

---

## Seventy-five years of estimating the force of infection from current status data

---

N. HENS<sup>1,2\*</sup>, M. AERTS<sup>1</sup>, C. FAES<sup>1</sup>, Z. SHKEDY<sup>1</sup>, O. LEJEUNE<sup>2</sup>, P. VAN DAMME<sup>2</sup>  
AND P. BEUTELS<sup>2</sup>

<sup>1</sup> Interuniversity Institute of Biostatistics and Statistical Bioinformatics, Hasselt University, Diepenbeek, Belgium

<sup>2</sup> Centre for Health Economics Research & Modelling Infectious Diseases; Centre for the Evaluation of Vaccination, Vaccine and Infectious Disease Institute, University of Antwerp, Antwerp, Belgium

(Accepted 18 August 2009; first published online 21 September 2009)

### SUMMARY

The force of infection, describing the rate at which a susceptible person acquires an infection, is a key parameter in models estimating the infectious disease burden, and the effectiveness and cost-effectiveness of infectious disease prevention. Since Muench formulated the first catalytic model to estimate the force of infection from current status data in 1934, exactly 75 years ago, several authors addressed the estimation of this parameter by more advanced statistical methods, while applying these to seroprevalence and reported incidence/case notification data. In this paper we present an historical overview, discussing the relevance of Muench's work, and we explain the wide array of newer methods with illustrations on pre-vaccination serological survey data of two airborne infections: rubella and parvovirus B19. We also provide guidance on deciding which method(s) to apply to estimate the force of infection, given a particular set of data.

**Key words:** Hugo Muench, incidence data, infectious diseases, serological data, statistical modelling.

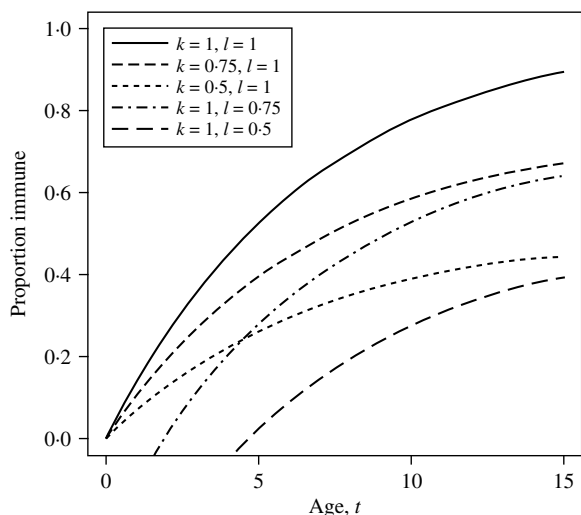
### INTRODUCTION

Although epidemics were already documented by Hippocrates (458–377 B.C.), it was not until 1760 that mathematical modelling of infectious diseases was first documented in a publication by Daniel Bernoulli [1], who used a mathematical method to study smallpox infections. These and subsequent models at the very start of the 20th century [2–5] focused on determining the infectious disease spread and the associated basic and effective reproduction number (see Appendix for terminology used). Hugo Muench first proposed the concept of estimating the force of infection (at that

time called the 'effective contact rate') as a key parameter in such mathematical models, in a publication entitled 'Derivation of rates from summation data by the catalytic curve' in the March 1934 issue of the *Journal of the American Statistical Association* [6]. His work became widely known only 25 years later with the publication of his book *Catalytic Models in Epidemiology* [7].

Before Muench's seminal work, series of physicians' case-reports, i.e. incidence data, which are often affected by underreporting and misdiagnosis, were used to study the 'effective contact rate' of a given disease in a given population [1, 8]. Muench suggested using a catalytic model on summation data to obtain a measure of the rate at which a susceptible acquires infection (and not necessarily disease). The name 'catalytic' was inspired by the similarity to the

\* Author for correspondence: Professor N. Hens, Interuniversity Institute of Biostatistics and Statistical Bioinformatics, Hasselt University, Agoralaan 1, Building D, B-3590 Diepenbeek, Belgium. (Email: niel.hens@uhasselt.be)



**Fig. 1.** Muench’s catalytic model for  $\lambda=0.15$  and various choices for  $k$  and  $l$  [see equation (1)].

equations used to study the processes that drive chemical reactions. In his 1959 monograph [7], Muench referred to this effective contact rate as the force of infection (FOI). The importance of Muench’s catalytic model is the capacity to use test results on, for example, serological or saliva samples, in addition to reported incidence data, in order to estimate the FOI, especially for infections that leave permanent markers of immunity in their surviving hosts.

More explicitly, Muench stated that the simplest approach is to assume a constant effective exposure rate  $\lambda$  per unit of time  $t$  and that this rate applies to the entire population at all times. Nonetheless, the immune proportion  $\pi$  increases only to the extent that previously uninfected individuals can incur new infections. In addition, the model allowed that some fraction of the population cannot be infected at all. Muench’s catalytic curve can be written as

$$\pi(t) = k(l - e^{-\lambda t}), \tag{1}$$

where  $k$  is the proportion of the population that can be infected,  $\pi$  is the proportion of all previously infected (and immune) individuals prior to age  $t$ , and  $l$  is the proportion of the population which may show evidence of exposure. Figure 1 illustrates the behaviour of Muench’s model for different choices for  $k$  and  $l$ . Note that the model gives negative estimates of the proportion immune for  $l < 1$  and thus interpretation is not straightforward in this case.

