r$ for    $t_r = \left\lfloor \frac{T}{|S_r| \lceil \log_2(K) \rceil} \right\rfloor$ times;
    let $\hat{p}_i^r$ be the empirical toxicity of dose $i$;
    let $S_{r+1}$ be the set of $\lceil |S_r|/2 \rceil$ arms with smallest $\hat{d}_i^r := |\theta - \hat{p}_i^r|$
Return $\hat{k}_T$ the unique arm in $S_{\lceil \log_2(K) \rceil}$
```

Limitations

- the error bound is only meaningful for large values of T
- uniform sampling in early phases may be unethical

Thompson Sampling for MTD Allocation

$\boldsymbol{p} = (p_1, \dots, p_K) \in [0,1]^K$: vector of toxicity probabilities

$\Pi^{(0)}$: prior distribution on $\boldsymbol{p}$
$\Pi^{(t)}$: posterior distribution after observing $(D_1, X_1, \dots, D_t, X_t)$

Thompson Sampling

Sample $(\tilde{p}_1(t), \dots, \tilde{p}_k(t)) \sim \Pi^{(t)}$ and allocate dose

$$D_{t+1} = \underset{k \in [K]}{\arg\min} \, |\tilde{p}_k(t) - \theta|$$

- **1st view**: play the optimal arm (= MTD) in a model sampled from the posterior distribution
- **2nd view**: a randomized design in which the probability to select dose k is the posterior probability that k is the MTD

Thompson Sampling for MTD Allocation

$\boldsymbol{p} = (p_1, \ldots, p_K) \in [0,1]^K$: vector of toxicity probabilities

$\Pi^{(0)}$: prior distribution on $\boldsymbol{p}$
$\Pi^{(t)}$: posterior distribution after observing $(D_1, X_1, \ldots, D_t, X_t)$

Thompson Sampling

Sample $(\tilde{p}_1(t), \ldots, \tilde{p}_k(t)) \sim \Pi^{(t)}$ and allocate dose

$$D_{t+1} = \underset{k \in [K]}{\arg\min} \, |\tilde{p}_k(t) - \theta|$$

- **1st view**: play the optimal arm (= MTD) in a model sampled from the posterior distribution
- **2nd view**: a randomized design in which the probability to select dose k is the posterior probability that k is the MTD

Independent Thompson Sampling

A simple, product prior

$\Pi^0 = \bigotimes_{i=1}^{K} \pi_k^0$, where $\pi_k^0 = \mathcal{U}([0,1])$

$\Pi^t = \bigotimes_{i=1}^{K} \pi_k^t$, where

$$\pi_k^t = \mathrm{Beta}\big(S_k(t) + 1, N_k(t) - S_k(t) + 1\big)$$

- $N_k(t)$: number of times dose k was given up to time t
- $S_k(t)$: number of times dose k was found toxic up to time t

Independent Thompson Sampling

$$\forall k \in [K],\ \tilde{p}_k(t) \sim \pi_k^t$$

$$D_{t+1} = \underset{k \in [K]}{\arg\min} \, |\theta - \tilde{p}_k(t)|$$

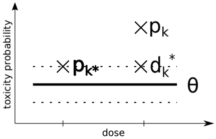
An asymptotically optimal algorithm

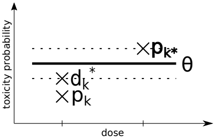
Upper bound on the number of allocations [Aziz et al., 2021]

For all $\varepsilon > 0$, there exists a constant $C_{\varepsilon,\theta,\boldsymbol{p}}$ s.t., for all $k \notin$ MTD

$$\mathbb{E}[N_k(T)] \leq \frac{1+\varepsilon}{\mathrm{kl}(p_k, d_k^*)} \log(T) + C_{\varepsilon,\theta,\boldsymbol{p}},$$

where $\mathrm{kl}(x, y)$ is the binary Kullback-Leibler divergence.





