**Webisode 2 – EPPI-Centre tools for collecting and using data**

Presenter: James Thomas (EPPI-Centre, UCL)

EPPI-Centre Evidence Tools, Products, and Projects – A series of webisodes from the Evidence for Policy and Practice Information and Co-ordinating (EPPI) Centre. Hosted by AIR’s Center on Knowledge Translation for Disability and Rehabilitation Research (KTDRR).

YouTube video: <https://youtu.be/FZX7nVrxFIE>

JOANN STARKS: Welcome to another session from EPPI-Centre Evidence Tools, Products and Projects. This series of brief webisodes will introduce the audience to several tools, products and projects of the Evidence for Policy and Practice Information and Coordinating Center or EPPI-Centre. Based at University College, London's Institute of Education, the EPPI-Centre focuses on the development of systematic reviews and studies the use of research evidence.

I'm Joann Starks from the Center on Knowledge Translation for Disability and Rehabilitation Research, or KTDRR, and American Institutes for Research. The Center on KTDRR it's sponsoring these webisodes with support received from the National Institute on Disability, Independent Living, and Rehabilitation Research, NIDILRR, in the US Department of Health and Human Services.

This session focuses EPPI-Centre Tools for Collecting and Using Data. Our presenter is James Thomas, Associate Director of the EPPI-Centre. James' interests include developing methods for research synthesis, in particular for qualitative and mixed methods reviews, and in using emerging information technologies such as text mining in research. Welcome, James. I'll now hand things over to you.

JAMES THOMAS: Thanks very much, and thank you for inviting me to give this short webisode. I'm going to be talking about EPPI-Centre tools collecting and using data. And I should acknowledge right at the outset that we use many different tools, which we can't possibly claim credit for our generating all of them.

And there are a wide variety of tools out there for use in systematic reviews. I'm going to refer to some of them in this little webisode, and also highlight some of the tools which we've either developed ourselves or were involved in co-developing.

So, I'm going to talk about three broad areas of tools for use in systematic reviews. I'm going to talk a little bit at the beginning about data and data management throughout the lifecycle of a review, and then think about data at specific stages of the review. In particular about identifying and selecting studies for inclusion in the review, capturing and extracting data to describe the studies, and how to synthesize study findings.

And I know there's an awful lot of interest now in automation technologies and systematic reviews. And that's an area of particular interest of mine. So I'm going to highlight some of the tools that are available now and some of the tools that are available in hopefully the near future to help make our lives easier when we're conducting systematic reviews.

So turning to probably quite a familiar slide to many of you, the common stages of a systematic review, from forming the review team, formulating the question, all the way through to communicating and engaging about the findings of the review.

And information management is necessary at very many stages of the review, in particular to do with searching and identifying studies, describing the studies looking at the quality and the reliability of the studies, and then synthesizing their findings. And we find tools-- information management tools to be essential for all of these detailed data management components of doing a systematic review.

So thinking in particular here at the beginning about collecting data for the review, we conduct many different searches of many databases when we're conducting systematic reviews. We search large online databases, such as PubMed is particularly large and famous one, and we also increasingly follow citation trails.

So these are where we've got a study which we're interested in, which looks as though it's relevant to our review. What we do is we look at the references the end of the paper. And then we also look to see which papers have cited that paper. Because following the citation trials can well identify more relevant studies for inclusion in the review.

And we often search these bibliographic databases and download thousands, sometimes tens of thousands of references. And so keeping track of all of these potentially vital pieces of information is quite a critical information management task. So what we do is we've got some software called EPPI Reviewer, which we upload all of our references into.

So this is just an example of one of the pages in that tool. This is a live review where we've uploaded references from the Web of Knowledge or Web of Science. And you can see in the middle of the screen there we've got the search string that we use. So we've got a nice detailed record of the way in which we search that database. You can see it's a Boolean search where we separate the different components of our search conceptually in searching the database.

