eIUS: Disease Control Experience Report 1

Interviewee profile

Research fellow in mathematical biology

Time spent in research

About 95 % (including preparing presentations, etc); a little bit of time I spend on teaching;

Research area and Research question(s)

His area of research is Epidemiology.

"The sorts of questions that we are addressing are: if you get an outbreak of a disease in the country, let's say in a particular farm: What is going to happen? How big is the outbreak? How big the area of the country that is going to be affected? What would be the best strategies for controlling that outbreak, or how do you prevented for being a large outbreak?"

Research Lifecycle

Start of the research process – Literature Review

"Every problem is different, I can say of avian influenza, what we first did was to [ask]: What kind of data do we have? We are talking about the disease spread about poultry farms, so we wanted to know where are they. We wanted to know things like the risk factors between the farms. Which ways the disease can be transmitted, and we wanted to know in relation to that, how industry works, the sort of movements that go between the farms, and in terms of getting this information, we got a list of what we need. The sources of information based on DEFRA [UK – Department for Environment and Rural Affairs], the experts that we had, and the poultry industry itself (..); fortunately, we have good contacts with those. We have questions, like how many farms a particular 'feed lobby' would visit in a particular trip, so we could ask them for typical values of this sort of thing."

"Some of the information was very definite, for example, DEFRA would do a review of where the poultry farms are, a poultry questionnaire, and give us set corners, so we have definitive information there. Other information, such as, which of the transmission was most risky, it is rather more subjective, and so, we will get the opinion of people like DEFRA, people from industry to give their opinion about which movement it is more likely to propagate the infection, and in addition, we look at work that has been done at transmitting infection in other areas, such as Europe or Asia [where the disease between farms was propagated]. So, we do literature [reviews] (..); [but] no two situations are the same, you can do maybe a comparison with science in Europe and the UK, but in Asia it is a little different, because you do not have these huge companies (..) [and rules how chicken] are distributed from each other (..). The routes of transmission [of diseases] are different, and the probability of transmission in those areas is greater."

"**PubMed** will be the one that it is most often used for this sort of review. I have to say that we split the project up." The interviewee has a physics background working in mathematical/systems biology; in **the project** there is another mathematician, while the other three people were vets. In terms of the analysis of papers to do with transmission or any other work that was done in the area they did not do what could be called a systematic review, which is a systematic analysis of every relevant possible journal or publication which is quite a task in itself: "This is not something that I was familiar with before. Starting in this field, this is something that particularly vets are keen on."

This means: in terms of data collection, information comes from different sources.

Data collection process and simulation model

"So, we have data collection (..) [based on] DEFRA, and there were sites over Internet which gave us information about things like poultry feed; other information comes from the literature as well, to know what assumptions the authors made and why did they make it, then, we are still left with gaps in our model, so there were things that we did not know the answer to (..) so we could go to a group of experts, or usually representatives from DEFRA or people in the poultry industry, and ask them for their opinion, and these would be clustered as 'expert opinion' (..)."

Mathematical model: Individual based stochastic simulation model

"Individual based stochastic simulation model and it basically means that you are representing every individual farm in the country, so each one of these have their own position in the model, spatially explicit, because these farms have a particular location and we know where they all are. Stochastic model because an epidemic propagated by a stochastic process [random process - sequence of events]. So an infected farm may have a risk to transmit to any other farm in a particular day, and this may or may not happen. So, a simulation may say, yes it did happen, and then, which were the subsequent events that follow from that. (..) So each time that you run these simulations, you get a different answer, and in a sense that echoes with the outbreak, because each times (..) a farm [is infected] you are going to get a different output, sometimes it would trigger the outbreak immediately, sometimes you detected it and take steps to prevent pass[ing] it on. Sometimes, it's passed on and causes a major outbreak. So, by running many different simulations, you might get the spectrum of possible scenarios and probabilities associated with it."

"The analysis takes place before it happens, this work was done before we have a single outbreak in the UK, apart from the case in X [which was not actually in a farm, it was a swan with the infection]".

On 4 June 2008, Defra confirmed that the Avian Influenza present in laying hens on a premises near Banbury in Oxfordshire is the highly pathogenic H7N7 strain: http://www.defra.gov.uk/animalh/diseases/notifiable/disease/ai/index.htm

"So, the work was done in advance of any outbreak in the poultry industry, only just before, actually. But the question addressed is, given an outbreak, given that a particular farm is infected, what is going to happen? So, it is not about preventing an initial outbreak, it is about once we have an outbreak what is the probability that it is going to infect 50 farms. What is the probability that is going to finish out and it is not

going anywhere; and then, if we introduce various strategies, such as the protection and surveillance that were published around the time, what is the impact that it's going to have; is this a good strategy? Because the government is faced with this problem of an infection coming into the country, and (.) wants to know the right policy (..). But it needs a range of potential policies that it could choose, and needs to know in advance which one is going to be the most effective."

