The effect of estimating prevalences on the population-wise error rate

Remi Luschei and Werner Brannath
University of Bremen, Germany
ADMTP Workshop, Basel 2023

The population-wise error rate (PWER, *Brannath et al. 2023*)

- The PWER is a type I error rate for clinical trials with multiple populations
- A multiplicity issue may occur here
- Control of the family-wise error rate may be too stringent
- The PWER is an "essential" error criterion, i.e. it only accounts for (multiple) type I errors that may affect future patients

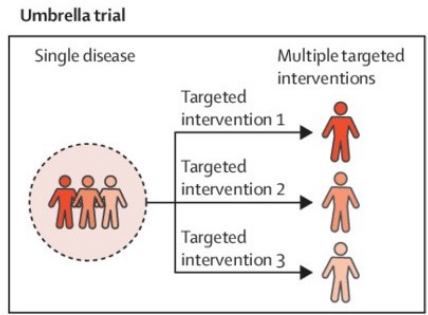
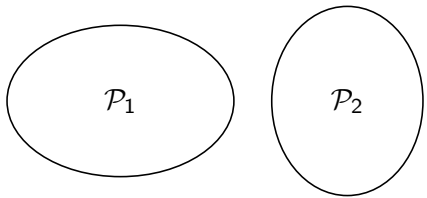


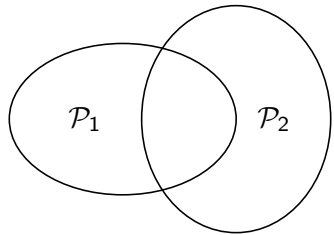
Figure: Example of a trial with multiple populations (*Park et al. 2021*)

Disjoint populations



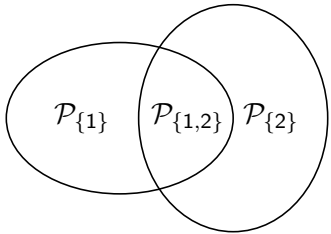
- We test the hypotheses $H_1 : \theta(T_1, \mathcal{P}_1) \leq 0$ and $H_2 : \theta(T_2, \mathcal{P}_2) \leq 0$
- A type I error at H_1 does not harm anyone in $\mathcal{P}_2$, and vice versa
- Unadjusted testing does not increase the chance for receiving inefficient treatments
- No multiplicity adjustment required for protection of patients

Intersecting populations



- We test the hypotheses $H_1 : \theta(T_1, \mathcal{P}_1) \leq 0$ and $H_2 : \theta(T_2, \mathcal{P}_2) \leq 0$
- The patients in $\mathcal{P}_1 \cap \mathcal{P}_2$ could now be affected by a type I error at H_1 or H_2.
- To bound the risk for future patients, we should account for this with some kind of multiplicity adjustment.
- The larger the intersection the more relevant is the adjustment.

Intersecting populations



- **Population-wise error rate (PWER):**

  Probability that a randomly selected, future patient will be exposed to an inefficient treatment:

$$\text{PWER} = \pi_{\{1\}}\mathbb{P}(\text{falsely reject } H_1) + \pi_{\{2\}}\mathbb{P}(\text{falsely reject } H_2)$$
$$+ \pi_{\{1,2\}}\mathbb{P}(\text{falsely reject } H_1 \text{ or } H_2)$$

General definition of the PWER

- Regard m patient populations $\mathcal{P}_i$ with corresponding null hypotheses $H_i : \theta_i \leq 0$, where $\theta_i = \theta(\mathcal{P}_i, T_i)$ is the efficacy of treatment T_i in $\mathcal{P}_i$ (in comparison to a control)
- Define the strata $\mathcal{P}_J := \left(\cap_{j \in J} \mathcal{P}_j\right) \setminus \left(\cup_{k \notin J} \mathcal{P}_k\right), J \subseteq I = \{1, \ldots, m\}$ with prevalences $\pi_J = \mathbb{P}(\mathcal{P}_J)$ among $\mathcal{P} = \bigcup_{i=1}^{m} \mathcal{P}_i$
- The population-wise error rate is defined by

$$\text{PWER} := \sum_{J \subseteq I} \pi_J \mathbb{P}(\text{falsely reject any } H_j \text{ with } j \in J)$$

- We have PWER $\leq$ FWER.

