RESEARCH ARTICLE

Maximising power with the minimum number of mosquitoes: Designing robust sample sizes for the WHO Cone Bioassay

Gemma Francesca Harvey, Frank Mechan, Giorgio Praulins, Rosemary Lees

Liverpool School of Tropical Medicine, Liverpool, England, L3 5QA, UK

v1 **First published:** 27 Aug 2025, **9**:64
https://doi.org/10.12688/gatesopenres.16361.1
Latest published: 27 Aug 2025, **9**:64
https://doi.org/10.12688/gatesopenres.16361.1

Abstract

Power calculations are an essential component of experimental design when evaluating vector control tools. Determining appropriate sample sizes for robustly detecting a difference between treatment groups in a bioassay (or any comparative experiment) is complicated by multiple sources of variation. While modern simulation-based methods exist to account for compounding sources of variation, uptake is slow due to limited availability of training and hardware. Here we present an accessible, user-friendly framework for performing sample size calculations for World Health Organisation (WHO) cone bioassays. Additionally, we conduct a literature review of studies published between 1998 and 2024 to identify sources of variability in WHO cone bioassay methodologies.

We use simulation-based methods to assess power in the WHO cone bioassay, utilising the 2013 WHO guidance for phase I laboratory testing of 'long-lasting' insecticidal nets as an illustrative example of how sample size impacts detection of differences in mosquito mortality between treatments. Futhermore, we establish plausible variability assumptions across three levels of bioassay variability for the testing of insecticide-treated nets: between nets of the same product, between net pieces from the same net and between replicates on a net piece.

We demonstrate that the biggest factor in determining the number of samples, and therefore mosquitoes, needed in an experiment is the effect size to be detected (mortality difference between treatments). Larger mortality differences (e.g. a 20% difference) are readily detected with the phase I guidance yet detecting a 10% difference requires more than triple this sample size. Here, we present a user-friendly browser application to allow researchers to easily design robust WHO cone bioassay experiments (link: Cone Bioassay Sample

Size app).

Keywords
Malaria, bioassay, mosquito, insecticide, sample size, power analysis, variability, vector control

Corresponding author: Frank Mechan (frank.mechan@lstmed.ac.uk)

Author roles: Harvey GF: Data Curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation; **Mechan F**: Conceptualization, Formal Analysis, Methodology, Software, Supervision, Writing – Original Draft Preparation; **Praulins G**: Data Curation, Formal Analysis, Investigation, Methodology, Visualization, Writing – Original Draft Preparation; **Lees R**: Funding Acquisition, Project Administration, Resources, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported, in whole or in part, by the Gates Foundation [INV-050591 – I2I 3 grant number]. The conclusions and opinions expressed in this work are those of the author(s) alone and shall not be attributed to the Foundation. Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 License has already been assigned to the Author Accepted Manuscript version that might arise from this submission. Please note works submitted as a preprint have not undergone a peer review process. Additional funding to support this research was provided by Vestergaard Sarl.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Harvey GF, Mechan F, Praulins G and Lees R. **Maximising power with the minimum number of mosquitoes: Designing robust sample sizes for the WHO Cone Bioassay [version 1; peer review: awaiting peer review]** Gates Open Research 2025, **9**:64 https://doi.org/10.12688/gatesopenres.16361.1

First published: 27 Aug 2025, **9**:64 https://doi.org/10.12688/gatesopenres.16361.1

Introduction

Insecticide-based interventions are a core strategy in malaria vector control (WHO, 2024d). In most malaria-endemic countries, the main vector control tools are insecticide-treated nets (ITNs) and indoor residual spraying (IRS). Between 2000 and 2015 it was estimated these interventions averted 542-753 million clinical *Plasmodium falciparum* cases, with ITNs suggested to have contributed between 62-72% to this reduction (Bhatt et al., 2015). However, the effectiveness of pyrethroids, the mainstay insecticide class on ITNs is jeopardised by the development of resistance in vector populations across Africa (Ranson et al., 2011, Moyes et al., 2020, Hancock et al., 2020), incentivising the development of ITNs containing novel compounds.

New insecticide-based products undergo a comprehensive evaluation of their efficacy, safety and quality testing as part of the World Health Organisation (WHO) prequalification (PQ) for Vector Control Products process to obtain a PQ listing (WHO, 2025). The initial stages of efficacy and quality testing are laboratory based, requiring demonstration of insecticidal effect before subsequent testing (such as regeneration studies, wash resistance studies, and semi-field experiments).

For evaluation of neurotoxic insecticides, laboratory assays conventionally force mosquito contact with the treated surface to assess bioefficacy of active ingredient/s (AI) (Skovmand et al., 2021). The WHO cone bioassay is a widely used laboratory bioassay to test new ITN and new IRS products (Mbwambo et al., 2022, Mbuba et al., 2023, WHO, 2013, 2024c). Within this study we focus on the WHO cone bioassays use for testing ITNs. The stated purpose of this bioassay is to demonstrate the "biological availability and potency of AI(s) on the surface of a test material" as per the 2024 WHO PQ of Vector Control Products Implementation guidance (Modules 3 and 5 Bioassay methods for insecticide-treated nets: Cone test) (WHO, 2024c). Additionally, WHO cone bioassays are used in durability studies, to investigate changes in ITN bioefficacy with operational use. Furthermore, with the development of novel AIs targeted against pyrethroid-resistant mosquitoes; the WHO cone bioassay is used to evaluate pyrethroid-nets supplemented with a second AI such as the synergist piperonyl butoxide (PBO).

The key outcomes of the WHO cone bioassay are 1 hr knockdown and 24 hr mortality. The sample size recommended in the existing WHO (2013) guidelines for laboratory and field-testing of 'long-lasting' insecticidal nets in phase I laboratory studies is fixed at four nets (with one net piece from each net and ten cones/replicates per net piece, for a total of 200 mosquitoes per ITN product being assessed). In the 2024 WHO PQ of Vector Control Products Implementation guidance (Modules 3 and 5, Bioassay methods for insecticide-treated nets: Cone test) gives recommendation to the number of mosquitoes per cone, exposure time, mosquito age, test room environmental conditions and cone angle (WHO, 2024c). Recommendations on sample size in this guidance are flexible, with the total sample size of nets dependent on the study goals and the product. Therefore, sample size calculations are the responsibility of the testing facility or manufacturer. In regeneration studies and wash resistance studies, the guidance does recommend a minimum of four ITN samples (a single net piece sample from four nets) should undergo bioassay testing (WHO, 2024a,b).

As with all methods including quality testing and the evaluation of new chemistries, it is important that the testing process is standardised between laboratories to ensure results are robust and reproducible. Moreover, for conclusions from these results to be valid, the limits of interpretation must be well-defined, which will depend on the sample size. Specifically, the researcher must know the smallest effect size (generally mortality difference between treatment groups) their sample size can reliably detect (Mechan et al., 2024). For example, if a sample size is minimally designed to detect a 10% mortality difference with 80% chance of detection a.k.a 'power' per convention, then an observed difference smaller than 10% is not a robust estimate or confirmation of the effect, regardless of p-value <0.05 (Mechan et al., 2024, Ioannidis, 2005). However, in practice the number of mosquitoes available for testing is often a key limiting factor in the number of replicates performed in WHO cone bioassays. To bridge these practical and statistical considerations, the pragmatic question when designing an experiment should be "what is the smallest difference I can robustly detect with the resources available to me?".

