Mathematical models in the evaluation of health programmes

Geoffrey P Garnett, Simon Cousens, Timothy B Hallett, Richard Steketee, Neff Walker

Modelling is valuable in the planning and evaluation of interventions, especially when a controlled trial is ethically or logistically impossible. Models are often used to calculate the expected course of events in the absence of more formal assessments. They are also used to derive estimates of rare or future events from recorded intermediate points. When developing models, decisions are needed about the appropriate level of complexity to be represented and about model structure and assumptions. The degree of rigor in model development and assessment can vary greatly, and there is a danger that existing beliefs inappropriately influence judgments about model assumptions and results.

Introduction

Although policy decisions in public health would ideally be based on evaluations that measure effect directly, modelling does, and often should, play a major part in large-scale evaluations. The purpose of our Review is to assist readers to assess critically and interpret appropriately the results of such modelling exercises. We first present practical and theoretical reasons why models are and should be used in large-scale evaluations. We then summarise the types of different modelling approaches and discuss how models and their outputs should be judged.

The role of models in programme evaluations

Public health programmes need to be evaluated on whether anticipated benefits are indeed happening and whether they are cost effective. For many programmes, a randomised controlled trial is not an option for ethical or practical reasons (eg, programme implementation is done by large groups, such as national health systems, which cannot be randomised). Alternatively, the evaluation question might not be whether an intervention is effective, but whether it is being successfully implemented. In these cases we have to observe trends in health outcomes with the programme in place and model what would have happened without it. Even when a community-randomised trial is feasible, other activities might interfere with the effect of the intervention, especially if the number of communities that can be randomised is small. In such circumstances modelling is likely to play an important part in the evaluation for one of several reasons (figure 1). First, we might wish to estimate the causal effect of the programme on the health outcome of interest in circumstances in which other changes, which might affect that outcome, are also taking place (figure 1). Second, we might wish to estimate changes in a final health outcome, which is difficult or costly to measure. In these circumstances, we might prefer to measure changes in a more easily measurable intermediate outcome and model the consequences of these changes on the final endpoint of interest (figure 1). Third, we might wish to explore the relation between intervention and outcome to better understand findings from field studies. The scope of the models depends upon whether all the factors that lead to a health outcome are described and their respective roles accounted for, or whether the model is used as a quantitative framework with which to translate intermediate outcomes into a health effect.

In the first case (attribution of effect), simply measuring changes in the health outcome of interest over time will not provide an estimate of the causal effect of the programme, because changes could have happened even in the absence of the programme. What is needed is an estimate of what would have happened in the absence of the programme (the counterfactual). This estimate can then be compared with what happened in the presence of the programme (the actual) to estimate the effect of the programme. Modelling can often be the only way to obtain an estimate of the counterfactual of interest.

An example is establishing the counterfactual for the trend in HIV prevalence in the context of a mature HIV epidemic. The effects of saturation and AIDS deaths mean that simply continuing the past trend is unlikely to show what would be seen in the absence of any intervention. For programmes that seek to change HIV risk behaviours, the recorded prevalence data can be compared with the prevalence that would be expected in the presence of saturation and population turnover but in the absence behavioural change. Such an approach was used in the recent Avahan and Global Fund evaluations, making the assumption that individual behaviour change can be attributed to the programme. In both of these exercises the uncertainty in the model parameter values and the data were used to create a counterfactual distribution, rather than a single value (figure 2).

There are many situations in which measuring the health outcome of interest is difficult or costly. Two common reasons for this are the outcome of interest is a rare event, or the outcome of interest will not happen for a long time. In both of these situations, evaluations can measure intermediate outcomes that are more frequent, or happen sooner, and then use a model to extrapolate to the outcome of interest.

A rare event, in this context, refers not only to conditions that only happen in a small number of individuals each year—eg, infection with the Ebola virus or variant Creutzfeldt-Jakob disease. Rather, many important public health outcomes are rare events when it comes to their measurement. For example, the primary indicators for the major health-related millennium development goals—the
Review

under-5’s mortality rate and the maternal mortality ratio—
cannot be measured with much certainty in the most recent
year in high-mortality countries because vital events are not
routinely and accurately registered. Even in large sample
demographic and health surveys (DHS) mortality in
children younger than 5 years is generally estimated for the
5 years before the survey. Maternal mortality, because it is a
less common event (often measured per 100000 live-births),
is even harder to measure reliably with such surveys.

