Improving incidence estimation in practice-based sentinel surveillance networks using spatial variation in general practitioner density
Cécile Souty, Pierre-Yves Boëlle

To cite this version:


HAL Id: hal-01400246
https://hal.sorbonne-universite.fr/hal-01400246
Submitted on 21 Nov 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License
Improving incidence estimation in practice-based sentinel surveillance networks using spatial variation in general practitioner density

Cécile Souty and Pierre-Yves Boëlle

Abstract

Background: In surveillance networks based on voluntary participation of health-care professionals, there is little choice regarding the selection of participants’ characteristics. External information about participants, for example local physician density, can help reduce bias in incidence estimates reported by the surveillance network.

Methods: There is an inverse association between the number of reported influenza-like illness (ILI) cases and local general practitioners (GP) density. We formulated and compared estimates of ILI incidence using this relationship. To compare estimates, we simulated epidemics using a spatially explicit disease model and their observation by surveillance networks with different characteristics: random, maximum coverage, largest cities, etc.

Results: In the French practice-based surveillance network – the “Sentinelle” network – GPs reported 3.6% (95% CI [3;4]) less ILI cases as local GP density increased by 1 GP per 10,000 inhabitants. Incidence estimates varied markedly depending on scenarios for participant selection in surveillance. Yet accounting for change in GP density for participants allowed reducing bias. Applied on data from the Sentinelle network, changes in overall incidence ranged between 1.6 and 9.9%.

Conclusions: Local GP density is a simple measure that provides a way to reduce bias in estimating disease incidence in general practice. It can contribute to improving disease monitoring when it is not possible to choose the characteristics of participants.

Keywords: Surveillance, Incidence estimation, General practitioners, Influenza-like illness, Surveillance networks, Epidemiological surveillance, GP density

Background

Surveillance networks for common acute conditions often rely on a sample of healthcare professionals who report disease cases seen in their practice [1, 2]. Various methods aim at optimizing the choice of these data providers and improving incidence estimates. For example, the locations of participating health professionals can be chosen to maximize population coverage [3, 4]. Another possibility is to make the monitored population representative in terms of age, sex or socio-economic status [5, 6]. It may also be possible to select participants so that estimates from the surveillance network are as close as possible to a reference epidemiologic signal, for example influenza-like hospitalizations when monitoring flu in the community [7].

Yet these a priori methods for selection implicitly assume that all health professionals or organizations would be willing to participate in surveillance upon proposal and not subject to turnover once recruited. These assumptions may indeed be reasonable in hospital or laboratory based surveillance networks, but not in surveillance networks based on general practitioners (GPs) in primary
care. Indeed, in our experience, only a few percent of practicing GPs agree to participate in surveillance networks [8] and their participation is most of the times for a couple of years only [9].

More subtly, it has not been considered that physician density (physician per population ratio) changes the number and motives for medical consultations [10–12]: typically, in areas of low GP density, GPs see more cases of acute diseases because the population served is larger and a higher percentage of consultations is for treating acute diseases. On the contrary, in areas of high GP density, the number of acute disease cases seen by a GP is smaller, not so much as a result of less activity, but because fewer consultations are for treating acute diseases [13]. Accounting for local variation in physician density will be important when computing disease incidence indicators in health systems where patients are not registered to a unique GP practice, as is the case in most European countries [1, 14, 15].

To examine further the effects of GPs selection in primary care surveillance networks, we use a simulation approach and compare incidence estimates for various surveillance network structures similar to the French GP-based Sentinelles network [16]. Our goal is to explore how the spatial sampling of data providers affects the estimates of incidence and to propose improved estimators. We finally report an empirical comparison of incidence estimation based on influenza-like illness (ILI) in France.

Methods

The French Sentinelles network

The French general practitioners Sentinelles network [17] is a real-time epidemiologic surveillance system based on approximately 500 GPs, corresponding to 1% of all French GPs located all over the country [18]. Sentinel general practitioners (SGPs) are recruited and participate in surveillance on a voluntary basis. Differences between participating and non-participating GPs have been reported elsewhere [8, 19]. They report cases observed in their practice population, such as ILI cases, using a web interface or a dedicated software [20]. Incidences are calculated in real-time from their reports [19] allowing to detect outbreaks [21].

Here, we included the 301 SGPs participating to the French Sentinelles network during the 2012/13 influenza season. In addition to data provided by SGPs, we also obtained data on consultation volumes for all practicing French GPs and separately for SGPs from the national health insurance system [19, 22].

