On Probabilistic Models
for Surveillance and Prediction of Disease Incidence
with Latent Processes:
Case Studies on Meningococcal Outbreaks,
Childhood Diabetes and Poliomyelitis

Jukka O. Ranta

Division of Biometry
Rolf Nevanlinna Institute

Faculty of Science
University of Helsinki

Academic Dissertation for the Degree of Doctor of Philosophy

To be presented, with the permission of the Faculty of Science
of the University of Helsinki, for public criticism in Auditorium IV,
Porthania, on January 27th 2001, at 10 a.m.

Jukka O. Ranta

Division of Biometry
Rolf Nevanlinna Institute

Faculty of Science
University of Helsinki

Research Reports A34
January 2001

Rolf Nevanlinna Institute
P.O.Box 4 (Yliopistonkatu 5)
FIX-00014 University of Helsinki, Finland
ISBN 952-9528-61-2
ISSN 0787-8338
YLIOPISTOPAINO
Helsinki 2001

PDF version
ISBN 952-9528-62-0
http://ethesis.helsinki.fi/
HELSINGIN YLIOPISTON VERKKOJULKaisut
Helsinki 2001
Acknowledgements

I would like to thank all the co-authors for their valuable contributions during the work, especially Elja Arjas for his guidance and for being my supervisor over the years. I also thank Antti Penttinen for sharing his ideas on spatial statistics. I thank the whole INFEMAT study group, the pioneers of mathematical modelling of infectious diseases at the department of vaccines, National Public Health Institute (KTL). Many thanks belong to Pirjo Mäkelä for her valuable insights on meningococcal epidemics. Furthermore, I owe many thanks to the whole DiMe study group at the diabetes research unit, KTL, and Marjatta Karvonen for providing the data and for the encouragement on disease mapping. I also owe thanks to Tapani Hovi at the department of virology, KTL, for sharing his interest in modelling the environmental surveillance of polioviruses. I thank my colleagues at the Rolf Nevanlinna Institute (RNI) for all the educational discussions and debates over the years. Working at RNI has been a continuous learning experience. Financial support from the Graduate School of Mathematical Analysis and Logic, RNI and KTL is gratefully acknowledged. Last, but not least, I thank my parents and my wife Riikka for their warm-hearted support.

Helsinki, December 2000

Jukka Ranta

List of original publications

The following original papers are referred in the text by their Roman numerals.


(II) Ranta J, Mäkelä PH, Arjas E. Predicting meningococcal disease outbreaks in structured populations. Submitted for publication.


(IV) Ranta J, Hovi T, Arjas E. Poliovirus surveillance by examining sewage water specimens. Studies on detection probability using simulation models. Submitted for publication.
Contents

1 Foreword .................................................. 5

2 A Bayesian prelude ..................................... 6

3 Concise notes on a complex theme .................. 7
  3.1 What should a model describe? ................. 7
  3.2 Practical compromises ......................... 7

4 Interpretation and transferability of parameters ... 8
  4.1 The sampling experiment ...................... 8
  4.2 Bayesian perspectives ......................... 9
  4.3 Towards observables .......................... 10
  4.4 Universality as consensus ..................... 10

5 Back to the future ..................................... 11
  5.1 Posterior predictive distribution .............. 11
  5.2 Temporal predictions of infectious disease. Example: meningococcal bacteriad ......................... 12
    5.2.1 Explanatory modelling ...................... 12
    5.2.2 An extension to the basic chain-binomial model 13
    5.2.3 The usefulness of meningococcal carriage data 15
  5.3 Spatial predictions of noninfectious chronic disease. Example: childhood type 1 diabetes .... 16
    5.3.1 Empirical modelling ....................... 16
    5.3.2 Georeferenced register data and loss of information 17
    5.3.3 The scope of spatial modelling .......... 19
    5.3.4 Disease mapping and GIS ................ 20
  5.4 Surveillance of infectious disease. Example: poliomyelitis ................................. 23
    5.4.1 Models of large scale dynamics ........ 23
    5.4.2 Scenarios of disease incidence ........ 23
    5.4.3 Active versus passive monitoring in poliovirus surveillance . 24

6 Concluding remarks ................................... 25
1 Foreword

This thesis consists of four papers demonstrating the application of probabilistic modelling techniques in three different epidemiological settings. Two of the papers, I & II, describe a prediction problem of the development of a largely asymptomatic epidemic caused by meningococcal bacteria in a military establishment. Paper III is an application of disease mapping and detection of excess incidence of insulin dependent diabetes mellitus in small geographical areas. Paper IV is concerned with the surveillance of polio epidemics by detection of polioviruses in sewage waters. Although the applications are epidemiologically very different, they share common aspects in modelling and inference. Firstly, all models are probabilistic, three of them employ Bayesian inference. Secondly, predictions are often used as an easily interpretable device for presenting results.

Three of the papers deal with infectious pathogens (meningococcus, poliomyelitis), but one is about a noncontagious and chronic disease (diabetes mellitus). Only little is known about the aetiology of the latter, thus an explorative statistical method is used to detect ‘unusual’ geographical patterns of incidence in large databases, in an effort to develop new hypotheses and increase aetiological understanding. Also surveillance and evaluation of emerging cluster alarms play a role in such applications. In the other applications, aetiology is well known, but the development of an outbreak is difficult to detect, predict and control. Hence, methods are needed to give formal structure for quantifying the expert’s knowledge. The reasons for methodological choices are mostly pragmatic, derived from the application of probability theory as a Bayesian inferential tool in a cross disciplinary (ad)venture.

Sections 2-4 contain philosophically oriented discussions on modelling and applied statistics related to the background of papers I-IV. Section 5 is concerned with the particular modelling of outbreaks of infectious diseases and subclinical infections in communities (meningococcus) of non-infectious chronic diseases with nationwide georeferenced data (diabetes mellitus), and of environmental surveillance of subclinical infections in large populations (poliomyelitis).
2 A Bayesian prelude

*It is not that paradigms are under-determined by facts; it is rather that there are no paradigm-independent facts we can use to justify the adoption of one rather than another.* Barry Gower.

Among statisticians there has been a long, and still ongoing, debate about the validity of subjective Bayesian analysis and the (supposed) objectivity of classical frequentist analysis [1]. Traditions are slowly changing and the spectrum of acceptable scientific methods grows wider. During the last decade, papers advocating Bayesian analysis have appeared in e.g. *the British Medical Journal* [2], [3], [4] and in *Trends in Genetics* [5]. Also, p-values have received methodological attention e.g. in *the American Journal of Epidemiology* [6]. The history of what might be called 'scientific method' has been vividly reviewed in a book by Barry Gower [7] and several books on Bayesian method have been published [8], [9], [10], [11], [12], [13]. In Bayesian analyses, some habits of the old practice of statistics are rendered virtually useless and this may cause some confusion. For example, common statistical tests usually require a minimum sample size in order to be valid. A posterior distribution is defined for any number of experiments, e.g. any number, $n$, of individuals in an epidemiological study. Although technically correct, presenting only lack of knowledge does not make sense. Sometimes a small $n$ is a result of the investigator stratifying the data excessively. The fewer data, the less information we gain and there is no magic trick around this fact. But probability theory yields consistent results: with no data we are left with our priors! This is only one - and by no means representative - example of the interesting challenges and opportunities in applying modern statistical methods. For a discussion on Bayesian sample size determination, see e.g. reference [14].