Evaluation of a measles vaccination programme, whose
goal is to reduce childhood deaths from measles, provides an
easy example of how we can use a model to extrapolate from
a more common intermediate event to a rarer, but more
important, outcome. For the programme to have any
effect, several conditions must be met. First, we need to
make sure that measles vaccine is available for distribution
within the country. Next, the vaccine needs to be distributed
to health facilities. Then, young children need to be
vaccinated at the right age (late enough to minimise
interference from maternal antibodies, but early enough
to protect them before they are exposed to the wild virus).

This sequence of events is a causal chain that explains how
measles vaccine reduces measles deaths (figure 3). Directly
measuring changes in the number of measles deaths in
children younger than 5 years is difficult, if not impossible,
in settings where measles remains a problem. Instead of
trying to measure measles deaths, we might measure the
number of measles cases in children and estimate the
effect of the measles vaccination programme on deaths
due to measles with a model.

We could further reduce data collection needs by
extending the model further back up the causal chain
(figure 3). For example, we could simply measure measles
vaccine doses available in a country, or the proportion of
children vaccinated, and use this information to estimate
the reduction in cases of measles and deaths. A recent
example of this approach estimated the effect of measles
vaccine on child mortality, as a starting point they used
the number of vaccine doses given in a country. However,
when intermediate outcomes are used as proxies, the
further removed the intermediate outcome is from the
outcome of interest the greater the uncertainty in the
estimated effect on the modelled outcome of interest.

For example, a model with measles vaccine availability at
country level as the intermediate outcome to estimate
measles deaths averted, would produce more uncertain
estimates than a model that uses the proportion of
children vaccinated.

The evaluation of human papillomavirus (HPV)
vaccination programmes is an illustration of a situation
in which there is a long delay between the intervention
and the outcome of interest. The health endpoint of
primary interest is cervical cancer, which is rare and can
take many years to develop. To assess HPV vaccination
programmes, rather than measuring the effect of
vaccination on cervical cancer, we measure vaccine
coverage and use models to predict the effect on infection
with HPV, the precancerous lesions that will be detected
in screening programmes, and the subsequent incidence
of cervical cancer. Again the intermediate outcomes,
HPV infections and lesions detected, could be measured
to get closer to the event of primary interest (cancer rates
and deaths).

Modelling is often used to aid interpretation of
randomised controlled trials, especially trials of complex
interventions. The endpoint of a trial is reached through
a causal pathway that can be influenced by many contextual
variables. To understand the recorded effect size and be
able to generalise findings to other contexts, models can
be used to link trial outcomes to different intermediate
and process measures. These measures can then be
compared with values from other settings to explore the
potential effect of the intervention. For example, the
effectiveness of circumcision to prevent the transmission
of HIV at the population level will probably depend upon
the levels of sexual risk behaviour in the population. On
the basis of measured efficacy of adult-male circumcision
in reducing the acquisition of the infection by circumcised
men, models have been used to predict the population
level effect of circumcision in a range of different types of
HIV epidemic.

In evaluations of complex programmes with many
interventions, models can be used to explore the
expected contribution of different elements to the overall
outcome of interest. In trials, there is a trade-off
between the isolation of an individual treatment being evaluated and maximising the expected effect size by combining
intervention activities and treatments. This trade-off
often generates circumstances in which models can be
used to estimate the role of the different components
and to identify a minimum package of treatments.