Given that the smallest effect size that can be detected with a sample size is intrinsically tied to the variability of the outcome, well supported assumptions about variability are essential to produce robust data and conclusions (Mechan et al., 2024). If the data is more variable than was assumed by the experimental design, the chance of detecting the difference falls and the researchers may fail to detect a genuine effect. This variability in mosquito mortality has multiple sources, some which can be minimised by controlling the conditions of the experiments (Mechan et al., 2024, Praulins et al., 2024). As part of this study, we reviewed literature on WHO cone bioassay methodologies from 1998 to 2024 to investigate sources of variability. These findings were used to refine recommendations for minimizing variability. Assuming the appropriate measures have been taken to minimise variability, the residual variation must be quantified and built into sample size calculations.

In the WHO cone bioassay there are three core units of variability within the experimental design: (1) different examples of the same net product or 'net variability', (2) different net pieces from the same net or 'net piece variability', and (3) different replicate cones on the same piece or 'replicate variability'. This does not exclude the occurrence of other sources of variability, such as biological or environmental variation. A robust sample size calculation must assume a defined amount of variability for each of these three units. This multilevel sample size approach is generally not accounted for in vector biology experiments as commonly used conventional formulas are not designed to handle multiple, hierarchically arranged, interacting sources of variation. However, recent developments in simulation-based sample size methods have overcome this challenge of considering multiple levels of variability in experimental designs (Mechan et al., 2024, Johnson et al., 2015), with associated app-based user interfaces making otherwise challenging statistical programming highly accessible to the end-user.

To generate the sample size guidance needed to produce robust WHO cone data, two key parameters must be determined: (1) the smallest difference in mosquito mortality which can be reliably detected between treatment groups and (2) the variability associated with the endpoint, which may be multiple values if considering the different layers of nets, net pieces, and replicates. If the mortality difference observed in the subsequent results is below the predetermined minimum, or the observed variability exceeds the estimate, then the risk of a type I error (false positive) is unacceptably high and thus the results cannot be considered robust (Ioannidis, 2005). Furthermore, it is important to note that the power to detect a given mortality difference (e.g. a 10% difference) is dependent on the actual values being compared (e.g. 50% vs 60% is not the same as 80% vs 90%) due to mortality being truncated below 0% and above 100% thus variance is maximised at 50% (Mechan et al., 2024).

The effect size threshold (minimum mortality difference to be detected) and variability assumed should be set as a pragmatic trade-off to detect as small an effect as is relevant with the resources available. To aid this decision making, we present a data analysis pipeline for extracting variability estimates from existing data and identifying the sample sizes needed to detect different effect sizes, as has been demonstrated previously for WHO tube bioassays, (Mechan et al., 2024). Thus, rather than a prescriptive or universal sample size number, end-users can readily identify the experimental design suitable for their purposes and resources, with full knowledge of the assumptions made and the limits of interpretation.

To illustrate these concepts and findings, we use the 2013 WHO cone bioassay sample size guidance for phase I laboratory testing of ITNs (WHO, 2013) as a practical example of how the different levels of the sample size (nets, net pieces, replicates) interact to generate power. Given the increased flexibility of more recent WHO cone bioassay guidance (WHO, 2024c), we demonstrate how modifying the sample size impacts power. To support this non-prescriptive approach, we present an app-calculator to empower researchers to calculate a robust sample size for WHO cone bioassay experiments to test ITNs.

Methods

Literature review

To investigate the use of the WHO cone bioassay for the bioefficacy evaluation of pyrethroid ITNs we conducted a literature review of studies published between 1998 and 2021. The search was conducted in 2021 using the keywords "bio-efficacy" or "cone bioassay tests" or "Insecticide treated nets" and "long lasting insecticidal nets" and later updated in 2024 to include any subsequently published papers. We used inclusion criteria to identify studies with *Anopheles* mosquitoes that reported 24 hr mortality. The search identified 2,487 titles (148 from PubMed and 2,338 from PubMed Central), which we screened to identify relevant papers. Duplicate papers already found in PubMed were removed from the PubMed Central search results leaving a total of 632 papers for screening. Further screening for papers that met the inclusion criteria narrowed the results down to 55. The selected papers were thoroughly analysed, and specific details were recorded to ensure a comprehensive evaluation of study methodologies. This structured approach ensured that all relevant methodological details were consistently captured across the reviewed studies. The information which was extracted is detailed in Table 1.

Table 1. Summary of investigated parameters for literature review on cone testing methodologies.

Category	Parameter
Reference Details	Title
	Author
	Year of Publication
Guidelines	Referenced Guidelines

Table 1. *Continued*

Category	Parameter
Net Characteristics	Type of Nets Tested
	Source of Nets (Purchased/Other)
	Net Hanging Status (Hung Before Testing)
	Net Aging Duration (If Aged)
Storage & Transport	Net Transport Conditions
	Sample Storage Conditions
Testing Setup	Net Sample Piece Size Used
	Number of Samples per Net/Net Side
	Total Number of Nets Tested
	Inclusion of Negative Control Netting
	Sample Size
	Number of Cones or Replicates per Net Section
	Number of Mosquitoes per Cone
Mosquito Parameters	Age of Mosquitoes
Exposure Conditions	Exposure Time to Netting
	Temperature During Testing
	Humidity During Testing

Using WHO cone bioassay data from multiple laboratories to measure variability in mortality

Existing WHO cone bioassay data was obtained based on pyrethroid-susceptible *Anopheles gambiae* exposed to the same six new, unwashed pyrethroid-treated nets which were circulated around seven test centre laboratories. The purpose of evaluating this data was to identify an appropriate range of variability estimates for the simulation study.

Outcomes assessed were between-net and between-net piece variability (between-replicate variability could not be calculated in most cases due to data aggregation). Variability between parameters of interest and outcome (24 hr mosquito mortality) were analysed using generalised linear mixed models (GLMMs) with binomial link function within the 'lme4' package in R (version-4.3.3). To quantify variability in the outcome in each laboratory, random effect terms for 'net' (different nets of the same product), 'net piece' (different pieces from the same net) were included in the GLMMs. When considering overall net and net piece variability across the laboratories, 'Testing lab' (variable for different laboratory test centres) was considered as a fixed effect term as well as the random effects listed.

Statistical significance of model parameters was assessed using log-likelihood ratio tests, comparing the explanatory power of different models with and without the inclusion of a given parameter. Variability values associated with 'nets' and 'net pieces' were extracted from random effects estimates of the GLMM, with standard deviation used as the unit of variation. To visualise the spread of data between different nets, violin plots with a bee swarm function were created using the 'ggbeeswarm' package.

Simulation study

Here we use the simulation principles of Johnson et al. (2015) and the sample size framework of Mechan et al. (2024) to perform simulation-based power analysis. To briefly summarise, this simulation-based power analysis process for calculating sample size involves simulating a very large number of plausible datasets from variability estimates. Subsequently, each simulated dataset is analysed individually using generalised linear models (GLM) to determine how often an effect is detected. The aim is to establish the minimum sample size (i.e. number of nets, net pieces, and replicates) to reliably distinguish a difference in mean mosquito mortality between different treatment groups/ITN products. The threshold at which a sample size was deemed sufficient was 80% power (meaning an 80% chance of detecting the effect, per convention).

The power of different experimental designs was evaluated. A plausible range of combinations of effect sizes (differences in mean mosquito mortality), sample size (number of nets, net pieces, and replicates), and variability assumptions (informed by our analysis of existing data sets) were assessed (Table 2). For each unique combination of simulation

Table 2. Range of parameters used in the simulation study.