There will always be arguments in favour of, and counter arguments against any methodology and paradigm. With modern day computers and algorithms, there are fewer computational and practical imperatives for immediately choosing a classical approach over its Bayesian alternative. One can choose a paradigm, but one can not do without one. It has been said [4] that "Most statisticians have become Bayesians or frequentists as a result of their choice of university". The full circle is complete: the classical direct sampling probabilities have their counterpart in Bayesian inverse probabilities. Hence, the two paradigms could be seen today as *completing* instead of *competing.*
3 Concise notes on a complex theme

Indeed, there is no such thing as a "stochastic process" in the sense that the individual events have no specific causes. One who views human diseases or machine failures as "stochastic processes" as described in some orthodox textbooks, would be led thereby to think that in gathering statistics about them he is measuring the one controlling factor; the physically real "propensity" of a person to get a disease or a machine to fail. E.T. Jaynes.

3.1 What should a model describe?

The incidence pattern of a multicausal disease, or an outbreak of a contagious pathogen, is a complex biological process. Sometimes we succeed in describing some aspects of these processes reasonably with linear regression arising merely from a statistical convention, sometimes with stochastic processes arising from elaborated descriptions of the phenomenon. There is a long tradition in statistics to start with a standard class, a toolbox, of statistical models, fit these to the empirical data, and then provide estimates and confidence intervals of model parameters. A more substance oriented approach would perhaps start with the properties of the phenomenon, then construct a formal description using any set of conditional distributions, and finally assess the quality of the whole description. The latter is sometimes called holistic, as opposed to analytic. It might also be called analytic in the sense that a problem is decomposed into a set of conditional distributions which define the joint distribution [12]. Such an approach possibly has the advantage of providing a deeper and more explicit insight into the whole problem. This includes both observed and hidden (latent) quantities, as well as uncertainties about these quantities and descriptions of their relations. Papers I, II and IV aim at holistic explanatory modelling of the phenomenon, while paper III is perhaps more about empirical modelling of the data.

3.2 Practical compromises

Unfortunately, there are many interesting phenomena of which we do not know enough in order to give a formal description of its real, or even hypothetical underlying biological mechanisms. Obtaining well defined measurements in order to provide data on all biologically relevant factors in a controlled experiment can be even further out of reach in many epidemiological applications. The less we know about the details of the study subjects, the more we need to confine ourselves to simple all-purpose statistical models as explorative descriptions of the data. On the other hand, if every individual in an epidemiological study is described in extreme detail, their attributes are hard to view as 'exchangeable' or 'identically distributed' random variables. In a sense, statistical models prosper between complete ignorance and excess of detail.

A historical review addressing empirical 'toolbox' modelling and explanatory 'substance' modelling can be found in reference [15]. The distinction between the two
approaches is not always clearcut. Usually, towards the former approach, a standard statistical method comes with software suited for quick and highly automated black box computation. Formal descriptions of an expert's knowledge about the biological properties of the phenomenon almost invariably lead to models without analytical solutions within a standard class of statistical models. Numerical computation techniques are thus needed for complex modelling. For Bayesian analysis, e.g. a toolbox called WinBUGS is being developed [16], [17]. While such general purpose software can greatly assist in computation, it also imposes some restrictions on what can be computed. Limitations of computer resources and the labour of programming have become an integral part of the work of modern statistical modelling. Implementing a complex model may require more intellectual efforts than constructing the model in the first place. The question is whether it is cost effective to laboriously search for complex model descriptions, or to settle for a standard summary of the data. Most likely, both are needed at different stages of research. Sometimes, an analysis may also be based solely or largely on expert knowledge without explicitly conditioning on real data from the actual situation addressed. Such types of probability models are gaining popularity due to their easy implementation with spreadsheets [18], [19], [20], [21]. In Bayesian interpretation, these models represent prior distributions.

4 Interpretation and transferability of parameters

The essence of the present theory is that no probability, direct, prior, or posterior, is simply a frequency. H. Jeffreys (1939).

4.1 The sampling experiment

As has been often emphasized [22], model parameters should have meaningful interpretations, i.e. a parameter should closely correspond with some real quantity of interest. It has also been discussed [23] that in all practical situations, such quantities are finite observations, whereas 'parameters' are rather abstract and idealized. But this is the very motivation of using models: we need to formulate our ideas of the physical world in order to make concise statements about it. It is often not only the population under study which we are interested in. We would usually like to "transfer" the estimated parameter value, e.g. the mortality ratio of patients with Neisseria Meningitidis, to describe other populations of patients which are reasonably 'similar'. For such learning to take place in a formal and well defined way, we need some apparatus to describe how our previous state of knowledge can be updated. For doing this, we need (i) models to describe how the finite observed quantities are related to each other and (ii) principles of performing inference on the common, though idealized parameters and forthcoming observations. Effectively, these are quite inseparable, but for simplification they might be viewed in two parts. The former task is the most difficult one, the specification of a model. The latter is a more technical question of inference. In any case, feedback between the two tasks is vital.
To start with a classical example of inference, the (static) probability $p$ of an arbitrary individual having a disease is the sampling probability of choosing one individual with a disease from a group of individuals, where $p \times 100\%$ have the disease. This is strictly the interpretation of probability in the spirit of John Venn 1888 [7], [24]. An obvious estimate of $p$ would be the fraction of disease cases in a random sample. It might be noted that even this interpretation involves subjective judgement of what individuals belong to the group, and of the equal (symmetric) chances of being selected in a sampling experiment. However, the modern use of probability theory in scientific reasoning is not restricted to cases where we have a population, upon which we can impose a real sampling experiment in order to make the conceptual connection between observed frequencies and probabilities. Such examples of (Bayesian) probabilistic inference were encountered in papers I-III.

4.2 Bayesian perspectives

The vocabulary of probabilistic argumentation ranges from measurement errors to subjective beliefs about unique events. For example, in ecological studies, there is often only one realization, a naturally occurring 'experiment', of a series of biological events we can observe. In geographical epidemiology, disease occurrence at a specific location during a specific period of time is studied. There is no population of realizations with a common spatiotemporal location at our disposal for estimating the chance of a particular realization of events from the observed frequencies. We can never collect additional samples of realizations under exactly the same conditions. Apart from well controlled and repeatable experiments in vitro, this is often true with observational data. The same remark applies to a single outbreak of meningococcical diseases in a garrison. No other epidemics in the history is a replicate of exactly the same biological conditions. Abstract probabilities are related with observations in a way which can resemble a sampling experiment. A frequentist interpretation of repeated experiments can reappear as a thought experiment that can be used to specify a prior distribution in Bayesian analysis, although this is not necessary. Subjective degrees of certainty can also be based on the judgement of symmetry and exchangeability of 'equally likely' events. For a further discussion, see e.g. references [7], [8].