---

Figure 1: Framework for programme evaluations involving modelling

A

Programme of interest

Other factors affecting endpoint of interest

Health outcome of interest

Scope of model

B

Programme of interest

Intermediate outcomes

Health outcome of interest

Scope of model

www.thelancet.com Vol 378 August 6, 2011
Figure 2: Counterfactual projection for generalised HIV epidemics
Projection of 2007 HIV prevalence based on data through 2002 under the assumption of no subsequent behaviour change (simulated data). The data from ten antenatal clinics (circle and lines: each circle shows one point estimate from one clinic and the lines connect estimates from the same clinic) and a large household survey in 2005 (blue square, A). The previous limit on prevalence in 1984 is shown as a vertical bar. Model projections assuming no behaviour change (B): each thin path shows one item in the posterior distribution and the thick path shows the best supported projection. Distribution of the expected change in prevalence between 2002 and 2007 with the counterfactual model (i.e., without behaviour change, C). The vertical line shows the naive null assumption of zero change. Distribution of the expected prevalence in 2007 with Model M0 (i.e., without behaviour change, D). The vertical line shows the prevalence estimate in the baseline survey in 2002. If future observations show HIV prevalence in 2007 outside of these ranges, then there would be evidence for behaviour change affecting the course of the epidemic. Model based on Hallett and colleagues.5,6

Figure 3: Causal pathway for measles vaccine effect on measles deaths

Models can also be used to assess how an intervention should be rolled out. In the case of the randomised trials of HPV vaccine, because of the efficacy of the vaccine, quite simple models show that vaccination is cost effective.25 However, models are also needed to assess how a vaccination programme should be designed (i.e., who should be vaccinated), because the possible options are too many to test in trials.6,7 In practice, evaluations can deal with both the attribution of effects in the presence of confounding factors and a lack of data on the health outcome of primary interest. In many instances, a planned assessment that involves the use of
modelling is the only practical option. In such circumstances we should ask whether the modelling approach used was appropriate and whether the uncertainty inherent in the modelling process is reasonably shown in the results.

Mathematical modelling methods
Mathematics provides a precise quantitative language to describe the relation between variables and changes in states, and in medicine we can represent mathematically the clinical course of disease, the distribution of disease across populations and over time, and the mechanisms that generate disease. The development of a model of infection, disease, or death requires us to precisely set out our assumptions about the parameters and processes influencing health, and enables us to calculate the expected effect of programmes. Here we discuss different approaches to modelling and identify a short list of key model characteristics (table).

We draw a distinction between purely statistical models, which describe associations between variables, and mathematical models, which set out a theoretical framework that represents the causal pathways and mechanisms linking exposures, interventions, and infection or disease. Broadly, statistical models are those used to derive parameter estimates from empirical data, and mathematical models are those used to make predictions on the basis of those parameter estimates.

The development of a mathematical model involves the definition of equations or sets of rules that describe the

<table>
<thead>
<tr>
<th>Basic model approach: mathematical or statistical</th>
<th>Use</th>
<th>Strengths</th>
<th>Weaknesses</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mathematical models</strong></td>
<td>Provides mechanistic representation for how disease burden is established</td>
<td>Predicts the incidence and prevalence of disease; understanding implications of causal pathways</td>
<td>Makes explicit assumptions; allows predictions beyond range of data; generates testable predictions</td>
<td>Relies upon a good understanding of pathways or assumptions; can be disassociated from data</td>
</tr>
<tr>
<td><strong>Statistical models</strong></td>
<td>Summarises association between variables, expressing the relation with parsimonious functional forms</td>
<td>Used to quantify the association between variables without assuming a pathway</td>
<td>Flexible approach that can be used quickly and routinely; only a few clear and testable assumptions made about form of data; properties well defined and understood</td>
<td>Provides only a description of the data not an exploration of how the system works (ie, shows associations, not explicit causal pathways)</td>
</tr>
<tr>
<td><strong>Feedback in the system: linear or non-linear</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Linear</strong></td>
<td>Represents the incidence rate of disease in different populations as constant or a linear function of environmental factors, including disease prevalence</td>
<td>Modelling progress to chronic diseases; often used in health economic analysis</td>
<td>Simple to construct and analyse model</td>
<td>Ignores indirect effects; in particular, for infectious diseases where averting one infection can terminate a chain that would otherwise lead to many more infections; thus linear models can underestimate or overestimate the effect of interventions</td>
</tr>
<tr>
<td><strong>Non-linear</strong></td>
<td>Transmission of infectious disease is expressed as function of current disease prevalence, number of susceptible individuals, and the risk of infection</td>
<td>Explores the indirect (ie, downstream consequences) effects of changes in the system; exploring dynamics of the system and its sensitivity to parameter values; making counterfactual projections for the course of epidemics without intervention</td>
<td>More complete description of the system; some dynamic effects can be important</td>
<td>Dynamic behaviours can depend on parameter values and assumptions in the model</td>
</tr>
<tr>
<td><strong>The role of chance: deterministic or stochastic</strong></td>
<td>Representation of model system is approximated for (infinitely) large population; every realisation is the same for a given set of parameter values</td>
<td>Represents the average behaviour of the system applying rates of change</td>
<td>Often easier to solve and analyse than stochastic models; in some systems gives the same results as the average of several analogous stochastic simulations</td>
<td>When chance plays an important role in system’s behaviour, deterministic model will generate unrealistic predictions</td>
</tr>
<tr>
<td><strong>Stochastic</strong></td>
<td>Stochastic simulations represent events (such as incident infection and death) happening to individuals with a certain chance; the model, run multiple times with the same parameters, will give different results</td>
<td>Studying systems where there are small populations (eg, in clinical trials, or the occurrence of infections in a hospitals) or where, because at times with small numbers of infections, stochastic variation makes an important contribution to the system’s overall behaviour</td>
<td>More accurate in some circumstances; allows the variation in predictions due to chance to be quantified</td>
<td>Analysis can be difficult and computationally intensive</td>
</tr>
</tbody>
</table>