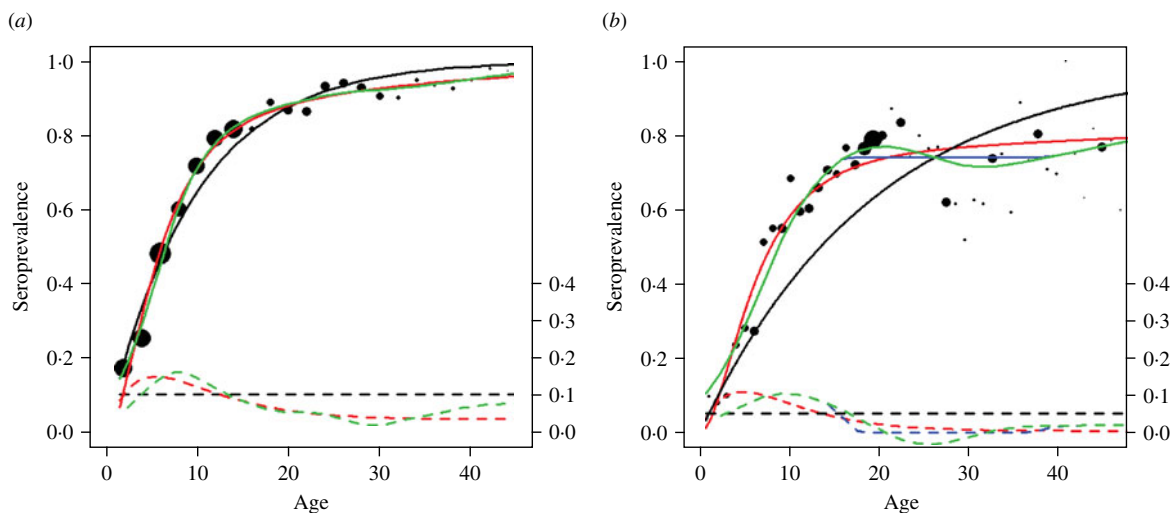
Using the method of moments to estimate these parameters, Muench illustrated his approach on several datasets. These included intraperitoneal mouse

protection test results for yellow fever in South America, a test which remains positive once an individual has been infected by yellow fever. He considered two regions; one in which the population was originally assumed entirely susceptible with frequent epidemics so that  $\lambda$  actually corresponded to a steady effective exposure rate. He compared these results with those of a second region where an outbreak of yellow fever occurred. He illustrated that these test results cannot show when individuals had been infected, only that it must have occurred at some time before taking the test. Indeed, by analysing summation data one estimates the probability of past infection at a certain time-point and derives the FOI using the analytical expression in equation (1). The latter is based on the untestable assumption of time homogeneity.

Testifying its importance, numerous later publications echoed this inherent limitation of not being able to use direct data on person-time incidence but rather summation data to derive the FOI under the assumption of time homogeneity. Muench already discussed other complications such as test reversion rates, differential mortality, and passive immunity through transfer of maternal antibodies.

Although applicable to cumulative incidence data too, we illustrate Muench’s method on cross-sectionally collected seroprevalence data because of the above limitations of reported incidence data. We use two serological surveys, the first on rubella in the UK, collected in 1986–1987 for males only [9]; and the second on parvovirus B19 in Belgium, collected in 2001–2003 [10]. We make a number of typical common assumptions, i.e. that no portion of the population is free from exposure, a perfect test is used, the time homogeneity assumption holds, disease-related mortality is negligible compared to all-cause mortality, infection confers lifelong immunity [in equation (1):  $l=k=1$ ] and, specifically for rubella in the UK, that the schoolgirl programme that was implemented had negligible effect on the seroprevalence of rubella in males. Whereas Muench originally used the method of moments to estimate the FOI, we cast his model in the maximum-likelihood framework. More specifically, we use a binomial likelihood to relate the age-specific prevalence (or cumulative incidence) to the age-specific observed proportion immune:

$$L(\lambda; \mathbf{y}, \mathbf{n}, \mathbf{a}) = \prod_{i=1}^N \pi(a_i)^{y_i} (1 - \pi(a_i))^{n_i - y_i}, \tag{2}$$



**Fig. 2.** (a) UK rubella and (b) Belgium parvovirus B19 infection by age (years): the observed seroprevalence per integer age value with size proportional to the sample taken (●), the fitted seroprevalence curve (—, upper curve) and the FOI curve (- - -, lower curve). Four different models were used: Muench’s constant FOI model (black curves); Farrington’s exponentially damped model (red curve), a spline model (green curve) and its monotonized version (green-blue-green curve) as applied by Hens *et al.* [10]. Note that, by definition, the latter curve (partly) overlaps with the green curve for rubella (parvovirus B19).

where  $a=(a_1, \dots, a_N)$  denotes the age vector of length  $N$ ;  $y=(y_1, \dots, y_N)$  and  $n=(n_1, \dots, n_N)$  denote the corresponding vectors of positive and total counts per age value. Assuming Muench’s model,  $\pi(a_i)$  in equation (2) is given by equation (1). Maximizing the likelihood in equation (2) with respect to  $\lambda$  yields the maximum-likelihood estimate  $\hat{\lambda}_{ML}$ .

The black curves in Figure 2 show the fitted seroprevalence and FOI curves together with the observed seroprevalence for both infections. The constant FOI was estimated as 0.104 and 0.053 for rubella and parvovirus B19, respectively. The assumption of a constant FOI seems visually appropriate for the rubella data but inappropriate for the parvovirus B19 data (cf. Fig. 2). Such observations led other researchers to develop new methods allowing for an age-dependent FOI.

