Consolidated responses to questions and comments received in response to public posting of a protocol and notice of intent to determine non-inferiority of insecticide-treated nets and IRS

Question: The definition of standard net and first in class are good, but they ignore the history of current products. If ITNs are taken as a class including LLIN and dipped nets, then the first in class was a dipped net and LLIN were not before recently used in controlled, randomized cluster trials. That makes the standard net and the first in class the same net. If we now split – which was not done when LLIN were recommended – then the first LLIN to get a preliminary and later full recommendation was Olyset. But using Olyset as first in class for LLIN is problematic, because it was originally not evaluated in phase I and II tests, these were not yet defined. When Olyset later were tested in Phase I, it failed and led to the development of the tunnel test for the reason that permethrin is very repellent. However, tunnel test are now used as fall back tests and net often pass in tunnel tests even they failed in cone test also for other pyrethroids than permethrin. It is probably safer to say that tunnel test work at lower insecticide dosages than cone tests due to a potentially much, much longer exposure time. So, some clarification is needed and I suggest a solution below.

For IRS products, what is first in class for e.g. OP products (it is explained in the text that preferably first in class should be same insecticide class).

Possible solution: It would be helpful and very practical if WHO simply made a table for ITN, LLIN and IRS that gives the name of the products that are first in class and that are standards (the latter if possible).

Answer: We shall develop and include a table on the first-in-class products that have established a new intervention class if/when it has been determined that non-inferiority testing is a routine procedure in the evaluation of vector control products. A decision on the role of non-inferiority testing will only be made once data on pyrethroid-PBO nets have been generated and assessed by an Evidence Review Group, and their recommendations have been assessed by the Malaria Policy Advisory Committee.

It is acknowledged that as new products come online with modes of action other than those of fast acting neurotoxicity, new/modified test procedures to evaluate them will need to be developed. Mention of the need to investigate alternative test procedures is made on the final page of the protocol.

Question: Chapter 3.2 and 3.3, primary and secondary endpoints, IRS. The text is not clear. 3.3 says that residual effect is a secondary endpoint, and 3.2 says that mortality during duration of the study is a primary endpoint. I hope it is meant in 3.2. that residual effect is determined using cone test with fully susceptible mosquitoes over the time claimed by the product and with 80 % as threshold. Then the secondary endpoint is with free flying mosquitoes and 50 % endpoint and this endpoint will of course depend on resistance mechanisms and intensity in the study area, so these must be known to give meaningful data. 3.3 is well explained, but 3.2. is not. However, 50 % control is probably not going to have epidemiological impact and I do not understand this threshold.

Answer: The primary endpoint is total mosquitoes killed over a specific time period. This means that products that may not reach the current 80% WHO mortality threshold can still be assessed and compared to a product with a similar claim of residual activity in areas where insecticide resistance may mean that 80% mortality cannot be reached with wild mosquito populations. The secondary endpoint is the duration of efficacy measured by cone test. This is still required to assess the duration of efficacy of products as they are applied under operational conditions-much in the same way that ITN durability is measured under user conditions. It is, however, acknowledged that for some actives with non-neurotoxic modes of action it may be necessary to develop new methods of measuring the residual activity of these compounds as an alternative to cone bioassays. The text has been modified to remove mention of 50% efficacy.

Question: Chapter 5.1.1. needs some clarification with regards to the mention of "products with synergists should also be evaluated without the synergists". I remember that the manufacturer of Olyset Plus did not agree to produce the product without PBO saying that to make that, it would be a totally new product development. The evaluation was instead conducted using Olyset and Olyset+, and Olyset+ became a first in class for pyrethroid+PBO nets, but WHO phase I and II studies as well as the field study of Protopopoff *et al.*[1] showed that the two nets are very different in permethrin "bleeding rate" and thus in the amount of permethrin exposed to mosquitoes and to wash off (quote from WHOPES 2012 working group report). It would therefore be good if WHO force the producers of a candidate to provide the same net just without the synergist. Even that might give a slightly different release rate and surface concentration, it is not going to be such a big difference (20 times) as between Olyset and Olyset Plus.

Answer: There is no necessity to demonstrate difference between constitute active ingredients of a product; a new product looking to be covered by existing policy for an established intervention class should be compared as a whole against the first-in-class product that established the class, or another suitable comparator as indicated by WHO.

Question: Section 5.3.1. mentions nets washed 20 times as surrogates for three-year-old net. This is in principle the philosophy of lab and semi field tests, but it has not been verified. In these two test protocols, wash off is the reason for loss of insecticide, but in three years use in most African countries, such nets are not washed more than three to six times in three years and evaporation and rubbing off are main causes for insecticide loss (see eg Kilians study of Permanet 2). Evaporation depends on vapor

1 Protopopoff N, Mosha JF, Lukole E, Charlwood JD, Wright A, Mwalimu CD, Manjurano A, Mosha FW, Kisinza W, Kleinschmidt I, Rowland M. (2018) Effectiveness of a long-lasting piperonyl butoxide-treated insecticidal net and indoor residual spray interventions, separately and together, against malaria transmitted by pyrethroid-resistant mosquitoes: a cluster, randomised controlled, two-by-two factorial design trial. Lancet. 391(10130):1577-1588.

pressure and temperature and vapor pressure is low for pyrethroids but higher for most other insecticides. So, if the surrogate may work out reasonable for pyrethroid nets, it is probably less correct for other insecticides that may have 10 to 1000 times higher vapor pressure.

Answer: We are aware of the limitation of using washing to simulate the loss of bio-efficacy under field conditions, and do not see the washing as a replacement for long-term durability studies. However, while ITN are being aged under field conditions, which takes time, it is thought that 20x washing, as is currently outlined in ITN testing guidelines, will at least give an approximation of how different nets compare as a result of 'aging', despite the limitations of this approach. As the guidelines for simulating aging of other chemistries are updated, these recommendations will be refined.

Question: Section 5.3.3. mentions washing interval should not be determined in the current regeneration time method. If mortality reach 100 % after one day because wash off was low and start dosage was high, it does not mean that regeneration has stopped, it just means that the mortality study cannot show it continues. It has been suggested that median knock down time should be used as an alternative method; chemical methods can also be used; the easiest may be to replace the fully susceptible strain used now with a strain that provides around 80 % mortality with the test product before and after washing. Doing that, the mortality does not go into saturation and we are not so close to 100 % that sample variation will kill the statistic.

Answer: These inputs will be taken into consideration as part of the overall revision of WHO testing guidelines. The non-inferiority protocol is not designed to question or revise these procedures, but to guide initial non-inferiority studies to generate data based on which the value of non-inferiority determination as part of the WHO evaluation process for new tools can be assessed.

Question: Section 5.4.2, study duration. Can this be made more flexible for IRS? Say a producer has a new product that he thinks will provide residual effect (primary endpoint as defined above) for two years, then he must sit on his bands for the two years of the study. Would it be possible for him to first claim one year, get a preliminary recommendation after that year, but let the study continue till the efficacy falls below 80%? Then get a new preliminary recommendation for say 18 months or two years or whatever the study shows.

Answer: WHO no longer provides interim or preliminary recommendations for vector control products. However, changes to the product's use and/or claims are indeed feasible once data to support these have been generated. The above scenario is therefore perfectly feasible under the WHO evaluation pathway, whereby a product would initially be assessed again the WHO specified testing thresholds and be prequalified if it meets these. The label claim can subsequently be amended if additional data show that the product does have a longer residual life than that initially demonstrated.

Question: BASF´s Interceptor® G2 is a first in class for dual nets containing pyrethroids and a new a.i. According to the guideline, this first in class will need to be included in all testing for non-inferiority. How will the supply of these first in class products be organized? Will WHO purchase Interceptor® G2 nets needed for these tests?