(Continues on next page)
spread of infection or the development of disease, where there is analogy between the mathematical representation and the biological system. The model’s predictions are obtained, either analytically, with the relation between inputs and outcomes known for all input parameter values, or numerically, with the results generated in steps for only certain sets of parameters. With numerical analysis, a range of different sets of parameters are chosen for only certain sets of parameters. With numerical

fundamentally non-linear processes. This arises because the rate of spread of infection is determined by the rate of contact between susceptible and infected individuals, and the abundance of infected individuals is itself a function of the rate of the transmission of the pathogen. Therefore, it does not follow that halving the risk of contracting an infectious disease in a population will reduce the incidence of that disease by half. Disease incidence could change by less than half (if the epidemic has the potential to spread extensively), more than half (if the epidemic is more fragile), or be eliminated (if the epidemic is close to its threshold for persistence). This complexity necessitates more involved descriptions of the relation between risk and incidence for infectious diseases.

Models can either be stochastic or deterministic. Stochastic models include chance explicitly in their results, with the different possible events selected with random number generators based on the probability that they will happen. The role of chance is important when examining systems with small numbers of individuals or events, such as the spread of meticillin-resistant *Staphylococcus aureus* in hospital wards or smaller geographical regions, but less so when common events in are modelled in large populations where deterministic models, which predict the average outcome, can be used. A deterministic model will

### Method of representing individuals: populations or individuals

| Population based | Track changes happening to groups within the population without specifying which individuals are involved; heterogeneity between individuals represented by dividing population into subgroups | Examining population-level processes and trends | Represents the experience of typical individuals; computationally efficient or analytically tractable | Misses individual level influences that might be important in certain circumstances | GOALS model,37 Avahan assessment,30,31 analysis of HIV prevalence trend data35

| Individual based | Model explicitly tracks sets of individuals over time and all events are tied to individuals, such that the life-course of each individual is known | Representing the life experience of individuals, exploring the interaction of individuals | Allows the history of every individual to be tracked, allows the importance of individual-level interactions and network structures to be explicitly represented | Models are very difficult to construct and analyze fully, a substantial amount of information is needed to establish parameters for the model, behaviour of system may not be fully understood, complexity of model can give false impression of credibility | Onchosim for exploring transmission and control of onchocerciasis,41 STDSim model that tracks individuals infection with multiple sexually transmitted infections and simulates complex pattern of sexual partnership formation between individuals,41,46 models of sex partner networks that simulate individuals have more than one sexual partner at the same time

### Changes modelled with respect to different variables: age and time

| Age | Models events experienced by individuals as they age, sometimes a single birth cohort sometimes multiple birth cohorts | Used to represent the typical experience of a cohort across time | Straightforward description of life experience with alternative interventions | Assumes that environment is stable | Cervical cancer screening,47 calculating cost-effectiveness of chlamydia screening

| Calendar time | Models the changing pattern of events over time. Allows for dynamic interactions between individuals over time | Used to model the spread of epidemics over time | Explores transient behaviour with introduction and uptake of interventions and the development of their effects over time | Often missing data on early periods; other dimensions, such as geographical and social space could be represented | GOALS model,37 Avahan assessment,30,31 analysis of HIV prevalence trend data35