**Standing on the shoulders of Muench**

Muench’s work has led to the development of many models to estimate the FOI. These models can broadly be divided into *parametric* and *non-parametric methods*. Whereas parametric methods rely on specific functional relationships (e.g. Muench’s model is parametric since it assumes a constant FOI), non-parametric models relax such assumptions. In this section we present an overview of these different approaches. Starting with Muench’s work, Table 1 summarizes the key contributions with respect to the

shape of the FOI, the focus of the analysis and the data used.

*Parametric methods*

Griffiths [11] was the first author to propose the use of a linear rather than a constant FOI to model measles incidence data from England and Wales for the period 1956–1969. With the above notation, his model could be formulated as

$$\pi(t) = 1 - e^{-\int_0^t \lambda(\omega) d\omega}, \tag{3}$$

where  $\lambda(t) = \gamma_0(t + \gamma_1)I_{\{t > \tau\}}$ , with  $\gamma_0, \gamma_1$  parameters to be estimated and  $\tau$  the threshold for inherited immunity.  $I_{\{t > \tau\}}$  denotes the indicator function taking value 1 if  $t > \tau$  and 0 otherwise. In general, equation (3) is the solution of the differential equation

$$dx(t)/dt = -\lambda(t)x(t) \tag{4}$$

(with initial condition  $x(0) = 1$  meaning that everyone is assumed susceptible at birth) that describes the changes in the proportion of susceptibles  $x(t) = 1 - \pi(t)$  with respect to time. Note that  $\int_0^t \lambda(\omega) d\omega$  denotes the cumulative FOI up to age  $t$  acting on susceptibles. Whereas Griffiths [11] described the estimation procedure outlined above as simple and straightforward to apply, he actually used an alternative method to model the measles incidence (rather than cumulative incidence) using a multinomial model and thereof estimated the attack rate.

Table 1. *Historical overview of key-contributions to the estimation of the force of infection (FOI) from summation data in the 20th century*

Reference	FOI	Focus	Data
Muench (1934) [6]	Constant	Estimating a constant FOI from summation data	Yellow fever protection test results; tuberculin test results; incidence data on whooping cough and chicken-pox
Wilson & Worcester (1941) [69]	Constant	Comparison of the model of Muench to the work of Collins on incidence data	Incidence data on measles
Griffiths (1974) [11]	Linear	Linear FOI	Incidence data on measles
Grenfell & Anderson (1985) [15]	Polynomial	Estimating an age-dependent FOI from incidence or serological data	Incidence data on measles and serological data on measles
Farrington (1990) [16]	Nonlinear: exponentially damped linear model	Estimating an age-dependent FOI	Serological data on measles, mumps and rubella
Keiding (1991) [30]	Non-parametric estimation using kernel methods	A statistical perspective on estimating incidence and prevalence from serological and case-notification data	Incidence data on hepatitis A, and incidence and prevalence of other non-infectious disease data

Note that the catalytic model as presented in equation (3) is actually a survival model and that the probability of past infection is the cumulative distribution function of the time to infection or alternatively  $1 -$  survival function. The (cumulative) FOI is the (cumulative) infection hazard.

In general the change in the susceptible portion could be both age- and time-dependent. It is under the assumption of time homogeneity that age can be used to determine the time of infection and thus age and time are identified and equation (3) is typically denoted in terms of age rather than time. Time homogeneity implies that neither the pathogen's transmissibility, nor susceptible people's receptiveness to infection (irrespective of their age), nor the frequency and intensity of interactions necessary for transmission to occur (e.g. social contact patterns for air- and saliva-borne infections, sexual intercourse for sexually transmitted infections) have changed substantially with time. In particular feco-orally transmitted infections (e.g. hepatitis A virus, cholera) are documented to be time heterogenous, due to the drastic improvements in sanitary conditions in many settings over the last century [12, 13]. Note that Schenzle *et al.* [13] argued that, for the case of hepatitis A, it is more appropriate to assume age homogeneity and time heterogeneity, and thus modelled time effects. The limitation of having to choose for either time or age may be overcome by use of data which is both age and time structured (see below).

Griffiths used maximum-likelihood theory to estimate the parameters in the catalytic linear model and for the first time addressed the goodness-of-fit to the data using Pearson's  $\chi^2$  test. Interestingly, Griffiths justified his choice of a linear FOI by using a non-parametric estimate for the FOI which was plotted against age and showed a linear trend. Note that Griffiths applied his model only up to age 10 years as by then most children had been infected with measles. Griffiths' work was later used by several other authors to model measles and other common childhood infections [14]. Griffiths' 1974 contribution should be seen as the completion of the basic building block for the estimation of the FOI.

Grenfell & Anderson [15] extended Griffiths' approach to encompass a general polynomial description of changes in the FOI with age and derived a stepwise maximum-likelihood method for parameter estimation from datasets consisting of either case notifications or serological information. This encompasses both Muench's and Griffiths' catalytic model. The appropriate polynomial degree can be found by minimizing the relative deviance. Grenfell & Anderson discussed the advantages and disadvantages of using case-notification data and serological data and stressed that availability is the key criterion for which type of data to use [15]. For case-notification data, the quality largely depends on the biases and inaccuracies of the case-notification system. Serological databases too can suffer from the

representativeness of the collected samples, as well as implicit assumptions for analysis, such as time homogeneity, lifelong immunity and the other complications listed above, duly noted by Muench [6]. Several of these issues gave rise to further research.

