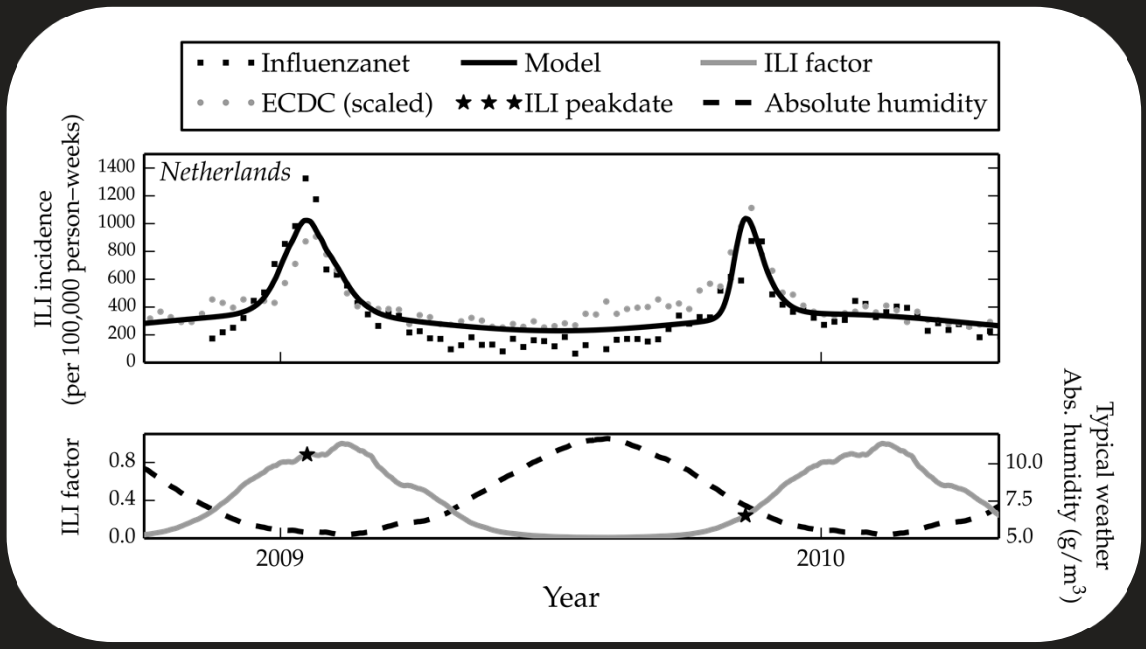
# Participatory Surveillance and Mathematical Models in Epidemiologic Research: Successes and Challenges

Sander Paul van Noort




Dissertation presented to obtain the Ph.D degree in Biology

Instituto de Tecnologia Química e Biológica António Xavier | Universidade Nova de Lisboa

Oeiras,
July, 2014

INSTITUTO
DE TECNOLOGIA
QUÍMICA E BIOLÓGICA
/UNL

Knowledge Creation

# Participatory Surveillance and Mathematical Models in Epidemiologic Research: Successes and Challenges

Sander Paul van Noort

Dissertation presented to obtain the Ph.D degree in Biology

Instituto de Tecnologia Química e Biológica António Xavier | Universidade Nova de Lisboa

Research work coordinated by:

FUNDAÇÃO CALOUSTE GULBENKIAN
Instituto Gulbenkian de Ciência

Oeiras,
July, 2014

INSTITUTO
DE TECNOLOGIA
QUÍMICA E BIOLÓGICA
/UNL

Knowledge Creation

Supervisor: Dr. M. Gabriela M. Gomes

The studies described in this thesis were funded by the European Commission (grants MEXT-CT-2004-14338 and EC-ICT-231807)

# CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# SUMMARY

Theoretical epidemiology aims to understand the dynamics of diseases in populations and communities. Biological and behavioral processes are abstracted into mathematical formulations which aim to reproduce epidemiological observations. In this thesis a new system for the self-reporting of syndromic data — Influenzanet — is introduced and assessed. The system is currently being extended to address greater challenges of monitoring the health and well-being of tropical communities. Mathematical transmission models for influenza and malaria are constructed and fitted to epidemiological observations, informing the design of future editions of self-reporting surveillance.

Influenzanet is an internet-based system to monitor influenza-like illness (ILI) throughout Europe using self-reporting volunteers. It consists of national websites where everybody is invited to register by completing an intake questionnaire, and participants are weekly reminded to complete a symptoms questionnaire. The ILI activity is determined by applying a clinical case definition.

Based on the reports of tens of thousands of volunteers across Europe, the ILI incidence as determined by Influenzanet closely follows the trends reported by traditional surveillance based on sentinel medical doctors. Since Influenzanet applies a single case definition uniformly across Europe and does not depend on a country-dependent rate at which people seek medical care,

the measured ILI activity can be directly compared across countries, which is not possible in tradition surveillance.

Based on the follow-up questions of the symptoms questionnaire, Influenzanet also monitors behavioral characteristics of participants with ILI in different countries and subgroups, such as how often and when they seek medical care. ILI risk factors and season dependent vaccine effectiveness are determined based on life style and demographic information supplied at registration. The flexibility of the system has allowed various extensions, such as questionnaires concerning stress, contact patterns, and vaccine side effects. Based on different case definitions other diseases such as dengue are monitored. Given the automated real-time data collection and analyses, in countries with sufficient participants the yearly onset of the ILI epidemic can be detected up to two weeks earlier than with the traditional surveillance.

According to the measured ILI activity by both Influenzanet and the sentinel doctors, during seasons when the activity starts early due to the emergence of a new strain, the epidemic is not necessarily more severe as one could naively expect. In this thesis it is hypothesized that this phenomenon is due to an increased probability for an influenza infected person to have ILI symptoms when the temperature and absolute humidity are lower. Including this so-called ILI factor in a transmission model significantly improves the fit to the observed ILI activity. The model further predicts that public health measures which aim to delay the onset of an epidemic or pandemic, such as school closures and global vaccination, could have the averse effect of increasing disease burden if this causes the epidemic to unfold in colder and dryer months.

The yearly influenza epidemics are sustained by antigenic drift, a phenomenon by which each year new strains of influenza A arise due to mu-

tations that enable to virus to escape population immunity and gradually replace older strains. Infectious diseases with antigenic diversity pose a challenge to modeling the transmission dynamics. *Plasmodium falciparum*, which causes the most severe forms of human malaria, is also characterized by immense antigenic diversity. During the blood phase of infection, the parasite expresses certain variant surface antigens (VSAs) on the surface of infected red blood cells. Of special importance is the PfEMP1 family, which constitutes a major target of the immune system. Each malaria parasite has the ability to express around 60 different VSAs and generally only one is predominantly expressed at a given time in a way that appears orchestrated by the host immune system. A relatively conserved subset of these VSAs is preferentially expressed in non-immune patients and related to severe malaria. Empirical evidence suggests that most parasites are able to express VSAs from this conserved subset.

In this thesis a mathematical model is constructed, which integrates both a between-host element regulating the transmission dynamics, and a within-host element regulating the VSA expression. The main hypothesis is that VSAs can be organized into blocks which are characterized by their dominance. An infected host will express the most dominant VSA block for which it has no prior immunity, and an infected host can be superinfected if the invading parasite can express a more dominant VSA block. This generates two opposing selective forces within the parasite population, favoring both high-dominance VSAs with the ability to superinfect, and increasingly diverse subsets of low-dominance VSAs to escape the immunity of the population. The model correctly reproduces the observed serological trends, as well as a parasite population where most parasites can express both the most and least dominant VSA blocks.

# SUMÁRIO

Epidemiologia teórica tem como objetivo compreender a dinâmica de doenças em populações e comunidades. Os processos biológicos e comportamentais são transcritos em formulações matemáticas que visam reproduzir observações epidemiológicas. Nesta tese um novo sistema para o auto-preenchimento de dados sindrômicos — Influenzanet — é introduzido e avaliado. O sistema está a ser estendido para dar resposta aos desafios de monitorizar saúde e bem-estar em comunidades tropicais. Modelos matemáticos de transmissão da gripe e malária são construídos e ajustados a observações epidemiológicas, informando o desenho de edições futuras de vigilância com base em auto-preenchimento voluntário.

Influenzanet é um sistema que monitoriza a síndrome gripal em toda a Europa, pela Internet, com base em participantes voluntários. Consiste em *websites* nacionais onde todos os habitantes estão convidados a inscrever-se preenchendo um questionário de adesão. Após a inscrição, os participantes passam a ser lembrados semanal de preencher um questionário de sintomas. A atividade da síndrome gripal é determinada aplicando uma definição de caso clínico.

Utilizando dezenas de milhares de relatórios de voluntários na Europa, a incidência da síndrome gripal determinada pelo Influenzanet segue estreitamente as tendências reportadas pelo sistema de vigilância tradicional baseado em redes de Médicos-Sentinela. Uma vez que o Influenzanet aplica uma única

definição de caso em toda a Europa e não depende da percentagem de pessoas que procuram assistência médica (que varia entre países) a atividade de síndrome gripal pode ser comparada diretamente entre países, o que não é possível no sistema de vigilância tradicional.

O questionário de sintomas também permite ao Influenzanet monitorizar o comportamento dos participantes com síndrome gripal, em particular no que respeita a procura de cuidados médicos. Fatores de risco da síndrome gripal e a eficácia da vacina — que varia de ano para ano — são determinados com base nas informações recolhidas pelo questionário de adesão. A flexibilidade do sistema tem permitido várias extensões, tal como questionários sobre stress, padrões de contato, e os efeitos secundários da vacina. Usando definições de caso diferentes, a sistema é usado para monitorizar outras doenças, como o dengue. A recolha de dados em tempo real e as análises automáticas permitem que, em países com participantes suficientes, o início da epidemia seja detectado precocemente, com uma antecipação que pode ir até duas semanas relativamente à vigilância tradicional.

Segundo dados do Influenzanet e da vigilância tradicional, a atividade da síndrome gripal parece menor em épocas em que a atividade começa cedo. Uma hipótese considerada nesta tese é que este fenómeno se deve a um aumento da probabilidade de uma pessoa infectada com influenza ter sintomas quando a temperatura e a humidade absoluta são mais baixos. Incluindo este fator num modelo de transmissão melhora significativamente o ajuste às séries temporais de síndrome gripal. O modelo prevê assim que medidas de saúde pública que visam retardar a epidemia ou pandemia, como o encerramento de escolas e vacinação em massa, podem ter efeitos adversos e aumentar o número de casos da doença se a epidemia for deslocada para meses mais frios e secos.

As epidemias de gripe anuais são sustentadas por um fenómeno evolutivo de deriva antigénica, segundo o qual em cada ano novas estirpes surgem por mutações que permitem ao vírus escapar a imunidade da população e substituir formas anteriores. As doenças infecciosas com diversidade antigénica representam um desafio importante na modelagem da dinâmica de transmissão. *Plasmodium falciparum*, que causa a forma mais severa da malária humana, é caracterizada por uma diversidade antigénica imensa. Durante a fase sanguínea da infecção, o parasita expressa antigénios variáveis (VSAs) na superfície dos glóbulos vermelhos infectados. Com especial importância temos a família PfEMP1, que constitui um alvo importante para o sistema imunológico. Cada parasita tem a capacidade de expressar cerca de 60 VSAs diferentes e geralmente apenas uma variante é predominantemente expressa de cada vez de forma que se julga orquestrada pelo sistema imune do hospedeiro. Um subconjunto relativamente conservado destes VSAs é expresso preferencialmente em pacientes não-imunes e aparece relacionado com formas mais severas da doença. Existe também evidência empírica de que a maioria dos parasitas é capaz de expressar VSAs deste subconjunto conservado.

Nesta tese um modelo matemático é construído, que integra tanto um elemento extra-hospedeiro que regula a dinâmica de transmissão como um elemento intra-hospedeiro que regula a expressão dos VSAs. A hipótese principal é que os VSAs podem ser organizados em blocos que são caracterizados pela seu grau de dominância. Uma pessoa infectada vai expressar VSAs do bloco mais dominante para qual o organismo não tenha imunidade prévia, e uma pessoa infectada pode ser superinfectada se o parasita invasor expressar um VSAs de dominância superior. Isto gera duas forças seletivas opostas na população de parasitas, favorecendo VSAs de alta dominância com ca-

pacidade de superinfecção por um lado, e subconjuntos diversos de VSAs de baixa dominância para escapar à imunidade da população por outro lado. O modelo reproduz corretamente as tendências sorológicas observadas em estudos de campo, e reproduz uma população de parasitas onde a maioria pode expressar blocos de VSAs nos extremos da escala de dominância.

## PRINCIPAL PUBLICATIONS

Chapter 2
Sander P. van Noort, Marion Mühlen, Helena Rebelo de Andrade, Carl E. Koppeschaar, José M.L. Lourenço, M. Gabriela M. Gomes. "Gripenet: an internet-based system to monitor influenza-like illness uniformly across Europe". *Eurosurveillance (2007). 12(7): E5-6*

Chapter 3
Sander P. van Noort, Cláudia Codeço, Carl E. Koppeschaar, Marc van Ranst, Daniela Paolotti, M. Gabriela M. Gomes. "Ten-year performance of Influenzanet: ILI time series, risks, and vaccine effects in the Grote Griepmeting, Gripenet, and Influweb cohorts". *Epidemics (2014). minor revision*

Chapter 4
Sander P. van Noort, Ricardo Águas, Sébastien Ballesteros, M. Gabriela M. Gomes. "The role of weather on the relation between influenza and influenza-like illness". *Journal of Theoretical Biology (2012). 298: 131–137*

Chapter 5
Sander P. van Noort, Marta C. Nunes, Gareth D. Weedall, Lars Hviid, M. Gabriela M. Gomes. "Immune selection and within-host competition can structure the repertoire of variant surface antigens in *Plasmodium falciparum*". *PLoS One (2010). 5(3): e9778*

Appendix A
Sander P. van Noort, Nico Stollenwerk. "From dynamical processes to likelihood functions: an epidemiological application to influenza". *Computational and Mathematical Methods in Science and Engineering (2008). ISBN 978-84-612-1982-7*

Sander P. van Noort, Nico Stollenwerk, Lewi Stone. "Analytic likelihood function for data analysis in the starting phase of an influenza outbreak". *Computational and Mathematical Methods in Science and Engineering (2009). ISBN 978-84-612-9727-6*

Detailed Influenzanet results
http://www.influenzanet.eu/results

# ACKNOWLEDGMENTS

Gabriela Gomes, without whose continuous trust and patience, this thesis would have never been completed.

Gabriela Gomes, Odo Diekmann, Jacco Wallinga, Sido Mylius, Cláudia Codeço, Gareth Weedall, Helena Rebelo de Andrade, Lars Hviid, Marc van Ranst, Marion Mühlen, Marta Nunes, Nico Stollenwerk, Ricardo Águas, Richard Marquet, Sébastien Ballesteros, Zé Lourenço, Bob Planqué, Carl Koppeschaar, Marc Peletier, Martin Takken, Ronald Smallenburg, Isabel Gordo, Jorge Carneiro, Maite Severins, Nuno Sepúlveda, Suani Pinho, Ana Franco, Cátia Bandeiras, Caetano Mendes, Delphine Pessoa, Dinis Gökaydin, Erida Gjini, Frank Hilker, Flávio Coelho, Greg King, Guilherme Conçalves, Jean-Baptiste André, João Lopes, Kirsten Schmitz, Maíra Aguiar, Marta Lopes, Natalia Mantilla Beniers, Patrícia Soares, Paula Rodrigues, Philip Gerrish, Vítor Faustino, Vívian, Joana, Catarina, Tim, Pa, Ma, Iris, Gerbrand, Sam, Stijn, Flávia.

*Misschien is het beter om iets niet te hebben,*
*dan om het gelijk al weer kwijt te zijn*

Acda en De Munnik, Als je me morgen ziet

*Soms is het beter iets moois te verliezen,*
*beter verliezen dan dat je het nooit hebt gehad*

Rowwen Hèze, Heilige Anthonius

# 1

# INTRODUCTION

*The grand aim of all science is to cover the greatest number of empirical facts by logical deduction from the smallest number of hypotheses or axioms.*

Albert Einstein

## 1.1 THEORETICAL EPIDEMIOLOGY

Theoretical epidemiology aims to understand the dynamics of diseases in populations and communities. Biological, medical and social processes are translated into mathematical models which aim to reproduce epidemiological observations, often based on public health data. The biological processes are usually individual oriented, such as the number of infectious contacts, whereas the epidemiological observations are usually measured in the population, such as the total number of infections.

A classic model in theoretical epidemiology is the so-called SIR-model, which describes the transmission of an infectious disease in a host population. The population (size $N$) is classified into three subgroups: those who are susceptible ($S$), those who are currently infected ($I$), and those who have recovered from an infection and are current immune ($R$). This infection process can be visualized as in Figure 1.1A.

To transform this infectious process into a mathematical model, the "infection" and "recovery" processes have to be defined. For the deterministic model, a common assumption is that infected hosts make $\beta$ infectious contacts per week and recover at a rate $\tau$. If each contact is made at random, so-called homogeneous mixing, the rate of new infections per week is then determined by the total number of infectious contacts per week ($\beta I$) and the probability that a contact is made with a susceptible host ($S/N$). In mathematical terms,

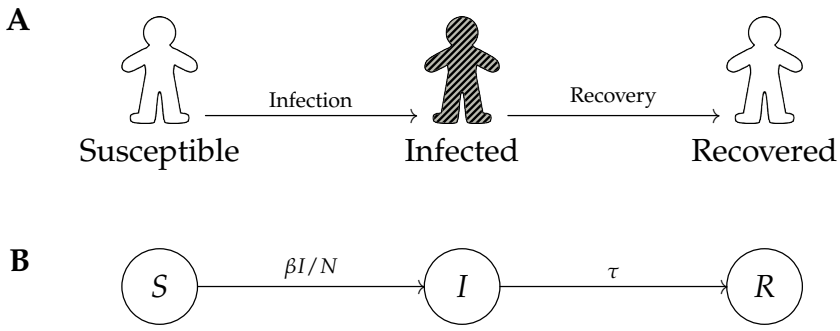



Figure 1.1: SIR MODEL

the model can be illustrated by Figure 1.1B and described by the following differential equations:

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta I \frac{S}{N}$$
$$\frac{\mathrm{d}I}{\mathrm{d}t} = \beta I \frac{S}{N} - \tau I$$

This system was first described by (KERMACK AND MCKENDRICK, 1927) and forms the basis for many transmission models of infectious diseases. Accessible books which introduce the mathematical modeling of infectious diseases are written by (ANDERSON AND MAY, 1991) and (DIEKMANN AND HEESTERBEEK, 2000), among others.

Some individual biological and behavioral processes are difficult to measure directly, such as the transmission rate of influenza in the previously described SIR model. Although certain elements of the transmission process can be measured directly, such as the viral shedding of infected persons (CARRAT ET AL., 2008) and the number and type of personal contacts made per person (MOSSONG ET AL., 2008), it is difficult to actually quantify the transmission rate in the real world based on all different aspects of transmission.

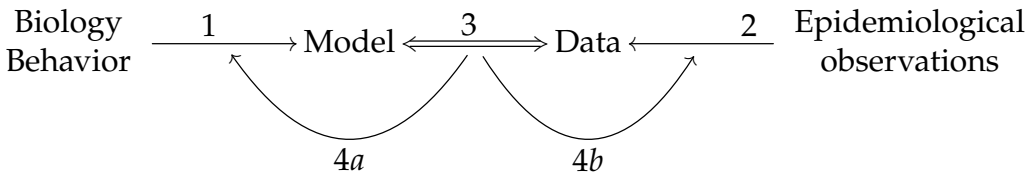



Figure 1.2: THEORETICAL EPIDEMIOLOGY

Theoretical epidemiology allows the indirect quantification of these individual processes, by fitting transmission models to epidemiological data and discovering the most likely elements and parameters. This process can also be applied in reverse: mathematical models can be used to estimate the outcome of the intervention in certain individual processes. For example, by assuming that closing schools or airports decreases the effective transmission of a disease, the expected decrease in infections can be estimated (COLIZZA ET AL., 2006). Mathematical models thus enable translation between individual biological and behavioral processes on one hand, and population-based epidemiological and serological observations on the other (Figure 1.2).

The construction of a mathematical model is usually an iterative process as visualized in Figure 1.2, and all steps are explored in this work.

1. Possible relevant biological and behavioral processes are selected and transcribed into a model.
2. Epidemiological observations are gathered into a database.
3. The model is fitted to epidemiological data.
4. Based on how well the model fits the data, if necessary:
    a) Selection of the relevant biological and behavioral processes (step 1) is revised.
    b) Data collection protocol (step 2) is updated.

## 1.2 MODEL FITTING

Although a transmission model describes the actual number of infected individuals in the population, the epidemiological data might contain observation errors. Different observation uncertainties can be distinguished, which will be illustrated based on a simple SIR model applied to influenza. At any time, the observed number of persons with influenza-like illness ($ILI_t$), is considered to depend on the number of influenza-infected ($I_t$) by the following equation:

$$ILI_t = p_t \cdot I_t + J_t + \epsilon_t \tag{1.1}$$

where $p_t$ denotes an observation bias, $J_t$ a confounding variable, and $\epsilon_t$ a measurement error.

OBSERVATION BIAS Consistent over- or under-quantification of the data is known as an observation bias, as indicated by $p_t$ in equation (1.1). Not all persons with influenza are observed since not all influenza infections are symptomatic, and not all symptomatic infected have all the symptoms to fit the ILI case definition. Furthermore, if the observed number of ILI cases are provided by general practitioners, only the persons with ILI who have visited their doctor are counted.

CONFOUNDING The observation of interest may be due to other causes, generating confounding, indicated by the fraction $J_t$ in equation (1.1). During the season a number of persons will have symptoms which fit the ILI case definition, independent of the presence of influenza, for example due to other severe respiratory infections such as respiratory syncytial virus (RSV).

MEASUREMENT ERROR A measurement error indicates that each observed value differs from the actual value in the population, indicated by the additional term $\epsilon_t$ in equation (1.1). A common assumption is that all errors $\epsilon_t$ are independent and follow from a normal distribution, such that the most likely parameters can be determined using the well-known least-squares method.

The measurement error may also be due to a sampling bias. The number of ILI cases is usually observed within a certain cohort, whereas the influenza transmission process occurs in the whole population. The expected number of ILI cases in a cohort of $P$ persons, given that a fraction $i$ of the population has ILI, follows from a binomial distribution $ILI \sim \text{Bin}(P, i)$, which is known as a binomial experiment.

## 1.3 DATA COLLECTION: ORIGINS OF INFLUENZANET

A common source of epidemiological data is the counting of disease cases over time in association with plausible risk factors. These data are traditionally collected via the public health system: medical doctors provide reports on patients they have diagnosed with a certain disease, either clinical or by laboratory diagnosis. Relatively recently, the meteoric rise of the Internet, in particular "web 2.0", where Internet users also actively participate and submit data, has created a new path to survey the population.

On 1 January 2002, twelve European countries introduced the euro. Euro coins have a European side, which is equal in every country, and a national side, which indicates in which country the coin is produced. In the Netherlands, a project "Eurodiffusie" (Euro diffusion) was setup which consisted of a website[1], where people could weekly report the nationality of the coins cur-

1 http://www.eurodiffusie.nl

rently in their wallet. The project monitored the diffusion of euro coins from various countries into the Netherlands over time (HEK ET AL., 2002).

The success of the "Eurodiffusie" project was overwhelming, which gave Carl Koppeschaar (popular science journalist) the idea to build a similar system to monitor the presence of flu and common cold in the population. Together with people from the "Eurodiffusie" project, Marc Peletier and Bob Planqué, a core team was formed consisting of Sander van Noort (mathematical modeler), Martin Takken (programmer), and Ronald Smallenburg (entrepreneur), together with the help of people from various areas such as medical doctors, teachers, biologists, journalists, and epidemiologists. The first action meeting took place on 17 September 2003, and 6 weeks later the website degrotegriepmeting.nl was launched, on 1 November 2003.

The primary objective of 'Grote Griepmeting" was to let people actively participate on an interesting health subject: flu and the common cold. Everybody can register by completing an intake questionnaire containing life style and demographic questions. Participants receive a weekly newsletter which directs them to a questionnaire to report any flu-like symptoms they had in the preceding week. Based on the collected data, the activity of influenza-like illness (ILI) is determined and graphics and maps are published online. The website was setup to actively involve participants into the system with real-time results, discussion forums, competitions and educational material.

In response to prime time media attention, over 25,000 Dutch and Belgian participants registered for "Grote Griepmeting" during the first season. This participation success and the collected data made the system also interesting from a public health point of view. The system was consequently implemented in Portugal as "Gripenet" (2005), in Italy as "Influweb" (2008), and in the United Kingdom as "FluSurvey" (2009). In 2009 an FP7 project (Epiwork)

started, with one of its goals to introduce the system in more European countries under a common name Influenzanet. Influenzanet was implemented in Sweden as "Influensakoll" (2010) in France as "Grippenet" (2011), in Spain as "Gripenet" (2012), in Ireland as "FluSurveyIE" (2012), and in Denmark as "Influmeter" (2013). The Portuguese team (coordinated by Gabriela Gomes) also helped to introduce the system in Latin America: in Brazil as "Gripenet" (2009), in Mexico as "Reporta" (2009), and in Salvador, Brazil as "Dengue-naweb" (2010), a similar system focusing on dengue. Independently, similar systems were implemented in Australia as "FluTracking" (2007), and in the United States as "FluNearYou" (2011).

## 1.4 MULTIPLE STRAIN DISEASES

The basic SIR model as illustrated in Figure 1.1 is a viable model for most childhood diseases, such as measles, mumps and rubella, where infection (or vaccination) leads to life-long protection. However, there are also various diseases for which multiple strains are in circulation, and infection by one strain only gives partial immunity against infection by other strains (GOMES ET AL., 2002).

For Influenza A H3N2, each year new strains arise due to mutation, which gradually replace the older strains. The circulating strains each season are closely related, generating a ladder-like phylogenetic tree. A recently infected person will have high crossimmunity against all circulating strains, but over the years the circulating strains continue to mutate, which leads to a decrease in immunity and to an eventual new infection. Various distinct mechanisms have been proposed to explain the characteristic shape of influenza phylogenetics (FERGUSON ET AL., 2003; VAN NOORT, 2005; KOELLE ET AL., 2006; PARISI

ET AL., 2013), but in this thesis this debate is circumvented and multi-season influenza is modeled by assuming full crossimmunity for a single season, whereas for each new season the model estimates independently how many people will have become susceptible again to the new circulating strains.

In the final chapter a more unusual type of multi-strain diversity is described based on *Plasmodium falciparum*, which causes the most severe forms of human malaria. Although for most infectious diseases, immunity depends on the specific strain a host has been infected with, the main immunity for *Plasmodium falciparum* is directed at specific variant surface antigens (VSAs) that are expressed by the parasite during infection. Each parasite has the capacity to express ~60 different VSAs, of which during a single infection only a subset are predominantly expressed to provoke an immune response. Immune selection is modeled as a two-level process: a between-host element which describes the disease dynamics of new infections, and a within-host element which describes which VSAs are expressed within an infected host.

## 1.5 THESIS OUTLINE

Chapter 2 introduces a new system for the collection of epidemiological data, Influenzanet, which determines ILI activity in the population based on self-reporting volunteers. The data collection method is explained and put in comparison with the traditional surveillance system based on health care-seeking patients (ECDC). Results for the Netherlands, Belgium, and Portugal during the season 2006–2007 are generated and compared with the ECDC data.

In Chapter 3 the collected Influenzanet data in eight participating countries for the seasons 2003–2013 are assessed, by performing a time series analyses on the ILI incidences of Influenzanet and ECDC, and by comparing the

determined ILI risk factors with those encountered in other epidemiological studies. Furthermore, the vaccine effectiveness for ILI is determined for each season, and how often and when participants with ILI seek medical care is compared across countries.

In Chapter 4 an influenza transmission model is fitted to ILI incidence from ECDC and Influenzanet in the Netherlands, Belgium, and Portugal for the seasons 2003–2013. The ILI factor, the probability that a person infected with influenza actually fits the ILI case definition, is introduced and linked directly to the measured absolute humidity and temperature.

In Chapter 5 the antigenic diversity of VSAs in the parasite population of *Plasmodium falciparum* is explored. To reproduce serological and epidemiological observations, the concept of dominance between difference VSA blocks is introduced, regulating which VSA block will be expressed upon infection. This chapter is concerned with malaria, and the remaining with influenza, exposing the generality of the adopted research procedures.

In Appendix A a new analytic framework for parameter estimation for stochastic transmission models is developed, and presented for theoretical insight, although in this thesis deterministic models are fitted to the data using numerical procedures.

## 1.6 REFERENCES

ANDERSON RM and MAY RM (1991). *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press.

CARRAT F, VERGU E, FERGUSON NM, LEMAITRE M, CAUCHEMEZ S, ET AL. (2008). Time lines of infection and disease in human influenza: a review of volunteer challenge studies. *Am J Epidemiol*, 167(7):775–785.

COLIZZA V, BARRAT A, BARTHÉLEMY M, and VESPIGNANI A (2006). The role of the airline transportation network in the prediction and predictability of global epidemics. *Proc Natl Acad Sci U S A*, 103(7):2015–2020.

DIEKMANN O and HEESTERBEEK J (2000). *Mathematical Epidemiology of Infectious Diseases*. John Wiley & Sons Ltd.

FERGUSON NM, GALVANI AP, and BUSH RM (2003). Ecological and immunological determinants of influenza evolution. *Nature*, 422(6930):428–433.

GOMES MGM, MEDLEY GF, and NOKES DJ (2002). On the determinants of population structure in antigenically diverse pathogens. *Proc Biol Sci*, 269(1488):227–233.

HEK G, NUIJENS M, PLANQUE B, and VAN DER PLOEG H (2002). Het grote internationale eurodiffusie-experiment. *Natuur en Techniek*, 70(11):56–62.

KERMACK W and MCKENDRICK A (1927). A contribution to the mathematical theory of epidemics. *Proc. Royal Soc. (A)*, 115:700–21.

KOELLE K, COBEY S, GRENFELL B, and PASCUAL M (2006). Epochal evolution shapes the phylodynamics of interpandemic influenza a (h3n2) in humans. *Science*, 314(5807):1898–1903.

MOSSONG J, HENS N, JIT M, BEUTELS P, AURANEN K, ET AL. (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med*, 5(3):e74.

PARISI A, LOPES JS, NUNES A, and GOMES MGM (2013). Heterogeneity in antibody range and the antigenic drift of influenza a viruses. *Ecological Complexity*, 14(0):157 – 165. ISSN 1476-945X.

VAN NOORT SP (2005). *The Genetic Drift of Influenza A*. Master's thesis, Utrecht University.

# 2

## INFLUENZANET: AN INTERNET-BASED SYSTEM TO MONITOR INFLUENZA-LIKE ILLNESS UNIFORMLY ACROSS EUROPE

Sander P. van Noort[1], Marion Mühlen[1], Helena Rebelo de Andrade[2], Carl E. Koppeschaar[3], José M. Lima Lourenço[1], M. Gabriela M. Gomes[1]

[1]Instituto Gulbenkian de Ciência, Oeiras, Portugal

[2]Instituto Nacional de Saúde Dr. Ricardo Jorge, Lisbon, Portugal

[3]Kennislink, Amsterdam, Netherlands

*Eurosurveillance (2007) Jul 1; 12(7): E5–6*

ABSTRACT

*Influenzanet has been monitoring the activity of influenza-like illness (ILI) with the aid of volunteers via the internet in the Netherlands and Belgium since 2003 and in Portugal since 2005. In contrast with the traditional system of sentinel networks of mainly primary care physicians collected via the European Influenza Surveillance Scheme (EISS), since 2008 coordinated by the European Centre for Disease Control and Prevention (ECDC), Influenzanet obtains its data directly from the population.*

*Any resident of the three countries can participate in Influenzanet by completing an application form on the appropriate websites — degrotegriepmeting.nl in the Netherlands and Belgium, gripenet.pt in Portugal — which contain various medical, geographic and behavioral questions. Participants report weekly on the website any symptoms they have experienced since their last visit. ILI incidence is determined on the basis of a uniform case definition.*

*In the 2006–2007 season, 19,623 persons participated in Influenzanet in the Netherlands, 7,025 in Belgium, and 3,118 in Portugal. The rise, peak and decline of ILI activity occurred at similar times according to Influenzanet and ECDC. However, ILI attack rates in the Netherlands (8.7%), Belgium (8.1%) and Portugal (7.3%) were remarkably more similar in Influenzanet than in ECDC (0.8%, 3.9%, and 0.6% respectively).*

*Monitoring ILI activity with the direct participation of volunteers provides similar incidence curves compared to the traditional system coordinated by ECDC. Whereas ECDC provides an established system whose data is validated by virology tests, Influenzanet is a fast and flexible monitoring system whose uniformity allows for direct comparison of ILI rates between countries. A current objective of Influenzanet is to engage more European countries.*

## 2.1 INTRODUCTION

During the winter 2003–2004 season, the Netherlands and Belgium launched a system to monitor the activity of influenza-like illness (ILI) with the help of volunteers via the Internet (degrotegriepmeting.nl). The success of this initiative, which attracted over 25,000 participants in the first year, inspired the establishment of a similar system in Portugal in 2005–2006 (gripenet.pt). Throughout this paper, the system is referred to as "Influenzanet".

Traditionally, influenza surveillance in Europe is monitored by the European Influenza Surveillance Scheme, since 2008 coordinated by the European Centre for Disease Control and Prevention (ECDC), a collaborative program of mainly primary care physicians, epidemiologists and virologists who actively collect clinical and virological data on influenza. In this paper, we argue that the Influenzanet monitoring system in which the data is gathered directly from the population offers some advantages over the established surveillance system based on the network of general practitioners (GPs). It has previously been shown that the participants of Influenzanet in 2003–2004 in the Netherlands were representative for the Dutch population (MARQUET ET AL., 2006). Here, we compare Influenzanet results in the three countries with the ECDC results from the same countries during the 2006–2007 season.

Table 2.1: INTAKE QUESTIONNAIRE FOR INFLUENZANET during the seasons 2003–2011.

| QUESTION | ANSWERS |
|---|---|
| Postal code | First digits |
| Birth date | Year and month |
| Gender | Male; Female |
| Daily routine | School; Work; Home; Reformed; Other |
| Daily means of locomotion | Bicycle; Motorcycle; By foot; Car; Public transport |
| Colds per year | <2; 2–5; >5 |
| Vaccination [1] | Yes; No |
| Reasons vaccinated[1,2] | GP recommendation; Protect me; Protect others; Other |
| Reasons not vaccinated[1,2] | GP recommendation; No protection; Causes flu; Side effects; Still planning; No risk group; Other |
| Chronic diseases | Asthma or lung disease; Diabetes; Heart disease; Kidney disorder; Immunodeficiency |
| Allergy | Hay fever; House dust mites; Pets |
| Smoking | Daily; Sometimes; Never |
| Fruits and vegetables[3] | Regularly; Sometimes; Hardly ever |
| Vitamin supplements[3] | Regularly; Sometimes; Hardly ever |
| Diet[3] | Vegetarian; Vegan; Low calorie; Other |
| Sports hours per week | <1; 2–4; >4 |
| Household | Alone; Only with adults; With children |
| Occupation of children | Home; Nursery; School |
| Pets at home | Cats; Dogs; Birds; Other |

[1] In 2009–2010 also for H1N1pdm vaccine
[2] Since 2007
[3] Not in Portugal

Table 2.2: SYMPTOMS QUESTIONNAIRE FOR INFLUENZANET during the seasons 2003–2011.

| QUESTION | ANSWERS |
|---|---|
| Symptoms since last visit | Nasal congestion; Cough; Sore throat; Headache; Myalgia (muscle or joint pain); Chest pain; Stomach ache; Diarrhea; Nausea; Chills; Eye irritation; Vomiting |
| Start date symptoms | |
| Measured temperature | Not measured; <37; 37–37.5; 37.5–38; 38–38.5, 38.5–39; 39–39.5, 39.5–40, >40 °C |
| Start date fever | |
| Fever started abruptly | Yes; No; Don't know |
| Visit to medical doctor[1] | Yes; No |
| Daily routine alteration[1] | Yes, stayed at home; Yes, but went to work/school; No |
| Days at home | |
| Vaccinated[2] | Yes; No |

[1] Only when symptoms or fever are present

[2] Every 4 weeks to participants who were not vaccinated

## 2.2 METHOD

### 2.2.1 *Influenzanet*

Influenzanet is a fully internet-based system, currently hosted on two websites: degrotegriepmeting.nl for the Netherlands and Belgium (Flanders), and gripenet.pt for Portugal. Any resident of these countries can register for Influenzanet by completing an online application form containing various medical, geographic, and behavioral questions (Table 2.1). Participants are mainly recruited via mass-media, which present information on the system and give regular updates of the latest results. Participation is further stimulated by email newsletters, online educational materials, competitions, presentations, and other activities. Once registered, participants receive a weekly email newsletter reminding them to complete their symptoms questionnaire. In this questionnaire participants are asked to select from a list of symptoms the ones they have experienced since their previous visit to the Influenzanet website (Table 2.2). If symptoms are reported, participants are asked to provide the date of onset, and whether these led to change of behavior and/or a GP consultation, and if so, the outcome of the consultation.

The incidence of ILI is determined based on the symptoms reported, using a uniform case definition. ILI is defined as acute onset of fever (a measured temperature $\geq$38 °C), plus muscle pain or headache, plus cough or sore throat. The day of fever onset determines the onset of ILI. A participant is considered to be active between the day of registration and the day of the last completed symptoms questionnaire. Only participants who have completed at least three symptoms questionnaires are included in the analysis. The daily incidence is determined by the number of participants with an onset of ILI on a given

day, divided by the number of active participants on that day. The weekly incidence for each day is determined by the total number of participants with an onset of ILI in the previous seven days, divided by the average number of active participants during those seven days (participant-weeks). The ILI attack rate for both Influenzanet and ECDC is defined as the cumulative incidence rate over the total surveillance period.

### 2.2.2 *ECDC*

During the 2005–2006 influenza season, 39 countries were members of European Influenza Surveillance Scheme (EISS), the predecessor for the European Influenza Surveillance Network (EISN) as coordinated by the ECDC, and the sentinel surveillance was carried out by 21,162 GPs, pediatricians and other physicians. The population under clinical surveillance by the sentinel networks represents at least a median number of 24.8 million inhabitants of Europe. The population under surveillance in the Netherlands accounts for 0.7% of the total population, in Belgium 0.4%, and in Portugal 0.7%. Although there are differences in the general characteristics of the sentinel systems in each of the countries, the majority collect weekly incidences of ILI cases per 100,000 inhabitants, as is the case in the Netherlands, Belgium, and Portugal (European Influenza Surveillance Scheme, 2007). The different case definitions used in these countries are shown in Table 2.3. Using historical data, several countries within ECDC introduced an influenza activity baseline. The intensity of influenza activity is determined by measuring the influenza activity against the baseline and its geographical spread. A proportion of the sentinel physicians additionally collect nose and/or throat swabs for virological surveillance according to a uniform swabbing protocol. The weekly

Table 2.3: INFLUENZA-LIKE ILLNESS CASE DEFINITIONS as used by Influenzanet and by the sentinel GPs reporting to ECDC in the Netherlands, Belgium, and Portugal. Influenzanet uses the same ILI case definition in all countries.

| SYSTEM | ILI CASE DEFINITION |
|---|---|
| Influenzanet | Acute onset, and measured temperature $\geq$38 °C, and muscle pain or headache, and cough or sore throat. |
| ECDC: Netherlands | An acute onset (i.e. at most a prodromal stage of three to four days), accompanied by a rise in rectal temperature of $>$38 °C, and at least 1 of the following symptoms: cough, coryza, sore throat, frontal headache, retrosternal pain, myalgia. (Pel criteria) |
| ECDC: Belgium | Sudden onset with fever, myalgia and respiratory symptoms (cough or thoracic pain) |
| ECDC: Portugal | 6 of the following criteria: sudden onset, fever, cough, chills, prostration and weakness, myalgia or general pain, rhinitis and/or pharyngitis, contact with a case. |

incidence covering the period from Monday to Sunday is published on the ECDC website the following Wednesday or Thursday. This number is usually updated one week later to include the latest available information.

## 2.3 RESULTS

We compared the Influenzanet and ECDC data from 15 December 2006 to 1 May 2007. Influenzanet data showed that in the Netherlands, 17,056 out of 19,623 participants (87%) completed at least three symptoms questionnaires, in Belgium 6,062 out of 7,025 (86%) and in Portugal 2,167 out of 3,118 (69%). The national participation rate was 0.1% in the Netherlands (total population 16.3 million), 0.1% in Flanders (6.2 million), and 0.02% in Portugal (10.5 mil-

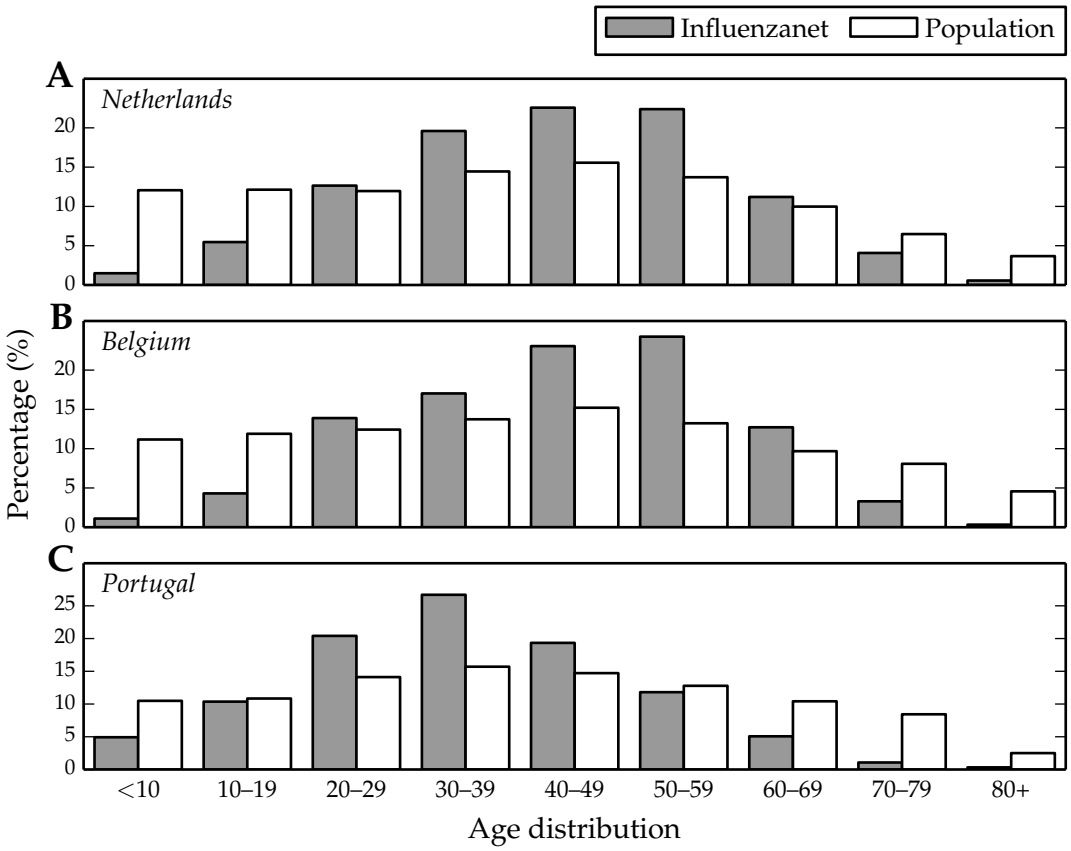



Figure 2.1: AGE DISTRIBUTIONS OF INFLUENZANET PARTICIPANTS compared with the national age distribution in A) Netherlands, B) Belgium, and C) Portugal (2006–2007). In the Netherlands and Belgium the age groups $<$20 and $\geq$70 years are underrepresented, with a very clear under representation in the age groups $<$10 and $\geq$80 years. In Portugal the age groups $<$ 10 and $\geq$60 years are underrepresented, with a very clear underrepresentation in the age group $\geq$70 years. Age distributions are downloaded from Eurostat (epp.eurostat.ec.europa.eu).

lion). In all three countries, the younger and older age groups are underrepresented (Figure 2.1). The geographical distribution of participants follows the patterns of population density, with higher concentration in the larger cities.

In the Netherlands, 31% participants matching the ILI case definition visited a GP (477 out of 1536), in Belgium 71% (321 out of 453) and in Portugal 56% (71 out of 127). Influenza activity in Europe was, in 2006–2007, mainly due to influenza A (H3N2) (European Influenza Surveillance Scheme, 2008). In all three countries, the incidence curves provided by Influenzanet show the same trends as the incidence curves of ECDC (Figure 2.2). For each country the shapes of the curves are similar, with peak incidence occurring in the same week and approximately equal onset and decline of ILI activity. However, incidences calculated from Influenzanet data are higher than those reported by ECDC. According to Influenzanet data, the ILI attack rate in the Netherlands was 8.7%, in Belgium 8.1%, and in Portugal 7.3%, while according to ECDC data, it was 0.8%, 3.9%, and 0.6% respectively.

## 2.4 DISCUSSION

Although there is an approximately simultaneous rise, peak and decline of ILI activity in the Influenzanet and ECDC epidemic curves, quantitatively the incidences obtained by Influenzanet are much higher than those provided by ECDC. This could be partially explained by the use of different denominators in the incidence calculations. Influenzanet participants are requested to fill in the questionnaire each week irrespective of whether they have experienced any symptoms, and the incidence of ILI is determined, considering only those participants who have filled in their symptoms questionnaire. Influenzanet is therefore independent of the rate at which people seek advice from a health

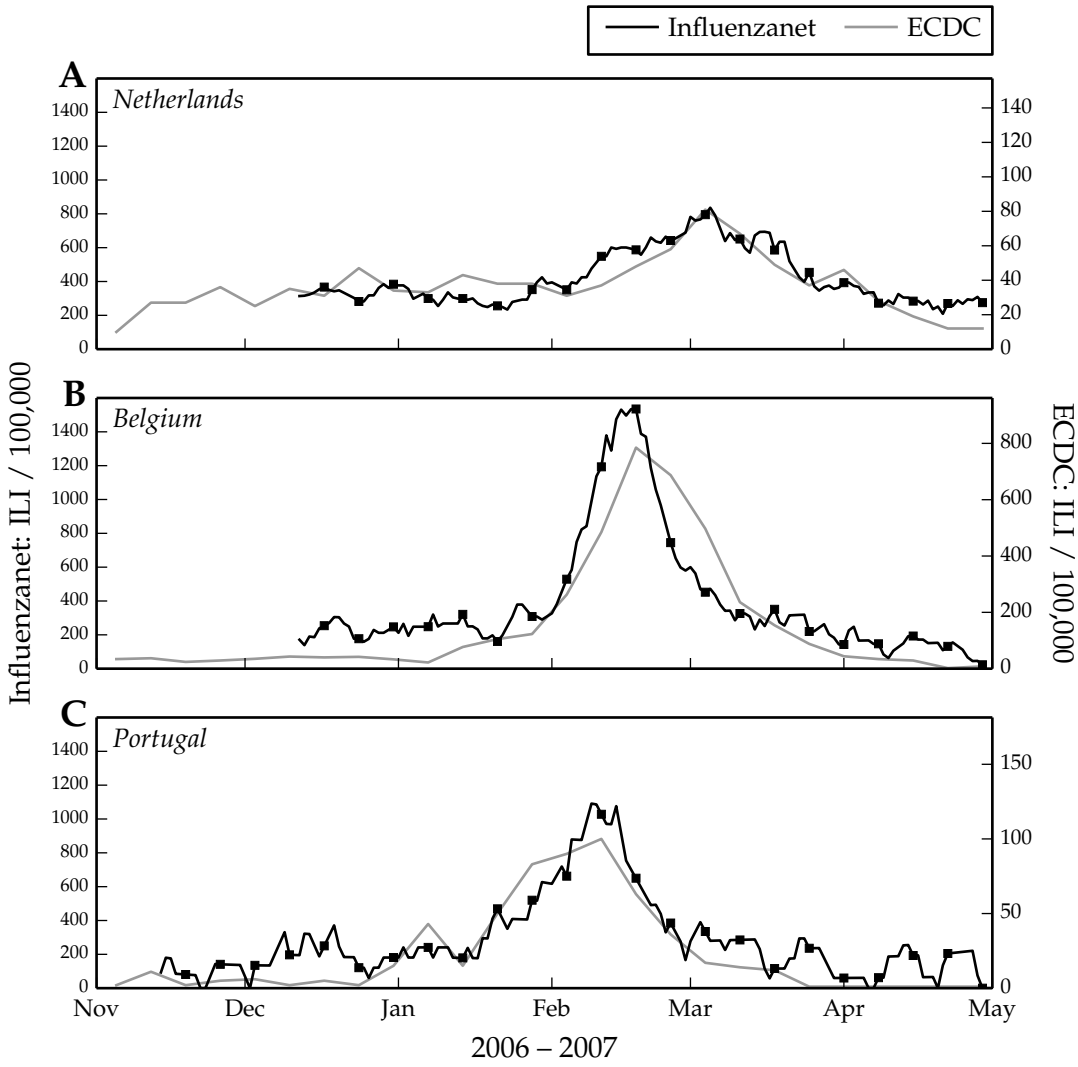



Figure 2.2: COMPARISON OF ILI INCIDENCE CURVES BETWEEN INFLUENZANET AND ECDC for A) Netherlands, B) Belgium, and C) Portugal (2006–2007). ECDC provides for each week (Monday–Sunday) the number of patients diagnosed with ILI, per 100,000 of the population under observation by the sentinel network. Influenzanet provides for each day the number of ILI onsets per 100,000 active participants (those who filled in their symptoms questionnaire for that period) in the preceding 7 days. The data points in the incidence curve of Influenzanet which monitor the same time period (Monday–Sunday) as ECDC are marked with squares. Note that Influenzanet only monitors the Northern Dutch-speaking part of Belgium (Flanders), whereas ECDC monitors the whole of Belgium.

professional. In contrast, the incidence determined by ECDC depends on the GP visiting rates which differ across countries (as can be seen also in the Influenzanet data), reflecting differences in the health care systems.

Other population-based surveillance systems for ILI have been tested (Marquet et al., 2006), but they often depend on the proportion of the people which seeks advice from a health professional when experiencing ILI symptoms. An interesting real-time monitoring system used data on symptoms reported through the NHS Direct service in the UK, a nurse-led telephone helpline for medical advice (Harcourt et al., 2001; Cooper et al., 2002). Although such systems may perform very well within one country, applying them in other countries can bring very different results. Social and cultural differences between countries may affect the tendency for people to seek advice when experiencing ILI symptoms, leading to differences in reported incidence rates.

The ILI attack rates measured by Influenzanet in winter 2006–2007 were very similar in the three countries (7.3–8.7%), while according to ECDC, the ILI attack rate in Belgium (3.8%) was five times higher than in the Netherlands (0.8%) and seven times higher than in Portugal (0.6%) (Figure 2.3). The reasons for this discrepancy could be the different case definitions used by Dutch, Belgian and Portuguese sentinel GPs (Table 2.3), the different GP visiting rates per country and the extent to which the population under observation is representative for the general population. ECDC is in the process of standardizing the case definitions for ILI used by the GPs in the different countries (Aguilera et al., 2003). The GP visiting rates, however, are not realistically controllable.

More data and analysis are needed to establish a baseline for Influenzanet. According to the Influenzanet data collected in 2006–2007 (Figure 2.2), the ILI

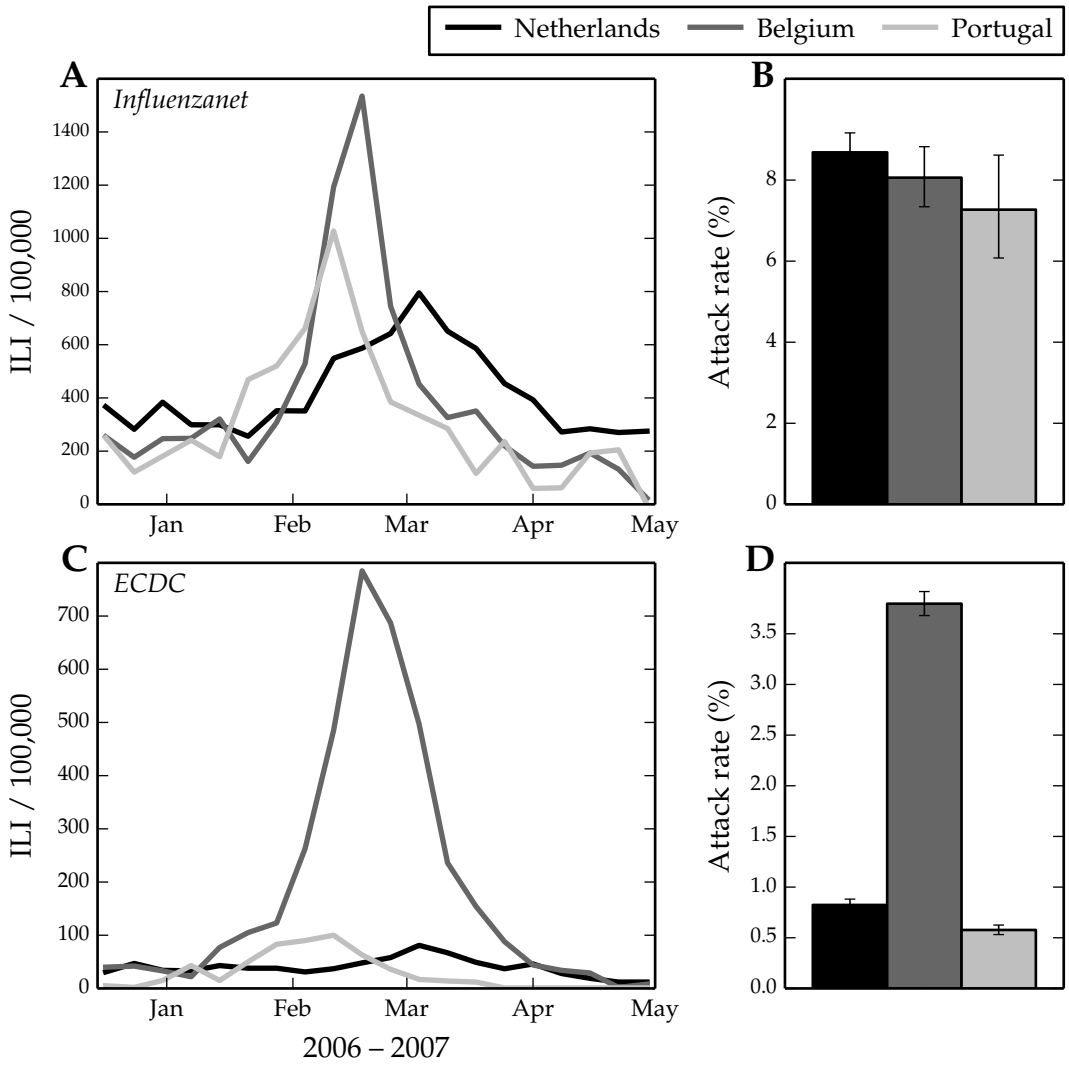



Figure 2.3: COMPARISON OF ILI ACTIVITY BETWEEN THE NETHERLANDS, BELGIUM, AND PORTUGAL for A,B) Influenzanet and C,D) ECDC (2006–2007). The ILI incidence by Influenzanet is defined as the number of ILI onsets per 100,000 participant-weeks. The peak of the ILI activity was first reached in Portugal, closely followed by Belgium and then in the Netherlands. According to Influenzanet data the height of the peak in ILI incidence in the Netherlands is slightly lower than in Belgium and Portugal, but the activity lasted longer. Therefore, ILI attack rates in the three countries are similar according to Influenzanet data (B), whereas according to ECDC data there is great variation (D).

incidence outside the epidemic peak in the Netherlands (~300 per 100,000) is different from the rates in Portugal and Belgium (~100–200 per 100,000). However, outside the influenza seasons, virology tests only rarely confirm influenza cases, rendering a system based on symptoms to low specificity. Hence the reported differences are not necessarily related to differences in influenza attack rates.