---

**Table: Types of models**
produce the same results each time it is run with a given set of parameters, whereas a stochastic model will produce different quantitative results each time due to the play of chance in the course of the simulation. In some cases, a stochastic model and the analogous deterministic model can produce qualitatively different results, for instance when chance events trigger repeated rises and falls in the spread of the disease.53,54

The distinction between stochastic and deterministic models is not synonymous with the level of detail in the model, and in many cases the same model structure can be implemented with either a numerical integration algorithm (for deterministic results), or Monte Carlo simulation techniques (for stochastic results).

Another important distinction is between individual-based (microsimulation) models and population models. In the former, each individual in the population is explicitly represented and tracked. By contrast, in a population model, individuals are only represented as members of a group, with groups explicitly represented and tracked. In an individual-based model of a small population, the characteristics of each individual can be different and this approach has been used to model household or institutional outbreaks of infection.56 Often, however, many individuals are ascribed the same characteristics, but with events tracked for specific individuals. This approach allows larger populations to be modelled as individuals more efficiently.57,58 Not all stochastic models are individual-based; however, all individual-based models will be stochastic. Individual-based models are powerful for two reasons. First, they allow an individual’s experiences to be tracked, and second, they allow a great deal of individual level detail to be included in the analysis. However, they can be computationally intensive to use and difficult to interpret. Therefore population models, including important heterogeneities by stratifying the population according to different criteria, are still often used.

In infectious disease, heterogeneity in the risk of acquiring and transmitting infection can allow an infection to invade a population when the same average risk in a homogeneous population would not allow this. However, this same heterogeneity that makes invasion more likely will also restrict the spread of infection. A classic example of such heterogeneity is provided by HIV, where a small fraction of a population with many contacts, such as sex workers and their clients, can acquire HIV infection while spread throughout the remaining population is limited.59,60

With such heterogeneity, there is the additional parameter of the form of contact between the groups, which could be completely random or not (eg, people in similar social and demographic groups might be more likely to be contacts of one another, often referred to as assortative mixing).56 For instance, if the population of sex workers mixed randomly with everyone else, the spread of HIV could be greater than if they exclusively had sex with a small group of regular clients.56 A homogeneous model (distinct from a model with a randomly mixing population divided into different risk groups) assumes that everyone has the same risk of infection or disease, and this is often a starting point for models since it is analytically the most easily solved.

The assumptions that a model makes about how the status of individuals within a population changes over the course of infection or disease can be described in flow diagrams. Figure 4 shows an example of a simplified flow diagram for hepatitis B in a homogeneous population.59

![Flow diagram describing the epidemiology of hepatitis B virus, which can be translated into a mathematical model](image-url)

**Figure 4:** Flow diagram describing the epidemiology of hepatitis B virus, which can be translated into a mathematical model. Susceptibles (S) acquire infection, which is acute (A), this then resolves and they either recover (R) or retain a chronic infection (C). Boxes represent these mutually exclusive states called state variables. The arrows represent flows into and out of the states. The parameter values control the rates of flow between the state variables. The ordinary differential equation model represented by this flow diagram is shown alongside in words and symbols—colours and numbers show the relation between the model terms and the arrows in the flow diagram. In the case of susceptibles, dS/dt means the rate of change in susceptible numbers with respect to time. This is given by the addition of births and numbers show the relation between the model terms and the arrows in the flow diagram. In the case of susceptibles, dS/dt means the rate of change in susceptible numbers with respect to time. This is given by the addition of births and deaths and infections, where λ is the per-susceptible incidence. This per-susceptible incidence can be a function of the infectious population made up of those with acute and chronic infection: for example,