And towards the top right there you can see a list of the other sources that we searched. And also, we've got a record of the number of items that we found each time we uploaded a search, and also whether or not the items were flagged as duplicate, so whether they were items that were already in our system.

So we try and keep a good audit trail of all of the data that enters our system so that we can be confident when we're looking through the information that we don't actually inadvertently lose vital studies for our review. Once we've uploaded the documents there's a task which we call sifting or screening studies for relevance in the review. And as that's an area of increasing interest in automation, I'm going to talk about that a little bit later.

But once we've identified the studies which are important and relevant for our review, we then look at what's called data extraction or classification or data collection. And here what we're trying to do is to identify information about each study which is going to help us either understand the context of the study, how the studies relate to one another, the participants, potentially the interventions if there are interventions in the studies, and the reliability of the study.

So we look at the methods that the study used in terms of its data collection, its analysis, its sampling. And all of these different pieces of information help us to understand what kind of reliability we should place on our study, how much reliance we should place on its findings, and how the study might fit into our emerging picture of answering our review question.

And so we have a number of what's highlighted on this site here, called standardized code sets. And you can see on the left there are a range of different code sets there, some which are just to do with screening on titles and abstracts and full texts, and some which are to do with risk of bias assessments. So you can see we is the Cochrane tool for risk of bias assessments, also the NICE quality appraisal checklists, qualitative and quantitative studies, the CAST checklist, and so on.

And we've also got some data extraction code sets in there. And what we're wanting to really highlight here is this balance that we need to strike each time we do a review between data reuse and using the same tools across lots of different reviews, which aids consistency, consistency of interpretation and reliability of the review itself. But also the way in which we need sometimes to customize tools for use in particular reviews.

So here we've got for example, the NICE quality appraisal checklist. You can see various checked various items that we need to check on when we're carrying out a quality appraisal for this type of study, which is a quantitative study.

Another tool I thought it would be useful to highlight was one to assess the quality of qualitative studies. And this is an EPPI-Centre tool, which we've used in many reviews now over the years. And we get quite a lot of inquiries about the availability of the tool and how to use the tool and whether people can use the tool. It's freely available in many of our reports and people have used it in different reviews.

And what we've tried to do when we've been developing this tool, when we developed this tool in the past, is to strike a balance between a very detailed tool-- there are some checklists for appraising the quality of qualitative studies which have more than 50 items, more than 50 questions to answer, and some which are very brief.

And what we have tried to do is to strike a pragmatic in-between stance on this where we don't have an exhaustive number of items to check but we do identify some of what we think are the critical components of extracting what we think might be quality or reliability of a qualitative study.

So you can see here in the tool we look at whether or not steps were taken to strengthen the rigor in the sampling of the study, whether the steps were taken to strengthen the rigor of the data collected and in the analysis of those data. And then we're particularly interested in looking to see whether or not the findings of the study are supported or grounded in the data which were collected.

When trusted in terms of breadth and depth this is not necessarily a quality issue. But certainly we do find that studies which are more insightful do tend to get deeper into the phenomenon that's being looked at. And the critical one for us often is question six there. We're looking to check to see whether or not the study has taken measures and the perspectives of the people carrying out the study have been such that they've really privileged the perspective of the participants-- in this case YP means Young People-- but certainly the perspectives and the experiences of the participants.

We've then got some overall questions on reliability, and particularly about relevance. So whilst some studies are beautifully carried out and you can't fault them in those senses, in terms of the utility in the reviewing question sometimes what we need to do is to ascertain the degree to which there's alignment between the original study's research questions and our reviews research questions.

So that's a little bit on one of the tools that we've developed for assessing the quality of qualitative studies in particular. And in terms of the application of these tools, what we tend to do is upload PDF files into the system and then we capture a range of data about the studies.

So you can see here on this little screen shot, we've got the PDF file there on the right and we've got the codes tool on the left there. And here you can see that we capture categorical information here. We've got checkboxes. So we're looking at the design of the evaluation. Was it a trial or was it post-test only? Was it pre- and post-test, et cetera? Was it other?