In a Bayesian framework, probability models are often written in terms of observable quantities and unobservable parameters. The distinction is that, in principle, we could find out the true value of an observable quantity by direct measurement. In practice, it can be only too difficult or too laborious to find out. Generally, missing data can be seen as observable quantities. They could have been observed, although by now they could have been destroyed already. A parameter does not possess similar physical existence in the past, present or future, and therefore it cannot be objectively observed, no matter what experiments we set up. In a Bayesian definition, a 'probability', or a 'chance', is an unobservable subjective degree of certainty, rather than a limiting frequency of real events. Another characteristic feature is that inference is seen as a process of updating subjective knowledge that is first expressed by prior (before data) and then by posterior (after data) distributions.
4.3 Towards observables

Traditionally, parameters (intensities, chances, theoretical means & variances), provide the principal means to summarize the data. Sometimes in applications, what we claim about the observables is eventually more important than parameters (if not always!). All papers I-IV exploit hierarchical modelling and three of them deal with Bayesian inference. Hierarchical models are constructed by means of a chain of several conditional distributions. In a single conditional distribution, the distinction between 'parameters' and 'random variables' is traditionally straightforward in notation; the distribution of a random variable is given by the theoretical parameter(s).

In hierarchical models, a 'random variable' of one conditional distribution is taken as a 'parameter' of another. For example, in papers I-II, the number of invasive meningococcal disease cases is treated as a random variable whose conditional distribution is determined by the chance of disease upon infection, and the number of new infections. The latter 'parameter' is a random variable of still another conditional distribution, describing the chance of new infections. Yet, only the observable quantities can become known as data and thus the distinction between observables and unobservables seems, at least, pragmatically justified. When assigning a prior to a parameter in a hierarchical model it can also be very enlightening to check what this means in terms of the observables. For example, the prior of avoidance probability, $q$, in an epidemic model implies (together with initial conditions) a prior predictive distribution on the weekly number of infections, which can be more meaningful in the application than the prior of $q$ alone.

Presenting a posterior distribution of a parameter is not strictly a claim about 'the location of a true value'. Yet, we sometimes willingly idealize the parameter to represent a 'true signal'. This should be borne in mind when interpreting disease maps, for instance. A smoothed incidence value on a map is not a 'true' value. Only by using self generated artificial data could we compare the smoothed value with a known true value. Instead of plotting point estimates of intensities, multiple imputation of maps [25] or marginal posterior distributions and credible intervals of intensities can be used to represent the uncertainty within a given model. Furthermore, spatial predictive distributions of the number of disease cases can be employed for a direct assessment of the observables, as in paper III. Likewise, the concrete number of new infections in a given time interval can be more relevant than the basic reproduction number in epidemic models. Predictions of the former could (in principle) be evaluated and falsified with some accuracy, unlike those of purely unobservable quantities.

4.4 Universality as consensus

Estimates of parameters may not readily lend themselves to other models with different assumptions, or to another population living in a different environment. Firstly, the confounding factors in different populations may differ so that the 'universality'
of our estimates derived from one population is limited. Secondly, when the posterior is sensitive to prior specifications, different investigators will arrive at different conclusions. Yet, a group of investigators, each with a different prior, can arrive at a consensus which is as close to objectivity (intersubjectivity, [8], p.236-7) as can be reached unless, eventually, the quantity under inspection can be revealed to them all for direct observation. Related discussion in the context of risk analysis can be found e.g. in reference [26]. Thirdly, some parameters appear to be more model dependent than others. For example, estimates of infection probabilities and basic reproductive numbers depend on the assumed contact structure of the model and the type of the model (e.g. SIS, SIR, SIRS). In contrast, the expected duration of infectious period in a model can be more directly associated with the observable mean of durations in a defined population. The final judgement of transferability is outside the scope of the statistical paradigm. Hence, the utility of distributions of parameters is in their potential use as an intermediate device in inference, classification and prediction within a given model.

5 Back to the future

Indeed, Sire, Monsieur Lagrange has, with his usual sagacity, put his finger on the precise difficulty with the hypothesis: it explains everything, but predicts nothing. – Laplace in conversation with Napoleon.

5.1 Posterior predictive distribution

It can be argued that predictions are the ultimate purpose of all modelling. Predictive performance is also the most pragmatic challenge for any particular model. Qualitatively, though, it is of some interest how the predictions are produced. A black box that predicts well, but does not provide insight is not always a satisfactory solution. If a hierarchical model is constructed using variables which have a clear interpretation, we can hope to attain a suitable level of transparency.

In papers I, II and III, predictions were derived from posterior predictive distributions. The underlying principle [10], [27] can be introduced as follows. Adopting a generic notation, \( \pi \), for probability distributions, the conditional distribution of an observable \( Y_i \) is written as \( \pi (Y_i \mid \theta) \). We summarize our beliefs about the unknown parameter with a prior \( \pi (\theta) \) and, utilizing the observations \( Y_{1,...,n} \), wish to predict the forthcoming observation \( Y_{n+1} \). The posterior distribution of \( \theta \) is

\[
\pi (\theta \mid Y_{1,...,n}) = \frac{\pi (Y_{1,...,n} \mid \theta) \pi (\theta)}{\int_\theta \pi (Y_{1,...,n} \mid \theta) \pi (\theta) \ d\theta}.
\]

Predictions of the forthcoming observation can now be generated by repeating a simple two step algorithm: (i) sample a value \( \theta^s \) from the above posterior distribution, then (ii) sample the predicted value \( Y_{n+1} \) from the distribution \( \pi (\cdot \mid \theta^s) \). If samples
from these distributions are not readily available, then both steps involve another algorithmic sampling procedure. By repeating these steps we generate a sufficiently large sample of predictions of \( Y_{n+1} \) which can be summarized by its empirical distribution, e.g. histogram, sample mean, etc. In fact, the sampling algorithm provides a way to approximate the posterior predictive distribution, which is the integral

\[
\pi(Y_{n+1} | Y_1, \ldots, Y_n) = \int \pi(Y_{n+1} | \theta) \pi(\theta | Y_1, \ldots, Y_n) \, d\theta.
\]

The posterior predictive approach has the novelty that we can assign direct probabilities to the predicted quantities and new data can be added sequentially to improve the predictions. Uncertainty in all the underlying parameters is accounted for by numerical integration (sampling) over the parameter space. In many applications, some of the observations are missing. These can be treated in a similar manner with other unknown parameters of the model.

In paper I, the weekly disease cases due to meningococcal bacteria in a military garrison were completely observed, but nothing was observed about the latent asymptomatic carriers of the bacteria. Obviously, predictions of an epidemic can then be fairly crude. In paper II, the missing data consist of unobserved carrier prevalences in the majority of the units in a military garrison. In a typical investigation, units are selected for inspection because one or more invasive cases of meningitis, or some other form of meningococcal disease, had occurred there. The probability of such inspection is then the probability of a nonzero number of disease cases, which depends on the latent carriage process, i.e. missing data. The observed data consist of both the disease cases in all units, and the number of carriers in selected units at different time points.