Answer: Whether Interceptor® G2 will be entitled a first-in-class product will be depend on the assessment of data on its epidemiological impact by WHO. In the absence of a policy recommendation for this class of net, we would like to discourage the use of the term 'first-in-class' in the context of Interceptor G2 and any other vector control products under evaluation.

The supply of first-in-class products to enable non-inferiority studies has been recognized as an issue that will need to be addressed if non-inferiority studies become a routine requirement of the evaluation process for vector control interventions. At present this has not been decided. To avoid the frequent re-evaluation of a first-in-class products in non-inferiority studies of second-in-class products, discussions are ongoing as to whether a second-in-class product could become an alternative active comparator for non-inferiority evaluation of new products once it has been shown to be non-inferior to the first in class product. It is envisaged that this arrangement would then make it easier to source an active comparator product for non-inferiority studies.

Question: In the glossary, the second in class for ITN is described as: In the case of ITNs, it second-in-class products also contain the same AI(s) as the first-in-class product. It is not mentioned how the technology for treatment is included. Will incorporated PE nets and coated PET nets be considered as equivalent under this rule? Meaning, if a first-in-class net is based on PET technology, is it a valid comparator for a second-in-class based on PE technology?

For dual nets containing a pyrethroid plus a non-pyrethroid insecticide, does this class also comprise an a.i. other than chlorfenapyr? Chlorfenapyr is the new a.i. on the first-in-class net Interceptor® G2. Or will a new product class be opened for every new non-pyrethroid a.i.?

Answer: Regarding technology for treatment it is presently envisaged that incorporated PE and coated PET nets would indeed be compared to each other, as is the case for the envisaged studies of pyrethroid-PBO nets. Regarding classes of ITNs, new nets with a pyrethroid and a different a.i. than chlorfenapyr are currently considered as a different class, and are required to undergo epidemiological studies to demonstrate public health value.

Question: On page 8, only the cone test is requested as a bioassay during a ITN hut trial. Can the tunnel test be used as alternative in case the cone test is not suitable to show the performance of an a.i. in case e.g. it is a metabolic toxin?

Answer: Investigators are encouraged to use other test methods **in addition to**, not as a replacement of the cone test.

Question: Aging of candidate nets in the field and assessing field durability for ITN. On page 8, the surrogate for a large field trial by using nets aged for three years is described.

- How will this aging be done for the candidate net as it is not part of a large field trial or does the second in class also need to do a three-year large field trial?
- How will nets for the active comparator (first-in class) and candidate LLIN be chosen from a large field trial to make sure that these nets are representative?

- Has this type of assays been done in the past and has it proven to give valid information?

The concern here is that in village trials the nets selected might be quite diverse concerning remaining a.i. content and number of holes. These two parameters are quite dependent on how the nets are treated and handled during these three years. The *WHO Guidelines for laboratory and field-testing of long-lasting insecticidal nets* is given as a reference describing the sampling process. These guidelines describe for the large-scale trial to sample enough LLINs to power the study. As an example, it is mentioned, that a total of 250 LNs per product will allow detection of a 10%-point difference in LN attrition rate. In an experimental hut trial only a restricted number of nets can be used. How do you select the mentioned sub-sample for the non-inferiority trails? The above-mentioned guidelines ask for six replicate nets per each treatment arm in a hut trial.

- Will the aged active comparator nets be available?

At one point the first in class will have passed all requested trials in addition to the large field trials. Once no further large field trials are done, how will these aged first in class nets be provided? Storage after the field trials is only an option for a short time as it is also known that during storage the a.i. content decreases. When storing these nets from field trials for a period longer than a year, they might not be a suitable comparator for the second in class anymore as the performance of this stored first-in-class would be weakened due to low a.i. content. The second-in class would look to positive when compared to a first-in-class with reduced a.i. content. To date, no data has been created on how long a net aged in a field trial can be stored without losing substantial amounts of a.i. Dissipation of insecticides is a dynamic process. Even freezing nets during that interval would possibly create issues, so although dissipation could be arrested, it is not a completely assured path for evaluation.

Answer: WHO has existing guidance for the evaluation of ITNs/LLINs (previously referred to as WHOPES phase 3 studies) using a household randomized prospective longitudinal cohort design over three years. It is suggested that this design is adhered to and nets are evaluated in hut trials at year three (and after year two, in the case of pyrethroid+PBO nets) to measure the non-inferiority of whole nets as a function of their fabric integrity and insecticidal content. It is recommended that new products are assessed against first in class products for durability assessment to ensure that second in class products are non-inferior to first in class products for the duration of their effective life.

Provided that non-inferiority testing becomes a routine evaluation procedure, WHO ITN testing guidelines will be updated to reflect the need for adequate sample size to detect differences between first in class and second in class products. It is envisaged that sample size calculations will be required to estimate the number of houses needed for follow up to detect 1) attrition, 2) proportion of serviceable nets and 3) the non-inferiority of second in class nets to first in class products in experimental huts or other suitable bioassays based on the mortality endpoint. The variability in efficacy of naturally aged ITNs is an active research question. Should high variability be found then the experiment hut trial design would require a large(er) number of ITNs sampled from the field to be included. Nets of a particular type (for both first in class and candidate nets) can be rotated daily if required throughout the study. For example, for the example with seven arms outlined in Section 5.3.6, then six different nets could be sampled each week (assuming collection done six days a week) so there would be 6×4=24 ITNs in a four-arm study with a single rotation, which should provide sufficient power. More individual ITNs could be sampled if the number of arms or numbers of rotations were increased.

Question: On page 10 the following is mentioned: *The ceiling should be left unsprayed.*

In the existing WHO "Guidelines for testing mosquito adulticides for Indoor Residual Spraying and treatment of mosquito nets" it is stated that the ceiling should be sprayed (see page 21, 3.2.1). We would request that WHO Guidelines be followed and that the ceiling be sprayed to avoid mosquitoes finding an untreated place to rest.

Answer: Yes, the current WHO guidelines for testing IRS do state this, but they are presently undergoing revision. Data on housing structure shows a move towards the use of steel roofs; in spray campaigns these are generally not sprayed. As the residual activity of an insecticide varies significantly depending on the substrate upon which it is applied, it is now considered as advantageous if only the walls are sprayed so it is possible to estimate product residual efficacy on one substrate only, that of the walls. The IRS guidelines will be updated to reflect this change in guidance, including if manufacturers or testing facilities choose to spray ceilings of huts, this should be clearly mentioned and be justified.

Question: Number of huts used per study arm in IRS trials. On page 9, it is requested that a minimum of four huts per treatment arm must be used for an IRS hut trial. As minimum three arms, although four arms are requested. So, one study site needs to have 12 to 16 huts. In our experience, it is typical that trial sites have 6-10 huts available.

- How many trial sites in Africa would be suitable to perform IRS trials with four huts per treatment arm?

Answer: High between hut variability in IRS efficacy indicates that substantial increases in precision are gained by increasing the number of huts from two to four for each arm. Further work is needed to confirm this heterogeneity in different sites. If a site can show the same precision using fewer huts, then this recommendation may be reconsidered. In an absence of such site-specific information, the use of four huts per arm is recommended but not compulsory.

Using one substrate in a study will require nine huts, if there is only one control arm (four first in class, four candidate IRS, one control). More than one control hut is, however, recommended. It is anticipated that additional huts may need to be built to allow for this design in existing experimental hut study sites. A compromise to get around building more huts would be to conduct the four repetitions in different sites, for example by conducting two repetitions in site (A) and two in site (B) or all four in site (C).

Question: Cone bioassays in IRS trials: Can the collection of dead mosquitoes be used as an alternative data set to evaluate the effectivity of an insecticide in case the cone test is not suitable to show the performance of an a.i., for example in the case of a metabolic toxin?

Answer: This question is not directly related to the non-inferiority protocol, but we have taken note of it to inform discussions on the overall revision of WHO testing guidelines. For now, investigators are encouraged to generate data using test methods other than cone bioassays in addition to, not as a replacement of the cone test, to inform discussions around the revision of WHO testing guidance.