And so we capture categorical information. So this is information which is known when we set up the coding tool when we identify the coding tool. We know that we're, for example, going to be ticking the checkbox ‘trial’ for some of the studies in the review.

But you can also see the little info buttons. And for every classification we make we might well want to qualify that or add additional information. So we also, as well as capturing the categorical information, the tick boxes, we have free text information which the reviewer can add to describe the study in more detail or to justify why they've made a particular classification.

And you can see on the right there also there's a section in the PDF that's been highlighted, just a couple of lines highlighted in blue. And that's additional information that we capture, again to justify and to give a transparent trail between the text that we read in the documents and the classification that we allocate to a particular study.

And the third form of data, which isn't on the screen here, is statistical data. So we also capture data--for example just the number of people in a study--but also data about outcomes. So we capture odds, ratios, standardized mean differences and that kind of thing, which describe the findings of the research in statistical terms.

When I'm thinking about the findings of the research of course, that brings us on to analyzing the data. I'm not going to go into great detail here. But there are a number of different things we do to analyze the data. For a start, when we're just beginning our synthesis, when we're beginning our analysis, what we find it really helpful to do is to map the extent of research activity that we've got in the review.

Sometimes we have a specific mapping stage interview. And I think another webisode has talked specifically about that. So what this is about is about describing research activity across a number of different dimensions. And you can see here that it's not just about doing this in words, but what we try and do is visualize research activity. So we can see that we might have frequencies of categories.

Again, these categories are those which we've used the checkboxes, and I've shown you on the previous screen. Now we've added up how many checkboxes for different domains here we've got in this pie chart here. And we've been able to visualize the range of research activity in this case about the type of community engagement.

And in the middle here you can see an evidence map, which are increasingly popular in some fields. Here we're able to visualize across a couple of dimensions the extent, the quantity of research in particular fields. And on the bottom right, this is a tool called VOSviewer, which analyzes the text in the papers in the review.

And here that's able to tell us where the-- it's a density plot, these things are called. So it tells us over in terms of all the topics in the map where are they most densely clustered. Where is the research activity most intense?

Again, looking at analytical tools, we also do statistical meta-analysis. You can see here with the setting up as a small meta-analysis, and we use R, and in particular the Metafor. But we use some other packages too, that were created by Wolfgang Viechtbauer.

So we use EPPI Reviewer as the data collection front-end, which then calls these packages which are written in R. I want to just go through these so you can see there's some R code on the right there. There's a forest plot there. And really what we're trying to do here is to make it a transparent link between the data, which we've captured, which we extracted in our data collection process, the R code which was executed in order to create the meta-analysis, and then the meta-analysis itself.

So again, what we're trying to do is to make sure that we've got everything in the same place. That we're able to see right from the moment that we upload a paper into the system all the way through to extracting information about the study, to looking at the scope of research activity in a map, right through into the statistical analysis. We've got what might be thought of as an audit trail all the way through, so that we make sure that we don't lose data and we treat data in a consistent way.

Now the final element that I was going to talk about today was automation in our systematic reviews, and the automation that we've got in our tools. And this is something which has been gaining a lot of interest and a lot of attention. And we got into looking at this originally simply because of the data deluge that we were suffering from when we were doing systematic reviews.

Increasing numbers of papers being published. There are more and more sources to search. And what that's meaning is that the time it takes to do a systematic review, we were finding, was increasing simply because of the volume of research, and also because we were getting better at finding research. So two things were going on there to increase burden in systematic reviews.

So I mentioned citation screening earlier. A critical phase in the review, because this is the moment where we find all we lose or we miss the studies which are going to make up our findings. So it's really important that we make sure that we find the right studies.

And in terms of automation, citation screening has probably received the most research and development in terms of computer science, machine learning, and that kind of thing. And we've found when we've done a systematic review looking at the various different methods, that it's possible to reduce workloads in systematic reviews using some of these technologies quite considerably. Certainly in excess of 30% in most cases, and in some cases up to 97%.