Control of the PWER

- Assume that each H_i is tested with a continuous test statistic Z_i.
- The PWER is often maximized under the global null hypothesis, for example when the Subset Pivotality assumption (*Westfall and Young 1993*) is fulfilled
- Example: Z- or t-distributed statistics for normally distributed data with homogeneous variance
- Exhaustive PWER-control is possible by computing the rejection boundary c from

$$\text{PWER} = \sum_{J \subseteq I} \pi_J \mathbb{P}_I \left(\cup_{j \in J} \{ Z_j > c \} \right) = \alpha,$$

because the parameters of the Z-/t-distribution only depend on the sample size vector $(n_J)_{J \subseteq I}$

Estimation of the prevalences

- The prevalences $\pi = (\pi_J)_{J \subseteq I}$ are usually unknown
- Estimate π by the MLE $\hat{\pi}$ of the multinomial distribution $\mathrm{M}(N, \pi)$
- Let n_J be the number of patients sampled from $\mathcal{P}_J$, with $\sum_{J \subseteq I} n_J = N$
- $\hat{\pi}_J = n_J / N$
- Control $\widehat{\mathrm{PWER}} = \sum_{J \subseteq I} \hat{\pi}_J \mathbb{P}\left(\cup_{j \in J} \{Z_j > \hat{c}\}\right)$
- Is the true PWER still controlled in this situation?
- Asymptotically, it is, because $\hat{\pi}$ is asymptotically consistent:

$$\mathrm{PWER} - \alpha = \sum_{J \subseteq I} (\pi_J - \hat{\pi}_J) \, \mathbb{P}\left(\cup_{j \in J} \{Z_j > \hat{c}\}\right) \xrightarrow[N \to \infty]{\text{a.s.}} 0$$

Setup of the simulations

- Define the true $\pi = (\pi_J)_{J \subseteq I}$ for populations $\mathcal{P}_j = \{B_j = 1\}$ with independent, binary biomarkers B_j (e.g. randomly generate the biomarker probabilities)
- Generate random sample sizes $(n_J)_{J \subseteq I}$
- Calculate $\hat{c}$ from

$$\widehat{\text{PWER}} = \sum_{J \subseteq I} \hat{\pi}_J \left(1 - F_{\Sigma_J}(\hat{c}, \ldots, \hat{c})\right) = \alpha.$$

- Compare the true PWER

$$\text{PWER}\,(\hat{c}) = \sum_{J \subseteq I} \pi_J \left(1 - F_{\Sigma_J}(\hat{c}, \ldots, \hat{c})\right)$$

with α

Distribution of the simulated values of PWER($\hat{c}$)

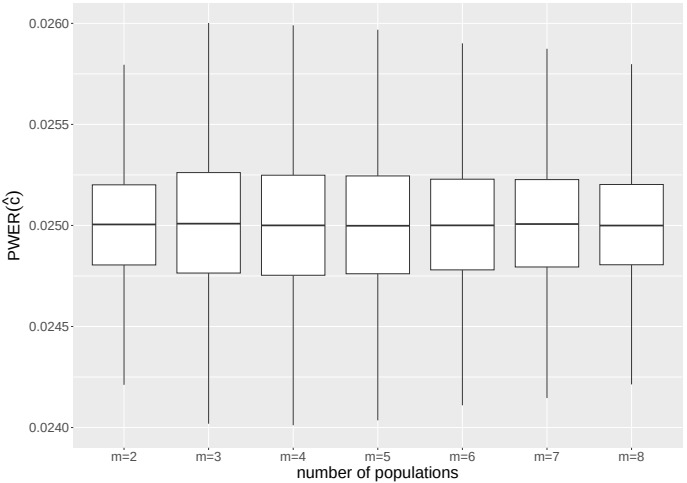


Figure: 10^4 simulated values of PWER($\hat{c}$) for $\alpha = 0.025$, $N = 500$ and different m