Parameter variable	Possible values
Mortality difference (effect size)	5-50%
Total number of nets of each treatment	2-10
Total number of net pieces of each treatment	1-5
Total number of replicates on each net piece	5-11
Net variability	0.10-1.60
Net piece variability	0.10-0.90
Replicate variability	0.10-0.90

parameters, 5,000 plausible datasets were simulated. Each simulated dataset was analysed using a binomial GLM with net type as fixed effect. Following this, p-values were extracted from each GLM. The power for each parameter combination was reported as the proportion of 5,000 analyses for which $p<0.05$. By comparing the power for different sample sizes, we identify the minimum number of samples (nets, net pieces, and replicates) needed to detect a given effect size (e.g. a 15% difference in mosquito mortality). Unless otherwise stated we report power for a given experimental design at the point of maximum uncertainty ('Umax'), meaning a 15% mortality difference represents the difference between 50% and 65%. To maximise accuracy of power estimates close to 80% (78-82%), the threshold at which a design is either sufficiently powered or not, we performed follow-up evaluations of these experimental conditions with 15,000 simulations. For the purposes of reporting here, power estimates values >79.5% were rounded up to 80%.

Evaluating the 2013 WHO guidance for phase I laboratory testing

As we cannot show all the possible designs for a WHO cone bioassay experiment here in this article, we focus the results on the practical example of the sample size recommended in the 2013 WHO guidance for phase I laboratory testing ("Guidelines for laboratory and field-testing of long-lasting insecticidal nets") (WHO, 2013). Phase I laboratory testing requires a sample size of four nets, one net piece, and ten replicates on each ITN treatment group (for a total of 200 mosquitoes per treatment group).

We evaluate the power of a WHO cone bioassay experiment to detect mortality differences with this fixed sample size, as well as showing the impact of increasing or decreasing sample size beyond this. This is not to say that the findings are limited to this context, but instead it is used for illustrative purposes. The open access associated WHO cone bioassay sample size Rshiny application (Cone Bioassay Sample Size app) allows the reader to investigate power for other plausible study designs. This is particularly beneficial for visualising the dynamic relationship between the different number of nets, net pieces, and replicates at the same time, which is challenging to communicate in a static format. This Rshiny application was created in R (version-4.3.3).

Results

Characterising the reporting and application of the cone method from the literature

The methodological details extracted from the selected papers identifies great inconsistency in how methods are referenced and how much methodological detail is reported, as well as the details of how the tests themselves were performed (Figure 1). Only some authors referenced guidelines in the Methods sections of the published studies and many referenced guidelines that were years out of date at the time of publication. Out of the 55 papers selected for analysis 15% did not reference any guidelines, while 5% referenced published papers instead of guidelines. Of the papers that did reference WHO guidelines or other associated documents, 4% used the 1998 document "Test procedures for insecticide resistance monitoring in malaria vectors, bio-efficacy and persistence of insecticides on treated surfaces: report of the WHO informal consultation, Geneva, 28-30 September 1998" (WHO, 1998), while 18% used the 2005 guidelines "Guidelines for laboratory and field-testing of long-lasting insecticidal nets" (WHO, 2005). There were no papers that referenced the 2006 guidelines "Guidelines for testing mosquito adulticides for indoor residual spraying and treatment of mosquito nets" (WHO, 2006) or the 2024 guidelines "WHO Prequalification of Vector Control Products Implementation guidance (Modules 3 and 5) Bioassay methods for insecticide-treated nets: Cone test" (WHO, 2024c). Only 7% referenced the 2011 guidelines "Guidelines for monitoring durability of long-lasting insecticidal nets under operational conditions" (WHO, 2011), while 47% referenced the 2013 guidelines "Guidelines for laboratory and field-testing of long-lasting insecticidal nets" (WHO, 2013), and only 2% referenced the 2016 guidelines "Test procedures for insecticide resistance monitoring in malaria vector mosquitoes: second edition" (WHO, 2016). A further 2% of studies referenced both WHO (2011) and WHO (2013) guidelines. In terms of how many years out of date the referenced guidelines were,

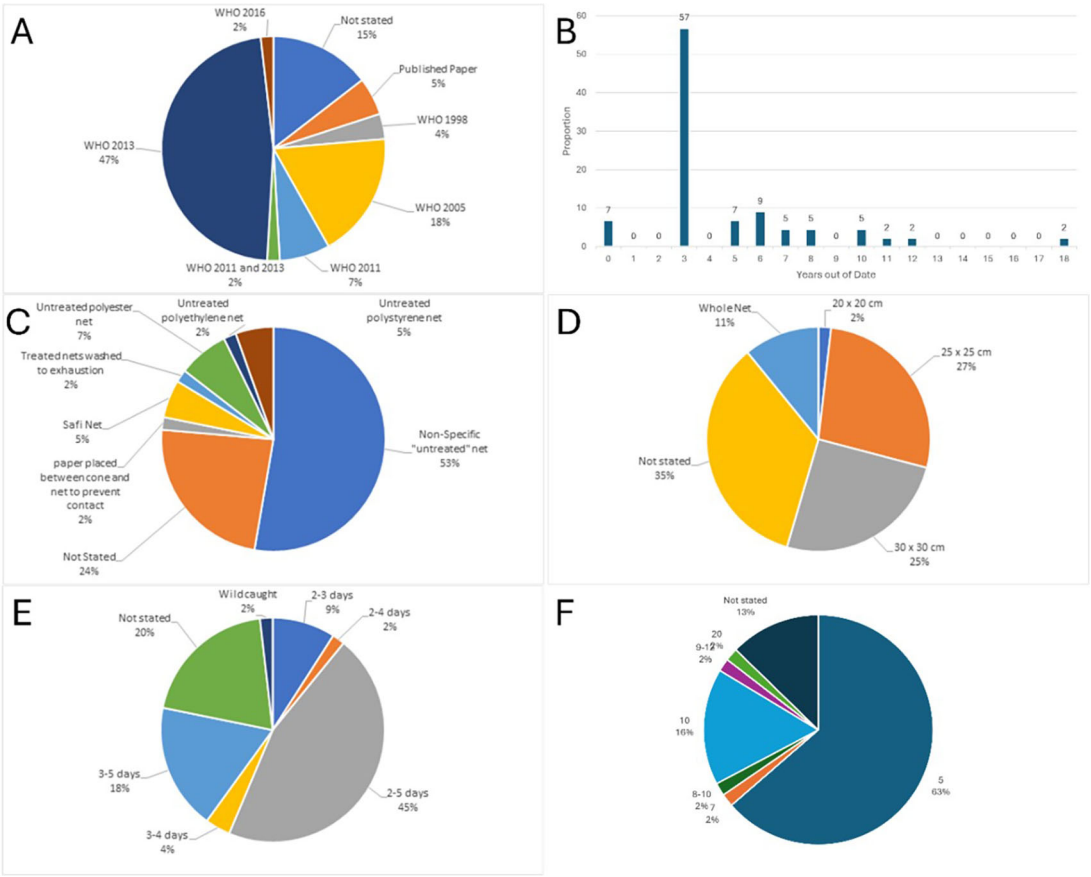

Figure 1. Standardisation in WHO cone bioassay testing. (A) Proportion of referenced guidelines used for cone bioassay testing. (B) Number of years since the referenced guidelines became outdated. (C) Types of netting used as a negative control. (D) Net sample piece sizes. (E) Testing age of mosquitoes. (F) Number of mosquitoes used per cone.

the selected papers varied widely, with 7% using guidelines that were up to date, and 57% used guidelines that were 3 years out of date. The remaining 36% of studies referenced guidelines that were 5-18 years out of date. There is a need for greater consistency in the referencing of guidelines and adherence to up-to-date guidelines to ensure the reliability and comparability of results across studies.

