# Stopping Criteria for Priority Screening

Stopping criteria is one of those endlessly debated topics. There are no hard and fast rules, though there are a few guidelines and some suggested algorithms to help make a decision.

In terms of using Priority Screening where only partial screening will take place, the purpose of this approach is to reduce the number of items that need to be manually screened, so it is particularly useful when the search yields are too large to screen everything given resource constraints. There is, however, a risk associated with this approach which varies from review to review. Where the prioritisation has worked well, then the risk of an eligible study appearing further down the prioritised list is very low. However, because we are not manually screening the whole corpus, we will never know for certain whether this is the case, and methods for determining how well the prioritisation has worked in a 'live' review are still in their infancy. As such, the choice to use this approach will largely depend on whether the risk of missing a relevant study is acceptable or not. *(Consider the context of the review; a quick overview of a topic requires less consideration than a rigorous examination of a specific review question. A brief look at attitudes to public transport would need less rigour than a thorough analysis of a particular medical treatment, where you wouldn't want to miss a single paper covering some contraindication.)* Thus the circumstances will dictate the importance of capturing all relevant papers, or the seriousness of missing a relevant paper.

So, there may be some reviews for which – either because of the nature of the topic, or the expectations of the funder / audience – the idea of missing even one study will be problematic. (Note that the risk of missing a study does not equate to actually missing a study!). We have developed some "checks" to help identify any eligible studies in the unseen part of the prioritised list, but these do not guarantee that all hidden eligible studies would be found.

Another thing to consider is whether your exclusion criteria are well-defined. If things are clear-cut, then the PS system will perform better.

Here are three potential ways of deciding when to stop screening. (Note that none are guaranteed. You may carry on screening and find another include, especially if it is atypical i.e. different from the includes that you've already found and that the PS system has learnt from / searched for. OR you may carry on screening for many more thousands of items and not find any more includes…)

1. Having a set criterion based on resources – time, number of refs etc. – do what is achievable in terms of screening given those available resources

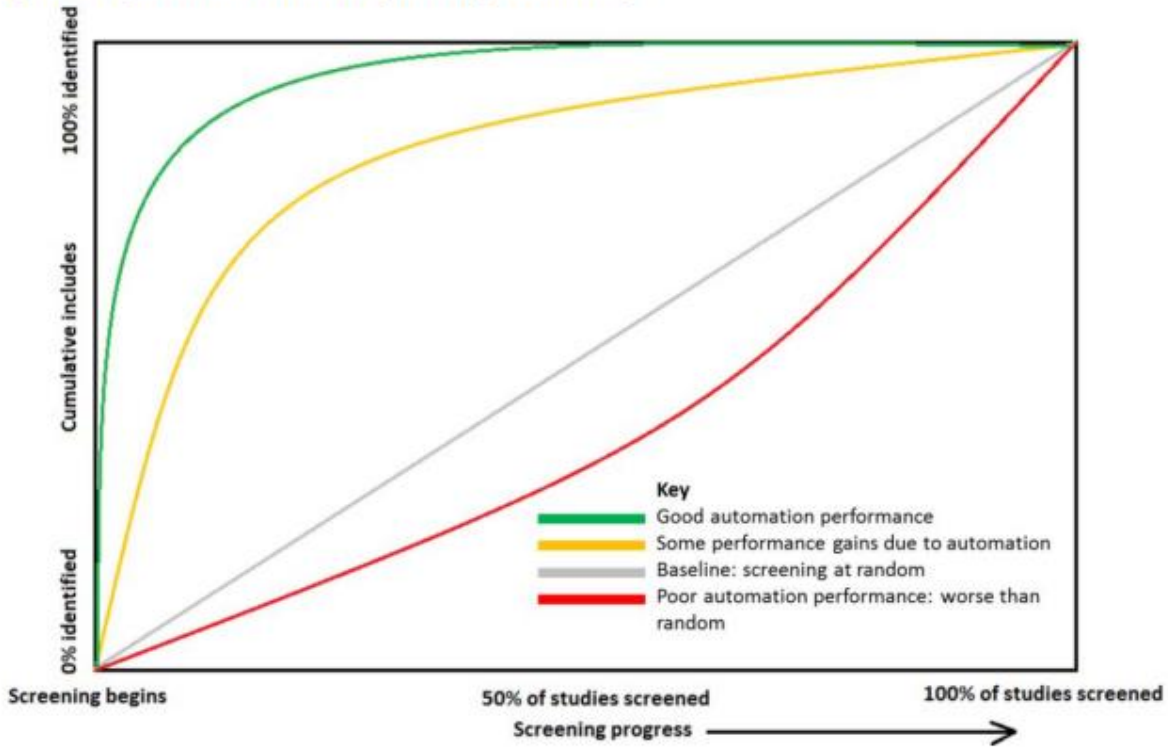
2. Having a set criterion based on reaching a 'plateau' where no new includes found (e.g. after 1,000 refs) – and accept there is the attendant risk of cutting off too early