Incidence estimation from cases reported by participating GPs

Incidence in the general population is computed from the number of patients with a specific combination of symptoms reported by the sample of participating SGPs [19, 23]. We present below estimators to compute incidence based on post-stratification. For most of the cases, incidence is first estimated by region (NUTS2 – Nomenclature of Territorial Units for Statistics level 2 [24]) and summed to estimate national incidence. Incidences estimates per population are computed as a ratio of estimated incidences divided by the area population (census data – National Institute of Statistics and Economics Studies [25]). A supplementary file contains all computational details (see Additional file 1).

Horvitz-Thompson estimator: proportion of SGPs among GPs

A typical approach to reduce bias in non-representative samples is to use a Horvitz-Thompson estimator [26], where sampling weights correspond to the inverse of the inclusion probability. In the French Sentinelles network, these weights are computed as the percentage of SGPs participating in surveillance by region [19]. Indeed, the inclusion probability for SGP $k$ is $\pi_k = nSGP_R/nGP_R$, where $nSGP_R$ is the number of SGPs and $nGP_R$ the total number of GPs in the region $R$ where the SGP $k$ practices. Thus, the national incidence estimator in period $t$ is:

$$I_n(t) = \sum_k \pi_k \cdot cases(k, t)$$

where $k$ runs over participating SGPs and $cases(k, t)$ is the number of cases reported by the SGP $k$ during period $t$.

Incidence estimator taking into account local GP density

All calculations are detailed in Additional file 1.

Direct estimate using local GP density

Assume that the number of cases seen by a GP decreases with increasing local GP density $m$ (number of GPs per population in the district or LAU1 - Local Administrative Units level 1 [24]) as $E(cases) = \lambda/m$, where $\lambda$ is the incidence per population. The national incidence estimator is:

$$I_D(t) = \sum_k \frac{m_k}{m_{D,k}} \cdot nSGP_R \cdot cases(k, t)$$

where $m_{D,k} = nGP_D(k)/pop_D(k)$ is the GP density at the departmental level (NUTS3 [24]) in SGP $k$’s department of practice.

Calibrated estimate using local GP density

Calibration is a generic method to improve estimation when auxiliary information associated with measurements is available [27]. We use here the inverse of local GP density (LAU1) $1/m$ as auxiliary information for
reported cases. The well-known ‘ratio-estimator’ [27] obtained with this approach has expression:

\[ \hat{I}_{Cm}(t) = \sum_k \frac{h_{R,k}}{m_{R,k}} \cdot \pi_{k} \cdot \text{cases}(k, t) \]

where \( m_{R,k} \) is the GP density in the region of practice of SGP \( k = n \times GP_{p}/\text{pop}_{R} \) and \( h_{R,k} \) is the harmonic mean of GP densities among SGPs in the region \( R \) where SGP \( k \) practices, i.e. \( h_{R,k} = n \times GP_{p}/\sum_{i} (1/m_{i}) \) where \( i \) runs over participating SGPs in \( R \).

**Direct and calibrated estimates using local GP density and consultation volume**

In a previous work [19], we found a positive association between the number of cases reported by SGPs and the number of consultations. Assuming joint proportionality of reported cases to GP density and consultation volume, a direct estimator of incidence is:

\[ \hat{I}_{Dmc}(t) = \sum_k \frac{m_{k}}{\rho_{R,k}(t)} \cdot \frac{1}{\pi_{R,k}(t)} \cdot \text{cases}(j, t) \]

where \( \rho_{R,k}(t) = c_{SGP_{p}}(t)/c_{GP_{p}}(t) \) is the percentage of consultations by SGPs \( c_{SGP_{p}}(t) \) among all consultations \( c_{GP_{p}}(t) \) in region \( R \) where SGP \( k \) practices, during period \( t \).

Using calibration proposed by Deville and Särndal [27], the incidence estimator calibrated on both local GP density and consultations is defined as follows:

\[ \hat{I}_{Cmc}(t) = \sum_k \left( \frac{\pi_{k} \cdot \text{cases}(k, t)}{T_{k} - t_{x}} \cdot \hat{I}_{Cm}(t) \right) \]

with \( T = \sum_k \pi_{k} \cdot x_{k} \cdot \pi_{k} \cdot x_{k} \) where the auxiliary information \( x_{k} = (1/m_{k}, \rho_{R,k}) \), \( t_{x} \) is the population total of \( x \), and \( t_{x} \) denote the Horvitz-Thompson estimator for the \( x \)-vector.

**Investigating the relationship between GP density and reported cases by SGPs**

We analysed the number of cases reported to the Sentinelle network during an epidemic by a SGP using Poisson regression with dependant variable the local GP density (district level – LAU1). We used the data over the last five seasons (2010/11 to 2014/15) for each SGP.