In paper III, spatial predictions of childhood diabetes were suggested for the purposes of public health surveillance. In studies of environmental aspects of health, a geographical small area may be selected for further inspection because the observed incidence seems unusually high. In that case, one is assessing a 'cluster alarm' that was self-generated by the data. There are two general approaches for avoiding post hoc evaluation of such alarms [28]. In predictive inference, one can monitor further cases and compare these with the predictive distribution. Another approach is to expand the study area. In this way, one can compute the probabilities for the disease occurrences over a larger region. Spatial predictive distributions were computed in paper III for the disease incidence in a small area, based on the incidence data of all the other areas.

### 5.2 Temporal predictions of infectious disease. Example: meningococcal bacteria

#### 5.2.1 Explanatory modelling

Observations of infectious diseases in human populations provide us with time series of incidences, prevalences and other health indicators. Data are gathered for the purposes of monitoring, controlling and for planning interventions. The problem of predicting is obviously connected to the problem of regression. In essence, predicting is explaining
a response variable with a mechanism utilizing some explanatory variables. In the simplest case of a time series we aspire to find a regression function \( f \), such that

\[
y_i = f(t_i) + \epsilon, \quad \epsilon \sim N(0, \sigma^2)
\]

where \( y_i \) is the observation at time \( t_i \), \( i = 1, 2, 3, \ldots \). Different functions \( f \) could be chosen to minimize the discrepancy between the observed and expected values of \( y \). The function could also depend on the previous values \( y_{i-1}, \ldots, y_1 \) and other covariates. Unfortunately, minimal error and perfect fit do not guarantee the correctness of the model in the light of new data. In paper I, the numbers of invasive meningococcal disease cases in a garrison during the first four weeks in service were

\[
y_1 = 0, \quad y_2 = 3, \quad y_3 = 7, \quad y_4 = 14.
\]

An exponential regression function would fit adequately, but predictions from such a model would grossly overestimate the number of invasive diseases on the fifth and the sixth weeks, which were: \( y_5 = 6 \) and \( y_6 = 0 \). A refined regression function could then be employed in order to achieve better predictive performance. Following the classical statistical approach, we could operate on the observed layer of the epidemic by assigning models for correlation and expected values instead of describing explicitly the underlying biological process. Departing from such tradition, the emphasis in the present approach to predicting outbreaks (papers I-II) is not only on 'modelling the observed data', but also on an attempt of 'modelling the phenomenon'. The model structures rely on background knowledge of the phenomenon under study. This kind of information can be a valuable aid and should be exploited whenever available.

### 5.2.2 An extension to the basic chain-binomial model

The aetiology of meningococcal infections is sufficiently well understood to facilitate interesting model developments. It is known that the infection is asymptomatic and can be transmitted in close contacts. This suggests a classical Reed-Frost [29],[22] type of model formulation in which a susceptible individual has to avoid infection from all infectives in order to remain susceptible. It is also known that an infection has a duration that can be several months long, and that the rare condition of an invasive disease (e.g. meningitis) is likely to occur very soon after the acquisition of the bacteria, if at all. Because the infectious period is long, it is not realistic to assume a basic chain binomial model in which the infections can last only over a constant time step. Concerning outbreaks within military garrisons, we also know that the finite population is stratified into subpopulations which are here called units. It is reasonable to assume more frequent and closer contacts between individuals belonging to a same unit than between those living in different units. In paper II, the probability of infection for each susceptible in a unit \( j \) during week \( i \) was modelled as

\[
1 - q_i^{L_j} q_i^{L_i \cdot I_{ij}}; \quad q_w \leq q_b,
\]

where \( q_w \) is the chance of avoiding infection from an infective within the same unit, and \( q_b \) is the chance of avoiding infection from an infective in an external unit. \( I_{ij} \) denotes
the number of infectives in unit \( j \) in the beginning of week \( i \), and a \( +' \) sign in the subscript denotes a summation over the index number. A picture of this probability as a function of the numbers of infectives \( I_{ji} \) and \( I_{-i} - I_{ji} \) is shown in Figure 1, drawn with parameter values \( q_w = 0.97 \) and \( q_b = 0.9999 \). The number of units is assumed to be 20, each of size 68. These same parameter values were used for simulating the artificial data set in paper II.

![Graph showing infection probability](image)

Figure 1: Model of infection probability for a susceptible in unit \( j \) with different numbers of infectives within the unit, \( I_{ji} \), and in the rest of the garrison, \( I_{-i} - I_{ji} \).

Different models for the infection probability could be considered as well. Another classical model is known as the Greenwood model [29] where the probability of infection is zero or a positive constant depending on whether there are infectives in the population. With a similar population structure this would be written as

\[
1 - \left( q_w 1_{\{I_{ji}>0\}} + 1_{\{I_{ji}=0\}} \right) \left( q_b 1_{\{I_{-i}-I_{ji}>0\}} + 1_{\{I_{-i}-I_{ji}=0\}} \right) ; \quad q_w \leq q_b ,
\]

In papers I-II the models were constructed following the Reed-Frost type of hypothesis (1) because it is less reasonable to assume that the infection probability does not depend on the actual number of infectives when the units consist of dozens of individuals. If the individuals consisting of the group of susceptibles (in a given unit \( j \) at the beginning of a given week \( i \)) are assumed exchangeable, then we obtain a binomial distribution for the number of new infections for each week, with the conditional probability in the expression (1). This is a conditional probability given the number of infectives in each unit and given the chances \( q_w \) and \( q_b \).

Following a similar reasoning, the probability of terminating an infection (carriage), and the probability of developing an invasive disease are modelled. The development of the whole epidemic can be described as a chain of binomial probabilities. However,
there are important differences between the standard chain binomial model [29], [22]
and the model used in papers I-II. Firstly, because of the long duration of infection,
the process is not modelled here as a sequence of nonoverlapping generations of infectives.
Secondly, there are two important biological conditions, both of which can be
observed: the invasive disease cases, and the asymptomatic infections (carrier). In
paper I, observations of the disease cases were available, although these were not unit
specific. In paper II, a simulated data set was constructed to demonstrate the model
in a situation where unit specific observations of numbers of disease cases are available,
and when the carriage prevalence in selected units is determined.