Question: On page 12 under 6.2.2, it is stated: *The candidate product is classified as better than the negative control or standard comparator in terms of mosquito blood-*

feeding if it has a significantly lower proportion of mosquitoes that have blood-fed at the 5% significance level (i.e., p-value < 0.05).

For non-repellent a.i., this will not be achievable when compared to a repellent a.i. With this statement non-repellent chemistry is defined as inferior although they might show higher mortality. This should be addressed in the document. Hut trials have shown that a combination of a repellent and a non-repellent a.i. in IRS lowers the bioefficacy of the non-repellent a.i. This would not be a solution to this dilemma.

Answer: Non-inferiority comparisons should not be made between products using different entomological modes of action, such as comparing a repellent IRS product with a non-repellent one. The text in the protocol has been changed to clarify the endpoints for IRS (mortality) and ITNs (mortality and / or feeding inhibition).

Question: Study power: On page 13 under 7.2 the following is stated: *The study should be powered to have a sensitivity of 80% (i.e., β=0.2).* What is the basis of the data that the 80% are chosen?

Answer: A power of 80% was selected by the ERG based on a compromise between accuracy and practicability and is standard for most power calculations from randomized control trials

Question: Glossary. Could a definition of non-inferiority be included?

Answer: A definition has been added to the protocol, as follows:

Non-inferiority trial

A non-inferiority trial aims to demonstrate that the test product is not worse than the comparator by more than a small pre-specified amount. This amount is known as the non-inferiority margin, or delta.

Question: In the glossary, second-in-class products contain the same AI(s) as the first in class. Shouldn't this be broadened to insecticide class or mode of action? Current pyrethroid-only LLINs and IRS products don't have the same AI yet are in the same class.

Answer: The glossary has been revised in an attempt to provide further clarity on this point. It now reads as follows: *Second-in-class product refers to products that have demonstrated the same entomological effect as the first-in-class product, but have not undergone epidemiological evaluation. In the case of ITNs, second-in-class products also contain AI(s) and/or synergists from the same chemical class.*

Question: Could the standard definitions of sensitivity and specificity be given, with a description of how this applies to a non-inferiority trial?

Answer: Additions to the protocol have been made as follows.

Sensitivity

The percentage of candidate products inducing a truly similar mortality as the first-in-class product that are correctly defined as non-inferior. High sensitivity means that most

candidate products that are non-inferior will be admitted to the class (and become second-in-class products).

Specificity

The percentage of candidate products inducing a truly lower mortality than the first-in-class product that are correctly defined as inferior. High specificity means that the majority of truly inferior candidate products are excluded from the intervention class.

Question: The background says that "the PQT evaluates each product in isolation against the thresholds set in the testing guidelines." Which testing guidelines are these, and would it make more sense to set thresholds based on the first-in-class product? What WHO body would evaluate the data generated by non-inferiority trials – presumably PQT?

Answer: In this section, reference is made to existing WHO testing guidelines that were developed under the former WHOPES process and remain in effect until their comprehensive revision has been completed. This topic has been clarified in the revised non-inferiority protocol. Regarding the setting of thresholds, this is a topic related to the revision of WHO testing guidance and will be address as part of it, not here. For now, it has not been decided which department in WHO would assess data from non-inferiority studies, if this method is introduced as a routine evaluation procedure. The data generated from non-inferiority studies on pyrethroid-PBO nets will be assessed by re-convening the Evidence Review Group, in which all three departments involved in the WHO vector control evaluation process (i.e. PQT, NTD and GMP) will participate.

Question: It appears that all generic/equivalent products produce what used to be called Phase I and II data in order to be considered part of a class covered by WHO policy. Is this correct interpretation or does this refer to follow-on products in a new product class? Consistency in language is required (equivalency or 'me-too' refers to products that use existing product specifications and by definition are exactly the same as the reference product; follow-on products to a new product class may not be exactly the same as the 'first in class' product).

Answer: For the current purpose of the protocol, which is to investigate the value of non-inferiority testing, it is envisaged that all products entering a product class will need to generate non-inferiority data regardless of whether or not they are equivalent to a product already in that class.

Question: How can it be assured that the active comparator is performing as well as it is expected to, and how is the target dose of AI on all nets verified? If the active comparator changes in formulation or specifications after generating epidemiological data, what should be used as the active comparator? What will happen if it's not possible to procure the active comparator product?

Answer: The verification of target dose of AI on the walls in the case of IRS or the ITNs is covered by existing WHO testing guidance, which should be adhered in non-inferiority trials. Regarding changes in the formulation or specifications of the active comparator, this 'new' product would then need to be used as the active comparator. Alternatively, it is envisaged that a second-in-class products that has proven to be non-inferior to the first in class product could be used as the active comparator. This area will be further discussed if non-inferiority testing is adopted as a standard test procedure.

WHO is aware of the challenges faced regarding the availability of the active comparator product and is investigating ways in which WHO could assist in overcoming this hurdle.

Question: Wild, free-flying mosquito populations can cause a large amount of variation in hut trial results – would it be possible to instead conduct large-cage trials of nets with similar methods and endpoints, with defined mosquito populations? This would also make population characterization more straightforward and avoid ethical issues with human participants in hut trials.

Answer: Non-inferiority studies should, for now, be conducted by means of experimental huts. Other test methods may be equally suitable for this purpose, but a decision to use one method over another should be informed by data to demonstrate it suitability. Resource permitting, the use of other test methods in addition to experimental huts is encouraged in the non-inferiority study protocol, to build the require evidence base to inform revision of current WHO guidance on the evaluation of ITNs and IRS.

Question: It could be argued that in RCTs, impacts on mosquito population age structure are one of the most important endpoints, and the endpoints of a hut trial may not reflect this community-based effect.

Answer: We agree with this assessment. Yet, RCTs cannot be conducted for each and every vector control products. The methodology proposed for non-inferiority determination is thus meant to be a pragmatic approach to generate evidence to inform whether second-in-class products should be covered by WHO policy that was developed based on data for a first-in-class product, while we are cognizant that this approach has its limitation. In this context it should be noted that non-inferiority studies are powered on mosquito mortality because it is recognized that changes in mosquito population age structure, which are extremely important predictors of transmission, are influenced by insecticide induced mosquito mortality.

Question: "Further endpoints should be included based on the manufacturer's claim." What if a manufacturer's claim is not yet public (submitted to PQ but not published). Would second-in-class/follow-on products only be able to be tested after the first-in-class has received a published PQ listing? What if follow-on products have product claims beyond the first in class product? This is within the remit of PQ so this should be reflected clearly in this document.

Answer: As indicated in earlier responses, the current protocol has been designed to generate data to inform further discussions on the potential value non-inferiority testing rather than to address all of the questions that may be associated with this method should WHO decide to make it a standard requirement. We do take note of the above and will provide clarity with regards to the questions raised above, if non-inferiority testing becomes a routine component of the evaluation of vector control products.

Question: For the IRS secondary endpoints – it's possible that a trial that measures mosquito mortality until it is 50% lower than the first month could be quite long. Why not measuring for as long as the label claim?

Answer: The text has been updated to reflect the manufacturer's label claim to be used to set the duration of efficacy used as the secondary endpoint.

Question: Under the protocol, candidate products must be no worse than the first-in-class, and superior to the current standard of care. If the first-in-class has already shown superiority over the standard of care, why does the standard of care need to be included for comparison, which could create more confusion? What happens if the first-in-class does not perform better than the standard of care in the non-inferiority trial? For example, given the recently reported results from the Olyset Plus trial in Tanzania showing significant decreases in PBO content, it is possible that nets washed 20 times with the proper regeneration time may have little to no PBO left and may not be superior to the current standard of care in a hut trial, although they showed epidemiological impact for the first two years of the RCT. How will the data be interpreted in that case? Given potential differences in performance by current PQ-listed pyrethroid-only nets, how will the brand that represents the current standard of care be chosen?