3. Establish what we expect to be a 'baseline inclusion rate' (screen a random sample of studies to get an expected include rate), then screen across the entire set of studies until a similar proportion are identified – the most rigorous method, which could be combined with the other two methods

   (e.g. If you screen 500 random items during the PS system's training phase, and find 20 includes, you'd expect to find a similar proportion over your entire reference set. So, given 5,000 items in total, the BIR would indicate you'd expect to find around 200 includes. If you reached that number and your chart had plateaued out, that would be two corroborating indicators that you could safely stop screening.)

The screening progress graph is a good guide to the effectiveness of priority screening in your review; ideally your curve would *follow the shape of a lower-case letter **r**,* with a steep upward incline followed by a plateauing out. You may have seen the following figure from our documentation online (https://eppi.ioe.ac.uk/CMS/Portals/35/machine_learning_in_eppi-reviewer_v_7_web_version.pdf, from https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=3899#ERWebML).

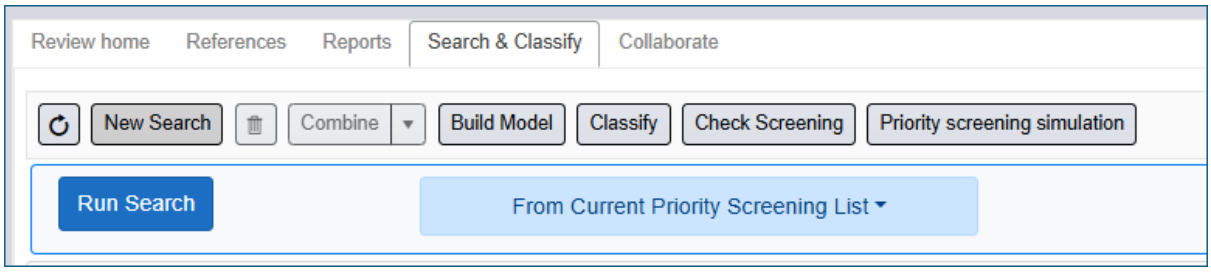
**Figure 3: possible results of priority screening**