A first complication is that under the assumptions stated above, the model for the prevalence should be monotonically increasing with age. The monotonicity issue was not relevant for Muench and Griffiths since a model with constant or linear FOI always estimates a monotone prevalence. However, this is not necessarily the case with high-order polynomials. In 1990, Farrington [16] placed the issue of monotonicity in the heart of the estimation problem. He noted that for measles, mumps and rubella, the FOI typically rises linearly from birth up to about 10 years of age, after which it drops off again for older ages. This qualitative form was also observed by Griffiths. Farrington considered a nonlinear model, i.e. an exponentially damped linear model, in age

$$\lambda(a) = (\gamma_0 a - \gamma_1) e^{-\gamma_2 a} + \gamma_1, \quad (5)$$

where  $\gamma_0$ ,  $\gamma_1$  and  $\gamma_2$  are positive parameters that can be estimated using maximum-likelihood and constrained optimization. The FOI in equation (5) is 0 for age 0, then shows a linear increase and ends in an exponential decay. Following Griffiths, Farrington [16] assessed the goodness-of-fit of his model using a non-parametric estimate of the FOI, i.e. a moving average, indicating again that a more formal non-parametric approach for the estimation of the FOI is needed, at least in the exploratory stage.

Note that for specific choices of  $\lambda(a)$ , such as those proposed by Muench and by Grenfell & Anderson (see also [17]), equation (3) corresponds to a generalized linear model (GLM [18]) with binomial response and log-link. Other parametric models fitted within the framework of GLMs have been proposed since then using different link-functions such as the logit and cloglog link [19–29]. Of those, the most popular parametric model employed is the piecewise constant FOI where for predetermined intervals a constant FOI is assumed. The choice of these intervals is usually inspired by the intuitive relevance of the ages of mixing groups in the population (e.g. pre-school, school, high school, etc.). A drawback for the model of Grenfell & Anderson and many other parametric models is the probable occurrence of a negative FOI for some age values or an unrealistic steep increase for higher age values because of the chosen functional relationship. Farrington's model is not a member of

the GLM family, and deals with these issues by constraining the model based on prior knowledge. However, if that knowledge is unavailable or questionable, non-parametric options could be explored.

#### *Non-parametric methods*

Although non-parametric techniques were used before to assess the fit of a parametric, possibly nonlinear, function [11], Niels Keiding [30] was the first to explicitly use a non-parametric technique to estimate the FOI from serological data, based on the isotonic Kaplan–Meier estimator of  $1 - \pi(a)$  which finds its origin in survival analysis. Keiding also addressed the issues of time homogeneity, monotonicity, and censoring.

Most of the various non-parametric methods in the GLM framework developed since Keiding, involved estimating the (sero)prevalence by a non-parametric technique and subsequently deriving the FOI by  $\hat{\lambda}(a) = \Delta \hat{\pi} / (1 - \hat{\pi}(a))$ . Note that this latter expression for the FOI is a discretized version of equation (4) with  $x(a) = 1 - \pi(a)$ .

Of these non-parametric applications, spline-based methods have become very popular [10, 31–34]. A spline can be seen as a concatenation of a rich number of local polynomials which are glued together in a 'smooth'/continuous way. In 1996, Keiding and colleagues proposed to replace the kernel smoother in his earlier work with a smoothing spline [19]. Subsequent work involved semi-parametric models [35–37], in which the age-specific prevalence is modelled non-parametrically and possible covariate effects such as gender are included in the parametric component of the model. The green (blue) curves in Figure 2 represent the (monotonized) estimated prevalence and FOI based on the spline methodology [10, 19, 37].

Looking further at the different methods as applied to the rubella and parvovirus B19 data in Figure 2, several observations can be made. First, Muench's model seems to fit the serological profile of rubella well, whereas it does not follow the pattern of the parvovirus B19 seroprofile. Farrington's exponentially damped linear model shows an improved performance for rubella whereas the fit to the parvovirus B19 data seems reasonable except that it is not able to capture the decay in seroprevalence at about 25 years of age. The spline model shows a similar fit to the seroprevalence data as the exponentially damped linear model for the rubella data, but some quantitative differences appear on the scale of the FOI.

Moreover, the spline model is able to capture the decaying seroprevalence at around the age of 25 years whereas its monotone version is a regularization to ensure a positive FOI. Indeed, when looking at parvovirus B19 in Figure 2*a*, the spline fit reveals a non-monotone pattern (green curve) whereas its monotonized version, applying a smooth-then-constrain approach, shows a monotone trend (green-blue-green curve) for the prevalence and a positive FOI.

Based on the non-parametric fits, one might question which of the underlying assumptions is violated for parvovirus B19 in Belgium. This could for instance be due to antibody titres declining with time post-infection, and eventually falling below the cut-off for positivity; or a time dependence in the FOI resulting in a cohort effect, or the use of a manifestly imperfect test. It is only by contrasting different methods and the graphical exploration of the goodness-of-fit that such distortions appear [21].

When estimating the FOI from summation data, phenomena like these are often not taken into account because of the lack of information in the data and thus result in what is often referred to as an average profile. The methods that have been used to investigate the potential mechanisms behind such features mostly rely on contrasting mathematical models with incidence data. It seems hardly possible and beyond the scope of the current paper to list all papers that may have used Muench's catalytic model as a starting point (without necessarily referencing it) to make extensions for such mathematical modelling studies. Nonetheless we further illustrate the vast influence of Muench's pioneering work by listing some of the main examples of such extensions: (1) models assuming infection confer no or no lasting immunity. Examples of such analyses include *Bordetella pertussis* [38] and *Haemophilus influenzae* type b [39], transmission models in which waning immunity was taken into account; (2) analyses describing the possible occurrence of seasonality [40–42] or regular epidemics [43]; (3) analyses on chronic infections (see e.g. [44–48]); (4) analyses taking into account vaccination programmes (see e.g. [48–50]); (5) further extensions of the more basic catalytic models have been used (see e.g. [51]).