And we've also found that you can use these tools in a number of different… On the right there I've got a summary of the conclusions we came to. The first one was screening prioritization, which is really using these tools to identify the most relevant studies at the beginning of the screening process. And I'll talk about that a bit in a moment.

We can also use these tools as a second screener. And what I mean by that is that sometimes some people doing reviews will look at everything in duplicate. They'll have two people screening every single record. In this case what they might be able to do is replace one of those people and say one person having to look at 10,000 records, for example, and have the machine do it instead.

And in the final case is automatic study exclusion, where it's possible to train the machine well enough to be able to replicate human decisions accurately enough so that actually we don't need to look-- even have one person looking at every single citation.

So our little slide on how the machine learns. This process has come from computer science and is known in that area as active learning. We've got some slight differences from it here. But essentially the process is that we upload all of our citations, all of our titles, abstracts, et cetera, into our database in stage one.

And then in stage two we undertake what would be quite familiar to many systematic reviewers, a phase of manual screening. So we look at a random sample of the citations that we uploaded, quite a small random sample hopefully, and that gives us just enough to begin the machine learning.

So hopefully what we identify-- and we need to keep screening at random and so we found some-- some examples of what is relevant and inevitably lots of examples of what's irrelevant. Once we've got examples of both classes of documents we can move on to the third stage, which is the machine learning.

And there what we do is we take the relatively small number of relevant and irrelevant studies and we build a machine learning model. And this machine learning model learns the combinations of terms that are used in those titles and abstracts that we've already looked at manually. And then what we use that model for is to look at all of the rest of the studies in our database and rank them in order of their similarity to the ones that we've seen that are relevant.

And so what we end up with is an audit list with those which the machine thinks are most relevant to our review at the top of the list. We then look, and move then on stage four, where the list of studies then-- the top of the list of studies is looked at manually. And the way that we have set this up in our system is that the top 25 are looked at manually.

So a reviewer will look through the top 25, and then behind the scenes when they get to number 25 the machine will go back to stage three. It will now have more data to work from because it's got an additional 25 on top of what it had before. It will rebuild its machine learning model. It will re-predict through the rest of the unseen citations which ones are likely to be relevant. And then as the review on stage four begins screening again, what they're actually doing is they're looking at the top of the new list.

And this cycle goes round and round with the machine learning model being created time and again, and the list in which the reviewer sees the item is being recreated. This all happens behind the scenes. They don't have to worry about that particular part of the process. And what that results in is a ranking of the citations in order of relevance.

And this is an example from some of the reviews from the Cochrane Heart Group. There's six reviews here. And looking at the top middle review for example, you can see that the x-axis is the screening progress. So on the far left of the x-axis nothing has been screened. On the far right everything in the review has been screened, in this case about 15,000 studies.

And then the y-axis is the cumulative number of relevant studies found. So if we were screening at random the proportion of relevant studies found would be in proportion to how far we've got through the screening process. So in this case it would simply be a diagonal line going from the bottom left to the top right.

But you can see in these graphs that we're not seeing that shape at all. What we're seeing is a very sharp and extremely sharp in some cases, rise to the top, at which point it flattens out. And what that's showing us is that the machine learning algorithm is able in some cases-- for example on the top in the middle-- very quickly to identify the relevant studies and present them to the reviewer. And then everything else in the screening process is irrelevant.

Then you can see, for example, on the top left that sometimes that learning process takes a little while. And again on the bottom right you can see that. So we're not seeing it like a right angle. It takes a little while and there are quite a number of irrelevant studies, which are found at the same time as relevant ones.

So, screening prioritization I referred to before. That's the process by which you would use this. And you would find the most relevant studies really quite quickly in the screening process. But you would still look through all the rest of the studies, because you'd want to be sure that someone had looked at all of them to double check them.

But you could also use this process in an automatic way to say once I've screened, for example, in these reviews-- any of the reviews-- if we'd screened manually 30% or so of the studies, all of the rest could have been discarded without any loss of relevant studies to the review. So there's potential here for a real workload savings.