- logarithmic number of allocations to sub-optimal doses
- lower bound proving its optimality... in an asymptotic regime

Outline

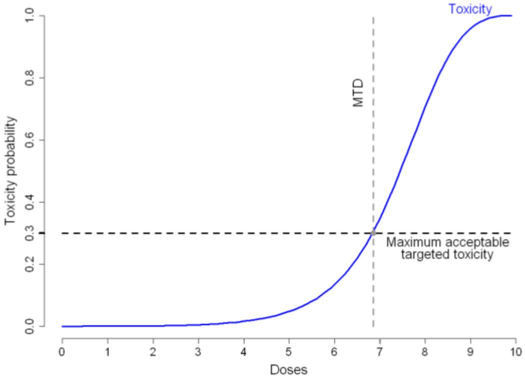
1. Solving (unstructured) MTD identification

2. Exploiting monotonicity assumptions

3. Beyond MTD identification

A structured bandit problem

For clinical trials involving a single agent, the toxicity is increasing with the dose :



How to incorporate this information in algorithms?

- [Garivier et al., 2019] : an identification algorithm
- this work: Thompson Sampling

A Bayesian model for increasing toxicities

Parametric assumption: given two parameters $\beta_0, \beta_1 \in \mathbb{R}$,

$$p_k(\beta_0, \beta_1) = \frac{1}{1 + e^{-\beta_0 - \beta_1 u_k}}$$

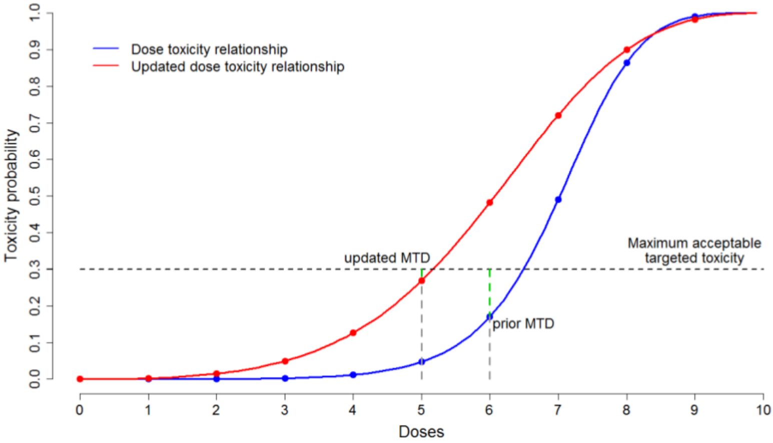
u_k: effective dose (some carefully chosen parameter)

Bayesian model: $(\beta_0, \beta_1) \sim \pi$, e.g.

$$\beta_0 \sim \mathcal{N}(0, 100) \;\text{ and }\; \beta_1 \sim \mathrm{Exp}(1).$$

➜ the posterior distribution π_t on (β_0, β_1) can be sampled from (e.g. using Hamiltonian Monte-Carlo methods)

Illustration of the posterior update



source: Marie-Karelle Riviere (PhD thesis)

Thompson Sampling versus the CRM

Thompson Sampling

$$\left(\tilde{\beta}_0(t), \tilde{\beta}_1(t)\right) \sim \pi_t,$$

$$D_{t+1}^{\text{TS}} \in \underset{k\in[K]}{\arg\min} \left|\theta - p_k\left(\tilde{\beta}_0(t), \tilde{\beta}_1(t)\right)\right|$$

Continual Reassesment Method (CRM) [O'Quingley et al., 1990]

$$\hat{\beta}_i(t) = \int_{\mathbb{R}} \beta_i d\pi_t(\beta_0, \beta_1) \quad \text{(posterior mean)}$$

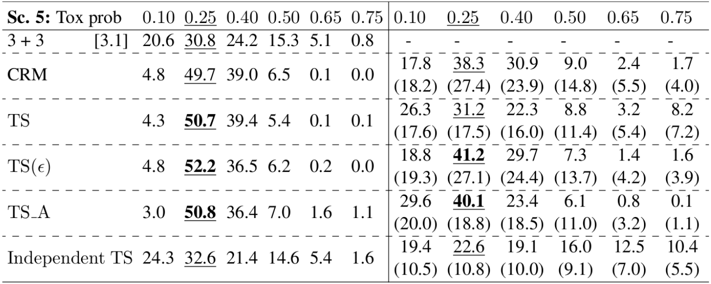
$$D_{t+1}^{\text{CRM}} \in \underset{k\in[K]}{\arg\min} \left|\theta - p_k\left(\hat{\beta}_0(t), \hat{\beta}_1(t)\right)\right|$$

➜ compared to the existing CRM, TS is adding exploration

Empirical performance

Too much exploration may be un-ethical ➜ two variants of TS restricting the set of doses that can be chosen

$\boldsymbol{T} = 36$ **patients** , $K = 6$ doses , $\theta = 0.3$



Sc. 5: Tox prob	0.10	0.25	0.40	0.50	0.65	0.75	0.10	0.25	0.40	0.50	0.65	0.75