Methodological details extracted from the papers included the sources of the nets used in the studies, which were diverse and included direct sourcing from the chemical company, local markets, homes, research institutes, storage facilities, and the Ministry of Health. However often it was not stated at all where the nets were sourced from. Additionally, the storage conditions for the nets varied widely, with some studies not stating the storage conditions and others using variously foil, plastic bags, or envelopes to store the nets, sometimes in the fridge.

The size of the net sample used in the bioassay varied, with most studies using 25 × 25 cm samples, while others used similar sized pieces at 30 × 30 cm or 20 × 20 cm samples, or instead did not break down nets and instead did cone bioassays on whole nets. A total of 35% of papers did not state the size of net piece tested. They reported the total number of nets used, which varied from one to 932. For each net, they indicated the number of net pieces collected, with a range of one to ten net pieces per net. Additionally, they detailed the number of cones in each net piece, which ranged from one to 16 cones.

The papers provided a range of information regarding their mosquito sampling. The age of the mosquitoes used in the bioassay also varied, with the majority (45%) of studies using 2–5 day old mosquitoes, while others used 2-4 day (2%), 2-3-day (9%), 3-4 day (4%) or 3–5 day (18%) old mosquitoes. Some studies (20%) did not state the age at all, and some used wild caught mosquitoes of unknown age (2%).

The number of mosquitoes exposed per cone varied, with most studies using five mosquitoes per cone (73%), while others used seven (2%), ten (19%), eight to ten (2%), nine to 12 (2%) or 20 (2%) mosquitoes per cone, or did not state the number (13%). The negative control netting used in the studies varied, with non-specific "untreated" nets being the most common. Finally, papers typically specified the overall sample size for the study, which varied widely from 20 to 20,000 mosquitoes.

Determining appropriate variability assumptions for the simulation study

Variation in 24 hr mortality between nets (of the same product) and between net pieces (from the same net) in the WHO cone bioassay was quantified using testing data with the same samples across multiple laboratories. With pyrethroid-susceptible *An. gambiae* Kisumu exposed to a pyrethroid-only ITN the between-net variation differed across laboratories, with standard deviation (SD) values ranging from 0.55 to 2.00 (Figure 2 and Table 3). Variability between pieces from the same net also differed between laboratories but to a lesser extent (SD range: 0.08 to 0.85) compared to between nets (Table 3).

When measuring overall variability across laboratories by combining all data, the between-net SD was 0.641 and between-piece SD was 0.249 (across laboratories in Table 3). Replicate variability could not be directly assessed due to the data being aggregated. As in practice many laboratories may have substantially higher variability, to be conservative we take these summarised variability values as a pragmatic minimum/'floor' variability estimate. We use this floor estimate, as well as the range of values observed to establish a conservative set of three categories of variabilities:

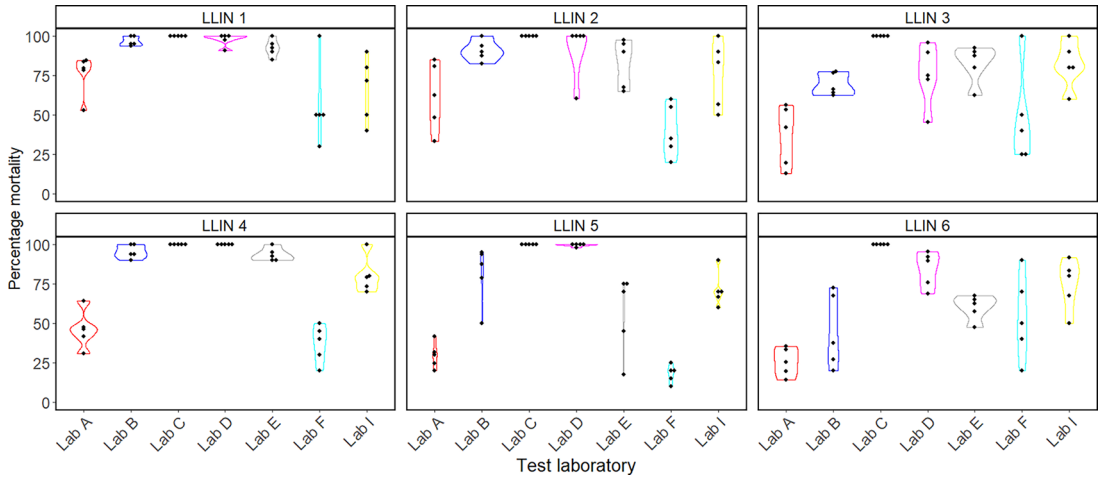

Figure 2. Percentage 24 hr mortality of ITNs tested across multiple laboratories. Plots highlighting 24 hr mortality of pyrethroid-only insecticide treated-nets (ITNs) (LLIN 1 to LLIN 6) samples tested across different laboratories (Lab A-F, I) with pyrethroid-susceptible An. gambiae s.s. mosquitoes. Black dots represent raw percentage mortality values from individual bioassays (one dot represents 25 mosquitoes).

Table 3. Mortality variability between-nets and between-net pieces across multiple laboratories.

Lab	Species tested	Net SD	Net piece SD
A	*An. gambiae*	1.023	0.087
B	*An. gambiae*	1.2365	0.3464
C*	*An. gambiae*	0	0
D	*An. gambiae*	1.9969	0.8516
E	*An. gambiae*	0.9837	0.6076
F	*An. gambiae*	0.5574	0.2956
I	*An. gambiae*	1.0748	0.3221

Variability of 24 hr mortality in pyrethroid-susceptible An. gambiae s.s. between nets of the same product and net pieces from each net (represented by standard deviation, SD).
*Lab C: variability values are 0 due to mortality being 100% across nets, thus these data were not used in determining simulation assumptions.

Table 4. Variability estimates for 24 hr mortality between-nets, between-net pieces and between-replicates.

Variability category	Between-net SD	Between-net piece SD	Between-replicate SD
'Moderate' (practical minimum)	0.60	0.30	0.30
'High'	1.20	0.60	0.60
'Very high'	1.60	0.90	0.90

Estimated 'moderate', 'high' and 'very high' variability values for 24 hr mortality (represented as standard deviation, SD) for nets of the same ITN product, net pieces from each net and replicates on each net piece. Values are based on quantified variability values across multiple laboratories (except for between-replicate SD, where variability is conservatively assumed to be the same as between-net piece).

'moderate', 'high', and 'very high' (Table 4). The preset variability assumptions are intended for use where robust estimates are not available prior to testing; it would be reasonable for researchers to assume lower variability for their own power calculations if indicated by robust laboratory-specific data.

Limits of interpretation of the fixed 2013 sample size guidance

Here we investigate mortality differences that can be reliably detected with different sample sizes using variability assumptions in the pre-set categories established above. We use the sample size guidance in the 2013 phase I laboratory testing documentation as a starting point, then demonstrate the impact of increasing or decreasing the number of samples beyond this. In the 2013 WHO guidance, sample size is fixed at four nets of each treatment group, with one piece from each net and ten replicate cones on this piece (for a total of 200 mosquitoes per treatment group).