5.2.3 The usefulness of meningococcal carriage data

Standard statistical methods are commonly applied to study the association between
disease cases and carriage prevalence in the epidemiological literature on meningococcal
outbreaks. The results have often been controversial [30], [31]. Occasionally, a high
prevalence is found among those in close contact with the disease cases, but exam-
pies of the opposite situation are also common. Recently, no clustering of the outbreak
strain was observed in the school classes of meningococcal disease cases in a Norwegian
study [32]. In Spain, the carriage prevalence in a population with high disease rate was
observed to be the same as that in a population of low disease rate [31]. Some authors
have concluded already in the 1970’s that there is no useful association between disease
incidence and the carriage prevalence as such [33] – it is rather the acquisition rate of
the bacteria that is important. Perhaps, this has led to the practice that observations
on carriers are no longer routinely collected. When the outbreak is seen as a dynam-
ic process evolving in time, we can recognize that temporal observations on carriage
prevalence do contain information on the acquisition rate and thus on the likelihood of
disease occurrences. The connection is not a straightforward correlation, because the
acquisition rate is the ‘slope’ of the carriage prevalence which is a nonlinear stochastic
function of time.

A thorough assessment of meningococcal disease incidence needs to account for the
chance of disease upon infection, the ‘initial’ state of the epidemic system, obser-
ations on the numbers of both carriers and susceptibles, and the contact structures
between them, among other important quantities. All these are specific to a given
population and its environment. Samples of carriers, when collected, are often class-
ified into several phenotypes of the bacteria. In papers I & II, multiple phenotypes
were not considered, but they could provide valuable information on the spread of the
epidemic and a challenge for further multivariate modelling. Apparently, only the
virulent strains of bacteria can cause disease cases, but carriage of other strains may
provoke some degree of immunity against disease [32]. Under any classification of the
bacteria, more than a single cross sectional set of observations in time needs to be
collected from each subpopulation under study in order to facilitate temporal analyses
of the acquisition rate.

Examples of temporal models of disease incidence are often connected with large scale
national surveillance data. For example, a time series analysis has been applied in forecasting meningococcal diseases in a large population, based on national records of diagnosed cases in Canada [34]. Likewise, a computational outbreak detection model has been suggested, utilizing a national notification scheme of salmonella infections in Australia [35]. These models are less suitable for describing a single outbreak in a semiclosed population. Thus far, (epidemic) models of longitudinal data have had little implications on the practical control of outbreaks in communities, such as military garrisons, schools, campus areas, refugee camps, etc. With the aid of detailed longitudinal data on both carriage and disease, it is hoped that some progress can be made in the prognoses of meningococcal disease risk in well defined communities, and in understanding the dynamics of an outbreak.

5.3 Spatial predictions of noninfectious chronic disease. Example: childhood type I diabetes

5.3.1 Empirical modelling

Much of the methodology in epidemiological research on detecting high risk geographical areas of disease occurrence has been focused on cluster testing. A range of statistical tests for different purposes are available. A common underlying assumption in many tests is complete spatial randomness (CSR) which is thought as the equivalent to 'absence of any systematic patterns'. Statistical clusters exist only with respect to a given model. This can be obscured if the results are to be interpreted as a bold declaration about the biological qualities of the population studied. Simple models of CSR can be refined to describe more complicated patterns and often CSR is known to be an oversimplified assumption in the first place. Moreover, in the theory of point processes, it is a well known result that the same point pattern can be generated from a model of first order effects as well as from a model of second order effects [36], [37]. If the aetiology of a disease is unknown, and we detect a 'cluster' by using a statistical test based on geographical coordinates, we can not conclude whether it was due to an infectious agent, due to a locally elevated environmental risk factor, or due to a genetically susceptible group of individuals. For such conclusions we would need to include data on these factors in our model. Clearly, a cluster test cannot suggest the underlying cause for departures from a simple CSR hypothesis. Of course, similar reasoning concerns space-time clustering.

In epidemiological studies, cluster testing can be seen as a part of explorative spatial data analysis. The diseases under inspection are often multicausal and their aetiology is quite unknown. Consequently, it is difficult to construct elaborated models based on biological and epidemiological knowledge external to data. Models can then serve conventionally as descriptions of the data rather than postulated descriptions of the phenomenon. In papers (I-II) on meningococcal epidemics, a rich (postulated) description was possible, for instance. In contrast, childhood diabetes (insulin-dependent diabetes mellitus) is a multicausal disease which has a genetic background. Several factors related to the environment are suggested to play a role, e.g. virus infections [38], [39], or nitrates in drinking water [40], [41], but the aetiology is still unclear. In
paper III, CSR was not assumed, but instead an intensity model was applied as a
description and for making inference on area specific incidence rates of type I diabetes.
In addition to maps of estimated intensities, model based predictions can then be used
to assess local disease incidence in a geographical region.

5.3.2 Georeferenced register data and loss of information
In general, spatial models are not written in terms of intensities or probabilities concern-
ing the individuals. Nevertheless, a large scale model can be derived from a fine
scale model. As an illustration, starting from the description of an individual belonging
to the population at risk of a chronic disease, a simple intensity model would be

\[ \eta_i(t) = I_{[t < \tau_i]}(t) \left( \phi_1 S_{i1}(t) + \ldots + \phi_J S_{iJ}(t) \right), \]

where \( I_{[t < \tau_i]}(t) \) is the indicator function of the disease occurrence, i.e. it is 1 if the
\( \tau_i \) individual has not yet developed the disease, and 0 otherwise. \( \tau_i \) denotes the time
of onset (e.g. diagnosis) of the disease. Similarly, \( S_{ij}(t) \) is the indicator function of
spatial exposure

\[ S_{ij}(t) = \begin{cases} 1, & \text{individual } i \text{ is within geographical area } j \text{ at time } t \\ 0, & \text{otherwise}. \end{cases} \]

Whenever an individual is located within the \( j \)th geographical area, he is assumed to
be under a risk, described by the intensity \( \phi_j \). This intensity model is naïve in the
sense that it has no memory of past exposures. For instance, the model could be better
used to describe the risk of being robbed in different parts of a city; once you walk out
from a dangerous area, you are safe from that place. Avoiding chronic diseases is not
quite like avoiding robbers. Exposures can cumulate over time, and a more realistic
model could be written as

\[ \eta_i(t) = I_{[t < \tau_i]}(t) f(H_i(t)), \]

where \( f \) is a function of the history \( H_i(t) \) of all spatial exposures of the \( i \)th individu-
ual up to time \( t \). A genuine environmental exposure model would require extensive
information concerning the life history of both cases and controls. In spatial analyses
of disease risk such data are seldom available, but the problem is acknowledged. In
medical geography it has been coined as 'time geography' [42]. We consider below only
the simple model without memory and show how this can be related to the common
Poisson model of spatial disease incidence.