3 + 3 [3.1]	20.6	30.8	24.2	15.3	5.1	0.8	-	-	-	-	-	-
CRM	4.8	49.7	39.0	6.5	0.1	0.0	17.8 (18.2)	38.3 (27.4)	30.9 (23.9)	9.0 (14.8)	2.4 (5.5)	1.7 (4.0)
TS	4.3	**50.7**	39.4	5.4	0.1	0.1	26.3 (17.6)	31.2 (17.5)	22.3 (16.0)	8.8 (11.4)	3.2 (5.4)	8.2 (7.2)
TS(ϵ)	4.8	**52.2**	36.5	6.2	0.2	0.0	18.8 (19.3)	**41.2** (27.1)	29.7 (24.4)	7.3 (13.7)	1.4 (4.2)	1.6 (3.9)
TS_A	3.0	**50.8**	36.4	7.0	1.6	1.1	29.6 (20.0)	**40.1** (18.8)	23.4 (18.5)	6.1 (11.0)	0.8 (3.2)	0.1 (1.1)
Independent TS	24.3	32.6	21.4	14.6	5.4	1.6	19.4 (10.5)	22.6 (10.8)	19.1 (10.0)	16.0 (9.1)	12.5 (7.0)	10.4 (5.5)

% of recommendation (left) and allocation (right)
(average over 2000 repetitions)

Outline

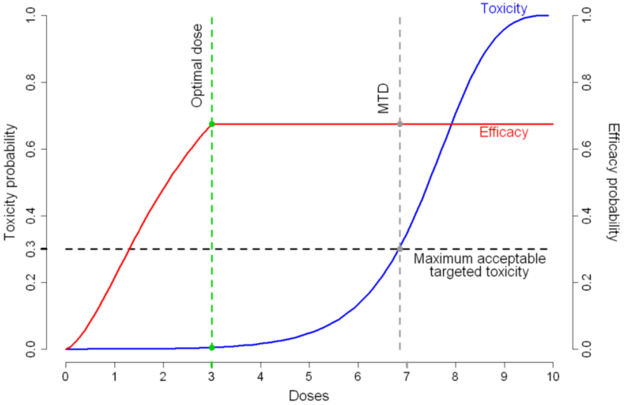
1. Solving (unstructured) MTD identification

2. Exploiting monotonicity assumptions

3. Beyond MTD identification

A two-dimensional structured bandit

For certain agents, a plateau of efficacy is observed, which motivates the search of the Minimal Effective Dose (MED)



$$k^\star = \min \left\{ k \in [K] : \mathrm{eff}_k = \max_{\ell : p_\ell \leq \theta} \mathrm{eff}_\ell \right\}$$

A Bayesian model

Toxicity: $p_k(\beta_0, \beta_1) = \frac{1}{1+e^{-[\beta_0+\beta_1 u_k]}}$

$$\beta_0 \sim \mathcal{N}(0, 100), \quad \beta_1 \sim \mathrm{Exp}(1)$$

Efficacy: τ indicates the beginning of the plateau

$$\mathrm{eff}_k(\gamma_0, \gamma_1, \tau) = \frac{1}{1 + e^{-[\gamma_0+\gamma_1(v_k \mathbb{1}(k<\tau)+v_\tau \mathbb{1}(k\geq\tau))]}}$$

$$\gamma_0 \sim \mathcal{N}(0, 100), \quad \gamma_1 \sim \mathrm{Exp}(1), \quad \tau \sim (1/K, \ldots, 1/K).$$

Thompson Sampling

$$\left(\tilde{\beta}_0(t), \tilde{\beta}_1(t), \tilde{\gamma}_0(t), \tilde{\gamma}_1(t), \tilde{\tau}(t)\right) \sim \pi_t,$$

$$D_{t+1}^{\mathrm{TS}} \in \mathrm{MED}\left(\tilde{\beta}_0(t), \tilde{\gamma}_0(t), \tilde{\beta}_1(t), \tilde{\gamma}_1(t), \tilde{\tau}(t)\right)$$

Results

Competitive results wrt. the state-of the art MTA-RA algorithm [Riviere et al., 2017]

$$\boldsymbol{T} = 60 \textbf{ patients}, \ K = 6 \text{ doses}, \ \theta = 0.35$$

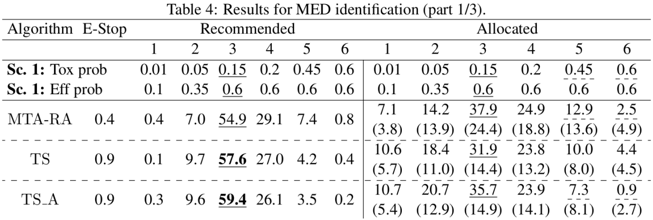


Table 4: Results for MED identification (part 1/3).