Although Influenzanet aims to attract a representative sample of the population, people who do not experience any ILI symptoms may not consider themselves suitable for participation. The ILI incidence based on just the first completed symptoms questionnaire of every participant, which contains data on symptoms a participant had upon registration, is significantly higher than the ILI incidence based on subsequent reports (Figure 2.4). To remove this selection bias, participants are only active from the day of registration onward, and the first symptoms questionnaire concerning the week before registration is thus not included in the analysis.

The non-representative nature of the Internet-using population results in a selection bias that is generally a major concern in web-based surveys (Marquet et al., 2006). Based on the data supplied by all participants upon application, the representativeness of the Influenzanet sample can be determined. (Marquet et al., 2006) showed that the demographic and health characteristics of the Influenzanet participants in 2003–2004 in the Netherlands were remarkably similar to those in the general Dutch population. Similar results were obtained for the Belgian population (Vandendijck et al., 2013). There is evidence, however, that the younger and older age groups are similarly underrepresented in all three countries (Figure 2.1).

Influenzanet seeks to monitor the representativeness of the participants in all three countries, and direct recruitment aims at targeting the underrepre-

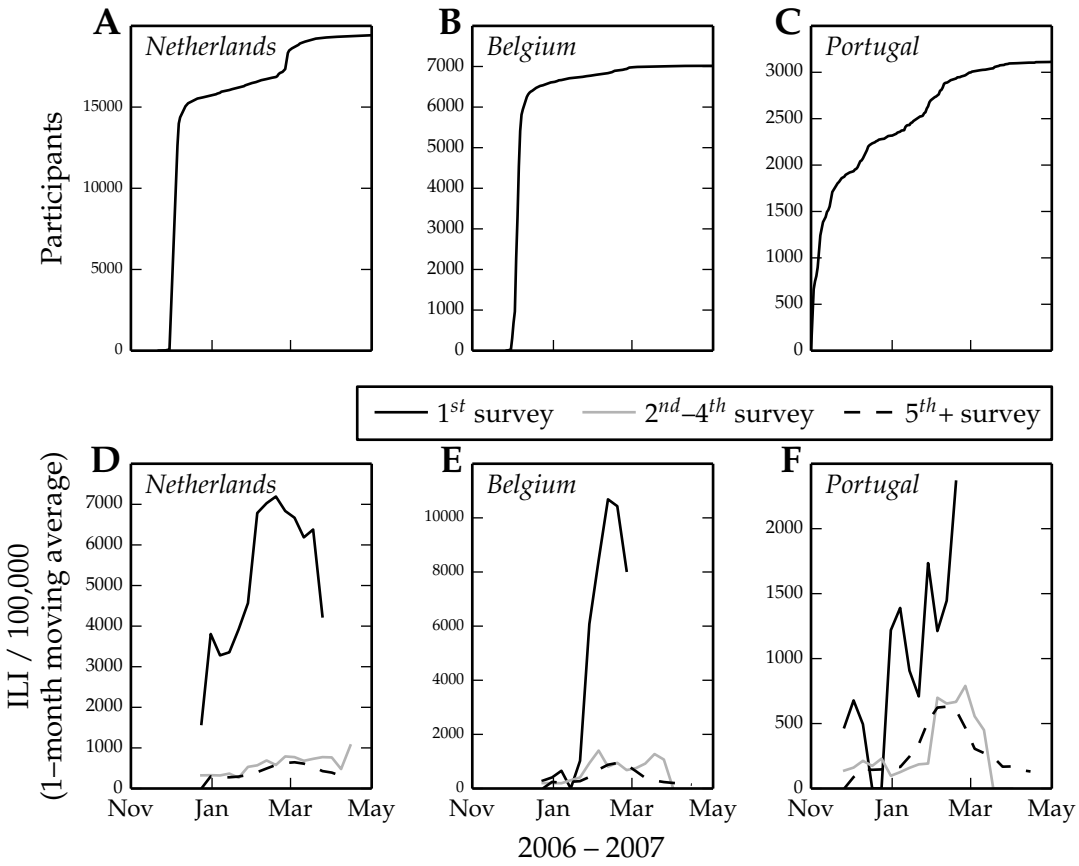



Figure 2.4: PARTICIPANT RECRUITMENT AND ILI INCIDENCE BASED ON SURVEY NUMBER in A,D) Netherlands, B,E) Belgium, and C,F) Portugal (2006–2007). A-C) The total number of registered participats during the season. Note that most participants join at the beginning of the season and during the onset of the epidemic. D-F) ILI incidence based on the survey number. To ensure that difference are not due to any bias in participants who only complete a few surveys, only participants who completed at least 8 surveys during the season are included.

sented sections of the population. The number of participants is critical for Influenzanet's success. A survey performed among 4,500 Dutch participants of Influenzanet at the end of the first season showed that most of them were recruited via radio or television (47%), via newspaper (21%) and via internet sites (16%)

Since the Influenzanet data are collected and analyzed in one place, results can be published in real time, whereas ECDC reporting each Thursday the ILI incidence for the previous Monday–Sunday period is four days behind. Influenzanet also has the capability to publish a daily incidence rate. However, participants with ILI might fill in their symptoms questionnaire earlier than participants without symptoms, which would lead to an overestimation of ILI incidence rates for the most recent days in real-time monitoring. The ILI incidence becomes increasingly more reliable as time passes and all participants complete their symptoms questionnaires. The advantage of Influenzanet, however, lies not only in its potential for an earlier assessment of weekly ILI incidences, but also in the possibility of observing the daily fluctuations in real time, thus allowing to detect early warning signals. Similar time advantages could be achieved by ECDC as well, if all GPs reported electronically in real time, as has been demonstrated by pilot projects in France and Germany (Lange and Schöttler, 2002; Carrat et al., 1998). The Influenzanet system gathers a variety of valuable data on ILI activity, however, only a fraction of these have been analyzed so far. For example, Influenzanet has the potential of monitoring the geographical spread of ILI, using the postal codes of the participants. Demographic data can also be used to monitor ILI activity in different subgroups of the population. Comparing participants with different behaviors could give indications on risk factors. Detecting an earlier rise

of ILI activity in certain subgroups could make Influenzanet an even faster early-warning system.

The uniformity of this monitoring system makes it possible to compare ILI activity between countries without further data standardization. This has important practical implications for studies concerning the global spread of ILI activity. Current efforts are directed at recruiting more European countries to join the Influenzanet. The strength of Influenzanet lies in the unique central control of every element of the monitoring system: the recruitment of participants, the questionnaires, the case definitions, the analysis of the data and the presentation of results. This makes the system not only efficient but also very flexible. If desired, any specific component can be altered without disturbing the overall system functionality. For example, the case definition can at any moment, even retrospectively, be adapted to include demographic variables. Further advantages and disadvantages of Influenzanet and ECDC are listed in Table 2.4. The two systems can complement each other to provide a better understanding of ILI activity in Europe.

## 2.5 CONCLUSION

Based solely on voluntary online reports from participants in the Netherlands, Belgium, and Portugal, Influenzanet detected an approximately simultaneous rise, peak and decline of the ILI activity as compared to ECDC during the 2006–2007 influenza season. In contrast to ECDC, however, Influenzanet uses a uniform monitoring system, allowing the direct comparison of ILI activity between countries, potentially offering a platform to monitor the geographical spread of ILI throughout Europe. We believe that the established system of ECDC, which is validated by laboratory results, could be complemented

Table 2.4: ADVANTAGES AND DISADVANTAGES OF INFLUENZANET AND ECDC

| | |
|---|---|
| ECDC: | + Established system |
| | + Combine clinical and virological data in the same population |
| | - Participating countries using different case definitions |
| | - Dependent on the GP visiting rate |
| Influenzanet: | + Uniform method, allows for direct comparison of ILI rates across countries |
| | + Flexible |
| | + Real-time monitoring |
| | + Channel to participants for influenza-related information |
| | - Self-selection bias of participants |
| | - Dependent of continuous participation of volunteers |

by the fast and flexible system of Influenzanet. Furthermore, Influenzanet could provide an important channel for influenza awareness and education in Europe. Our current strategy is to extend Influenzanet to include more European countries, thus increasing the value of its results and its impact. Those interested in implementing Influenzanet are encouraged to contact the corresponding author.

## ACKNOWLEDGMENTS

This research was funded by the Gulbenkian Foundation (FCG), the Portuguese Research Council (FCT) and the European Commission (grant MEXT-CT-2004-14338). We thank the European Influenza Surveillance Scheme, the Portuguese Influenzanet team and the Grote Griepmeting team for providing us with the ILI incidence data.

### AUTHOR CONTRIBUTIONS

Sander van Noort performed the data analyses and wrote the paper. Carl Koppeschaar was the initiator of the Influenzanet system. José Lourenço programmed the Portuguese website gripenet.pt. Helena Rebelo de Andrade wrote the part of the paper concerning the ECDC system. Marion Mühlen revised the paper. Gabriela Gomes contributed to the data analyses and revised the paper.

### SIGNIFICANT CHANGES COMPARED TO PUBLISHED ARTICLE

The applied ILI case definition has been updated to be consistent throughout the thesis which presents updated numbers and figures. The surveillance systems are always referred to as Influenzanet and ECDC. The determined attack rate is visualized in Figure 2.3. The bias in the symptoms reported during the first survey is visualized in Figure 2.4.

## 2.6 REFERENCES

Aguilera JF, Paget WJ, Mosnier A, Heijnen ML, Uphoff H, et al. (2003). Heterogeneous case definitions used for the surveillance of influenza in europe. *Eur J Epidemiol*, 18(8):751–754.

Carrat F, Flahault A, Boussard E, Farran N, Dangoumau L, et al. (1998). Surveillance of influenza-like illness in france. the example of the 1995/1996 epidemic. *J Epidemiol Community Health*, 52 Suppl 1:32S–38S.

Cooper DL, Smith GE, Hollyoak VA, Joseph CA, Johnson L, et al. (2002). Use of nhs direct calls for surveillance of influenza–a second year's experience. *Commun Dis Public Health*, 5(2):127–131.

EUROPEAN INFLUENZA SURVEILLANCE SCHEME (2007). *Annual report: 2005-2006 influenza season*. NIVEL, Utrecht, the Netherlands.

EUROPEAN INFLUENZA SURVEILLANCE SCHEME (2008). *Annual report: 2006-2007 influenza season*. NIVEL, Utrecht, the Netherlands.

HARCOURT SE, SMITH GE, HOLLYOAK V, JOSEPH CA, CHALONER R, ET AL. (2001). Can calls to nhs direct be used for syndromic surveillance? *Commun Dis Public Health*, 4(3):178–182.

LANGE W and SCHÖTTLER M (2002). Real-time influenza surveillance in germany–results of a pilot project. *Med Microbiol Immunol*, 191(3-4):139–144.

MARQUET RL, BARTELDS AIM, VAN NOORT SP, KOPPESCHAAR CE, PAGET J, ET AL. (2006). Internet-based monitoring of influenza-like illness (ili) in the general population of the netherlands during the 2003-2004 influenza season. *BMC Public Health*, 6:242.

VANDENDIJCK Y, FAES C, and HENS N (2013). Eight years of the great influenza survey to monitor influenza-like illness in flanders. *PLoS One*, 8(5):e64156.

# 3

# TEN-YEAR PERFORMANCE OF INFLUENZANET: ILI TIME SERIES, RISKS, AND VACCINE EFFECTS IN THE GROTE GRIEPMETING, GRIPENET, AND INFLUWEB COHORTS

Sander P. van Noort[1], Cláudia T. Codeço[2], Carl E. Koppeschaar[3], Marc van Ranst[4], Daniela Paolotti[5], M. Gabriela M. Gomes[1]

[1]Instituto Gulbenkian de Ciência, Oeiras, Portugal

[2]Oswaldo Cruz Foundation, Rio de Janeiro, Brazil

[3]De Grote Griepmeting, Amsterdam, Netherlands

[4] University of Leuven, Leuven, Belgium

[5] ISI Foundation, Turin, Italy

*Epidemics (minor revision)*

ABSTRACT

*Recent public health threats have propelled major innovations on infectious disease monitoring, culminating in the development of innovative syndromic surveillance methods. Influenzanet is an internet-based system that monitors influenza-like illness (ILI) in cohorts of self-reporting volunteers in European countries since 2003. We investigate and confirm coherence through the first ten years in comparison with ILI data from the European Influenza Surveillance Network and demonstrate country-specific behavior of participants with ILI regarding medical care seeking. Using regression analysis, we determine that chronic diseases, being a child, living with children, being female, smoking and having pets, are all independent predictors of ILI risk, whereas practicing sports and walking and bicycling for locomotion are associated with a small risk reduction. No effect for using public transportation or living alone was found. Furthermore, we determine the vaccine effectiveness for ILI for each season.*

## 3.1 INTRODUCTION

Recent concerns with emerging infectious diseases have exposed deficiencies in disease surveillance systems and impelled radical rethinking on how to monitor population health and detect anomalies in real time (BUTLER, 2006). In this context, new approaches in syndromic surveillance — the collection and interpretation of data for public health before laboratory or clinical confirmation is available (LAZARUS ET AL., 2001; MANDL ET AL., 2004) — have emerged. Several systems are in evaluation, showing a large diversity of data sources and methodologies employed, such as telephone-based health information services (COOPER ET AL., 2008), automated medical records (LAZARUS ET AL., 2001; VAN DEN WIJNGAARD ET AL., 2008), pharmacy sales and absenteeism (CHRETIEN ET AL., 2008), queries to online search engines (GINSBERG ET AL., 2009), and telephone-based self-reporting in cohorts of randomly selected participants (MERK ET AL., 2013). Syndromic surveillance is complementary to traditional public health surveillance in disease reporting (HENNING, 2004; LIPSITCH ET AL., 2009).

Influenzanet is a monitoring system for influenza-like illness (ILI) in voluntary cohorts of internet users. It was initially conceived to make scientific information accessible to a broad public and to kindle students' enthusiasm for science, and was launched in the Netherlands and Belgium (degrotegriepmeting.nl) in 2003 and in Portugal (gripenet.pt) in 2005. The system was consecutively implemented in Italy (influweb.it) in 2008, United Kingdom (flusurvey.org.uk) in 2009, Sweden (halsorapport.se) in 2011, France (grippenet.fr) and Spain (gripenet.es) in 2012, and Denmark (influmeter.dk) and Ireland (flusurvey.ie) in 2013. Similar systems have been implemented outside Europe, most no-

tably in Australia (flutracking.net) in 2007, Mexico (reporta.c3.org.mx) in 2009, and the United States (flunearyou.com) in 2011.

Based on single-season analysis, previous studies established good correlations between ILI incidences as determined by Influenzanet and by the clinical surveillance by sentinel General Practitioners (GPs) as coordinated by the European Centre for Disease Prevention and Control (ECDC) (Friesema et al., 2009; Marquet et al., 2006; van Noort et al., 2007; Paolotti et al., 2014; Vandendijck et al., 2013). The absolute ILI incidence as reported by Influenzanet is, however, much more consistent across countries according to Influenzanet than reported by the ECDC, due to country specific medical care seeking rates and disparities in ILI case definitions used by GPs in different countries (van Noort et al., 2007). This uniformity in rates reported across European countries facilitates the geographical analysis and modeling of epidemics (van Noort et al., 2012). By integrating serological data sources, ILI rates reported by Influenzanet have been converted to estimates of influenza attack rates (Patterson-Lomba et al., 2014).

Here we aim to further establish the Influenzanet system as a valid sentinel for ILI surveillance, by confirming that both the timing and relative intensities of epidemics are consistent with those reported by ECDC, and that the identified risks factors for ILI are consistent with those in published literature. The analysis is based on data collected over the first 10 seasons (2003–2013) from the countries in which Influenzanet was implemented for at least 5 seasons: Netherlands, Belgium, Portugal, and Italy. Time series analyses are applied to compare ILI incidences from Influenzanet and ECDC, whereas regression analysis is used to determine individual risk factors based on personal characteristics and vaccination status. Furthermore, based on the health seeking

behavior as reported to Influenzanet, differences in ILI incidence by Influenzanet and ECDC are explained.

## 3.2 METHOD

DATA COLLECTION Influenzanet participants are recruited from the general population by completing an intake questionnaire on one of the national websites, containing various demographic and life style questions. During the influenza season, participants receive a weekly newsletter by email in which they are directed to an online questionnaire about a number of symptoms that they might have experienced since their last report. The ethics committee of Instituto Gulbenkian de Ciência approved the study.

PARTICIPANTS An active population of participants is essential for the consistency of the system. An important cornerstone for success is the feedback of information to keep the participants involved and motivated. The websites contain a wealth of information on influenza, ILI and common cold, while the educational and scientific aims of the project are explained in direct mailings to schools, in repeated interviews on television and radio, and in newspapers. Schools are provided with educational material on influenza to promote incorporation of ideas of disease surveillance in science classes. At the beginning of each season, all participants from previous seasons are sent an email inviting them to participate again by completing an intake questionnaire for the new season. Based on a unique user id, participants can be tracked over multiple seasons.

BIAS Public health statistics such as asthma, diabetes and influenza vaccination rates in the Influenzanet participants have been shown to be similar for the Dutch (MARQUET ET AL., 2006) and Belgian (VANDENDIJCK ET AL., 2013) populations. Although younger and older age groups are underrepresented in Influenzanet, these differences did not seem to have an impact on the observed ILI trends (VAN NOORT ET AL., 2007). To minimize the selection bias in recruiting participants who already have ILI, any symptoms that started before or on the registration date are excluded from the analysis. Only participants who participate at least 3 times during a season are included in the analyses.

ILI INCIDENCE ILI is defined as the acute onset (within a few hours) of fever (a measured temperature $\geq$38 °C), together with muscle pain or headache, and cough or sore throat. The day of fever onset determines the day of ILI onset. Participants are considered active between registration date and the date of their last completed symptoms questionnaire. ILI incidence is determined by dividing the number of ILI onsets per week by the number of active participants. If participants fit the ILI case definition in consecutive questionnaires, this is considered as a single ILI episode.

EUROPEAN INFLUENZA SURVEILLANCE NETWORK The clinical surveillance of influenza in the European Influenza Surveillance Network (EISN, formally EISS), coordinated by the ECDC, is generally based on reports made by sentinel GPs. The ILI incidence for each country is determined by the number of patients who visit their (sentinel) GP and fit the (country-specific) ILI case definition (AGUILERA ET AL., 2003), divided by the total number of people assigned to the participating GPs.

CROSS-CORRELATION For each country, the crosscorrelation between ILI incidence rates as reported by ECDC and Influenzanet is determined. Since both time series are autocorrelated and share a common seasonal trend, this direct crosscorrelation could give a misleading indication of their relationship (BLOOM ET AL., 2007). Therefore, both time series are also prewhitened by fitting seasonal autoregressive integrated moving average (ARIMA) models using the Box-Jenkins approach (ALLARD, 1998), where the model with the lowest Akaike information criterion is selected (HYNDMAN AND ATHANASOPOULOS, 2014). The detrended time series are obtained by filtering each time series by the selected model, and for each country the crosscorrelation between the detrended time series from Influenzanet and ECDC is determined.

Since the ILI incidence from Influenzanet is based on the reported day of onset and the ILI incidence from ECDC is determined by the week a patient visited their GP, it would be expected that the reported ILI incidence from Influenzanet precedes the ILI incidence as determined by ECDC. Since in Influenzanet not only the week of onset but the actual day is recorded, the weekly ILI incidence from Influenzanet can be shifted by single days, where a shift of zero days indicates that the ILI incidence from both systems is compared for the period Monday–Sunday.

MEDICAL CARE SEEKING BEHAVIOR Each participant who reports (ILI) symptoms, is asked some follow-up questions, such as whether the participant visited a medical doctor. This allows the determination of the percentage of participants with ILI who seek medical care. Since participants could seek medical care after they have reported their symptoms to Influenzanet, a reported visit within 15 days after a reported ILI onset is still considered. Since season 2011–2012, Influenzanet participants who reported to have visited a

medical doctor are also asked how many days elapsed between the onset of symptoms and the visit.

RISK FACTOR ANALYSIS We use regression analysis to explore the association between several individual covariates and the occurrence of at least one ILI episode during a season. These covariates are selected beforehand, consisting mostly of characteristics which have been identified in other studies regarding influenza risk, and some extra characteristics which are not normally analyzed. Most covariates are considered equal across all seasons: age group (<15, 15–49, 50–64, 65+), household situation (alone, with children or with only adults), gender, chronic disease (asthma, diabetes, heart disease, and/or immunocompromised), smoking, sports (at least 1 hour per week), possession of pets (dogs, cats, and/or birds), and primary mode of daily locomotion (bicycle / foot, car, or public transport). The covariate "risk group (others)" includes those participants who report to belong to a risk group, but are younger than 65 years (60 years in the Netherlands since 2008) and did not report any of the chronic diseases. The effect of vaccination is considered as a season-dependent covariate. For the season 2009–2010, the vaccine status is based on the pandemic vaccine. Country of residence and season (indirectly) are two extra covariates.

Only ILI onsets during the weeks when influenza strains were circulating in the population are considered. These periods are defined for each season and country as the weeks when the number of influenza-confirmed samples as reported by ECDC was at least 15% of the maximum for that season (moving average over 3 weeks) (Figure B.13). All participants are considered independent between two different seasons, and participants who were not active for the complete influenza period are excluded. Since Influenzanet is a cohort

study in which healthy individuals (without ILI) are recruited and the possible onset of ILI is monitored over a fixed period of time, Influenzanet can determine the risk ratios for all the covariates. For each covariate a univariate risk ratio is determined. Adjusted risk ratios are determined by a multivariate log-binomial regression model including all global and season-dependent covariates, analyzed by the general linear model in R software (R DEVELOPMENT CORE TEAM, 2014). The variance inflationary factor (VIF) for each covariate is determined to check for collinearity in the multivariate regression model.

## 3.3 RESULTS

ILI INCIDENCE The ILI incidence as determined by Influenzanet correlates well over multiple seasons with the ILI incidence as reported by ECDC (Figure 3.1). However, Influenzanet measures ILI incidence in all countries on the same scale, while the incidences reported by ECDC are in general lower and vary in scale between countries.

The crosscorrelation between the raw ILI incidences from Influenzanet and ECDC is significant (Figure 3.2A,C,E,G). The ILI incidences show a high level of autocorrelation and some degree of seasonality (Figures B.5–B.12A, Appendix B, page 211). We fitted to each time series a seasonal ARIMA model (Table B.1, Appendix B, page 206), and filtered the time series by each model to obtain a detrended time series (Figures B.1–B.4, Appendix B, page 207). The detrended time series are no longer autocorrelated (Figures B.5–B.12B, Appendix B, page 211), as confirmed by the Ljung-Box test (Table B.1, Appendix B, page 206), The detrended time series of Influenzanet and ECDC also show a significant level of crosscorrelation at a lag of zero weeks (Fig-

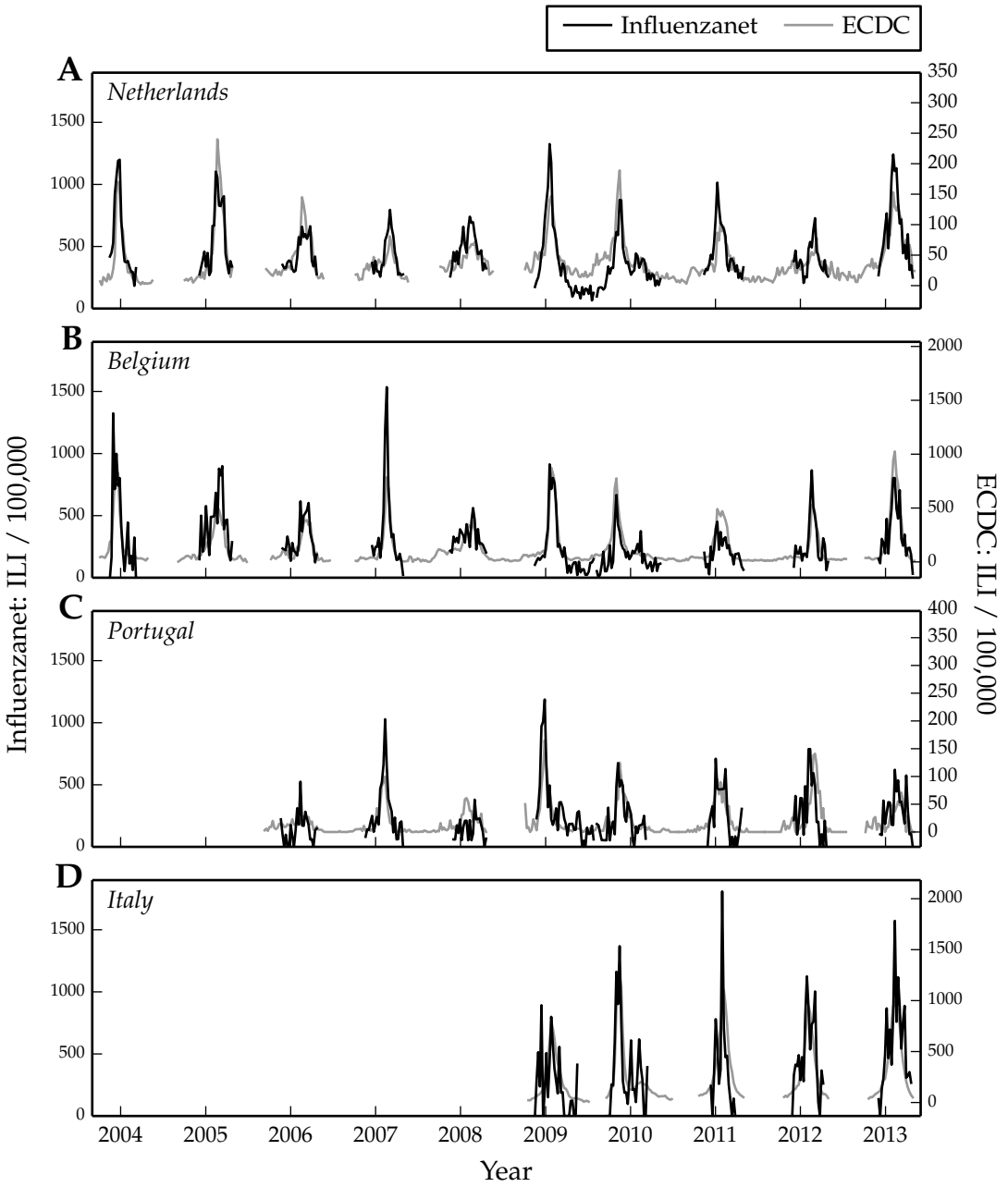



Figure 3.1: ILI INCIDENCE FOR INFLUENZANET AND ECDC in A) Netherlands, B) Belgium, C) Portugal, and D) Italy (2003–2013). The ILI incidence by Influenzanet is defined as the number of ILI onsets per 100,000 participant-weeks. The vertical axes for the ILI incidence are scaled based on a linear regression between the ILI incidences of Influenzanet and ECDC.

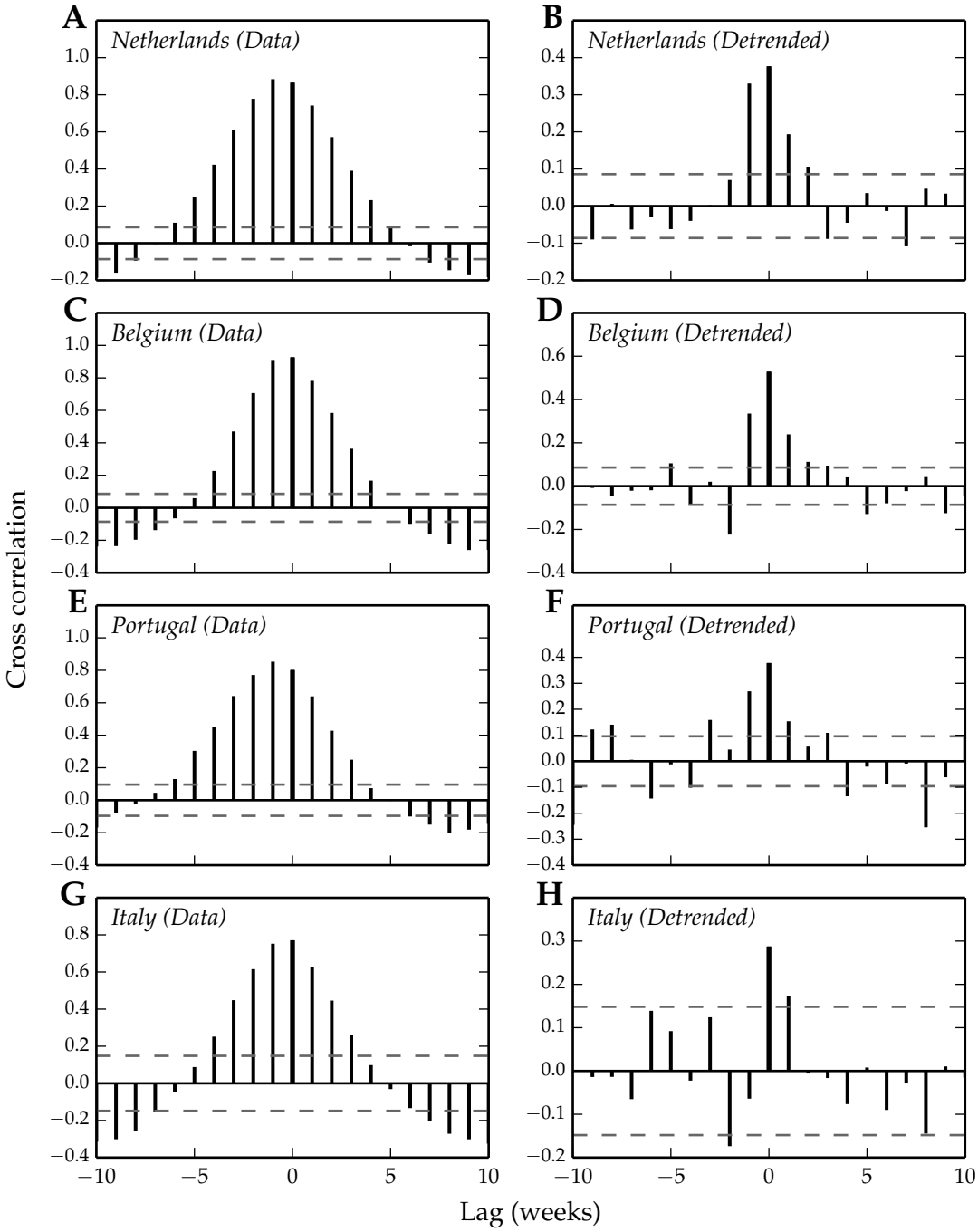



Figure 3.2: CROSSCORRELATION BETWEEN THE ILI INCIDENCES OF INFLUENZANET AND ECDC using either the raw ILI incidences (left panel) or the detrended time series (right panel) in A,B) Netherlands, C,D) Belgium, E,F) Portugal, and G,H) Italy (2003–2013).

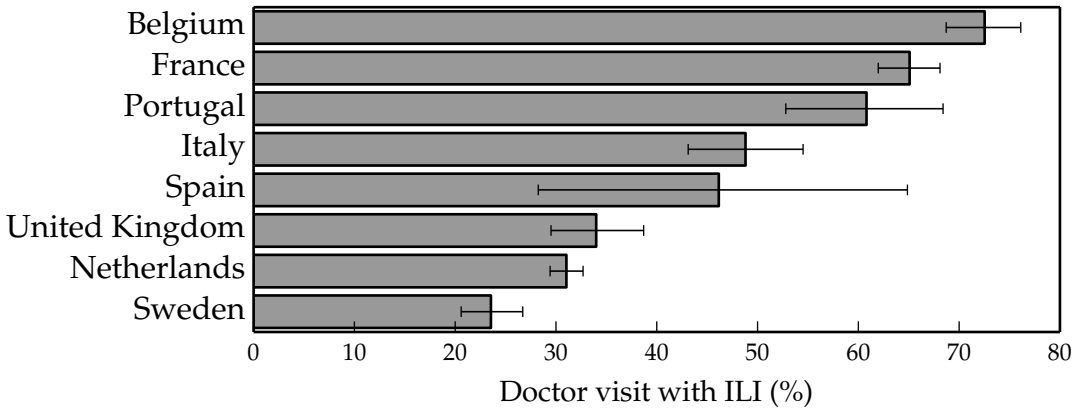



Figure 3.3: INFLUENZANET PARTICIPANTS WITH ILI WHO VISITED A MEDICAL DOCTOR specified by country (2011–2013).

ure 3.2B,D,F,H), for the Netherlands (0.38), Belgium (0.53), Portugal (0.38), and Italy (0.29).

MEDICAL CARE SEEKING BEHAVIOR The percentage of participants with ILI who sought medical care varies greatly by country (Figure 3.3). Similar differences are observed in the number of days between the onset of symptoms and visiting the doctor (Figure 3.4). The observed patterns do not change if only working adult participants are considered in the analysis.

The crosscorrelation between the detrended time series from Influenzanet and ECDC is maximum when a shift of 4 days is applied in the Netherlands, 1–2 days in Belgium, 1 day in Portugal, and no shift in Italy (Table 3.1). This corresponds well with the median delay between ILI onset and seeking medical care as reported during the seasons 2011–2013: 4 days in the Netherlands, 2 days in Belgium, 1 day in Portugal and 1 day in Italy.

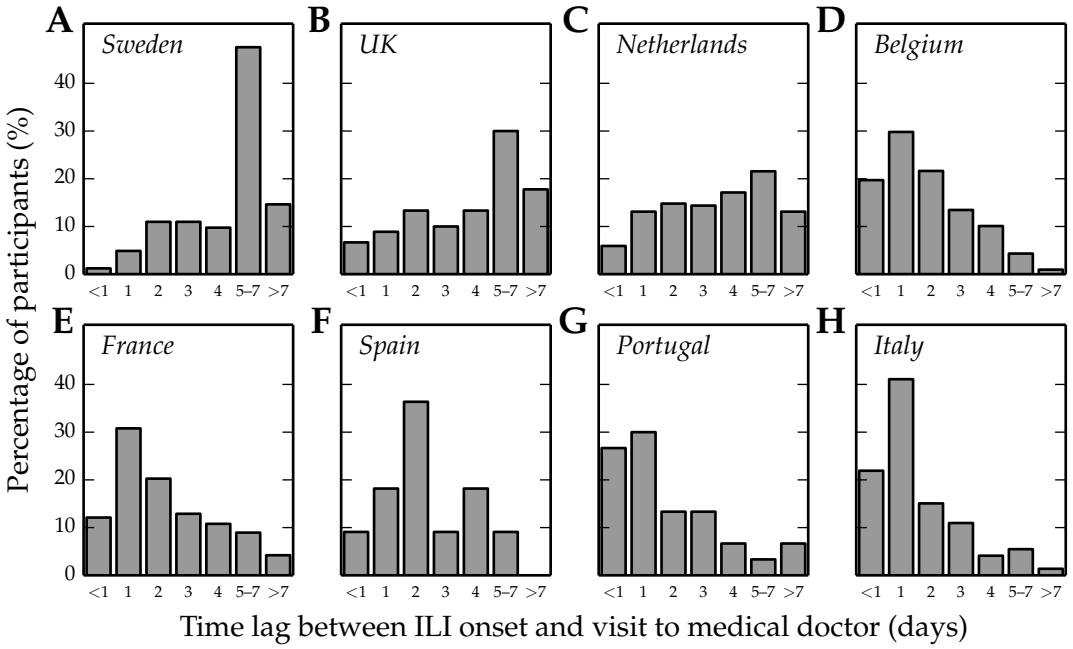



Figure 3.4: TIME LAG BETWEEN ILI ONSET AND VISIT TO MEDICAL DOCTOR in A) Sweden, B) United Kingdom, C) Netherlands, D) Belgium, E) France, F) Spain, G) Portugal, and H) Italy (2011–2013).

Table 3.1: CROSSCORRELATION BETWEEN INFLUENZANET AND ECDC FOR DAILY LAGS based on the detrended time series (2003–2013). Values in bold indicate the crosscorrelation at a lag equal to the median number of days between ILI onset and visit to a medical doctor.

| LAG | NETHERLANDS | BELGIUM | PORTUGAL | ITALY |
|---|---|---|---|---|
| –6 | 0.40 | 0.42 | 0.23 | –0.01 |
| –5 | 0.40 | 0.45 | 0.32 | 0.09 |
| –4 | **0.47** | 0.52 | 0.32 | 0.19 |
| –3 | 0.46 | 0.50 | 0.35 | 0.24 |
| –2 | 0.39 | **0.57** | 0.35 | 0.22 |
| –1 | 0.32 | 0.57 | **0.42** | **0.24** |
| 0 | 0.38 | 0.53 | 0.38 | 0.29 |
| 1 | 0.34 | 0.46 | 0.22 | 0.24 |
| 2 | 0.35 | 0.44 | 0.22 | 0.20 |
| 3 | 0.34 | 0.37 | 0.24 | 0.18 |
| 4 | 0.31 | 0.40 | 0.13 | 0.16 |
| 5 | 0.35 | 0.32 | 0.17 | 0.17 |
| 6 | 0.34 | 0.24 | 0.06 | 0.16 |
| 7 | 0.19 | 0.24 | 0.15 | 0.18 |

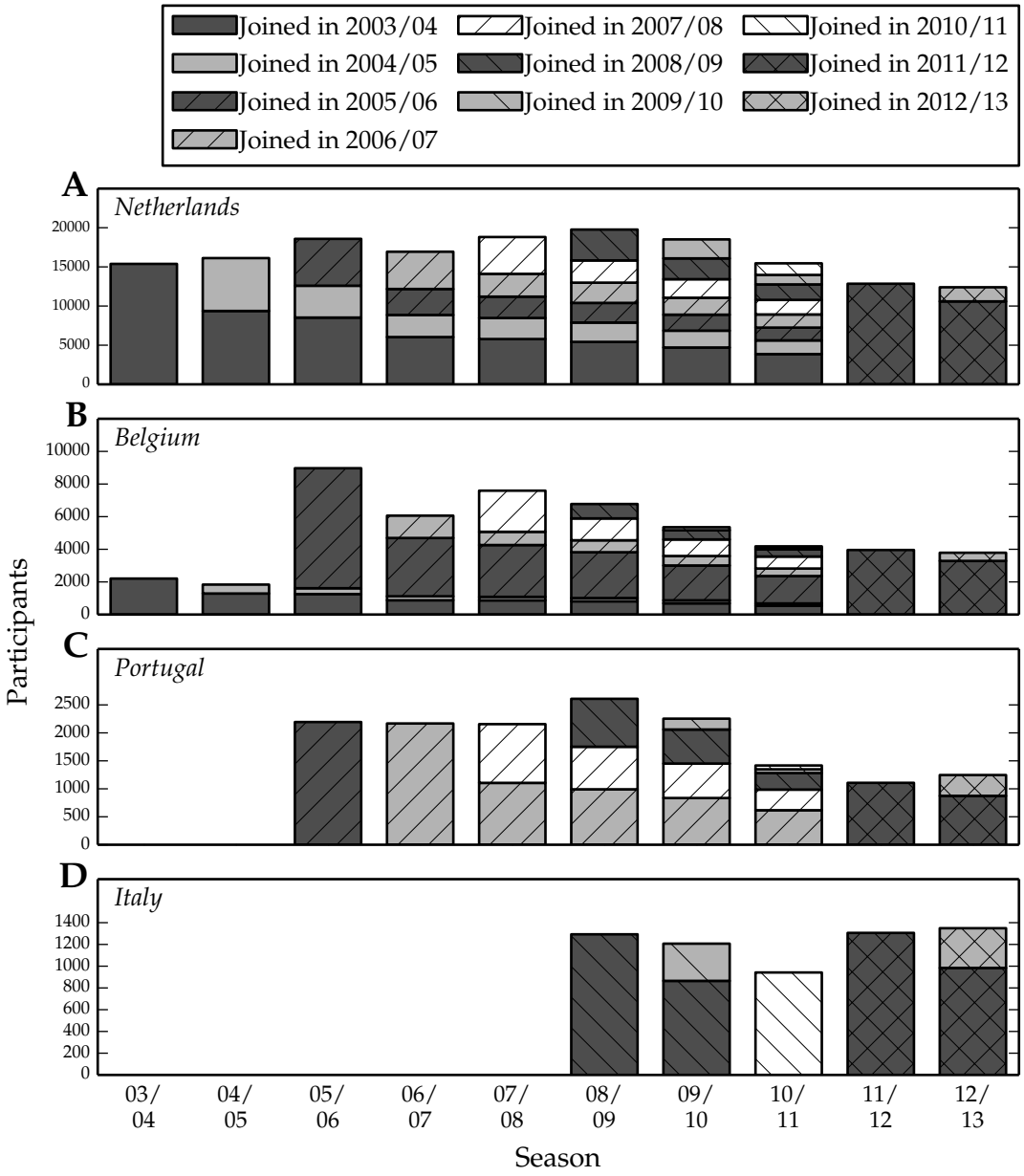



Figure 3.5: NUMBER OF INFLUENZANET PARTICIPANTS who completed at least 3 symptoms questionnaires in A) Netherlands, B) Belgium, C) Portugal, and D) Italy (2003–2013). The participants are differentiated by the season in which they registered. In Portugal in season 2006–2007, in Italy in season 2010–2011, and in all countries in season 2011–2012 a new database structure was introduced, providing a reset of participant ids.

PARTICIPATION The Netherlands has most participants (on average 19491 per season, of which 16481 completed at least 3 symptoms questionnaires), followed by Belgium (6001; 5072), Portugal (2871; 1894), and Italy (1882; 1219) (Figure 3.5). This corresponds to 0.1% of the population in the Netherlands and Belgium (Flanders, since basically all participants are from Flanders), 0.02% in Portugal, and 0.003% in Italy. Of all participants who completed at least 3 symptoms questionnaires during a season, in the Netherlands $76 \pm 8\%$ participated again in the following season, $74 \pm 12\%$ in Belgium, $69 \pm 12\%$ in Portugal, and $70 \pm 4\%$ in Italy.

RISK FACTORS The univariate risk ratios are listed in Table B.2, Appendix B, page 217, whereas the adjusted risk ratios from the multivariate regression are listed in Table 3.2. The variance inflationary factor (VIF) for the covariates varies between 1.3 and 2.8 in the multivariate regression (Table B.2, Appendix B, page 217), reassuring that model specification is not compromised by undesirable collinearities (O'Brien, 2007).

According to the adjusted risk ratios, having a chronic disease (asthma, diabetes, heart disease and/or immunocompromised), living with children, being female, belonging to a younger age group, having pets (cats and/or dogs), and being a smoker, were all independent predictors of the risk of having at least one ILI episode during a flu season. A small risk reduction was observed in participants who primarily use bike or foot for locomotion (compared to a car) and participants who practice more than 1 hour of sports per week. No significant effect was observed for participants who live with other adults (compared to living alone), participants who have birds at home, and participants who use public transportation (compared to using a car).

Table 3.2: RISK FACTORS AND VACCINE EFFECTIVENESS FOR ILI in the Netherlands, Belgium, Portugal, and Italy (2003–2013).

| QUESTION | ANSWER | RR (ADJUSTED) |
|---|---|---|
| Age | <18 | 1.59 (1.46 – 1.74) |
| | 18–49 | * |
| | 50–64 | 0.82 (0.78 – 0.86) |
| | 65+ | 0.46 (0.41 – 0.51) |
| Household | Alone | * |
| | Only with adults | 0.99 (0.92 – 1.05) |
| | With children | 1.31 (1.22 – 1.40) |
| Gender | Male | * |
| | Female | 1.22 (1.17 – 1.28) |
| Chronic diseases | Asthma or lung disease | 1.58 (1.47 – 1.69) |
| | Diabetes | 1.27 (1.15 – 1.41) |
| | Heart disease | 1.29 (1.13 – 1.47) |
| | Kidney disorder | 1.23 (0.80 – 1.90) |
| | Immunodeficiency | 1.23 (1.02 – 1.49) |
| Unknown risk group[a] | | 1.23 (1.06 – 1.41) |
| Smoking | | 1.16 (1.10 – 1.22) |
| Pets at home | Dogs | 1.15 (1.09 – 1.22) |
| | Cats | 1.07 (1.02 – 1.12) |
| | Birds | 1.03 (0.94 – 1.13) |
| Sports | $\geq$ 1 hour per week | 0.95 (0.90 – 1.00) |
| Daily means of locomotion | Bicycle / Foot | 0.95 (0.90 – 1.00) |
| | Car | * |
| | Public Transport | 0.97 (0.89 – 1.05) |
| Vaccination | 2003–2004 | 1.07 (0.78 – 1.47) |
| | 2004–2005 | 1.10 (0.94 – 1.28) |
| | 2005–2006 | 0.97 (0.83 – 1.12) |
| | 2006–2007 | 1.00 (0.86 – 1.15) |
| | 2007–2008 | 0.81 (0.70 – 0.94) |
| | 2008–2009 | 0.80 (0.71 – 0.90) |
| | 2009–2010[b] | 0.87 (0.74 – 1.03) |
| | 2010–2011 | 0.67 (0.58 – 0.78) |
| | 2011–2012 | 0.97 (0.82 – 1.16) |
| | 2012–2013 | 0.80 (0.71 – 0.91) |

[a] Classified by GP as risk group due to factors not specified on this table
[b] Vaccination for the new H1N1pdm strain

The vaccine effectiveness for influenza-like illness varies from season to season. A significant reduction in ILI due to vaccination was observed in the seasons 2007–2008, 2008–2009, 2010–2011, and 2012–2013, while no significant effect was observed in other seasons. The vaccine effectiveness for season 2009–2010 is possibly underestimated, since the vaccine only became available when the ILI activity was already epidemic.

## 3.4 DISCUSSION

Based on single-season analysis, previous studies established excellent correlations between ILI incidences as determined by Influenzanet and by ECDC (Friesema et al., 2009; Marquet et al., 2006; van Noort et al., 2007). A question remained on whether this consistency would persist for multiple-season data streams. We showed that during 10 seasons in the Netherlands and Belgium (2003–2013), 8 seasons in Portugal (2005–2013), and 5 seasons in Italy (2008–2013), the ILI trends from Influenzanet and ECDC are consistent in both timing and relative magnitude, with a significant crosscorrelation between both time series as lags of zero weeks. The signal from Influenzanet precedes ECDC by a few days, corresponding approximately to the median number of days between ILI onset and seeking medical care. However, this does not necessary indicate that in real-time monitoring Influenzanet would detect ILI trends earlier, since this depends on when the data becomes available and the statistical uncertainties in the data (Section 6.3.3)

Although both time series are correlated over the full 10-year period, there are localized discrepancies between the data streams, which could be attributed to the different methodology and composition of the cohorts in both systems. As an example, young children are largely underrepresented in In-

fluenzanet (VAN NOORT ET AL., 2007), whereas young children visit relatively more often a medical doctor. This could explain why for the season 2007–2008 in Portugal, dominated by the influenza B strain affecting mostly children, a small epidemic was reported by ECDC which went mostly undetected by Influenzanet. Another local discrepancy is the relatively high ILI incidence as reported by ECDC in the Netherlands during the months preceding the 2009 ILI pandemic, which might be attributed to an increase in both awareness by medical doctors and patients due to a global concern about the new H1N1 influenza strain (KERAMAROU ET AL., 2011).

The presence of multiple independent sources encourages the development of integrative methods that explore the specific strengths of each system (REIS ET AL., 2007). Having multiple independent systems could uncover aspects of influenza transmission that would go unnoticed if only one data stream was available. Another cross-country data source for ILI incidences is Google Flu Trends (GINSBERG ET AL., 2009), which determines ILI incidence based on the frequency of ILI-related search terms. However, Google Flu Trends is not a strictly independent data source, since their algorithms rely on the ECDC data streams for calibration.

Patterns in medical care seeking behavior suggest cultural difference between northern and southern Europe. In southern Europe (France, Italy, Portugal, and Spain) participants generally visit a medical doctor within 1–2 days after the onset of symptoms, whereas in northern Europe (Sweden, United Kingdom, and the Netherlands) participants seek medical care generally only 5–7 days after the onset of symptoms. Belgium (Flanders) seems an exception to the suggested pattern, most likely because according to Belgium law, an employer can require from their employee a medical statement within 24 hours to justify work absenteeism. A similar pattern is observed in the per-

centage of participants with ILI who seek medical care, which is lower in the northern Europe (except Belgium) than in southern Europe. The two patterns could be associated by considering that in countries where participants wait longer before seeking medical care, many participants would no longer feel sufficiently ill to warrant a visit to a medical doctor.

This variation in medical care seeking rates across countries is one of the reasons why ILI incidences reported by ECDC cannot be compared directly (VAN NOORT ET AL., 2007). Variations in medical care seeking could also affect the determined ILI incidence by ECDC within a country, if certain subgroups of the population visit a doctor at different rates. Influenzanet does not only serve as an independent source for ILI activity, but could also be used to calibrate ILI data as collected by GP sentinel systems.

A crucial element in the success of Influenzanet, is having a sufficiently large cohort of participants. In the Netherlands (on average 16481 active participants), Belgium (5072), Portugal (1894), and Italy (1219) the Influenzanet cohort was large enough to detect similar ILI epidemics as ECDC in all seasons, with the exception of season 2007–2008 in Portugal. Larger cohorts would lead to lower statistical noise such that epidemics could be detected earlier and even small ILI epidemics could be distinguished from baseline ILI activity. Furthermore, in larger cohorts, different subgroups, for example based on age or vaccine status, could be monitored separately. Of all active participants during a certain season, $73 \pm 11\%$ participated again in the following season. Although this shows impressive loyalty of participants, each season an effort should be made to recruit new participants to at least replace those who have left.

Risk factors estimated from the Influenzanet cohort are consistent with the influenza literature. Higher risk of ILI in children and in those living with

children was observed, in consistency with observational studies (CAUCHEMEZ ET AL., 2009; MONTO AND ROSS, 1977; MONTO, 2004; VIBOUD ET AL., 2004). The increased ILI risk in women compared to men, which may be due to more intensive contact between women and children, has also been previously recognized (MONTO AND ROSS, 1977). We found a significantly reduced risk of ILI among participants over 65. This is not due the higher vaccine uptake in seniors, since vaccine status is already included as a separate covariate in this multivariate analyses. Seniors are generally considered a risk group for influenza, not because of a higher probability for infection, but due to their greater risk for complications (MONTO, 2004). Having a chronic disease, such as asthma, diabetes, heart disease or immunocompromised, was a strong predictor of ILI in the Influenzanet cohort. People with these chronic diseases are generally advised to take an influenza vaccine. Increased risk of influenza has been observed in children with asthma in clinical cohort studies (GORDON ET AL., 2009), while diabetes is known to be strongly associated with complications due to influenza infections (IRWIN ET AL., 2001). An increased risk of ILI was observed among the Influenzanet participants who smoke, as has been confirmed by other studies (ARCAVI AND BENOWITZ, 2004).

The Influenzanet system is flexible to the extent that questions of interest can easily be added or removed, allowing for the estimation of risk factors which are not usually considered. In this study, we found a small but significant protective effect of walking or bicycling as a primary means of locomotion in comparison with traveling by car, while no significant risk of traveling by public transportation was observed, nor in participants who live with other adults in comparison with adults who live alone. A small increase in risk was observed in participants who have pets at home. Practicing sports for at least

one hour per week was associated with a small but significant decrease on the ILI risk.

Not only extra questions could be included in the intake questionnaire, entire new questionnaires could be added in particular seasons enabling further studies. A stress-related questionnaire released in the Netherlands in season 2004–2005 revealed significant trends between stress/personality and ILI self-reporting (Smolderen et al., 2007), and a simple questionnaire related to contact behavior, showed that changes in contact patterns could explain changes in disease incidence (Eames et al., 2012).

In only 4 out of 10 seasons Influenzanet estimated a significant reduction in ILI due to vaccination, and the direct effectiveness of vaccination varied between a significant 33% in season 2010–2011 and a non-significant -10% in season 2004–2005. A relatively low vaccine effectiveness against ILI is to be expected, since vaccination targets specifically the influenza virus, and not other influenza-like illnesses. A double-blind, randomized, placebo-controlled trial measured within the same cohort a vaccine efficacy for serologically confirmed influenza of respectively 50% (1997–1998) and 86% (1998–1999), but a vaccine effectiveness for ILI of -10% (1997–1998) and 33% (1998–1999) (Bridges et al., 2000). According to a large meta-study based on 48 reports on vaccine effectiveness in healthy adults, inactivated parenteral vaccines were 30% effective against ILI, and 80% efficacious against influenza when the vaccine matched the circulating strain and circulation was high, but this decreased to an effectiveness against ILI of 12% and efficacy against influenza of 50% when it did not (Demicheli et al., 2009).

For two seasons (2003–2004 and 2004–2005) Influenzanet estimated a negative although non-significant vaccine effectiveness. Both seasons were characterized by a poor vaccine match (Jin et al., 2005; Belongia et al., 2009). A

negative vaccine effect can be due to original antigenic sin, the tendency for antibodies produced in response to exposure to influenza vaccine antigens to suppress the maturation of antibodies with high affinity to the actual virus (GUPTA ET AL., 2006).

In an observational study of vaccine effectiveness, any preexisting bias between vaccinated and unvaccinated participants could distort the results. The univariate risk ratios (Table B.2, Appendix B, page 217) indicate on average a 10% higher ILI reduction in vaccinated participants than the adjusted risk ratios (Table 3.2). Participants over 65 years of age have a lower ILI rate and a relatively high vaccination rate, and the multivariate model estimates that a part of the reduction in ILI in vaccinated participants is due to their age. Although the multivariate regression analysis aims to correct for these biases, it is possible that other biases not represented by any of the risk factors listed in Table 3.2 exist.

A cohort study of 72,527 seniors over 65 years of age followed during an 8 year period, found that vaccinated seniors already had a reduced risk of death and pneumonia hospitalization in the periods before the influenza season, and that the risk reduction actually decreased during the influenza season (JACKSON ET AL., 2006). Such a preferential receipt of vaccine by relatively healthy seniors could lead to overestimation of the vaccine effectiveness in observational studies. It is plausible that most elderly Influenzanet participants are relatively healthy and that this selection bias is less present in Influenzanet, leading to relatively lower estimates of vaccine effectiveness than in the average literature. Because of global recommendations for influenza vaccination, placebo-controlled trials, which could clarify the effects of influenza vaccines in individuals, are no longer considered possible on ethical grounds (JEFFERSON ET AL., 2010).

Since the Influenzanet participants are not a random sample of the overall population, care should be taken in extrapolating the estimated risks to the overall population in the respective countries. However, the observed consistency in risk factors for ILI between Influenzanet and those reported by studies in community settings further establishes that the Influenzanet population is a valuable sentinel for ILI surveillance in the population, in addition to the merits of engaging the participants in public health research and promoting risk awareness.