Assuming model (3), and that there are \( N \) individuals over the whole region, the
conditional probability of \( Y(t) \) disease cases in time \( t \) takes the binomial form

\[ P\left( Y(t) \mid N, \phi_{j}, j = 1, \ldots, J, H_i(t), i = 1, \ldots, N \right) \]

\[ = \binom{N}{Y(t)} P(\tau_i < t)^{Y(t)} P(\tau_i \geq t)^{N-Y(t)}, \]
where

\[ P(\tau_i \geq t) = \exp\left(-\int_0^t \eta_i(s) \, ds\right) \]

is the survival probability. This formulation does not provide us probabilities on where the disease cases occur since we have not determined a model on spatial transitions of the individuals. Assuming then that \( N_j \) individuals stay within the \( j \)th area over the whole study interval and are exposed to an environmental risk, the binomial model can be written for the disease counts \( Y_j(t), \ldots, Y_j(t) \) in each geographical area, and the survival probability of an individual within the \( j \)th area simplifies to

\[ P(\tau_i \geq t \mid i \in A_j) = \exp(-\phi_j t). \]

Hence, the conventional binomial model of disease incidence, with disease probability \( p_j = 1 - \exp(-\phi_j t) \) can be obtained. The maximum likelihood estimate of the binomial probability is \( \hat{p}_j = Y_j(t)/N_j \), and the corresponding estimate for the intensity is \( \hat{\phi}_j = (\ln(N_j) - \ln(N_j - Y_j(t)))/t \). The binomial model is often approximated further by the Poisson model

\[ Y_j(t) \sim \text{Bin}(p_j, N_j) \approx \text{Poisson}(p_j N_j), \]

and the approximation is accurate if the disease is rare. Usually, the Poisson model is adopted as a starting point in spatial models of disease incidence. In paper III, the spatial model was written as

\[ Y_j(t) \sim \text{Poisson}\left( \int_0^t \int_{A_j} \lambda(s, a) n(s, a) \, ds \, da \right), \]

where \( n \) is the population density function, and \( \lambda \) is the intensity function of the (approximating) Poisson process of events of disease diagnosis. The interpretation of the population density is that it aims to describe the 'instantaneous' population mass at a spatiotemporal location \((s, a)\), resulting from the spatial movements of all individuals. In practice, it is convenient to discretize the model in space and in time, because both population counts and disease counts can be observed at some time steps and in some geographical areas. Essentially, we attempt to approximate the person time at risk, i.e. the integral of \( n \), and it is not of primary interest to find a particular model for \( n \) as a continuous function. A population model could be refined to account for births, deaths and migration if these were considered substantial, provided that we have some data on these. In paper III, the population was simply described statistically as a time dependent random variable, year by year, for each geographical area. In fact, the observed population counts were available for every second year, in each cell of the spatial lattice used.

To summarize, in the absence of data on the actual environmental exposure histories of each individual, we can compute estimates of the spatial intensities \( \lambda_j \) using some approximation of the person time at risk, and the observed disease counts. It was illustrated that such a model of 'area intensities' is a very coarse summary of a model describing the disease risk due to environmental exposure of individuals. Typically adopted approximations were distinguished:
1. Spatial exposure is a surrogate of the unmeasured environmental exposures.
2. Spatial exposure history (memory) is not accounted for.
3. Person times at risk within areas are approximated from residential records.

5.3.3 The scope of spatial modelling

Spatial models based on coarse health related data, as described above, are best suited for surveillance, dissemination and evaluation of geographical disease incidence. It was emphasized that environmental exposure is, at best, a surrogate for the physical dose exposure of harmful substances. Careful estimation of dose effects requires individual based data. These are often not available, or it can be too expensive and laborious to obtain such exposure history of both cases and controls. Instead, register based data on various statistics concerning geographical areas are readily available and these are used in many fields of research. Geographical information systems (GIS) with large data sets further expand the scope for modelling. Disease mapping is a commonly used device for dissemination and analysis of health related geographical data [43], [44], [45], [46], [47]. Standardized mortality or morbidity rates (SMR) - or standardized incidence ratios (SIR) - can be computed as a preliminary analysis. In areas of low population density, a map of SMR’s

\[
\frac{Y_j(t)}{E_j(t)} = \frac{\text{#Cases in the } j\text{th area during time } t}{\text{Expected #cases in the } j\text{th area during time } t}
\]

is notoriously dominated by the sampling variability. The same variability affects the maximum likelihood estimates of a Poisson intensity \( \lambda_j = Y_j(t)/E_j(t) \), where \( E_j(t) \) is the known or approximated count of the population at risk during the study interval of length \( t \). Comparison between area specific observed rates is difficult when some of the areas contain only low population counts. Traditional statistical techniques are not valid unless we are confident that there are no spatial confounding effects. Therefore, smoothing methods are needed to account for the unobserved spatial confounders. Beyond smoothing, the models can also be used for spatial predictions, as in paper IV. Posterior predictive distributions provide the means for assessing whether the observed disease count in one area is exceptional, given the rest of the data.

As a statistical convention, confounding is modelled either as heterogeneous or as locally dependent random effects, or both. Effectively, area specific disease intensities \( \lambda_j \) are smoothed towards a global or a local mean, or both. In Bayesian models, pairwise difference priors are employed to define the smoothing structure e.g. as a Markov random field (MRF). A pairwise difference prior is not a proper probability distribution, although the posterior distribution is. Hence, we cannot generate a prior predictive distribution from a MRF model in order to make comparisons with the posterior predictive distribution. A different approach would be to specify proper spatial prior distributions, or perhaps to use partially ordered Markov models [48] instead of MRF models.
5.3.4 Disease mapping and GIS

Modern tools of Geographical Information Systems (GIS) and georeferenced databases of statistical registers enable us to present data as a high resolution map of grid cells or even 'pixels'. As a result, it is easy to draw noisy images of 'raw data'. These fine grained images are increasingly used as visual evidence and for drawing conclusions. Apart from data manipulation and visualization, extracting reliable information can be difficult without suitable statistical techniques [49]. It has been noted, e.g. [50], that 'hypotheses generated using GIS often are not expressed as falsifiable predictions'. It can be argued, though, that there is no need for statistical methods because all data are observed and available within GIS, neither is there any need to make inferences beyond these data. A similar view has been expressed also in the context of data mining [51]. Whenever the goal is to predict or to make inference, it is obvious that a single realization of a spatial process does not imply knowledge of the underlying phenomenon without some inherent uncertainty, which calls for modelling.

Aggregations of georeferenced data, even without missing values, are subject to sampling variability, especially when ratios of variables are computed. A map of observed incidences of a disease can be completely useless for visual evaluation of risk patterns in geographically small areas. With suitable modelling techniques, better evaluation of data, and the underlying phenomenon, can be achieved. Some techniques can be used for better descriptive analysis, as in paper III. The map of observed incidences of childhood diabetes during 1987-1996 in Finland is shown below (a), accompanied by maps of (b) smoothed incidence values (posterior means), (c) posterior probabilities of exceeding the overall incidence, and (d) the posterior coefficient of variation (posterior standard deviation divided by posterior mean).