As would be expected, the probability of detecting a difference increases as the values being compared against moves away from 50% (in either direction) (Figure 3). Thus 60% vs 70% requires less samples to reliably detect a given difference than 50% vs 60%. Assuming our 'moderate' estimate (between-net SD of 0.6, and a between-replicate SD of 0.3) the smallest difference that can be detected by this sample size guidance with, at least 80% power at the point of maximum uncertainty ('Umax', where the reference value is 50% mortality), is 17.5% (power: 90%). A 10% difference is

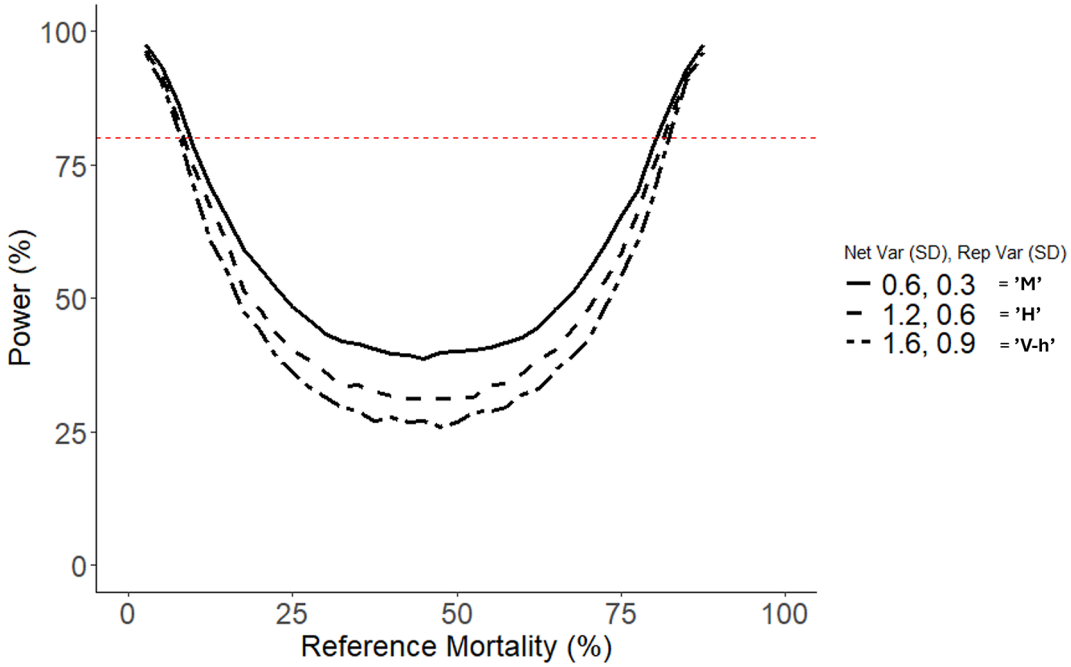

Figure 3. Power to detect 10% mortality differences between treatments in cone bioassays. Power to detect a 10% difference in 24 hr mortality between pyrethroid-treated nets at determined variability levels in the WHO cone bioassay following the 2013 WHO guidelines (four nets, one net piece per net and ten replicates per net piece). The levels of variability (Var) assessed are: 'moderate' ('M') standard deviation (SD) values = Net 0.6, Replicate (Rep) 0.3, 'high' ('H') SD values = Net 1.2, Rep 0.6 and 'very high' ('V-h') SD values = Net 1.6, Rep 0.9. The red dashed horizontal line represents 80% power threshold.

only reliably detectable with this sample size guidance and 'moderate' variability when reference mortality is 82.5% or above (i.e. 82.5% vs 92.5% can be reliably detected).

As shown in Figure 3, assuming greater variability leads to less power to detect a 10% difference. Assuming 'high' variability (between-net SD of 1.2, and a between-replicate SD of 0.6) still allows a 17.5% difference to be detected at the 'Umax' (power: 84%). However, if assuming 'very high' variability (between-net SD of 1.6, and a between-replicate SD of 0.9) the smallest mortality difference between treatment groups that can be detected by this sample size at 'Umax' is 20% (power: 89%).

How does changing the sample size impact power?

Here we explore how changes to the sample size guidance would impact the power to detect different mortality differences. We investigate power using 'moderate' and 'high' variability estimates mentioned above, assessing how power decreases as the variability increases.

Net number

'As outlined above the smallest difference that can be reliably detected with the 2013 WHO guidance at the 'Umax', with 'moderate' net variability assumption, is 17.5% (power: 90%) (Figure 4A). If instead just three nets were used, a 17.5% difference could still be detected (power: 80%). Detecting a 15% difference requires five nets (power: 88%), with a 10% difference outside the scope of our simulations (ten nets: 79% power). However, as reference mortality approaches 100%, a 10% difference becomes more feasible to detect with four nets. The threshold at which a 10% difference becomes detectable (under the same net piece, replicate, and variability assumptions) is approximately 82.5%.

Under our 'high' net variability assumptions, a 20% difference can be reliably detected at the 'Umax' using three nets (power: 84%), and a 15% difference with six nets (Figure 4B). A 10% difference is outside the scope of simulations but becomes detectable with four nets (as per the 2013 WHO guidance) when reference mortality is 82.5% or higher.

Net piece number

The power to reliably detect a difference between treatment groups increases with the number of net pieces, with the largest increase from one to two net pieces but the additional benefit diminishing noticeably above three net pieces (Figure 5A). We fix the number of nets at four and replicates at ten as per the 2013 WHO guidance.

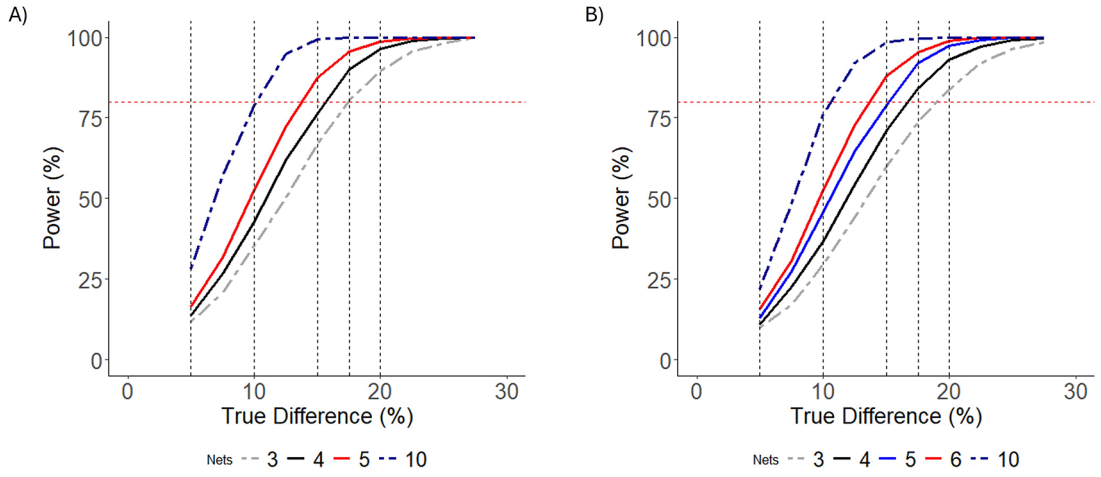

Figure 4. Power to detect mortality differences between treatments when alternating nets. The power to detect different 24 hr mortality effect sizes between pyrethroid-treated nets considering alternate numbers of nets, when net variability and replicate variability is assumed 'moderate' (A) and is assumed 'high' (B). Black solid line represents power of the 2013 WHO guidance (four nets) across different mortality differences at the maximum point of uncertainty ('Umax'). 'Moderate' standard deviation values: between-net = 0.6 and between-replicate = 0.3, 'high' standard deviation values: between-net = 1.2 and between-replicate = 0.3 keeping the number of net pieces fixed at one and number of replicates fixed at ten as per the 2013 WHO guidance. Red dashed horizontal line represents 80% power threshold, black dashed vertical lines represent true mortality difference of 5%, 10%,15%, 17.5% and 20%.