The system presented here stands on a concept for syndromic surveillance that depends on intense activity in science communication, public awareness and sufficient levels of Internet penetration. It has reported ILI activity in a consistent way for over 10 seasons in multiple countries. Influenzanet reports ILI trends consistent with GP sentinel surveillance (ECDC), and can complement these systems by providing valuable information about medical care seeking behavior. Based on reported symptoms, Influenzanet can be extended to detect diseases other than influenza, including those in developing settings. Influenzanet as an Internet monitoring system based on voluntary participants might therefore develop into an important weapon to fight influenza as well as other contagious diseases globally.

## ACKNOWLEDGMENTS

We thank the national Influenzanet teams who provided data for this study. This research was funded by the European Commission (EC-ICT-231807). MVR acknowledges the support of the Institute for the Promotion of Innovation by Science and Technology (IWT) in Flanders (strategic basic research project SIMID). MGMG thanks ICyT projects 60/2010 and 343/2010. CTC was sup-

ported by a fellowship from the Brazilian Research Council (CNPq) during this study.

### AUTHOR CONTRIBUTIONS

Sander van Noort performed the data analyses and wrote the paper. Cláudia Codeço contributed to the data analyses and revised the paper. Carl Koppeschaar was the initiator of the Influenzanet system. Marc van Ranst is responsible for the Belgian Influenzanet, and Daniela Paolotti for the Italian Influenzanet. Gabriela Gomes is responsible for the Portuguese Influenzanet, contributed to the data analyses, and revised the paper.

## 3.5 REFERENCES

Aguilera JF, Paget WJ, Mosnier A, Heijnen ML, Uphoff H, et al. (2003). Heterogeneous case definitions used for the surveillance of influenza in europe. *Eur J Epidemiol*, 18(8):751–754.

Allard R (1998). Use of time-series analysis in infectious disease surveillance. *Bull World Health Organ*, 76(4):327–333.

Arcavi L and Benowitz NL (2004). Cigarette smoking and infection. *Arch Intern Med*, 164(20):2206–2216.

Belongia EA, Kieke BA, Donahue JG, Greenlee RT, Balish A, et al. (2009). Effectiveness of inactivated influenza vaccines varied substantially with antigenic match from the 2004-2005 season to the 2006-2007 season. *J Infect Dis*, 199(2):159–167.

Bloom RM, Buckeridge DL, and Cheng KE (2007). Finding leading indicators for disease outbreaks: filtering, cross-correlation, and caveats. *J Am Med Inform Assoc*, 14(1):76–85.

Bridges CB, Thompson WW, Meltzer MI, Reeve GR, Talamonti WJ, et al. (2000). Effectiveness and cost-benefit of influenza vaccination of healthy working adults: A randomized controlled trial. *JAMA*, 284(13):1655–1663.

Butler D (2006). Disease surveillance needs a revolution. *Nature*, 440(7080):6–7.

Cauchemez S, Donnelly CA, Reed C, Ghani AC, Fraser C, et al. (2009). Household transmission of 2009 pandemic influenza a (h1n1) virus in the united states. *N Engl J Med*, 361(27):2619–2627.

Chretien JP, Burkom HS, Sedyaningsih ER, Larasati RP, Lescano AG, et al. (2008). Syndromic surveillance: adapting innovations to developing settings. *PLoS Med*, 5(3):e72.

Cooper DL, Smith GE, Regan M, Large S, and Groenewegen PP (2008). Tracking the spatial diffusion of influenza and norovirus using telehealth data: a spatiotemporal analysis of syndromic data. *BMC Med*, 6:16.

Demicheli V, Jefferson T, Al-Ansary LA, Ferroni E, Rivetti A, et al. (2009). Vaccines for preventing influenza in healthy adults. *Cochrane Database Syst Rev*, 3.

Eames KTD, Tilston NL, Brooks-Pollock E, and Edmunds WJ (2012). Measured dynamic social contact patterns explain the spread of h1n1v influenza. *PLoS Comput Biol*, 8(3):e1002425.

Friesema IHM, Koppeschaar CE, Donker GA, Dijkstra F, van Noort SP, et al. (2009). Internet-based monitoring of influenza-like illness in the general population: experience of five influenza seasons in the netherlands. *Vaccine*, 27(45):6353–6357.

Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, et al. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014.

Gordon A, Ortega O, Kuan G, Reingold A, Saborio S, et al. (2009). Prevalence and seasonality of influenza-like illness in children, nicaragua, 2005-2007. *Emerg Infect Dis*, 15(3):408–414.

Gupta V, Earl DJ, and Deem MW (2006). Quantifying influenza vaccine efficacy and antigenic distance. *Vaccine*, 24(18):3881–3888.

Henning KJ (2004). What is syndromic surveillance? *MMWR Morb Mortal Wkly Rep*, 53 Suppl:5–11.

Hyndman R and Athanasopoulos G (2014). Forecasting: principles and practice.

Irwin DE, Weatherby LB, Huang WY, Rosenberg DM, Cook SF, et al. (2001). Impact of patient characteristics on the risk of influenza/ili-related complications. *BMC Health Serv Res*, 1(1):8.

Jackson LA, Jackson ML, Nelson JC, Neuzil KM, and Weiss NS (2006). Evidence of bias in estimates of influenza vaccine effectiveness in seniors. *Int J Epidemiol*, 35(2):337–344.

Jefferson T, Di Pietrantonj C, Al-Ansary LA, Ferroni E, Thorning S, et al. (2010). Vaccines for preventing influenza in the elderly. *Cochrane Database Syst Rev*, (2):CD004876.

Jin H, Zhou H, Liu H, Chan W, Adhikary L, et al. (2005). Two residues in the hemagglutinin of a/fujian/411/02-like influenza viruses are responsible for antigenic drift from a/panama/2007/99. *Virology*, 336(1):113–119.

Keramarou M, Cottrell S, Evans MR, Moore C, Stiff RE, et al. (2011). Two waves of pandemic influenza a(h1n1) 2009 in wales–the possible impact of media coverage on consultation rates, april-december 2009. *Euro Surveill*, 16(3).

Lazarus R, Kleinman KP, Dashevsky I, DeMaria A, and Platt R (2001). Using automated medical records for rapid identification of illness syndromes (syndromic surveillance): the example of lower respiratory infection. *BMC Public Health*, 1:9.

Lipsitch M, Hayden FG, Cowling BJ, and Leung GM (2009). How to maintain surveillance for novel influenza a h1n1 when there are too many cases to count. *Lancet*, 374(9696):1209–1211.

Mandl KD, Overhage JM, Wagner MM, Lober WB, Sebastiani P, et al. (2004). Implementing syndromic surveillance: a practical guide informed by the early experience. *J Am Med Inform Assoc*, 11(2):141–150.

Marquet RL, Bartelds AIM, van Noort SP, Koppeschaar CE, Paget J, et al. (2006). Internet-based monitoring of influenza-like illness (ili) in the general population of the netherlands during the 2003-2004 influenza season. *BMC Public Health*, 6:242.

Merk H, Kühlmann-Berenzon S, Bexelius C, Sandin S, Litton JE, et al. (2013). The validity of self-initiated, event-driven infectious disease reporting in general population cohorts. *PLoS One*, 8(4):e61644.

Monto AS (2004). Occurrence of respiratory virus: time, place and person. *Pediatr Infect Dis J*, 23(1 Suppl):S58–S64.

Monto AS and Ross H (1977). Acute respiratory illness in the community: effect of family composition, smoking, and chronic symptoms. *Br J Prev Soc Med*, 31(2):101–108.

O'Brien RM (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, 41(5):673–690. ISSN 0033-5177.

Paolotti D, Carnahan A, Colizza V, Eames K, Edmunds J, et al. (2014). Web-based participatory surveillance of infectious diseases: the influenzanet participatory surveillance experience. *Clin Microbiol Infect*, 20(1):17–21.

PATTERSON-LOMBA O, VAN NOORT SP, COWLING BJ, WALLINGA J, GOMES MGM, ET AL. (2014). Utilizing syndromic surveillance data for estimating levels of influenza circulation. *Am J Epidemiol*, 179lh(11):1394–1401.

R DEVELOPMENT CORE TEAM (2014). *R: A language and environment for statistical computing*. R foundation for statistical computing, Vienna, Austria.

REIS BY, KOHANE IS, and MANDL KD (2007). An epidemiological network model for disease outbreak detection. *PLoS Med*, 4(6):e210.

SMOLDEREN KGE, VINGERHOETS AJJM, CROON MA, and DENOLLET J (2007). Personality, psychological stress, and self-reported influenza symptomatology. *BMC Public Health*, 7:339.

VAN NOORT SP, MUEHLEN M, REBELO DE ANDRADE H, KOPPESCHAAR CE, LOURENÇO JML, ET AL. (2007). Gripenet: an internet-based system to monitor influenza-like illness uniformly across europe. *Euro Surveill*, 12(7):E5–E6.

VAN NOORT SP, ÁGUAS R, BALLESTEROS S, and GOMES MGM (2012). The role of weather on the relation between influenza and influenza-like illness. *J Theor Biol*, 298:131–137.

VANDENDIJCK Y, FAES C, and HENS N (2013). Eight years of the great influenza survey to monitor influenza-like illness in flanders. *PLoS One*, 8(5):e64156.

VIBOUD C, BOËLLE PY, CAUCHEMEZ S, LAVENU A, VALLERON AJ, ET AL. (2004). Risk factors of influenza transmission in households. *Br J Gen Pract*, 54(506):684–689.

VAN DEN WIJNGAARD C, VAN ASTEN L, VAN PELT W, NAGELKERKE NJD, VERHEIJ R, ET AL. (2008). Validation of syndromic surveillance for respiratory pathogen activity. *Emerg Infect Dis*, 14(6):917–925.

# 4

# THE ROLE OF WEATHER ON THE RELATION BETWEEN INFLUENZA AND INFLUENZA-LIKE ILLNESS

Sander P. van Noort[1], Sébastien Ballesteros[2], Ricardo Águas[1], M. Gabriela M. Gomes[1]

[1]Instituto Gulbenkian de Ciência, Oeiras, Portugal

[2] Universidade de Lisboa, Lisbon, Portugal

*Journal of Theoretical Biology (2012) 298: 131–137*

ABSTRACT

*Influenza epidemics, enabled by viral antigenic drift, occur invariably each winter in temperate climates. However, attempts to correlate the magnitude of virus change and epidemic size have been unsatisfactory. The incidence of influenza is not typically measured directly, but rather derived from the incidence of influenza-like illness (ILI), a clinical syndrome. Weather factors have been shown to influence the manifestation of influenza-like symptoms.*

*We fitted an influenza transmission model to time series of influenza-like illness as monitored from 2003 to 2013 by two independent symptomatic surveillance systems (Influenzanet and ECDC) in three European countries. By assuming that seasonality only acts upon the manifestation of symptoms, the model shows a significant correlation between the absolute humidity and temperature at the time of infection, and the proportion of influenza infections fulfilling the clinical ILI case definition, the so-called ILI factor.*

*When a weather-dependent ILI factor is included in the model, the epidemic size of influenza-like illness becomes dependent not only on the susceptibility of the population at the beginning of the epidemic season but also on the weather conditions during which the epidemic unfolds. The combination reduces season-to-season variation in epidemic size and, interestingly, leads to a non-monotonic trend whereby the largest ILI epidemic occurs for moderate initial susceptibility.*

## 4.1 INTRODUCTION

Influenza epidemics occur invariably each winter in temperate climates, exhibiting impressive resilience to demographic, social and medical change (HOPE-SIMPSON, 1981). The continuous viral evolution (SMITH ET AL., 2004) and global migration (NELSON ET AL., 2007) keeps host populations susceptible for recurring epidemics. However, attempts to directly correlate the magnitude of virus change and the epidemic size have been unsatisfactory (CHOWELL ET AL., 2008), and this could be partly attributed to interference with seasonal factors.

The seasonality of influenza is governed by numerous factors and occupies researchers from a multitude of disciplines. Weather effects on viral survival (SHAMAN AND KOHN, 2009) and host susceptibility (DOWELL, 2001), and the effect of social contacts on transmission (CAUCHEMEZ ET AL., 2008), favor the transmission of influenza during the winter months (LIPSITCH AND VIBOUD, 2009). Transmission models for influenza, which include seasonality in transmission, usually apply a seasonal sinusoidal transmission rate (KOELLE ET AL., 2006; BACAËR AND AIT DADS, 2011), or link the transmission rate directly to weather factors such as the absolute humidity (SHAMAN ET AL., 2010).

Transmission models for influenza are not typically fitted to incidence data for influenza, but rather to incidence data for influenza-like illness (ILI) or excess pneumonia and influenza (P&I) death rates (CHOWELL ET AL., 2008; SHAMAN ET AL., 2010; MILLS ET AL., 2004). Determining the presence of the influenza virus in the population is a costly process, only performed regularly on symptomatic persons to determine the circulating strains (MEERHOFF ET AL., 2004). Instead, cohort based syndromic surveillance systems determine the weekly burden of ILI based on a clinical case definition (MANDL ET AL.,

2004). Various case definitions are in use by different surveillance systems, requiring the simultaneous presence of multiple influenza-related symptoms, to exclude persons with non-influenza respiratory syndromes. A typical case definition, as applied by the Influenzanet system, requires the simultaneous occurrence of a fast rise of fever, muscle pain and at least one respiratory symptom (VAN NOORT ET AL., 2007). Recent studies have shown that experimentally infected persons do not regularly expose all symptoms (LAU ET AL., 2010), indicating that the influenza activity in the population could be significantly higher than the reported ILI incidence.

Influenza transmission models fitted to ILI incidence, typically assume a constant asymptomatic rate (FERGUSON ET AL., 2003) or constant reporting rate (CAUCHEMEZ ET AL., 2008) to transform the influenza incidence to an ILI incidence. Seasonal variation in these parameters could, however, significantly influence the expression of clinical symptoms in influenza infected persons (BAETJER, 1967; ASSAAD AND REID, 1971; ECCLES, 2002; FUHRMANN, 2010), and thus the proportion of influenza infections that fit the ILI case definition.

We will explore the effect of a seasonal ILI factor, defined as the proportion of influenza infections that fit the ILI case definition. We fit a simple transmission model to ILI incidence data corresponding to ten seasons (2003–2013) in the Netherlands, Belgium, and Portugal from two independent ILI surveillance systems: European Influenza Surveillance Scheme as coordinated by the ECDC and Influenzanet. The transmission rate is set to be constant during the winter months throughout all seasons. Although seasonality is likely to coact in the transmission rate and season-to-season variation of influenza subtypes, the model aims to solely explore the seasonal variation of the ILI factor, and link it directly to the measured weather variables in each country.

## 4.2 METHOD

### 4.2.1 *Model*

We divide the population into three subgroups: hosts with symptomatic influenza infection ($I$), hosts who are susceptible to symptomatic infection ($S$) and hosts who are immune to symptomatic infection. Infected hosts recover at a rate $\tau$, and make $\beta$ infectious contacts per week while infected. A constant low rate of infection per week, $m$, is caused by external contacts, most notably persons being infected abroad. The system is described by the following equations:

$$\begin{aligned} \frac{\mathrm{d}S}{\mathrm{d}t} &= -\lambda S \\ \frac{\mathrm{d}I}{\mathrm{d}t} &= \lambda S - \tau I \\ \lambda &= \beta I + m \end{aligned} \tag{4.1}$$

where $\lambda$ denotes the force of infection.

A proportion $p$, the ILI factor, of all symptomatic infections fits the ILI case definition, while a time-dependent but season invariant proportion $J$ of the population fits the ILI case definition without an influenza infection. The measured ILI prevalence is thus given by:

$$\mathrm{ILI}(t) = p(t)I(t) + J(t) \tag{4.2}$$

The ILI incidence is estimated by multiplying the ILI prevalence by the recovery rate $\tau$. All symptomatic infected persons are assumed to transmit equally, independently of whether they fit the ILI case definition.

### 4.2.2 *Model including asymptomatics*

If a constant proportion of all infected is asymptomatic, and possibly less infectious, the above system of equations still applies, but the interpretation of the transmission rate $\beta$ is transformed. We divide the population into four subgroups: hosts with symptomatic influenza infection ($I$), hosts with asymptomatic infection ($A$), hosts who are susceptible to either symptomatic or asymptomatic infection ($\hat{S}$) and hosts who are immune to both symptomatic and asymptomatic infection. Upon infection, a proportion $\alpha$ of the susceptibles gets a symptomatic infection ($I$), while a proportion $1-\alpha$ gets an asymptomatic infection ($A$). Both symptomatic and asymptomatic infected recover at a rate $\tau$, where symptomatic make $\hat{\beta}$ infectious contacts per week while infected, and asymptomatic make $\phi\hat{\beta}$ infectious contacts, where $0 \leq \phi \leq 1$. A constant low rate of infection per week, $m$, is caused by external contacts. This system is described by the following equations:

$$\begin{aligned}
\frac{\mathrm{d}\hat{S}}{\mathrm{d}t} &= -\lambda\hat{S} \\
\frac{\mathrm{d}I}{\mathrm{d}t} &= \alpha\lambda\hat{S} - \tau I \\
\frac{\mathrm{d}A}{\mathrm{d}t} &= (1-\alpha)\lambda\hat{S} - \tau A \\
\lambda &= \hat{\beta}I + \phi\hat{\beta}A + m
\end{aligned}$$

It follows that $I = \alpha(I + A)$, and substitution of $A = \frac{1-\alpha}{\alpha} I$ lead to the following system:

$$\begin{aligned} \frac{\mathrm{d}(\alpha\hat{S})}{\mathrm{d}t} &= -\lambda(\alpha\hat{S}) \\ \frac{\mathrm{d}I}{\mathrm{d}t} &= \lambda(\alpha\hat{S}) - \tau I \\ \lambda &= \left(1 + \phi\frac{1-\alpha}{\alpha}\right)\hat{\beta}I + m \end{aligned}$$

which is equivalent to the system (4.1), where $S = \alpha\hat{S}$ denotes hosts that are susceptible to symptomatic infection, and the transmission parameter $\beta$ is transformed by:

$$\beta = \left(1 + \phi\frac{1-\alpha}{\alpha}\right)\hat{\beta}$$

An SIAR model with transmitting asymptomatics thus has the same fundamental dynamics as the basic SIR model. The underlying parameters $\alpha$ and $\phi$ are not identifiable when only symptomatically infected hosts are observed.

### 4.2.3 *Data*

Two independent symptomatic surveillance systems, which measure the ILI activity in multiple countries in Europe, are currently active. The European Influenza Surveillance Network as coordinated by the European Centre for Disease Control and Prevention (ECDC) measures the ILI activity based on the number of ILI patients who visit a GP belonging to one of the national sentinel networks. Influenzanet measures the ILI activity based on weekly online symptoms questionnaire completed by voluntary participants. Google Flu Trends (Ginsberg et al., 2009), another cross-country monitoring system,

is not strictly independent since their algorithms rely on ECDC data streams for parameter calibration.

We apply the model to those countries and seasons for which both Influenzanet and ECDC provide ILI incidence data: the Netherlands, Belgium, and Portugal (2003–2013) from September to May. The ILI incidence data as determined by Influenzanet is uniform across countries and seasons, due to the application of a single ILI case definition in all countries, and the independence of country- and season-specific rate at which influenza-infected persons seek medical attention (VAN NOORT ET AL., 2007). We therefore scale the ILI incidence data from ECDC to the same order as the ILI incidence data from Influenzanet based on the linear regression between both data sets (VAN NOORT ET AL., 2014). The model is fitted to the average of the ILI incidence from Influenzanet and the scaled ILI incidence from ECDC. For weeks in which data from only one system is available, most notably due to the absence of Influenzanet data in Portugal 2003–2005, the ILI incidence is based on only one data source.

Weather variables for each country are based on measurements from a central location in the country: Utrecht in the Netherlands, Brussels in Belgium, and Coimbra in Portugal[1]. We assess, independently, the influence of temperatures and absolute humidity on syndromic reporting. For both factors we consider two alternative implementations: (1) integrate weekly weather measurements straight into the model; and (2) construct a periodic function based on a 40-year average (1970–2010) of the 1-month moving average of the weather factor of interest. Implementation (2) has advantages for long-term prediction.

1 http://www.noaa.gov

### 4.2.4 *Fitting*

The baseline level of non-influenza ILI activity $J(t)$ is determined by fitting a general function

$$J(t) = A \cdot \sin\left((t - H) \cdot \frac{2\pi}{365}\right) + V$$

with parameters amplitude ($A$), horizontal shift ($H$) and vertical shift ($V$), to all historic non-epidemic ILI incidences. ILI activity outside the influenza periods is considered non-epidemic, where the influenza period for each season is defined as the weeks when the number of positive samples as collected by ECDC is at least 15% of the maximum number of positive samples in that season (for more details see Section 6.3.3).

The recovery rate $\tau$ is set to $7/5$, corresponding to an average of 5 days during which influenza-infected persons excrete the virus (Lau et al., 2010). The transmission coefficient $\beta$, summarizing contact rates, viral persistence, and all other social and environmental factors that affect transmission, is assumed to be country-specific but invariant across seasons. The proportion of susceptibles at the beginning of each season $S_0$ is estimated through the fitting, as an independent variable, which captures the cumulative effect of antigenic drift and shift, births and deaths and possible vaccination campaigns between the end of the preceding influenza season and the start of the next. The start of the influenza season is set to 1 September, corresponding roughly to the start of schools after summer vacation. The migration rate $m$ is set to $10^{-7}$ per week for all countries. Simulations with a migrations factor an order of magnitude higher or lower do not significantly alter the goodness-of-fit of the models or the qualitative conclusions of the study (data not shown). All model fittings are performed by a least squares minimization fitting procedure implemented

in Python 2.7. For each fit the goodness-of-fit ($R^2$) is determined by a linear regression between the model output and the ILI incidence.

## 4.3 RESULTS

The main focus of this study is the estimation and interpretation of the proportion of influenza infections that fit the ILI case definition, the so-called ILI factor $p$. We begin by estimating directly from the data the contribution of non-influenza cases to the weekly ILI incidence, and proceed by estimating indirectly (based on a model) the ILI factor under a series of approximations of increasing refinement.

### 4.3.1 *Non-influenza ILI*

Figure 4.1 shows for each country the best fit of a baseline incidence to non-epidemic ILI activity. The seasonal variation in non-influenza ILI cases is more marked in northern countries.

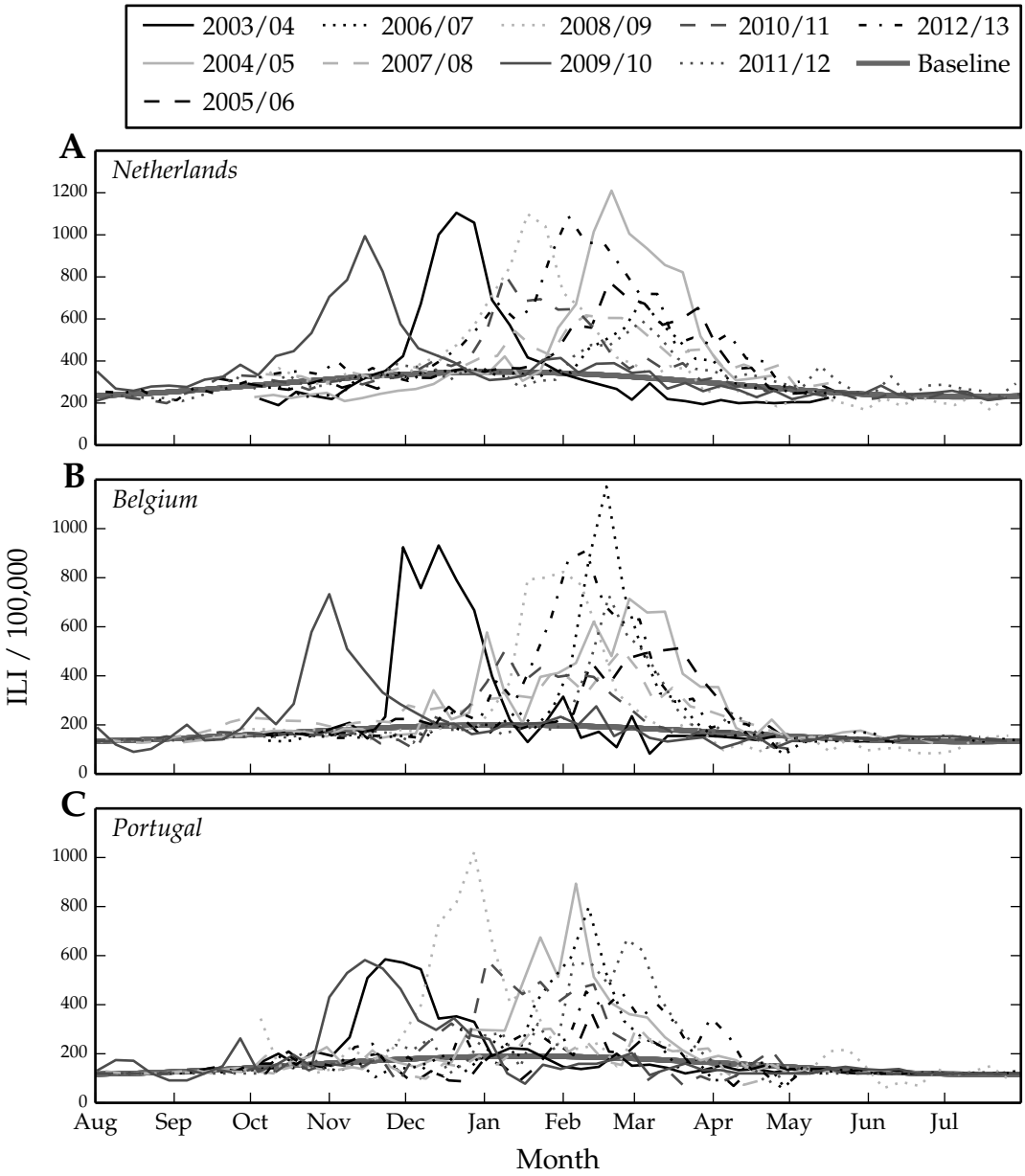



Figure 4.1: ILI INCIDENCE AND A FITTED BASELINE in A) Netherlands, B) Belgium, and C) Portugal (2003–2013). The ILI incidence by Influenzanet is defined as the number of ILI onsets per 100,000 participant-weeks. The ILI incidence by ECDC is defined as the weekly number of patient visits with ILI per 100,000. The plotted ILI incidence is based on the average incidence of both systems.

### 4.3.2 *First approach: ILI factor constant*

As a first approximation, we assume that the ILI factor $p$ is constant throughout the seasons. The system of (4.1) and (4.2) can then be written as:

$$\begin{aligned} \frac{\mathrm{d}(pS)}{\mathrm{d}t} &= -\lambda(pS) \\ \frac{\mathrm{d}(pI)}{\mathrm{d}t} &= \lambda(pS) - \tau(pI) \\ \lambda &= (\frac{\beta}{p})(pI) + m \\ \mathrm{ILI}(t) &= (pI) + J(t) \end{aligned} \tag{4.3}$$

leading to a system with variables $pS$, $pI$ and $\frac{\beta}{p}$.

For each country we estimate a constant transmission rate $\frac{\beta}{p}$ and season dependent parameters $pS_0$. Figure 4.2 shows the estimated ILI incidence curves and the estimated parameters are listed in Table 4.1 ($p$ constant). The model is able to capture the timing of the epidemics, but cannot capture the magnitude of the ILI incidence accurately, most notably due to an underestimate of the ILI incidence during the seasons 2004–2005, 2008–2009, and 2012–2013, and an overestimate during the pandemic season 2009–2010.

### 4.3.3 *Second approach: ILI factor per season*

As an intermediate step in the identification of temporal trends in the ILI factor, we assume the ILI factor $p$ to be constant during each season, but variable from season to season. Since the ILI factor only influences the observation of ILI and not the influenza dynamics itself, the estimated parameters $p$ can

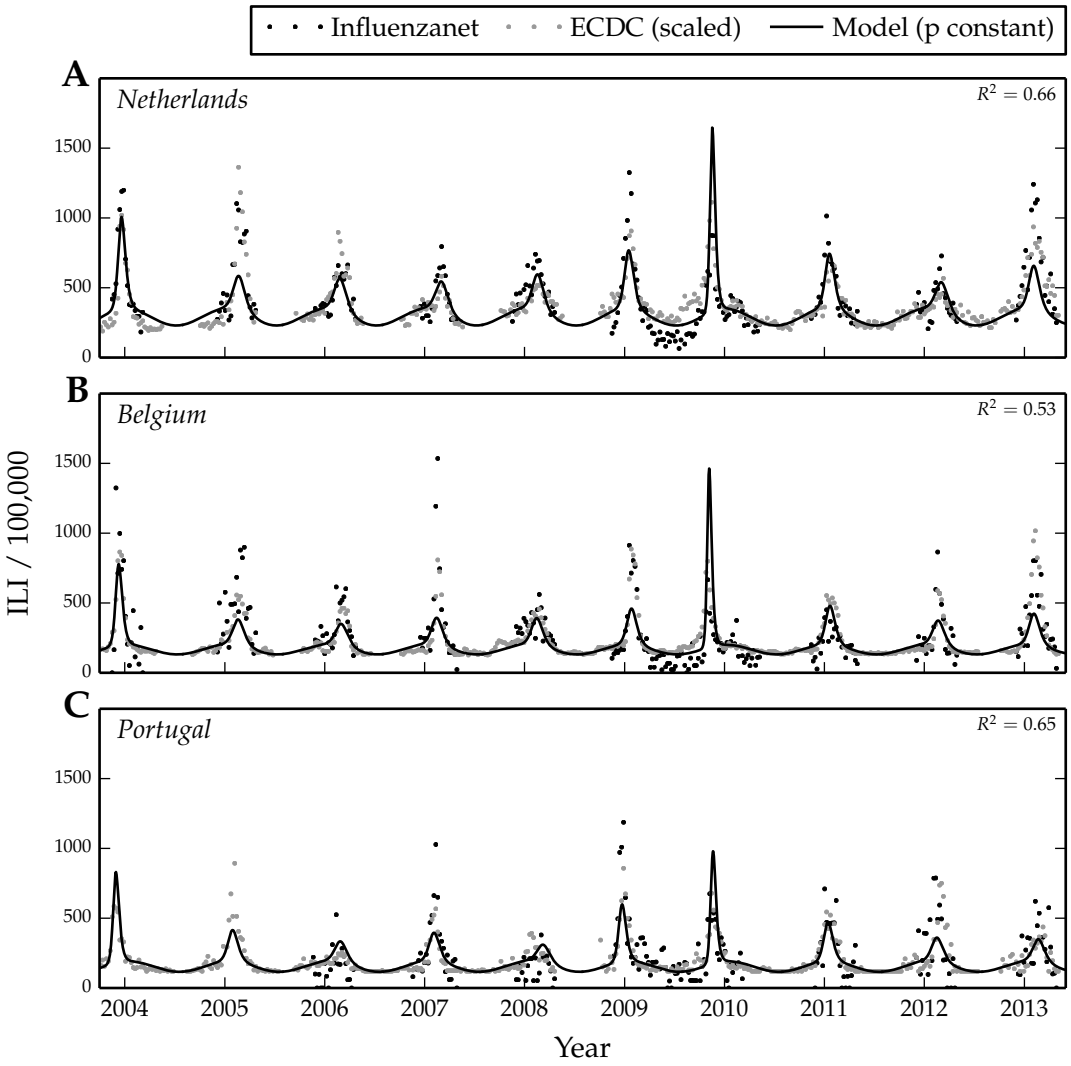



Figure 4.2: BEST FIT FOR THE REFERENCE MODEL based on the first approach: the ILI factor $p$ constant throughout all seasons in A) Netherlands, B) Belgium, and C) Portugal (2003–2013). The ILI incidence by Influenzanet is defined as the number of ILI onsets per 100,000 participant-weeks.

Table 4.1: ESTIMATED PARAMETERS FOR THE THREE MODEL APPROACHES

| | $p$ CONSTANT | $p$ PER SEASON | | $p$ BY CLIMATE | | |
|---|---|---|---|---|---|---|
| | $pS_0$ | $S_0^*$ | $p^*$ | $S_0^*$ | $p^*(W)$ | $W$ |
| NETHERLANDS | | | | | | |
| 2003/04 | 0.048 | 0.14 | 0.38 | 0.09 | 0.69 | 5.8 |
| 2004/05 | 0.040 | 0.11 | 1.00 | 0.07 | 0.98 | 5.3 |
| 2005/06 | 0.039 | 0.11 | 0.59 | 0.07 | 0.92 | 5.4 |
| 2006/07 | 0.039 | 0.11 | 0.36 | 0.07 | 0.87 | 5.4 |
| 2007/08 | 0.040 | 0.12 | 0.40 | 0.07 | 0.92 | 5.4 |
| 2008/09 | 0.043 | 0.13 | 0.55 | 0.08 | 0.88 | 5.4 |
| 2009/10 | 0.058 | 0.18 | 0.16 | 0.12 | 0.24 | 7.2 |
| 2010/11 | 0.043 | 0.12 | 0.35 | 0.08 | 0.89 | 5.4 |
| 2011/12 | 0.039 | 0.11 | 0.28 | 0.07 | 0.85 | 5.5 |
| 2012/13 | 0.041 | 0.12 | 0.84 | 0.08 | 0.99 | 5.3 |
| All seasons | $\beta/p = 50$ | $\beta^* = 17$ | | $\beta^* = 27$, $k = -0.73$, $\epsilon^* = 3.8$ | | |
| BELGIUM | | | | | | |
| 2003/04 | 0.036 | 0.10 | 0.49 | 0.08 | 0.60 | 5.9 |
| 2004/05 | 0.029 | 0.08 | 0.96 | 0.07 | 0.95 | 5.3 |
| 2005/06 | 0.028 | 0.08 | 0.67 | 0.06 | 0.87 | 5.4 |
| 2006/07 | 0.029 | 0.08 | 0.96 | 0.07 | 0.99 | 5.2 |
| 2007/08 | 0.029 | 0.08 | 0.50 | 0.06 | 0.94 | 5.3 |
| 2008/09 | 0.030 | 0.09 | 0.79 | 0.07 | 0.91 | 5.4 |
| 2009/10 | 0.046 | 0.14 | 0.14 | 0.12 | 0.17 | 7.7 |
| 2010/11 | 0.031 | 0.09 | 0.34 | 0.07 | 0.92 | 5.3 |
| 2011/12 | 0.029 | 0.08 | 0.61 | 0.07 | 0.94 | 5.3 |
| 2012/13 | 0.030 | 0.08 | 1.00 | 0.07 | 0.99 | 5.2 |
| All seasons | $\beta/p = 68$ | $\beta^* = 24$ | | $\beta^* = 30$, $k = -0.72$, $\epsilon^* = 3.8$ | | |
| PORTUGAL | | | | | | |
| 2003/04 | 0.035 | 0.08 | 0.30 | 0.06 | 0.54 | 9.4 |
| 2004/05 | 0.027 | 0.06 | 0.94 | 0.04 | 0.95 | 8.2 |
| 2005/06 | 0.026 | 0.06 | 0.13 | 0.04 | 0.90 | 8.4 |
| 2006/07 | 0.027 | 0.06 | 0.82 | 0.04 | 0.94 | 8.3 |
| 2007/08 | 0.025 | 0.06 | 0.13 | 0.04 | 0.83 | 8.5 |
| 2008/09 | 0.031 | 0.07 | 0.73 | 0.05 | 0.74 | 8.8 |
| 2009/10 | 0.037 | 0.09 | 0.23 | 0.06 | 0.39 | 10.1 |
| 2010/11 | 0.029 | 0.07 | 0.59 | 0.05 | 0.94 | 8.3 |
| 2011/12 | 0.026 | 0.06 | 1.00 | 0.04 | 0.90 | 8.4 |
| 2012/13 | 0.026 | 0.06 | 0.62 | 0.04 | 0.90 | 8.4 |
| All seasons | $\beta/p = 76$ | $\beta^* = 32$ | | $\beta^* = 47$, $k = -0.48$, $\epsilon^* = 3.9$ | | |

Table 4.1: (Previous page) The reference model ($p$ constant) assumes the ILI factor $p$ to be constant throughout all seasons. The intermediate model ($p$ per season) assumes the ILI factor $p$ constant throughout a season, but changing from season to season. The final model ($p$ by climate) assumes a correlation between the typical absolute humidity and the ILI factor. The parameter $W$ is the typical absolute humidity at the peak of each epidemic, with the associated ILI factor $p(W)$. Parameters indicated with a ($^*$) are scaled due to the normalization of the ILI factor $p$.

be interpreted as the average proportion of influenza cases that fitted the ILI case definition during the epidemic in that specific season.

System (4.1) with a constant transmission rate $\beta$ and season dependent parameters $S_0$ and $p$ is mathematically equivalent to applying system (4.3) to each season independently. From the estimated parameters $pS_0$ and $\frac{\beta}{p}$ for each season, we separate a constant transmission rate $\beta$ and a normalized season dependent ILI factor $p$. Figure 4.3 shows the estimated ILI incidence curves and the estimated parameters are listed in Table 4.1 ($p$ per season).

Figure 4.4 shows the relation between the estimated ILI factor $p$ for each epidemic, and two weather factors (temperature and absolute humidity) during the epidemic period. The epidemic period is defined as the two weeks before and two weeks after the epidemic peak. For both weather factors a similar correlation with the logarithm of the ILI factor $p$ is observed, whether we consider the actual measurements for each year or 40-year averages. Located in the European continent, the three countries included in this study exhibit typical northern weather patterns, whereby temperature and humidity decrease between August and February, to start increasing again until August, completing the annual cycle. As the epidemic peaks occur when the weather factors above are going through the decreasing phase, we also obtain a correlation

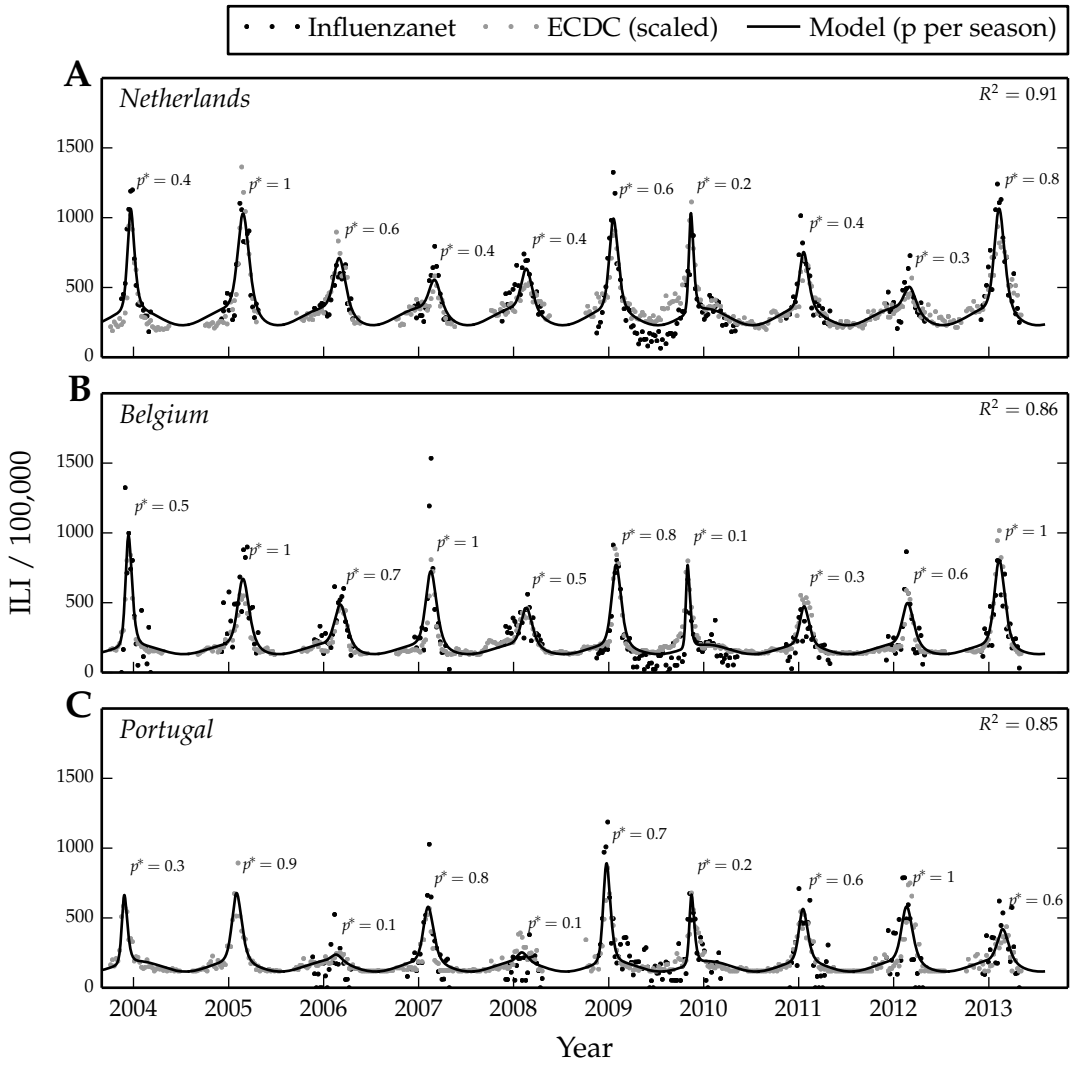



Figure 4.3: BEST FIT FOR THE INTERMEDIATE MODEL based on the second approach: the ILI factor $p$ estimated independently for each season for A) Netherlands, B) Belgium, and C) Portugal (2003–2013). The ILI incidence by Influenzanet is defined as the number of ILI onsets per 100,000 participant-weeks.

Table 4.2: GOODNESS-OF-FIT OF THE MODELS TO THE DATA for the reference model ("Constant") and the final models using either the The weekly absolute humidity ("Abs hum"), the typical absolute humidity based on a 40-year average ("Typ abs hum"), the weekly temperature ("Temp"), and the typical temperature ("Typ temp").

| COUNTRY | CONSTANT | TYP ABS HUM | ABS HUM | TYP TEMP | TEMP |
|---|---|---|---|---|---|
| Netherlands | 0.66 | 0.82 | 0.77 | 0.79 | 0.74 |
| Belgium | 0.53 | 0.81 | 0.70 | 0.76 | 0.66 |
| Portugal | 0.65 | 0.76 | 0.74 | 0.78 | 0.76 |

between the ILI factor and the timing of the epidemic activity (Figure B.14, Appendix B, page 218).

### 4.3.4 *Third approach: ILI factor by climate*

The observed correlation between the weather variable ($W$) and the ILI factor $p$, suggests the following relation:

$$\log p(t) = k \cdot W(t) + \epsilon \tag{4.4}$$

where the parameters $k$ and $\epsilon$ are the linear regression coefficients. Instead of a season-dependent ILI factor, we will now apply the above relation between the ILI factor and a weather variable into the model. The constant transmission rate $\beta$, the season dependent initial susceptibility $S_0$ and the parameters $k$ and $\epsilon$ were initially approximated by adopting the averaged weather variable over a period of four weeks around the epidemic peaks. Using a continuously changing weather variable $W$ throughout the seasons, we will readjust the model using these parameters as initial conditions.

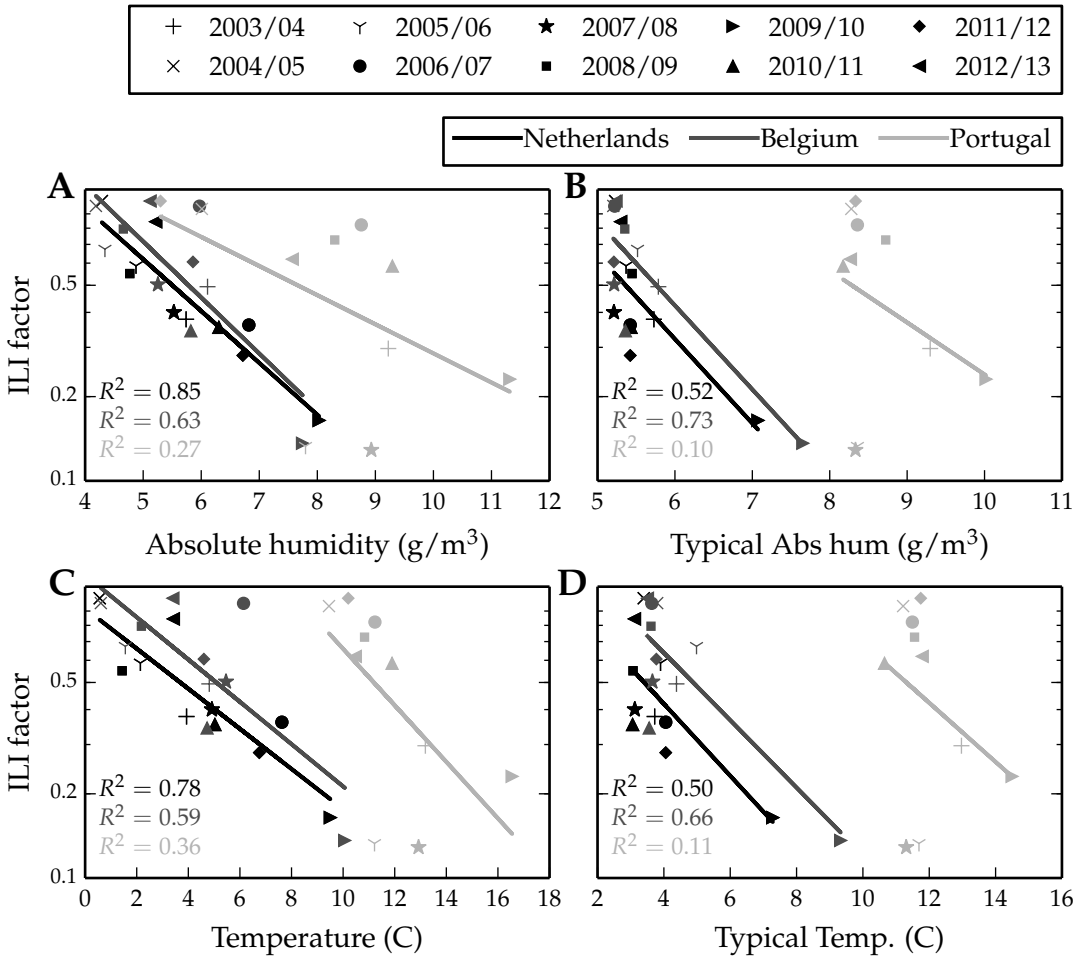



Figure 4.4: Relation between ILI factor and weather for A) absolute humidity, B) typical absolute humidity based on a 40-year average, C) temperature, and D) typical temperature. The ILI factor $p$ is estimated by the second approach, where it is assumed constant throughout the season but changes from season to season.

Using the typical average humidity as the weather variable, Figure 4.5 shows the estimated ILI incidence and the estimated parameters are listed in Table 4.1 ($p$ by climate). Estimated ILI incidence curves based on the other weather variables are available in the supplementary material (Figures B.15–B.17, Appendix B, page 219). The goodness-of-fit improves significantly by applying a seasonal ILI factor instead of a constant ILI factor, using either temperature or absolute humidity (Table 4.2). Although the goodness-of-fit of the intermediate model (as measured by $R^2$) may appear more favorable, we must note that the weather model only requires two parameters ($k$ and $\epsilon$) and the weather measurements to determine the time-dependent ILI factor $p$, whereas in the intermediate model the number of parameters is as large as the number of epidemic seasons included in the fit (ten in this case). The F-test statistic confirms that both models perform significantly better than the original constant-$p$ model. Since in the three countries absolute humidity and temperature are highly correlated (Figure 4.6), it is expected that both weather variables perform equally well.

### 4.3.5 *Influenza and ILI*

Figure 4.7 shows the dynamics of four characteristic seasons for each country, the full line representing ILI incidence according to the best fitting model and the dashed lines the extrapolated influenza frequencies by applying the ILI factor $p$. Early epidemics, which occur when the initial proportion of susceptibles is high, are characterized by high frequency of influenza infections, but not necessarily a higher burden of ILI (Panels A–C) as the epidemic activity occurs when the ILI factor is still low (Panels D–F).

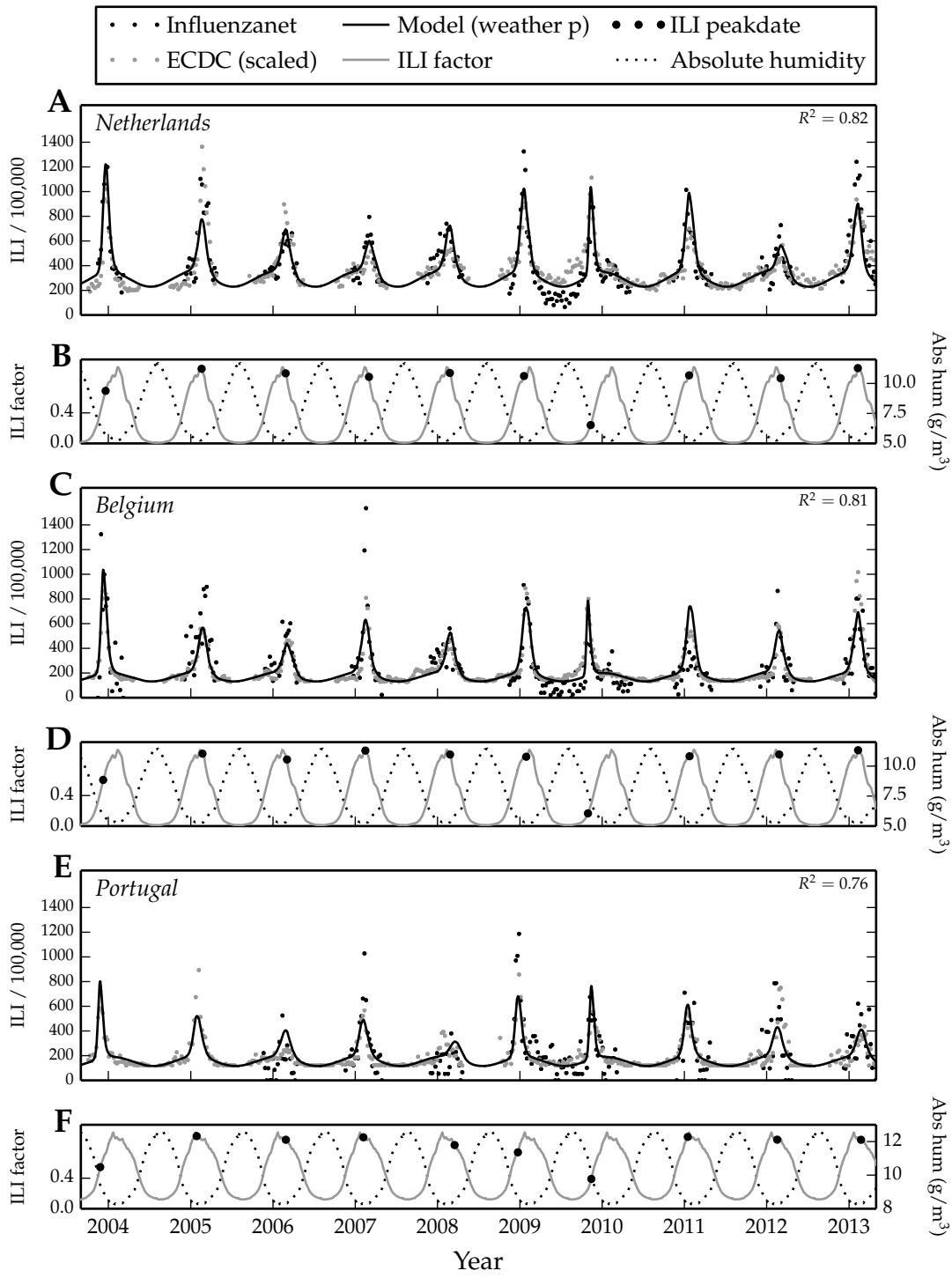



Figure 4.5: Best fit for the final model

Figure 4.5: (Previous page) Based on the third approach: the ILI factor $p$ determined by the typical absolute humidity. The ILI incidence by Influenzanet is defined as the number of ILI onsets per 100,000 participant-weeks. A,C,E) ILI incidence curves and B,D,F) the typical absolute humidity and the corresponding ILI factor $p$ for A,B) Netherlands, C,D) Belgium, and E,F) Portugal (2003–2013). The ILI incidence by Influenzanet is defined as the number of ILI onsets per 100,000 participant-weeks. The black dots on the ILI factor curve indicate the time of the epidemic peaks.

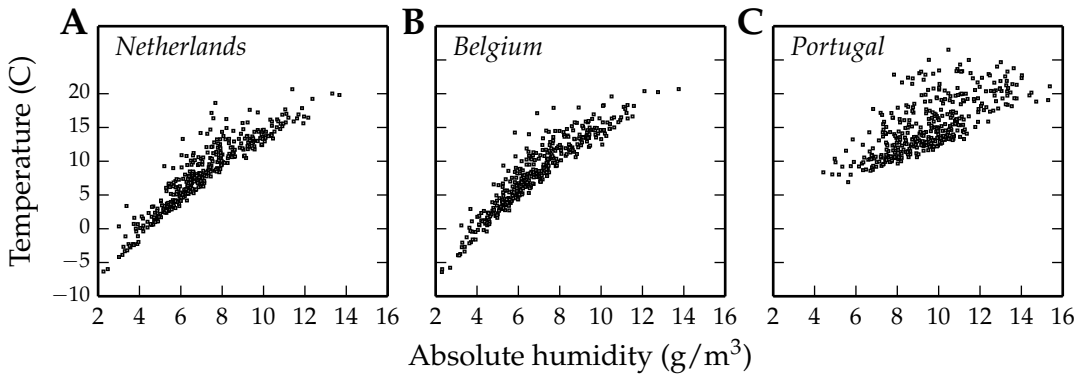



Figure 4.6: ABSOLUTE HUMIDITY AND TEMPERATURE in A) Netherlands, B) Belgium, and C) Portugal (2003–2013).

Figure 4.8 plots the relation between the initial proportion of susceptibles and the peak ILI incidence generated by the model with estimated parameters for each country. The peak ILI incidence initially increases with the initial proportion of susceptibles, but the tendency is inverted at some point. Increasing the proportion of susceptibles above a certain level will lead to a lower peak ILI incidence, due the milder weather conditions at the time in which an early epidemic takes place.