Historically, smoothing methods originate from the developments in image analysis [52] where the underlying true image is often known, to some extent, and the analysis only aims to remove random noise from the imprecise picture elements. Smoothing means accounting for distorting random noise, rather than modelling the intrinsic phenomenon that gives rise to the data. The latter would require explanatory type of modelling. With GIS technology, one can freely define arbitrary geographical areas for modelling by choosing a suitable set of pixels. It is difficult to go beyond descriptive analysis if there is no substantive information available to suggest elaborated explanatory model structures. For comparison, in an epidemic model of a garrison, stratification into units is natural since these represent biologically relevant contact structures. A neighbourhood structure of pixels in a spatial prior distribution of the underlying area effects is seldom so clearly defined, because it is only intended to account for unobserved confounding. If these confounders were observed, we could define an advanced model with covariates. As surrogates of covariates, pixel wise defined areas may provide a meaningful partition of space in some applications [53]. Yet, important unobserved spatial confounders are often known to be present, which should be properly addressed in the analysis.

There are intriguing similarities between structured population models and spatial
models, in which further development, e.g. in hierarchical multivariate modelling, is likely to take place. Spatial statistics and GIS technology have slowly started to merge, but the development holds promises which are yet to be fulfilled [54].

Figure 3: Left (c): Posterior probability of exceeding overall observed incidence (38/100,000/year) of childhood diabetes, 1987-1996. Scale: \(<0.1, 0.1 - 0.25, 0.25 - 0.5, 0.5 - 0.75, 0.75 - 0.9, > 0.9\) from blue to red. Right (d): Posterior coefficient of variation of childhood diabetes, 1987-1996. (Posterior SD / posterior mean intensity). Scale: \(<0.25, 0.25 - 0.26, 0.26 - 0.27, 0.27 - 0.3, 0.3 - 0.36, > 0.36\) from blue to red.
5.4 Surveillance of infectious disease. Example: *poliomyelitis*

5.4.1 Models of large scale dynamics

Classical population models of epidemics have been actively studied in order to describe and quantify the assumed temporal dynamics of infectious diseases and the effects of vaccination campaigns [55], [56]. These descriptions stem from the tradition of differential equation models of dynamical systems, although some of the models can be stochastic. The stochastic approach is often called for to describe 'stochastic fade out' of an epidemic, i.e. extinction of the infectious pathogen by chance when only a small number of people are infected in a population. In small groups, natural variation by chance can play an important role, but also large scale phenomena often exhibit stochastic behaviour that can not be satisfactorily described by a deterministic model. In a more physical and deterministic interpretation, 'natural variation by chance' could rather be understood as our imperfect knowledge about all the details of the real biological process studied. This corresponds to the Bayesian interpretation of randomness as uncertainty. Simulation techniques can be used to incorporate uncertainty to the models, and very complicated models can be constructed from a set of simple stochastic transition rules.

Deterministic or not, a large population epidemic model is usually less concerned with simultaneous estimation of all model parameters, but rather with theoretical descriptions of "would be"-worlds based on given assumptions. Possibly, parameter estimates could be drawn from altogether separate sources, but these (point) estimates can be highly unreliable and model dependent, and thus, difficult to apply to a particular population in a unique situation. In such cases, parameter estimates could be replaced by distributions of plausible parameter values. Multiple data sources can be exploited probabilistically in a Bayesian framework (e.g. in papers I & II) to construct prior predictive distributions which are then updated by data drawn from the particular population that is being addressed. With fewer data, predictions become increasingly dependent on priors. Models of belief networks [13], [57], are closely related to such quantification of possible realizations. For an example of a deterministic model (for tuberculosis) with a slightly different uncertainty analysis, see e.g. [58]. Eradication of poliovirus is hoped to occur in the near future as a result of large vaccination campaigns worldwide [59]. On the other hand, polioviruses can persist unnoticed over long times due to silent subclinical infections [60]. A stochastic simulation model was applied in paper IV to describe the chances of detecting polioviruses under different scenarios.

5.4.2 Scenarios of disease incidence

Once a stochastic model is constructed from a set of postulated premises it is relatively straightforward to simulate even large systems of conditional distributions by computer, and finally record the results for analysis. The ensemble of simulated results can be interpreted as prior predictive distributions under the given premises (priors). The outcome generating mechanism is then throughout stochastic:

\[
\text{prior} \Rightarrow \text{parameters} \Rightarrow \ldots \Rightarrow \text{outcome, stoc} \quad \text{stoc} \quad \text{stoc}
\]
where 'outcome' could be the prevalence function \( I(t) \) of infected individuals at time \( t \), for instance. In a deterministic model, uncertainty about the model parameters implies uncertainty about the model predictions. If these fixed parameters are replaced with a distribution describing the uncertainty, we can compute distributions of predictions also from a deterministic model. Although such predictions may not be generally titled as 'prior predictive distributions', the probabilistic interpretation is essentially the same as with the stochastic models:

\[
\text{prior } \Rightarrow \text{ parameters } \Rightarrow \text{ det} \ f(\text{parameters}) = \text{outcome},
\]

only now the intermediate conditional distributions have been replaced with a deterministic function. For large hierarchical models with several levels, it is quite possible to find different structures and parameter values which produce the same distribution of outcomes. This leads to problems in estimation and identifiability of unknown quantities, which can only be solved by collecting data related simultaneously to all or most levels of the hierarchy. It would be hard to justify 'inference from data' when the result is largely dependent on priors and little, if at all, on the data. While inference from the data may be limited, theoretical considerations still can play an enabling role for another purpose.

If, rather, we use models to generate distributions of outcomes, e.g. incidence or prevalence of polioinfecions in a large population, we can make conditional inference on the relative efficiency of different observational designs. If, under most simulated scenarios, one observational scheme outperforms the other, we may conclude that this is likely the case also in real situations. Similarly, possible effects of different vaccination strategies have been theoretically evaluated by conditioning on various epidemic scenarios, focusing on the reproduction number as a threshold parameter [55], [61]. Of course, this is different from actually conditioning on observed data, as with posterior distributions. The relation between epidemic models and data has been discussed at length, e.g. in [62], although the probabilistic (Bayesian) approach was largely omitted at the time.

### 5.4.3 Active versus passive monitoring in poliovirus surveillance

The underlying 'main scenario' in paper IV was that polioviruses have been almost, perhaps even completely, eliminated from a large population. The practical task would then be to continue the surveillance in order to detect the possible remaining few carriers of poliovirus, and to detect externally introduced new epidemics as early as possible. Two surveillance schemes were computationally evaluated under two epidemic scenarios and one endemic scenario. The epidemic scenarios aimed to describe the early stages of an outbreak with basic reproduction number \( R_0 \) close to one. Model based inference is often driven to estimate \( R_0 \) as a parameter of interest, under the given model scenario [63], [64]. In paper IV, we were not intrinsically interested in \( R_0 \) which was used only as a tuning parameter in order to simulate epidemics with desired magnitude and allowing for stochastic extinction. Large values of \( R_0 \) would generate some realizations with a very large number of infections. This scenario is not realistic in a finite population, nor in line with our underlying hypothesis of nearly eliminated
poliovirus.