### A practical guide to estimate the FOI

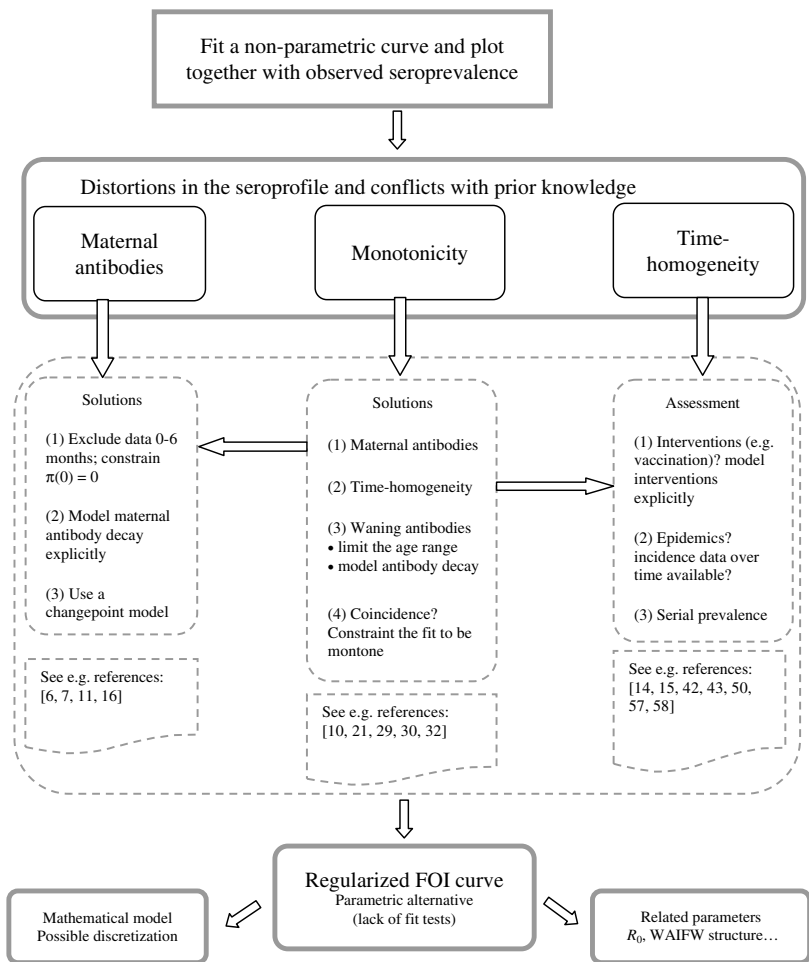
The diversity of methods to estimate the FOI raises the question, given a particular dataset, which method

should be used? Figure 3, presents a guiding flow chart, which starts from an exploratory stage in which a non-parametric model is fitted to the data. A graphical representation of such a fit is given in Figure 2. Given the epidemiology of the infectious disease under consideration and specific complications, such as diagnostic uncertainty, the shape of the non-parametric estimate and the goodness-of-fit to the data could show distortions with respect to monotonicity, maternal immunity and time homogeneity. For each of these distortions, different remedial techniques are available, which are listed in the flow chart (Fig. 3). These techniques enable obtaining a 'regularized' estimate of the FOI, which can be studied in detail, parametrized [52–56] and in turn can be used for the estimation of related parameters, such as the basic reproduction number [9, 16].

### Muench today: a road map to another 75 years?

Although, Muench's model was overly simplistic in assuming a constant FOI, his 1934 paper already raised concerns with respect to model assumptions and data constraints, which are still today important research topics. There are a number of areas where recent research has produced interesting results, and in this section we highlight four of these.

- (1) Time homogeneity is often assumed, as a necessary condition, if one can only use a single cross-sectional serological survey for estimating the FOI. In some cases (e.g. parvovirus B19 in Fig. 2), this seems questionable and might be untenable. If so, one needs longitudinal type of data (see e.g. [57]) or alternatively several prevalence surveys at different points in time [57, 58]. Since similar distortions were observed for parvovirus B19 in four other countries [59], other hypotheses such as waning of naturally acquired antibody levels with time post-infection need further investigation.
- (2) Antibody levels are commonly used to classify an individual's sample as positive, negative or inconclusive based on a given test-specific threshold. This allows estimation of the proportion of susceptible people at each age, and to derive from these proportions the FOI by age. Mixture models are a natural alternative for this type of data. Gay [60] was the first to model age-stratified serological data using two-component mixture models with age-dependent mixing probabilities and age-dependent mixture components. Using this mixture approach, Bollaerts *et al.* [61]



**Fig. 3.** Flow chart of a practical guide to estimate the force of infection (FOI) from seroprevalence data with reference to the literature on what to do and how to do it.