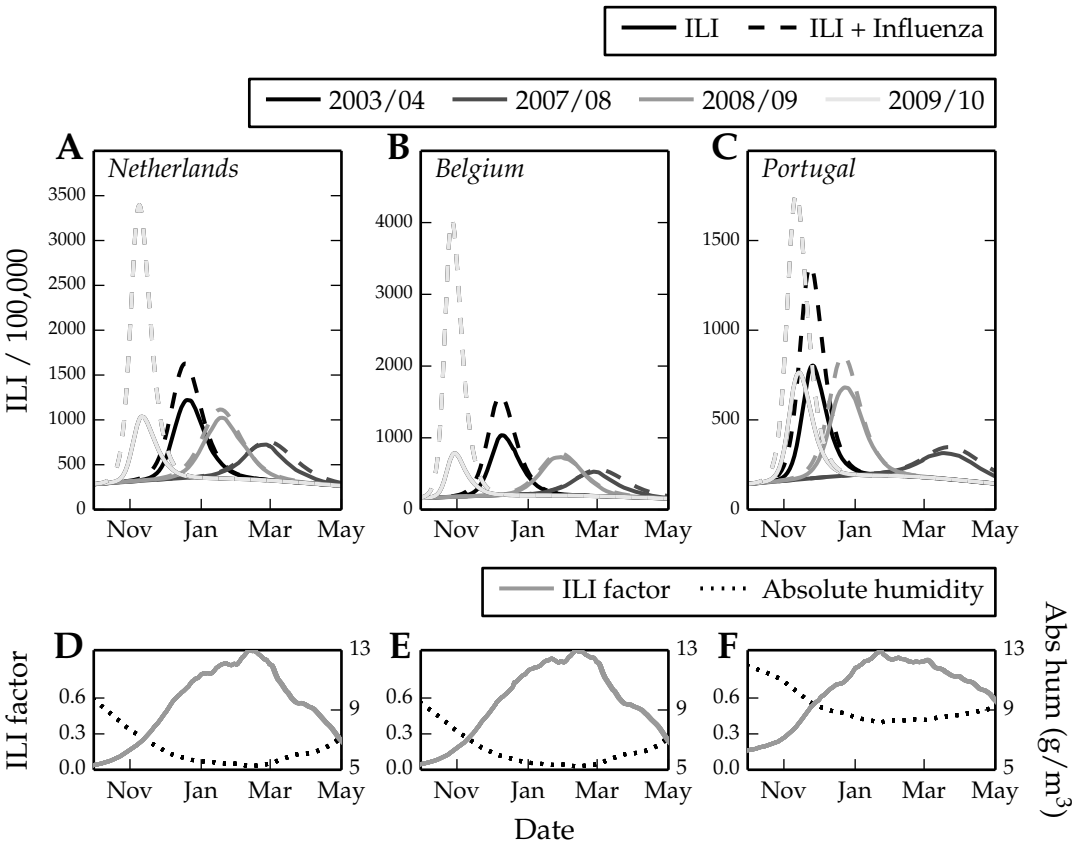



Figure 4.7: Fitted ILI incidence and estimated influenza incidence for four characteristic seasons in A,D) Netherlands, B,E) Belgium, and C,F) Portugal. A,B,C) The ILI incidence by Influenzanet is defined as the number of ILI onsets per 100,000 participant-weeks(solid line). The influenza incidence (dashed line) is determined by the estimated ILI factor $p$ and the ILI incidence, including the non-influenza ILI cases (baseline). D,E,F) The typical absolute humidity and the corresponding ILI factor $p$.

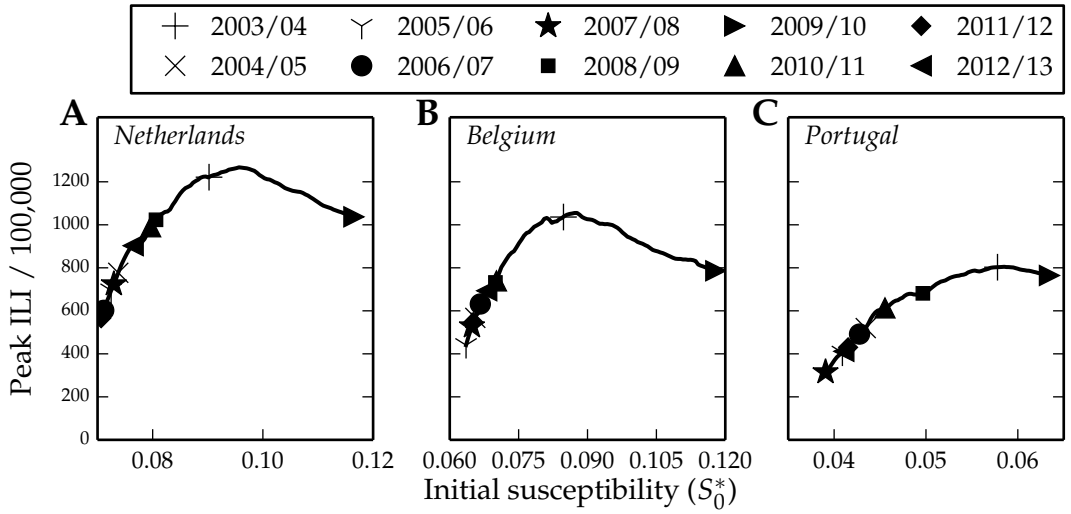



Figure 4.8: PEAK ILI INCIDENCE AND INITIAL SUSCEPTIBILITY in A) Netherlands, B) Belgium, and C) Portugal. The ILI incidence by Influenzanet is defined as the number of ILI onsets per 100,000 participant-weeks. The final model with estimated parameters are simulated for a range of initial susceptible proportions. The values corresponding to seasons included in this study are specifically marked.

## 4.4 DISCUSSION

By parameterizing multiple time series of influenza-like illness, we unlock a potential mechanism for the modest variation of ILI rates across seasons in temperate climates despite large variations in the initial susceptibility of the population. Focusing on the northern hemisphere (more specifically, three European countries) we simulate a minimal transmission model that runs for 9 months each year (September–May) over 10 seasons (2003–2013). The proportion of the host population susceptible to infection by the dominant virus ($S_0$) is a specificity of each influenza season since it depends on the antigenic novelty of the emerging virus. Variability in population susceptibility leads to variability in the timing of onset and peak of the epidemic — epidemics peak later when susceptibility is low. We find a significant correlation between the temperature and absolute humidity at time of infection and the proportion of infected persons fitting the ILI case definition, the ILI factor $p$.

By making the ILI factor $p$ directly dependent on weather variables, a single model simulation is able to capture the ILI incidence of all seasons 2003–2013, which includes the 2009–2010 season characterized by a new influenza A pandemic strain, the 2003–2004 season characterized by an antigenic cluster jump (SMITH ET AL., 2004), and the seasons 2005–2006, 2007–2008, 2010–2011, and 2012–2013 characterized by co-circulating influenza B strains.

Influenza transmission models which include weather variables typically incorporate seasonal variation on the transmission rate. We demonstrate that seasonal variation in the ILI factor alone significantly improves the ability of a simple model to reproduce the observed ILI incidence curves. It is also well recognized that each epidemic is caused predominantly by a different influenza virus, which could be evoked in defense of season-to-season variation

in the model parameters. These three factors — seasonality in transmission, seasonality in the ILI factor, and season-to-season variation due to changes in viral phenotype — are totally compatible and there is nothing precluding their combination in a more elaborate model formulation; however, more data types would be required for parameter estimation.

A more direct approach in determining the seasonal variation of the ILI factor $p$, for example, could be the confirmation of the presence of ILI symptoms in influenza infected persons (LAU ET AL., 2010) under various weather conditions. Season specific information about the circulating influenza viruses could also be included directly into the model when available. Figure 4.4 suggests that the ILI factor was relatively lower in Portugal during the seasons 2005–2006 and 2007–2008, seasons with significant circulation of influenza B, which could lead to milder symptoms (KIM ET AL., 1979). The migration rate $m$ of influenza from abroad, is assumed to be constant throughout all seasons, but the global surge of pandemic influenza A in 2009, could have increased the migration rate.

Although the model predicts positive correlation between the initial susceptibility and the incidence of influenza infections, after applying the ILI factor we are left with no correlation between the initial susceptibility and the incidence of ILI cases (CHOWELL ET AL., 2008). Public health measures taken to delay the onset of the epidemic, such as school closures and global vaccination, might lead to an increased ILI burden by shifting the epidemic to colder and dryer months.

## ACKNOWLEDGMENTS

We thank Flávio Coelho, Cláudia Codeço, Katia Koelle and Graham Medley for valuable discussions, and Jessica King for help with text editing. We thank the European Centre for Disease Control and Prevention (ECDC), and the national Influenzanet teams from the Netherlands, Belgium (Grote Griepmeting) and Portugal (Gripenet) for the data. Sander van Noort and Gabriela Gomes thank the Mathematical Biosciences Institute, Ohio State University, for hosting during the final stages of this work. This research was funded by the European Commission (EC-ICT-231807).

## AUTHOR CONTRIBUTIONS

Sander van Noort developed the model, analyzed the data, and wrote the paper. Ricardo Águas and Sébastien Ballesteros contributed to previous versions of the model. Gabriela Gomes contributed to the development of the model and revised the paper.

## SIGNIFICANT CHANGES COMPARED TO PUBLISHED ARTICLE

The analyses has been extended to include the seasons 2003–2013. The period over which the typical temperature and absolute humidity is determined is extended to 1970–2010. The correlation between temperature and absolute humidity is visualized in Figure 4.6. The F-test is only used to compare nested models. The model output (ILI prevalence) is converted to an ILI incidence rate, and all analyses and figures use the ILI incidence. The method

to determine the baseline ILI incidence is based on Section 6.3.3. The supplementary text on the model including asymptomatics has been included into the Method section.

## 4.5 REFERENCES

Assaad FA and Reid D (1971). Some factors influencing mortality from influenza. *Bull World Health Organ*, 45(1):113–117.

Bacaër N and Ait Dads EH (2011). Genealogy with seasonality, the basic reproduction number, and the influenza pandemic. *J Math Biol*, 62(5):741–762.

Baetjer AM (1967). Effect of ambient temperature and vapor pressure on ciliamucus clearance rate. *J Appl Physiol*, 23(4):498–504.

Cauchemez S, Valleron AJ, Boëlle PY, Flahault A, and Ferguson NM (2008). Estimating the impact of school closure on influenza transmission from sentinel data. *Nature*, 452(7188):750–754.

Chowell G, Miller MA, and Viboud C (2008). Seasonal influenza in the united states, france, and australia: transmission and prospects for control. *Epidemiol Infect*, 136(6):852–864.

Dowell SF (2001). Seasonal variation in host susceptibility and cycles of certain infectious diseases. *Emerg Infect Dis*, 7(3):369–374.

Eccles R (2002). An explanation for the seasonality of acute upper respiratory tract viral infections. *Acta Otolaryngol*, 122(2):183–191.

Ferguson NM, Mallett S, Jackson H, Roberts N, and Ward P (2003). A population-dynamic model for evaluating the potential spread of drug-resistant influenza virus infections during community-based use of antivirals. *J Antimicrob Chemother*, 51(4):977–990.

Fuhrmann C (2010). The effects of weather and climate on the seasonality of influenza: What we know and what we need to know. *Geography Compass*, 4(7):718–730.

Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, et al. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014.

HOPE-SIMPSON RE (1981). The role of season in the epidemiology of influenza. *J Hyg (Lond)*, 86(1):35–47.

KIM HW, BRANDT CD, ARROBIO JO, MURPHY B, CHANOCK RM, ET AL. (1979). Influenza a and b virus infection in infants and young children during the years 1957-1976. *Am J Epidemiol*, Apr;109(4):464–79.

KOELLE K, COBEY S, GRENFELL B, and PASCUAL M (2006). Epochal evolution shapes the phylodynamics of interpandemic influenza a (h3n2) in humans. *Science*, 314(5807):1898–1903.

LAU LLH, COWLING BJ, FANG VJ, CHAN KH, LAU EHY, ET AL. (2010). Viral shedding and clinical illness in naturally acquired influenza virus infections. *J Infect Dis*, 201(10):1509–1516.

LIPSITCH M and VIBOUD C (2009). Influenza seasonality: lifting the fog. *Proc Natl Acad Sci U S A*, 106(10):3645–3646.

MANDL KD, OVERHAGE JM, WAGNER MM, LOBER WB, SEBASTIANI P, ET AL. (2004). Implementing syndromic surveillance: a practical guide informed by the early experience. *J Am Med Inform Assoc*, 11(2):141–150.

MEERHOFF TJ, PAGET WJ, AGUILERA JF, and VAN DER VELDEN J (2004). Harmonising the virological surveillance of influenza in europe: results of an 18-country survey. *Virus Res*, 103(1-2):31–33.

MILLS CE, ROBINS JM, and LIPSITCH M (2004). Transmissibility of 1918 pandemic influenza. *Nature*, 432(7019):904–906.

NELSON MI, SIMONSEN L, VIBOUD C, MILLER MA, and HOLMES EC (2007). Phylogenetic analysis reveals the global migration of seasonal influenza a viruses. *PLoS Pathog*, 3(9):1220–1228.

SHAMAN J and KOHN M (2009). Absolute humidity modulates influenza survival, transmission, and seasonality. *Proc Natl Acad Sci U S A*, 106(9):3243–3248.

SHAMAN J, PITZER VE, VIBOUD C, GRENFELL BT, and LIPSITCH M (2010). Absolute humidity and the seasonal onset of influenza in the continental united states. *PLoS Biol*, 8(2):e1000316.

SMITH DJ, LAPEDES AS, DE JONG JC, BESTEBROER TM, RIMMELZWAAN GF, ET AL. (2004). Mapping the antigenic and genetic evolution of influenza virus. *Science*, 305(5682):371–376.

VAN NOORT SP, CODEÇO CT, KOPPESCHAAR CE, VAN RANST M, and GOMES MGM (2014). Ten-year performance of influenzanet: ili time series, risks, and vaccine effects in the grote griepmeting, gripenet, and influweb cohorts. *Submitted*.

van Noort SP, Muehlen M, Rebelo de Andrade H, Koppeschaar CE, Lourenço JML, et al. (2007). Gripenet: an internet-based system to monitor influenza-like illness uniformly across europe. *Euro Surveill*, 12(7):E5–E6.

# 5

# IMMUNE SELECTION AND WITHIN-HOST COMPETITION CAN STRUCTURE THE REPERTOIRE OF VARIANT SURFACE ANTIGENS IN *Plasmodium falciparum* — A MATHEMATICAL MODEL

Sander P. van Noort[1], Marta C. Nunes[2], Gareth D. Weedall[1], Lars Hviid[3], M. Gabriela M. Gomes[1]

[1] Instituto Gulbenkian de Ciência, Oeiras, Portugal

[2] University of the Witwatersrand, Johannesburg, South Africa

[3] University of Copenhagen and Rigshospitalet, Copenhagen, Denmark

*PLoS One (2010) Mar 22; 5(3): e9778*

ABSTRACT

*The evolutionary mechanisms structuring the expression pattern of variant surface antigen (VSA) families that allow pathogens to evade immune responses and establish chronic and repeated infections pose major challenges to theoretical research. In Plasmodium falciparum, the best-studied VSA family is erythrocyte membrane protein 1 (PfEMP1). Each parasite genome encodes about 60 PfEMP1 variants, which are important virulence factors and major targets of host antibody responses. Transcriptional switching is the basis of clonal PfEMP1 variation and immune evasion. A relatively conserved subset of PfEMP1 variants tends to dominate in non-immune patients and in patients with severe malaria, while more diverse subsets relate to uncomplicated infection and higher levels of pre-existing protective immunity.*

*Here, we use the available molecular and serological evidence regarding VSAs, in particular PfEMP1, to formulate a mathematical model of the evolutionary mechanisms shaping VSA organization and expression patterns. The model integrates the transmission dynamics between hosts and the competitive interactions within hosts, based on the hypothesis that the VSAs can be organized into so-called dominance blocks, which characterize their competitive potential. The model reproduces immunological trends observed in field data, and predicts an evolutionary stable balance between inter-clonally conserved dominance blocks that are highly competitive within-host and diverse blocks that are favored by immune selection at the population level.*

*The application of a monotonic dominance profile to VSAs encoded by a gene family generates two opposing selective forces and, consequently, two distinct clusters of genes emerge in adaptation to naive and partially immune hosts, respectively.*

## 5.1 INTRODUCTION

Although people living in malaria endemic regions typically carry *Plasmodium falciparum* parasites throughout life, clinical symptoms decrease markedly with age (DOOLAN ET AL., 2009). Naturally acquired immunity to the disease involves many components and their relative importance is only partially understood (LANGHORNE ET AL., 2008). However, antibodies undoubtedly form a critical component of immunity to the asexual blood stages (COHEN ET AL., 1961), and the parasite-encoded variant surface antigens (VSAs) exported to the surface of infected erythrocytes (IEs) are important targets (MARSH AND HOWARD, 1986; HVIID, 2005). *P. falciparum* parasites possess several VSA families, of which the best characterized is *P. falciparum* erythrocyte membrane protein 1 (PfEMP1) encoded by approximately 60 var genes per genome (GARDNER ET AL., 2002). The level of diversity among var genes varies greatly both between and within individual genomes (KYES ET AL., 2007). PfEMP1 variants mediate adhesion of IEs to different host endothelial receptors, and different binding properties have been associated with distinct patterns of sequestration and pathogenesis (MILLER ET AL., 2002). The importance of PfEMP1 in malaria pathogenesis has motivated the development of theoretical models of diversity and immune selection (GATTON AND CHENG, 2004; RECKER ET AL., 2004; BUCKEE ET AL., 2009).

Individual IEs express only a single PfEMP1 variant at a time (SCHERF ET AL., 2008). Early in blood stage infection after liver release, many var genes are transcribed by the various IEs, but gradually this pattern changes and particular subsets of var genes are predominantly expressed (PETERS ET AL., 2002; LAVSTSEN ET AL., 2005), while others may still be expressed at low frequency due to transcriptional switching (ROBERTS ET AL., 1992; HORROCKS

ET AL., 2004). Variants that predominate in the early phase of infection probably have higher effective multiplication rates (possibly due to more efficient endothelial sequestration rates) or higher on-switching rates. In any case, the history of PfEMP1 expression is recorded in the antibody repertoires that accumulate in individual hosts, regardless of the molecular basis of the sequence of expression (MARSH AND HOWARD, 1986; BULL ET AL., 1998; GIHA ET AL., 1999; OFORI ET AL., 2002).

There is evidence that there is a threshold of PfEMP1 expression necessary for induction of an immune response (KRAUSE ET AL., 2007). If so, low-level or heterogeneous expression of PfEMP1 variants, such as in the early stages of infection, might not be sufficient to induce immunity. As the immune system disables IEs expressing the dominant VSA, the parasite is either cleared from the host or parasites expressing an antigenically distinct VSA will come to dominate the infection (STAALSOE ET AL., 2002). When a VSA is no longer expressed, antibody levels against it decrease, but immunological memory persists and antibody levels can be rapidly restored upon re-exposure (O'NEIL-DUNNE ET AL., 2001; STAALSOE ET AL., 2001; NIELSEN ET AL., 2005).

Here we investigate the role of variation in adhesion properties and cumulative antibody repertoires in selecting for the observed patterns in expression (BULL ET AL., 1999; NIELSEN ET AL., 2002). Integrating these individual-level processes into a mathematical model of *P. falciparum* transmission, we refine the requirements for the emergence of realistic variation in VSA expression at the population level. The model allows for multiple genotype infections, encapsulating a form of within-host competition that gives selective advantage to parasites expressing more dominant VSAs (PHIRI ET AL., 2009).

Although global var gene diversity is immense (BARRY ET AL., 2007), there is increasing evidence that there exist restricted subgroups of antigenically simi-

lar VSAs that have a selective advantage in naive hosts and are associated with severe disease (called "high-dominance" VSAs here), whereas other more diverse PfEMP1 variants (called "low-dominance" VSAs) are more common in the uncomplicated and sub-clinical infections of more immune hosts (Nielsen et al., 2002; Bull et al., 2000; Kirchgatter and del A Portillo, 2002; Kaestli et al., 2006; Kyriacou et al., 2006; Rottmann et al., 2006; Normark et al., 2007; Bull et al., 2008). Our model suggests that within-host competition selects for a relatively conserved repertoire of high-dominance VSAs, while a diverse repertoire of low-dominance forms is maintained by their ability to remain unrecognized by host immunity for extended periods allowing chronicity of infections. We propose this mechanism of two-level selection as an evolutionary explanation for the subdivision of large VSA families such as PfEMP1.

## 5.2 METHOD

On the basis of available experimental evidence summarized above, we hypothesize that the global repertoire pool of variants within a given VSA family can be ordered into a dominance hierarchy that determines the order in which they are expressed in an infection. The dominance hierarchy is considered the aggregated result of a variety of selective factors, including adhesion avidity, receptor availability, metabolic cost, gene switching rates, and immunodominance, to name a few. The parasites in an infection will therefore tend to express the most dominant variant to which the host does not have pre-existing immunity. As immunity to the initially dominant variant is acquired, continued parasite survival depends on the ability to switch away from this variant, and switching back to the originally expressed variant will be unsuc-

cessful as long as protective levels of antibody with specificity for this variant persist.

Before constructing a model we must devise a scheme to aggregate the immense VSA diversity in a way that is both biologically meaningful and mathematically tractable. As we are mainly interested here to explore host population-level processes, we define "dominance blocks" as groupings of undefined numbers of consecutive variants. Since dominance reflects preferential expression, it is assumed here that antigenic switching during a single infection occurs among consecutive variants within a dominance block, and is therefore not visible at the scale of blocks. Dominance blocks are thus a convenient unit for the construction of transmission models at this level.

Estimates of the duration of a *P. falciparum* infection are in the order of 200 to 700 days (MacDonald, 1950; Sama et al., 2004; Aguas et al., 2008), with peaks in parasitaemia every 20–25 days (Collins and Jeffery, 1999). If peaks in parasitaemia correspond to clonal replacement of one variant by another, and in the absence of substantial cross-reactive immunity among intraclonal variants (Joergensen et al., 2006), between eight (an infection that lasts 200 days with peaks every 25 days) and 35 (peaks every 25 days of a 700 day infection) different VSAs can be assumed to be expressed in the course of an untreated monoclonal infection. This collection of VSAs corresponds to a dominance block. The model will be formulated in terms of blocks of VSAs, indexed such that a lower index represents a higher dominance (Figure 5.1A, top). Each parasite can thus be assumed to possess between approximately two (at 35 variants per block) and seven (at eight variants per block) dominance blocks of a given VSA family such as the 60-member PfEMP1 family, and should therefore be able to re-infect a given host at least a corresponding number of times (each time expressing variants from a new block) before all

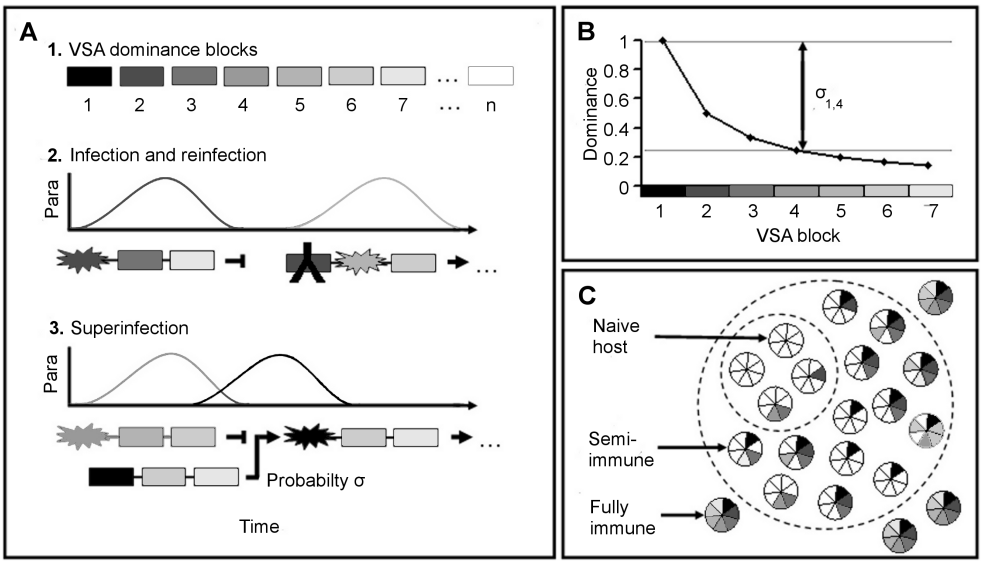



Figure 5.1: SCHEMATIC REPRESENTATION OF INFECTION DYNAMICS . A) VSA variants are organized into dominance blocks, where dominance ranges from the highest (VSA block 1) to the lowest (VSA block n) (A, top). Each parasite genotype contains VSAs from a fixed number of VSA blocks. In a naive host, a parasite clone (illustrated by a set of 3 VSA blocks) expresses VSAs from the most dominant block. Immunity is illustrated by an antibody and expression by a star. The host mounts an immune response to the expressed VSAs and eventually the infection is cleared. On a subsequent infection the host is already immune to VSAs from previously expressed blocks, leading to the expression of VSAs belonging to the next most dominant VSA block (A, middle). When an infected host is exposed to a new parasite which encodes VSAs from a more dominant block, the resident parasite can be replaced with a probability $\sigma$ (A, bottom). B) Implementation of dominance hierarchy of 7 VSA blocks, such that $\sigma_{1,4}$ is the probability of superinfection when VSA block 1 invades a host with a resident parasite expressing VSA block 4. C) Heterogeneous immune repertoires among hosts. Hosts acquire specific immunity with exposure, represented by colors matching VSA blocks in previous panels. The immune repertoire may contain gaps if a host has not been exposed to a particular variant block. Fewer hosts are susceptible to VSA 1 (small dashed circle) than to VSA 7 (large dashed circle).

family members have been expressed (Figure 5.1A, middle). In reality, the number may well be higher, as levels of variant-specific antibodies often decline fairly rapidly once exposure to the variant ceases (O'Neil-Dunne et al., 2001; Staalsoe et al., 2001; Nielsen et al., 2005).

Multiple-clone infections are very common in malaria endemic areas. We hypothesize that VSA dominance not only determines the order of VSA expression of a single clonal lineage, but also the dynamics of multiple-clone infections. When an already infected host is exposed to a new parasite, we assume that the transmission potential of the invader in relation to the resident increases with the dominance difference between the respective VSAs. Others (Gog and Grenfell, 2002) have shown that such weighted processes can be mathematically simplified while maintaining the essence of the model dynamics, by assuming the polarized view that the invader replaces the resident parasite (superinfection) with a probability $\sigma$ (Figure 5.1A) and is cleared otherwise. The probability $\sigma$ is formally defined as a function of the difference between the dominance blocks expressed by invader and resident parasites (Figure 5.1B).

As hosts in the population gradually acquire immunity to individual dominance blocks, they remain susceptible only to parasites which express blocks of lower dominance (Figure 5.1C). A correlation between dominance and disease severity is implicit in the model, and is used to evaluate its performance. Immunity to specific VSA blocks wanes over time. The mathematical formalism of the model is provided in Section 5.5.

## 5.3 RESULTS

We will describe equilibrium results from simulations of model realizations where parasites are described by four dominance blocks drawn from a pool of seven. This greatly simplifies the description of the model output but retains the generality of the model performance.

The vast majority of hosts without any pre-existing VSA-specific antibodies is predicted by the model to be infected by parasites expressing VSAs belonging to dominance block 1. As host repertoires of VSA-specific antibodies broaden, the probability that their infections will be dominated be parasites expressing VSAs from dominance blocks lower in the hierarchy increases, and the ability to predict which block is expressed in a given host decreases (Figure 5.2A). If we associate high-dominance VSAs with more severe forms of malaria, and note that the antibody repertoire broadens with age, this output fits the observations that overall malaria severity decreases with age in endemic areas (DOOLAN ET AL., 2009), and that low immunity and young age are associated with infections dominated by serologically similar VSAs, whereas VSA expression in older, more immune individuals with uncomplicated infections is much more diverse (NIELSEN ET AL., 2002; BULL ET AL., 2000). The host's capacity to clear an infection before exhausting the VSA repertoire of the infecting parasite (see Model outline above) and the non-random sequential expression of variants from high- to low-dominance, furthermore leads the model to predict a population-level gradient from high prevalence of hosts with antibodies against high-dominance VSAs to low prevalence of hosts with antibodies against low-dominance VSAs (Figure 5.2B). The model also predicts a negative correlation between the size of the antibody repertoire and the seroprevalence against the expressed VSA, meaning that VSAs

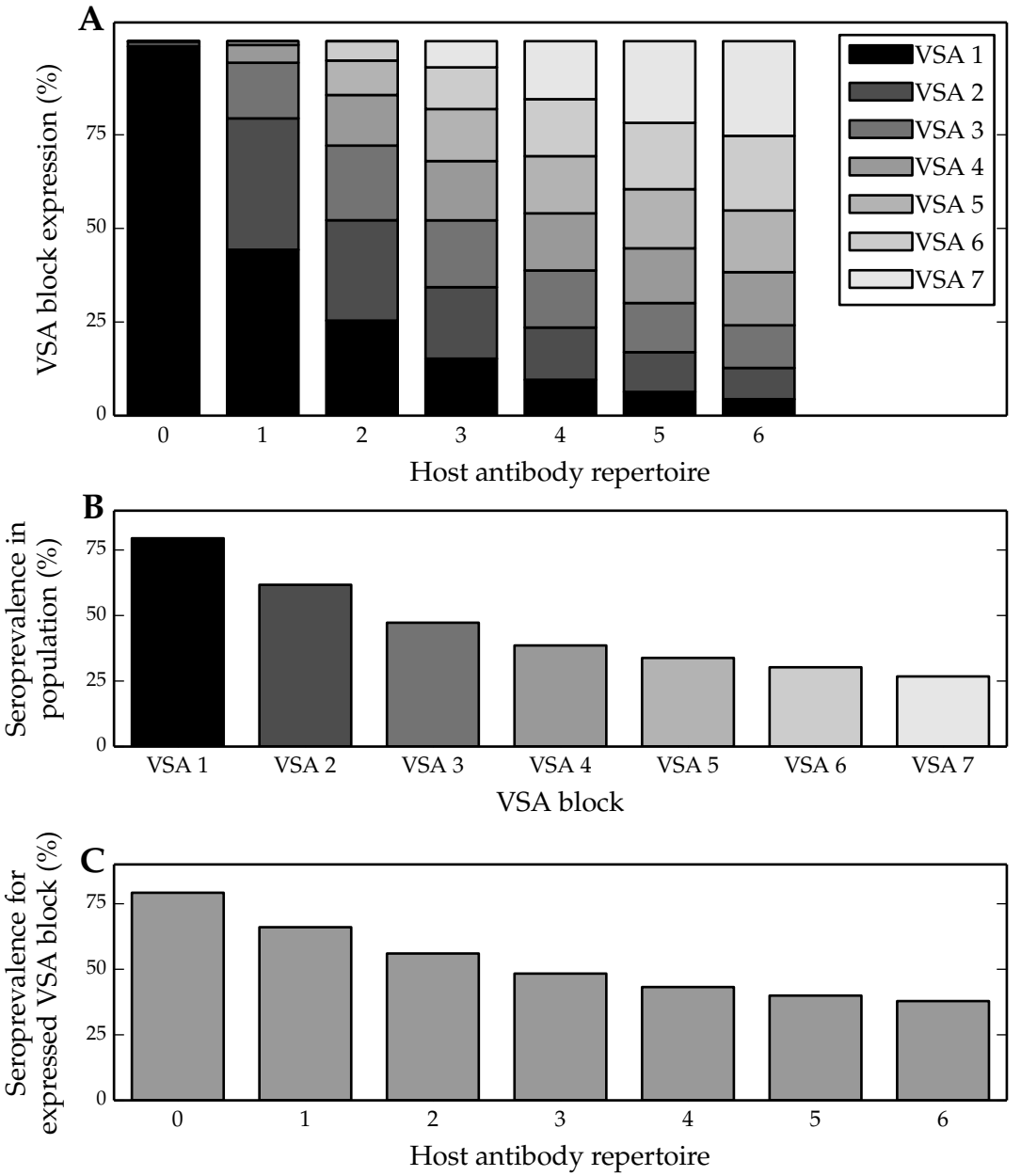



Figure 5.2: EXPRESSION AND SEROPREVALENCE OF VSAs at endemic equilibrium. A) VSA expression within hosts with different levels of past exposure (measured by host antibody repertoire). B) Seroprevalence for the different VSA blocks (trend displayed in Fig. 3 of (BULL ET AL., 2000)). C) Seroprevalence for the expressed VSA as a function of host antibody repertoire (trend displayed in Fig. 2c of (BULL ET AL., 2005)).

expressed in hosts with broad antibody repertoires are less recognized in the host population than VSAs expressed in hosts with narrow antibody repertoires (Figure 5.2C). The trends in Figure 5.2B,C correspond well with field data (BULL ET AL., 1999; NIELSEN ET AL., 2002; BULL ET AL., 2005).

The model predicts that the high prevalence of antibodies with specificity for dominant VSAs favors parasites encoding low-dominance variants while parasites encoding the high-dominance variants are simultaneously favored because of their ability to superinfect and displace resident parasites from infected hosts (Figure 5.3, solid red line). These opposing selective forces shape the parasite population such that a typical parasite genome will contain VSAs from both high- and low-dominance VSAs, while variants of intermediate dominance will be the least frequent in the parasite population. Indeed, the selection for high-dominance VSAs becomes weaker when the ability to superinfect is removed from the model, while assigning only a single VSA block to each parasite weakens the selection for low-dominance VSAs. Parasite genomes encoding multiple copies of the same VSAs are outcompeted by parasite genomes encoding the maximum number of distinct VSAs, consistent with high inter-locus diversity seen in the genome (GARDNER ET AL., 2002).

Finally, we explore the dependence of VSA distributions on the size of the available pool of variants. By increasing the size of the VSA pool from 5 to 8 blocks, the model predicts that the prevalence of high-dominance VSAs is essentially independent on the global size of the pool, while low dominance VSAs tend to become more heterogeneously distributed throughout the parasite population (Figure 5.4). In reality, the number of antigenically distinct VSAs, and thereby the number of possible dominance blocks, is likely to be very large, leading to a restricted set of highly dominant and serologically

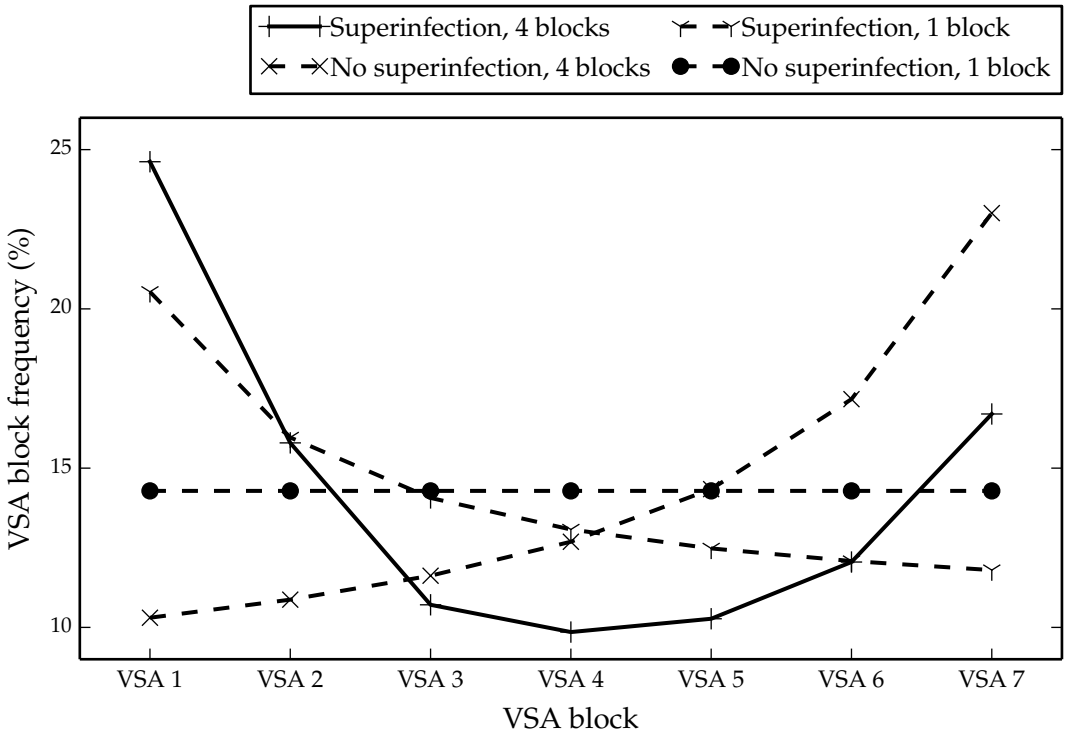



Figure 5.3: FREQUENCY DISTRIBUTIONS OF VSAS IN THE PARASITE POPULATION at endemic equilibrium. Opposing selection pressures favor VSA blocks at each end of the dominance hierarchy (solid red line), suggesting a mechanism behind the two clusters displayed in Fig. 3 of (BULL ET AL., 2008). Superinfection selects for high-dominance VSAs (compare solid red line and dotted blue line). When parasites contain multiple VSA blocks, acquired immunity selects for low-dominance variants (compare solid red line and dotted green line). Without superinfection and only a single VSA block per parasite, there is no frequency difference between the different VSA blocks (dotted yellow line).

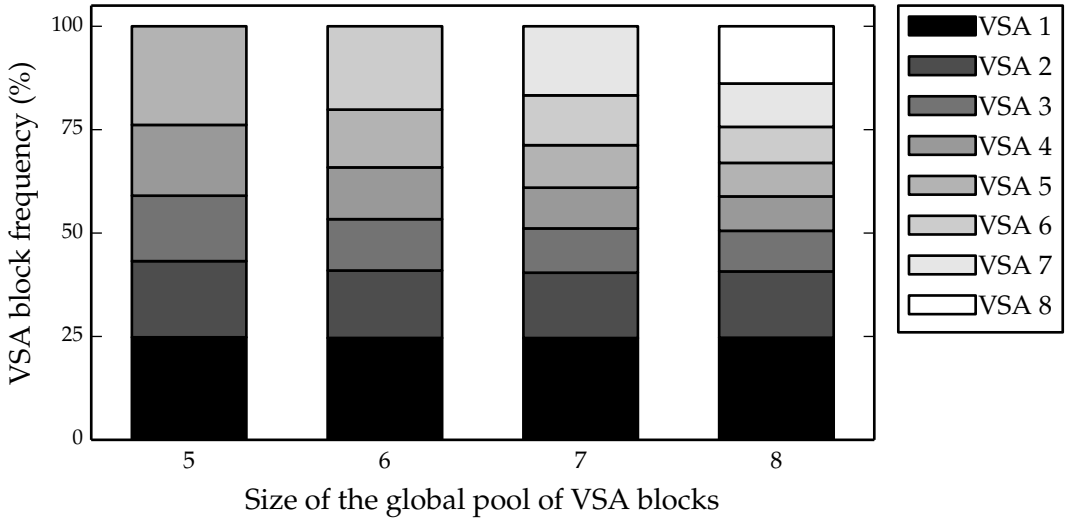



Figure 5.4: FREQUENCY DISTRIBUTIONS OF VSAs FOR INCREASING SIZES OF THE GLOBAL POOL at endemic equilibrium. Increasing the size of the global pool leaves the frequency of high-dominance VSAs unchanged, and diversity accumulates among the low-dominance VSAs. The frequency distribution for a global pool of 7 blocks is also shown in Figure 5.3 (solid red line).

similar VSAs, and a much larger set of serologically diverse VSAs each with low prevalence.

## 5.4 DISCUSSION

Immunity to malaria following natural exposure to *P. falciparum* is developed over several to many years, and sterile immunity is probably never achieved. However, immunity to severe disease develops much faster than protection from uncomplicated disease and asymptomatic parasitaemia (GUPTA ET AL., 1999). It has been suggested that this epidemiological pattern is due to the importance of VSA-specific immunity for clinical protection, to the non-random order in which immunity to specific VSAs is acquired, and to the association between particular VSAs with specific disease syndromes (HVIID, 2005, 2004).

We present a mathematical model which implements the interplay between two opposing selection pressures; one that favors virulent (high-dominance) VSAs in non-immune hosts, the other facilitating non-virulent (low-dominance) VSAs that allow chronic infections in individuals with substantial VSA-specific immunity. Our results complement previous models that have addressed expression patterns (RECKER ET AL., 2004; GATTON ET AL., 2003) and acquisition of immunity (RECKER ET AL., 2008).

We introduce the concept of dominance blocks to describe the competitive interactions among the different VSAs of a single parasite (intraclonal variation) and among the VSAs of different parasite clones (inter-clonal diversity). The model dynamics of VSA expression, within and between hosts, all follow from the dominance hierarchy. The preferential expression of high-dominance VSAs in naive hosts leads to host immunity and quickly exhaust the pool of susceptible hosts available to them. Low-dominance VSAs, on the other hand,

allow persistent infections and thereby ensure efficient transmission from host to host. The ability of parasites encoding high-dominance VSAs to superinfect, favors their conservation. The output of our model is compatible with field observations.

The existence of different levels of immune selection acting on VSAs has been suggested through the application of network approaches to serological data (BUCKEE ET AL., 2009). It is reinforced here with a model that is the first to combine intra-host competition and inter-host transmission to investigate the combined effects of selection at multiple levels. The model suggests a mechanism for the observed structuring of VSA into distinct clusters, reproducing important features of serological observations from the field. Although the results are general, it would be interesting to investigate how they might be modulated by transmission intensity and crossimmunity.

Molecular studies have shown that var genes, previously classified into five major groups (A–E), could be organized into two broad clusters (BULL ET AL., 2008; TRIMNELL ET AL., 2006). A relatively conserved cluster consists of restricted subsets of structurally related variants transcribed by parasites obtained from individuals with limited or no immunity preferentially transcribe (LAVSTSEN ET AL., 2005; WARIMWE ET AL., 2009), and parasites selected *in vitro* for reactivity of IEs with IgG from children with limited immunity (JENSEN ET AL., 2004). Transcription of two of these subsets (Group A and Group B/A) has repeatedly been associated with severe disease (KIRCHGATTER AND DEL A PORTILLO, 2002; KAESTLI ET AL., 2006; KYRIACOU ET AL., 2006; ROTTMANN ET AL., 2006; NORMARK ET AL., 2007). A much more diverse cluster contains Group C which has been largely associated with asymptomatic infections (KAESTLI ET AL., 2006).

In summary, we present a model that identifies the mechanisms that might be driving the evolution of separate clusters of VSAs, as seen for the var gene subfamilies of *P. falciparum*. The hypothesis predicts a restricted subset of high-dominance VSAs associated with severe malaria, and genetically and immunologically diverse low-dominance VSAs related to uncomplicated and asymptomatic infection.

## 5.5 MATHEMATICAL IMPLEMENTATION

The model, constructed as a system of ordinary differential equations, integrates dynamics at two levels (pathogen competition at the individual level as immunological memory accumulates, and pathogen transmission at the population level) in a form that is inevitably dense. In the interest of clarity we construct the model in a stepwise manner.

PARASITES WITH A SINGLE DOMINANCE BLOCK We write a first version where each parasite in characterized by only one dominance block, and then generalize for multiple blocks. Consider a pathogen population comprising a diversity of VSA blocks, indexed by the set $N = 1, 2, \ldots, n$, ordered by inverse dominance. Hosts are classified into uninfected ($S$) or infected ($I$) and by the subset of blocks to which they have immunological memory ($h$). Infected

hosts are further classified by the blocks that they are currently infected with ($p$). This system is written as

$$\begin{aligned}\frac{\mathrm{d}S^h}{\mathrm{d}t} &= \sum_{h=\tilde{h}+q_1} \gamma I^{\tilde{h}}_{q_1} - \sum_{q_2 \notin h} \lambda_{q_2} S^h + \mu\delta_{h=\emptyset} \\ &\quad - \mu S^h + W(S^h), \qquad \text{for all } h \subset N \\ \frac{\mathrm{d}I^h_p}{\mathrm{d}t} &= \lambda_p S^h - (\gamma+\mu) I^h_p + W(I^h_p) \\ &\quad + X(I^h_p), \qquad \text{for all } h \subset N, p \notin h\end{aligned}$$

where $q_1$ indicates the VSA block by which the host was previously infected while having immunological memory $\tilde{h}$, $q_2$ indicates the VSA block by which the susceptible host will be infected, $\gamma$ is the rate of recovery from infection, $\mu$ is the rate of birth and death, $\delta_{h=\emptyset}$ is a delta function indicating that individuals have no immunity at birth, $\lambda_p = \sum_{p \notin h} \beta I^h_p$ is the force of infection of variant $p$, and $W$ and $X$ are functions that determine waning immunity (5.1) and superinfection (5.2), respectively. For simplicity of notation, we write $h + p$, instead of $h + \{p\}$, even though $h$ is a set and $p$ is an element.

Waning immunity is implemented as

$$\begin{aligned}W(S^h) &= \sum_{q_1 \notin h} \alpha S^{h+q_1} - \sum_{q_2 \in h} \alpha S^h \\ W(I^h_p) &= \sum_{q_1 \notin h} \alpha I^{h+q_1}_p - \sum_{q_2 \in h} \alpha I^h_p\end{aligned} \tag{5.1}$$

where $q_1$ indicates the VSA block for which the host previously had immunity, and $q_2$ indicates the VSA block for which the host is losing immunity. Superinfection is implemented such that hosts currently infected by a block, $p$, can be superinfected by a higher dominance block, $q_2$, with a force of infection $\sigma_{q_2 p}$. By superinfection, we mean that hosts become infectious with the new block

while the old variant is cleared and added to the repertoire of immunological memory. This is formalized as

$$X(I_p^h) = \sum_{\substack{h=\tilde{h}+q_1 \\ p\notin\tilde{h}}} \sigma_{pq_1}\lambda_p I_{q_1}^{\tilde{h}} - \sum_{q_2\notin h} \sigma_{q_2p}\lambda_{q_2} I_p^h \tag{5.2}$$

where $q_1$ indicates the VSA block by which an infected host was previously infected while having immunological memory $\tilde{h}$. The coefficient $\sigma_{q_2p}$ (and equivalently $\sigma_{pq_1}$) is defined such that the rate of superinfection increases with the difference in dominance between blocks (see Figure 5.1B)

$$\sigma_{q_2p} = \begin{cases} 1/q_2 - 1/p, & \text{if } q_2 < p \\ 0, & \text{otherwise.} \end{cases} \tag{5.3}$$

PARASITES WITH MULTIPLE DOMINANCE BLOCKS The system is readily generalizable to a scheme where each parasite is characterized by multiple VSA blocks. Consider a parasite, $p = (v_1, v_2, \ldots, v_m)$, characterized by $m$ blocks drawn from a pool $N = (1, 2, \ldots, n)$, and ordered by inverse dominance. Upon infection by a parasite, $p$, a host with immunological memory, $h$, will express the most dominant block for which the host does not have immunity, $v = \min(p \setminus h)$. The system is written as

$$\begin{aligned} \frac{\mathrm{d}S^h}{\mathrm{d}t} = & \sum_{\substack{h=\tilde{h}+v_1 \\ v_1=\min(q_1,\tilde{h})}} \gamma I_{q_1}^{\tilde{h}} - \sum_{q_2\notin h} \lambda_{q_2} S^h + \mu\delta_{h=\emptyset} \\ & - \mu S^h + W(S^h), \qquad \text{for all } h \subset N \\ \frac{\mathrm{d}I_p^h}{\mathrm{d}t} = & \lambda_p S^h - (\gamma+\mu) I_p^h + W(I_p^h) \\ & + X(I_p^h), \qquad \text{for all } h \subset N, p \not\subset h \end{aligned}$$

where $q_1$ indicates a parasite expressing VSA block $v_1$ by which the susceptible host was previously infected while having immunological memory $\tilde{h}$, $q_2$ indicates the parasite by which a susceptible host will be infected, and $\lambda_p = \sum_{p \not\subset h} \beta I_p^h$ is the force of infection of parasite $p$. Waning immunity is implemented as

$$W(S^h) = \sum_{v_1 \notin h} \alpha S^{h+v_1} - \sum_{v_2 \in h} \alpha S^h$$
$$W(I_p^h) = \sum_{v_1 \notin h, v_1 \notin p} \alpha I_p^{h+v_1} - \sum_{v_2 \in h, v_2 \notin p} \alpha I_p^h$$

where $v_1$ indicates the VSA block for which the host previously had immunity, and $v_2$ indicates the VSA block for which the host is losing immunity.

Superinfection is determined by

$$X(I_p^h) = \sum_{\substack{h=\tilde{h}+v_1 \\ v=\min(p \setminus h) \\ v_1=\min(q_1 \setminus \tilde{h})}} \sigma_{vv_1} \lambda_p I_{q_1}^{\tilde{h}} - \sum_{\substack{q_2 \not\subset h \\ v=\min(p \setminus h) \\ v_2=\min(q_2 \setminus (h+v))}} \sigma_{v_2 v} \lambda_{q_2} I_p^h$$

where $q_1$ indicates the parasite expressing VSA block $v_1$ by which the host was previously infected while having immunological memory $\tilde{h}$, and $q_2$ indicates the parasite expressing VSA block $v_2$ by which the host will be superinfected, The coefficients $\sigma$ are as in equation (5.3).

The principal steps in this process are represented diagrammatically in Figure 5.5. The parameters describing the rates of transition between compartments take values in accordance with previous studies (AGUAS ET AL., 2008):

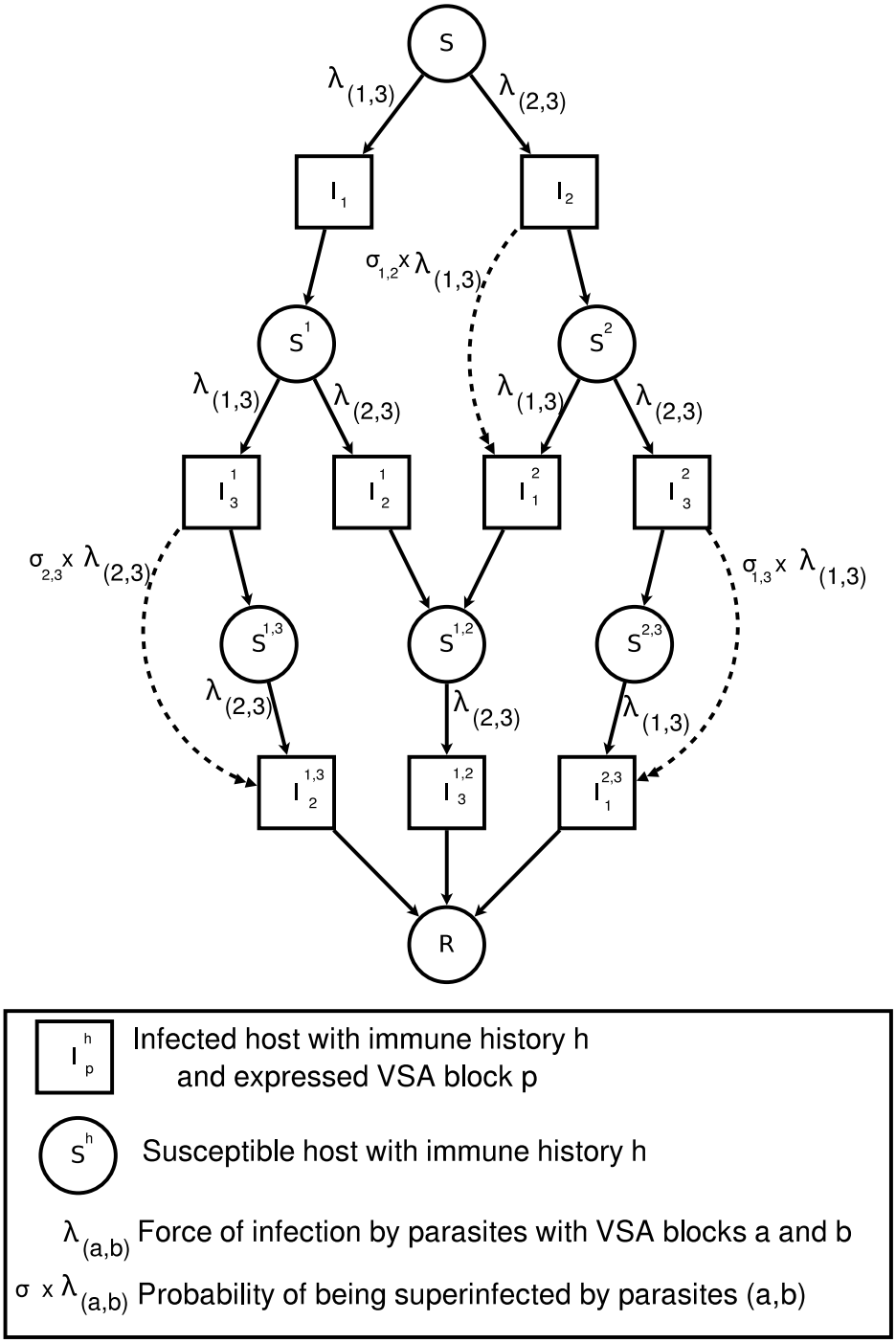



Figure 5.5: Flow diagram for two parasites: (1,3) and (2,3).

birth and death ($\mu = 1/50$); recovery from infection ($\gamma = 6$); loss of immunity ($\alpha = 0.8$); transmission ($\beta = 10$). The time unit is one year.

ACKNOWLEDGMENTS

We thank Jorge Carneiro and Isabel Gordo for valuable comments.

AUTHOR CONTRIBUTIONS

Sander van Noort developed the model, analyzed the data, and wrote the paper. Marta Nunes, Gareth Weedall, and Lars Hviid revised the paper. Gabriela Gomes contributed to the development of the model and revised the paper.

## 5.6 REFERENCES

Aguas R, White LJ, Snow RW, and Gomes MGM (2008). Prospects for malaria eradication in sub-saharan africa. *PLoS One*, 3(3):e1767.

Barry AE, Leliwa-Sytek A, Tavul L, Imrie H, Migot-Nabias F, et al. (2007). Population genomics of the immune evasion (var) genes of plasmodium falciparum. *PLoS Pathog*, 3(3):e34.

Buckee CO, Bull PC, and Gupta S (2009). Inferring malaria parasite population structure from serological networks. *Proc Biol Sci*, 276(1656):477–485.

Bull PC, Buckee CO, Kyes S, Kortok MM, Thathy V, et al. (2008). Plasmodium falciparum antigenic variation. mapping mosaic var gene sequences onto a network of shared, highly polymorphic sequence blocks. *Mol Microbiol*, 68(6):1519–1534.

Bull PC, Kortok M, Kai O, Ndungu F, Ross A, et al. (2000). Plasmodium falciparum-infected erythrocytes: agglutination by diverse kenyan plasma is associated with severe disease and young host age. *J Infect Dis*, 182(1):252–259.

BULL PC, LOWE BS, KORTOK M, and MARSH K (1999). Antibody recognition of plasmodium falciparum erythrocyte surface antigens in kenya: evidence for rare and prevalent variants. *Infect Immun*, 67(2):733–739.

BULL PC, LOWE BS, KORTOK M, MOLYNEUX CS, NEWBOLD CI, ET AL. (1998). Parasite antigens on the infected red cell surface are targets for naturally acquired immunity to malaria. *Nat Med*, 4(3):358–360.

BULL PC, PAIN A, NDUNGU FM, KINYANJUI SM, ROBERTS DJ, ET AL. (2005). Plasmodium falciparum antigenic variation: relationships between in vivo selection, acquired antibody response, and disease severity. *J Infect Dis*, 192(6):1119–1126.

COHEN S, MCGREGOR IA, and CARRINGTON S (1961). Gamma-globulin and acquired immunity to human malaria. *Nature*, 192:733–737.

COLLINS WE and JEFFERY GM (1999). A retrospective examination of the patterns of recrudescence in patients infected with plasmodium falciparum. *Am J Trop Med Hyg*, 61(1 Suppl):44–48.

DOOLAN DL, DOBAÑO C, and BAIRD JK (2009). Acquired immunity to malaria. *Clin Microbiol Rev*, 22(1):13–36, Table of Contents.

GARDNER MJ, HALL N, FUNG E, WHITE O, BERRIMAN M, ET AL. (2002). Genome sequence of the human malaria parasite plasmodium falciparum. *Nature*, 419(6906):498–511.

GATTON ML and CHENG Q (2004). Modeling the development of acquired clinical immunity to plasmodium falciparum malaria. *Infect Immun*, 72(11):6538–6545.

GATTON ML, PETERS JM, FOWLER EV, and CHENG Q (2003). Switching rates of plasmodium falciparum var genes: faster than we thought? *Trends Parasitol*, 19(5):202–208.

GIHA HA, STAALSOE T, DODOO D, ELHASSAN IM, ROPER C, ET AL. (1999). Nine-year longitudinal study of antibodies to variant antigens on the surface of plasmodium falciparum-infected erythrocytes. *Infect Immun*, 67(8):4092–4098.