$$\lambda = \beta A \cdot \psi C / N$$

where β and β’ are transmission coefficients, representing contact rates and transmission likelihoods combined, for acute and chronic infections respectively. Acute infection resolves at a rate δ and a proportion δ develop immunity and (1-δ) develop chronic infection. Infection is the non-linear term in this model as the susceptible are multiplied by β and β’ representing contact rates and transmission likelihoods combined, for instance, if the population of sex workers mixed randomly with everyone else, the spread of HIV could be greater than if they exclusively had sex with a small group of regular clients.56
The boxes describe individuals in mutually exclusive health states and the arrows describe movements into or out of these states. This can then be used to derive the model equations. The rate at which the number of individuals in each box changes (the left hand side of the model equations) is determined by the arrows, which are the terms in the right-hand side of our equations. The crucial detail of the model is the content of these terms. In our example the infection term moving individuals from susceptible to infected has an important influence and includes a non-linear term representing the interaction between infectious and susceptible individuals.

Our hepatitis B model (figure 4) is constructed to look at changes with respect to time, but it could have been constructed to look at changes over age instead (in this case it would be necessary to assume a chance of infection that is fixed in time), or over both. Ordinary differential equations are used to specify the model when only tracking changes in the states over time. If it is important to track changes over other factors too, such as age (modelling progress to cancer), geographical space (modelling the range of vectors), or time since an event (from an infection to cancer); partial differential equations can be used. Alternatively, discrete categories representing age groups or geographical space might be represented. Differential equations represent time and age as continuous variables but many models, especially in demography, represent changes in discrete steps with difference equations.

In deciding which models are appropriate to answer a particular evaluation question it should be remembered that all models are a simplification of reality. Sometimes difficult decisions must be taken about how a system should be represented.

Assessing models

Panel 1 provides a summary checklist of items that should accompany the most rigorous model analyses. None of these indicators of model quality guarantees that the model produces accurate outputs, but the presence of this information will help readers assess the appropriateness of the model.

It is essential to have a clear understanding of the model’s structure to judge model results and outputs. We believe that in addition to the full technical description of the model, there should also be a clearly presented and easily accessible lay summary, which includes flow diagrams of the way the disease is represented, a table showing how the model parameters were assigned, and the sources of information for the parameter values.

The specification of model parameters is not simply a listing of what parameters are included (that should be evident from the flow diagrams or the equations), but precise definitions of the parameters, as well as their values and ranges. It is informative to identify where the parameter values came from, and whether or not they are based on observation, are extracted from previous modelling studies, varied depending on other parameters in the model, or estimated in the model analysis through fitting to data.

Model validation implies that a model has been tested repeatedly across the full range of scenarios or settings to which it is supposed to apply, and has been shown to reproduce closely the independently recorded relation between input, intermediate, and outcome variables. Few, if any, models will have undergone sufficient testing across a range of settings and situations to be deemed fully validated. At a minimum, however, presentation of a model should be accompanied by an assessment of its goodness-of-fit to observed data and, ideally, how well the model predicts what was actually recorded (out-of-sample prediction). The distinction between how well a model fits recorded data and how well a model is able to predict what will be seen is an important one. Goodness of fit refers to an assessment of how well the model can replicate empirical observations when the parameters in the model are adjusted to achieve the closest match. A range of methods for fitting and assessing goodness of fit.
are available, and the method used for this assessment should be clearly reported. When the data available to estimate the parameters are sparse and resulting estimates have wide ranges of acceptable values a model can fit the data equally well with several different sets of parameters. This problem can be benign if the model behaves the same way with different parameter sets. However, in some situations, it is possible to get very different predictions with different, equally well fitting, sets of parameter values (figure 5). This uncertainty is often hidden, but should be explicitly addressed when presenting model fits. One approach to quantifying this uncertainty is to identify several different sets of well fitting parameter combinations and to run the same analysis for each of these combinations. The model results can then be presented as the distribution of outputs from these analyses.

Figure 5: Uncertainty in model parameterisation of epidemic context perpetuates to uncertainty in intervention effectiveness
Several different parameter sets can be fitted to the same data for urban Zimbabwe (A). The effect of the same circumcision intervention in women, uncircumcised men, circumcised men, and overall varies according to the set of parameter used (B).19

Out-of-sample prediction involves estimating model parameters with one set of data, with these parameter values within the model to make predictions and then comparing these predictions with observed data not used in the parameter estimation process. If, in a large number of instances and over a wide range of scenarios, the model predictions match the observation, this can be a powerful validation of the model (panel 2).