- proposed a direct estimator for an age-dependent FOI (i.e. without using a predefined threshold to distinguish susceptible from immune people).
- (3) As sera are often tested for more than one antigen, joint analysis for diseases with similar transmission routes can lead to new insights. Hens *et al.* [10] introduced the age-dependent joint and conditional FOI, a framework allowing formal statistical tests such as testing the assumption of separable mixing. Earlier work on bivariate models [9, 62] focused on the estimation of the basic reproduction number from serological survey data incorporating the effect of individual heterogeneity (see also [63]) and testing for separable mixing [9].
  - (4) A fundamental concept for the estimation of the basic reproduction number is the mass action principle, relating the FOI with the transmission rate, i.e. the *per capita* rate at which an individual of a particular age makes an effective contact with

a person of a specific age, per year. The transmission rates populate the so-called WAIFW matrix [64], which is traditionally imposed to be of a particular structure (rather *ad hoc*, albeit inspired by, e.g. social and schooling systems). Wallinga *et al.* [65] were the first to conceptually link seroprevalence data with data on conversational contacts per person, whilst assuming that transmission rates are proportional to rates of conversational contact. Using data from a social contact survey, Ogunjimi *et al.* [66] and Goeyvaerts *et al.* [67] proposed the disentanglement of the WAIFW matrix into two components: the surface of conversational contacts and an age-dependent proportionality factor.

**DISCUSSION**

Although the basic reproduction number is a very powerful and elegant summarizing parameter at the

heart of infectious disease transmission dynamics, the paucity of opportunities, and difficulties to obtain direct estimates of  $R_0$ , make the FOI extremely relevant. Indeed, it may often be the parameter, which through its estimation allows all other parameters to be estimated.

In disease burden estimates, it determines estimates of the occurrence of infections, of clinical cases and all consequences arising from these cases (e.g. hospitalizations, deaths, life-years lost, disability adjusted life years). Moreover, in studies estimating the impact of infectious disease interventions, such as vaccination, in terms of effectiveness and cost-effectiveness, the FOI will time and again rank high if not first of the most influential parameters determining the outcomes, and hence the public health policy measures based on these outcomes. Therefore

estimating the FOI as accurately as possible, is essential.

Muench’s seminal works published 75 and 50 years ago, are still inspirational today for the conceptual basis he provided for the FOI, and the pitfalls and problems he identified. We have described not only the various extensions those standing on the shoulders of Muench have proposed, but also noted that we are still trying to deal with the same pitfalls and problems already identified by Muench.

There appears to have been an increased interest in developing approaches for the estimation of the FOI since the mid-1980s [68]. We hope the guidance we have provided will be useful for researchers to decide, after exploration with a non-parametric method, which of the parametric methods is best suited for the dataset under consideration.

**APPENDIX. Glossary of words in alphabetical order**

Terminology	Description	Common synonyms
Basic reproduction number, $R_0$	The number of secondary cases one typical infectious individual produces during his/her entire infectious period in a <i>completely susceptible</i> population	<ul style="list-style-type: none"> <li>● Basic reproductive number, basic reproductive rate</li> </ul>
Censoring	When the event time is not exactly known, it is either left-, right- or interval-censored, according to whether the event has occurred, has not occurred or occurred in a specific time interval	
Effective reproduction number, $R$	The number of secondary cases one typical infectious individual infects during his/her entire infectious period in a <i>given</i> population	<ul style="list-style-type: none"> <li>● Effective reproductive number, effective reproductive rate</li> </ul>
Force of infection	<i>Per capita</i> rate at which a susceptible individual acquires infection	<ul style="list-style-type: none"> <li>● Effective contact rate (as Muench initially called it)</li> <li>● Person time incidence rate</li> <li>● Acquisition rate</li> <li>● Infection hazard</li> <li>● Attack rate in susceptibles</li> </ul>
Generalized linear model	A flexible generalization of ordinary regression models where the distribution of the outcome variable is linked to the <i>linear predictor</i> through a function called the link function	<ul style="list-style-type: none"> <li>● The binomial regression model is a specific type of generalized linear model</li> </ul>
Monotonicity assumption	The property of preserving an increasing order, i.e. being a non-decreasing function	<ul style="list-style-type: none"> <li>● Monotonically increasing</li> </ul>
Relative deviance	Goodness-of-fit measure defined as the deviance divided by the degrees of freedom, i.e. the difference of the number of observations and the number of parameters used in the model	
Time homogeneity	Not depending on time; time-invariant (not to be confused with age independency)	<ul style="list-style-type: none"> <li>● Steady state</li> <li>● Endemic equilibrium</li> </ul>
WAIFW matrix	‘Who acquires infection from whom matrix’: a matrix of transmission rates by categories of infectious and susceptible individuals (usually age-structured)	<ul style="list-style-type: none"> <li>● Transmission matrix</li> <li>● Beta-matrix</li> </ul>



## ACKNOWLEDGEMENTS

We thank the editor and both referees for their valuable suggestions that have led to an improved version of the manuscript. This work was supported by research project (MSM 0021620839), funded by 'SIMID', a strategic basic research project funded by the institute for the Promotion of Innovation by Science and Technology in Flanders (IWT) (project number 06008); by the Fund of Scientific Research (FWO, Research Grant G039304) in Flanders, Belgium; and by the IAP research network (no. P6/03) of the Belgian Government (Belgian Science Policy). The R-code used to analyse the datasets in this manuscript is available from the authors.

## DECLARATION OF INTEREST

None.