Other treated cases (with similar results)

- Random or fixed true prevalences
- Dependent or independent biomarker probabilities
- Known or unknown variances of the response
- Population-specific treatments or one single treatment to be tested in all populations

Distribution of PWER($\hat{c}$) with one prevalence fixed at $\pi_J = 0.5$

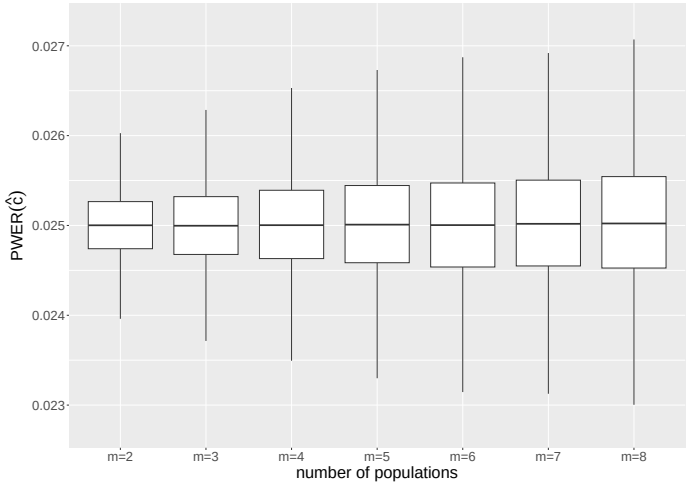


Figure: 10^4 simulated values of PWER($\hat{c}$) for $\alpha = 0.025$, $N = 500$ and different m

Distribution of the simulated values for different total sample sizes

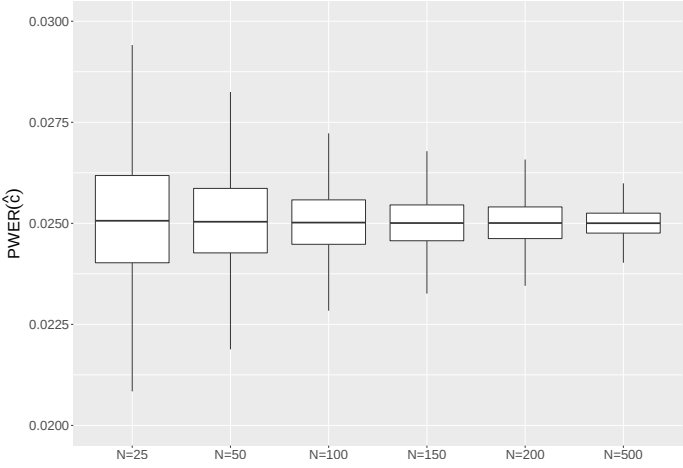


Figure: 10^4 simulated values of PWER($\hat{c}$) for $\alpha = 0.025$, $m = 3$ and different N

The maximal strata-wise FWER

- Unter PWER-control, we are also interested in controlling the strata wise error rates $\text{FWER}_J(c) := \mathbb{P}(\max_{j \in J} Z_j > c)$
- We see in our simulations that the estimated FWER_J also approximate the true ones very well
- *Brannath et al. 2023* give some upper bounds for $\max_{J \subseteq I} \text{FWER}_J(\hat{c})$ whose quality depend on different factors like π_J or $|J|$