GOG JR and GRENFELL BT (2002). Dynamics and selection of many-strain pathogens. *Proc Natl Acad Sci U S A*, 99(26):17209–17214.

GUPTA S, SNOW RW, DONNELLY CA, MARSH K, and NEWBOLD C (1999). Immunity to non-cerebral severe malaria is acquired after one or two infections. *Nat Med*, 5(3):340–343.

HORROCKS P, PINCHES R, CHRISTODOULOU Z, KYES SA, and NEWBOLD CI (2004). Variable var transition rates underlie antigenic variation in malaria. *Proc Natl Acad Sci U S A*, 101(30):11129–11134.

Hviid L (2004). The immuno-epidemiology of pregnancy-associated plasmodium falciparum malaria: a variant surface antigen-specific perspective. *Parasite Immunol*, 26(11-12):477–486.

Hviid L (2005). Naturally acquired immunity to plasmodium falciparum malaria in africa. *Acta Trop*, 95(3):270–275.

Jensen ATR, Magistrado P, Sharp S, Joergensen L, Lavstsen T, et al. (2004). Plasmodium falciparum associated with severe childhood malaria preferentially expresses pfemp1 encoded by group a var genes. *J Exp Med*, 199(9):1179–1190.

Joergensen L, Turner L, Magistrado P, Dahlbäck MA, Vestergaard LS, et al. (2006). Limited cross-reactivity among domains of the plasmodium falciparum clone 3d7 erythrocyte membrane protein 1 family. *Infect Immun*, 74(12):6778–6784.

Kaestli M, Cockburn IA, Cortés A, Baea K, Rowe JA, et al. (2006). Virulence of malaria is associated with differential expression of plasmodium falciparum var gene subgroups in a case-control study. *J Infect Dis*, 193(11):1567–1574.

Kirchgatter K and del A Portillo H (2002). Association of severe noncerebral plasmodium falciparum malaria in brazil with expressed pfemp1 dbl1 alpha sequences lacking cysteine residues. *Mol Med*, 8(1):16–23.

Krause DR, Gatton ML, Frankland S, Eisen DP, Good MF, et al. (2007). Characterization of the antibody response against plasmodium falciparum erythrocyte membrane protein 1 in human volunteers. *Infect Immun*, 75(12):5967–5973.

Kyes SA, Kraemer SM, and Smith JD (2007). Antigenic variation in plasmodium falciparum: gene organization and regulation of the var multigene family. *Eukaryot Cell*, 6(9):1511–1520.

Kyriacou HM, Stone GN, Challis RJ, Raza A, Lyke KE, et al. (2006). Differential var gene transcription in plasmodium falciparum isolates from patients with cerebral malaria compared to hyperparasitaemia. *Mol Biochem Parasitol*, 150(2):211–218.

Langhorne J, Ndungu FM, Sponaas AM, and Marsh K (2008). Immunity to malaria: more questions than answers. *Nat Immunol*, 9(7):725–732.

Lavstsen T, Magistrado P, Hermsen CC, Salanti A, Jensen ATR, et al. (2005). Expression of plasmodium falciparum erythrocyte membrane protein 1 in experimentally infected humans. *Malar J*, 4(1):21.

MacDonald G (1950). The analysis of malaria parasite rates in infants. *Trop Dis Bull*, 47(10):915–938.

Marsh K and Howard RJ (1986). Antigens induced on erythrocytes by p. falciparum: expression of diverse and conserved determinants. *Science*, 231(4734):150–153.

MILLER LH, BARUCH DI, MARSH K, and DOUMBO OK (2002). The pathogenic basis of malaria. *Nature*, 415(6872):673–679.

NIELSEN MA, GREVSTAD B, A-ELGADIR TME, KURTZHALS JAL, GIHA H, ET AL. (2005). Differential induction of immunoglobulin g to plasmodium falciparum variant surface antigens during the transmission season in daraweesh, sudan. *J Infect Dis*, 192(3):520–527.

NIELSEN MA, STAALSOE T, KURTZHALS JAL, GOKA BQ, DODOO D, ET AL. (2002). Plasmodium falciparum variant surface antigen expression varies between isolates causing severe and nonsevere malaria and is modified by acquired immunity. *J Immunol*, 168(7):3444–3450.

NORMARK J, NILSSON D, RIBACKE U, WINTER G, MOLL K, ET AL. (2007). Pfemp1-dbl1alpha amino acid motifs in severe disease states of plasmodium falciparum malaria. *Proc Natl Acad Sci U S A*, 104(40):15835–15840.

OFORI MF, DODOO D, STAALSOE T, KURTZHALS JAL, KORAM K, ET AL. (2002). Malaria-induced acquisition of antibodies to plasmodium falciparum variant surface antigens. *Infect Immun*, 70(6):2982–2988.

O'NEIL-DUNNE I, ACHUR RN, AGBOR-ENOH ST, VALIYAVEETTIL M, NAIK RS, ET AL. (2001). Gravidity-dependent production of antibodies that inhibit binding of plasmodium falciparum-infected erythrocytes to placental chondroitin sulfate proteoglycan during pregnancy. *Infect Immun*, 69(12):7487–7492.

PETERS J, FOWLER E, GATTON M, CHEN N, SAUL A, ET AL. (2002). High diversity and rapid changeover of expressed var genes during the acute phase of plasmodium falciparum infections in human volunteers. *Proc Natl Acad Sci U S A*, 99(16):10689–10694.

PHIRI H, MONTGOMERY J, MOLYNEUX M, and CRAIG A (2009). Competitive endothelial adhesion between plasmodium falciparum isolates under physiological flow conditions. *Malar J*, 8:214.

RECKER M, ARINAMINPATHY N, and BUCKEE CO (2008). The effects of a partitioned var gene repertoire of plasmodium falciparum on antigenic diversity and the acquisition of clinical immunity. *Malar J*, 7:18.

RECKER M, NEE S, BULL PC, KINYANJUI S, MARSH K, ET AL. (2004). Transient cross-reactive immune responses can orchestrate antigenic variation in malaria. *Nature*, 429(6991):555–558.

ROBERTS DJ, CRAIG AG, BERENDT AR, PINCHES R, NASH G, ET AL. (1992). Rapid switching to multiple antigenic and adhesive phenotypes in malaria. *Nature*, 357(6380):689–692.

ROTTMANN M, LAVSTSEN T, MUGASA JP, KAESTLI M, JENSEN ATR, ET AL. (2006). Differential expression of var gene groups is associated with morbidity caused by plasmodium falciparum infection in tanzanian children. *Infect Immun*, 74(7):3904–3911.

SAMA W, KILLEEN G, and SMITH T (2004). Estimating the duration of plasmodium falciparum infection from trials of indoor residual spraying. *Am J Trop Med Hyg*, 70(6):625–634.

SCHERF A, LOPEZ-RUBIO JJ, and RIVIERE L (2008). Antigenic variation in plasmodium falciparum. *Annu Rev Microbiol*, 62:445–470.

STAALSOE T, HAMAD AA, HVIID L, ELHASSAN IM, ARNOT DE, ET AL. (2002). In vivo switching between variant surface antigens in human plasmodium falciparum infection. *J Infect Dis*, 186(5):719–722.

STAALSOE T, MEGNEKOU R, FIEVÉT N, RICKE CH, ZORNIG HD, ET AL. (2001). Acquisition and decay of antibodies to pregnancy-associated variant antigens on the surface of plasmodium falciparum-infected erythrocytes that protect against placental parasitemia. *J Infect Dis*, 184(5):618–626.

TRIMNELL AR, KRAEMER SM, MUKHERJEE S, PHIPPARD DJ, JANES JH, ET AL. (2006). Global genetic diversity and evolution of var genes associated with placental and severe childhood malaria. *Mol Biochem Parasitol*, 148(2):169–180.

WARIMWE GM, KEANE TM, FEGAN G, MUSYOKI JN, NEWTON CRJC, ET AL. (2009). Plasmodium falciparum var gene expression is modified by host immunity. *Proc Natl Acad Sci U S A*, 106(51):21801–21806.

# 6

# DISCUSSION

*When things get too complicated, it sometimes makes sense to stop and wonder: Have I asked the right question?*

Enrico Bombieri

## 6.1 SYNDROMIC SURVEILLANCE: METHODS AND STATISTICS

### 6.1.1 *Influenzanet and traditional surveillance in Europe*

Influenzanet and ECDC run independent syndromic surveillance systems, which report the activity of influenza-like illness (ILI) based on a clinical case definition. ILI activity is traditionally being determined based on the rate of health care seeking patients diagnosed with ILI. This data is collected via national networks of sentinel general practitioners (GPs), currently coordinated in Europe by the European Influenzanet Surveillance Network (EISN) within the ECDC structure. Influenzanet runs national websites, where cohorts of volunteers weekly report their symptoms (Chapters 2 and 3) and the ILI activity is based on the proportion of participants whose symptoms fit the ILI case definition. Most assertions done in this section on Influenzanet would apply to any self-reporting surveillance system, whereas most assertions for ECDC hold for any syndromic surveillance system which depends on health care seeking. The ILI activity calculation in both systems is described by:

$$\text{Influenzanet:} \quad \frac{\text{Participants whose self-reported symptoms fit ILI case definition}}{\text{Active participants in Influenzanet}}$$

$$\text{ECDC:} \quad \frac{\text{Patients visiting GP and diagnosed with ILI}}{\text{Patients registered to sentinel GP}}$$

An important difference between the two systems concerns the bias introduced in determining the ILI activity. For Influenzanet a selection bias could be present in the active participants: people who decide to participate in Influenzanet might not be representative for the whole population. Since the sentinel GPs are generally selected to have a representative registration of patients, such a selection bias is less prevalent in ECDC.

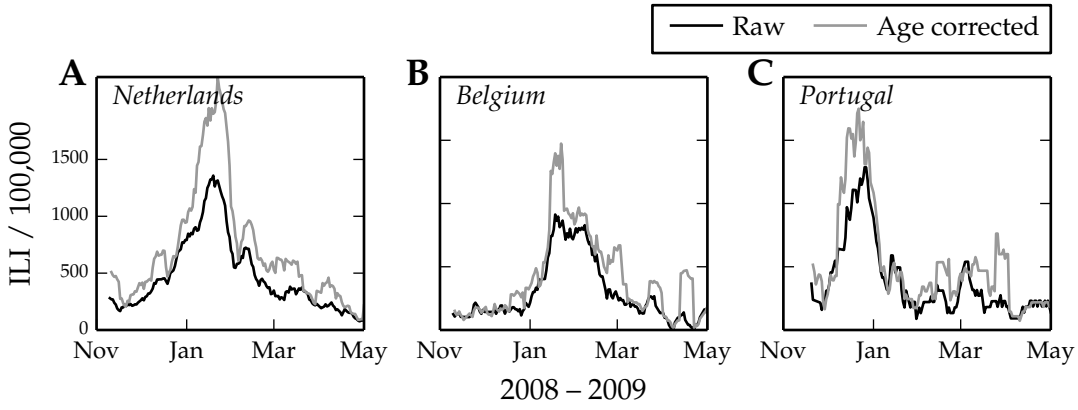



Figure 6.1: AGE CORRECTED ILI INCIDENCE compared to raw ILI incidence in A) Netherlands, B) Belgium and C) Portugal (2008–2009). The ILI incidence by Influenzanet is defined as the number of ILI onsets per 100,000 participant-weeks.

For ECDC a reporting bias could be present in the people who decide to seek medical care if they suffer from ILI symptoms, since they might not be representative for a typical person with ILI. Since in Influenzanet all active participants report weekly, such a reporting bias is less prevalent in Influenzanet.

BIASES IN INFLUENZANET Based on the intake questionnaire, some selection biases in the Influenzanet participants can be detected. For some characteristics it has been shown that the Influenzanet participants are representative, such as the presence of asthma and diabetes in Dutch (MARQUET ET AL., 2006) and Belgian (VANDENDIJCK ET AL., 2013) participants. With respect to characteristics for which the Influenzanet participants are not representative for the whole population, a corrected ILI incidence can be determined.

For most countries the younger and older age groups are underrepresented (Figure 2.1, Chapter 2, page 21), and an age-corrected ILI incidence can be calculated by:

$$\text{ILI (corrected)} = \frac{\sum_{\text{age groups}} \left( \text{ILI}_{\text{age group}} \times \text{Population}_{\text{age group}} \right)}{\text{Population}}$$

Figure 6.1 shows the raw and age-corrected ILI incidences for the Netherlands, Belgium and Portugal in season 2008–2009.

Some selection biases might be hard to detect and therefore hard to correct. For example, Influenzanet might attract persons who are more susceptible for flu-like symptoms than an average person, a so-called hidden bias (Rosenbaum, 1991). Since participation is voluntary, active recruitment strategies targeting underrepresented groups (Rehn et al., 2014) does not prevent the presence of these hidden selection biases.

In contrast with most traditional surveillance systems, a total increase in participants does only marginally increase the costs of running the system. Moreover, having more participants does facilitate the recruiting of new participants due to increased media attention.

BIASES IN ECDC Influenzanet data shows that across Europe only 25–75% of people with ILI actually seek medical care (Figure 3.3, Chapter 3, page 44), such that the ILI activity as determined by ECDC is structurally underestimated. Note that the determined percentages depend on the applied ILI case definition. More critically, there could be a reporting bias in those people with ILI who decide to seek medical care. Since people wait various days before a possible visit to a medical doctor (Figure 3.4, Chapter 3, page 45), persons who are sick for a longer time might be overrepresented. Furthermore, changes in

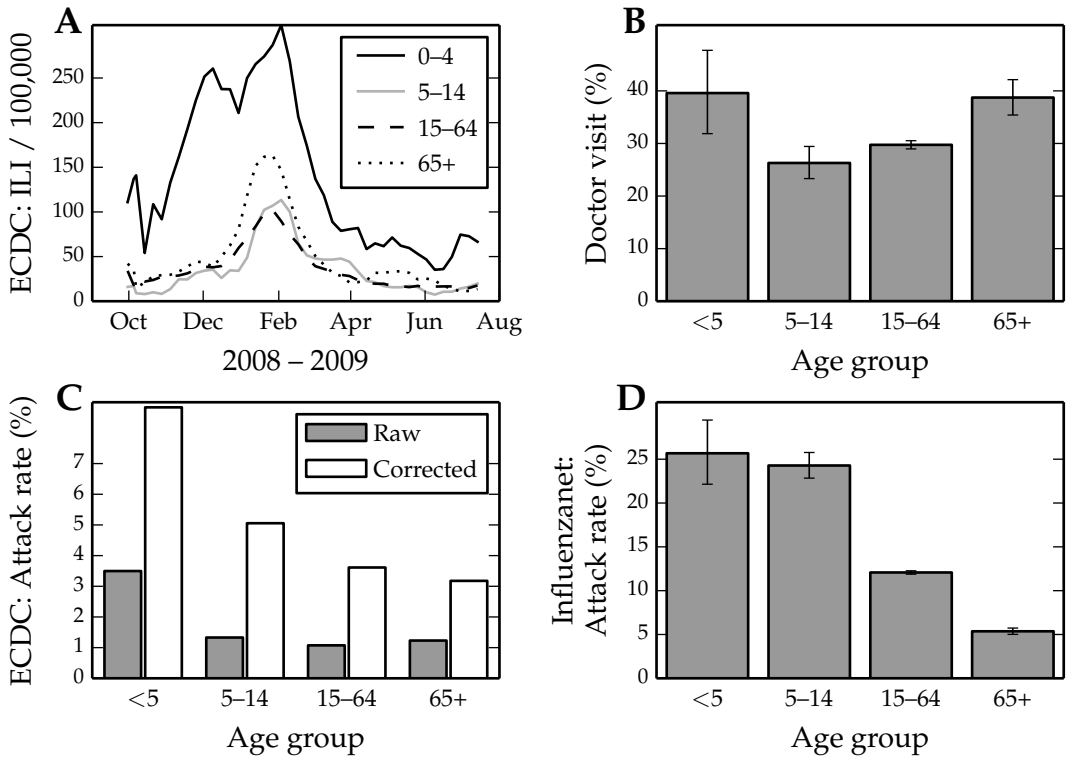



Figure 6.2: ILI ACTIVITY BY AGE GROUP ACCORDING TO ECDC in the Netherlands (2008–2009). The ILI incidence by ECDC is defined as the weekly number of patient visits with ILI per 100,000. A) ILI activity by age group. B) Percentage of Influenzanet participants with ILI who sought medical care by age group. C) Raw and corrected ILI attack rate by age group. D) ILI attack rate according to Influenzanet data.

Table 6.1: PARTICIPATION RATE IN INFLUENZANET for various countries (2003–2013). The number of participants is the average over all seasons the system was active. Active participants completed at least 3 symptoms questionnaires. The participation rate is based on the number of active participants.

| COUNTRY | POPULATION | PARTICIPANTS | ACTIVE | PARTICIPATION |
|---|---|---|---|---|
| Netherlands | 17M | 14121 | 12397 | 0.074% |
| Belgium (Flanders) | 6M | 4127 | 3788 | 0.061% |
| Portugal | 10M | 1527 | 1246 | 0.012% |
| Italy | 59M | 2071 | 1349 | 0.002% |
| United Kingdom | 64M | 5909 | 3911 | 0.006% |
| Sweden | 10M | 2954 | 2411 | 0.025% |
| France | 66M | 6000 | 4754 | 0.007% |
| Spain | 46M | 632 | 476 | 0.001% |

medical care seeking rate over time could disturb the determined ILI activity. The impact of such reporting biases could be especially significant in countries where fewer people seek medical care.

According to the ILI activity as reported by ECDC, Dutch people over 65 years old suffer slightly more ILI than participants 5–14 and 15–64 years old (Figure 6.2A). However, according to Influenzanet data, participants over 65 years old with ILI are much more likely to seek medical care (Figure 6.2B). A corrected ILI attack rate for each age group can be estimated by dividing the ILI attack rate by the medical care seeking rate. According to the corrected ILI attack rate, people over 65 years old actually have the lowest attack rate from all age groups (Figure 6.2C), consistent with Influenzanet data (Figure 6.2D).

#### 6.1.1.1 *Advantages of traditional surveillance*

1. COHORT SIZE The sentinel GPs aim to represent 1–5% of the physicians in a country,[1] although in 2005–2006 the population under surveillance in the Netherlands (0.7%), Belgium (0.4%), and Portugal (0.7%) is significantly lower (VAN NOORT ET AL., 2007). In Influenzanet, the country with the highest participation rate only monitors 0.07% of the population, with significantly lower participation rates in some other countries (Table 6.1). Even though the actual population under surveillance within the ECDC system should incorporate the medical care seeking rate, the population under surveillance by ECDC is still significantly larger than by Influenzanet. This could be one of the reasons why the ILI activity as determined by ECDC has less statistical noise than the ILI incidence by Influenzanet (Figure 3.1, Chapter 3, page 42).

2. HISTORIC DATA AND CONTINUITY The national sentinel networks which report to ECDC have been in operation for many decades,[1] whereas the first Influenzanet system has been operational since 2003. Furthermore, for the foreseeable future many people will continue to visit a medical doctor when they are sick, so provided enough funding a system based on sentinel doctors will continue to function. Although Influenzanet participants have proven to be quite loyal during the first 10 years of the system, with around 75% of participants remaining in a following season (Figure 3.5, Chapter 3, page 47), each season an effort is required to recruit new participants. Furthermore, low participation rate in some countries (Table 6.1) indicates that it may be difficult to obtain a high participation rate. Funding for Influenzanet is still vital for the maintenance, in particular to keep participants active by

1 http://ecdc.europa.eu/en/activities/surveillance/EISN/surveillance/Pages/sentinel_surveillance.aspx

sending them weekly newsletters, and recruiting new participants by continuing to provide the media with the latest news and results, or other public initiatives.

3. VIRAL CONFIRMATION In most countries sentinel GPs are requested to take a swab of some of their patients with ILI, which are tested for the presence of the influenza virus. Based on the detected strains worldwide, the WHO decides each year the composition of the influenza vaccine. Furthermore, the results of these laboratory tests are usually considered in determining whether the detected ILI incidence can be considered the start of the influenza epidemic. Recent studies have shown that self-sampling by lay people and shipment of nasal swabs via regular mail is feasible (COOPER ET AL., 2008; ELLIOT ET AL., 2009). During the 2011–2012 season, Swedish partners of Influenzanet performed a pilot study in which people could self-sample with nasal swabs, which were then sent to laboratory for viral confirmation (Plymoth et al., unpublished).

#### 6.1.1.2 *Advantages of Influenzanet*

1. ILI ONSET DAY RECORDED For Influenzanet not only the day a participant completed the symptoms questionnaire is recorded, but each participant also reports on which day the symptoms and/or fever started. The ILI incidence for Influenzanet is usually determined based on the day the fever started, whereas ECDC determines the ILI incidence based on the week a person visited the GP. Since people usually wait a few days before seeking medical care (Figure 3.4, Chapter 3, page 45), the ILI incidence by Influenzanet better resembles the actual ILI activity in the population. Although this does not necessarily indicate that Influenzanet is faster in real-time reporting

(Section 6.3.3), it can be important when linking ILI activity with weather variables (Chapter 4) or school closures (EAMES ET AL., 2012).

2. AUTOMATED The Influenzanet system operates in a relatively automatic manner. Participants are weekly sent a reminder, usually accompanied by a newsletter, and report their symptoms online. All questionnaire data are stored in a central database and the latest results are generated and published daily.[2]

For ECDC, sentinel GPs send their latest data once a week to a national coordinator, who reports the latest (provisional) results a few days later to the ECDC. This delay currently allows Influenzanet, — in countries with sufficient participants — to detect the yearly onset of the ILI epidemic up to two weeks earlier than ECDC (see Section 6.3.3). However, this reporting delay is not inherent to the ECDC methodology, since in principle a system could be developed for sentinel GPs to report their data in real-time (LANGE AND SCHÖTTLER, 2002; CARRAT ET AL., 1998).

3. FLEXIBLE Modifications in the Influenzanet system, for example the inclusion of an extra question or symptom, requires only a change on the central server, after which all participants will be presented with the updated questionnaires. Since 2011 all European systems use the same platform (PAOLOTTI ET AL., 2014), which facilitates the application of European-wide changes.

For ECDC it is much harder to implement (European-wide) changes. Although an objective of the ECDC is to unify all sentinel GP systems, progress is slow. The most noticeable difference between countries is the application of different ILI (or ARI, acute respiratory illness) case definitions (AGUILERA

2 http://www.influenzanet.eu/results

ET AL., 2003). It is understandable that countries would be reluctant to implement changes, to avoid discontinuities in their historic data.

4. CONSISTENT ACROSS COUNTRIES Due to the application of single ILI case definition across countries, Influenzanet is able to measure the ILI incidence rate consistently across multiple countries (Figure 2.3A, Chapter 2, page 25). In contrast, the ILI incidence as reported by ECDC may vary significantly by country (Figure 2.3B, Chapter 2, page 25), due to the different rates at which people seek medical care (Figure 3.3, Chapter 3, page 44), and the different national ILI case definitions (Table 2.3, Chapter 2, page 20).

5. RISK FACTORS AND VACCINE EFFECTS The intake questionnaire which is completed by each participant at the beginning of every season, contains various demographic, medical, socioeconomic and lifestyle questions. Together with the determined ILI attack rate, Influenzanet can estimated risk factors for ILI (Table 3.2, Chapter 3, page 49). The determined risk factors are in correspondence with those found in the literature, but also allow for the assessment of risk factors not previously identified, such as having pets (dogs or cats) at home. Furthermore, based on the vaccination status a season-dependent vaccine effectiveness for ILI can be determined (Table 3.2, Chapter 3, page 49, and Section 6.3.2).

6. UNAFFECTED BY MEDIA HYPE A common misconception about Influenzanet is that an increase in media interest for influenza would lead to more ILI reports and distort the results. However, since Influenzanet determines the ILI incidence among all its active participants, an increase in participation would increase both the numerator and denominator for the ILI

incidence. Although people who have ILI are more likely to register on Influenzanet (Figure 2.4, Chapter 2, page 27), this selection bias is removed in the analyses by excluding all symptoms a participant had upon joining Influenzanet. Furthermore, if necessary all participants who register during the season could be completely excluded from the analyses, removing this selection bias completely.

In contrast, if people who do not normally seek medical care for ILI symptoms would decide to do so due to increased media attention, the ILI incidence as reported by ECDC would be artificially increased. During the spring and summer of 2009 in Wales, a significant increase in ILI rate was reported by the sentinel GPs, whereas the percentage of people who were tested positive for influenza was relatively low. The concern about the new H1N1pdm influenza strain might have been largely responsible for phenomenon (KERAMAROU ET AL., 2011).

7. MONITOR BEHAVIOR Influenzanet participants not only report their symptoms, but in case of symptoms they also receive some follow-up questions. Influenzanet has determined the rate of participants with ILI who seek medical care (Figure 3.3, Chapter 3, page 44) and when (Figure 3.4, Chapter 3, page 45), but also monitors when they stay at home, for how many days, and whether they used any over-the-counter drugs.

8. MONITOR OTHER DISEASES Based on the self-reported symptoms and a syndromic case definition, other syndromes besides ILI could be monitored. Since 2011 Influenzanet monitors in all countries ILI, gastroenteritis, common cold, and hay fever. In Salvador, Brazil, a system based on the original Portuguese Gripenet platform was implemented to monitor dengue

(denguenaweb.br). Although in principle any syndrome could be monitored, care should be taken in interpreting data from diseases with high incidence. Anecdotal evidence suggests that participants who regularly have the same symptoms, such as a headache, might stop reporting them to avoid getting the weekly follow-up questions.

9. EXTRA QUESTIONNAIRES The way in which the Influenzanet system is constructed — a website where volunteers can register and complete (weekly) questionnaires — allows the collection of data that can extend far beyond the collection of weekly ILI cases. An extended questionnaire released in the Netherlands (2004–2005) revealed significant trends between stress/personality and ILI self-reporting (SMOLDEREN ET AL., 2007). Influenzanet in the United Kingdom includes a questionnaire about inter-personal contacts which, included in a transmission model, inform how changes in contact patterns result in a fall in the reproduction number of influenza during the school holidays (EAMES ET AL., 2012).

Extra questionnaires which are not directly related to ILI surveillance could be included as well. In 2009–2010 in the Netherlands and Belgium a questionnaire was added for participants (and non Influenzanet participants) to report any side effects from their influenza vaccine. Differences in effects were detected between the standard seasonal flu shot, the pandemic vaccine in Belgium (*Pandemrix*) and the pandemic vaccine in the Netherlands (*Focetria*).[3] In Portugal in 2009, a questionnaire was included on pandemic awareness.[4]

3 http://www.degrotegriepmeting.nl/nl/artikelen/groot-aantal-meldingen-bijwerkingen-griepprik/

4 http://www.gripenet.pt/pt/resultados/estudo-pandemia-2009/

### 6.1.2 *Data mining*

Although Influenzanet performs on the Internet and without any input of medical doctors, it should not be confused with data mining. In data mining surveillance, data which has been collected for other purposes, such as Google searches (GINSBERG ET AL., 2009) or Twitter messages (SIGNORINI ET AL., 2011), are transformed to estimated ILI rates. Based on historic data from traditional surveillance systems, algorithms are calibrated to translate the frequency of flu-related search terms or messages to estimated ILI rates. These algorithms are then able to estimate ILI rates based on new data. Since these data sets are usually not medically related, they are often made freely available, leading to the application of novel analytic methods.

Data mining methods such as Google Flu Trends have been shown to be able to detect changes in ILI trends earlier than the traditional surveillance based on sentinel GPs (GINSBERG ET AL., 2009). Several features could enable earlier detection based on data mining. People with ILI (or others around them) might search on Google or post on Twitter before they seek medical care. The traditional surveillance system might also have a longer delay between data collection and publication than a fully automatic algorithm. Finally, the algorithms can be calibrated such that they are more sensitive to early signals in ILI trends (CHRISTAKIS AND FOWLER, 2010).

Extreme caution should be applied to surveillance reports that rely on these secondary data sources (BUTLER, 2013). To monitor influenza activity in the population, the most reliable way would be the confirmation of real influenza infections. However, since this is a costly process, and arguably clinical illness is a more interesting statistic than actual infections, syndromic surveillance is based on clinical symptoms. Both Influenzanet and ECDC assess the activity

Box 6.1: Beware of self-enforcing systems

*A younger inexperienced Indian chief was wondering how much firewood he needed to gather for the winter. He was not like the chiefs in the pas that could tell from the clouds and stuff like that. He decided to make his people gather tons of firewood, more than they usually gather just to be safe. The young chief was still curious though so he decided to call the weather service people. They said that it was supposed to be a pretty cold winter, colder than most years. So the young chief made his people gather more firewood. Again the chief called the weather service and they said that it was suppose to be even colder. So the Indians went back to wood cutting. Once again the chief called, and the weather service said that there may be another ice age. The chief asked him how they could tell all of this and he simply replied, "Because the Indians are gathering firewood like crazy!".*

of influenza-like illness based on clinical symptoms and an ILI case definition. Data mining surveillance adds one more proxy to the surveillance, in this case behavioral, how people react to having (or seeing) ILI-like symptoms, such as searching for information on Google or writing about it on Twitter. Depending on underlying behavioral processes, changes in ILI trends estimated by data mining, may not only be due to changes in the presence of ILI symptoms in the population, but also due to the higher or lower rate at which people search or talk about it (Box 6.1). The algorithms should therefore be regularly re-calibrated, such that data mining methods can only serve as secondary systems.

### 6.1.3 *ECDC, Influenzanet, and data mining in perspective*

The ILI trends as determined by Influenzanet and ECDC are very similar (Figure 3.1, Chapter 3, page 42). Divergence of the ILI trends in certain periods

Table 6.2: DIFFERENCES INFLUENZANET AND ECDC

| | INFLUENZANET | ECDC |
|---|---|---|
| ILI: | Self-reported symptoms | GP diagnosis |
| Bias: | Participation in Influenzanet | Visit GP if sick |
| Advantages: | 1. ILI onset day recorded<br>2. Automated<br>3. Flexible<br>4. Consistent across countries<br>5. Risk factors and vaccine effects<br>6. Unaffected by media hype<br>7. Monitor behavior<br>8. Monitor other diseases<br>9. Extra questionnaires | 1. Size of the cohort<br>2. Historic data and continuity<br>3. Viral confirmation |

might be attributed to differences in biases and methodology. Although divergence of the trends could lead to uncertainty and confusion about the current state of influenza activity in the country, this might actually be an important benefit of having several independent systems. If only one system would be active, such uncertainty would not be known and could lead to a false confidence in the data. The differences and advantages of Influenzanet and ECDC are summarized in Table 6.2.

In summary, ECDC delivers a robust weekly ILI incidence accompanied by an integrated system for viral confirmation. Influenzanet provides an independent ILI surveillance system, which also provides related information on for example behavior, ILI risks, and contact patterns. Influenzanet allows the development of more sophisticated surveillance methods based on the detailed personal and symptoms information on all participants (Section 6.3.4). The often freely available "big data" has led to considerable innovation on data mining surveillance, but the adopted algorithms continue to depend on

regular calibration based on independent systems such as ECDC and Influenzanet.

## 6.2 LINKING MODELS AND DATA

Influenza epidemics invariably occur during winter in temperate climates, and various theories have been put forward to explain this phenomenon. Although no conclusive explanation has been found, studies have pointed at decreased immunity, increased transmission, increased virulence during winter, and changes in contact patterns (FUHRMANN, 2010). More specifically, immunity appears lower in winter due to lower sunlight-induced vitamin D production (CANNELL ET AL., 2006) and increased melatonin secretion (DOWELL, 2001). Transmission appears higher during winter, due to less UV radiation which inactivates influenza viruses (SAGRIPANTI AND LYTLE, 2007), the longer survival time of influenza particles in lower absolute humidity (SHAMAN ET AL., 2010), and increased aerosol transmission at lower temperature (LOWEN ET AL., 2008). Influenza infected individuals may expose more symptoms during winter, due to the colder temperature (ECCLES, 2002), and decreased vitamin D production (CANNELL ET AL., 2006). School closure has been linked to a decrease of infectious contacts between children (CAUCHEMEZ ET AL., 2008), and in most countries children have a long summer holiday.

Surveillance during the summer is usually low, but in the spring and summer of 2009 the virological surveillance increased due to the detection of a new pandemic influenza strain. This revealed that several lineages of seasonal influenza strains continued to circulate for prolonged periods of time, suggesting that unseasonal transmission of influenza A viruses might be more widespread than is usually supposed (GHEDIN ET AL., 2010).

### 6.2.1 *Time-varying ILI factor or transmission*

Most studies have focused on the impact of weather variables on the seasonality of influenza, with British general practitioner Robert Edgar Hope-Simpson stating:

> *The method by which the influence of season is mediated so as to control influenza outbreaks is as yet unidentified, but its importance cannot be doubted. It is concerned in most of the phenomena [concerning influenza evolution and epidemiology] and may indeed provide the key to understanding them*
>
> *(*Hope-Simpson, *1979).*

Weather elements might not only play a role in regulating the seasonality of influenza, but also in the relative severity of the yearly epidemics of influenza-like illness (ILI).

Figure 6.3A,B shows the ILI incidences in the Netherlands for seasons 2003–2013 according to Influenzanet and ECDC. Both systems indicate that the earliest ILI epidemic occurred in the season 2009–2010, characterized by the new pandemic strain (Garten et al., 2009), whereas the second-earliest was in 2003–2004, the season when the Fuijan strain started circulating, a strain characterized by an antigenic cluster jump (Smith et al., 2004). ILI epidemics that occur early in the season seem to be characterized by influenza strains for which the immunity in the population is low. These early ILI epidemics were however not the most severe according to either system.

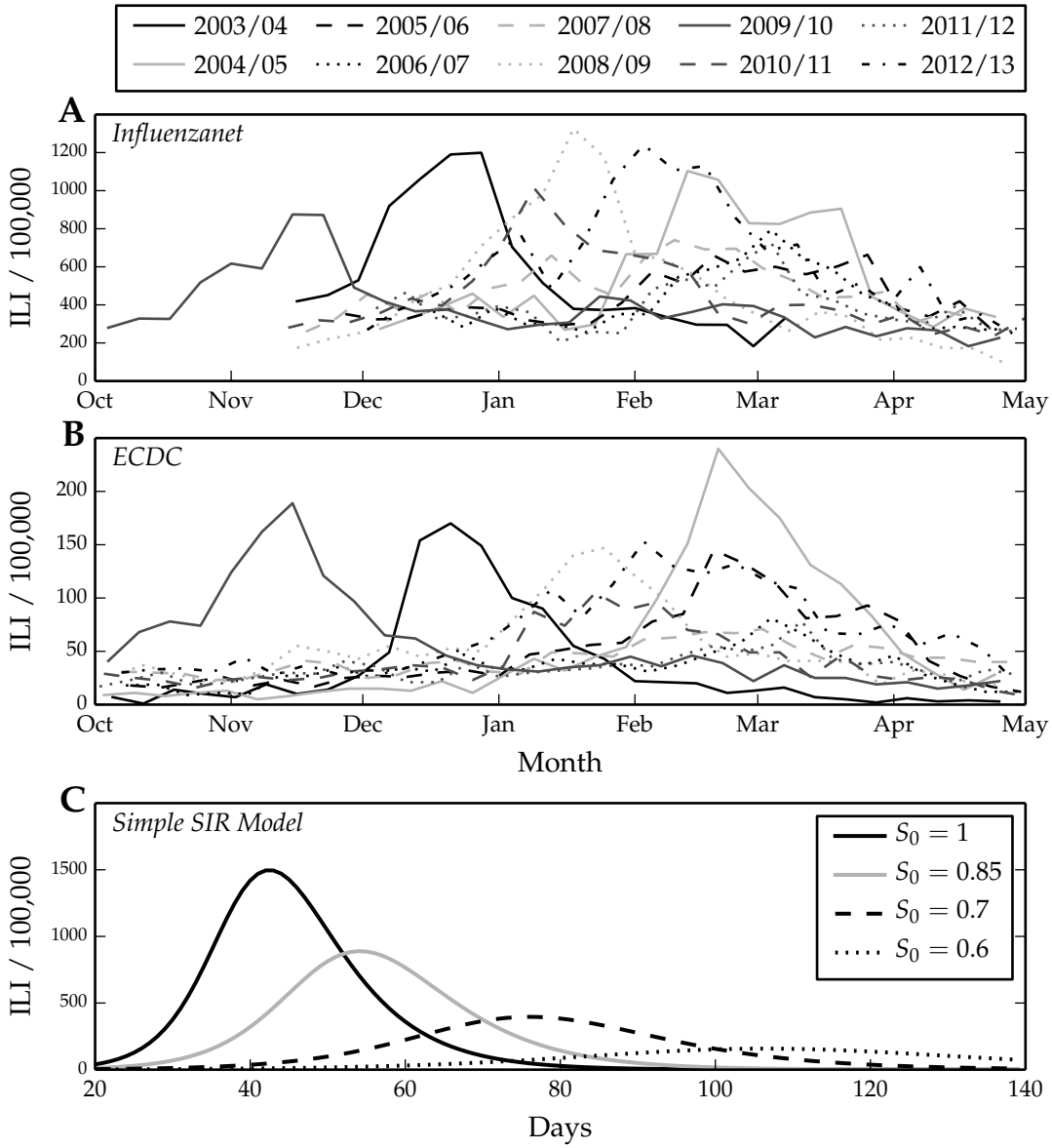



Figure 6.3: ILI INCIDENCE DATA COMPARED TO BASIC SIR MODEL . A) ILI incidence for Influenzanet in the Netherlands (2003–2013). The ILI incidence by Influenzanet is defined as the number of ILI onsets per 100,000 participant-weeks. B) ILI incidence for ECDC in the Netherlands (2003–2013). The ILI incidence by ECDC is defined as the weekly number of patient visits with ILI per 100,000. C) Simple SIR model with various initial fraction of susceptibles ($S_0 \in [1, 0.85, 0.7, 0.6]$), whereas the other parameters are fixed ($\beta = 0.44$, $\tau = 0.2$, $m = 10^{-5}$, $p = 0.08$).

One of the most basic deterministic transmission models for influenza is given by the following SIR system:

$$\begin{aligned}\frac{\mathrm{d}S}{\mathrm{d}t} &= -(\beta I + m)S \\ \frac{\mathrm{d}I}{\mathrm{d}t} &= (\beta I + m)S - \tau I \\ ILI &= p \cdot I\end{aligned}$$

where $S$ denotes the fraction of susceptibles, $I$ the fraction of influenza infected and $ILI$ the fraction of persons with ILI. The parameters are the transmission coefficient $\beta$, the recovery rate $\tau$, and the migration rate $m$. The ILI factor $p$, introduced in Chapter 4, denotes the fraction of persons infected with influenza who have sufficient symptoms to fit the ILI case definition.

Figure 6.3C shows simulations of this model for various initial fractions of susceptibles. A higher initial fraction of susceptibles results in ILI epidemics which are not only earlier, but also higher. This does not correspond to the actual measured ILI data (Figure 6.3A,B), which indicates that this basic SIR model would not be able to accurately capture the multi-season dynamics of influenza (see also Figure 4.2, Chapter 4, page 73).

Since the most severe ILI epidemics appear to occur during the months January–March, when winter conditions are most severe, the same weather variables which regulate the seasonality of influenza, might also regulate the ILI epidemics during each winter. Two possible candidates are seasonal transmission and seasonal ILI factor. Based on the relation (4.4) (Chapter 4, page 77), the time-varying variables during the winter are implemented as:

$$\begin{aligned}p(t) &= p_{\min} + (p_{\max} - p_{\min})e^{-10(1-\sin\frac{2\pi t}{365})} \\ \beta(t) &= \beta_{\min} + (\beta_{\max} - \beta_{\min})e^{-10(1-\sin\frac{2\pi t}{365})}\end{aligned}$$

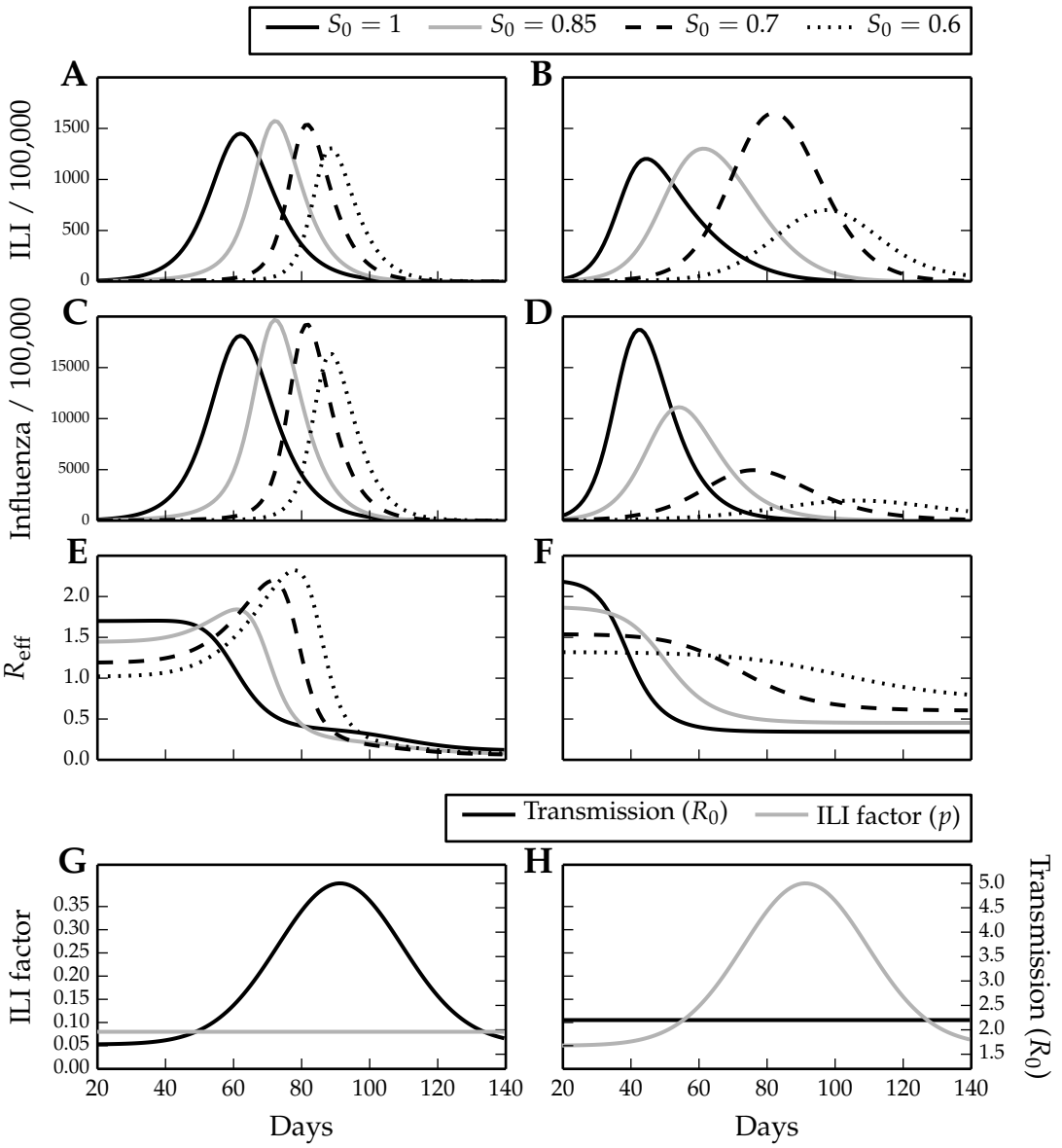



Figure 6.4: INFLUENZA TRANSMISSION MODELS WITH SEASONALITY IN TRANSMISSION OR ILI FACTOR . Both models have parameters $\tau = 0.2$, $m = 10^{-5}$ and initial fraction of susceptibles $S_0 \in [1, 0.85, 0.7, 0.6]$. On the left (A,C,E,G) is a model simulation with seasonal transmission and parameters $\beta_{\text{min}} = 0.34$, $\beta_{\text{max}} = 1$, and $p = 0.08$. On the right (B,D,F,H) is a model simulation with seasonal ILI factor and parameters $p_{\text{min}} = 0.05$, $p_{\text{max}} = 0.4$, and $\beta = 0.44$. A,B) ILI incidence, C,D) Influenza incidence, E,F) Effective reproduction ratio, and G,H) ILI factor and transmission.

Figure 6.4 (left panel) shows an illustrative simulation with time-varying transmission, and Figure 6.4 (right panel) mimics a a time-varying ILI factor. In both simulations, a higher initial fraction of susceptibles leads to earlier onset of the ILI epidemic, but not necessarily to higher ILI incidence (Figure 6.4A,B). There are some differences in the dynamics of both simulations. For the model with time-varying transmission, the number influenza infected follows the same pattern as the number of ILI cases (Figure 6.4A,C), whereas for the model with time-varying ILI factor the patterns are different (Figure 6.4B,D). The effective reproduction number $R_{\text{eff}}$ can increase during the winter in the model with time-varying transmission (Figure 6.4E), whereas it is strictly decreasing in the model with a time-varying ILI factor (Figure 6.4F). So in principle the two processes can be disentangled given appropriate data.

### 6.2.2 *Fitting the model to data*

When a mathematical model has been constructed for disease transmission, it has to be determined whether the model output accurately describes the epidemiological observations. This can be done either qualitatively, meaning that the model simulations do (visually) reproduce the same trends as the observed data, or quantitatively, meaning that model parameters have to be determined to fit the observed data as accurately as possible. In the previous section it has been shown that a model with a time-varying transmission or time-varying ILI factor could qualitatively reproduce the observation that earlier ILI epidemics are not necessarily more severe.

In Chapter 4 the actual ILI data from Influenzanet and ECDC for three countries and the seasons 2003–2013 are quantitatively fitted to a deterministic transmission model in which a time-varying ILI factor is directly related to

the actual absolute humidity and temperature over time. The seasonality in the transmission is modeled as a discrete process, with a constant high transmission in the winter and an insignificant low transmission during the summer. The model furthermore includes a seasonal baseline ILI incidence.

Using the F-test it is confirmed that the inclusion of the seasonal ILI factor significantly improved the fit. Although models with seasonal transmission varying continuously in time are common (KOELLE ET AL., 2006; BACAËR AND AIT DADS, 2011), sometimes directly linked to measured values such as the absolute humidity (SHAMAN ET AL., 2010), Chapter 4 is probably the first effort to explicitly define a time-varying or weather-dependent ILI factor to fit an influenza transmission model to ILI data.

Although quantitatively fitting the model to data is a stronger argument for the validity of the model, in this case the weather-dependence of the ILI factor, caution must taken in interpreting the determined model parameters. Some parameters for the model with a time-varying ILI factor (Table 4.1, Chapter 4, page 74) are actually a composition of other unidentifiable parameters. The estimated transmission factor $\beta$ is a composition of the transmission rate of symptomatics, the percentage of asymptomatics and the transmission rate of asymptomatics, and a normalization of the ILI factor (see Chapter 4). Furthermore, the estimated parameters depend on the model assumptions. As an example, the model assumes a constant transmission rate throughout the winter, but incorporating a time-varying transmission rate into the model would have an impact on the other estimated parameters.

#### 6.2.2.1 *Stochastic models*

The procedure just described is deterministic, meaning that each simulation is uniquely determined by the model parameters. In reality, most processes

are usually stochastic in nature. As an example, the deterministic model considers that all infected hosts make exactly $\beta$ infectious contacts per week. In a stochastic model, the number of infectious contacts per host is not fixed, but usually follows from a certain distribution, such that the number of contacts differs from week to week and from person to person. Each simulation of a stochastic model with the same parameters is expected to generate a different output.

If the number of infected hosts is high, the output of a stochastic model generally approaches the output of a deterministic model, such that the transmission factor $\beta$ can be interpreted as the average number of infectious contacts per week per person. Just as a mathematical model can be considered a simplification of the real biological processes, a deterministic model can be interpreted as a simplification of processes that are inherently stochastic. For each model it should be determined whether this simplification is justified.

In Appendix A a method for parameter estimation of stochastic models is introduced based on master equations, and this system is applied to various simple stochastic transmission processes. For the simplest infection processes, it is possible to give an explicit analytic formula for determining the most likely parameters, based on the observed data. However, for even moderately elaborate models, such as the SIR model, the method leads to analytically unsolvable partial differential equations (PDEs). An open question remains as to whether for more complicated models for which the stochastic nature is considered important, it would be better to estimate parameters by solving these PDEs numerically, or whether numerical procedures applied direct to the model simulations, such as Markov chain Monte Carlo or other integrated methods (Bretó et al., 2009), are preferable.

### 6.2.3 *Perspectives*

Although the model from Chapter 4 was able to accurately capture the ILI dynamics for three countries during ten seasons with the introduced seasonal ILI factor, the assumption of constant transmission throughout the winter for all symptomatic influenza infected might be too strong. Furthermore, it is assumed that all symptomatic infections transmit equally, independent of the severity of symptoms and whether they fit the ILI case definition. Although persons with more symptoms might have higher viral shedding (Lau et al., 2010), they might reduce contact with others by staying home and therefore decrease their effective transmission. Future research is required to disentangle (He et al., 2013) the importance of seasonality in the transmission and the ILI factor.

Although it is difficult to directly measure seasonality in transmission, it might be possible to determine seasonality in the ILI factor by other methods. Some cohort studies detect mild and asymptomatic infections by testing for seroconvergence between the beginning and the end of the season (Mann et al., 1981). A relation between the weather conditions during the ILI epidemic, the severity of the epidemic, and the rate of seroconvergence (see Figure 6.4B,D), might indicate the impact of a seasonal ILI factor. In (Lau et al., 2010) household members for families with one confirmed index case for influenza were regularly monitored for 1–2 weeks, both for symptoms and (RT-PCR) confirmed influenza cases. Such a study under various weather conditions might expose a weather-associated ILI factor.

The possible importance of the ILI factor was inspired by the selection of an ILI case definition for Influenzanet. Although many ILI case definitions lead to very similar trends, the inclusion or exclusion of a few symptoms from the

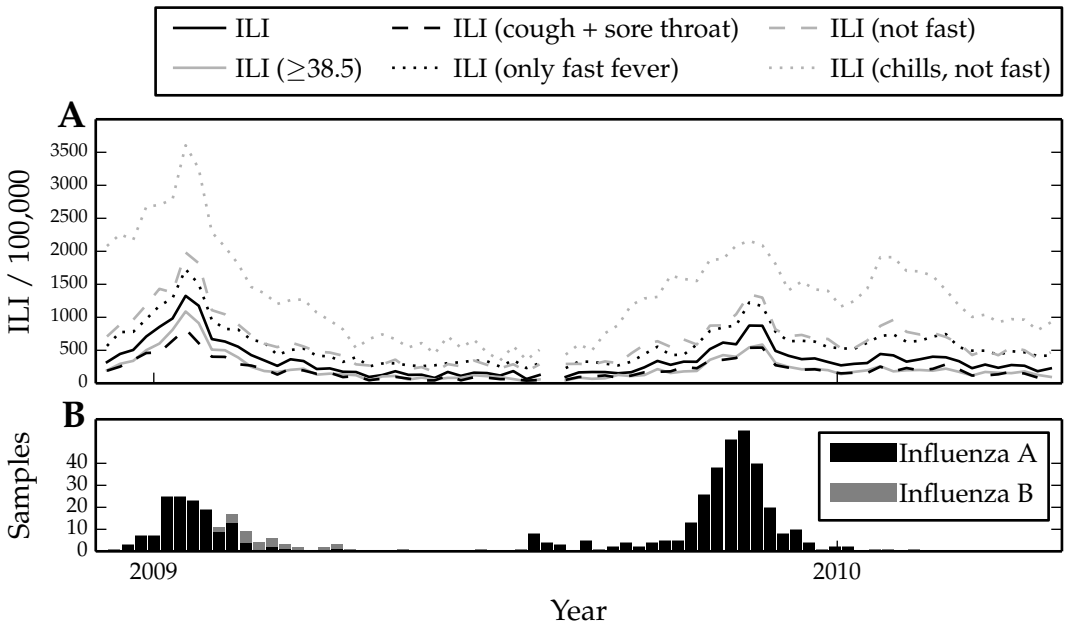



Figure 6.5: ILI INCIDENCES BASED ON VARIOUS CASE DEFINITIONS in the Netherlands (2008–2010). A) ILI incidences according to Influenzanet. The ILI incidence by Influenzanet is defined as the number of ILI onsets per 100,000 participant-weeks. B) Confirmed influenza cases according to ECDC. The case definitions are listed in Table 6.3. Note that in January–March 2010 many participants match the "weaker" ILI case definitions, despite the lack of confirmed influenza infections.

Table 6.3: VARIOUS ILI CASE DEFINITIONS

| CASE DEFINITION | SYMPTOMS |
|---|---|
| ILI | Acute onset, and temperature $\geq$38 °C, and muscle pain or headache, and cough or sore throat |
| ILI ($\geq$38.5) | Acute onset, and temperature $\geq$38.5 °C, and muscle pain or headache, and cough or sore throat |
| ILI (cough + sore throat) | Acute onset, and temperature $\geq$38 °C, and muscle pain or headache, and cough, and sore throat |
| ILI (only fast fever) | Acute onset, and temperature $\geq$38 °C. |
| ILI (not fast) | Temperature $\geq$38 °C, and muscle pain or headache, and cough or sore throat |
| ILI (chills, not fast) | Temperature $\geq$38 °C or cold shivers / chills, and muscle pain or headache, and cough or sore throat |

case definition could lead to significant decrease or increase in the absolute value for the ILI incidence (Figure 6.5). It is possible that during warm or humid weather some ILI symptoms are less likely to be experienced (BAR-OR ET AL., 1977; BOLLAG, 2009; ECCLES, 2002; CANNELL ET AL., 2006) by persons infected with influenza.

Vaccination is not incorporated explicitly into the model, but is implicitly included in the estimated fraction of susceptibles at the beginning of each season. Since the vaccination campaign is usually at the beginning of the season, this is generally an accurate simplification of the model. However, during the pandemic season 2009–2010 the vaccine only became available when the ILI activity was already epidemic in all countries, so the model might be improved by including explicitly the vaccination campaign for the pandemic vaccine in 2009.

## 6.3 INFLUENZANET: REMAINING CHALLENGES

### 6.3.1 *The challenge of accuracy of self-reported symptoms*

Although Influenzanet participants in general seem to be quite able to evaluate and report their own symptoms, care should be taken with the applied ILI case definition. The ILI case definition used throughout this thesis requires a measured body temperature $\geq$38 °C, which thus requires participants to actually measure their temperature. Since 2011 participants in all countries can also report an unmeasured fever, which allows for a case definition for suspected ILI which includes a measured or an unmeasured fever (acute onset, and unmeasured fever or temperature $\geq$38 °C, and muscle pain or headache, and cough or sore throat). The percentage of participants with suspected ILI who had a measured temperature $\geq$38 °C, and thus fit the standard ILI case definition, varies by country (Figure 6.6). This is probably due to cultural differences, where participants in for example the United Kingdom are much less likely to actually measure their body temperature.

For an applied ILI case definition which requires a measured temperature, these differences could affect the measured ILI incidences. The differences between the Netherlands (60%), Belgium (52%), and Portugal (49%) are relatively small, leading to similar ILI attack rates (Figure 2.3A, Chapter 2, page 25). For cross-country analyses not including data before 2011, it might be preferable to use an ILI case definition which does not require a measured temperature.