## REFERENCES

- Bernoulli D. An attempt at a new analysis of the mortality caused by smallpox and of the advantages of inoculation to prevent it [in French]. *Histoire et Mémoires de l'Académie des Sciences* 1760; **2**: 1–79.
- Hamer WH. Epidemic disease in England. *Lancet* 1906; **i**: 733–739.
- Brownlee J. Statistical studies in immunity: the theory of an epidemic. *Proceedings Royal Society of Edinburgh* 1906; **26**: 484–521.
- Ross R. *The Prevention of Malaria*. London: John Murray 1911.
- Kermack WO, McKendrick AG. Contributions to the mathematical theory of epidemics. *Proceedings of the Royal Society of London, Series A* 1927; **115**: 700–721.
- Muench H. Derivation of rates from summation data by the catalytic curve. *Journal of the American Statistical Association* 1934; **29**: 25–38.
- Muench H. *Catalytic Models in Epidemiology*. Harvard University Press: Boston, 1959.
- Collins SD. Age incidence of the common communicable diseases of children. *Public Health Reports* 1929; **44**: 14.
- Farrington CP, Kanaan MN, Gay NJ. Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data. *Journal of the Royal Statistical Society, Series C* 2001; **50**: 251–292.
- Hens N, *et al.* Modelling multi-sera data: the estimation of new joint and conditional epidemiological parameters. *Statistics in Medicine* 2008; **27**: 2651–2664.
- Griffiths D. A catalytic model of infection for measles. *Journal of the Royal Statistical Society, Series C* 1974; **23**: 330–339.
- Gay NJ. A model of long-term decline in the transmissibility of an infectious disease: implications for the incidence of hepatitis A. *International Journal of Epidemiology* 1996; **25**: 854–861.
- Schenzle D, Dietz K, Frösner GG. Antibody against hepatitis A in seven European countries. Statistical analysis of cross-sectional surveys. *American Journal of Epidemiology* 1979; **110**: 70–76.
- Anderson RM, May RM. Directly transmitted infectious diseases: control by vaccination. *Science* 1982; **215**: 1053–1060.
- Grenfell BT, Anderson RM. The estimation of age-related rates of infection from case notifications and serological data. *Journal of Hygiene* 1985; **95**: 19–36.
- Farrington CP. Modeling forces of infection for measles, mumps and rubella. *Statistics in Medicine* 1990; **9**: 953–967.
- Massad E, Raimundo SM, Silveira ASB. A continuous function model for the age-related force of infection. *Mathematical and Computer Modelling* 1990; **13**: 101–112.
- McCullagh P, Nelder JA. *Generalized Linear Models*. Chapman & Hall, 1989.
- Keiding N, *et al.* Estimation from current status data in continuous time. *Lifetime Data Analysis* 1996; **2**: 119–129.
- Shkedy Z, *et al.* Modeling age dependent force of infection from prevalence data using fractional polynomials. *Statistics in Medicine* 2006; **5**: 1577–1591.
- Hens N, *et al.* Modelling distortions in seroprevalence data using change-point fractional polynomials. *Statistical Modelling* (in press).
- Hens N, *et al.* Handling missingness when modelling the force of infection from clustered seroprevalence data. *Journal of Agricultural, Biological and Environmental Statistics* 2007; **12**: 1–16.
- Faes C, *et al.* Population-averaged versus herd-specific force of infection. *Journal of the Royal Statistical Society, Series C* 2006; **55**: 595–613.
- Van Effelterre T, *et al.* Contact patterns and their implied basic reproductive numbers: an illustration for varicella-zoster virus. *Epidemiology and Infection* 2009; **137**: 48–57.
- Nardone A, *et al.* The comparative sero-epidemiology of varicella zoster virus in 11 countries in the European region. *Vaccine* 2007; **25**: 7866–7872.
- Nardone A, Miller E. Serological surveillance of rubella in Europe: European sero-epidemiology network (ESEN2). *Eurosurveillance* 2004; **9**: 5–7.
- Edmunds WJ, Pebody RG, Aggerback HEA. The sero-epidemiology of diphtheria in Western Europe. European sero-epidemiology network (ESEN). *Epidemiology and Infection* 2000; **125**: 113–125.
- Edmunds WJ, *et al.* The prevaccination epidemiology of measles, mumps and rubella in Europe: implications for modeling studies. *Epidemiology and Infection* 2000; **125**: 635–650.
- Beutels P, *et al.* Hepatitis B in St Petersburg, Russia (1994–1999): incidence, prevalence and force of infection. *Journal of Viral Hepatitis* 2003; **10**: 141–149.