Algorithm	E-Stop	Recommended						Allocated					
		1	2	3	4	5	6	1	2	3	4	5	6
Sc. 1: Tox prob		0.01	0.05	0.15	0.2	0.45	0.6	0.01	0.05	0.15	0.2	0.45	0.6
Sc. 1: Eff prob		0.1	0.35	0.6	0.6	0.6	0.6	0.1	0.35	0.6	0.6	0.6	0.6
MTA-RA	0.4	0.4	7.0	54.9	29.1	7.4	0.8	7.1 (3.8)	14.2 (13.9)	37.9 (24.4)	24.9 (18.8)	12.9 (13.6)	2.5 (4.9)
TS	0.9	0.1	9.7	**57.6**	27.0	4.2	0.4	10.6 (5.7)	18.4 (11.0)	31.9 (14.4)	23.8 (13.2)	10.0 (8.0)	4.4 (4.5)
TS_A	0.9	0.3	9.6	**59.4**	26.1	3.5	0.2	10.7 (5.4)	20.7 (12.9)	35.7 (14.9)	23.9 (14.1)	7.3 (8.1)	0.9 (2.7)

% of recommendation (left) and allocation (right)
(average over 2000 repetitions)

Conclusion

Thompson Sampling is a flexible algorithm for which we gave examples of applications in

- phase I trials (one critrion: toxicity)
- phase I/II trials (two criteria: toxicity and efficacy)

➜ what if there are more than two criteria?
(e.g. multiple indicators of efficacy)

A big **gap between theory and practice**:

- theoretical guarantees for an independent prior
- prior distributions leveraging extra information used in practice (with only some consistency guarantees for the CRM)

Some open questions:

- Do we need exploration?
- How to appropriately balance allocation (=treatment) and recommendation (=identification)?

References

Aziz, M., Kaufmann, E., and Riviere, M. (2021).
On multi-armed bandit designs for dose-finding clinical trials.
Journal of Machine Learning Research, 22(14):1–38.

Garivier, A., Ménard, P., and Rossi, L. (2019).
Thresholding bandit for dose-ranging: The impact of monotonicity.
In International Conference on Machine Learning, Artificial Intelligence and Applications.

O'Quingley, J., Pepe, M., and Fisher, L. (1990).
Continual reassessment method: A practical design for phase I clinical trials in cancer.
Biometrics, 46(1):33–48.

Riviere, M.-K., Yuan, Y., Jourdan, J.-H., Dubois, F., and Zohar, S. (2017).
Phase i/ii dose-finding design for molecularly targeted agent: Plateau determination using adaptive randomization.
Statistical Methods in Medical Research.

Details on the variants of Thompson Sampling

TS(ε) outputs a dose that belongs to the set

$$\left\{k \in [K] : \left|p_k(\hat{\beta}_0(t), \hat{\beta}_1(t)) - p_{\mathrm{MTD}(\hat{\beta}_0(t), \hat{\beta}_1(t))}(\hat{\beta}_0(t), \hat{\beta_1(t)})\right| \leq \varepsilon\right\}$$

$$(\varepsilon = 0.05)$$

TS_A outputs a dose that belongs to the set

$$\left\{k \in [K] : \mathbb{P}_{(\beta_0,\beta_1)\sim\pi_t}\left(p_k(\beta_0, \beta_1) > p_{\mathrm{MTD}(\beta_0,\beta_1)}(\beta_0, \beta_1)\right) \leq c_1\right\}$$

$$(c_1 = 0.8)$$