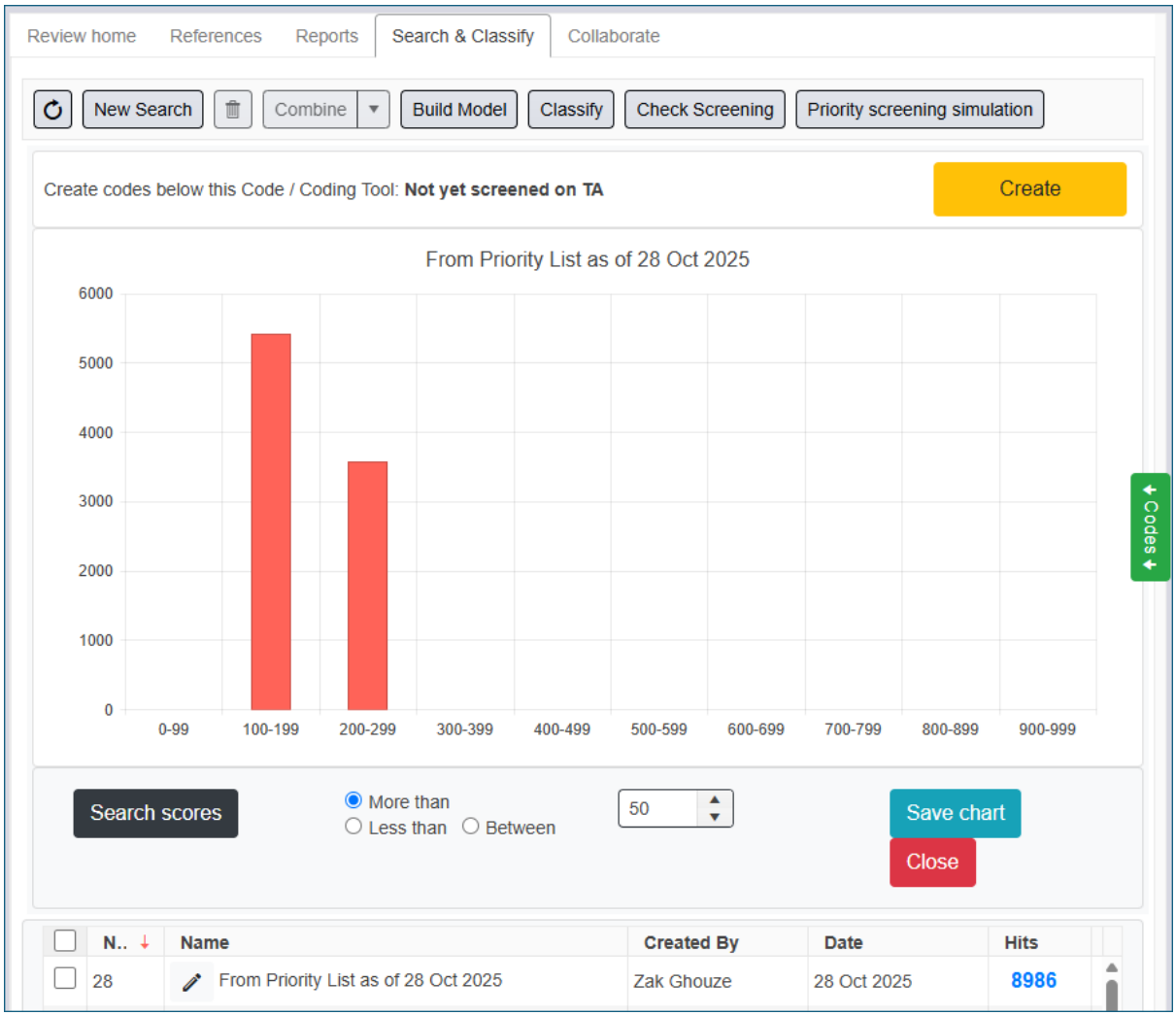
Note that the PS system (and classifier models) learn from examples of includes & excludes, and will therefore identify similar items. You could be screening and suddenly find a “new seam” of includes to “mine”, different from those you’d previously found. The PS system would learn from these new includes and reorganise your screening list so as to prioritise this new type of include and bring them to the top of your list of items left to screen. This is why we sometimes see “jumps” or sudden upward inclines in PS progress charts. If your includes are all quite similar, your chart should be fairly smooth.)

Bear in mind that any suggested numbers (e.g. screening 500 items in succession without finding another include) are simply suggestions. One does have to bear in mind overall numbers too. If you had 1,000,000 items to screen, and you had screened 500, that’s a comparatively small proportion of your total. The PS system may still be learning; there may still be atypical includes, as yet unscreened, for it to use as training material; you may want to screen more items in random mode than you usually do before switching to priority mode…