The maximal strata-wise FWER

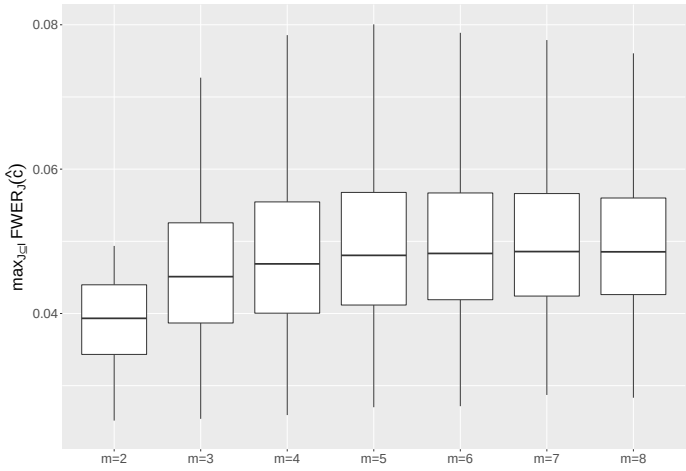


Figure: 10^4 simulated values of $\max_{J \subseteq I} \mathrm{FWER}_J(\hat{c})$ for $\alpha = 0.025$, $N = 500$ and different m

Introduction of a minimal prevalence for neglected strata

- In case of small strata $\mathcal{P}_J$ it could happen that no patient is recruited in $\mathcal{P}_J$.
- Then the PWER does not account for the multiplicity in $\mathcal{P}_J$.
- This is especially a problem when the strata are intersections of many different populations
- Solution: introduce a minimal prevalence π_{min} for $\mathcal{P}_J$ in the PWER
- In the simulations, in cases of neglected intersections of many populations, for $\pi_{\text{min}} = 1/(2^m - 1)$ we get $\max_J \text{FWER}_J \leq 0.055$

Introduction of a minimal prevalence for neglected strata

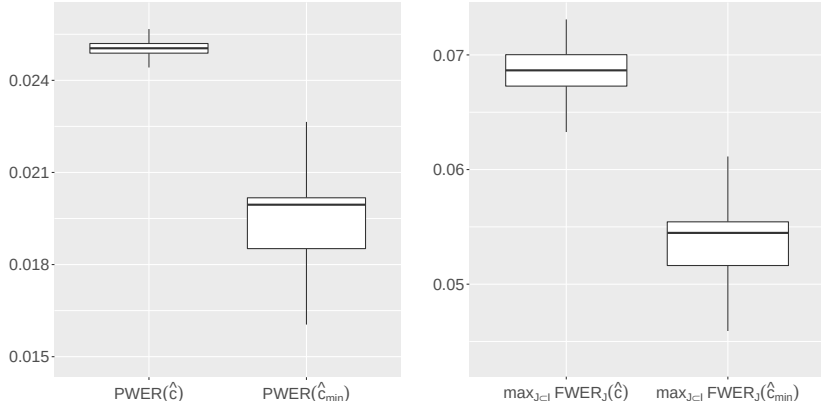


Figure: True PWER (left) and maximal strata-wise FWER (right) with and without an upscaling by π_{min} (in a setting with $m = 3$ and $N = 500$)

Summary

- The PWER is type I error criterion adapted to clinical trials with overlapping populations
- The PWER gives the average probability of exposing a future patient to an inefficient treatment: $\text{PWER} = \sum_{J \subseteq I} \pi_J \text{FWER}_J$
- The prevalences π_J must be estimated, e.g. by MLE of multinomial distribution
- This does not cause a significant inflation of the true PWER
- Asymptotically, the true PWER is perfectly controlled
- $\max_J \text{FWER}_J$ is bounded by 2α in our simulation setting
- Introduction of π_{min} reduces the FWER of neglected strata

References

- Brannath, Werner, Charlie Hillner, and Kornelius Rohmeyer (2023). "The population-wise error rate for clinical trials with overlapping populations". In: *Statistical Methods in Medical Research* 32.2, pp. 334–352.
- Park, J.J.H. et al. (2021). "Randomised trials at the level of the individual". In: *The Lancet Global Health* 9.5, e691–e700.
- Westfall, P. H. and S. S. Young (1993). *Resampling-based multiple testing. Examples and methods for p-value adjustment*. New York: Wiley.