The presentation of sensitivity analyses and estimates of the uncertainty around modelled outputs is also important. The results of sensitivity analyses, in which parameter values are varied to establish how model outputs change, can provide important insights into the behaviour of the model. Sensitivity analyses enable readers to identify which parameter values are most influential in the model and which parameters should be most closely scrutinised. Sensitivity analyses, which can explore the effect of extreme parameter values, should be distinguished from uncertainty analyses, which seek to derive a range of credible results on the basis of an exploration of the range of reasonable parameter values. In sensitivity and uncertainty analyses, changes in a single parameter value, a limited number of parameter values or all parameter values could be explored. Parameters can be varied systematically, or chosen at random. Latin Hypercube Sampling for example, increases the efficiency with which multiple parameter values are sampled.48,68 The appropriate methods for uncertainty analysis will vary according to the ambition of the exercise and constraints imposed by model complexity and computing power. The important point is that the choice of method should be presented and justified.

Uncertainty analyses that explore the range of input parameters go only part of the way to the assessment of total uncertainty in the analysis, because uncertainty will also derive from the necessary decisions that were taken in the construction of the model itself. Two important elements in the assessment of the appropriateness of a model’s structure are qualitative and somewhat subjective. The first is construct validity: does the model’s structure and its explanation accord with the biological system? Have the important causal processes been included? For example, in a model of hepatitis B does the pattern of progression, from acute infection through to chronic infection to disease, make sense? Is the contact process modelled in a way appropriate to the known spread of this virus? The second qualitative test of a model is whether a convincing explanation of the results can be provided. For example, if a model predicts that low coverage of vaccination against rubella could increase the incidence of congenital rubella syndrome, this might reasonably be explained by a delay in the average age at which girls are infected such that they are more likely to be infected when pregnant. If the model structure makes sense but the model results do not match recorded patterns of infection and disease, then either our understanding of the infection
or disease is inadequate, or inappropriate parameter values have been used. Modelling can thus be used to review our understanding of an infection or disease. However, it is all too easy for the modeller to believe erroneous model results rather than challenge the validity of their model.69

In judging construct validity, it is important to understand the scientific rationale for the modelling decisions taken at each stage in the analysis. There is always a trade-off in the level of detail and complexity in models. Models described as simple can be easier to interpret and explain or better scientifically grounded than the models where the levels and causes of mortality are rather different from the settings used in the analyses? How well would the model perform for countries where AIDS is an important cause of child death? (In none of the settings examined was AIDS an important cause of child death.) These are questions which cannot, at present, be answered. Thus, although it might have been reasonable to use LiST in the Global Fund assessment, the model-derived predictions of deaths averted should be interpreted with caution.

Panel 2: The Life-Saved Tool

The Life-Saved Tool (LiST),32,64 is a model to estimate the effect of an increase in the coverage of a set of interventions on child mortality, and was used in the Global Fund assessments of the scale up of interventions to combat HIV, tuberculosis, and malaria.4 Most countries covered by the assessment do not have reliable data on malaria mortality trends but they do have measures of coverage of malaria interventions such as ownership of long-lasting insecticide treated nets. LiST was used to estimate the number of malaria deaths averted by scaling up malaria interventions, including insecticide treated nets, indoor residual spraying, and treatment. This is an example of a model used to extrapolate to an outcome of primary interest. Some work that assessed how well LiST predicts changes in mortality has been done.33,35 In general these comparisons found that the changes in mortality predicted by LiST fell within the confidence bounds of the measured changes in mortality.36

Was it appropriate to use LiST in the Global Fund assessment? Although the assessment used studies from three countries in south Asia and three countries in sub-Saharan Africa, this is a very small sample of settings and LiST cannot be deemed to have been fully validated. How well would the model perform for countries in Latin America where the levels and causes of mortality are rather different from the settings used in the analyses? How well would the model perform for countries where AIDS is an important cause of child death? (In none of the settings examined was AIDS an important cause of child death.) These are questions which cannot, at present, be answered. Thus, although it might have been reasonable to use LiST in the Global Fund assessment, the model-derived predictions of deaths averted should be interpreted with caution.