Another way in which it's possible to save work in a review and to use machine learning automation in the review is using large bespoke classifiers that do a particular job. In this case, I've built to classifier which is able to identify randomized controlled trials. And this is built thanks to the Cochrane Crowd, on 280,000 records, which members of the Cochrane Crowd have looked at and determined whether they're RCTs or not.

And we can see from the graphs here and from the text that actually this is a very accurate classifier. For example, if we trust our classifier to discard those references which it's very confident are not randomized controlled trials, we only potentially lose 0.1% of relevant trials.

And I think one of the questions that we as a systematic review community have to think about is whether that's good enough. Some people would say, well no it's not. What we need is 100%. On the other hand, some of the search filters that are used currently and widely used in systematic reviews only have a calibrated recall of 98.5%.

So some of our current methods don't reach 100%. We know that. One of the things we need to do as a community is to consider what are the acceptable thresholds for the use of these tools. An example of using the classifier as an EPPI Reviewer is here. This is a fairly typical search. I've applied the RCT classifier and I've got a range of scores along the bottom here.

So I've got a number of studies. The number of studies are shown in the y-axis. And I've got the scores, which the classifier has shown. And this is very typical of most searches in that if I was looking for randomized controlled trials I put in a lot of material which is not randomized controlled trials.

And you can see here the large majority are not randomized controlled trials. They're scored between 0% and 9% likely of being RCTs. And really, the classifier is accurate there and I would be quite safe in discarding all of those if I was looking for trials. And I wouldn't be losing relevant trials.

Another area which we're currently working on in terms of automation is the extraction of data from graphs. And this can be really quite problematic when you're doing a systematic review if you encounter the results that presented not in numeric form but as a graph. And what people have to do typically then is to recreate the data by using rulers and the like measure across on the graph and to recreate the data.

So what we're looking to do here is some automation, some user interface work, which will help us to do this more accurately and more quickly. One of the really quite exciting developments over recent years has been what we're starting to understand as a connected ecosystem of services and of data.

And this is just a schematic showing how the new EPPI Reviewer 5 that we're working on with NICE is architected; where in the middle there we've got EPPI Reviewer, our systematic review software, which is made up of a number of connected data services and data stores.

We've got an index-- an open access index of studies in there. And we've got a number of services, which include machine learning, de-duplication services, statistical services, and services which IBM Research is currently developing on prediction and recommendation based on the studies which are in a particular review or particular domain.

I won't go through it all in detail, but I will leave that one there. You can press pause if you like and have a look at the different components there. One of the things we're planning to do with this is to open source the software so that we can encourage and facilitate collaborations. Because this is a job which is much bigger than one or two organizations can do on their own. If we're to really realize what is becoming a vision of open access and open data material, we need to make sure that we're developing systems which are able to speak to one another.

So I'll just finish on a “coming soon.” We've got EPPI Reviewer 5 coming soon. And we'll be writing a blog on that soon. And we'll be wanting to put it out there for people to try out and give us feedback on. And really, we're very open to collaborations and to new ways of thinking about how we might work with the data which go up to make systematic reviews.

What we're finding is that our understanding of what a systematic review is, is slowly changing over time. And we're becoming increasingly aware that what we need to be doing is starting to build and maintain systems which survey the evidence base or our surveillance systems of the evidence base, so that when systematic reviews are needed we're in a better place to be able to identify the research more quickly than we have currently been in the past.

So on that, I think probably that's the end of my slides. I'd like to say thank you for the invitation to talk about all of these things, and please do get in touch if you've got questions.

JOANN STARKS: Well, thank you very much, James, for sharing this information on EPPI-Centre tools for collecting and using data, including that introduction to some of the automated processes that you're working on. We also want to thank our funding agency, NIDILRR, for supporting this and other webcast activities. Please look for the other sessions in this series on the EPPI-Centre Evidence, Tools, Products and Projects.