Answer: In the recently reported results from the Olyset Plus trial in Tanzania[1] the Olyset Plus ITN was superior to Olyset nets i.e. the standard of care. It may therefore be interpreted that this first in class product is a benchmark against which new dual active pyrethroid PBO ITNs with the label claim of superior efficacy against resistant mosquitoes be tested. Standard of care products always need to be included as the presence of insecticide resistance may mean that the efficacy of the first-in-class diminishes relative to the standard of care. In such a scenario an inferior product would be non-inferior to a poorly performing first-in-class product but would be shown to be inferior if evaluated in a site when the first-in-class product was performing better (i.e. there was a difference between standard of care and the first-in-class product). Ensuring the candidate product is superior to the standard-of-care prevents this from happening.

Question: What is the time window for characterizing the resistance profile of the mosquito population in the hut trial site? How close to the trial do previously-characterized sites need to be re-evaluated? Are biochemical assays or target site resistance genotyping acceptable for resistance profiling?

Answer: In section 5.1 of the protocol this is addressed: Mosquito species composition and their susceptibility to the insecticides under evaluation should be determined twice during the hut trial, preferably at the beginning and at the end of the study. A discriminating concentration bioassay should be used to assess the frequency of resistance to all insecticides under investigation. For insecticides that show <80% induced mortality in a discriminating dose bioassay, the intensity of resistance should be quantified using procedures outline in existing WHO guidance. Target site resistance genotyping is not currently acceptable for resistance profiling alone although it is useful supporting evidence and to build up the evidence base for their use in the future.

Question: Is it necessary to do molecular assays to determine proportion of mosquitoes in the trial that fed on humans – isn't it generally high in hut trials?

Answer: This has been clarified in section 4.5: In short, in areas where vectors may feed on animals outside of huts and enter to rest, a sample of mosquitoes from human-baited huts at each site should be identified by species, with assessment of the blood meals to determine the proportion that fed on humans.

Question: The protocol calls for two replicate trials, with a request for a 3rd trial if the studies are inconsistent. If a product has inconsistent results between the two trials, what will happen to its PQ listing in the interim until a 3rd trial is completed?

Answer: This topic will be discussed in further detail if non-inferiority testing becomes standard practice. For now, the aim is to generate a set of data from two trials on pyrethroid-PBO nets, to inform further discussion on the potential value of non-inferiority testing.

Question: There seems to be some confusion between parametric and non-parametric tests. For example, a single hut with enough mosquitoes (on the order of a thousand) would produce fairly narrow confidence limits for any percentage calculated. What is the guidance in the event the mosquito numbers in the single digits are collected each night?

Answer: The analysis requires enough mosquitoes to be collected to be statistically valid. It will be impossible to show non-inferiority with very low numbers of mosquitoes, which is why power calculations are essential prerequisite for non-inferiority analyses. Sites with low numbers of mosquitoes per hut per night can still be used to generate data for the evaluation of products, but require modification of the protocol. For example, for ITNs the numbers of arms used (increasing the length of the study required as a full LSD rotation is necessary) or doubling the rotation may be required to get sufficient power. For IRS more timepoints may need to be included. Conversely, it is good if there are very high numbers of mosquitoes in a hut each night. Though mortality estimates will have tight confidence interval estimates this does not mean that the experimental design can be cut back, because of the necessity for a full rotation (for ITNs) or the need to demonstrate residual activity over time (for IRS). We are unclear about the reference to parametric vs non-parametric analyses. As the data is count data we would recommend that the analysis was always parametric.

Question: In the WHO guidelines for testing ITNs, aren't sleepers rotated each week?

Answer: Guidance in this area is provided in section 5.3.2 of the non-inferiority protocol: "Treatments will initially be allocated randomly to the experimental huts. To avoid any potential bias due to hut position, the treatments should be rotated through all experimental huts using a randomized LSD for both ITN types. For example, the trial outlined above with seven arms would require a 7x7 LSD and take a minimum of 56 nights to perform (49 nights of data collection with one day between each experimental block for cleaning). Sleepers will rotate to a different hut on a nightly basis. Treatments will be rotated between huts once each sleeper has slept for one night under each net. A window of at least 24 hours is required between treatment rotation so that the huts can be thoroughly cleaned and aired. This will minimize any carry-over effects between treatments."

In 5.4.1 a sentence has been added: Because IRS treatment cannot be rotated between huts, a minimum of four huts per treatment arm must be used. Sleepers will rotate to a different hut on a nightly basis.

With regards to the statistical analysis, the importance of heterogeneity is discussed in section 7.1

Question: How is the required mosquito strain chosen for net wash testing, and does it need to be the same for all second-in-class nets? Are we assuming that existing Phase I data for PBO nets, which did not measure PBO regeneration time, would no longer be valid? It would be far easier if nets were washed according to the washing procedure defined in Phase 1.

Answer: For candidate nets with a claim against resistant mosquitoes it will be appropriate to evaluate nets at Phase I using a resistant strain. This area will be covered in the revised ITN guidelines that are currently under development. Planned non-inferiority studies on pyrethroid-PBO nets present an opportunity to close existing data gaps on this class of ITNs.

Question: Should Abbott's formula be included for mortality correction?

Answer: The use of control corrected mortality is appropriate and this is covered in the existing WHO ITN and IRS testing guidelines.

Question: Given that the same nets used in different settings can have huge variability in durability, how can data from nets used in the field for three years be compared? If the first-in-class and second-in-class were used in different settings, wouldn't external factors bias the results?

Answer: The two types of nets would need to be used in the same setting, alongside each other, to generate fully comparable data.

Question: Since the study is powered to the non-inferiority margin, approximately what effect size would be seen when comparing the second-in-class to the standard comparator for superiority? Conceivably with a large enough sample size, even a marginal increase over the standard comparator could be statistically significant at the 5% significance level.

Answer: There is no requirement for the minimum effect size for the candidate product over the standard of care. This is because any statistically significant improvement over the standard of care would likely have epidemiological benefit. Whether that benefit is cost effective is a programmatic issue that is beyond the scope of this protocol.

Question: When determining sample size, how does the discriminating dose bioassay relate to the mortality caused by the products being tested? How would an appropriate mortality range be determined?

Answer: The discriminating concentration of an insecticide corresponds to twice the LC99 against a fully susceptible population. This concentration on impregnated papers closely translates into field application doses of insecticides. This means that if the wild population is fully susceptible, the application dose will yield 100% mortality. Resistant population will provide lower mortality depending upon the intensity. Bioassay on an ITN, however, involve exposure to treated netting and not on the impregnated paper treated with a discriminating dosage of an active ingredient. Nevertheless, bioassay on ITN are considered to give sufficient correlation with 24-hour hut mortality. Scientific articles have reported such association between bioassay mortality and 24 hour mortality in experimental hut trials (for example, Churcher *et al. eLife* 2016). It is appreciated that there is considerable measurement error in both these metrics, but as long as WHO procedures for discriminating dose bioassays are adhered to and the assay is repeated multiple times within a site the data generated should be sufficient for statistically robust power calculations.

Question: Shouldn't step iv (and others, such as adverse effects questionnaire for participants) be aligned with standard WHO hut trial protocols?

Answer: Yes, the sampling protocol should follow standard WHO protocol (both past and in this document) as a minimum, but additional numbers of (for example rotations) may be required to generate the requisite power for the non-inferiority analysis. This has been clarified in the revised protocol.

Question: We find the protocol presents practical challenges and raises concerns as to whether the protocol will ultimately be suitable for the assessment of PBO LLIN products.

5.3 ITN evaluation. While LLINs washed 20 times is accepted as a surrogate for natural aging, through internal studies conducted during the development of PermaNet® 3.0 and subsequent field evaluations, Vestergaard has observed that the main loss of PBO is not through washing of the net. PBO is instead lost through evaporation and mechanisms related to the natural, real-life use of the PBO LLIN. The conduct of non-inferiority studies on PBO LLINs will not suitably assess the longevity of PBO and distinguish between PBO LLIN products.