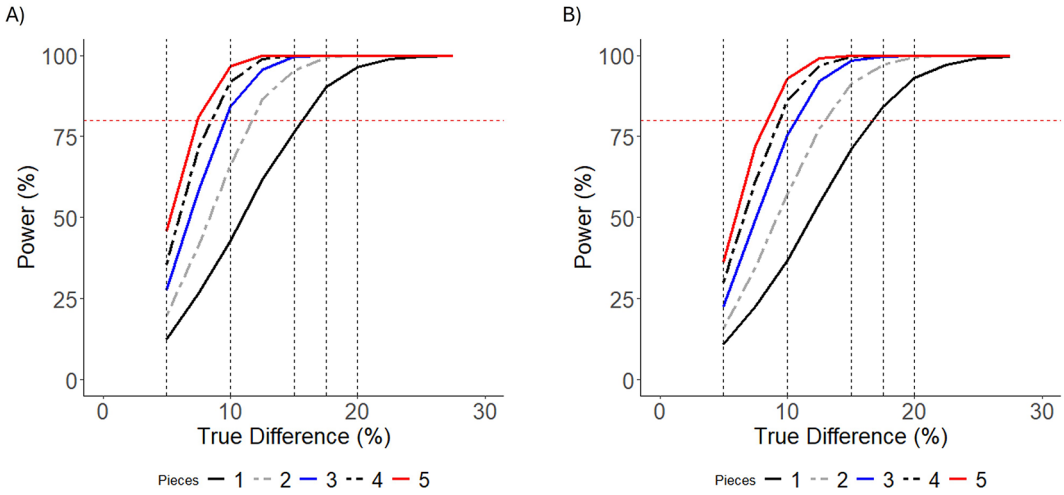

Figure 5. Power to detect mortality differences between treatments when alternating pieces. The power to detect different 24 hr mortality differences between pyrethroid-treated nets across different numbers of net pieces, when net variability and net piece variability is assumed 'moderate' (A) and assumed 'high' (B). Black solid line represents power of the 2013 WHO guidance (one net piece) across different mortality differences at the maximum point of uncertainty ('Umax'). 'Moderate' standard deviation values: between-net = 0.6, between-net piece = 0.3 and between-replicate = 0.3, 'high' standard deviation values: between-net = 1.2, between-net piece = 0.6 and between-replicate = 0.3, keeping the number of nets fixed at four and number of replicates fixed at ten as per the 2013 WHO guidance. Red dashed horizontal line represents 80% power threshold, black vertical lines represent true difference % at 5%, 10%,15%, 17.5% and 20%.

With the standard sample of one net piece, a 17.5% difference can be detected with the 2013 WHO guidance sample size at 'Umax' when the variability is 'moderate' (Figure 5A). A 15% difference can be detected using two net pieces (power: 95%) under these assumptions. A 10% difference can be detected with three net pieces (power: 84%).

With 'high' variability assumptions (keeping replicate variability 'moderate'), the 2013 WHO guidance can detect a 17.5% difference at 'Umax' (Figure 5B). A 15% difference can be detected with two net pieces (power: 91%) and 10% with four net pieces (power: 86%).

Replicate number

Assuming a 'moderate' level of net and replicate variability, the 2013 WHO guidance can detect a 17.5% difference between treatment groups (power: 90%) (Figure 6A). Keeping the number of nets fixed at four and pieces fixed at one, detecting a 15% and 10% difference is outside of the range of scope of simulations at 'Umax' when altering the number of replicates. However, if the reference mortality is just slightly higher than the 'Umax' (55%), a 15% difference becomes detectable with ten replicates. Furthermore, a 10% difference becomes detectable when the reference mortality is 82.5%. It is noteworthy that under these assumptions a 20% difference is still detectable with fewer replicates than the 2013 WHO guidance, with just six replicates giving 82% power.

Assuming a 'high' variability, the 2013 guidance can detect a 17.5% difference between treatment groups (power: 84%) (Figure 6B). Reliably detecting a 10% and 15% difference under the 2013 WHO guidance is outside of the range of scope of simulations at 'Umax'. However, if the reference mortality threshold is increased to approximately 62.5% a 15% difference becomes detectable. A 10% difference becomes detectable at a reference mortality at approximately 82.5% or higher. As with 'moderate' variability assumptions, at 'high' variability with fewer replicates than the 2013 WHO guidance, a 20% difference is detectable with just seven replicates giving 82% power (Figure 6B).

Optimising sample size of nets, net pieces and replicates to minimise the number of mosquitoes needed for testing

As shown above there is a trade-off between the ability to detect small differences and the size of the sample needed. For researchers performing a WHO cone bioassay the limiting resource is typically testing-age female mosquitoes. Identifying the optimal trade-off between detecting a sufficiently small difference and the mosquitoes needed to achieve this can minimise the resources needed for an experiment. Excessive effort can be avoided if identifying very small

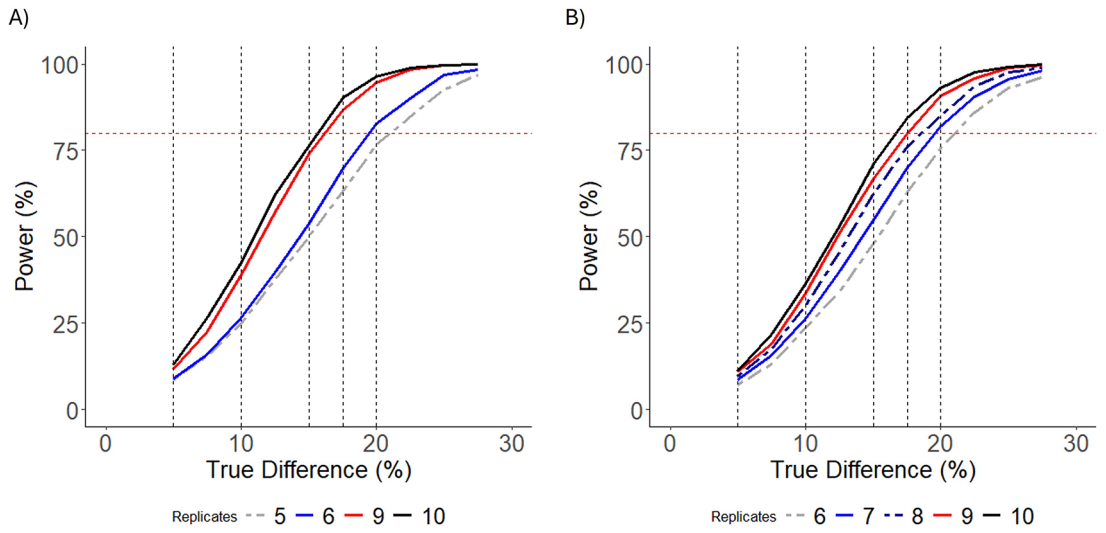

Figure 6. Power to detect mortality differences between treatments when alternating replicates. The power to detect differences in 24 hr mortality between pyrethroid-treated nets across different numbers of replicates when net variability and replicate variability is assumed 'moderate' (A) and is assumed 'high' (B). Black solid line represents power of the 2013 WHO guidance (ten replicates) across different mortality differences at the maximum point of uncertainty ('Umax'). 'Moderate' standard deviation values: between-net = 0.6 and between-replicate = 0.3, 'high' standard deviation values: between-net = 1.2 and between-replicate = 0.6, keeping the number of nets fixed at four and numbers of net pieces fixed at one as per the 2013 WHO guidance). Red dashed horizontal line represents 80% power threshold, black vertical lines represent true difference % at 5%, 10%, 15%, 17.5% and 20%.

differences is not necessary to answer the research question, and wasted effort avoided by not performing experiments underpowered to answer the question.

As shown above the 2013 WHO guidance to use 200 mosquitoes per treatment group is powered to detect a 17.5% difference at 'Umax', assuming 'moderate' to 'high' variability. However, decreasing the number of replicates from ten to six still gives sufficient power to detect a 20% difference under the same conditions, reducing the number of mosquitoes required from 200 to 120. This highlights the trade-offs between power and sample size.