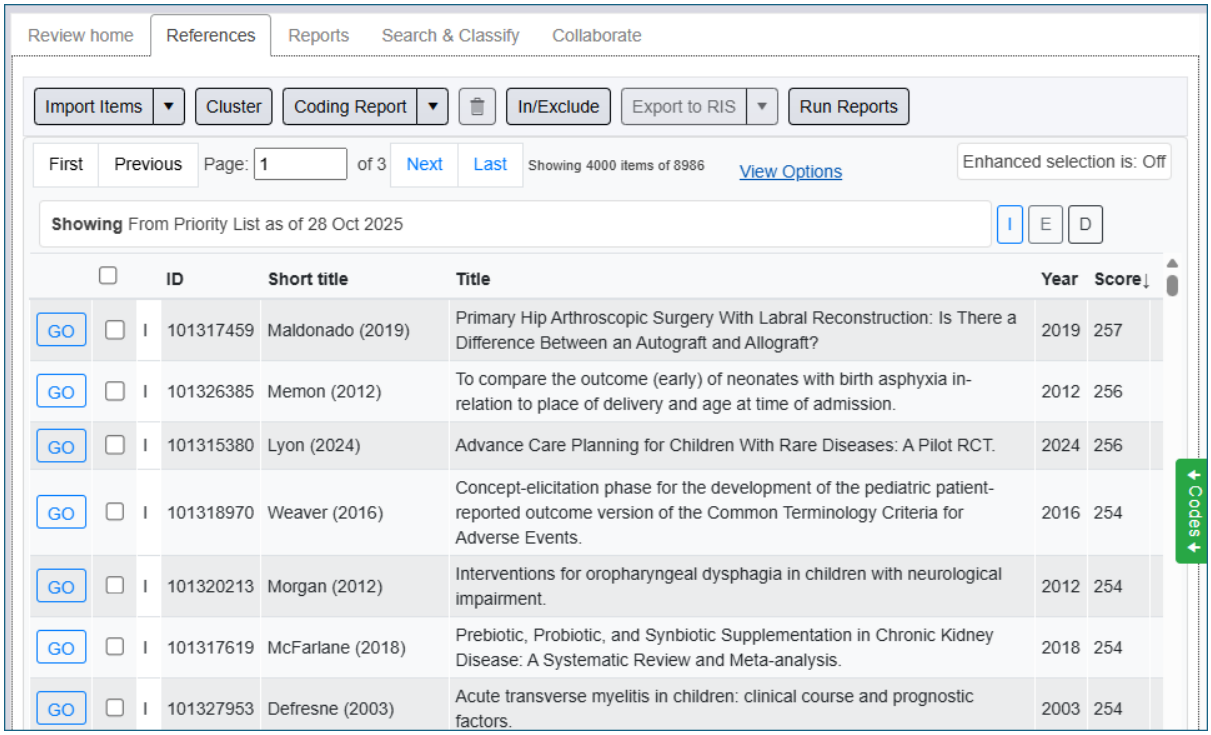
We do now have tools in EPPI Reviewer to look at the probable relevance of all items in the PS list, so we can look over those yet to be screened (as opposed to having to just work through the list) and see their probable relevance score.




e.g.



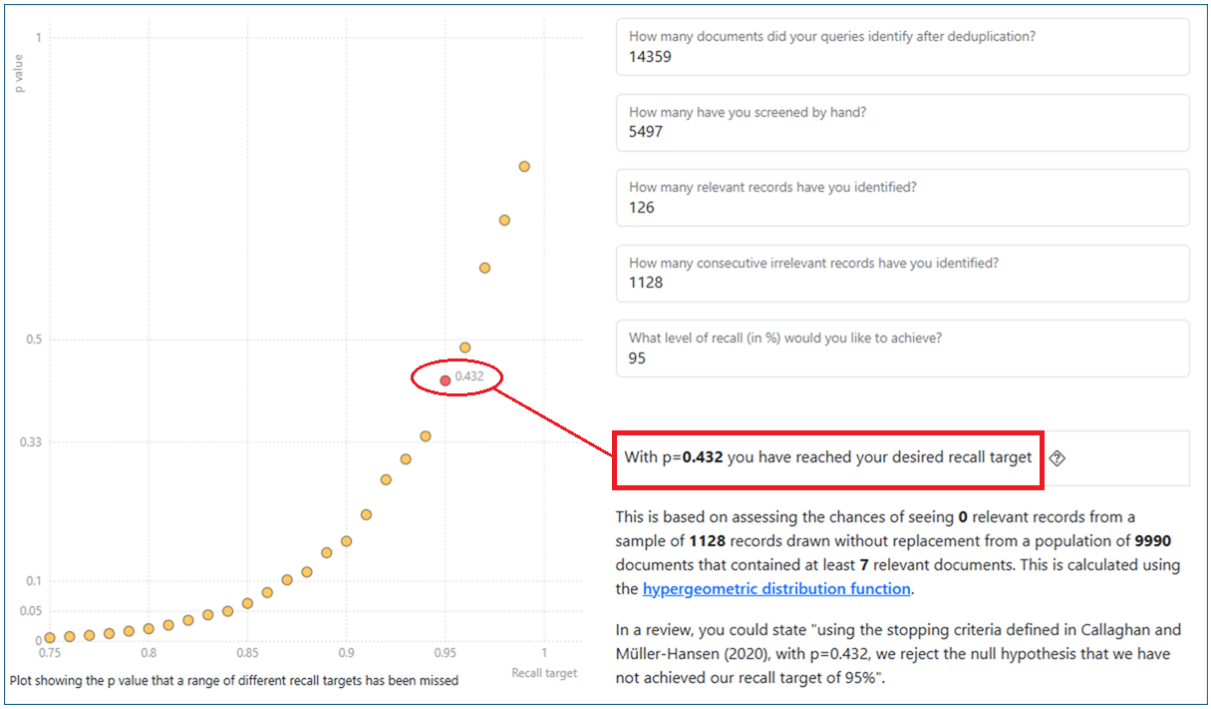
If you click on the number of hits (8986) you can see items listed with their score (assuming you have your ***View Options*** set to show the score column). In fact, from this example we can see the highest probable relevance is only 25.7% / 0.257. Those with higher probabilities have already been screened.