However, it is important to realize that this involves a compromise. Applying an ILI case definition which accepts unmeasured fever (possible since 2011), may introduce another "self-diagnosis" bias, if certain groups are more

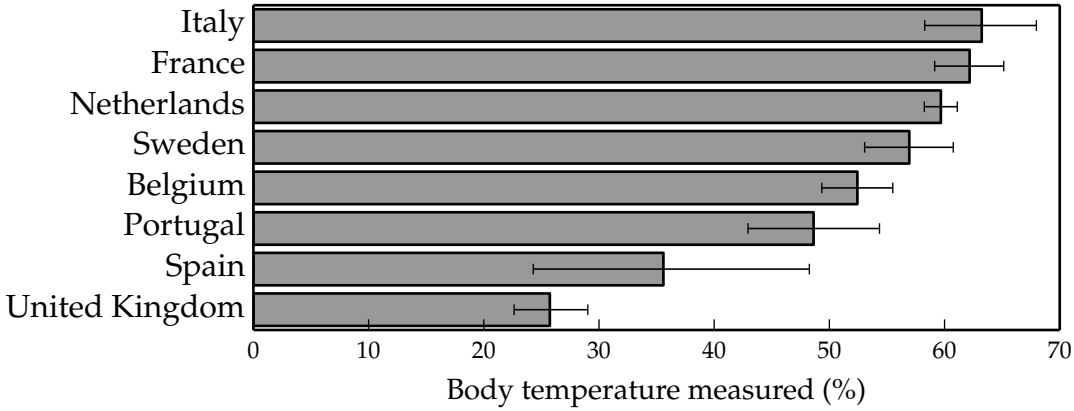



Figure 6.6: PARTICIPANTS WITH SUSPECTED ILI WHO MEASURED THEIR BODY TEMPERATURE and had a temperature $\geq$38 °C, classified by country (2011–2013). Suspected ILI is defined by the standard ILI case definition, but with a measured or unmeasured fever (acute onset, and fever (unmeasured or temperature $\geq$38 °C), and muscle pain or headache, and cough or sore throat).

likely to think they have a fever. A Hong Kong study in which children and adults were asked whether they have fever before their actual temperature was measured, revealed that children were much more likely than adults to claim they have fever for any measured temperature, and that a significant quantity of people with low temperature claimed to have fever (PATTERSON-LOMBA ET AL., 2014).

The determined ILI risks (Table 3.2, Chapter 3, page 49) can also depend on the applied ILI case definition. For a risk analysis during the seasons 2011–2013 for the same covariates, the determined ILI risks for children and women are lower when using the case definition for suspected ILI (accepting an unmeasured fever), than with the standard ILI case definition (requiring measured temperature $\geq$38 °C) (data not shown). Further research is needed to determine whether these differences are due to different habits in actually

measuring the body temperature, or to different self-evaluations for having a fever.

Similar to fever, the evaluation of other symptoms could also be different in various subgroups. For fever a reasonable objective standard exists — the actual measured temperature — but for many symptoms such as a headache or muscle pain this is much harder to evaluate. However, these problems are not unique to self-reporting systems such as Influenzanet, but could also affect the diagnosis by medical doctors.

### 6.3.2 *The challenge of estimating vaccine effectiveness*

With Influenzanet the incidence of influenza-like illness (ILI) in a cohort of self-reporting participants is measured, and from those participants it is known for each season whether they were vaccinated for influenza. Therefore, Influenzanet would seem like a powerful system to measure the effectiveness of the influenza vaccine, even in real-time. In Chapter 3 the vaccine effectiveness for all seasons 2003–2013 is determined.

There are two problems which complicate the determination of the vaccine effectiveness based on Influenzanet data. First of all, Influenzanet measures the incidence of influenza-like illness (ILI), a syndromic case definition, whereas the influenza vaccine only protects against infections with the influenza virus. Not all people infected with the influenza virus will have sufficient symptoms to fit the ILI case definition, and not all ILI cases are due to an infection with the influenza virus. A second problem is that vaccinated and unvaccinated participants cannot be directly compared. Participants who decide to take a vaccine most often do this because they belong to a risk group. Differences in ILI rate between vaccinated and unvaccinated participants can be either due to the vaccine, or due an a priori different risk between both groups. Here the vaccine effectiveness for the season 2008–2009 in the Netherlands is estimated based on various methods which try to minimize these problems. The vaccine effectiveness is determined by the relative risk reduction in vaccinated participants.

#### 6.3.2.1 *Influenza and ILI*

SURVEILLANCE PERIOD Although ILI onsets are reported during the whole season, only during certain weeks within each season the actual influenza

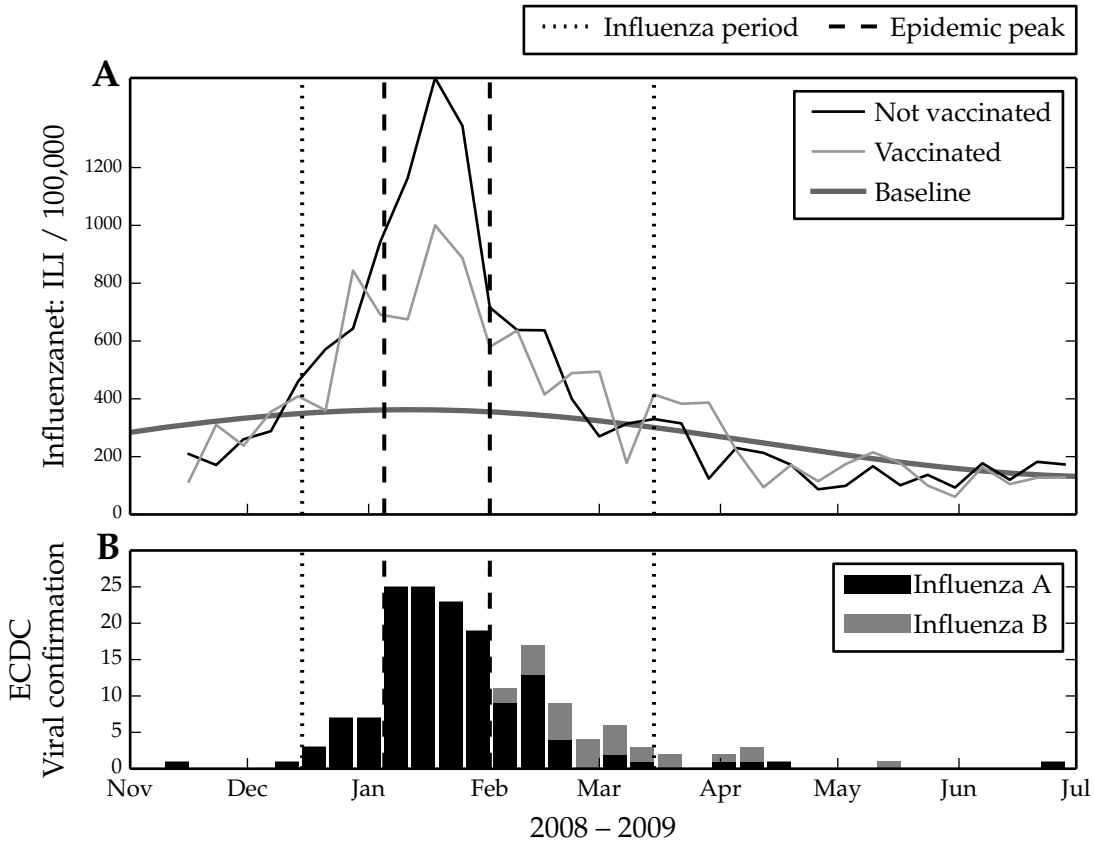



Figure 6.7: ILI INCIDENCE IN VACCINATED AND UNVACCINATED PARTICIPANTS in the Netherlands (2008–2009). A) ILI incidence in vaccinated and unvaccinated participants. The ILI incidence by Influenzanet is defined as the number of ILI onsets per 100,000 participant-weeks. The baseline is the typical number of ILI cases measured in the absence of circulating influenza viruses. B) Number of weekly confirmed laboratory cases as reported by the ECDC. The dotted vertical lines indicate the influenza period when the number of confirmed influenza cases is at least 15% (average over 3 weeks) of the maximum. The dashed vertical lines indicate the epidemic when the number of confirmed influenza cases is at least 70% (average over 3 weeks) of the maximum.

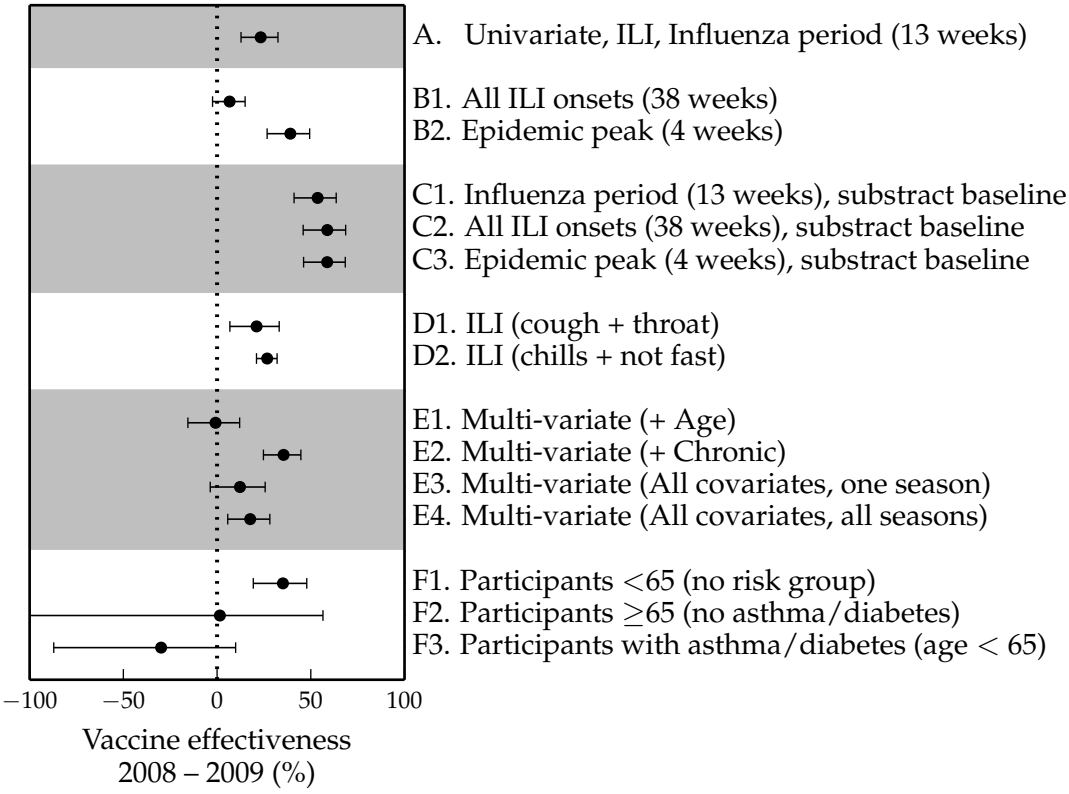



Figure 6.8: VACCINE EFFECTIVENESS BASED ON VARIOUS ASSUMPTIONS in the Netherlands (2008–2009).

virus is detected in significant quantities by virological tests. Figure 6.7B shows the number of confirmed influenza samples as reported by ECDC in the Netherlands during the season 2008–2009. The influenza period (between dotted lines) is defined as the weeks that the number of positive samples is at least 15% of the maximum number of positive samples (moving average over 3 weeks), and the epidemic peak (between dashed lines) as the weeks that the number of positive samples is at least 70%. During the 2008–2009 season the influenza period was from 15 December–15 March, and the epidemic peak from 5 January–2 February.

Figure 6.7A shows the ILI incidence in vaccinated and non-vaccinated participants. By default the vaccine effectiveness is estimated by only considering ILI onsets during the influenza period, which leads to an estimated vaccine effectiveness of 23% (CI 13–32%) (Figure 6.8A). The ILI curves of vaccinated and unvaccinated participants diverge mostly during the epidemic peak, when the ILI incidence among unvaccinated participants is clearly higher than among vaccinated participants. The estimated vaccine effectiveness based only on the ILI onsets during the epidemic peak is 39% (CI 27–49%) (Figure 6.8B1). Including all ILI onsets during the full season leads to a non-significant estimate for the vaccine effectiveness of 7% (CI -2–15%) (Figure 6.8B2).

INFLUENZA-RELATED ILI The baseline in Figure 6.7A is determined as the typical ILI incidence among all participants outside the influenza periods (see Section 6.3.3). It is plausible that most ILI cases under this baseline are not due to influenza infections. By considering only ILI cases above the baseline, the estimated vaccine effectiveness to influenza-related ILI is 54% (CI 41–64%) during the influenza period, 59% (CI 46–69%) during the full season, and 59% (CI 46–68%) during the epidemic peak (Figure 6.8C). Interestingly, by excluding the baseline ILI incidence, the estimated vaccine effectiveness for influenza-related ILI does depend strongly on the surveillance period.

ILI CASE DEFINITIONS The vaccine effectiveness can be estimated based on various ILI case definitions. The estimated vaccine effectiveness for the stronger "ILI (cough + sore throat)" case definition (Table 6.3) is 21% (CI 7–33%), whereas for the weaker "ILI (chills + not fast)" case definition it is 27% (CI 21–32%) (Figure 6.8D). During the season 2008–2009, the estimated vaccine effectiveness does not appear sensitive to the applied ILI case definition, but

the the confidence intervals are much smaller for weaker ILI case definitions, since it is determined based on more cases. For seasons with low activity or countries with low participation, this could be the difference between a significant or a non-significant estimate for vaccine effectiveness.

#### 6.3.2.2 *Multivariate analyses*

Many participants are vaccinated because they belong to a risk group. In the Netherlands people belong to a risk group if they are over 65 years old (over 60 since 2010) or if they have a chronic disease such as asthma or diabetes. Risk groups consist of people who either have a higher risk of influenza infection or a higher risk of complications if infected. Figure 6.9 shows the vaccination rate in the Netherlands in 2008–2009, structured by age and chronic diseases, together with reported reasons for vaccination.

A multivariate regression analysis determines the influence of various personal characteristics, including vaccination status, on the probability that a participant has at least one ILI onset. For a multivariate regression model which includes vaccine status and age group (<18, 18–49, 50–64, 65+) as risk factors, the estimated vaccine effectiveness is -1% (CI -16–12%) (Figure 6.8E1). The main reason for the estimated vaccine effectiveness to be lower than with the univariate model (Figure 6.8A), is that the model estimates participants over 65 to have a lower risk for ILI and infers that the lower ILI rate in participants over 65 is not due to their vaccination but rather due to their age.

An opposite effect is observed by a multivariate regression model which includes vaccine status and the presence of a chronic disease (asthma and/or diabetes) as risk factors, which estimates a vaccine effectiveness of 35% (CI 25–45%) (Figure 6.8E2). The estimated vaccine effectiveness is higher than with the univariate model Figure 6.8A), since the model estimates that participants

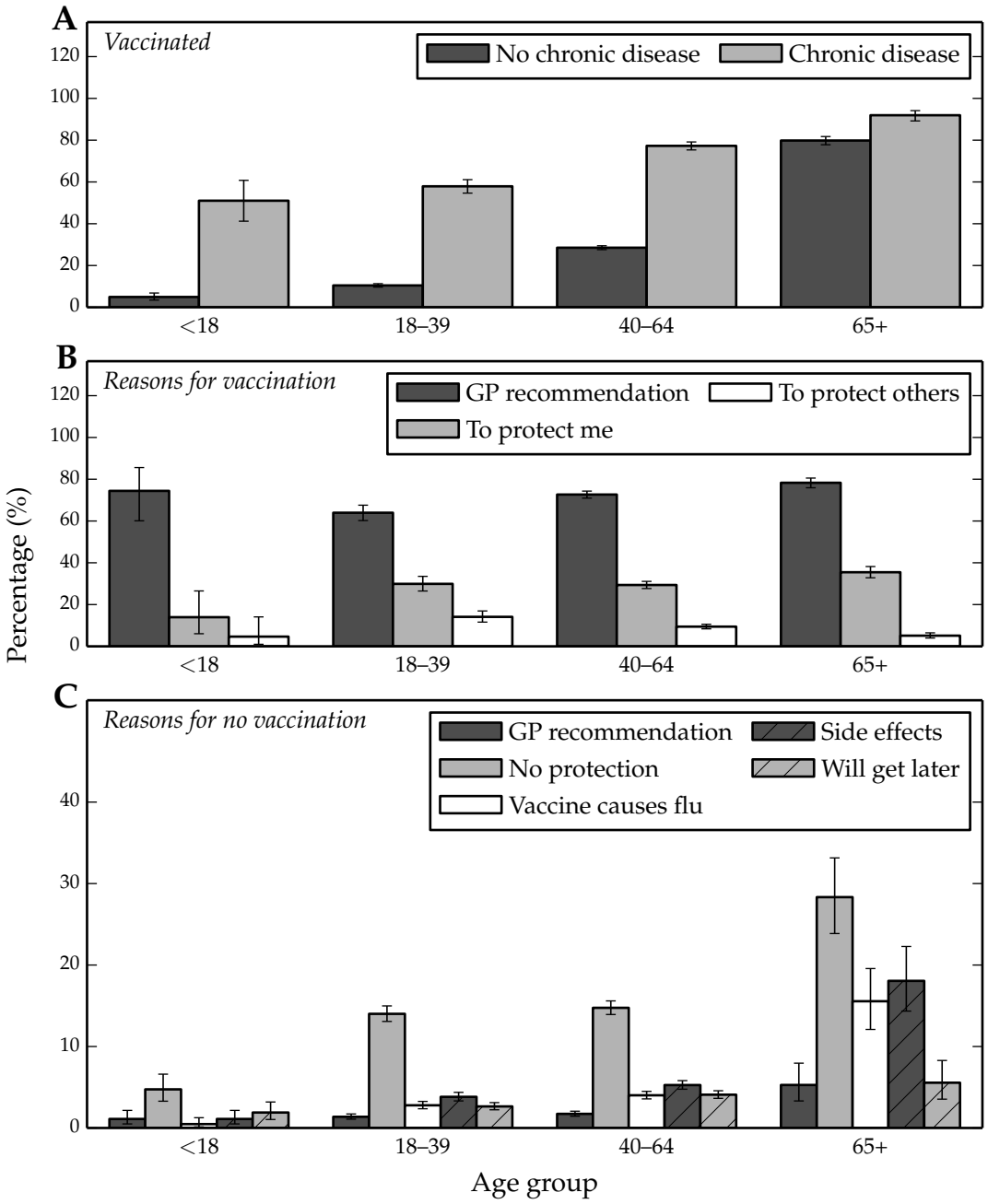



Figure 6.9: VACCINATION RATE AND REASONS FOR VACCINATION by age group in the Netherlands (2008–2009). A) Vaccination rate in participants with and without chronic disease. B) Reasons for vaccination (multiple answers possible). C) Reasons for no vaccination (multiple answers possible).

with asthma or diabetes have a higher risk for ILI, and infers a higher vaccine effectiveness.

For a multivariate regression model which includes various covariates (including chronic disease and age group, see Table 3.2, Chapter 3, page 49), the estimated vaccine effectiveness is non-significant 12% (CI -4–26%) (Figure 6.8E3). A multivariate regression model can also be applied to data from all seasons 2003–2013 simultaneously. Although the vaccine effectiveness is considered to be season dependent, all other variables are considered invariant across the seasons. The inclusion of data from all seasons 2003–2013 leads to different estimates for the season-independent covariates, and the model adjusts the vaccine effectiveness for the season 2008–2009 accordingly. According to this multi-season regression the estimated vaccine effectiveness for the season 2008–2009 is 18% (CI 6–28%) (Figure 6.8E4). This model is applied to determine the ILI risk factors in Table 3.2, Chapter 3, page 49, although for multiple countries simultaneously.

#### 6.3.2.3 *Different groups*

For a multivariate regression, it is assumed that certain personal characteristics (such as chronic disease and age) could lead to an (a priori) higher or lower risk of ILI, whereas the effect of a vaccine is equal for all. However, it is also possible, that the vaccine effect depends on these personal characteristics. For participants who do not belong to a risk group (age under 65, no asthma, no diabetes and no vaccine recommendation from their GP), the estimated vaccine effectiveness is 35% (CI 19–48%). For participants who do belong to a risk group, the estimated vaccine effectiveness is a non-significant 1% (CI -123–56 %) for participants over 65 without asthma or diabetes, and

a non-significant -30% (CI -87–10%) for participants under 65 with asthma or diabetes (Figure 6.8F).

#### 6.3.2.4 *General remarks*

Since ILI is not a perfect proxy for influenza infections, vaccination is voluntary and Influenzanet participation is voluntary, the Influenzanet system is not well suited to estimate the absolute vaccine effectiveness. Based on Influenzanet data for the Netherlands 2008–2009, it can be argued that there is no significant overall reduction in ILI cases among vaccinated participants (Figure 6.8B), and that the estimated vaccine effectiveness among participants belonging to a risk group is even negative although insignificant (Figure 6.8F3). On the other hand, the same data could be used to argue for a significant reduction of 54–59% in influenza-caused ILI among vaccinated participants (Figure 6.8C). Although the absolute value for the vaccine effectiveness depends strongly on the assumptions, the season-to-season variation (not shown) in vaccine effectiveness is more robust. Influenzanet might thus provide a valuable tool for determining the relative vaccine effectiveness in real-time.

The vaccine effectiveness is regularly determined by I-MOVE (Influenza Monitoring Vaccine Effectiveness in Europe) within the ECDC. For all patients with ILI who visit a sentinel GP and are swabbed, the vaccine effectiveness is determined based on laboratory confirmed influenza. According to Influenzanet data, vaccinated participants in the Netherlands are twice as likely to seek medical care if they have ILI than unvaccinated participants (Figure 6.10). Such a bias might affect the determined vaccine effectiveness by I-MOVE.

According to large meta-studies the estimated vaccine effects vary widely (Demicheli et al., 2009; Jefferson et al., 2010), and even the interpretation of the same data can be controversial (Beyer et al., 2013). As shown,

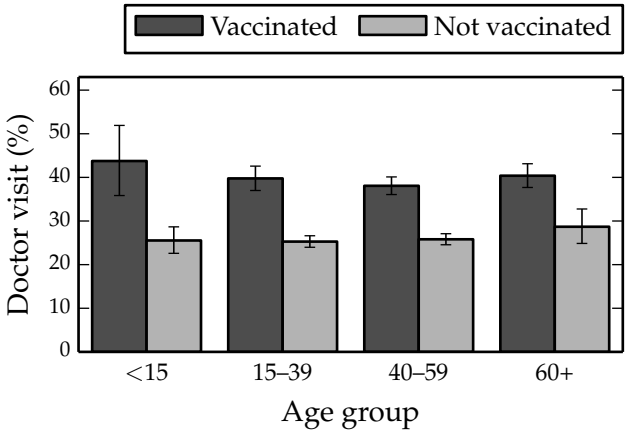



Figure 6.10: MEDICAL CARE SEEKING RATE IN VACCINATE AND UNVACCINATED PARTICIPANTS in the Netherlands (2003–2013).

the vaccine effectiveness does not only vary from season to season (Table 3.2, Chapter 3, page 49), but the estimated vaccine effectiveness depends strongly on the used method and assumptions (Figure 6.8). Since various studies on vaccine effects, see for example (BEYER ET AL., 2013), have long lists of competing interests with the pharmaceutical industry, ambiguity in determining vaccine effectiveness is a reason for concern. According to a study on newspapers in the UK in 2009–2010, academics with competing interests assessed the risk for the new H1N1pdm influenza strain higher and had increased advocacy for pharmaceutical products to counter this risk (MANDEVILLE ET AL., 2014).

### 6.3.3 *The challenge of real-time monitoring*

It is often suggested that self-reporting surveillance systems might be able to detect changes in disease activity earlier than the traditional surveillance systems, such as for Influenzanet in the Netherlands (Friesema et al., 2009), FluNearYou in the U.S.A. (Chunara et al., 2012), and FluTracking in Australia (Dalton et al., 2009). For Google Flu Trends, another well known online ILI surveillance system, it has been shown that it could consistently detect changes in ILI trends 1–2 weeks earlier during the 2007–2008 season in the U.S.A. than the traditional clinical surveillance by CDC (Ginsberg et al., 2009). However, such early detection has never been shown for any of the Influenzanet-like systems. Here the capabilities for real-time detection of ILI activity by Influenzanet and sentinel GP surveillance system as coordinated by ECDC are compared for the Netherlands, Belgium and Portugal (seasons 2003–2013).

Run-time detection of disease activity above baseline should not be confused with detection of a newly emerging disease. A cohort-based system such as Influenzanet, which so far has been able to recruit at maximum 0.07% of the population (Table 6.1), is not a viable system for early warning for the first cases of a new disease. This section is dedicated to the determination of whether the ILI activity, which is present throughout the whole season, is significantly higher than normal, signaling the beginning of the yearly influenza epidemic.

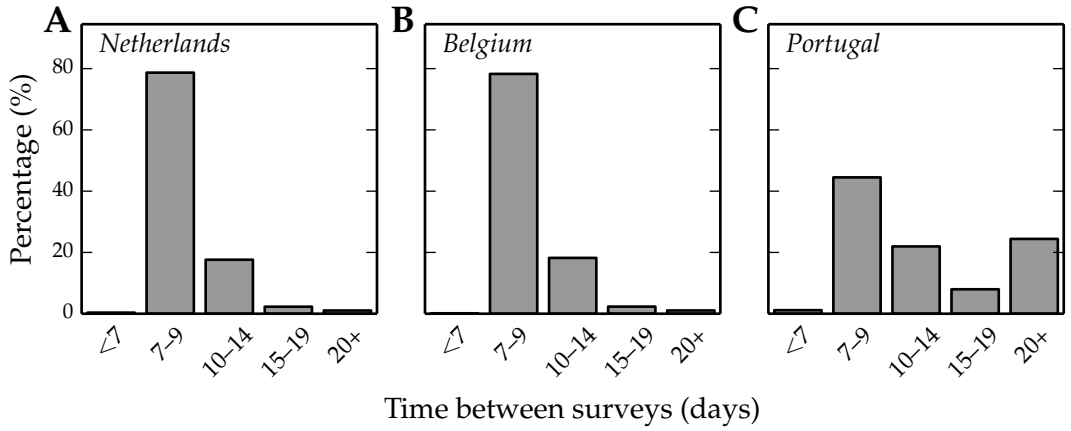



Figure 6.11: FREQUENCY OF REPORTING TO INFLUENZANET in A) Netherlands, B) Belgium, and C) Portugal (2008–2009).

#### 6.3.3.1 *ILI activity*

To determine the ILI activity in Influenzanet, both retrospectively and in real-time, it has to be determined how the number of ILI cases are counted (numerator), and what is the cohort size (denominator).

The simplest way to determine the number of ILI cases (numerator), is to count each week the number of completed questionnaires which fit the ILI case definition, leading to a so-called ILI reporting rate. The denominator can be either determined as the total number of registered participants, or the number of completed questionnaires each week. Since Influenzanet participation is online and voluntary, some participants lose interest in the system during the season and stop reporting. Determining the ILI reporting rate based on the number of registered participants therefore underestimates the ILI activity. Despite the weekly reminders sent by email, participants sometimes skip reporting during a certain week (Figure 6.11). Determining the ILI reporting rate based on the weekly number of completed questionnaires therefore overestimates the ILI activity.

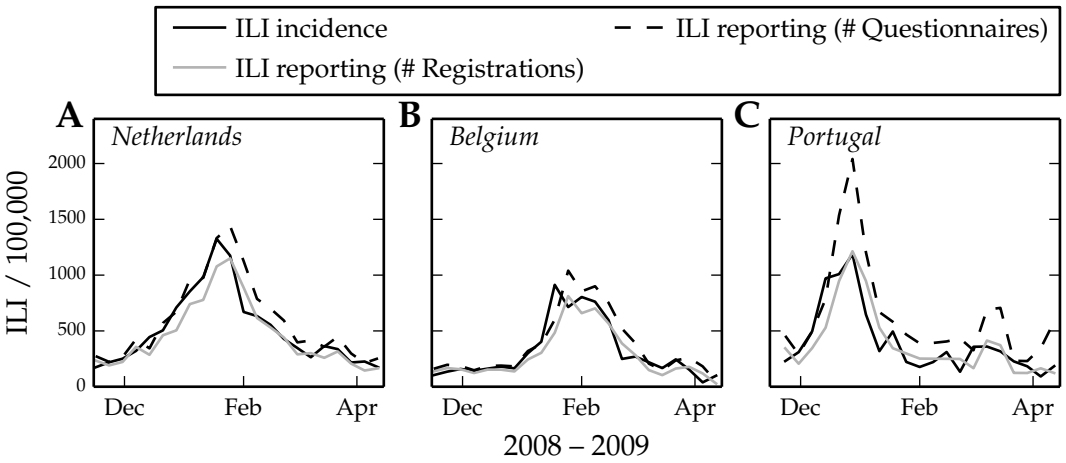



Figure 6.12: Comparison of ILI incidence and reporting rate for A) Netherlands, B) Belgium, and C) Portugal (2008–2009). The denominator for the ILI reporting rate is based on either all registered participants, or on the number completed questionnaires each week. The ILI incidence by Influenzanet is defined as the number of ILI onsets per 100,000 participant-weeks.

Since Influenzanet participants also report when their symptoms started, a more appropriate way to determine the ILI activity in Influenzanet is based on the onset day, leading to a so-called ILI incidence. The number of ILI cases (numerator) is determined as the total number of participants who fit the ILI case definitions, for which the onset day (based on the fever onset day) was in that week. The denominator is determined each week as the number of participants who have completed a symptoms questionnaire covering that week. Throughout the thesis the ILI activity by Influenzanet is always determined as the ILI incidence.

In Figure 6.12 the three alternatives for determining the ILI activity are plotted: the ILI reporting rate based on all registered participants, the ILI reporting rate based on the number of completed questionnaires each week, and the ILI incidence based on the day of ILI onset and the number of active

participants. The ILI incidence precedes the ILI reporting rate by a few days since it is based on the onset day of the symptoms, and not on the day of reporting. The amount of over- and underestimation of the ILI reporting rates is also visualized.

Since the Influenzanet ILI incidence is based on the day the symptoms started, whereas the ECDC ILI incidence is based on the day a patient visited a GP the ILI incidence more closely resembles the real ILI activity in the population. This can be important when linking ILI activity with weather variables (Chapter 4) or school closures (EAMES ET AL., 2012). Although sentinel GPs might start registering the onset day for patients they have diagnosed with ILI, this can only be used to determined an ILI incidence at the end of the season. Since ECDC cannot determine a dynamic denominator for the cohort size, in real-time reporting the ILI incidence based on the onset day would be consistently under estimated by ECDC.

In real-time monitoring, for participants who have not (yet) reported for a certain week, it is still unknown whether they had an onset of ILI or not. These participants are neither included into the numerator nor the denominator for the ILI incidence. Participants who only reported over a part of the week are counted as a fraction for the denominator. The determined ILI incidence is therefore always preliminary in real-time monitoring, especially over the most recent weeks, and will approach its final value when participants continue to report in the following weeks.

#### 6.3.3.2 *Data collection*

Influenzanet can update the ILI incidence in real-time as participants report their symptoms online, orchestrated by the reminders sent to them on a weekly basis. The email reminder can be sent to all participants on the same

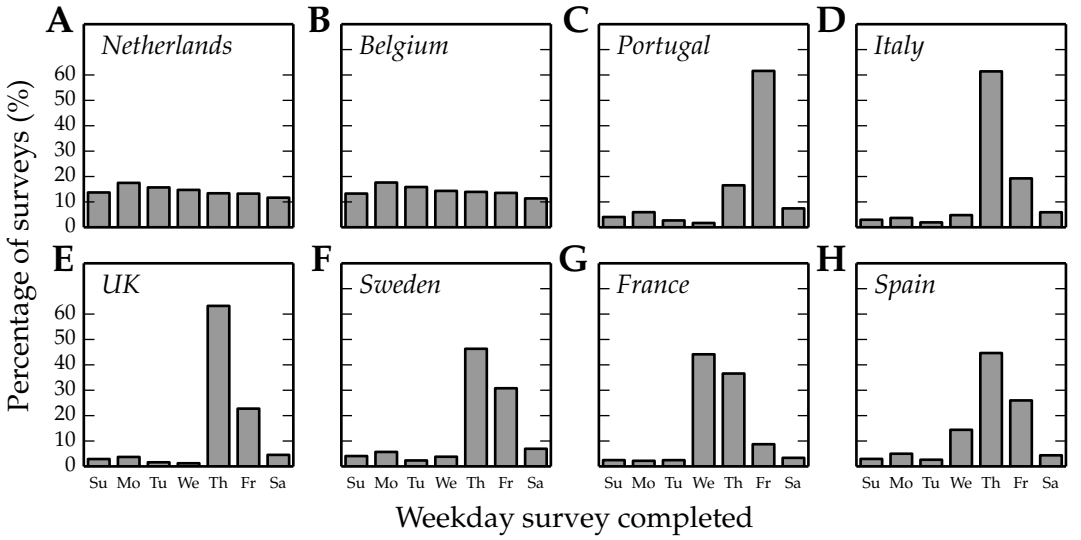



Figure 6.13: Weekday on which Influenzanet participants report for various countries (2012–2013).

day, or spread throughout the week. As Figure 6.13 shows, in the Netherlands and Belgium the reminders are sent throughout the week, whereas in Portugal, Italy, United Kingdom, Sweden and Spain the email reminders are sent once a week. Which method is used has implications for the time when data becomes available.

If the email reminder is sent to all participants on the same day, most symptoms questionnaire will be completed within a day. Influenzanet is therefore able to significantly update the ILI incidence once every week with a delay of 1 day based on complete data from most participants. If all email reminders are sent throughout the week, every day new information becomes available, and every day the latest weekly ILI incidence over the preceding 7 days can be updated. However, the determined ILI incidence over the most recent days is preliminary, since for many participants it is still unknown whether they had an onset of symptoms or not.

Sentinel doctors report at the end of every week the number of patients they diagnosed with ILI, and the weekly ILI incidence is reported and published by ECDC on the following Wednesday or Thursday. The reported incidence is usually slightly updated one week later to correct for late-reporting GPs. Due to the 2–5 days between the onset of ILI and the GP visit (Figure 3.4, Chapter 3, page 45), and the 3–4 days the ECDC needs to collect and publish, Influenzanet will be able report changes in activity 5–9 days earlier than ECDC, and based on the daily preliminary ILI incidence even earlier. This advantage would mostly disappear, however, if sentinel doctors would start to report daily (CARRAT ET AL., 1998; LANGE AND SCHÖTTLER, 2002).

#### 6.3.3.3 *Epidemic onset detection*

Having a smaller time lag in data collection does not necessary indicate that Influenzanet can detect changes in ILI trends earlier than ECDC. In real-time monitoring each system has to be able to determine whether changes in the detected ILI activity are due to real changes in ILI activity in the population, or are caused by stochastic uncertainty or measurement error. One way to evaluate the real-time power of a system, is to evaluate when each system is able to detect the onset of the yearly influenza epidemic.

To determine when the measured ILI activity is epidemic, first it has to be determined what ILI activity can be considered non-epidemic. One way to determine the baseline non-epidemic ILI incidence is to fit a general function to the historic non-epidemic ILI incidences:

$$B(t) = A \cdot \sin\left((t - H) \cdot \frac{2\pi}{365}\right) + V$$

with parameters amplitude ($A$), horizontal shift ($H$) and vertical shift ($V$). Such general function was first applied to epidemic detection by (SERFLING, 1963). Here no linear term for a secular trend is considered and only one Fourier term is included.

Various methods have been applied to choose which historic ILI incidences can be considered non-epidemic, such as only considering non-epidemic years (SERFLING, 1963), removing the 25% highest ILI activities (VIBOUD ET AL., 2004), removing the ILI rate above a certain threshold (COSTAGLIOLA ET AL., 1991), and removing the period when the ILI activity increased significantly (VEGA ET AL., 2013).

Since influenza-like illness is used as a proxy for influenza infections, it is also possible to look at the number of laboratory confirmed influenza cases to determine when ILI activity can be considered epidemic. In (COWLING ET AL., 2006) the non-epidemic periods are defined as those weeks when the percentage of samples tested positive for influenza is below a certain threshold. Based on the yearly reports published by ECDC, which only list the number of positive samples per week, the influenza period for each season is defined as those weeks when the number of positive samples is at least 15% of the maximum number of positive samples in that season (3-week average). For baseline fitting only the ILI incidences outside these determined influenza periods are included.

Figure 6.14B shows the absolute number of samples taken by the sentinel GPs and tested positive for influenza in the Netherlands 2007–2011, indicating the influenza period. In Figure 6.14A the ILI incidence as determined by Influenza is plotted, as well as the best fit for the baseline ILI incidence. The somewhat subjective method to choose the weeks when the ILI activity can be considered non-epidemic, will lead to different ILI baselines.

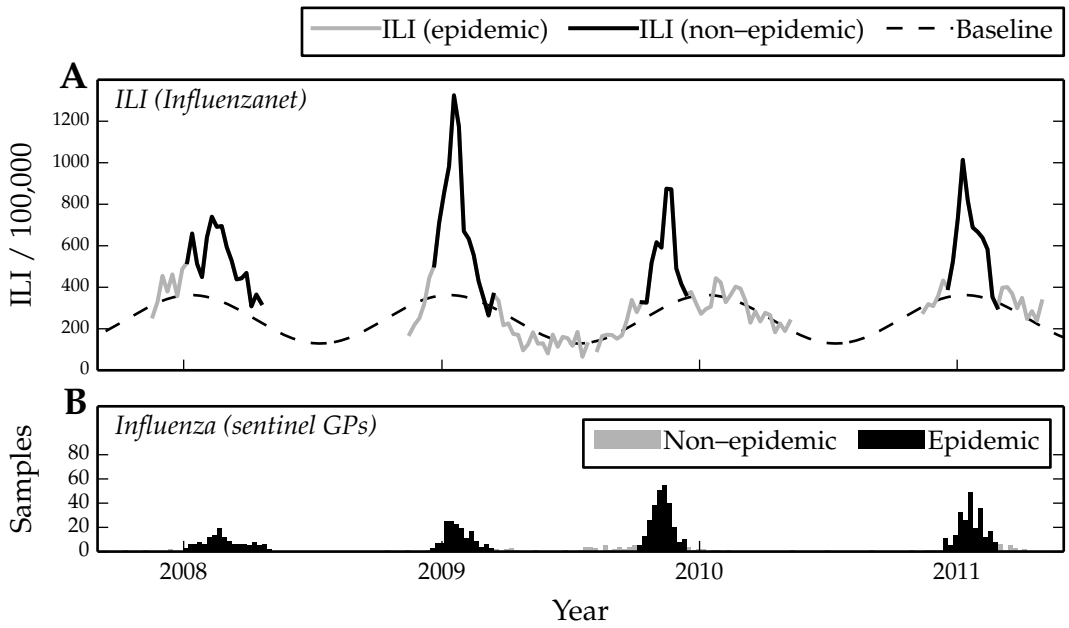



Figure 6.14: INFLUENZANET ILI INCIDENCE, BASELINE, AND CONFIRMED INFLUENZA CASES in the Netherlands (2007–2011). The ILI incidence by Influenzanet is defined as the number of ILI onsets per 100,000 participant-weeks. The epidemic periods are defined by the influenza periods, the weeks when the number of positive samples is at least 15% of the maximum number of positive samples in a season (3-week average).

Next it has to be determined whether the determined ILI incidence is significantly above baseline, usually by defining a certain threshold for epidemic activity. A common approach is to determine the variance in the historic non-epidemic ILI incidence, and determine the epidemic threshold such that the probability that the ILI activity is above this threshold is less than 5% (SERFLING, 1963; PELAT ET AL., 2007). In (VEGA ET AL., 2013) the threshold for epidemic activity is defined by fitting a constant baseline function only to the highest ILI rates of the pre-epidemic periods.

These methods to determine the threshold for epidemic activity do not work very well for Influenzanet. The number of participants can change from season to season, and in real-time monitoring the preliminary ILI incidence is determined by a lower number of participants as the data are still being entered. A lower number of participants leads to more uncertainty in the determined ILI incidence, and thus to a higher threshold. Therefore, the threshold for epidemic activity in Influenzanet cannot be based (only) on the the variance of historic Influenzanet ILI incidences.

Here we assume that the uncertainty in the measured ILI incidence by Influenzanet is only due to the sampling bias. Suppose that during a certain week $N_t$ Influenzanet participants are active and the baseline incidence is estimated by $b_t$. The number of ILI onsets during a non-epidemic period can be considered a binomial experiment

$$I_t \sim \text{Bin}(N_t, b_t)$$

which for large a number of participants can be approximated by the Poisson distribution with parameter $b_t N_t$ with a continuous approximation in the Gamma distribution with parameters $(b_t N_t, 1)$.

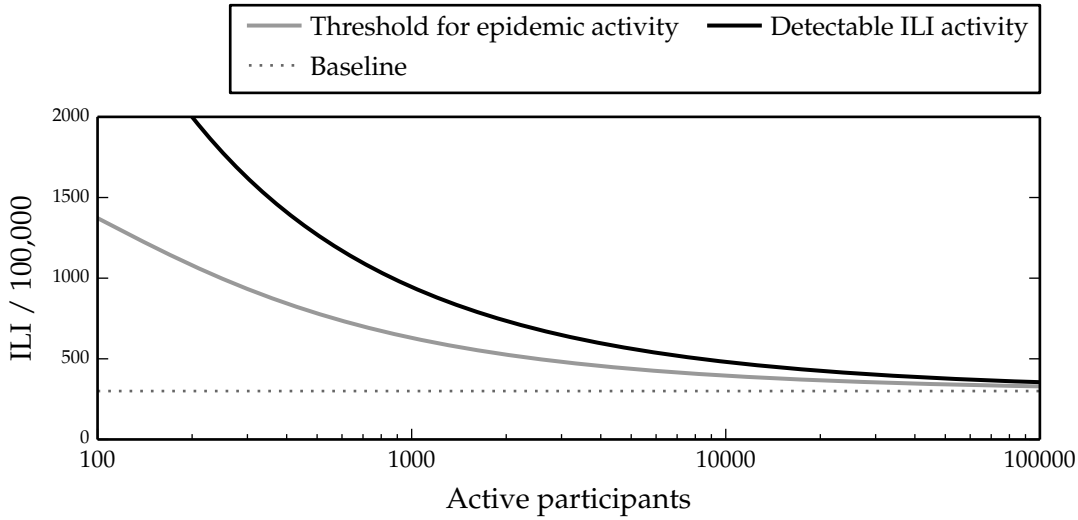



Figure 6.15: THRESHOLD FOR EPIDEMIC DETECTION AND DETECTABLE ILI ACTIVITY for various number of active participants. The ILI incidence by Influenzanet is defined as the number of ILI onsets per 100,000 participant-weeks. The baseline ILI incidence is fixed at 300 per 100,000.

The threshold $T_t = X_t/N_t$ is determined, such that the probability to observe more than $X_t$ cases is less than 5%

$$X_t = \inf\{x_t : P(I_t > x_t) < 0.05\}$$

A higher number of participants leads to a lower threshold (Figure 6.15).

A different question is given that the actual ILI incidence in the whole population is $p_t$, what is the minimal number of active participants needed to detect that this ILI activity is significantly above baseline. Based on (ROSNER, 2006), the minimal sample size to distinguish the output of a Bernoulli trial from a known proportion can be estimated by:

$$N_t = \frac{\left(z_{1-\alpha}\sqrt{b_t - b_t^2} + z_{1-\beta}\sqrt{p_t - p_t^2}\right)^2}{(p_t - b_t)^2}$$

with power $1 - \beta$ and significance level $\alpha$ for two-tailed and $2\alpha$ for one-tailed test. Figure 6.15 shows for various ILI incidences in the population the minimal number of active participants needed to determine with power 80% that the measured ILI incidence is significantly (level 5%) above a baseline of 300 per 100,000.

#### 6.3.3.4 *Detection*

For ECDC, each season the onset of the ILI epidemic is detected if the measured ILI activity is at least two consecutive weeks above the threshold for epidemic activity. The threshold has been taken from (VEGA ET AL., 2013), where it has been determined for various European countries based on the moving epidemic method (MEM). Although the published ILI activity by ECDC is usually slightly updated one week later, it is assumed that the final ILI incidence has always been published on the following Wednesday.

For Influenzanet, the ILI epidemic is detected if the measured ILI incidence is for 14 consecutive days above the threshold for epidemic activity. The threshold for each day is based on the baseline incidence and the number of active participants on that day. Real-time monitoring is simulated by creating a snapshot of the database as it existed on each day of the season. The detection day is the first day on which based on the available data the epidemic would have been detected.

In real-time monitoring, the preliminary ILI incidence over the most recent weeks is based on fewer participants, and therefore the threshold is higher. For countries where most participants report every week (Netherlands and Belgium, Figure 6.11), the preliminary ILI incidence does not contain a consistent bias compared to the final ILI incidence, but in countries where many

participants do not report every week, the preliminary ILI incidence is on average over estimated (data not shown).

Table 6.4 lists for all seasons when the epidemic was detected by Influenzanet and by ECDC. In the Netherlands, Influenzanet detected the onset of the ILI epidemic on average two weeks earlier than ECDC, and in the season 2011–2012 no epidemic was detected by ECDC. In 2007–2008 the ILI epidemic was detected 6 weeks earlier than ECDC, since Influenzanet measured two ILI peaks, whereas ECDC only measured the second.

In the Netherlands and Belgium the onset of the ILI epidemic was detected significantly earlier by ECDC during the pandemic season 2009–2010. This could be related to increase anxiety in the population, leading to an increased rate at which participants visit their GP (Keramarou et al., 2011). In (de Lange et al., 2013) the ILI detection capacity for Influenzanet, ECDC, and three other systems are compared during the season 2009–2010, and in (Vega et al., 2013) the performance of the moving epidemic method (MEM) is tested for all countries for the season 2009–2010. Table 6.4 indicates that the pandemic season might not have been the most typical season to evaluate the performance of ILI surveillance systems.

In Belgium, Influenzanet and ECDC detected the ILI epidemic on average in the same week, and during in the season 2010–2011 no epidemic was detected by Influenzanet. During the seasons 2003–2005 Influenzanet had significantly less participants (Figure 3.5, Chapter 3, page 47), which explains the later detection of the onset of the ILI epidemic.

In Portugal, Influenzanet detected the ILI epidemic later than ECDC, and during three out of eight seasons (2005–2006, 2007–2008 and 2012–2013) no onset of the ILI epidemic was detected by Influenzanet.

Table 6.4: DETECTION TIMES OF THE ILI EPIDEMIC by Influenzanet and ECDC in the Netherlands, Belgium, and Portugal (2003–2013).

| COUNTRY | SEASON | INFLUENZANET | ECDC | DIFFERENCE |
|---|---|---|---|---|
| Netherlands | 2003/04 | Fri 12 Dec | Wed 24 Dec | 12 |
| | 2004/05 | Tue 8 Feb | Wed 16 Feb | 8 |
| | 2005/06 | Thu 16 Feb | Wed 15 Feb | –1 |
| | 2006/07 | Thu 22 Feb | Wed 14 Mar | 20 |
| | 2007/08 | Fri 11 Jan | Wed 27 Feb | 47 |
| | 2008/09 | Thu 25 Dec | Wed 14 Jan | 20 |
| | 2009/10 | Tue 3 Nov | Wed 21 Oct | –13 |
| | 2010/11 | Mon 3 Jan | Wed 19 Jan | 16 |
| | 2011/12 | Fri 24 Feb | * | * |
| | 2012/13 | Tue 1 Jan | Wed 16 Jan | 15 |
| Belgium | 2003/04 | Tue 9 Dec | Wed 26 Nov | –13 |
| | 2004/05 | Mon 7 Feb | Wed 2 Feb | –5 |
| | 2005/06 | Sun 19 Feb | Wed 22 Feb | 3 |
| | 2006/07 | Thu 15 Feb | Wed 14 Feb | –1 |
| | 2007/08 | Tue 8 Jan | Wed 23 Jan | 15 |
| | 2008/09 | Sun 18 Jan | Wed 21 Jan | 3 |
| | 2009/10 | Wed 4 Nov | Wed 7 Oct | –28 |
| | 2010/11 | * | Wed 29 Dec | * |
| | 2011/12 | Fri 17 Feb | Wed 22 Feb | 5 |
| | 2012/13 | Fri 1 Feb | Wed 16 Jan | –16 |
| Portugal | 2005/06 | * | Wed 1 Mar | * |
| | 2006/07 | Fri 2 Feb | Wed 31 Jan | –2 |
| | 2007/08 | * | Wed 16 Jan | * |
| | 2008/09 | Sat 13 Dec | Wed 10 Dec | –3 |
| | 2009/10 | Sun 8 Nov | Wed 11 Nov | 3 |
| | 2010/11 | Thu 20 Jan | Wed 22 Dec | –29 |
| | 2011/12 | Tue 14 Feb | Wed 15 Feb | 1 |
| | 2012/13 | * | Wed 30 Jan | * |

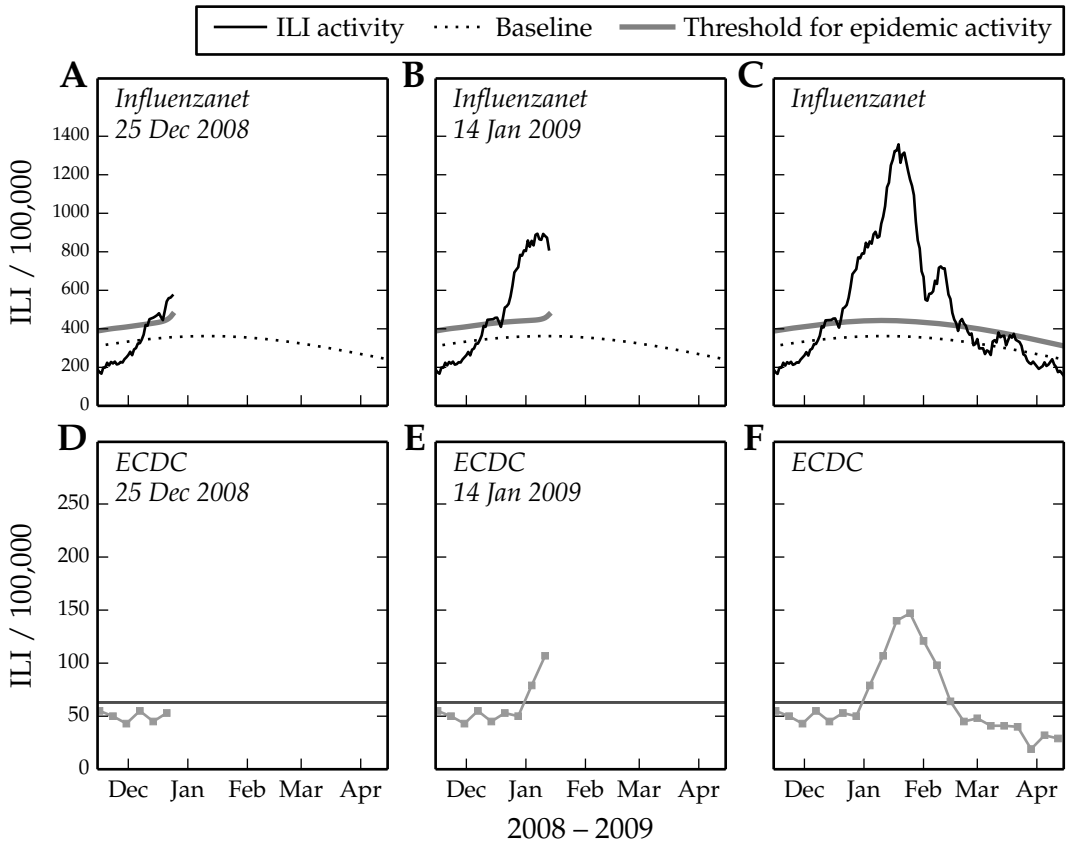



Figure 6.16: ILI EPIDEMIC DETECTION IN THE NETHERLANDS (2008–2009) by A–C) Influenzanet, and D–F) ECDC. The ILI incidence by Influenzanet is defined as the number of ILI onsets per 100,000 participant-weeks. The ILI incidence by ECDC is defined as the weekly number of patient visits with ILI per 100,000. A,D) date the ILI epidemic was detected by Influenzanet. B,E) date the ILI epidemic was detected by ECDC. C,F) Complete time series at the end of the season. Note that for Influenzanet the threshold for epidemic activity is slightly higher for the most recent days in real-time monitoring.

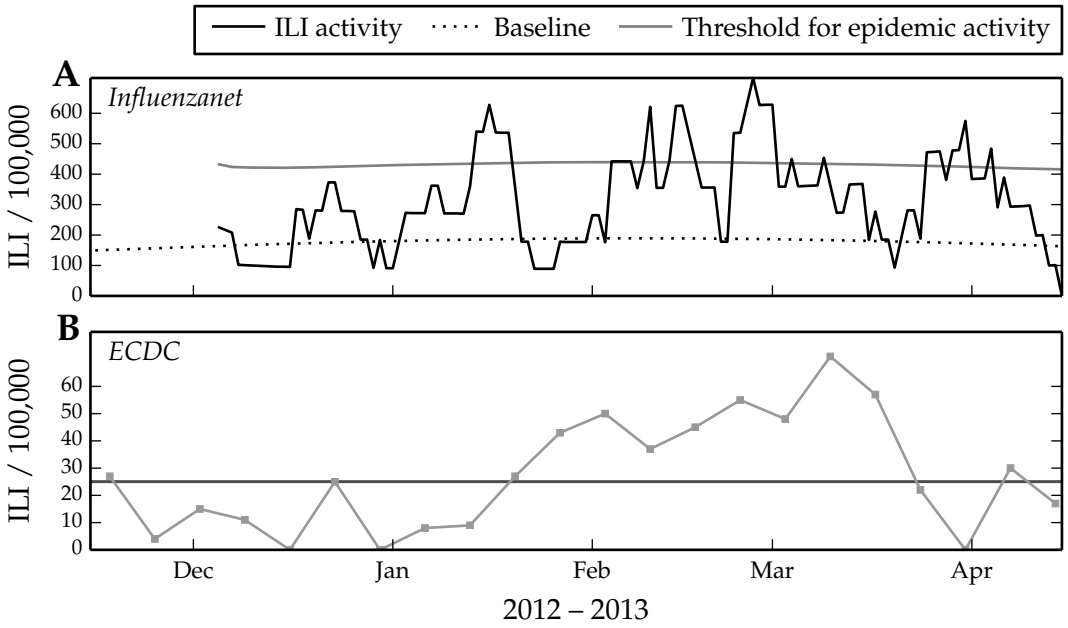



Figure 6.17: Detection of the ILI epidemic in Portugal (2012–2013) by A) Influenzanet, and B) ECDC. The ILI incidence by Influenzanet is defined as the number of ILI onsets per 100,000 participant-weeks. The ILI incidence by ECDC is defined as the weekly number of patient visits with ILI per 100,000. The ILI epidemic was not detected by Influenzanet.

Figure 6.16 shows the measured ILI activity by both systems in the Netherlands during the season 2008–2009 — a typical season in Table 6.4 — with snapshots of the available data on the days the ILI epidemic was detected by either system. Influenzanet was able to detect the onset of the ILI epidemic 2–3 weeks earlier than ECDC.

Figure 6.17 shows the measured ILI activity by both systems in Portugal during the season 2012–2013. Although Influenzanet measured increased ILI activity during the same time an ILI epidemic was detected by ECDC, the ILI activity was never for longer than two weeks above the threshold for epidemic activity and could therefore not be distinguished significantly from baseline ILI incidence.