That means the whole point of the model is to incorporate all the possible information available, analyse and interpret it and then answer these questions by running the simulations, finding the best control stop strategy in that context. The system is very complex: "(..) in a sense, these are emerging properties of the system, so you can not really answer the question of whether a ten kilometre surveillance area is going to have any effect, unless you actually put it in the model (..)."

QUESTION: 'Is the government strategy for controlling the disease an effective one, or is it a waste of money? You have to balance the effect of the strategy against the cost of the strategy."

"The **first outbreak** in a farm was around that time [he is not totally sure, but he recalled to be few days after the paper was published online 24th of October]. So 'it was quite nice', and the outbreak did no go anywhere [infected a farm and stayed where it was], and this was really our conclusion, that the majority of randomly seeded infections were not actually propagated actually to anywhere. So, it was quite rewarding to find out."

"In this case, the model was consistent with what actually happened. Once you have an outbreak (..) you start to query everything in this model."

"Since then, there had been two or three outbreaks (..) and each of them actually did not actually propagate any further."

Validating a model: "How do you actually establish (..) [this]? You can justify each of the assumptions you made, and you can say this is reasonable, that is reasonable, etc but unless it actually corresponds to reality, then there is no real validation. Because it is a statistical model, unless you have a thousand of outbreaks at least, then, you are not really going to be able to match the predictions of the statistical model properly in reality, but it is as far as you can validate the model I suppose."

Data Analysis and simulation model

Implementation of the model (code written in MatLab)

"It was done completely from the scratch. We have no real prior experience of this sort of modelling. So, I wrote the code in MatLab. I had quite a previous experience in MatLab, and what I did, it was basically to determine all of the components that we want in the model. Things like latency time, we also want to incorporate protection and surveillance zones. We want direct contact tracing, so, when you have an infected farm, we want to find out all the farms that have been associated with that farm, or in contact with it, to trace those and isolate those. So, [I] have the list of all of the components, and then, I just set up about writing the computer code, to do that; and it was a substantial piece of code, because I knew that I'd have to run many simulations, I tried to optimise that code as much as possible. So quite a lot of time was spent on making sure that it was right from the beginning and also that it was as fast as possible. (..) We had six months time to have this up and running."

“At the point at which we started this, and the subsequent four months, we did not have any data because the poultry registry data was not fully completed and compiled. So, we were doing this without any data; it has to be designed in such a way, that it was as adaptable as possible. Because, we did not know the questions that we wanted to ask. We could not be sure that we knew that we wanted direct contact tracing, or we did not know that we didn’t, therefore, the important thing at the first stage was to highlight all of the things that we wanted, and also, all of the things that we may want, to make sure that the code was adoptable enough (..). Then, we could do that without too much hard work. But the problem is, once you have a code in place and it is very much optimised for a particular problem, then, actually putting in additions to it (..) is difficult. So, I was very clear at the beginning that we need an outline of the code in the most general form (..).“

MatLab running in Condor pools

“Because it was such a complicated code, despite trying to optimise as much as possible, I knew that to get any sensible output I had to run millions of simulations and this in a single computer was not going to be feasible.”

“I can not remember when I come across this condor form, maybe I just looked at the university Website, and seen something about ‘High Throughput Computing’, So I knew that this thing was available, and I knew it has to do with ‘trivially parallel code’, so that is the key feature. So, basically if you run my simulation on one computer, and then, running the same simulation on another computer it is going to be twice fast. (..) So, with two computers you will double the speed. With 100 computers, you will have 100 times the speed. So, I knew that this would be a real solution. So I looked at the details of it, but I did not realise that they were issues with MatLab. So, I contacted the computer services department, and this was earlier, it may be 2 months or 2 months and a half into the thing, I had not finished the code. I said I am working on this code and at some point I am going to have to run millions of simulations with it, is it at all feasible with MatLab? They said: ‘no, it is not’. There were issues.”

“The first thing was a licensing issue, as MatLab needs licenses to run. We could get round that because you can compile MatLab to get a stand alone version of it, which does not need a license. So, you can run it on any computer without a License. So, the idea was that we could get a compiled version, and then, submit that to Condor. But condor has a requirement that the executable file (..) has to be a single file; while MatLab has the executable file and the whole package of additional run-time libraries that come along with that. So, it was not in the right format for condor. So, I was told by computer scientist that this was not compatible. I carried on working, because I will get results anyway, even if I could run just a few thousands, or a hundred thousand of simulations. I would still get some nice results, and it seemed to me that the most significant point was to get the model as accurately as possible, and never mind that I could not run it as much as I would hope to run it; and then, a couple of months later, someone from the computer science department contacted me – I had not been aware that they would have been still working on the problem – to say that they got it sorted. They solved the issue of how to run MatLab on Condor.”