Several published experimental huts studies conducted on Olyset® Plus and PermaNet® 3.0 showed retention of PBO after 20 standard WHO washes. 1, 2, 3 Whereas, Olyset® Plus was found to have nearly zero PBO left after two to three years of use.4 On the contrary, the WHOPES Phase III studies on PermaNet® 3.0, which are currently being finalised, indicate that PBO is retained in PermaNet® 3.0 after three years of natural, real-life use.5 It is only in field studies (minimum one year) where we have begun to see clear differences between PBO LLIN products. In the development of a product claim, requirements for durability and efficacy data can be nested within field studies (i.e., village studies, cluster randomised control trials, Phase III studies on durability and acceptability, operational research, post market surveillance activities).

Answer: We appreciate the limitations of washing as a method of aging nets; hence the requirement for field-aged nets to be retested. Whether or not non-inferiority is a suitable method for distinguishing between pyrethroid+PBO nets (please note that WHO does not use the term LLIN in the context of these nets) remains to be seen once data to inform such discussion are available.

Question: *5.3.6 Assessing field durability.* The protocol states that a product's inclusion to the class would be re-examined once an experimental hut trial on a three-year-old LLIN was conducted. If the first examination of active ingredient(s)/synergist longevity in natural use is only when a three-year-old net is made available, a situation could arise where millions of substandard LLINs are distributed before the issue can be detected and rectified.

Moreover, while the protocol recognises that sampling of a three-year-old LLIN would need to be conducted in a way that minimises bias, it does not suggest how this would be done in practice. Phase III studies on durability conducted by WHOPES and other organisations can attest to the variability in bioavailability and durability of used nets depending on environmental and use conditions. It is difficult to imagine how a single three-year-old net would be selected in a representative and unbiased manner, and how the data from one used net would sufficiently represent a product's performance. Again, as suggested earlier, there are other methods to obtain data on field durability in support of a product claim.

Answer: Non-inferiority is a means of analyzing data that is generated following standard WHO guidance for experimental hut studies on ITNs and IRS, while ensuring studies are adequately powered to ensure that reliable estimates of the true efficacy of products is obtained. It is envisaged that an adequate sample of ITNs obtained from prospective longitudinal cohort studies will be used to determine the non-inferiority of field aged nets in experimental huts using WHO recommended methodology. The protocol allows for more than one age net to be tested to account for variability in the aging process. Power calculations to select the correct number of nets from the field to capture true variability both among and between ITN brands will be required.

Question: *6.2.1 Non-inferiority test.* The protocol is inconsistent on whether an LLIN product would need to be within the non-inferiority margin for mosquito mortality and blood feeding inhibition, or mosquito mortality and/or blood feeding inhibition. Should an LLIN product be required to show non-inferiority on both mortality and blood feeding? The protocol has not clearly described how the data will be interpreted if only one of the endpoint is met. The recognised variability of local, wild mosquito populations at experimental hut sites will also generate endpoints that vary with geography, season, mosquito species, and insecticide resistance status.

Answer: Studies are powered to detect differences in mortality endpoints. However, as with existing WHO guidance the non-inferiority of ITNs will be considered on both the mortality and feeding inhibition; nets that are non-inferior on either endpoint will be considered non-inferior.

Question: The protocol lacks reference to Good Laboratory Practice. GLP conformity will require active collaboration of the first in class product manufacturer and the manufacturer of the standard comparator LLIN to provide new nets accompanied by certificates of analysis.

Answer: GLP is in the process of being rolled out, but few sites have so far been accredited. If non-inferiority studies become a standard requirement, all studies will have to be conducted at accredited sites to GLP standards. During 2019, however, testing at sites not yet accredited but on the list of being under development for GLP accreditation would be acceptable. A general communication on the topic of acceptable data from WHO PQT is under preparation.

Question: The choice of demonstrating with 95% probability that there is less than a 0.7 odds ratio appears subjective and there is a need to check that this would not exclude effective products (e.g. a product that gave 5% lower average mortality might be epidemiologically effective but not pass this criterion (as the 0.7 odds ratio represents only 9% difference in mortality and the replication required to demonstrate the difference between 9% and 5% with 95% confidence could be excessive)).

Answer: The cut off odds ratio of 0.7 represents a balance between practicalities and desire to not recommend a truly inferior product. This compromise was reached by consensus of the ERG as there is currently no practical evidence of what this value should be; the optimum will vary depending on setting and goal of the malaria control programme.

Question: As the Olyset Plus PBO concentration apparently fell in the 3rd year of the epidemiology trial that has been carried out and no benefit over a pyrethroid net

was found at that point, this calls into question its suitability as a 3-year PBO net and therefore as the active comparator in this protocol and particularly in the 2nd set of studies (treatment arm three on page 9).

Answer: We agree that with regards to testing pyrethroid-PBO nets, studies should be conducted after two years and – depending on the findings - after three years. The protocol has been amended accordingly.

Question: The glossary on page 6 indicates that second-in-class nets should contain the same ai(s) as first-in-class nets. We understand that for this protocol all nets should contain PBO and pyrethroid, but we do not agree that the pyrethroid has to be permethrin (to be compared with Olyset Plus). (It should, however, be noted that PBO may synergize deltamethrin and permethrin differently depending on the exact resistance mechanisms in place in a trial location, and that needs to be considered when evaluating nets for non-inferiority as nets that are non-inferior in one location may produce different results in another location where the underlying pyrethroid resistance mechanisms are different.)

Answer: The protocol has been edited from active ingredient to specify that this should be the same chemical class.

Question: This study design will occupy a substantial number of experimental huts (seven per trial) over a significant period of time (5 weeks). It will be important to check that a single round of the proposed seven treatments can be expected to provide sufficient power to demonstrate non-inferiority in the intended trial locations, otherwise huts would be occupied for a longer period. Availability of huts for this work also needs to be checked.

a) There may be situations where cattle are more appropriate than humans as bait, for some zoophilic vectors whose numbers would be too low to power the study adequately unless animal bait were used.

b) We agree with the need to check the regeneration time of the nets with respect to their PBO effect (5.3.3 on page 8). This may be several days, in which case the time involved for 20 washings prior to the trial needs to be planned in.

Answer: These points pertaining to study design/management are valid and text to qualify these points has been added to the protocol.

Question: We support the proposal to test the approach towards evaluation of candidate PBO nets in 2019 and evaluate the outcome before deciding whether to adopt it or generalise it to other product evaluations. An alternative approach should be considered which involves checking that performance of a candidate product meets or exceeds an absolute threshold in terms of mortality (e.g. 80%) and blood-feeding inhibition. A first-in-class product can be included in this test but only to check that the assay system is working and correctly demonstrating that it also passes the threshold criteria. Without this there is a danger that the first-in-class product performs unexpectedly well (e.g. >95% mortality) or poorly (e.g. <40% mortality): in this event a candidate ITN demonstrating strong efficacy (e.g. 85% mortality) in the first case could be considered 'inferior' and rejected whereas one with poor efficacy (e.g. 50% mortality) in the second case could be considered 'non-inferior' and accepted.

Answer: The alternative approach to add an absolute threshold in terms of mortality has not been used to measure the relative performance between products and between sites where the resistance profile of local vector mosquitoes and consequently the absolute mortality of different products may vary. As an illustration, a PBO LLIN would fail to reach a defined threshold of 80% mortality in many sites in Africa with pyrethroid resistance yet Olyset Plus was shown to have an epidemiological benefit in a recent RCT. Variability in response is always a consideration in experimental hut trials, but allowing a direct comparison with the first-in-class product reduces the main source of heterogeneity which is the difference between experimental hut trial sites in time and space. Power calculations should also account for between hut heterogeneity in efficacy.