We are still aiming to introduce an automated tool to suggest a stopping point to users, but this has to be done with caution. We don't want to give the impression that the tool is guaranteed to be right, and users can often take such tools as guaranteed… As always, these things are circumstance-dependent. They will work better with certain sets of refs / topics / screening criteria… (And there is still debate on the exact calculations involved!)

One external tool (not found within EPPI Reviewer) that you could try is Max Callaghan's BUSCAR utility at https://mcallaghan.github.io/buscar-app/. (See https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-020-01521-4 for some related background info.) Max has worked with us at the EPPI Centre on many occasions and we hope to have a similar system to this within ER in the future. One has to have a "perfect" set of numbers to use it though i.e. you can't change coding tool or item set midway through proceedings, etc.

You simply enter the number of items in your review, the number already screened, number of includes, number of consecutive excludes you found (i.e. how many items you have screened without finding an include), etc. Given a particular recall level (you want to get 95% of includes, for example), the app will give you some justification for stopping…



A more manual method for doing a similar power calculation is described below -:

When not all citations will be screened, firstly screen a random sample of studies. This step is to inform stopping rules. It involves estimating an inclusion rate, then testing this by ensuring a sufficient sample of studies has been screened to obtain this rate. The steps are as follows:

1) Screen a random allocation of references.
    a. The priority screening mode in EPPI-Reviewer can be set to random, or
    b. Use EPPI Reviewer's collaboration tools to create a random set of items and assign them to reviewers as wanted (one reviewer per item, multiple reviewers screening the same items and comparing results, etc.)

**NB: The screened references need to be complete. Studies will automatically be set as complete within single-screening mode. Studies that have been screened in double-screening mode and not completed will not be included within this step.**

2) Determine the proportion of includes from the screened references – this is the initial predicted inclusion rate.