Increasing the number of net pieces per net has the greatest mosquito cost, yet using only a single net piece means a core source of variability (between-net piece) is not accounted for. Up to this point, this paper has considered the impact of varying net, net piece, and replicate sample sizes independently, yet in practice all three may be adjusted at the same time to optimise power. We show here that for different research objectives in terms of the effect size to be detected, there are optimal combinations of net, net piece, and replicates that maximise power for the fewest mosquitoes (refer to Table 5).

Table 5. Minimal sample sizes per treatment to detect different mortality differences across variability estimates.

	Smallest effect size (24 hr mortality) detected with 80% power		
Assumed Variability (net, net piece, replicate)	**10%**	**15%**	**20%**
'Moderate' (0.6, 0.3, 0.3)	540 mosquitoes (4n/3p/9r, pwr = 80%) (6n/2p/9r, pwr = 80%)	240 mosquitoes (3n/2p/8r, pwr = 80%) (4n/2p/6r, pwr = 80%) (6n/2p/4r, pwr = 81%)	140 mosquitoes (2n/2p/7r, pwr = 83%)
'High' (1.2, 0.6, 0.6)	700 mosquitoes (7n/2p/10r, pwr = 80%)	300 mosquitoes (3n/2p/10r, pwr = 80%) (4n/3p/5r, pwr = 80%) (5n/3p/4r, pwr = 80%) (6n/2p/5r, pwr = 80%)	180 mosquitoes (3n/2p/6r, pwr = 84%) (3n/3p/4r, pwr = 84%)

Table 5. *Continued*

	Smallest effect size (24 hr mortality) detected with 80% power		
Assumed Variability (net, net piece, replicate)	**10%**	**15%**	**20%**
'Very high' (1.6, 0.9, 0.9)	825 mosquitoes (5n/3p/11r, pwr = 80%)	350 mosquitoes (5n/2p/7r, pwr = 80%)	200 mosquitoes (2n/2p/10r, pwr = 82%) (4n/2p/5r, pwr = 82%) (5n/2p/4r, pwr = 82%)

The minimum number of mosquitoes and associated sample(s) of nets, net pieces and replicates per treatment group requiredin a WHO cone bioassay to detect different effect sizes across 'moderate', 'high', and 'very high' variability at the maximum point of uncertainty ('Umax'). Target power = 80%. (n=number of nets/p=number of net pieces/r=number of replicates, pwr= power estimate).

Supporting sample size calculations for the WHO cone bioassay with a user-friendly application

To allow end-users to readily design their own experiments with WHO cones, we provide a browser-based Rshiny application. This app allows researchers to take advantage of the simulation-based framework outlined above, without the need for access to computer programmes or hardware. Users can input their minimum effect size and level of variability (for net, net piece, and replicate number), with the output all of the viable experiments they could perform that would provide robust data. Additionally, the app provides an interactive visualisation of how power changes with different numbers of net, net pieces, and replicates, across a fine-scale range of variability values (allowing tailoring to a laboratory's level of variation). Click here for access to the application for interactive visualisation to explore other combinations of sample size to support experimental design for the WHO cone bioassay: Cone Bioassay Sample Size app https://fmechan1.shinyapps.io/who_cone_app/.

Discussion

The core pragmatic question when designing a WHO cone experiment to compare ITNs is: "what is the smallest difference between groups I can reliably detect with the resources available to me?". The mortality difference detectable is dependent on sample size and the variability of the outcome, with differences smaller than the predetermined minimum not a robust finding, regardless of reported p-values. Here we use existing data to demonstrate how power to detect mortality differences in WHO cone bioassays changes with sample size, using the 2013 WHO phase I laboratory studies guidance as a starting point and expanding this to represent the more flexible 2024 guidance for WHO PQ of Vector Control Products Implementation (Modules 3 and 5, Bioassay methods for insecticide-treated nets: Cone test) (WHO, 2013, WHO, 2024c). We show that across our range of plausible variability estimates and sample sizes, general rules of thumb emerge. We show that a 20% mortality difference is readily detectable in almost all practical circumstances (even with very high variability assumptions) yet that conversely a 10% difference is challenging to detect in most practical circumstances, though more feasible when reference mortality is >80%. This leaves 15% difference as an achievable threshold for an effect size that can be detected across the entire range of reference values with sample sizes only slightly larger than the 2013 WHO guidelines. With the level of variability we measured in existing data from multiple laboratories, mortality differences smaller than 10% would not be a plausible goal unless the true values were very close to 100%. Consequently, by defining the minimum effect size in advance, the optimal sample size can be implemented thus avoiding resources being wasted.

As with previous simulation studies on WHO bioassays (Mechan et al., 2024), we demonstrate that statistical power is dependent not just on the absolute mortality difference itself (e.g. 10%) but on the actual values being compared (e.g. 50% vs 60%), with power increasing as the values move away from 50% (towards 0% or 100%). We also observe here, as highlighted previously for WHO tubes (Mechan et al., 2024), that increasing sample size has diminishing returns in terms of power, with each additional net, net piece, or replicate having a smaller impact on power than the last. Thus, sample size calculation is always an optimisation problem.

Here we highlight the importance of incorporating reasonable assumptions about variability when designing experiments. It is important for a test laboratory to have either a robust estimate of their variability or, if estimates are unavailable, to be sufficiently conservative in their assumptions. In the absence of available variability estimates, we would recommend using the 'high' estimate (net SD = 1.2, net piece SD = 0.6, and replicate SD = 0.6), which we have made the default in the associated app. Additionally, we advise against testing with only a single net piece per net, as this leaves a key source of variability unaccounted for. However, this does not necessarily need to come at the expense of greatly increased sample size as we show that reducing the number of replicates per net piece is an acceptable trade-off in most circumstances.

To minimise the sample size needed, it is important to reduce the variability of the bioassay data you are generating as much as possible. Here, our review of published studies using WHO cone bioassays highlighted widespread inconsistencies in the application of WHO cone bioassay methodology. For example, the use of outdated guidelines—some over a decade old—and the absence of any cited guideline in 15% of studies revealed a critical gap in standardisation. These inconsistencies extend to procedural details, such as the wide range of mosquito ages, net sample sizes, and numbers of mosquitoes per cone, all of which influence bioassay outcomes. The literature review underscores that these methodological discrepancies likely exacerbate the variability observed in mortality data and limit the comparability of results across studies. By linking these findings to the need for variability-driven sample size calculations, the review highlights how more consistent adherence to up-to-date, standardised protocols could reduce variability, improve data reliability, and ensure that studies are appropriately powered to detect meaningful differences in mosquito mortality outcomes.

Conventional formula-based power calculations are inflexible to multiple sources of variation, instead aggregating this together into an overall estimate of variability. However, WHO cone bioassays have multiple defined sources of variation, arranged in a hierarchical structure of nets, net pieces, and replicates. These different levels of variability can (and do, as we have demonstrated) have different amounts of variability associated with them that influence the optimal experimental design. For example, if there is high variability between net pieces then this is more efficiently addressed by testing more pieces per net rather than adding additional nets.

This study focused on relatively new ITN samples, that have been tested repeatedly but not delivered to, or deployed in, the field. We observed that most variability was between nets, with lower variation between net pieces from the same net. However, for nets that have been used in the field, such as in durability studies, it is reasonable to assume that variability between both nets and net pieces from the same net would be higher than observed here. Consequently, for these power estimates to be extended to post-deployment nets it is essential to adjust the variability values appropriately using relevant data from field nets, estimated using the same analysis methods described here. Furthermore, as we lacked sufficient data for estimating variability between replicates on the net piece, we conservatively assumed it to be as variable as different net pieces. Additional data on between-replicate variability would allow more precise assumptions. While here we focused on variability in 24 hr mortality, knockdown is also a common endpoint in the WHO cone bioassay. To be applicable to knockdown, researchers would require estimates of knockdown variability.