#### 6.3.3.5 *Perspective*

Not too much emphasis should be put on the exact differences in detection times between both systems (Table 6.4), since the thresholds for epidemic activity are determined by different methods. However, it seems that given enough participants, such as in the Netherlands, Influenzanet is able to detect the onset of the ILI epidemic up to 2 weeks earlier than ECDC. For countries with relatively few participants, such as in Portugal, Influenzanet cannot detect the onset of an ILI epidemic earlier.

The capacity for earlier detection of the onset of the ILI epidemic based on identifying a threshold, is probably not the most consistent advantage for a self-reporting system such as Influenzanet (Section 6.1). If sentinel GPs would start to report the latest ILI data daily and electronically, most time-advantage for Influenzanet would probably disappear. However, based on more sophisticated methods, taking full advantage of all available symptoms data and personal characteristics, Influenzanet might be able to detect signals

which predict the imminent onset of an ILI epidemic in the whole population (see also Section 6.3.4).

### 6.3.4 *The challenge of establishing a paradigm shift in clinical surveillance*

In standard analyses, all symptom data are basically summarized to determining if a participant has ILI or not, based on a certain ILI case definition. The ILI rate can be corrected for the whole population based on individual data from the intake questionnaire. However, to reduce all the data collected by Influenzanet to a single weekly ILI incidence, hides a lot of the system's potential.

Various different ILI case definitions can be applied in parallel and retrospectively. Although case definitions are generally only based on symptoms, even more sophisticated definitions can be applied. For example, age and chronic conditions can be included, and even external factors such as current weather (see Chapter 4) can be considered. Furthermore, the ILI incidence can be determined independently in various critical subgroups, such as those who are vaccinated and persons with high contact rates.

Influenzanet and ECDC monitor influenza-like illness, since it is easier (and less expensive) to determine the incidence of a clinical syndrome. However, the incidence of influenza infections is arguably more interesting. Since not all persons infected with influenza exhibit the same symptoms (LAU ET AL., 2010), and infections with other viruses such as RSV could lead to symptoms similar to influenza, the ideal clinical case definition for influenza that applies universally, does not exists. Many different ILI case definitions are in use by the sentinel GPs throughout Europe (AGUILERA ET AL., 2003). By applying a relatively weak ILI case definition, many people who fit the ILI case definition might be infected with something other than influenza, whereas by applying a relatively strong ILI case definition, persons with a mild influenza infection would not fit the case definition.

In (PATTERSON-LOMBA ET AL., 2014) the attack rate of real influenza infections is estimated for the Netherlands in 2012–2013. It assumes that only ILI cases above a determined baseline in Influenzanet are due to influenza infection, and that from all participants with influenza only a fraction fits the ILI case definition. This fraction is estimated from a study which monitored daily symptoms in persons with RT-PCR confirmed influenza infections (LAU ET AL., 2010),

However, instead of relying on a single ILI case definition, Influenzanet has the power to determine, in real-time and retrospectively, the incidence of various case definitions $I_i$ simultaneously. For example, it might be possible to choose these case definitions such that any participant will fit one and only case definition, such that the real influenza incidence $Flu(t)$ can be estimated by:

$$Flu(t) = I_1(t)P(Flu|I_1, t) + \ldots + I_n(t)P(Flu|I_n, t)$$

where $P(Flu|I_n, t)$ is the probability that a participant in week $t$ fitting the case definition for $I_n$ has an actual influenza infection. In general, the more ILI-symptoms a participant reports, the more likely that this participant has a real influenza infection, whereas during weeks when RSV is circulating, this probability is lower. Although determining these various case definitions with corresponding probabilities could be an enormous challenge, the results collected by (LAU ET AL., 2010) and provided online[5] make substantial progress in that direction.

Such a probabilistic approach to surveillance, for which each participant is given a probability of having the infection of interest (in this case, influenza),

5 http://web.hku.hk/~bcowling/influenza/HK_NPI_study.htm

might give a more accurate estimation of the real attack rate of influenza in the population than any single clinical case definition.

## 6.4 INTEGRATION OF PATHOGEN DIVERSITY WITH DISEASE EPIDEMIOLOGY

For both influenza H3N2 and *Plasmodium falciparum*, the pathogen diversity is critical in understanding the epidemiology of these infectious diseases.

For Influenza A H3N2, each year new strains arise due to mutation, which gradually replace the older strains and generating a ladder-like phylogenetic tree. Various distinct mechanisms have been proposed to explain the characteristic shape of influenza phylogenetics, such as a short-term full cross-immunity among strains (FERGUSON ET AL., 2003), a low prevalence during summer (VAN NOORT, 2005), a specific genotype-to-phenotype mapping (KOELLE ET AL., 2006), and a narrow immune response in children (PARISI ET AL., 2013). All proposed models have in common that they implement a mutation rate for the influenza virus, cross-immunity among strains, and seasonality in the transmission rate. Although the models aim to reproduce seasonal epidemics, none of the models is actually fitted to epidemiological observations, due to the complexity of required statistics.

The objective of the model in Chapter 4 and Section 6.2 was not to explain the typical antigenic diversity of influenza H3N2, but to determine the role of weather elements on the yearly epidemics of influenza-like illness. Influenza transmission models are fitted to actual observations of influenza-like illness and meteorological data on temperature and absolute humidity (Figure 4.5, Chapter 4, page 80). The typical antigenic diversity of influenza is already included into the model, although indirectly, by assuming full cross-immunity for a single season whereas for each new season the model estimates independently how many people will have become susceptible again to the new circulating strains.

*Plasmodium falciparum* is also characterized by immense antigenic diversity. During the blood phase of infection, the parasite expresses certain variant surface antigens (VSAs) of the PfEMP1 family on the surface of infected red blood cells. Each malaria parasite has the ability to express around 60 different VSAs and generally only one is predominantly expressed at a given time in a way that appears orchestrated by the host immune system. After a specific immune response has been mounted, the parasite is either cleared or switches expression to another VSA. A relatively conserved subset of these VSAs is preferentially expressed in non-immune patients and related to severe malaria, whereas more diverse subsets are related to uncomplicated malaria. Empirical evidence suggests that most parasites are able to express VSAs from this conserved subset.

Studies which aim to reproduce epidemiological observations can implement some effects of antigenic diversity of *Plasmodium falciparum* already into the model. In (AGUAS ET AL., 2008) a model is fitted to age-dependent attack rates of hospitalizations of (severe) malaria in regions of various transmission intensity. In the model a naive host which is exposed to the parasite will develop severe malaria, which corresponds to the expression of a VSA of the restricted subset. Each subsequent exposure will cause mild malaria, which corresponds to the expression of a VSA from the more diverse subsets.

The objective of Chapter 5 was to generate the typical antigenic diversity in the parasite population while reproducing observed serological trends. The main hypothesis for the model is that all VSAs in the parasite population are organized into blocks of antigenically related variants, characterized by their dominance, and that each parasite can express a limited number of these VSA blocks. Upon infection the parasite will express the most dominant VSA block

for which the host has no prior immunity, and expression of a more dominant VSA block increases the probability of severe malaria.

This initial model was able to reproduce the serological observations that infected erythrocytes from people with either severe malaria (dominant VSA block) or with a small antibody repertoire, are more often recognized by serum of other people in the population. Due to the increased immunity in the population against the dominant VSA blocks, the model predicts a selective force against these dominant VSA blocks, creating a competitive advantage for parasites which only encode for low-dominance VSAs.

Experimental studies show that mice infected with a low virulence strain of *Plasmodium chabaudi* can be superinfected if exposed to a more virulent strain (DE ROODE ET AL., 2005*a*,*b*). This inspired the extension of the model, such that an already infected host can be superinfected if exposed to a parasite which is able to express a more dominant VSA block, where the probability of superinfection depends on the difference in dominance. This mechanism generates a competitive advantage for parasites which are able to express high-dominance VSAs.

The two opposing selective forces generate a parasite population in which most parasites contain both a restricted block of high dominance VSAs and a increasingly diverse repertoire of low-dominance VSAs (Figure 5.4, Chapter 5, page 103), while still reproducing the serological observations (Figure 5.2B,C, Chapter 5, page 100).

The VSAs from the selected subset related to severe malaria — the dominant VSA block in the model — are being considered important targets for vaccine development (HVIID, 2010). Targeting only specific strains could have an impact on the disease epidemiology. As an example, pneumococcal disease is caused by ~90 different *Streptococcus pneumoniae* serotypes, and 7 of

the most commonly found to cause invasive disease in children are targeted by the PCV7 vaccine. The introduction in 2000 of the PCV7 vaccine not only led to a reduction in the rate of carriage of the vaccine serotypes, but the incidence of the non-vaccine serotypes also increased (HICKS ET AL., 2007). Mathematical models can play an important role in predicting how vaccination might impact the diversity of multi-type pathogens and disease.

## 6.5 REFERENCES

AGUAS R, WHITE LJ, SNOW RW, and GOMES MGM (2008). Prospects for malaria eradication in sub-saharan africa. *PLoS One*, 3(3):e1767.

AGUILERA JF, PAGET WJ, MOSNIER A, HEIJNEN ML, UPHOFF H, ET AL. (2003). Heterogeneous case definitions used for the surveillance of influenza in europe. *Eur J Epidemiol*, 18(8):751–754.

BACAËR N and AIT DADS EH (2011). Genealogy with seasonality, the basic reproduction number, and the influenza pandemic. *J Math Biol*, 62(5):741–762.

BAR-OR O, NEUMAN I, and DOTAN R (1977). Effects of dry and humid climates on exercise-induced asthma in children and preadolescents. *J Allergy Clin Immunol*, 60(3):163–168.

BEYER WEP, MCELHANEY J, SMITH DJ, MONTO AS, NGUYEN-VAN-TAM JS, ET AL. (2013). Cochrane re-arranged: support for policies to vaccinate elderly people against influenza. *Vaccine*, 31(50):6030–6033.

BOLLAG U (2009). Asthma data from the swiss sentinel surveillance network, 1989-2005 - from monitoring to research. *Swiss Med Wkly*, 139(39-40):571–575.

BRETÓ C, HE D, IONIDES EL, and KING AA (2009). Time series analysis via mechanistic models. *Ann. Appl. Stat.*, 3:319–348.

BUTLER D (2013). When google got flu wrong. *Nature*, 494(7436):155–156.

CANNELL JJ, VIETH R, UMHAU JC, HOLICK MF, GRANT WB, ET AL. (2006). Epidemic influenza and vitamin d. *Epidemiol Infect*, 134(6):1129–1140.

Carrat F, Flahault A, Boussard E, Farran N, Dangoumau L, et al. (1998). Surveillance of influenza-like illness in france. the example of the 1995/1996 epidemic. *J Epidemiol Community Health*, 52 Suppl 1:32S–38S.

Cauchemez S, Valleron AJ, Boëlle PY, Flahault A, and Ferguson NM (2008). Estimating the impact of school closure on influenza transmission from sentinel data. *Nature*, 452(7188):750–754.

Christakis NA and Fowler JH (2010). Social network sensors for early detection of contagious outbreaks. *PLoS One*, 5(9):e12948.

Chunara R, Aman S, Smolinksi M, and Brownstein JS (2012). Flu near you: An online self-reported influenza surveillance systems in the usa. *Online Journal of Public Health Informatics*, 5(1):e133.

Cooper DL, Smith GE, Chinemana F, Joseph C, Loveridge P, et al. (2008). Linking syndromic surveillance with virological self-sampling. *Epidemiol Infect*, 136(2):222–224.

Costagliola D, Flahault A, Galinec D, Garnerin P, Menares J, et al. (1991). A routine tool for detection and assessment of epidemics of influenza-like syndromes in france. *Am J Public Health*, 81(1):97–99.

Cowling BJ, Wong IOL, Ho LM, Riley S, and Leung GM (2006). Methods for monitoring influenza surveillance data. *Int J Epidemiol*, 35(5):1314–1321.

Dalton C, Durrheim D, Fejsa J, Francis L, Carlson S, et al. (2009). Flutracking: a weekly australian community online survey of influenza-like illness in 2006, 2007 and 2008. *Commun Dis Intell Q Rep*, 33(3):316–322.

de Lange MMA, Meijer A, Friesema IHM, Donker GA, Koppeschaar CE, et al. (2013). Comparison of five influenza surveillance systems during the 2009 pandemic and their association with media attention. *BMC Public Health*, 13:881.

Demicheli V, Jefferson T, Al-Ansary LA, Ferroni E, Rivetti A, et al. (2009). Vaccines for preventing influenza in healthy adults. *Cochrane Database Syst Rev*, 3.

Dowell SF (2001). Seasonal variation in host susceptibility and cycles of certain infectious diseases. *Emerg Infect Dis*, 7(3):369–374.

Eames KTD, Tilston NL, Brooks-Pollock E, and Edmunds WJ (2012). Measured dynamic social contact patterns explain the spread of h1n1v influenza. *PLoS Comput Biol*, 8(3):e1002425.

Eccles R (2002). An explanation for the seasonality of acute upper respiratory tract viral infections. *Acta Otolaryngol*, 122(2):183–191.

ELLIOT AJ, POWERS C, THORNTON A, OBI C, HILL C, ET AL. (2009). Monitoring the emergence of community transmission of influenza a/h1n1 2009 in england: a cross sectional opportunistic survey of self sampled telephone callers to nhs direct. *BMJ*, 339:b3403.

FERGUSON NM, GALVANI AP, and BUSH RM (2003). Ecological and immunological determinants of influenza evolution. *Nature*, 422(6930):428–433.

FRIESEMA IHM, KOPPESCHAAR CE, DONKER GA, DIJKSTRA F, VAN NOORT SP, ET AL. (2009). Internet-based monitoring of influenza-like illness in the general population: experience of five influenza seasons in the netherlands. *Vaccine*, 27(45):6353–6357.

FUHRMANN C (2010). The effects of weather and climate on the seasonality of influenza: What we know and what we need to know. *Geography Compass*, 4(7):718–730.

GARTEN RJ, DAVIS CT, RUSSELL CA, SHU B, LINDSTROM S, ET AL. (2009). Antigenic and genetic characteristics of swine-origin 2009 a(h1n1) influenza viruses circulating in humans. *Science*, 325(5937):197–201.

GHEDIN E, WENTWORTH DE, HALPIN RA, LIN X, BERA J, ET AL. (2010). Unseasonal transmission of h3n2 influenza a virus during the swine-origin h1n1 pandemic. *J Virol*, 84(11):5715–5718.

GINSBERG J, MOHEBBI MH, PATEL RS, BRAMMER L, SMOLINSKI MS, ET AL. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014.

HE D, DUSHOFF J, DAY T, MA J, and EARN DJD (2013). Inferring the causes of the three waves of the 1918 influenza pandemic in england and wales. *Proc Biol Sci*, 280(1766):20131345.

HICKS LA, HARRISON LH, FLANNERY B, HADLER JL, SCHAFFNER W, ET AL. (2007). Incidence of pneumococcal disease due to non-pneumococcal conjugate vaccine (pcv7) serotypes in the united states during the era of widespread pcv7 vaccination, 1998-2004. *J Infect Dis*, 196(9):1346–1354.

HOPE-SIMPSON RE (1979). Epidemic mechanisms of type a influenza. *J Hyg (Lond)*, 83(1):11–26.

HVIID L (2010). The role of plasmodium falciparum variant surface antigens in protective immunity and vaccine development. *Hum Vaccin*, 6(1):84–89.

JEFFERSON T, DI PIETRANTONJ C, AL-ANSARY LA, FERRONI E, THORNING S, ET AL. (2010). Vaccines for preventing influenza in the elderly. *Cochrane Database Syst Rev*, (2):CD004876.

KERAMAROU M, COTTRELL S, EVANS MR, MOORE C, STIFF RE, ET AL. (2011). Two waves of pandemic influenza a(h1n1) 2009 in wales–the possible impact of media coverage on consultation rates, april-december 2009. *Euro Surveill*, 16(3).

KOELLE K, COBEY S, GRENFELL B, and PASCUAL M (2006). Epochal evolution shapes the phylodynamics of interpandemic influenza a (h3n2) in humans. *Science*, 314(5807):1898–1903.

LANGE W and SCHÖTTLER M (2002). Real-time influenza surveillance in germany–results of a pilot project. *Med Microbiol Immunol*, 191(3-4):139–144.

LAU LLH, COWLING BJ, FANG VJ, CHAN KH, LAU EHY, ET AL. (2010). Viral shedding and clinical illness in naturally acquired influenza virus infections. *J Infect Dis*, 201(10):1509–1516.

LOWEN AC, STEEL J, MUBAREKA S, and PALESE P (2008). High temperature (30 degrees c) blocks aerosol but not contact transmission of influenza virus. *J Virol*, 82(11):5650–5652.

MANDEVILLE KL, O'NEILL S, BRIGHOUSE A, WALKER A, YARROW K, ET AL. (2014). Academics and competing interests in h1n1 influenza media reporting. *J Epidemiol Community Health*, 68(3):197–203.

MANN PG, PEREIRA MS, SMITH JW, HART RJ, and WILLIAMS WO (1981). A five-year study of influenza in families. joint public health laboratory service/royal college of general practitioners working group. *J Hyg (Lond)*, 87(2):191–200.

MARQUET RL, BARTELDS AIM, VAN NOORT SP, KOPPESCHAAR CE, PAGET J, ET AL. (2006). Internet-based monitoring of influenza-like illness (ili) in the general population of the netherlands during the 2003-2004 influenza season. *BMC Public Health*, 6:242.

PAOLOTTI D, CARNAHAN A, COLIZZA V, EAMES K, EDMUNDS J, ET AL. (2014). Web-based participatory surveillance of infectious diseases: the influenzanet participatory surveillance experience. *Clin Microbiol Infect*, 20(1):17–21.

PARISI A, LOPES JS, NUNES A, and GOMES MGM (2013). Heterogeneity in antibody range and the antigenic drift of influenza a viruses. *Ecological Complexity*, 14(0):157 – 165. ISSN 1476-945X.

PATTERSON-LOMBA O, VAN NOORT SP, COWLING BJ, WALLINGA J, GOMES MGM, ET AL. (2014). Utilizing syndromic surveillance data for estimating levels of influenza circulation. *Am J Epidemiol*, 179lh(11):1394–1401.

PELAT C, BOËLLE PY, COWLING BJ, CARRAT F, FLAHAULT A, ET AL. (2007). Online detection and quantification of epidemics. *BMC Med Inform Decis Mak*, 7:29.

Rehn M, Carnahan A, Merk H, Kühlmann-Berenzon S, Galanis I, et al. (2014). Evaluation of an internet-based monitoring system for influenza-like illness in sweden. *PLoS One*, 9(5):e96740.

de Roode JC, Helinski MEH, Anwar MA, and Read AF (2005*a*). Dynamics of multiple infection and within-host competition in genetically diverse malaria infections. *Am Nat*, 166(5):531–542.

de Roode JC, Pansini R, Cheesman SJ, Helinski MEH, Huijben S, et al. (2005*b*). Virulence and competitive ability in genetically diverse malaria infections. *Proc Natl Acad Sci U S A*, 102(21):7624–7628.

Rosenbaum PR (1991). Discussing hidden bias in observational studies. *Ann Intern Med*, 115(11):901–905.

Rosner B (2006). *Fundamentals of Biostatistics*. Brooks/Cole, Cengage Learning.

Sagripanti JL and Lytle CD (2007). Inactivation of influenza virus by solar radiation. *Photochem Photobiol*, 83(5):1278–1282.

Serfling RE (1963). Methods for current statistical analysis of excess pneumonia-influenza deaths. *Public Health Rep*, 78(6):494–506.

Shaman J, Pitzer VE, Viboud C, Grenfell BT, and Lipsitch M (2010). Absolute humidity and the seasonal onset of influenza in the continental united states. *PLoS Biol*, 8(2):e1000316.

Signorini A, Segre AM, and Polgreen PM (2011). The use of twitter to track levels of disease activity and public concern in the u.s. during the influenza a h1n1 pandemic. *PLoS One*, 6(5):e19467.

Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, et al. (2004). Mapping the antigenic and genetic evolution of influenza virus. *Science*, 305(5682):371–376.

Smolderen KGE, Vingerhoets AJJM, Croon MA, and Denollet J (2007). Personality, psychological stress, and self-reported influenza symptomatology. *BMC Public Health*, 7:339.

van Noort SP (2005). *The Genetic Drift of Influenza A*. Master's thesis, Utrecht University.

van Noort SP, Muehlen M, Rebelo de Andrade H, Koppeschaar CE, Lourenço JML, et al. (2007). Gripenet: an internet-based system to monitor influenza-like illness uniformly across europe. *Euro Surveill*, 12(7):E5–E6.

Vandendijck Y, Faes C, and Hens N (2013). Eight years of the great influenza survey to monitor influenza-like illness in flanders. *PLoS One*, 8(5):e64156.

VEGA T, LOZANO JE, MEERHOFF T, SNACKEN R, MOTT J, ET AL. (2013). Influenza surveillance in europe: establishing epidemic thresholds by the moving epidemic method. *Influenza Other Respi Viruses*, 7(4):546–558.

VIBOUD C, BOËLLE PY, PAKDAMAN K, CARRAT F, VALLERON AJ, ET AL. (2004). Influenza epidemics in the united states, france, and australia, 1972-1997. *Emerg Infect Dis*, 10(1):32–39.

# A

## FROM DYNAMICAL PROCESSES TO LIKELIHOOD FUNCTIONS

Sander P. van Noort[1], Nico Stollenwerk[2]

[1]Instituto Gulbenkian de Ciência, Oeiras, Portugal

[2] Universidade de Lisboa, Lisbon, Portugal

*Based on:*

*Proceedings of CMMSE 2008, ISBN 978-84-612-1982-7*

*Proceedings of CMMSE 2009, ISBN 978-84-612-9727-6*

ABSTRACT

*In epidemiological transmission processes, the stochasticity in the observed data can be due to the observation process or due to the dynamics of the system. In case of the latter, the master equation framework can be applied to deduce likelihood functions for parameter estimation.*

*For more simple epidemiological processes, we derive complete analytic expressions for the likelihood function and the best parameter estimators. For more elaborate models, such as the full SIR model, we show how the likelihood function can be expressed in terms of the solution to a partial differential equation.*

*Likelihood functions are commonly simulated numerically. The use of full or partial analytic solutions, could increase the accuracy of the parameter estimation.*

## A.1 INTRODUCTION

To model the stochasticity in epidemiological dynamic processes, the master equation framework (VAN KAMPEN, 1992; TOMÉ AND DE OLIVEIRA, 2001) can be used to determine likelihood functions. Solutions for the master equation can be achieved via generating functions or characteristic functions, the choice between the two often determined by ease or difficulty of solving the PDE and the subsequent back transformation to the original probabilities (VAN KAMPEN, 1992; BHARUCHA-REID, 1960). The likelihood functions can be maximized for parameter estimation and data fitting (HONERKAMP, 1993).

The likelihood functions for epidemiological processes are commonly determined by complete numerical simulation (STOLLENWERK AND BRIGGS, 2000). Also approximations such as the Poisson approximation (GUSTAFSSON AND STERNAD, 2007) over small time intervals are applied, which becomes less and less accurate for longer integration times.

Evaluation of the fitting process can be obtained via Fisher information (STOLLENWERK ET AL., 2001) or in the Bayesian framework via the posterior distribution for the parameters (SIVIA, 1996). Hypothesis testing and model selection can be achieved by the Kolomgorov-Smirnov test, either numerically (PRESS ET AL., 1992) or analytically (RÉNYI, 1962). A comparison of various models to one time series in a population biological context is demonstrated in (STOLLENWERK ET AL., 2001).

In this work, we will give the master equation for various epidemiological processes, starting with the complete SIR model to the more simple processes. We show how the likelihood function can be expressed in terms of the solution to a partial differential equations, and derive a complete analytic solution for the likelihood function when the PDE is analytically solvable.

## A.2 METHOD AND RESULTS

For various epidemiological processes, we will give the reaction scheme and the corresponding master equation. Based on the generating function, we will derive a partial differential equation, whose solution can be used to solve the master equation and the likelihood function. For the processes for which the PDE is solvable, we derive the solution and the complete expression for the likelihood function, and the best parameter estimators where feasible. Solutions from the more complex processes are used as a starting points for the more simple processes.

The dynamics of the mean values are the starting point of deterministic modeling. In the final section, we show how the dynamics of the mean follows directly from the master equation and its solution.

### A.2.1 *SIR model*

We first consider the basic epidemic process known as the SIR system, an important system for the analysis of recurrent outbreaks in seasonal influenza (CASAGRANDI ET AL., 2006). Susceptible individuals $S$ become infected on contact with already infected $I$ with infection rate $\beta$, and recover with rate $\gamma$ into the recovery $R$ class. The SIR system is given by the reaction scheme:

$$S + I \xrightarrow{\beta} I + I$$
$$I \xrightarrow{\gamma} R$$

for a fixed population size $S + I + R = N$. The master equation for the variables $S$ and $I$ is given by:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\, p(S,I,t) = &\frac{\beta}{N}(I-1)(S+1)\, p(S+1,I-1,t)\\ &+\gamma(I+1)\; p(S,I+1,t)\\ &-\frac{\beta}{N}SI\; p(S,I,t)\\ &-\gamma I\; p(S,I,t)\end{aligned} \tag{A.1}$$

where $p(S,I,t)$ is the probability to have $S$ susceptibles and $I$ infected at time $t$, with initial condition $p(S,I,t_0) = \delta_{S=S_0} \cdot \delta_{I=I_0}$.

To solve the master equation, we define the generating function $f$ as:

$$f(x,y,t) = \langle x^I y^S \rangle = \sum_{I=0}^{N}\sum_{S=0}^{N} x^I y^S\; p(S,I,t) \tag{A.2}$$

The probability $p(I,S,t)$ can be written in terms of the generating function as:

$$p(I,S,t) = \frac{1}{I!\; S!}\, \frac{\partial^{(I+S)} f(x,y,t)}{\partial x^I \partial y^S}\bigg|_{x=0,y=0} \tag{A.3}$$

which follows from the Taylor's expansion of the generating function (A.2) at $(x,y) = (0,0)$:

$$f(x,y,t) = \sum_I \sum_S \frac{x^I y^S}{I!\; S!}\, \frac{\partial^{(I+S)} f(x,y,t)}{\partial x^I \partial y^S}\bigg|_{x=0,y=0}$$

To find a solution for the generating function $f$, we first note that the first and second-order partial derivatives are given by:

$$\frac{\partial f(x,y,t)}{\partial x} = \sum_{I=0}^{N}\sum_{S=0}^{N} I x^{I-1} y^S\; p(S,I,t) \tag{A.4}$$

and

$$\frac{\partial f(x,y,t)}{\partial y} = \sum_{I=0}^{N}\sum_{S=0}^{N} Sx^I y^{S-1}\ p(S,I,t) \tag{A.5}$$

and

$$\frac{\partial^2 f(x,y,t)}{\partial x \partial y} = \sum_{I=0}^{N}\sum_{S=0}^{N} ISx^{I-1} y^{S-1}\ p(S,I,t) \tag{A.6}$$

The time dynamics for the generating function $f$ is given by

$$\frac{\partial}{\partial t} f(x,y,t) = \sum_{I=0}^{N}\sum_{S=0}^{N} x^I y^S\ \frac{\mathrm{d}}{\mathrm{d}t} p(S,I,t)$$

and by inserting the master equation (A.1) and substituting the partial derivatives (A.4), (A.5) and (A.6) we arrive at the following partial differential equation for the generating function:

$$\frac{\partial f}{\partial t} = \frac{\beta}{N} x \cdot (x-y)\frac{\partial^2 f}{\partial x \partial y} + \gamma(1-x)\frac{\partial f}{\partial x} \tag{A.7}$$

with initial condition $f(x,y,t_0) = x^{I_0} y^{S_0}$.

For observed data points $S$ and $I$ at times $t_0, \ldots, t_n$, the likelihood function $L(\beta,\gamma)$ is defined as the joint probability of the observations:

$$\begin{aligned} L(\beta,\gamma) &= P(S_n, I_n, t_n, \ldots, S_0, I_0, t_0) \\ &= \prod_{\nu=0}^{n-1} p(S_{\nu+1}, I_{\nu+1}, t_{\nu+1} | S_\nu, I_\nu, t_\nu) \cdot p(S_0, I_0, t_0) \end{aligned} \tag{A.8}$$

which can be maximized to obtain the most likely parameter values given the observed data. No analytic solution for the PDE (A.7) exists, so numerical methods have to be applied.

### A.2.2 *Onset of the epidemic*

We will now explore a simpler epidemic process in which there is an abundant number of susceptible individuals $S^*$. This model is an approximation of the initial exponential phase of an SIR epidemic, in which the number of infected individual is still relatively small, such that the pool of susceptibles can be considered constant. The reaction scheme is given by:

$$S^* + I \xrightarrow{\beta} I + I$$
$$I \xrightarrow{\gamma} R$$

where $S^*$ denotes the constant size of the pool of susceptibles. The master equation is given by:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}p(I,t) = \frac{\beta}{N}S^*(I-1)p(I-1,t) + \gamma(I+1)\ p(I+1,t) \\ - \left(\frac{\beta}{N}S^*I + \gamma I\right)\ p(I,t)\end{aligned}$$

This process is similar to the SIR system, where the probability $p$ is truncated by:

$$p(S,I,t) = \delta_{S=S^*}\ p(I,t)$$

such that:

$$f(x,y,t) = y^{S^*}\sum_{I=0}^{N} x^I\ p(I,t)$$

By defining a new generating function $g$ as

$$g(x,t) = \langle x^I \rangle = f(x,y,t)|_{y=1}$$

the PDE for the SIR system (A.7) is simplified to:

$$\frac{\partial g}{\partial t} = (1-x)(\gamma - \tilde{\beta}x)\frac{\partial g}{\partial x} \tag{A.9}$$

where $\tilde{\beta} = \frac{\beta S^*}{N}$ and initial condition $g(x, t_0) = x^{I_0}$.

We can solve this PDE analytically by introducing an arbitrary function $\Phi$ such that $g(x,t) = \Phi(z)$, and the separation of variables $z(x,t) = u(x) \cdot v(t)$:

$$g(x,t) = \Phi(z) = \Phi(u(x) \cdot v(t))$$

The direct partial derivative to time gives:

$$\begin{aligned}\frac{\partial}{\partial t}g(x,t) &= \frac{\mathrm{d}\Phi}{\mathrm{d}z} \cdot \frac{\partial z}{\partial t} \\ &= \frac{\mathrm{d}\Phi}{\mathrm{d}z} \cdot u(x)\frac{\partial v}{\partial t}\end{aligned}$$

while inserting the PDE (A.9) and determining the partial derivative to time gives:

$$\begin{aligned}\frac{\partial}{\partial t}g(x,t) &= (1-x)(\gamma - \tilde{\beta}x)\frac{\partial f}{\partial x} \\ &= (1-x)(\gamma - \tilde{\beta}x)\frac{\mathrm{d}\Phi}{\mathrm{d}z}\frac{\partial z}{\partial x} \\ &= (1-x)(\gamma - \tilde{\beta}x)\frac{\mathrm{d}\Phi}{\mathrm{d}z} \cdot \frac{\partial u}{\partial x}v(t)\end{aligned}$$

This separates the PDE (A.9) into two ODEs for $v(t)$ and $u(x)$:

$$\begin{aligned}\frac{\mathrm{d}v}{\mathrm{d}t} &= v(t) \\ \frac{\mathrm{d}u}{\mathrm{d}x} &= \frac{u(x)}{(1-x)(\gamma - \tilde{\beta}x)}\end{aligned}$$

and arbitrary function $\Phi(z)$.

Integration of the ODEs gives the special solutions

$$v(t) = e^t$$
$$u(x) = \left(\frac{x-1}{x-\frac{\gamma}{\tilde{\beta}}}\right)^{\frac{1}{\tilde{\beta}-\gamma}}$$

and as solution for the separation of variables

$$z(x,t) = \left(\frac{x-1}{x-\frac{\gamma}{\tilde{\beta}}}\right)^{\frac{1}{\tilde{\beta}-\gamma}} e^t \tag{A.10}$$

and the inverse relation

$$x = \frac{\frac{\gamma}{\tilde{\beta}}\left(ze^{-t}\right)^{\tilde{\beta}-\gamma} - 1}{\left(ze^{-t}\right)^{\tilde{\beta}-\gamma} - 1}$$

From the initial condition $g(x,t_0) = x^{I_0}$, it follows that

$$g(x,t_0) = \Phi(z(x,t_0)) = \left(\frac{\frac{\gamma}{\tilde{\beta}}\left(z(x,t_0)e^{-t}\right)^{\tilde{\beta}-\gamma} - 1}{\left(z(x,t_0)e^{-t}\right)^{\tilde{\beta}-\gamma} - 1}\right)^{I_0}$$

Now we take $z(x,t)$ from (A.10) for all times $t$ to obtain the general solution:

$$g(x,t) = \Phi(z(x,t)) = \left(\frac{\frac{\gamma}{\tilde{\beta}}(x-1)e^{(\tilde{\beta}-\gamma)(t-t_0)} - \left(x-\frac{\gamma}{\tilde{\beta}}\right)}{(x-1)e^{(\tilde{\beta}-\gamma)(t-t_0)} - \left(x-\frac{\gamma}{\tilde{\beta}}\right)}\right)^{I_0} \tag{A.11}$$

From the solution of the generating function, we can obtain the solution for the master equation similar to (A.3) as:

$$p(I,t) = \frac{1}{I!}\left.\frac{\partial^I g(x,t)}{\partial x^I}\right|_{x=0}$$

No explicit analytic solution for the $n^{th}$ derivative of the generating function $g$ exists. For calculating the likelihood function (A.8) and the maximum-likelihood, the derivative operator $\partial^{I_{\nu+1}}/\partial x^{I_{\nu+1}}$ should either by calculated numerically or by symbolic transformation computer programs. To determine the maximum likelihood for the two parameters, 2-dimensional Newton's method could be used.

### A.2.3 *Constant force of infection*

We will now explore the epidemic process, where a pool of susceptibles is exposed to a constant force of infection $\lambda$:

$$S \xrightarrow{\lambda} I$$

where the master equation is given by:

$$\frac{\mathrm{d}}{\mathrm{d}t}\, p(S,t) = \lambda(S-1)\; p(S-1,t) - \lambda S\; p(S,t)$$

This system is equivalent to the previous described SI model with a constant pool of susceptibles, setting $\beta = 0$ and swapping the variables $I$ and $R$ to $S$ and $I$, respectively. For a generating function $h(x,t)$ defined by

$$h(x,t) = \langle S^I \rangle$$

the PDE (A.7) is simplified to:

$$\frac{\partial h}{\partial t} = \lambda(1-x)\frac{\partial h}{\partial x}$$

while its solution (A.11) to

$$h(x,t) = \left(1 - (1-x)e^{-\lambda(t-t_0)}\right)^{S_0}$$

Via complete induction it follows that:

$$\frac{\partial^S f(x,t)}{\partial x^S} = \frac{S_0!}{(S_0-S)!}\left(e^{-\lambda(t-t_0)}\right)^S \left(1-(1-x)e^{-\lambda(t-t_0)}\right)^{S_0-S}$$

such that from (A.3) it follows that:

$$\begin{aligned} p(S,t) &= \frac{1}{S!}\left.\frac{\partial^S f(x,t)}{\partial x^S}\right|_{x=0} \\ &= \binom{S_0}{S}\left(e^{-\lambda(t-t_0)}\right)^S \left(1-e^{-\lambda(t-t_0)}\right)^{S_0-S} \end{aligned} \tag{A.12}$$

For observed data points in a time series $(S_0, \ldots, S_n)$ at times $(t_0, \ldots t_n)$, the likelihood function (A.8) is given by:

$$L(\lambda) = \prod_{\nu=0}^{n-1} \binom{S_\nu}{S_{\nu+1}} \left(e^{-\lambda(t_{\nu+1}-t_\nu)}\right)^{S_{\nu+1}} \left(1-e^{-\lambda(t_{\nu+1}-t_\nu)}\right)^{S_\nu - S_{\nu+1}}$$

and the log-likelihood $\ell = \log L$ by:

$$\begin{aligned} \ell(\lambda) = \sum_{\nu=0}^{n-1} \big(&\log\Gamma(S_\nu+1) - \log\Gamma(S_{\nu+1}+1) - \log\Gamma(S_\nu - S_{\nu+1}+1) \\ &-\lambda S_{\nu+1}(t_{\nu+1}-t_\nu) + (S_\nu - S_{\nu+1})\log(1-e^{-\lambda(t_{\nu+1}-t_\nu)})\big) \end{aligned}$$

using the gamma function $x! = \Gamma(x+1)$. To obtain the most likely parameter values given the data, the log-likelihood function has to be maximized: $\mathrm{d}\ell/\mathrm{d}\lambda = 0$.

If we assume that the time steps are equally distributed with step $\Delta t$, the derivative to $\lambda$ is given by:

$$\frac{\mathrm{d}\ell}{\mathrm{d}\lambda} = \frac{1}{1 - e^{-\lambda \Delta t}} \sum_{\nu=0}^{n-1} (S_\nu - S_{\nu+1}) \cdot e^{-\lambda \Delta t} \cdot \Delta t - \sum_{\nu=0}^{n-1} (S_{\nu+1}) \cdot \Delta t$$

and from setting $\frac{\mathrm{d}\ell}{\mathrm{d}\lambda} = 0$, the best estimator $\hat{\lambda}$ is given by:

$$\hat{\lambda} = \frac{1}{\Delta t} \log \left( \frac{\sum_{\nu=0}^{n-1} S_\nu}{\sum_{\nu=0}^{n-1} S_{\nu+1}} \right)$$

From $S + I = N$, the estimator $\hat{\lambda}$ can also be expressed as a function of the cumulative number of infected $I$:

$$\hat{\lambda} = \frac{1}{\Delta t} \log \left( \frac{N - \frac{1}{n} \sum_{\nu=0}^{n-1} I_\nu}{N - \frac{1}{n} \sum_{\nu=0}^{n-1} I_{\nu+1}} \right) \tag{A.13}$$

### A.2.4 *Constant influx*

A very simple epidemic process is constant influx of new infected, represented as a Poisson counting process with constant rate. This can be interpreted as the noise floor of infected individuals before a seasonal outbreak, where the force of infection is constant and there is an abundant number of susceptibles. The reaction scheme is given by

$$\xrightarrow{\mu} I$$

where the master equation is given by:

$$\frac{\mathrm{d}}{\mathrm{d}t}\, p(I,t) = \mu\, p(I-1,t) - \mu\, p(I,t)$$

This system is equivalent to system with a constant force of infection and an abundant number of susceptibles $S^*$, such that $S^* \gg I$ and $\mu = \lambda S^*$. The probability function (A.12), by noting that $1 - e^{-\mu} \approx \mu$, simplifies to:

$$p(I, t|I_0, t_0) = \frac{(\mu(t - t_0))^{I-I_0}}{(I - I_0)!} e^{-\mu(t-t_0)}$$

The best estimator as a function of the data points (A.13) simplifies to:

$$\hat{\mu} = \frac{I_n - I_0}{t_n - t_0}$$

### A.2.5 *Solutions of the mean*

For the SIR model in section A.2.1, we obtain the moments of the variables $I$ and $S$ by evaluating the partial derivatives of the generating function at point $(x, y) = (1, 1)$,

$$\left.\frac{\partial f(x, y, t)}{\partial x}\right|_{x=1,y=1} = \sum_{I=0}^{N} \sum_{S=0}^{N} I \cdot p(S, I, t) = \langle I \rangle$$

and

$$\left.\frac{\partial f(x, y, t)}{\partial y}\right|_{x=1,y=1} = \sum_{I=0}^{N} \sum_{S=0}^{N} S \cdot p(S, I, t) = \langle S \rangle$$

and for the correlations

$$\left.\frac{\partial^2 f(x, y, t)}{\partial x \partial y}\right|_{x=1,y=1} = \sum_{I=0}^{N} \sum_{S=0}^{N} SI \cdot p(S, I, t) = \langle SI \rangle \quad .$$

The dynamics of the mean value for $S$ and $I$ can be obtained via the generating function (A.2) and its PDE (A.7):

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\langle S\rangle &= \frac{\mathrm{d}}{\mathrm{d}t}\left(\left.\frac{\partial f(x,y,t)}{\partial y}\right|_{x=1,y=1}\right)\\ &= \left.\left(\frac{\partial}{\partial y}\left(\frac{\partial f}{\partial t}\right)\right)\right|_{x=1,y=1}\\ &= -\frac{\beta}{N}\langle SI\rangle + \alpha(N - \langle S\rangle - \langle I\rangle)\end{aligned}$$

and similarly

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\langle I\rangle &= \frac{\mathrm{d}}{\mathrm{d}t}\left(\left.\frac{\partial f(x,y,t)}{\partial x}\right|_{x=1,y=1}\right)\\ &= \frac{\beta}{N}\langle SI\rangle - \gamma\langle I\rangle\end{aligned}$$

By inserting the mean field approximation $\langle SI\rangle - \langle S\rangle\langle I\rangle \approx 0$, which neglects higher moments, we arrive at the well known closed system ordinary differential equations:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\langle S\rangle &= -\frac{\beta}{N}\langle I\rangle\langle S\rangle\\ \frac{\mathrm{d}}{\mathrm{d}t}\langle I\rangle &= \frac{\beta}{N}\langle I\rangle\langle S\rangle - \gamma\langle I\rangle\end{aligned}$$

which are the starting point of deterministic modeling.

For the onset of the epidemic in section A.2.2, from (A.11) we can calculate the solution for the mean value as

$$\langle I(t)\rangle = \left.\frac{\partial f(x,t)}{\partial x}\right|_{x=1} = I_0\, e^{(\tilde{\beta}-\gamma)(t-t_0)}$$

giving an exponential time dependence.

## A.3 DISCUSSION

For the standard susceptible-infected-recovered (SIR) epidemic process (Section A.2.1), we have defined the likelihood function in terms of the solutions to a partial differential equation (PDE) of second order, for which no analytic solution exists. For simpler epidemic processes, such as the onset of the epidemic (section Section A.2.2), a constant force of infection (Section A.2.3) and the constant influx of infected (Section A.2.4), the master equation led to a solvable PDE and we determined the likelihood functions analytically.

The analytic expressions obtained here can be treated in future numeric work. For epidemiological models in which the partial differential equation can not be solved analytically we often have to use numerical methods to obtain likelihood functions and their maximums (Stollenwerk and Briggs, 2000; Stollenwerk et al., 2001; Stollenwerk and Mikolajczyk, 2006) on the basis of numerically simulating the respective master equations (Gillespie, 1976, 1978).

The determined likelihood function for the constant influx of infected has been used to distinguish the start of the yearly influenza outbreaks and the normal fluctuations in the noise floor of sporadic influenza cases (van Noort and Stollenwerk, 2008). The determined likelihood function for the epidemic onset of a new unknown pathogen, could lead to more accurate estimates of its infectiousness than the traditional deterministic model or numerical simulations.

Analytic solutions for the likelihood functions can only be determined for relatively simple epidemiological processes and assume no observation errors. Most epidemiological processes are governed by more complex models and have observational noise in the data, so rather than the direct application

of the likelihood functions, they could be incorporated into more elaborate algorithms for parameter estimation.

## ACKNOWLEDGMENTS

This work has been supported by the European Union under the Marie Curie grant MEXT-CT-2004-14338. We thank Luis Sanchez and Maíra Aguiar, both Lisbon, and Gabriela Gomes, Oeiras, for scientific support, Lewi Stone and Haggai Katriel, Tel Aviv, Frank Hilker, Bath, and Friedhelm Drepper, Jülich, for valuable discussion.

## SIGNIFICANT CHANGES COMPARED TO PUBLISHED ARTICLES

This appendix combines the analytic results from two articles published in conference proceedings. Instead of deriving the solutions to each model separately, the intermediate results from the more elaborate models are used as a starting point for the more simple models. All applications of the simplified models to Influenzanet data are removed.

## AUTHOR CONTRIBUTIONS

Sander van Noort and Nico Stollenwerk derived the analytic solutions. The original articles as published in the proceedings were written by Nico Stollenwerk and revised by Sander van Noort. Sander van Noort selected the relevant parts of both articles for this appendix, including a significant restructuring of the method/results and a newly written discussion.

## 1.4 REFERENCES


Bharucha-Reid AT (1960). *Elements of the theory of Markov Processes and their applications*. McGraw-Hill, New York.

Casagrandi R, Bolzoni L, Levin SA, and Andreasen V (2006). The sirc model and influenza a. *Math Biosci*, 200(2):152–169.

Gillespie DT (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22:403–434.

Gillespie DT (1978). Monte carlo simulation of random walks with residence time dependent transition probability rates. *Journal*, 28:395–407.

Gustafsson L and Sternad M (2007). Bringing consistency to simulation of population models–poisson simulation as a bridge between micro and macro simulation. *Math Biosci*, 209(2):361–385.

Honerkamp J (1993). *Stochastic Dynamical Systems: Concepts, Numerical Methods and Analysis*. VCH Publishers, Heidelberg, New York.

van Kampen NG (1992). *Stochastic Processes in Physics and Chemistry*. North-Holland, Amsterdam.

Press WH, Teukolsky SA, Vetterling WT, and Flannery BP (1992). *Numerical Recipes in C, The Art of Scientific Computing 2nd edition*. Cambrid.

Rényi A (1962). *Wahrscheinlichkeitsrechnung*. VEB Deutscher Verlag der Wissenschaften, Berlin.

Sivia DS (1996). *Data analysis: A Bayesian tutorial*. Oxford University Press.

Stollenwerk N and Briggs KM (2000). Master equation solution of a plant disease model. *Physics Letters A*, 274:84–91.

Stollenwerk N, Drepper F, and Siegel H (2001). Testing nonlinear stochastic models on phytoplankton biomass time series. *Ecological Modelling*, 144:261–277.

Stollenwerk N and Mikolajczyk R (2006). Algorithm for parameter estimation in nosocomial infection. *Mathematical Modeling of Biological Systems*, 2:23–34.

Tomé T and de Oliveira MJ (2001). *Dinâmica estocástica e irreversibilidade*. Editora da Universidade de São Paulo, São Paulo.

VAN NOORT SP and STOLLENWERK N (2008). From dynamical processes to likelihood functions: an epidemiological application to influenza. In *Proceedings of CMMSE 2008, ISBN 978-84-612-1982-7*.

# B

## SUPPLEMENTARY MATERIAL

*Science is facts. Just as houses are made of stones, so science is made of facts. But a pile of stones is not a house and a collection of facts is not necessarily science.*

Henri Poincaré

## B.1 TIME SERIES ANALYSES

This section contains the time series analyses for ILI activity from Influenzanet and ECDC, as a supplement to Chapter 3. For both the weekly ILI incidence of Influenzanet and ECDC, the week 27 is considered the first week of each season, and each season consists of 52 weeks. Since Influenzanet in Italy did never measure the ILI incidence during the summer weeks, in Italy the week 38 is considered the first week of each season, and each season consists of 35 weeks. Only seasons for which both Influenzanet and ECDC data are available are considered: for the Netherlands and Belgium seasons 2003–2013, for Portugal seasons 2005–2013 and for Italy seasons 2008–2013. The time series are plotted in Figures B.1–B.4A,C.

The Box-Jenkins method is used to find for each time series the Autoregressive integrated moving average (ARIMA) model which best fits the time series. As a first step, each time series is made stationary by differencing. Since the time series of Influenzanet and ECDC both show seasonality, each time series is first seasonally differenced. Next, each time series is tested for stationarity by both the augmented Dickey-Fuller (ADF) test and the Kwiatkowsky-Philips-Schmidt-Shin (KPSS) test. Since the (seasonally differenced) ILI series for Influenzanet in Italy fails the ADF test, this time series is also once ordinary differenced. The autocorrelation and partial autocorrelation plots for the raw and integrated data are plotted in Figures B.5–B.12A,B,D,E.

For the next step, for each time series various seasonal ARIMA models with different orders ($p$, $q$, $P$, and $Q$) are iteratively fitted using maximum likelihood. The model with the lowest Akaike information criterion (AIC) for which the residuals resemble white noise is selected. The Ljung-Box test is used to determine whether the residuals resemble white noise, where the

lag for the Ljung-Box test is set to $\min(10, T/5)$, with T being the frequency of the time series.[1] The parameters and order for the best model are listed in Table B.1. The residuals are plotted in Figures B.1–B.4B,D, and the autocorrelation and partial autocorrelation plots for the residuals are plotted in Figures B.5–B.12C,F.

All analyses were done using R version 3.1.0. The ARIMA fitting was done using the function *Arima* from the R package *forecast*. The ADF and KPSS tests were done using the functions *adf.test* and *kpss.test* from the R package *tseries*. The Ljung-Box test was done using the function *Box.test* from the R package *stats*.

1 http://robjhyndman.com/hyndsight/ljung-box-test/

Table B.1: ESTIMATED PARAMETERS FOR THE FITTED SEASONAL ARIMA MODELS . Parameters with a * are significant. SEE is the standard error of estimate.

| SOURCE | MODEL | SEE | AIC | LJUNG–BOX |
|---|---|---|---|---|
| Influenzanet (nl) | (2,0,1) $(0,1,1)_{52}$ | 116 | 2523 | 0.99 (NS) |
| ECDC (nl) | (3,0,0) $(1,1,1)_{52}$ | 15 | 2962 | 0.93 (NS) |
| Influenzanet (be) | (1,0,0) $(0,1,1)_{52}$ | 160 | 2627 | 0.31 (NS) |
| ECDC (be) | (4,0,0) $(0,1,1)_{52}$ | 51 | 4340 | 0.98 (NS) |
| Influenzanet (pt) | (3,0,0) $(0,1,1)_{52}$ | 126 | 1941 | 0.57 (NS) |
| ECDC (pt) | (3,0,1) $(0,1,1)_{52}$ | 12 | 2571 | 0.65 (NS) |
| Influenzanet (it) | (0,1,2) $(0,1,1)_{35}$ | 245 | 1097 | 0.82 (NS) |
| ECDC (it) | (2,0,4) $(1,1,1)_{35}$ | 47 | 1295 | 0.66 (NS) |

| SOURCE | AR1 | AR2 | AR3 | AR4 | MA1 | MA2 | MA3 | MA4 |
|---|---|---|---|---|---|---|---|---|
| Influenzanet (nl) | 1.8* | −0.84* | | | −1* | | | |
| ECDC (nl) | 1* | 0.016 | −0.19* | | | | | |
| Influenzanet (be) | 0.63* | | | | | | | |
| ECDC (be) | 1.3* | −0.38* | −0.19* | 0.072 | | | | |
| Influenzanet (pt) | 0.6* | 0.24* | −0.16* | | | | | |
| ECDC (pt) | 1.3* | −0.17 | −0.25* | | −0.43* | | | |
| Influenzanet (it) | | | | | −0.65* | −0.35* | | |
| ECDC (it) | 1.7* | −0.77* | | | −0.11 | −0.14 | −0.19* | −0.57* |

| SOURCE | SAR1 | SMA1 |
|---|---|---|
| Influenzanet (nl) | | −0.7* |
| ECDC (nl) | −0.014 | −0.75* |
| Influenzanet (be) | | −0.64* |
| ECDC (be) | | −1* |
| Influenzanet (pt) | | −1* |
| ECDC (pt) | | −0.78* |
| Influenzanet (it) | | −1* |
| ECDC (it) | −0.063 | −1* |

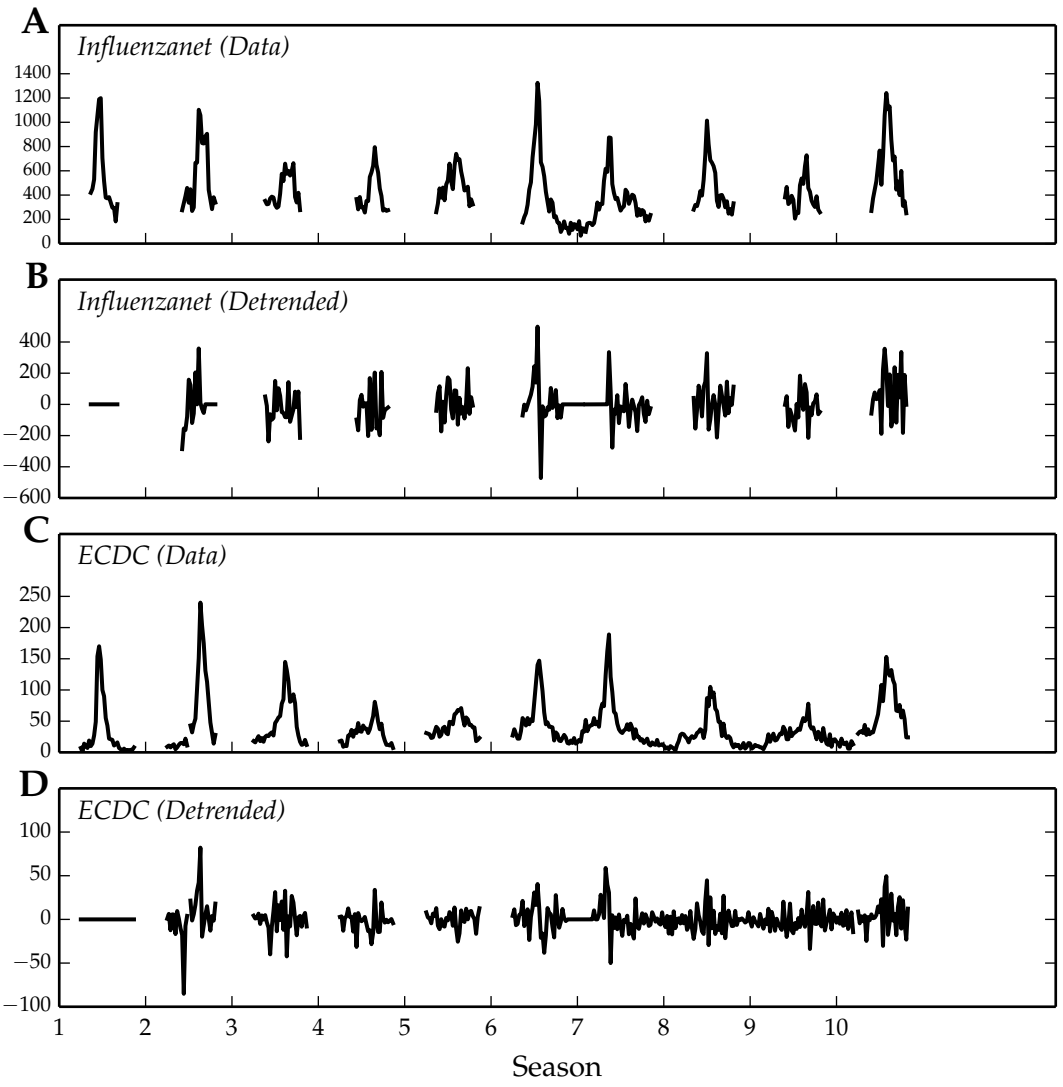



Figure B.1: ILI TIME SERIES IN THE NETHERLANDS (2003–2013) A) Influenzanet, B) Influenzanet (detrended), C) ECDC, and D) ECDC (detrended). The ILI incidence by Influenzanet is defined as the number of ILI onsets per 100,000 participant-weeks. The ILI incidence by ECDC is defined as the weekly number of patient visits with ILI per 100,000.