Question: The need to replicate hut treatments for IRS studies is recognised, provided WHO has evidence of significant inter-hut variability. However, there is no justification provided for the need for four replicates as opposed to three or two. With the proposed four replicate huts per IRS arm, and minimum of three treatments per trial, it would be impossible to get through the full programme of testing needed to progress new IRS products using the present arrays of huts available in many sites. So this would be a bottleneck until additional land could be acquired and huts built, with associate time and cost. Therefore, it is critical that the necessary level of replication is carefully reviewed and justified.

Answer: Data presented at the ERG demonstrated that four huts will provide higher quality data with greater sensitivity than when one or two huts are used. It is incumbent upon testing facilities to generate data that is adequately powered to provide WHO with data that is definitive and adequately powered to detect differences between products with the acceptable non-inferiority margins. Therefore, trial sites should perform power calculations and justify the number of huts used for IRS trials in their study protocols submitted to WHO. Data from different experimental hut trial sites can be analysed together to improve feasibility, as long as the distribution of arms across sites is representative (i.e. each site must have equal repetitions of different arms).

Question: We recognise the need for manufacturers to test their new IRS products on the variety of wall and ceiling surfaces that will be encountered in the areas where the product will be used. With respect to the efficacy evaluation requirements for PQ listing we interpret the protocol to be indicating that this evaluation should be carried out on a representative or predominant wall type. However, if several wall types had to be tested under this protocol with four or more hut replicates of each of three or four treatments per trial, the two trials needed for evaluation of a single new candidate IRS product would occupy an excessive and unrealistic number of huts.

Answer: It is not necessary to test more than one wall type in the non-inferiority trial since it is a direct comparison between the candidate product and the first in class product. The selection of substrates should be justified in protocols submitted to WHO and should reflect the predominant wall type of the area where the hut trial is being carried out. It is important that all studies are adequately powered and if necessary, trial sites may have to increase the number of experimental huts that they have available.

Question: For IRS we question the rationale for the two different treatment lists set out on page 9 of the protocol proposal. If a new insecticide is in the same chemical class as an existing WHO-prequalified IRS product or not it is in the same policy class (IRS

products) and therefore the appropriate standard reference product to use would be an effective WHO-prequalified IRS product, irrespective of its chemical class.

Answer: If a new insecticide does not have another insecticide in the same class then it is necessary to compare the new product to 1) a product with a similar duration of residual efficacy as per the label claim in addition to 2) the current standard of care in the region to demonstrate whether the candidate IRS may offer improved control for resistance management. It is agreed that, provided that the mode of action of an IRS product is mortality, then the insecticide class of the active comparator may be different to that of the candidate IRS.

Question: The proposals make reference to first-in-class IRS products. Please clarify which products these are given that the IRS class was not established through epidemiology RCTs.

Answer: WHO's policy recommendation is underpinned by a systematic review of RCT data.[2] The selection of a relevant active IRS comparator has been addressed in an earlier response. In essence, a number of options exist and more than one comparator may be selected. Discussion with WHO on this topic is encouraged prior to initiating trials. It should also be noted that, for now, the focus lies on investigating the value of non-inferiority studies by means of studies on pyrethroid-PBO nets. While the current protocol aims to also tackle the issue of IRS evaluations, guidance on this is likely to be further refined once non-inferiority data have been generated and reviewed by WHO.

Question: We advocate the alternative approach using absolute thresholds, as mentioned above, for the future evaluation of both LLIN and IRS products.

Answer: The alternative approach to add an absolute threshold in terms of mortality has not been used to measure relative performance between products and between sites where the resistance profile of local vector mosquitoes and consequently the absolute performance of different products may vary. Incorporating absolute thresholds will possibly exclude useful products that may have a role in insecticide resistance management strategies.

Question: We see no reason for each new chemical class of LLIN to be considered to be a new policy class. This is inconsistent with IRS insecticides which are all in the same policy class. It also creates unnecessary complexity and additional need for expensive and time-consuming epidemiology testing. The only exception to this would be 'insecticides' that do not kill or prevent blood feeding but work by reducing fecundity, adult life-span or other method not linked to mortality.

Answer: This comment is not related to the notice-of-intent nor to the non-inferiority protocol, and hence not addressed here. We do, however, take note of it and will reflect on it when appropriate, as part of broader discussions on the evaluation process for vector control tools.

Question: For insecticides in different chemical classes that have different characteristic levels of knockdown, excito-repellency or speed of kill, the free flying hut study system

2 WHO 2019. Guidelines for malaria vector control. https://www.who.int/malaria/publications/atoz/9789241550499/en/

should be adequate to evaluate them without the need to separate them into distinct policy classes.

Answer: The entomological mode of action of IRS is mortality. Experimental hut studies will be adequate to measure non-inferiority of insecticides that have mortality as the primary mode of action. For insecticides with alternate modes of action it is necessary to consult with VCAG on the determination of the product testing pathway.

Question: Where some insecticides or mixtures claim effectiveness against pyrethroid resistant mosquitoes, the hut studies can be carried out in areas where the wild mosquito population is pyrethroid resistant and provided specimens are collected and characterised for resistance this will still enable this study system to be used for evaluation without the need for defining new product classes or requiring non-inferiority studies.

Answer: It is necessary to evaluate products in areas where the vector population is suitable for assessment of the validity of the product label claim. The definition of new product classes and appropriate evaluation pathway is determined by WHO. Where an existing first in class product is available, the demonstration of no-inferiority of a second-in-class product will allow the inference that a second-in-class product will also provide acceptable public health value.

Question: Some products may provide additional public health value in combating pyrethroid resistant mosquitoes, as well as pyrethroid susceptible mosquitoes, and this can be recognised by categorising them separately once their performance on resistant mosquitoes is demonstrated. It may be possible to achieve this in terms of pyrethroid + PBO or via a new insecticide class or new insecticide class mixture. In all these cases it should be recognised that they are providing additional public health value not just 'public health value' as stated in the 2nd sentence of page 1 of the protocol proposal.

Answer: Public health value is defined as follows: '*A product has public health value if it has proven protective efficacy to reduce or prevent infection and/or disease in humans.*' New products are, in the first instance, required to demonstrate such epidemiological impact, rather than additional impact over another tool. Additional impact over the current standard of care would certainly be desirable, but is not a pre-requisite for a policy recommendation of a new tool.

Question: The protocol indicates that cone bioassays should be performed on nets during the study (5.3.5 on page 8) and on walls for IRS treatments (5.4.3). This may not be appropriate for some products, e.g. containing chlorfenapyr.

Answer: We do appreciate that other bioassay methods may need to be used to assess other new tools and may need to be amended/expanded if it becomes standard practice. For now, this protocol deals only with WHO recommended bioassay methods and products covered by a WHO policy recommendation.

Question: It is not clear from 3.2.2 on page 3 whether the primary endpoint for IRS is mortality assessed by cone bioassay or in a free-flying hut test. The former would pose problems for some new chemistry.

Answer: The primary endpoint is mortality measured from total mosquito mortality in free flying experimental hut tests for the duration of the evaluation that is dictated by

the label claim of the manufacturer. The residual efficacy of a product tested by cone bioassay is listed as a secondary endpoint.

Question: For IRS or LLIN products, what would be the appropriate comparators for mixtures of two active ingredients from different chemical classes under the proposed protocol?

Answer: For now, the focus is on generating data for pyrethroid-PBO nets for which the active comparator has been designated to be Olyset Plus. Should non-inferiority studies become a routine component of the evaluation process for vector control interventions, further communication on the appropriate comparator will be provided by WHO. Meanwhile, investigators planning to conduct non-inferiority studies on products containing a mixture are advised to directly contact WHO for further discussion on this topic.

Question: How will discriminating doses be defined in terms of mixtures of active ingredients in IRS and LLIN products (e.g. Fludora Fusion)?