In the first surveillance scheme, polioinfections are detected from the emerging cases of acute flaccid paralysis (AFP). In a sense, this is a 'passive' mode of investigation where polioviruses will eventually reveal their presence by emerging disease cases. AFP surveillance with an endemic scenario has been previously studied by Gary et al [35]. Using a stochastic model with a variable population size (births and deaths are simulated) Eichner and Dietz [60] have shown that the case free period should be at least three years before poliovirus can be declared eliminated from a large (closed) population. They also noted that this result depends on the case-to-infection ratio. If it is lower than the assumed ratio of 1/200, the case free period should be significantly longer before we can be assured that the virus is eliminated. In paper IV, AFP surveillance was simulated with similar and lower case-to-infection ratios, with endemic and epidemic scenarios.

In the second scheme, polioviruses are actively sought from sewage water samples according to a sampling protocol. The efficiency of the sampling scheme can be improved by increasing the number of samples and/or by improving the laboratory techniques used. With a sensitivity analysis it was shown that under reasonable assumptions it is likely that the environmental surveillance can detect polioviruses faster than AFP surveillance if the outbreak is not very large, and similarly, if the endemic prevalence is low. All this still depends on the case-to-infection ratio. In a vaccinated population this ratio would be very small, which supports the environmental surveillance scheme.

6 Concluding remarks

*Inference not followed by decision is largely idle, and no natural scientist worthy of the name would undertake the labor of conducting inference unless it served some purpose.*

E. T. Jaynes.

All of the four applications in papers I-IV deal with surveillance and control of diseases, but nothing was said about making actual decisions about interventions. In probabilistic inference, the posterior probability distribution summarizes our prior knowledge and the information drawn from data. Essentially, the same approach is suited for computing predictive distributions. We can fully quantify probabilistically how confident we are that a specific outcome will appear, or that a specific 'state of affairs' exists, conditionally on all we know and all we have observed. If the quantified variables are truly observables, we can always evaluate how right we were in the end of the day. Yet, the posterior distribution does not directly provide us with decisions. Decision theoretic considerations would vastly enlarge the scope of the present discussion. It might be necessary to explicitly define a loss function for each action-outcome pair in some form or another, see e.g. reference [12].

In decisions about control strategies of outbreaks, the predictive distribution would be of direct relevance. The current practice in decision making about interventions of
meningococcal outbreaks relies partly on rules of thumb. Several thresholds have been proposed, e.g. doubling of the weekly disease cases, or more than 1 case per 1000 population per week [66], [67]. A stochastic model which describes both latent carriage and observed invasive disease cases provides the means to quantify the partially observed epidemic together with dynamic predictions. Decisions could then be guided, e.g. by posterior probabilities of the unwanted outcomes. The likely effects of immunizations could be assessed within the same model, by treating a desired number of susceptibles as immune in the simulations. Clearly, there can be no simple dichotomous threshold that could capture all essential information both temporally and quantitatively. Another kind of decision making problem could be encountered in an empirical analysis of multicausal noninfectious diseases, e.g. childhood diabetes. Guided by a model based screening of georeferenced register data, further epidemiological case control studies could be planned and targeted. Finally, as a part of a follow up scheme, models can be used to theoretically evaluate alternative surveillance methods, e.g. in the case of poliovirus eradication. All these examples demonstrate models as a part of a complex (political) decision making process.

Little has been mentioned here about computing algorithms, although the programming has indeed been a major task. Most of the algorithms used are not yet easily accessible to a common lap-top user, but it is perhaps not too daring to predict that the development of a general purpose computational software will eventually find its way to the wider public. This has happened in the past with classical statistical methods. Links to available software can be found e.g. in reference [13].

Sensitivity analysis is a very important part of model aided decision making. The models should not be taken as given, but rather as a formalized framework that enables us to quantify our knowledge. Different prior distributions should be tested, especially when the data are sparse or otherwise limited. After all, the priors enable us to make explicit what we otherwise may assume implicitly a priori.

References


[38] Szopa TM, Titchener PA, Portwood ND, Taylor KW. Diabetes mellitus due to viruses - some recent developments. Diabetologia. 1993; 36: 687-95.


[60] Eichner M, Dietz K. Eradication of poliomyelitis: when can one be sure that poliovirus transmission has been terminated? Am J Epidemiol. 1996; 143: 816-22.


Summaries of the original papers

(I) Ranta J, Mäkelä PH, Takala A, Arjas E. Predicting the course of meningococcal disease outbreaks in closed subpopulations. Epidemiology and Infection 1999, 123: 359-71. A stochastic discrete time epidemic model was applied to meningococcal disease outbreaks in 'closed' populations such as military garrisons and schools. The weekly numbers of asymptomatic carriers were described as a hidden Markov model. Observations were available only about the weekly numbers of invasive disease cases. These were assumed to occur, with some disease probability, shortly after the acquisition of meningococcal carriage. Pre-knowledge about the disease probability and the duration of carriage was derived from other data sets in order to construct prior distributions. Predictions of the outbreak were summarized by the posterior predictive distribution, given the number of disease cases observed during the previous weeks. It was shown that the number of disease cases provides relatively weak information about the course of the epidemic.

(II) Ranta J, Mäkelä PH, Arjas E. Predicting meningococcal disease outbreaks in structured populations. Submitted for publication. A stochastic discrete time epidemic model was applied to predict the development of outbreaks of meningococcal disease in 'closed' populations which are further divided into subgroups called 'units'. Unit specific numbers of invasive diseases were assumed to be observed. In addition, the numbers of asymptomatic carriers were assumed to be observed from (a) units where invasive disease had been observed, and (b) both from some units where invasive disease cases had not occurred and from those where they had occurred. The performance of the model was assessed by using simulated data. It was shown that the observations of asymptomatic carriers together with the observations of invasive diseases can be a valuable combination for making better temporal predictions of outbreaks.

(III) Ranta J, Penttinen A. Probabilistic small area risk assessment using GIS-based data: a case study on Finnish childhood diabetes. Statistics in Medicine 2000, 19: 2345-59. A hierarchical spatial model was applied to describe the regional incidence of insulin dependent diabetes mellitus among the under 15-year-olds in Finland. The spatial data set originated from the combination of geographic information systems (GIS) and national data bases of both the disease cases and the population at risk. A simple Markov chain population model was constructed to describe missing population observations for every second year in each grid cell of 10 x 10 km². Assessment of disease incidence in geographically small areas was demonstrated by using a cross validation technique based on posterior predictive distributions. Maps of estimated disease intensities showed clear variation of disease risk across the country. Temporal changes were studied by comparing the maps describing the periods 1987-1991 and 1992-1996.
(IV) Ranta J, Hovi T, Arjas E. Poliovirus surveillance by examining sewage water specimens. Studies on detection probability using simulation models. Submitted for publication. Three transmission models of poliovirus infections in a large population were defined for simulating the number of infected individuals and the amount of excreted viruses. Environmental factors and the sampling techniques were accounted for in the simulation model. Finally, acute flaccid paralysis (AFP) surveillance and the sewage water sampling scheme were compared under the epidemic/endemic scenarios. It was shown that when the case-to-infection ratio is small (< 1/200), e.g. in vaccinated populations, it is likely that environmental surveillance can outperform AFP surveillance.