“The computer scientist wrote a Web page so that anybody could go along and actually use it. They actually asked me to follow that Web page and see if I could actually implement it with that. So, yes, I did that. I think it was about a month and a

half before I finished the code. So, I did not do it immediately. I started looking at the Web page, and then, later on I started to implement my code when it was running fully in condor, and yes, it basically solved the problem."

Number of simulations and paper publication: All together the interviewee runs many millions of simulations. About 50-60 million simulations have gone into the paper, but, obviously, he run a lot more than that to analyse the system, and then, to determine what to put in the paper.

"The information for using condor was put on the website, and then, what I did was to implement a piece of code that the computer scientist has written, which was basically to wrap up all the files that were needed into a single file, and then, I could submit that file to condor from my own computer, and then, the script that the computer scientist has written was to unwrap itself and run itself, so this is how we went around the issue of multiple files. So, the way that we submit to condor is to have a 'submit host', which the computer scientist controls and I have access to, so I could log into this computer, and have my files on it, and then, I could simply submit these files to condor, and then it would do all the work and it would produce all the hosted output files [one for each simulation], and then, I would write another piece of code, and would join all these outputs together in order to have one single output file. This is the way it worked."

"In terms of speed increase, (..) I submitted a file that contained maybe 50 or a hundred of simulations each; so it would run for a period of an hour or something each submission, and then, I may submit a thousand of those, to see how to multiply the numbers together; to have a total number of simulations. If you have 300 computers, then in a day you are going to do a year's work, or what it would take me a year on my laptop or desktop. The benefit is massive. This was a proper time, this condor system; taking a few numbers of computers at the university, in principle by now, it will be over a thousand."

"I could have run it, just on my computer, left it going for a year, and come back. In that sense, it is fully automatic. It is just waiting a year for what I could have done in a day, it is not very sensible, isn't it? (..) It is about the fact of being able to run 400 simulations at the same time instead of one. That is the thing."

Collaboration

Discussion – Collaborate with colleagues (coordinate with them, etc)

As the project is running over a relatively short time period the work is quite intensive. Initially it was a collaboration between a Maths department and a vet School from the interviewee's university and a second university. But eventually the project was split (the second university followed a more statistical approach in comparison to the simulation approach presented here), leaving the two departments at one institution to work together. "In terms of the grant application, it was good to put things together; however the diversion was a healthy thing."

"What we have were meetings as when we need them, I guess on average every other week, and sometimes twice a week. Other times less. Just as whenever we need to get together, we agree on the way forward. So, the way it worked was: so we have a meeting and we decide what we want to do in the model, and then I went away and programming it, and then in the process of programming it, I had a lot of questions, so

we go back, and have another meeting. Basically, systematically go through all these questions, agree on answers to them, so that these meeting can go for a whole day; however we always managed by the end of these to reach an agreed consensus. And get back, and maybe come back with some more questions or another meeting. If it can not be easily done by e-mail, then we would have another meeting."

"It worked like that, and it was absolutely vital that we meet up regularly, and I think that a big problem with other projects that I have been involved in, it might be that the ones that are being called collaborators are being in a distance of hundred of thousand of miles away, which makes impractical as far as I can see, to have a close working connection."

"In this project it was absolute vital and very important that mathematicians meet up, and agree and they can see and understand our problems, and we could understand them, in order to reach to a consensus."

"I was involved primarily with the computing, how to get condor working, writing the MatLab code, and all that side of things. The vets took on the task of interfacing with DEFRA, getting the data, the data arrive in a completely unusable form, and some sites where duplicated, and data has to be clean and that sort of things. So they took on that and also, they had good contacts with the poultry industry. So they arrange a meeting were we actually could talk to them and ask our questions. This was very valuable."

Regular face-to-face meetings are important, complemented by phone calls and e-mails – the latter being effective especially after "we knew what each of us was doing". "(..) but certainly at the beginning of the project, we needed to have weekly meetings at least, and these were essential."

Dissemination

"There was a journal paper out of it. Journal of Mathematical Biology by Springer."

Journal paper + related data into a repository (confidential data; model structure): The interviewee is generally quite open to the concept of connecting the underlying source data with the publication itself and thinks that journals should demand this more in the future. In the project at hand the data had a confidential character and therefore could not used in such an open way. The structure of the usedf model further complicated the publication of data: "We had a time frame of six months to do this work in, actually, extended rather beyond that. The model is highly complicated piece of code, so I would want to address it, of course, before putting into Internet, and I simply do not have the time to do that. I do not actually think that a model of this complexity, other people can take it or get a great deal of interest in it. It is something that in an ideal world I would have done."