Answer: As indicated in the above comment, the evaluation of mixtures by means of non-inferiority studies is currently not seen as a priority. In any case, discriminating concentrations relate to an insect's response against an individual AI. Test papers are not to be impregnated with mixtures. Note: Some LLIN have PBO+AI, but for synergist bioassays, mosquitoes are first exposed to PBO and then to AI.

Question: The protocol indicates on page 2 that it will be necessary to monitor the bioavailability of the partner AI in field-aged nets over a period of three years. Currently there is no standardized method for doing this for next generation nets including Interceptor G2.

Answer: It is recommended that whole nets are taken from the field and evaluated in experimental huts (Section 5.3.6). As indicated in an above response, is recommended that such evaluation should be conducted after two and three years for pyrethroid-PBO nets.

Question: For IRS studies the protocol indicates that ceilings are to be left unsprayed. We note that in the real world a significant element of the killing effect of IRS comes from spraying thatch ceilings.

Answer: It is recommended that ceilings of experimental huts are left unsprayed so that the longevity of the IRS is calculated for the wall substrate only. If the manufacturer wishes to measure the efficacy of the IRS on ceilings as well as walls this may be added to the study as a separate study arm.

Question: Could you kindly tell us the difference between the non-inferiority study and phase 2 study, can we add the first-in-class product as a positive control directly in the PQ phase2 study?

Answer: Non-inferiority studies, as currently envisaged, would indeed be 'phase 2' studies but explicitly include a first-in class product as the comparator and be powered to allow a comparison of the second-in-class product under investigation with the first-

in-class one. We would like to draw your attention to the power calculation outlined in the protocol, which is likely to differ from that of a conventional phase 2 study.

Question: For the equivalence product, whether need to do non-inferiority study?

Answer: Yes, all pyrethroid-PBO net products are requested to generate this information. If non-inferiority testing becomes the norm, all related testing guidance and communication around it will be updated accordingly, and be in line with WHO guidance on testing products claiming equivalence.

Question: We hope you could kindly understand that it's very difficult for our manufacturers get the first-in-class product, so we hope PQ/GMP can kindly establish a way to obtain the first-in-class product

Answer: Yes, we have been made aware of these difficulties. WHO will look into ways in which this can be facilitated.

Question: The primary concern is the reliance on hut studies to prove 'non-inferiority', particularly with the know heterogeneity in species composition, resistance profiles densities and other factors such as hut design and SOPs that have raised questions about controlling for these approaches in the past. Simply put, there is significant variation that will be difficult to control between sites with hut trials comparing performance against a first in class. One could imagine a situation whereby any subsequent net would be preferentially trialled at the site of the first in class test, although even that could provide differing results due the likely increase in resistance intensity over time. It is difficult to see with such heterogeneity how clear comparisons would be made without the need for a significant amount of scientific judgement and/or modelling.

Answer: Experimental huts continue to be the most widely used approach used to test the entomological efficacy on vector control tools. It is acknowledged that they are not perfect, but until a better assay has been developed and validated they will remain the gold standard. As indicated in the study protocol, investigators are encouraged to explore potential alternative assays alongside experimental huts – assuming budget is available for this – to inform discussions around potential alternatives.

Heterogeneity is well recognized as a fact that needs to be taken into account. This is why the study design articulated in the protocol aims for a side-by-side comparison of a second-in-class tool with a first-in-class one (i.e. at the same time and in the same location). It is required that at least two studies with the same effect i.e. demonstrated non-inferiority or superiority to the first-in-class product at a minimum of two study sites is demonstrated. The current protocol (a direct comparison between candidate and first-in-class) enables tests to be carried out in areas with insecticide resistance without penalizing the candidate product.

Question: While the proposed approach may work for the current impasse around PBO nets, it is unclear how effective it would be for any other product within a class in the future. There is a risk of setting a precedent with 'non-inferiority' testing that simply does not work for other classes and may serve to stymie innovation. Moreover, the proposed approach raises a number of issues that should be recognised, clarified or resolved before a decision on whether to continue along this path is taken.

It is unclear whether this guidance replaces the current limited field studies guidance for nets and IRS. If so, how will they be reflected when the guidelines are eventually updated and, considering these guidelines pertain to products that are not first in class, should these guidelines not be developed by the PQ team as they are entomological in nature?

Answer: Reference is made to existing WHO testing guidelines. This has been clarified in the revision of the protocol. Regarding the setting of thresholds, this is a topic related to the revision of WHO testing guidance and will be address as part of it, not here. For now, it has not been decided which department in WHO would assess data from non-inferiority studies, if this is introduced as a routine procedure. The data generated from non-inferiority studies on pyrethroid-PBO nets will be assessed by re-convening the Evidence Review Group, in which all three departments involved in the WHO vector control evaluation process will participate.

Question: Will these studies need to be carried out under GLP as other semi-field studies need to be? If so, do the sites have capacity to accommodate them (feedback suggests not)? If not, why not as this would then appear to be a separate requirement from what is needed for a PQT dossier? The suggestion that these trials be carried out in areas of intended use seems to point to non-GLP testing or would involve very expensive and complicated GLP supervised studies in countries with no GLP facility. This needs to be clarified.

Answer: Non-inferiority studies, as envisaged, would indeed be phase 2 studies as required for PQT dossier, but explicitly include a first-in class product as the comparator and be powered to allow a comparison of the second-in-class product under investigation with the first-in-class one. We would like to draw your attention to the power calculation outlined in the protocol, which is likely to differ from that of a conventional phase 2 study. It is suggested that these trials be carried out in areas of intended use for instance a West African or a South American testing facility rather than in each country where the product may be sold. It is envisaged that testing will ultimately be conducted at GLP accredited sites, but in the absence of a sufficient number of these data from sites that are on the list of undergoing GLP accreditation will suffice.

Question: It is unclear why a negative control is required when the intent of these guidelines is to compare non-inferiority to the first in class. The would appear to add more complexity and resources for little value.

Answer: a negative control is a requirement in all product evaluations as an experimental quality control to ensure that observed mortality or feeding inhibition is induced by the insecticide under evaluation and not due to poor experimental practice or environmental conditions.

Question: There is no incentive for the manufacturer of a first in class product to engage in this process once its product is on the market and there is an assumption that a manufacturer would provide products to act as the first in class which is likely to not be the case as they are under no obligation to provide anything. Why would any innovator allow their products to be tested repeatedly against the competition that will eat away at their market? Under such guidance a first in class product would be tested over and over again against their competition while any comparator product would only need to

be tested once. It simply would not make sense for a product to be first in class under such conditions. We are already seeing issues with PBO nets and it will continue for other products. One can foresee significant problems obtaining first in class nets for this testing which would stymie market development and exacerbate volume concerns.

Answer: Yes, we do understand the difficulties encountered in obtaining the first-in-class product for testing purposes. WHO will look into ways in which this can be facilitated. As is current practice with IRS and pyrethroid ITNs, it is envisaged that appropriate second in class products (e.g. similar label claim) that have demonstrated non-inferiority to first-in-class products may be used as the active comparator for future on-inferiority evaluations of other second-in-class products.

Question: Linked to the above, it is unclear why only the first in class product should be the comparison. Surely if a product has been proven to be 'non-inferior' to the first in class, and has a comparable entomological dataset, it should be viable as a comparator. Thus, any product included in the class, not just the first in class, should be seen as a comparator. This is how SRA regulators manage their testing and it may lessen the disincentives to be a first in class.

Answer: This is indeed the case and it is envisaged that non-inferior second-in-class products could be used as active comparators for future studies to support market development and reduce issues related to obtaining fist in class active comparators for each new study.

Question: The proposed methodology continually mentions 'standard of care' as being a pyrethroid LLIN, however, this is likely to change in the near future; particularly in countries where they are already using PBO nets or 'next gen' nets. Moreover, it seems somewhat reductive to only have to prove efficacy against pyrethroid only nets as the evidence suggests (at least from and entomological perspective) that they are much less effective than next generation LLINs. The notion of 'standard of care' is thus pretty hard to pin down it is unclear.