**Influenza epidemics simulations**

Influenza epidemics were simulated using a spatially explicit age-structured model including population and commuting data in France [16]. Simulations were performed at the district level (LAU1 - 3708 districts in France) and yielded weekly incidence data. Influenza cases in the 98 districts with no GPs were allocated to the neighbouring districts.

In each district, the average number of cases seen by a GP in 1 week was computed as the ratio of number of new cases in the district divided by the number of GPs in that district. We assumed that the actual number of cases reported by a GP had a Poisson distribution about this average.

**Evaluation of sample weights for incidence estimation**

**Simulated GPs networks**

To investigate the impact of geographical location of GPs participating to surveillance, we simulated networks by choosing data providers from the 60,000 practicing French GPs. Network sizes was limited to 300 to be similar to the existing French Sentinelle network.

We first investigated 1000 “random networks” including 300 GPs chosen at random from all GPs. Next, we selected a network including one GP taken in each of the districts of the administrative centres at the NUTS3 level (\( n = 96 \) departments in France). We also selected low GP density network, with GPs from the three districts with the lowest GP density in each department, and high GP density network where GPs were taken from the three districts with the highest GP density. Finally, we used the maximum coverage algorithm proposed by Polgreen et al. [3] to maximize the population covered by participating GPs using 300 GPs and a distance of 14 kilometres (8.7 miles – the average diameter of a district).

**Comparing estimated incidence to real incidence**

To compare incidence estimates based on the various surveillance networks, we computed the weekly average relative difference in percent and the root mean squared error (RMSE) between estimated and simulated incidence. Estimates were computed using the Horvitz-Thompson estimator \( \hat{I}_{Cm} \) and the estimators taking into account local GP density \( \hat{I}_{Dmc} \) and \( \hat{I}_{Cmc} \).

To account for variability in number of reported cases by participating GPs (Poisson distribution), we replicated estimations 100 times and reported the average.

**Empirical assessment: subsampling SGPs in the Sentinelle network**

To investigate empirically the characteristics of the estimators on real data, we used a subsampling approach. We estimated incidence using 75% of all participating SGPs only, and varied randomly the 75% selected. We computed the three estimators \( \hat{I}_{Cm}, \hat{I}_{Dmc}, \) and \( \hat{I}_{Cmc} \) as defined above (Horvitz-Thompson and both defined with local GP density) during the 2012/13 influenza epidemic (2012 week 51th to 2013 week 11th). Using results from subsampling, we estimated the standard deviation of estimated incidence and the coefficient of variation.

Finally, we applied all estimators to estimate ILI incidence over the 5 years period from 2009 week 32th to
2014 week 31th. Comparisons between estimators were done during epidemic periods as defined by the Sentinelles network [17, 21].

Results
Spatial distribution of GPs according to population in France
Overall, the average GP density in France was 96 GPs per 100,000 inhabitants, ranging from 12 to 526 GP per 100,000 inhabitants (except in 98 districts (2.6%) without a GP) (Fig. 1). The 301 SGPs of the Sentinelles network came from 260 districts all over the country. In the districts of practice of the SGPs, the average GP density was slightly lower than in overall France (94 GPs per 100,000 inhabitants).

Number of cases per SGP and GP density using data from the French Sentinelles network
SGPs practicing in places with high GP density, reported fewer ILI cases over the duration of an epidemic (Fig. 2). On average, the number of cases was reduced by 3.6% (95% CI [3; 4]) as GP density increased by 10 GP per 100,000 inhabitants \((p < 2.10^{-16})\). In other words, during a typical flu epidemic, a GP practicing in a district with 75 GP/100,000 inhabitants (first quartile of GP density in SGPs) reported 42 cases while a GP practicing in a district with 125 GP/100,000 (third quartile of GP density in SGPs) reported only 34.

Incidence estimates using simulated GPs networks
The geographical location of GPs included in the simulated surveillance networks did not show specific spatial patterns (Fig. 3) except for fewer GPs included in the administrative centres network (96 instead of 300). Visually, the actual Sentinelles network looked patchier than the maximum coverage, high/low GP density and administrative centres networks but similar to random networks.