Panel 3: The UNAIDS/WHO estimates of HIV and AIDS

UNAIDS and WHO developed models to estimate adult prevalence of HIV and AIDS, along with AIDS deaths, based on HIV testing of pregnant women attending for antenatal care in countries in sub-Saharan Africa. UNAIDS organised an external reference group to guide the development of the model, published descriptions of their methods and did extensive work to investigate the appropriate parameter values to use in the model. This was an on-going process, where new data were compiled and assumptions were reviewed and updated on an on-going basis. Uncertainty bounds were presented and sensitivity analyses done. In general, these estimates and the modelling on which they were based adhered to all five of the indicators of model quality. However, when nationally representative data on adult HIV prevalence in countries with generalised epidemics became available from household surveys it became clear that the model overestimated HIV prevalence.

So what was wrong with the UN-led modelling on HIV prevalence and AIDS mortality? The major problems resulted from little data on which to base assumptions, especially with regard to the relation between prevalence in pregnant women in the surveillance system and all women, and the prevalence ratio between women and men. The limited data that were available had been compiled and the values for the parameters in the models were based on these data but these values were not representative of the values found in later national surveys. UNAIDS and its partners have continued to use the same general approach to their modelling work, but now with more data available to refine the parameter values used in the modelling.

work. Inclusion of detail that has no relevance should be avoided: for example, in most circumstances there would be little purpose in including the sex of the child in a model of chickenpox transmission,70 but in a model of rubella, where the outcome of interest is congenital rubella syndrome, sex is an important detail.71

Even if a model and its application have all or most of these quality indicators, it does not guarantee that the modelled outputs are correct. The UNAIDS/WHO model of the worldwide HIV epidemic before 2004 (panel 3) had all the hallmarks of a high-quality model, but it produced results that were subsequently shown to be overestimates. Here, limited information about the biases in the estimates of HIV prevalence in pregnant women attending antenatal clinics in the surveillance system relative to the general population caused an overestimation of the scale of the epidemic. Thus, following most rigorous model analysis standards, although necessary, is not sufficient to guarantee that the model results are accurate and using parameter estimates and modelling approaches that establish conservative estimates might be particularly appropriate to improve credibility in many situations.
Discussion

With recent increases in resources committed to improve global health, there is growing demand for accountability and efficiency in programme implementation and the need for good evaluation. In many situations, available data and appropriate modelling techniques can clarify, within a causal framework, the relation between programme inputs and effect. Modelling exercises that are well done can provide credible evidence of the value of programmes and guide the roll out and improvement of interventions. Unfortunately, mathematical modelling can be a confusing or daunting term and is often called and treated as a black box. We have examined the types of mathematical models that include the most common applications, although it is not exhaustive. We have attempted to provide guidance on the standards that readers should expect from a modelling paper, to promote more balanced and carefully considered view of modelling work in the hope that this enables modelling work to be more useable, more widely accepted and appropriately influential in this field.

It is important to remember that in initial programme planning a model might be most useful when good data is not available, and yet public health policy decisions must be made—as at the beginning of an outbreak of a new respiratory pathogen. The model is valuable here for the very reason that good data is not available, and stands alone as our only method in such a setting because there is nothing else. By the time good data is available it might be too late to implement certain policies. In such a setting a good modeller will emphasise the limitations, but we cannot avoid decision making in uncertain situations.

As modelling moves from an academic exercise in mathematicians to the public health field it is essential that modellers work to explain models clearly, are rigorous in quality assurance, and provide full documentation. Modellers need to be held to a high standard in terms of communication and technical skills, providing transparent analyses that can be replicated by others. To benefit public health, non-modellers need to formulate their questions systematically, engage with modelling and hold modellers to account. Concomitantly, due consideration needs to be given to the range of choices and technical difficulties that can be involved in modelling programmes and the time and resources needed for good quality modelling as part of evaluation, rather than modelling being an inconvenient and rushed add-on.

Contributors
All authors were involved in the conceptualisation, writing, and editing of this Review.

Conflicts of interest
We declare that we have no conflicts of interest.

References