**Initial predicted inclusion rate = Number of citations that meet inclusion / number of citations screened**

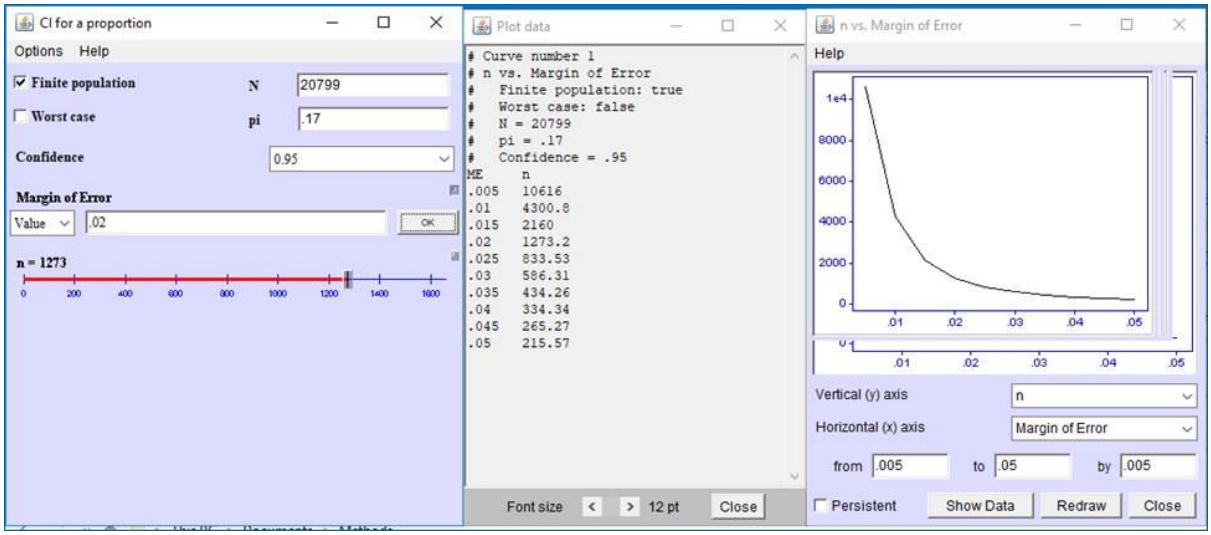
3) Use a power calculation to obtain **n**, the number of references required to be screened randomly in order to be a sufficient sample for obtaining that inclusion rate.

Use the initial predicted inclusion rate (e.g. 5%) and choose a margin of error of that predicted inclusion rate, e.g. 2% if you think the inclusion rate is most likely to be somewhere between 3 and 7% (i.e., 5±2%). You also need to set your confidence interval, which is usually fine to leave at the default of 95%.

4) Once **n** records are screened, check the inclusion rate again. If the predicted inclusion rate is within the margin of error then this is established as the baseline inclusion rate (BIR) and screening prioritisation can be turned on. The power calculation can also be re-run to check this. If the predicted inclusion rate is not within the margin of error, it suggests that your initial predicted inclusion rate was very inaccurate and you need to re-calculate the number of studies needed to be screened using the new predicted inclusion rate. Continue until a sufficient number of studies has been screened to achieve the predicted rate and a BIR is established.

See example below of a table of working out and screenshot of using the power calculator, which is useful for transparency and record-keeping purposes.

| **Number of items to be screened in the review:** | | 20, 799 | | | |
|---|---|---|---|---|---|
| **Stage** | **Number Screened (S)** | **Number of includes (I)** | **Estimated inclusion rate I/S** | **Power calculation informs number items to screen (n)** | **Margin of error for n @95% Confidence interval** |
| 1) Screening of a random allocation of references | 391 | 65 | 0.166 | 1251 | 0.02 (2%) |
| 2) Screening citations, to at least n (from stage 1) | 1767 | 304 | 0.172 | 1273 | 0.02 (2%) |
| **Decision** | As n for stage 2 is less than the number of items Screened for stage 2, AND the estimated inclusion rate is within the initial range predicted, this suggests a sufficient sample of studies have been randomly screened prior to screening prioritisation. We can therefore establish 17.2% as the Baseline Inclusion Rate. | | | | |




Once the BIR has been verified, the Priority Screening system can be changed from Random Mode to Priority mode.

(If you used a random assignment instead, then switch to using the Priority Screening system in Priority mode.

Setting up a rule like stopping screening after you have not included anything in the last X items screened is a feasible way to operate this in some cases although not particularly robust. Two methods that would be more reliable are:

1) Calculating the baseline inclusion rate using a random sample of manually screened items and applying the proportion to the whole pool of studies in your review - described here

**https://doi.org/10.1002/jrsm.1093Digital Object Identifier (DOI)**

2) Calculating statistical stopping criteria - described here https://doi.org/10.1186/s13643-020-01521-4

---

In general, it is your call how you calculate the stopping criteria in your review - it mostly depends on how essential it is for you to pick up all possible includes in the sample and how confident you want to be that nothing is getting missed. In some rapid/scoping reviews, people tend to be more flexible around it (I have seen reviews where teams simply decided to screen first X references using Priority Screening and stop after that or stopping based on how their screening graph looks - when the curve flattens out); in other cases, you have to be more statistically robust and report actual stats that allowed you do stop without arriving to the end of the screening list.