The precision of simulation-based power estimates is dependent on the number of simulations. Here we use 5,000 simulations to assess power, followed by a 15,000-simulation confirmation to assess power for estimates close to 80% (78-82), to minimise misclassification. Furthermore, given we conducted effect size in increments of 2.5% in the simulation analyses in this study, when it is stated that the minimum reference mortality for detecting a 10% difference is 82.5%, we are more accurately referring to a range of 82.5-85.0.

The browser application linked here allows users to use preset variability assumptions to design WHO cone experiments or manually input variability values relevant to their laboratories. By removing the need for software, programming knowledge, or computer hardware, this tool provides democratised access to powerful simulation methods for WHO cone experiments. By using a common framework, supports robust uniformity and comparisons between studies.

This simulation framework could be readily applied to using WHO cone bioassays for testing IRS products, given well-supported estimates of variability in that context. Furthermore, the simulation-based power analysis framework outlined is readily adaptable to durability studies, given plausible assumptions about variability on field-collected nets. Moreover, the process could be adapted for WHO bottle and WHO tunnel tests, as has been done previously for WHO tube assays (Mechan et al., 2024).

Data availability

Underlying data

This project contains the following underlying data: https://doi.org/10.5281/zenodo.15130209 (Harvey and Mechan, 2025b).

This project contains the following underlying data:

- Anonymised_Multilab_An.gambiae_only_dataset.csv

Software availability

Source code available from: https://github.com/i2i-Data-Repository/WHO-cone-bioassay-sample-size.

Archived source code is available from: https://doi.org/10.5281/zenodo.15130413 (Harvey and Mechan, 2025a).

Acknowledgements

We thank the following institutions for their data contribution to support this work: CREC/LSHTM Collaborative Programme (CREC)/PAMVERC Benin; Centre for Research in Infectious Diseases (CRID); Noguchi Memorial Institute of Medical Research (Vestergaard Noguchi Vector Labs (VNVL)); KCMUCo- PAMVERC, Moshi Tanzania; Institut de Recherche en Sciences de la Santé (IRSS); Kenya Medical Research Institute-Centre for Global Health Research (KEMRI-CGHR)/Research World Ltd and Centre Suisse de Recherches Scientifiques en Côte d'Ivoire (CSRS).

References

Bhatt S, Weiss D, Cameron E, *et al.*: **The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015.** *Nature.* 2015; **526**: 207–211.
PubMed Abstract | Publisher Full Text | Free Full Text

Hancock PA, Hendriks CJ, Tangena J-A, *et al.*: **Mapping trends in insecticide resistance phenotypes in African malaria vectors.** *PLoS Biol.* 2020; **18**: e3000633.
PubMed Abstract | Publisher Full Text | Free Full Text

Harvey G, Mechan F: **Cone Bioassay Simulation Calculator.** *Zenodo: Zenodo.* 2025a. [Accessed].
Publisher Full Text

Harvey G, Mechan F: Existing WHO cone bioassay data from multiple labs (Version 1). [Data set]. *Zenodo: Zenodo.* 2025b. [Accessed].
Publisher Full Text

Ioannidis JP: **Why most published research findings are false.** *PLoS Med.* 2005; **2**: e124.
PubMed Abstract | Publisher Full Text | Free Full Text

Johnson PC, Barry SJ, Ferguson HM, *et al.*: **Power analysis for generalized linear mixed models in ecology and evolution.** *Methods Ecol. Evol.* 2015; **6**: 133–142.
PubMed Abstract | Publisher Full Text | Free Full Text

Mbuba E, Odufuwa OG, Moore J, *et al.*: **Multi-country evaluation of the durability of pyrethroid plus piperonyl-butoxide insecticide-treated nets: study protocol.** *Malar. J.* 2023; **22**: 30.
PubMed Abstract | Publisher Full Text | Free Full Text

Mbwambo SG, Bubun N, Mbuba E, *et al.*: **Comparison of cone bioassay estimates at two laboratories with different Anopheles mosquitoes for quality assurance of pyrethroid insecticide-treated nets.** *Malar. J.* 2022; **21**: 214.
PubMed Abstract | Publisher Full Text | Free Full Text

Mechan F, Praulins G, Gillespie J, *et al.*: **Power calculation for mosquito bioassays: Quantifying variability in the WHO tube bioassay and developing sample size guidance for the PBO synergism assay using a Shiny application.** *Gates Open Res.* 2024; **8**: 96.
PubMed Abstract | Publisher Full Text | Free Full Text

Moyes CL, Athinya DK, Seethaler T, *et al.*: **Evaluating insecticide resistance across African districts to aid malaria control decisions.** *Proc. Natl. Acad. Sci.* 2020; **117**: 22042–22050.
PubMed Abstract | Publisher Full Text | Free Full Text

Praulins G, Murphy-Fegan A, Gillespie J, *et al.*: **Unpacking WHO and CDC Bottle Bioassay Methods: A Comprehensive Literature Review and Protocol Analysis Revealing Key Outcome Predictors.** *Gates Open Res.* 2024; **8**: 56.
PubMed Abstract | Publisher Full Text | Free Full Text

Ranson H, N'guessan R, Lines J, *et al.*: **Pyrethroid resistance in African anopheline mosquitoes: what are the implications for malaria control?** *Trends Parasitol.* 2011; **27**: 91–98.
PubMed Abstract | Publisher Full Text

Skovmand O, Dang DM, Tran TQ, *et al.*: **From the factory to the field: considerations of product characteristics for insecticide-treated net (ITN) bioefficacy testing.** *Malar. J.* 2021; **20**: 1–13.
Publisher Full Text

WHO: *Test procedures for insecticide resistance monitoring in malaria vectors, bio-efficacy and persistence of insecticides on treated surfaces: report of the WHO informal consultation, Geneva, 28-30 September 1998.* World Health Organization; 1998.

WHO: *Guidelines for laboratory and field testing of long-lasting insecticidal mosquito nets.* World Health Organization; 2005.

WHO: *Guidelines for testing mosquito adulticides for indoor residual spraying and treatment of mosquito nets.* World Health Organization; 2006.

WHO: *Guidelines for monitoring the durability of long-lasting insecticidal mosquito nets under operational conditions.* World Health Organization; 2011.

WHO: *Guidelines for laboratory and field-testing of long-lasting insecticidal nets.* World Health Organization; 2013.

WHO: *Test procedures for insecticide resistance monitoring in malaria vector mosquitoes.* World Health Organization; 2016.

WHO: *Implementation guidance (Module 3). WHO Prequalification of Vector Control Products: Bioassay methods for insecticide-treated nets: Regeneration study for ITN fabric.* World Health Organisation; 2024a.

WHO: *Implementation guidance (Module 3). WHO Prequalification of Vector Control Products: Bioassay methods for insecticide-treated nets: Wash resistance study for ITN fabric.* World Health Organisation; 2024b.

WHO: *Implementation guidance (Modules 3 and 5). WHO Prequalification of Vector Control Products: Bioassay methods for insecticide-treated nets: Cone test.* World Health Organisation; 2024c.

WHO: *World malaria report 2024: addressing inequity in the global malaria response.* World Health Organization; 2024d.

WHO: *Prequalified Vector Control Products – WHO.* World Health Organisation; 2025. [Accessed 01/04/2025 2025].
Reference Source