30. **Keiding N.** Age-specific incidence and prevalence: a statistical perspective (with discussion). *Journal of the Royal Statistical Society, Series A* 1991; **154**: 371–412.
31. **Nagelkerke N, et al.** Semi-parametric estimation of age-time specific infection incidence from serial prevalence data. *Statistics in Medicine* 1999; **18**: 307–320.
32. **Shkedy Z, et al.** Modelling forces of infection by using monotone local polynomials. *Journal of the Royal Statistical Society, Series C* 2003; **52**: 469–485.
33. **Greenhalgh D, Dietz K.** Some bounds on estimates for reproductive ratios derived from the age-specific force of infection. *Mathematical Biosciences* 1994; **12**: 49–57.
34. **Shiboski SC.** Generalized additive models for current status data. *Lifetime Data Analysis* 1998; **4**: 29–50.
35. **Hastie TJ, Tibshirani RJ.** Generalized additive models: some applications. *Journal of the American Statistical Association* 1987; **82**: 371–386.
36. **Hastie TJ, Tibshirani RJ.** *Generalized Additive Models*. New York: Chapman and Hall, 1990.
37. **Namata H, et al.** Estimation of the force of infection from current status data using generalized linear mixed models. *Journal of Applied Statistics* 2007; **34**: 923–939.
38. **van Boven M, et al.** Waning immunity and sub-clinical infection in an epidemic model: implications for pertussis in The Netherlands. *Mathematical Biosciences* 2000; **164**: 161–182.
39. **Coen PG, et al.** Mathematical models of *Haemophilus influenzae* type B. *Epidemiology and Infection* 1998; **120**: 281–295.
40. **Fine PEM, Clarkson JA.** Measles in England and Wales – I: an analysis of factors underlying seasonal patterns. *International Journal Epidemiology* 1982; **11**: 5–14.
41. **Grassly NC, Fraser C.** Seasonal infectious disease epidemiology. *Proceedings of the Royal Society of London, Series B* 2006; **273**: 2541–2550.
42. **Zaaijer HL, Koppelman MH, Farrington CP.** Parvovirus B19 viraemia in Dutch blood donors. *Epidemiology and Infection* 2004; **132**: 1161–1166.
43. **Whitaker HJ, Farrington CP.** Estimation of infectious disease parameters from serological survey data: the impact of regular epidemics. *Statistics in Medicine* 2004; **23**: 2429–2443.
44. **Mathei C, et al.** Evidence for a substantial role of sharing of injecting paraphernalia other than syringes/needles to the spread of hepatitis C among injecting drug users. *Journal of Viral Hepatitis* 2006; **13**: 560–570.
45. **Sutherland I, Svandova E, Radhakrishna SE.** The development of clinical tuberculosis following infection with tubercle bacilli. *Tubercle* 1982; **62**: 255–269.
46. **Sutherland I.** Recent studies in the epidemiology of tuberculosis, based on the risk of being infected with tubercle bacilli. *Advances in Tuberculosis Research* 1976; **19**: 1–63.
47. **Vynnycky E and Fine PEM.** The natural history of tuberculosis: the implications of age-dependent risks of disease and the role of re-infection. *Epidemiology and Infection* 1997; **119**: 183–201.
48. **Becker NG.** *Analysis of Infectious Diseases Data*. London: Chapman and Hall, 1989.
49. **Hens N, et al.** Estimating the impact of vaccination using age–time-dependent incidence rates of hepatitis B. *Epidemiology and Infection* 2008; **136**: 341–351.
50. **Amaku M, et al.** Vaccination against rubella: analysis of the temporal evolution of the age-dependent force of infection and the effects of different contact patterns. *Physical Review E* 2003; **67**: 5.
51. **Zhang YX.** A compound catalytic model with both reversible and two-stage types and its applications in epidemiological study. *International Journal of Epidemiology* 1987; **16**: 619–621.
52. **Beutels P, et al.** Hepatitis B in St Petersburg, Russia (1994–1999): incidence, prevalence and force of infection. *Journal of Viral Hepatitis* 2003; **10**: 141–149.
53. **Morris MC, et al.** Sero-epidemiological patterns of Epstein-Barr and herpes simplex (HSV-1 and HSV-2) viruses in England and Wales. *Journal of Medical Virology* 2002; **67**: 522–527.
54. **Nardone A, et al.** The comparative sero-epidemiology of varicella zoster virus in 11 countries in the European region. *Vaccine* 2007; **25**: 7866–7872.
55. **Shkedy Z, et al.** Modeling age dependent force of infection from prevalence data using fractional polynomials. *Statistics in Medicine* 2006; **5**: 1577–1591.
56. **Thiry N, et al.** The seroepidemiology of primary varicella-zoster virus infection in Flanders (Belgium). *European Journal of Pediatrics* 2002; **161**: 588–593.
57. **Whitaker HJ, Farrington CP.** Infections with varying contact rates: application to varicella. *Biometrics* 2004; **60**: 615–623.
58. **Ades AE, Nokes DJ.** Modeling age- and time-specific incidence from seroprevalence: toxoplasmosis. *American Journal of Epidemiology* 1993; **137**: 1022–1034.
59. **Mossong J, et al.** Parvovirus B19 infection in five European countries: seroepidemiology, force of infection and maternal risk of infection. *Epidemiology and Infection* 2008; **136**: 1059–1068.
60. **Gay NJ.** Analysis of serological surveys using mixture models: application to a survey of parvovirus B19. *Statistics in Medicine* 1996; **15**: 1567–1573.
61. **Bollaerts K, et al.** Estimating the force of infection directly from antibody levels. *Statistics in Medicine* (in press).
62. **Kanaan MN, Farrington CP.** Matrix models for childhood infections: a Bayesian approach with applications to rubella and mumps. *Epidemiology and Infection* 2005; **133**: 1009–1021.
63. **Coutinho FAB, et al.** Modelling heterogeneities in individual frailties in epidemic models. *Mathematical and Computer Modelling* 1999; **30**: 97–115.
64. **Anderson RM, May RM.** *Infectious Diseases of Humans: Dynamics and Control*. Oxford: Oxford University Press, 1991.
65. **Wallinga J, Teunis P, Kretzschmar M.** Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious

- agents. *American Journal of Epidemiology* 2006; **164**: 936–944.
66. **Ogunjimi B, et al.** Using empirical social contact data to model person to person infectious disease transmission: an illustration for varicella. *Mathematical Biosciences* 2009; **278**: 80–87.
67. **Goeyvaerts N, et al.** Estimating transmission parameters and the basic reproduction number using social contact data and serological data on varicella zoster virus in Belgium. *Journal of the Royal Statistical Society, Series C* (in press).
68. **Hens N, et al.** *Modeling Infectious Disease Parameters Based on Serological and Social Contact Data: A Modern Statistical Perspective*. New York: Springer-Verlag (in press).
69. **Wilson EB, Worcester J.** Contact with measles. *Proceedings of the National Academy of Sciences USA* 1941; **27**: 7–13.