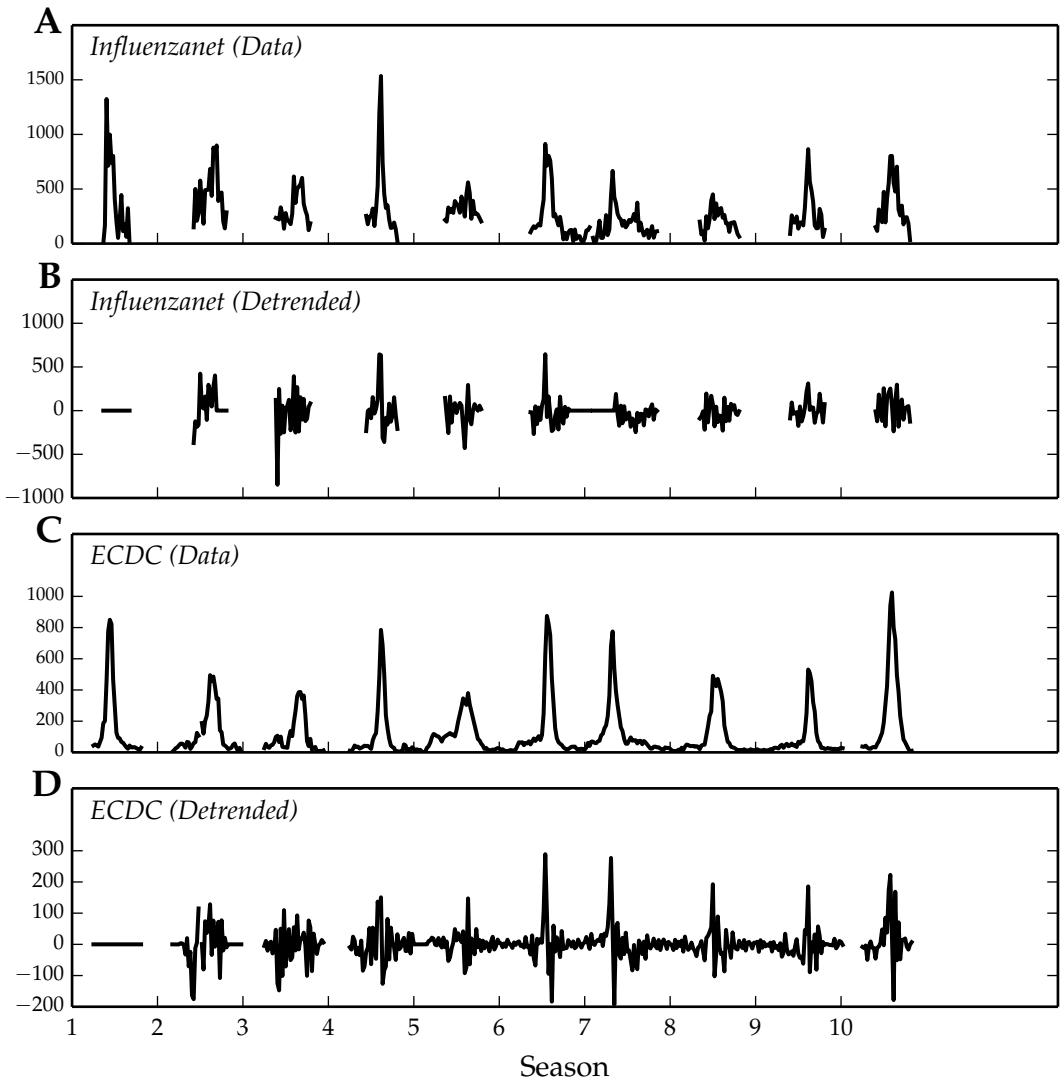



Figure B.2: ILI time series in Belgium (2003–2013) A) Influenzanet. The ILI incidence by Influenzanet is defined as the number of ILI onsets per 100,000 participant-weeks. B) Influenzanet (detrended) C) ECDC. The ILI incidence by ECDC is defined as the weekly number of patient visits with ILI per 100,000. D) ECDC (detrended).

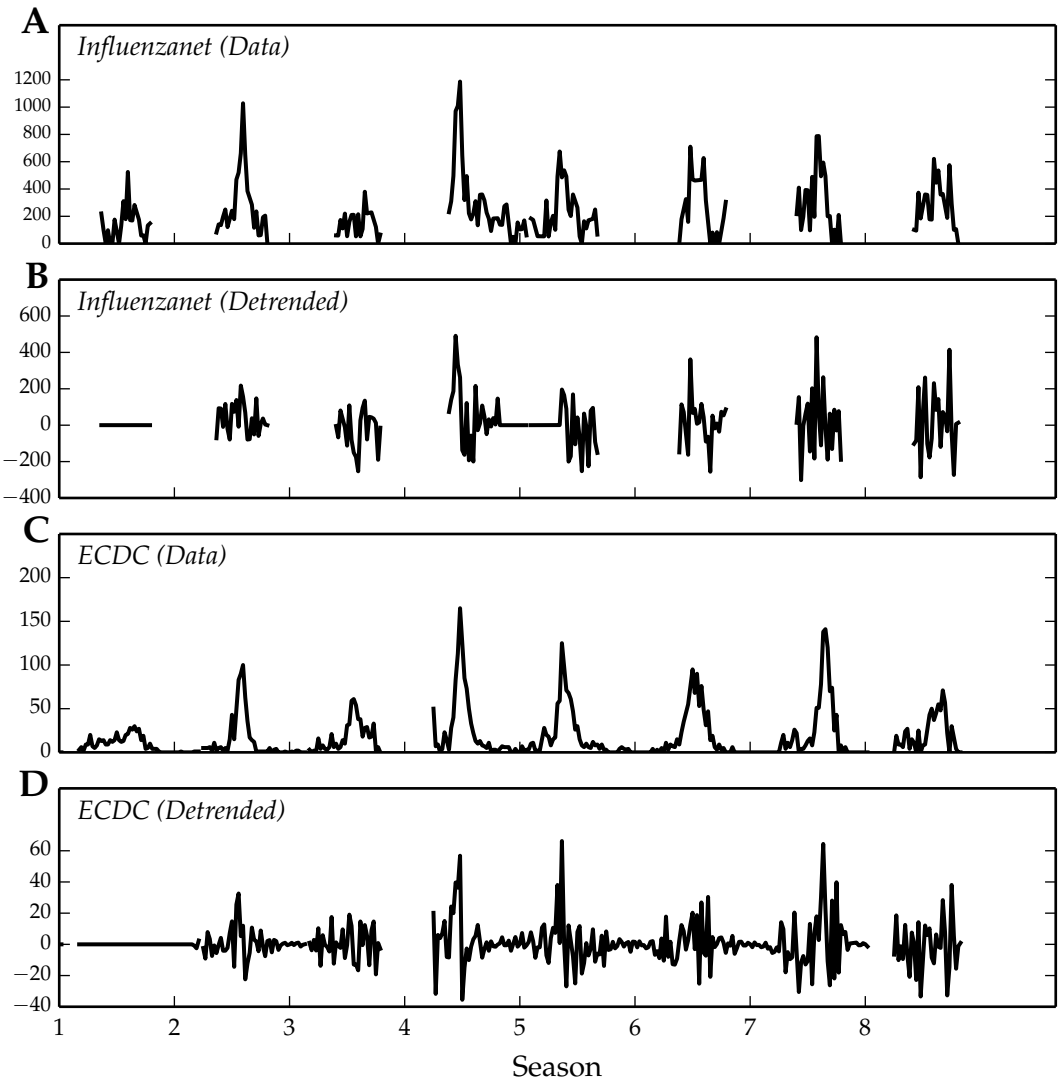



Figure B.3: ILI TIME SERIES IN PORTUGAL (2005–2013) A) Influenzanet, B) Influenzanet (detrended), C) ECDC, and D) ECDC (detrended) The ILI incidence by Influenzanet is defined as the number of ILI onsets per 100,000 participant-weeks. The ILI incidence by ECDC is defined as the weekly number of patient visits with ILI per 100,000.

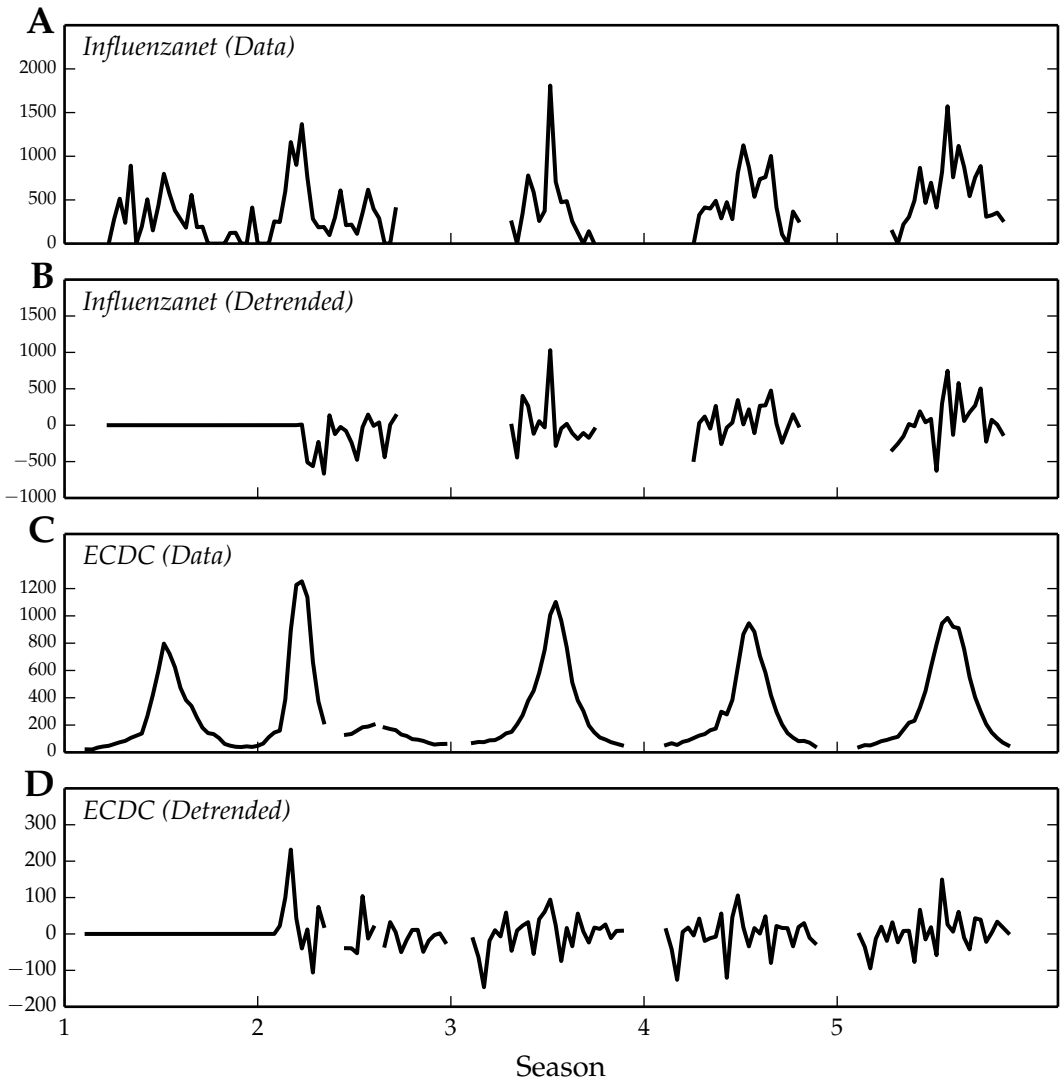



Figure B.4: ILI time series in Italy (2008–2013) A) Influenzanet, B) Influenzanet (detrended), C) ECDC, and D) ECDC (detrended) The ILI incidence by Influenzanet is defined as the number of ILI onsets per 100,000 participant-weeks. The ILI incidence by ECDC is defined as the weekly number of patient visits with ILI per 100,000.

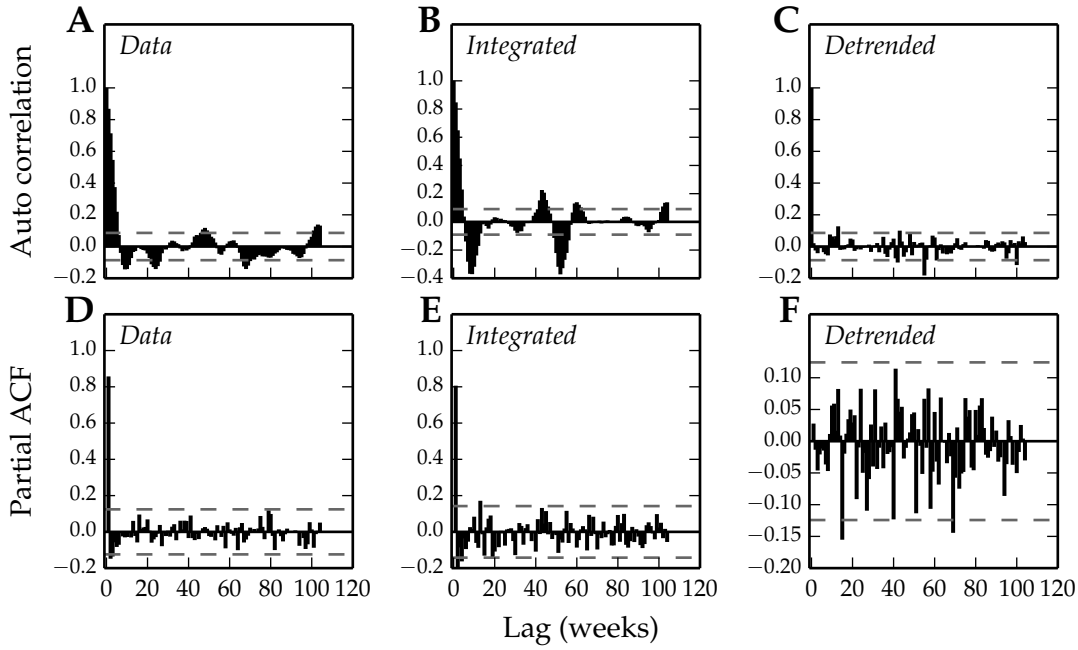



Figure B.5: AUTOCORRELATION FOR INFLUENZANET DATA IN THE NETHERLANDS . A–C) Autocorrelation, D–F) Partial autocorrelation. A,D) original series B,E) integrated series, and C,F) detrended series.

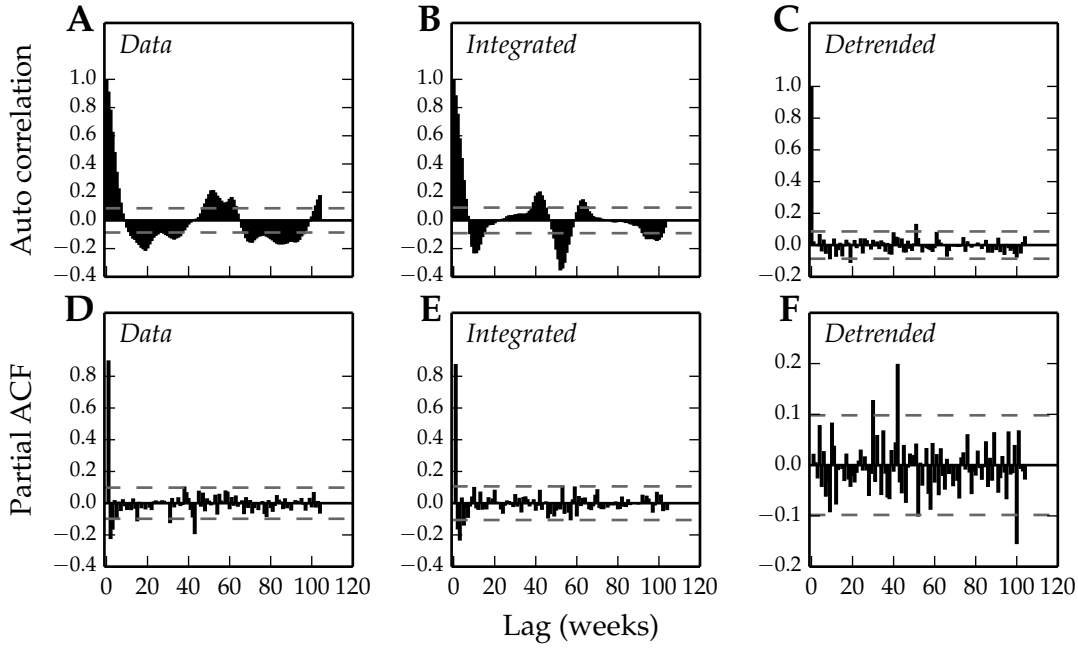



Figure B.6: AUTOCORRELATION FOR ECDC DATA IN THE NETHERLANDS

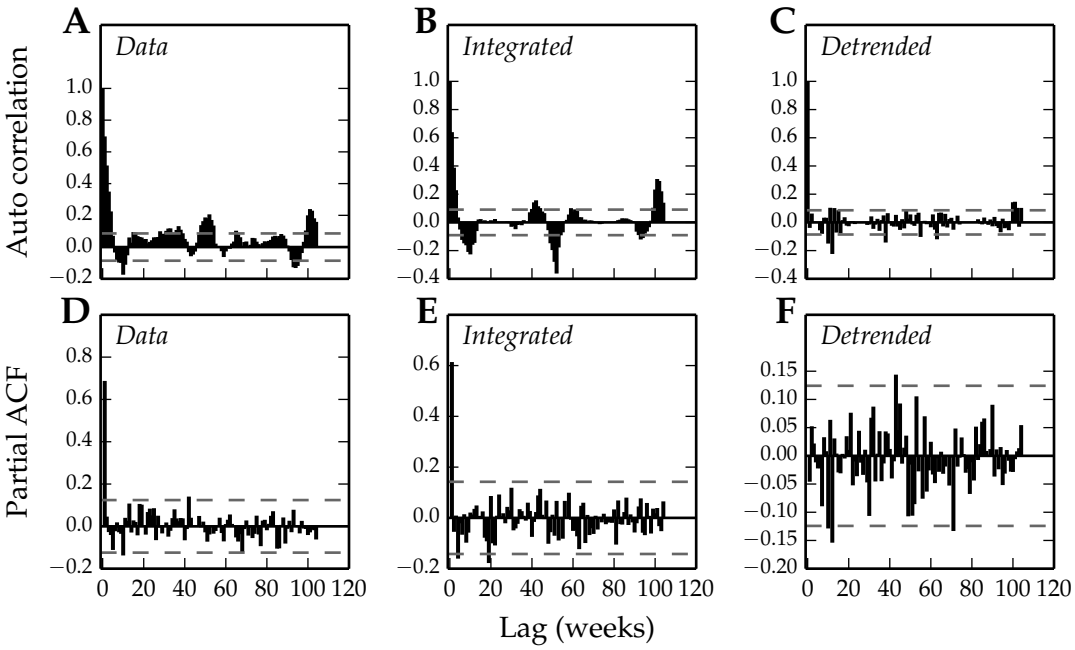



Figure B.7: Autocorrelation for Influenzanet data in Belgium

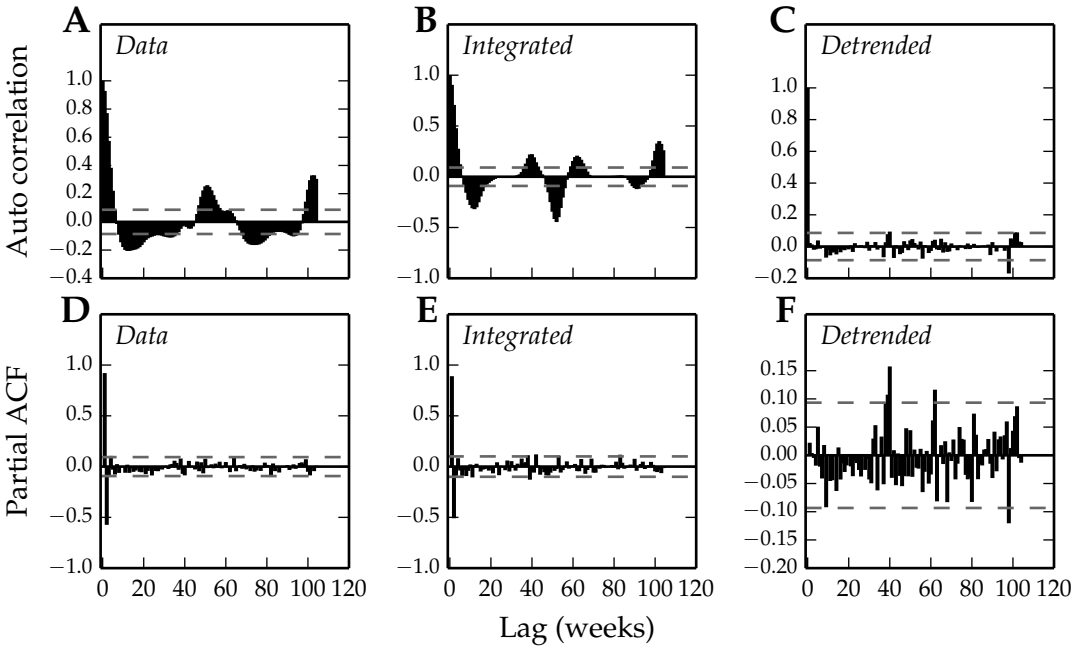



Figure B.8: Autocorrelation for ECDC data in Belgium

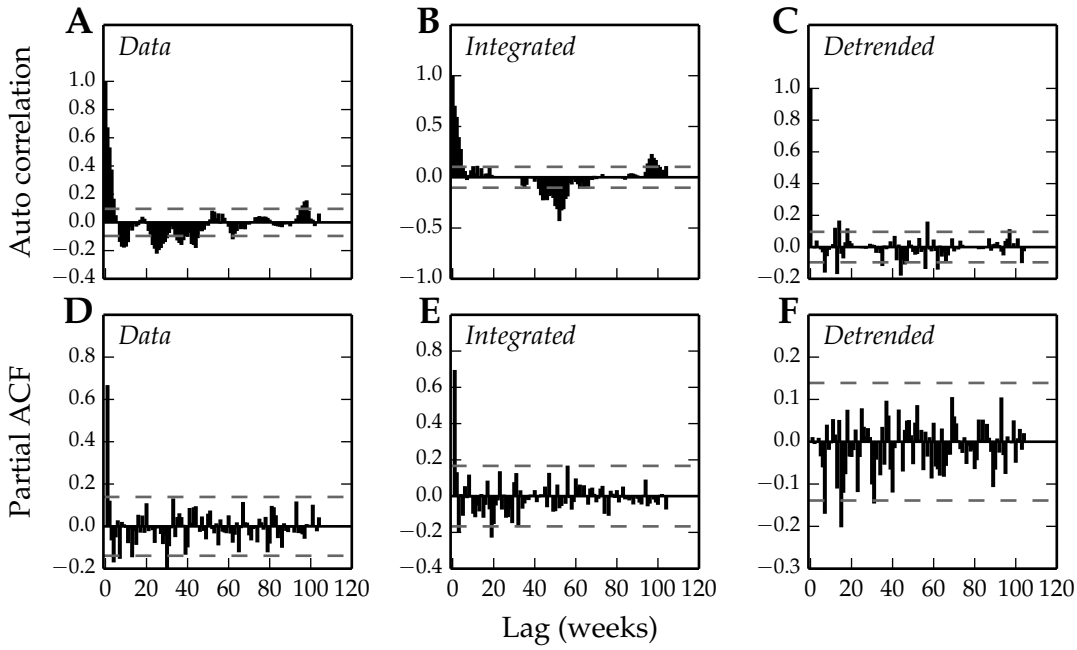



Figure B.9: AUTOCORRELATION FOR INFLUENZANET DATA IN PORTUGAL

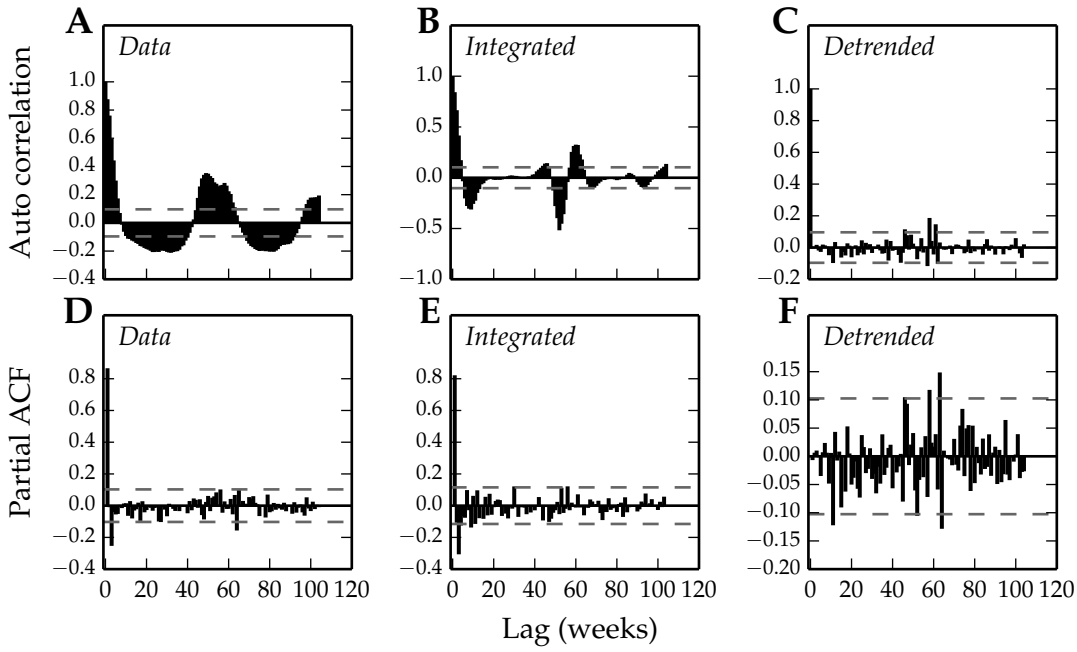



Figure B.10: AUTOCORRELATION FOR ECDC DATA IN PORTUGAL

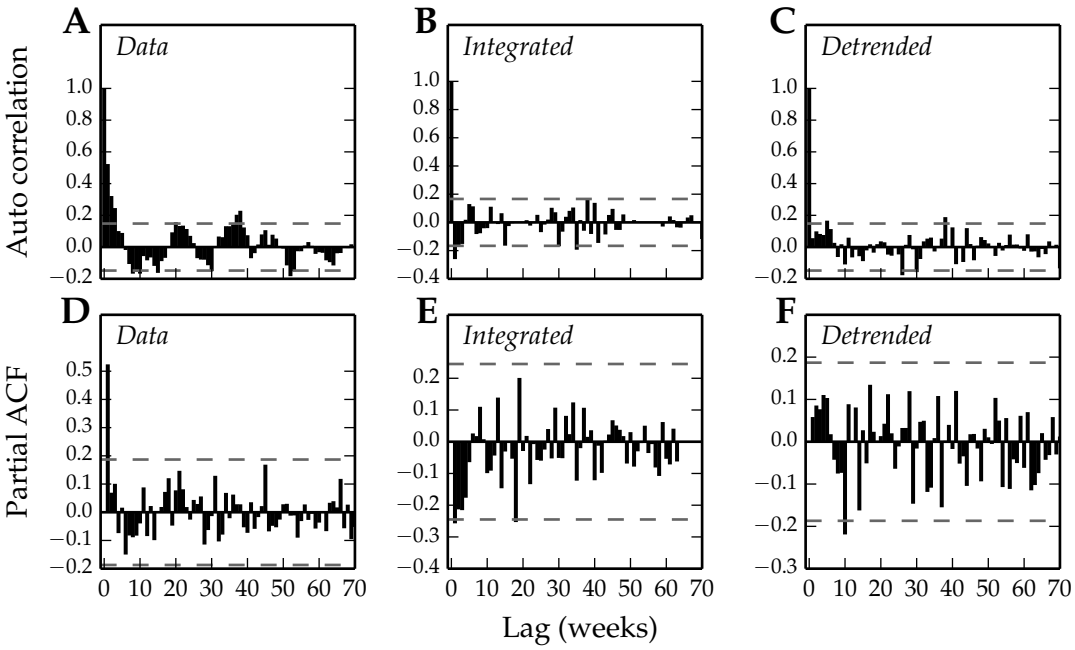



Figure B.11: Autocorrelation for Influenzanet data in Italy

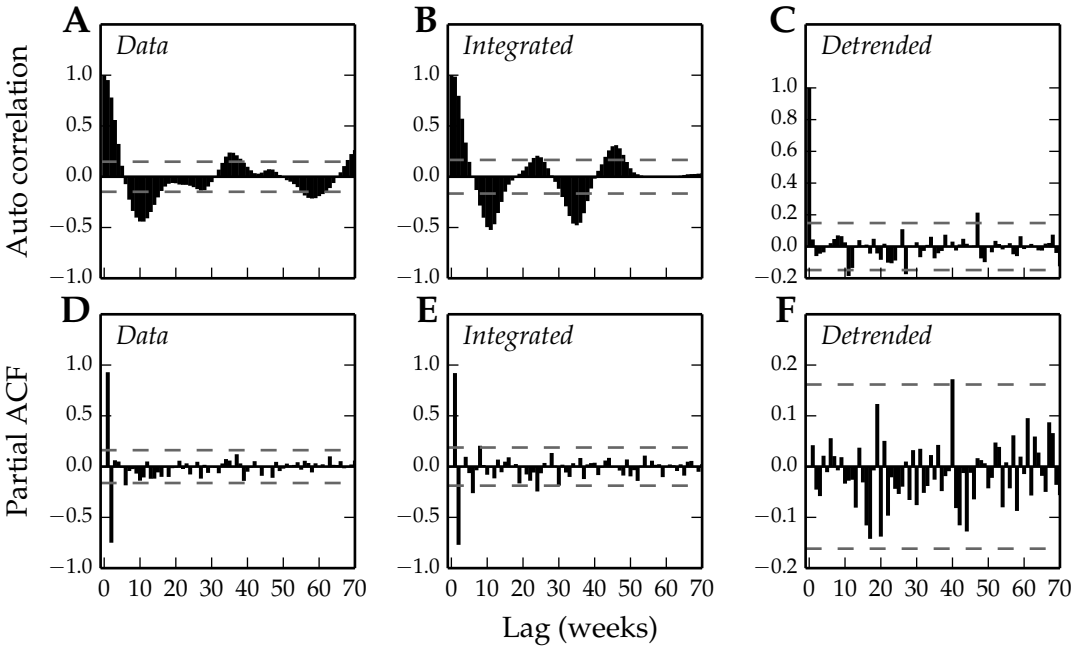



Figure B.12: Autocorrelation for ECDC data in Italy

## B.2 EXTENDED RISK ANALYSES

This section contains an extended risk analyses, as a supplement to Chapter 3. Table B.2 lists the extended results from the risk analyses. For each country the average percentage of participants who belongs to each covariate is listed. For each covariate the univariate risk ratio is determined. The variance inflation factors (VIF) for each covariate in the multivariate analyses is determined.

Figure B.13 shows the periods for which the risk analyses was performed. These periods are defined for each season and country as the weeks when the number of influenza-confirmed samples as reported by ECDC was at least 15% of the maximum for that season (moving average over 3 weeks). For each season, only ILI onsets during these periods are included, and only participants who were active during the full period are included.

All analyses were done using R version 3.1.0. The multivariate analyses was done using function *glm* from the R package *stats*. The VIF are determined using the function *vif* from the R package *rms*.

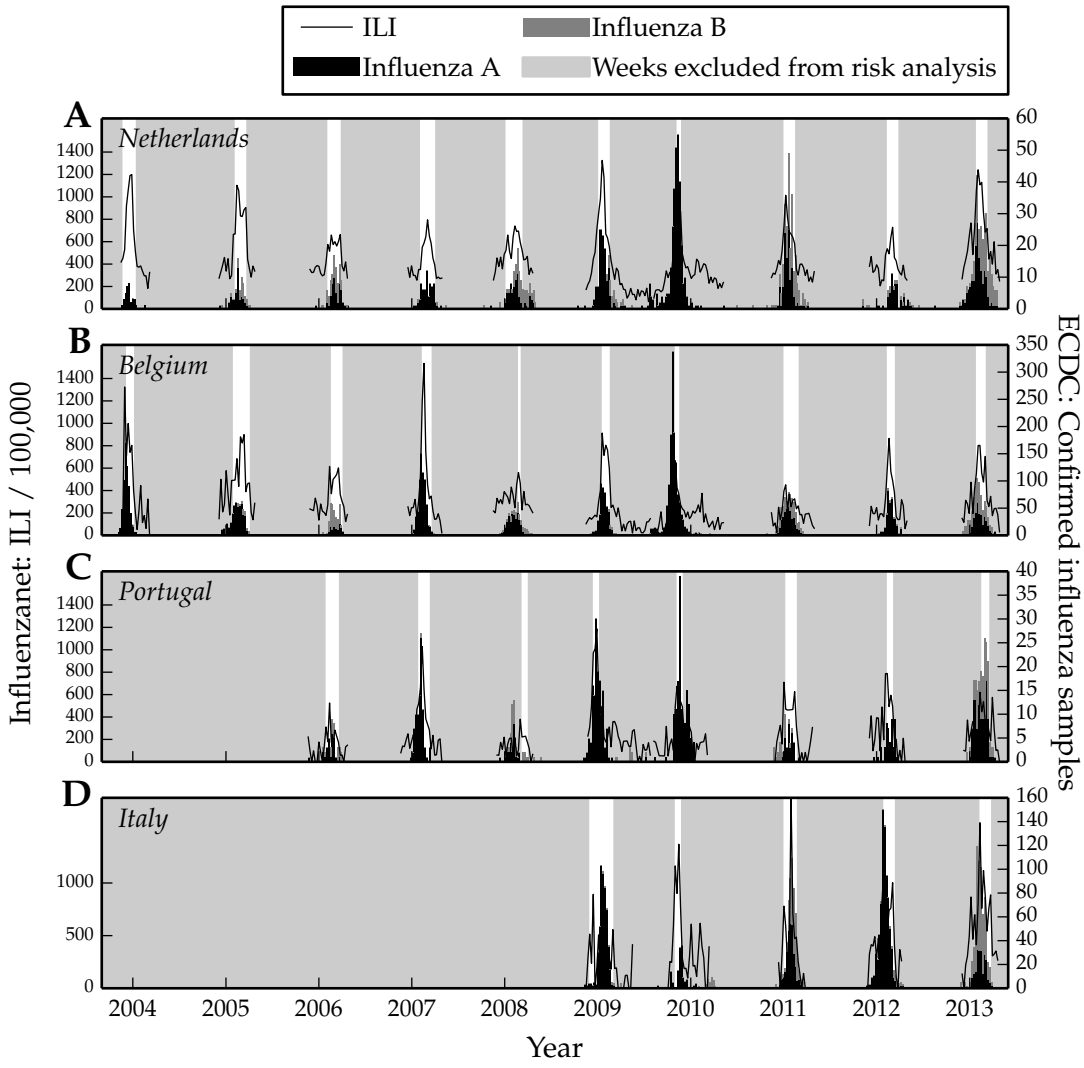



Figure B.13: Periods for the ILI risk analyses in A) Netherlands, B) Belgium, C) Portugal, and D) Italy (2003–2013). Periods which are grayed out are excluded.

Table B.2: RISK ANALYSES (2) . Extension to Table 3.2, Chapter 3, page 49, including univariate analyses, VIF and percentage per country (2003–2013).

| QUESTION | ANSWER | NL | BE | PT | IT | RR (UNIVARIATE) | VIF |
|---|---|---|---|---|---|---|---|
| Age | <18 | 4 | 2 | 5 | 7 | 1.64 (1.50 – 1.78) | 1.1 |
| | 18–49 | 49 | 47 | 67 | 52 | * | – |
| | 50–64 | 35 | 37 | 21 | 28 | 0.77 (0.73 – 0.80) | 1.3 |
| | 65+ | 13 | 13 | 6 | 13 | 0.39 (0.36 – 0.43) | 1.3 |
| Household | Alone | 18 | 13 | 11 | 10 | * | – |
| | Only with adults | 51 | 57 | 51 | 65 | 0.94 (0.88 – 1.00) | 2.5 |
| | With children | 32 | 30 | 39 | 26 | 1.54 (1.44 – 1.65) | 2.7 |
| Gender | Male | 44 | 56 | 49 | 62 | * | – |
| | Female | 56 | 44 | 51 | 38 | 1.39 (1.33 – 1.45) | 1.1 |
| Chronic | Asthma / Lung | 10 | 5 | 6 | 5 | 1.60 (1.51 – 1.71) | 1.2 |
| | Diabetes | 5 | 4 | 2 | 3 | 1.01 (0.92 – 1.12) | 1.1 |
| | Heart disease | 8 | 8 | 10 | 14 | 0.96 (0.85 – 1.09) | 1.2 |
| | Kidney disorder | 0.6 | 0.5 | 0.6 | 0.4 | 1.08 (0.70 – 1.68) | 1.0 |
| | Immunodeficiency | 3 | 3 | 1 | 0.5 | 1.19 (0.98 – 1.43) | 1.0 |
| Unknown risk group | | 4 | 5 | 4 | 2 | 1.05 (0.92 – 1.19) | 1.2 |
| Smoking | | 18 | 16 | 17 | 18 | 1.16 (1.10 – 1.22) | 1.0 |
| Pets | Dogs | 17 | 20 | 24 | 18 | 1.24 (1.18 – 1.31) | 1.1 |
| | Cats | 26 | 26 | 21 | 20 | 1.24 (1.18 – 1.30) | 1.1 |
| | Birds | 5 | 7 | 10 | 3 | 1.07 (0.98 – 1.17) | 1.0 |
| Sports | ≥1 hour/week | 54 | 46 | 46 | 44 | 1.01 (0.96 – 1.06) | 1.3 |
| Locomotion | Bicycle / Foot | 45 | 23 | 12 | 25 | 1.06 (1.01 – 1.12) | 1.3 |
| | Car | 47 | 66 | 70 | 59 | * | – |
| | Public Transport | 9 | 10 | 18 | 17 | 0.98 (0.90 – 1.06) | 1.1 |
| Vaccination | 2003–2004 | 25 | 32 | – | – | 0.97 (0.71 – 1.33) | 1.3 |
| | 2004–2005 | 24 | 32 | – | – | 0.97 (0.83 – 1.13) | 1.3 |
| | 2005–2006 | 29 | 41 | 18 | – | 0.85 (0.73 – 0.99) | 1.4 |
| | 2006–2007 | 29 | 38 | 17 | – | 0.90 (0.78 – 1.04) | 1.4 |
| | 2007–2008 | 32 | 41 | 18 | – | 0.75 (0.65 – 0.87) | 1.4 |
| | 2008–2009 | 37 | 43 | 17 | 22 | 0.74 (0.66 – 0.82) | 1.5 |
| | 2009–2010 | 43 | 31 | – | 26 | 0.86 (0.73 – 1.02) | 1.6 |
| | 2010–2011 | 42 | 48 | 24 | 19 | 0.63 (0.54 – 0.72) | 1.5 |
| | 2011–2012 | 40 | 44 | 24 | 19 | 0.93 (0.78 – 1.10) | 1.6 |
| | 2012–2013 | 38 | 46 | 24 | 17 | 0.74 (0.66 – 0.83) | 1.6 |

## B.3 MODEL FITS BASED ON OTHER WEATHER VARIABLES

This section contains model fits of other weather variables, as a supplement to Chapter 4. Figure B.14 shows the correlation between the estimated ILI factor for each season, and the timing of the peak of the ILI epidemic. In Figure B.15 a transmission model is fitted to ILI data using the actual measured absolute humidity for each week, in Figure B.16 using the typical measured temperature for each week, and in Figure B.17 using the actual measured temperature for each week.

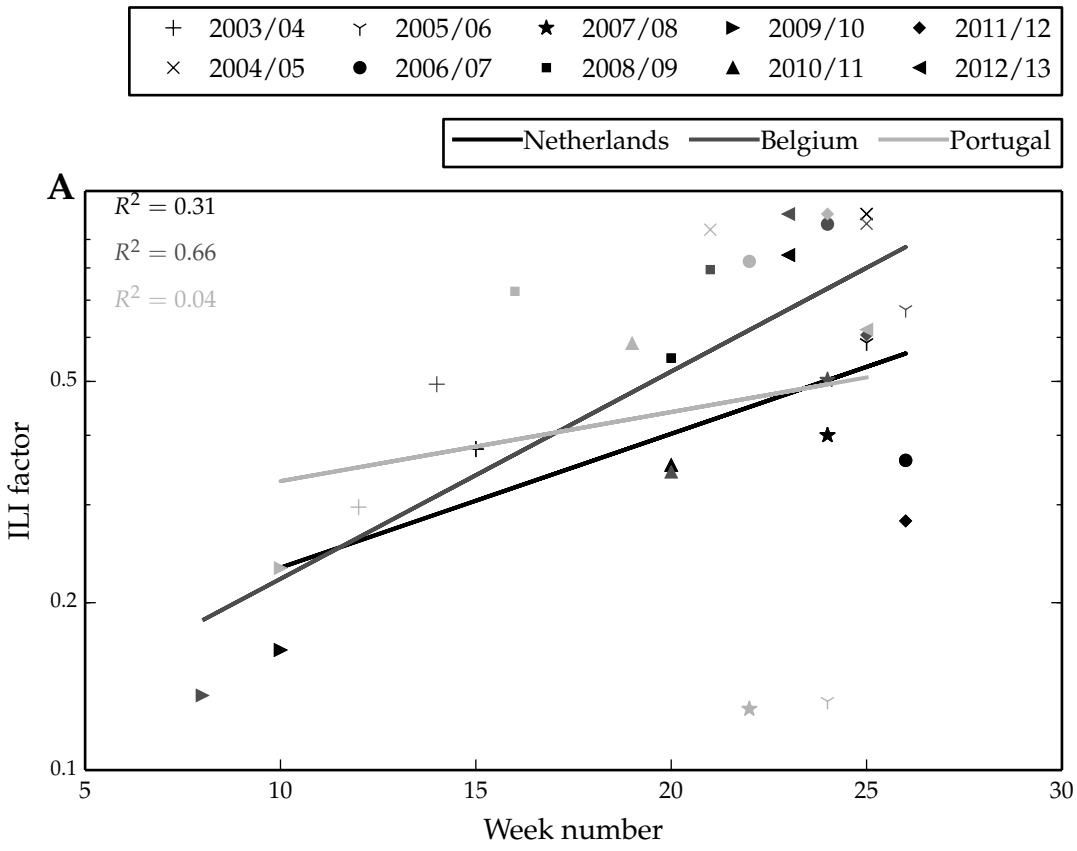



Figure B.14: ILI FACTOR AND TIMING OF ILI EPIDEMIC based on the peak date of the ILI epidemic.

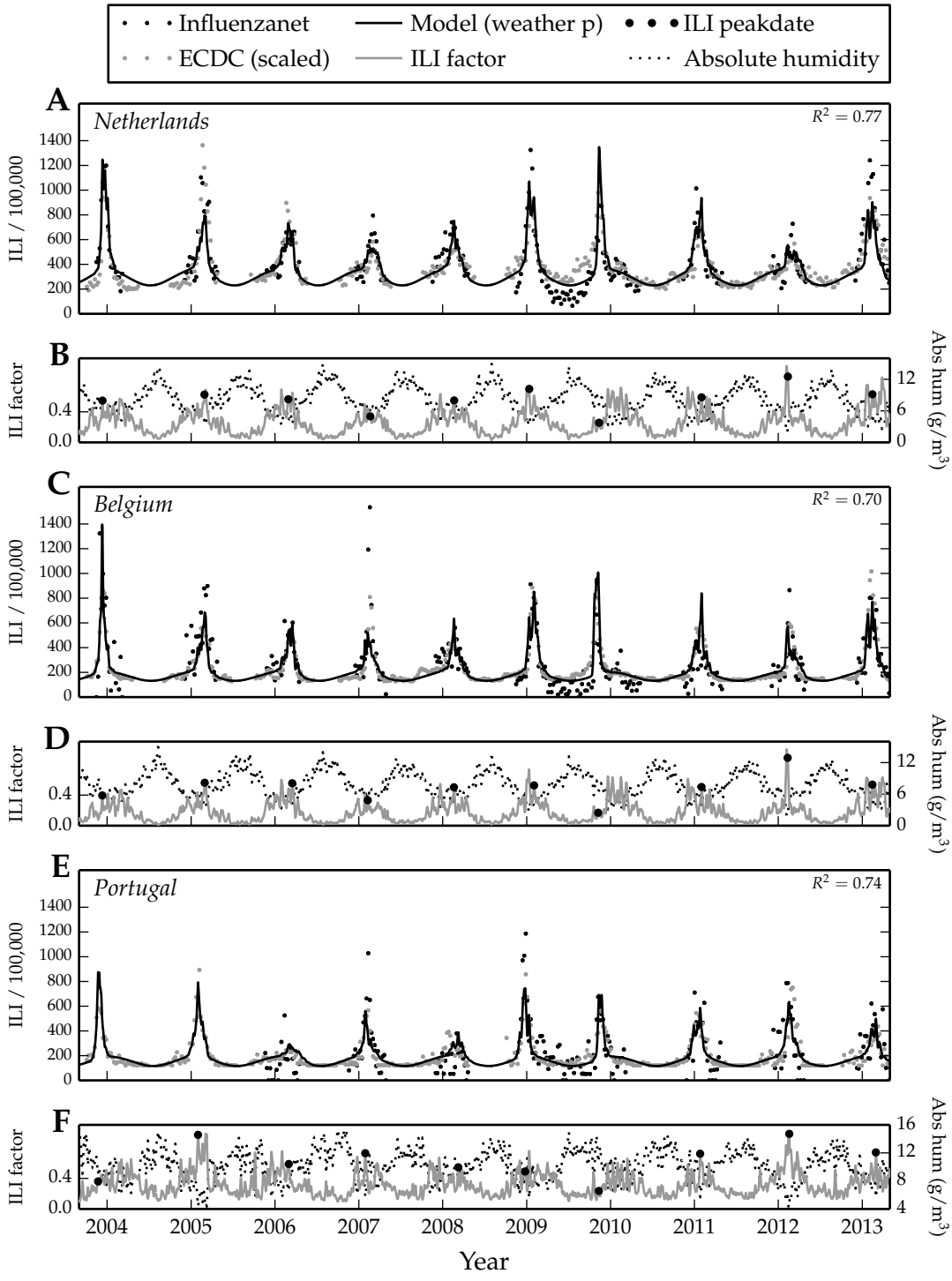



Figure B.15: MODEL BASED ON THE ACTUAL ABSOLUTE HUMIDITY

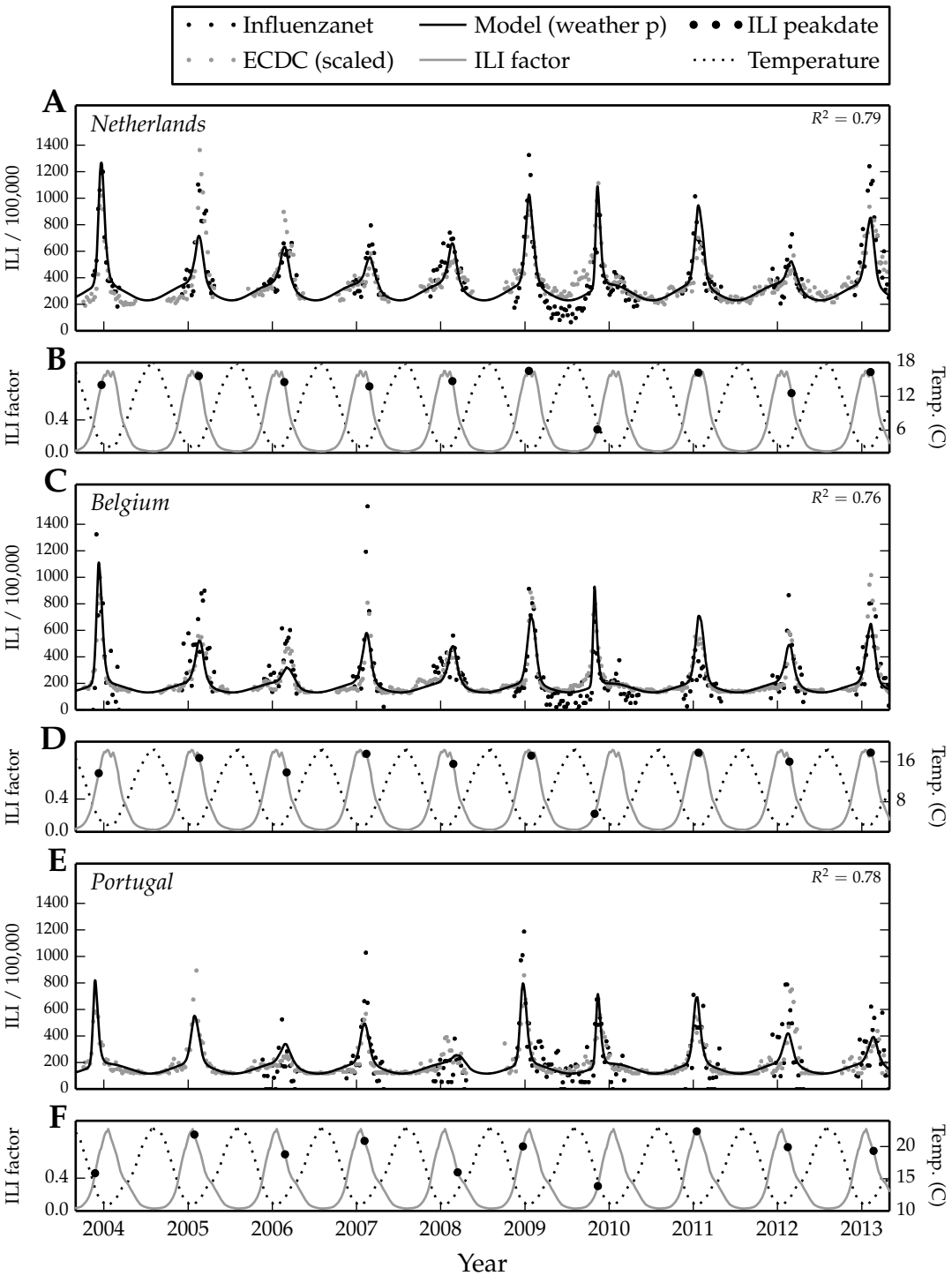



Figure B.16: Model based on the typical temperature

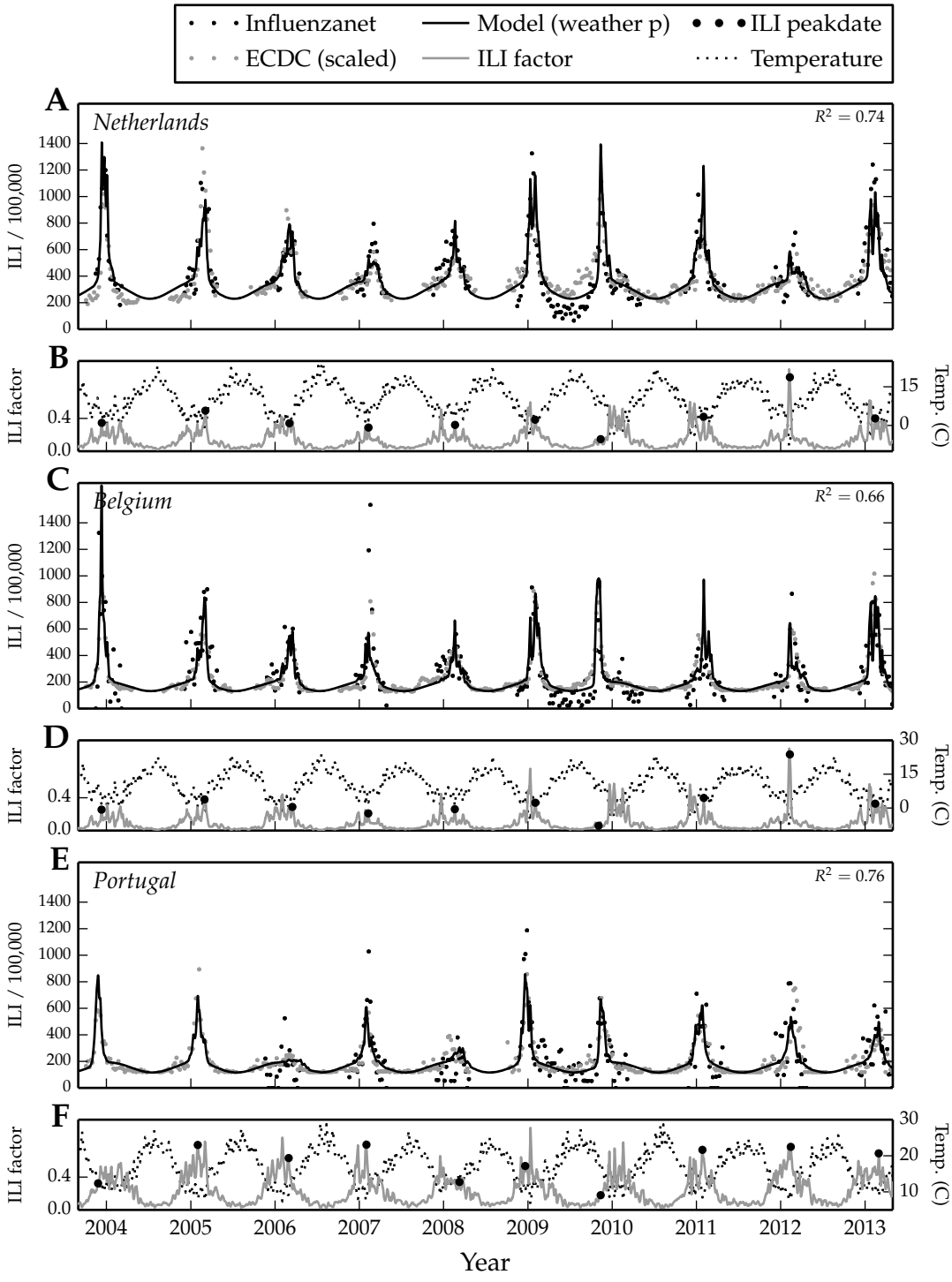
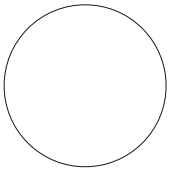
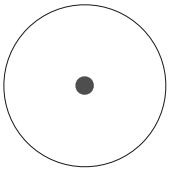



Figure B.17: MODEL BASED ON THE ACTUAL TEMPERATURE

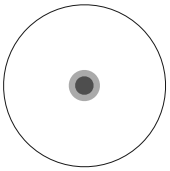
**1.** Imagine a circle that contains all of human knowledge.



**2.** By the time you finish elementary school, you know a little.



**3.** By the time you finish high school, you know a bit more.

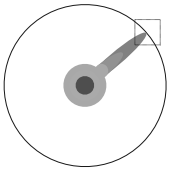


**4.** With a bachelor's degree, you gain a specialty.

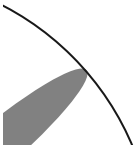
**5.** A master's degree deepens that specialty.

**6.** Reading research papers takes you to the edge of human knowledge.

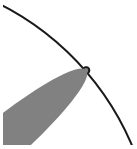
**7.** Once you're at the boundary, you focus.



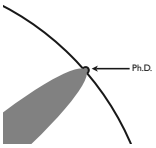
**8.** You push at the boundary for a few years.



**9.** Until one day, the boundary gives way.

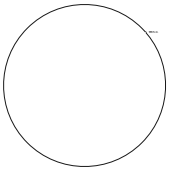


**10.** And, that dent you've made is called a Ph.D.




**11.** Of course, the world looks different to you now.

**12.** So, don't forget the bigger picture.



Keep pushing.*

* http://matt.might.net/articles/phd-school-in-pictures/

ITQB-UNL | Av. da República, 2780-157 Oeiras, Portugal
Tel (+351) 214 469 100
Fax (+351) 214 411 277

www.itqb.unl.pt