Answer: Standard of care is defined as "The type of insecticide-treated net (ITN) or IRS product predominantly used by the national malaria control programme in the country where the study will be implemented". In most countries this currently means a pyrethroid only LLIN. A pyrethroid only LLIN is thus required for comparison to ensure that first-in-class products demonstrate superior efficacy to pyrethroid only LLINs against the wild vector populations found at that testing facility. In the future it is anticipated that the standard of care will change.

Question: It is still unclear why LLINs are treated differently to IRS. Products should be evaluated on their entomological effect not their chemical components.

Answer: The entomological effect of LNs is mortality and feeding inhibition while the entomological mode of action of IRS is mortality only.

Question: There is a significant risk that, should this process be employed further than PBO nets, innovation will be stymied. If a product only has to prove 'non-inferiority' to the first in class, there is little incentive to make claims of elevated efficacy to that product. Moreover, this protocol appears to assume that the first in class should set the standard for the rest of the class; this may not be true due to superior products that may

come second in class and outperform the first in class. How then will WHO manage such a situation? Will the parameters for the class be altered or are they fixed? If they are altered, what does that mean for the original first in class? What if a manufacturer wants to prove superiority to the first in class? are they able to or do they just settle for 'non-inferiority'?

Answer: We would like to remind the reader that the current protocol has been developed to explore the value of non-inferiority testing, which will be further discussed once data to inform such discussion are available. The present protocol was not intended to rank the efficacy of different products within the same product class nor to answer all of the questions posed under this comment. If non-inferiority testing is adopted as a standard component of the evaluation of vector control tools, then due consideration will be given to address these and other questions relevant to this evaluation approach.

Question: It would be advantageous for WHO to consider taking a lead from stringent regulatory authorities who have been reviewing similar products for a number of years. While their standards for epidemiological evidence are not as exacting, they have well defined safety, quality and entomological efficacy systems that have been honed over decades and are understood by applicants and country regulators. The core components of such a system for efficacy are:

- Classes defined by 'use case' and never by AI or claims e.g. resistance.
- Entry into the class is defined by clear thresholds that are established by the first in class and iterated on with growing experience of the product in the field. These thresholds are laboratory based and would constitute the former 'phase 1' testing for laboratory efficacy. They are carried out to test the product against highly controlled situations and can test a wide range of vectors and resistance profiles in order to build claims.

Answer: The class of vector control products is defined as outlined in WHO/HTM/GMP/2017.13. A product class in vector control is a group of products that share a common entomological effect by which it reduces pathogen transmission and thus reduces infection and/or disease in humans. Currently, laboratory testing does include performance thresholds as outlined in WHO guidelines.

Question: Semi field studies are used to ensure the products can prove efficacy against defined field populations. This is not done in on a 'non-inferiority' basis, but again on thresholds. A know product from the class (e.g. the first in class, although any product in a class should be viable) is used as a positive control to calibrate the test, but to test subject is against compared to established thresholds (e.g. the former phase 2 guidance). The resolves the issue of constantly retesting the first in class and lessens the concerns of its manufacturer as their product is being used to calibrate, not to prove a competitor is superior or non-inferior. It is standard practice with SRA regulators and well understood by manufacturers and regulators.

In this way, the first in class is not continually re-tested against its competition and products can make claims, that would be reflected on their label, that could exceed the threshold for entry to the class. Using this methodology, WHO could much more clearly define a product class based on entomological effect and 'use case' rather than the current situation. Thresholds, set by a first in class, would be linked to the epidemiological data so a reviewer could clearly understand that a product meeting (or exceeding) them would have an epidemiological impact. Such a methodology would

be clearly understandable to member state regulators as it would be in line with OECD dossier requirements, to which many are familiar.

Answer: As mosquito resistance increases, the established thresholds of 80% mortality are not reachable in all but those test sites with susceptible mosquitoes. Setting a standard, i.e. a new product with proven public health value, allows the comparison of products within an acceptable range against the performance threshold of a proven intervention. In addition, once a second-in-class product has demonstrated non-inferiority to the first-in-class then it may be used as an active comparator for the evaluation of other second-in-class products.

Question: While this approach may solve the issues with PBO nets due to inconsistent and missing data, it is still unclear as to how this may be applied consistently to a broader range of products in the future. Moreover, it is unclear how the suggested methodology fits with the new PQT assessment mechanism as it would appear to replace current hut trial testing, although it is unclear how or whether this is the case.

Answer: As indicated in the notice of intent this work is indeed exploratory, and it has hence not yet been defined whether or how it will be applied to the broader range of products. These discussions will need to be informed by experience of implementing the approach in a select area, in this case the PBO nets. Based on this experience and the resulting data, decisions will be made on whether / how to take non-inferiority testing further. With regards to data packages required by PQT to assess vector control products, it is envisaged that the non-inferiority design would be complementary, using the very same studies required to generate the data package, but designed so that the product under investigation can be compared to a first-in-class product already covered by WHO policy.

Question: The data requirements and protocol for determining non-inferiority are welcomed, and we see a great value in laying out the considerations around non-inferiority that also will impact possible new vector control classes like Spatial Repellents. As the developer and manufacturer of what we expect to be the first-in-class Spatial Repellent product, we have an interest in understanding the data requirements for candidate second-in-class products (both our own and competitor products) to be considered under an existing policy recommendation, and we were surprised to read in the report from the meeting of the Evidence Review Group that the key audience for this document was not industry. The non-inferiority data requirements are of primary importance to us – not just researchers – as they will clarify the success criteria for new products under development and guide product development processes.

Answer: We apologize if the meeting report did not explicitly state industry as a key audience. WHO does consider industry to be one of the key audiences for this communication.

Question: We would like to confirm that a candidate second-in-class product cannot claim superiority over the first-in-class product, only superiority over a negative control or a standard comparator, which is in another/older class. If this is the case, it should be stated more explicitly in the document.

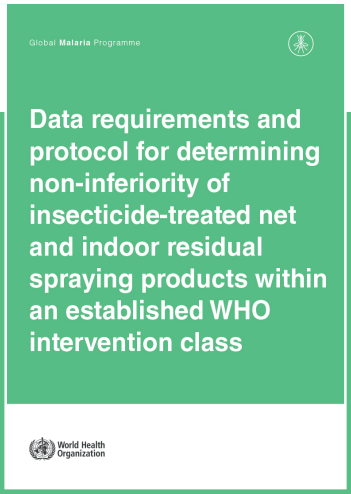
Answer: Indeed, a second-in-class product cannot claim entomological superiority over a first-in-class product under the current evaluation process, nor is this envisaged if non-inferiority evaluations are adopted as general practice.

Question: Data requirements and protocols for demonstrating non-inferiority should be an element within the WHO efficacy testing guidelines for each class of vector control products. For new/future vector control product classes (like Spatial Repellents), will data requirements and protocols for demonstrating non-inferiority be included when these guidelines are revised or written, rather than as a subsequent and separate document? This is ideal, as it would prevent unnecessary delays in assessing candidate second-in-class products (many of which are already under development) within these new product classes.

Answer: With regards to data packages required by PQT to assess vector control products, it is currently envisaged that the non-inferiority design would be complementary to the generation of data required by PQT. The same studies could be used to generate the data package for PQT assessment, but be designed so that the product under investigation can be compared to a first-in-class product already covered by WHO policy. Again, the area of non-inferiority trials is currently considered exploratory. If it does come into action as general practice, then WHO will ensure that the guidance gets integrated and the approach streamlined in every possible way.

ONLINE RESOURCES

Data requirements and protocol for determining non-inferiority of insecticide-treated net and indoor residual spraying products within an established WHO policy class. Geneva: World Health Organization; 2019 (https://www.who.int/malaria/publications/atoz/non-inferiority-protocol/en/)



WHO/CDS/GMP/2019.06