For the same simulated epidemic, estimated incidence was highly dependent on data provider selection using the Horvitz-Thompson estimator \(\hat{I}_n\) (Fig. 4). In random networks, incidence was estimated almost without bias (+4%: RMSE = 59 cases per 100,000 inhabitants) but with substantial variation from one network to the next. With the administrative centres network, incidence was underestimated by an average of 24% (RMSE = 306). The largest underestimates were obtained for the high GP density network, where incidence was one third less than the actual values (−37%; RMSE = 467). Conversely, the
low GP density network led to the largest overestimates, with estimated incidence more than one and half the actual incidence (+164%; RMSE = 2370). The maximal coverage network, while covering 79% of the French population with 300 GPs, overestimated incidence by one third (+33%; RMSE = 425). Finally, in the network corresponding to the actual Sentinelles network, incidences estimates were on average 10% higher than expected (RMSE = 133).

Introducing local GP density in incidence estimators
Taking into account local GP density allowed substantial bias reduction in all simulated networks: estimates were always closer to the actual incidence by comparison with the Horvitz-Thompson estimator. In random networks, the ratio estimator \( \hat{I}_{cm} \) and the direct estimator \( \hat{I}_{dm} \) performed similarly (RMSE = 26). For other network choices, the situations were contrasted: the ratio estimator \( \hat{I}_{cm} \) performed better for low GP density and maximum coverage networks and the direct estimator \( \hat{I}_{dm} \) for administrative centres and high GP density networks (Fig. 5).

Subsampling SGPs from the French Sentinelles network
We next compared estimators \( \hat{I}_n \), \( \hat{I}_{dm} \) and \( \hat{I}_{cm} \) on the Sentinelles data for 2012/13 influenza epidemic. The average standard deviation decreased from 24 cases per 100,000 inhabitants for \( \hat{I}_n \) to 19 for \( \hat{I}_{dm} \) and \( \hat{I}_{cm} \). Likewise, the coefficients of variation were 5.2% when using \( \hat{I}_{dm} \) and 5.3% with \( \hat{I}_{cm} \), less than with \( \hat{I}_n \) (5.9%).

In general, modifications of the original \( \hat{I}_n \) led to lower incidences estimates, as expected given the average GP density was slightly lower for the Sentinelles network than for overall France. Using the Horvitz-Thompson estimator as a reference with whole SGPs’ data, the estimated incidence was on average 1.6% less for the direct estimator \( \hat{I}_{dm} \) and 8.4% less for the ratio estimator \( \hat{I}_{cm} \). Finally, calibration on both local GP density and consultation volume led to similar performance compared with calibration on local GP density alone, with a reduction by 9.9% using \( \hat{I}_{cmc} \) and by 4.7% using \( \hat{I}_{dmc} \).

Discussion
Reducing bias in disease incidence estimates from GP surveillance networks is a challenge when participants are volunteers rather than carefully chosen by administrators. Here, we showed that bias due to variations in local GP density around participants could be reduced using weighted estimators, and supported this conclusion with empirical evidence. This approach could be used to introduce more flexibility for developing algorithms to identify the best placement of data providers [3–7].

We found that local GP density changed the reports of GPs to surveillance systems, with less ILI cases reported as local GP density increased. This was also true with acute diarrhea and chickenpox, two other diseases monitored by the Sentinelles network. For example, the number of acute diarrhea cases reported by SGPs during the epidemic was reduced by 3.4% (95% CI [2.9;3.9]) (\( p < 0.0001 \)) and the number of yearly chickenpox cases was reduced by 6.4% (95% CI [5.9;6.8]) (\( p < 0.0001 \)) when the GP density increase by 10 GP per 100,000 inhabitants. There is indeed a known effect of GP density on their activity which reflects competition between physicians [11]. While the number of consultations may be slightly reduced with increasing GP density, it is mostly the nature of consultations that changes [13]. More consultations are for prevention, or follow-up of chronic diseases in high GP density areas, leading to fewer acute disease cases per GP and possible bias in incidence estimation.

Surveillance networks in general practice must satisfy a number of constraints: sensitivity to change in incidence; large spatial coverage; robustness regarding dynamic changes in the network composition; and efficient use of resources. In this last respect, seemingly reasonable solutions may have undesirable features: we found here that selecting one GP in each of the 100 largest French towns would lead to large underestimation of incidence. Several approaches have been described to the

---

**Fig. 2** Average cumulative numbers of influenza-like illness (ILI) cases reported by French sentinel general practitioners during the 2012/13 influenza epidemic versus general practitioners (GP) density (number of GPs per 100,000 inhabitants) by district (LAU1), France
Fig. 3 Spatial spread of general practitioners (GPs) involved in the simulated practice-based surveillance networks, France; number of GPs involved in each network is reported in brackets.

Fig. 4 Influenza-like illness incidence rates from simulations (real) and estimated by Horvitz-Thompson estimator using various general practitioners (GPs) networks in France. Number of GPs involved in each network was reported in brackets.
optimal selection of data providers in surveillance, taking into account spatial coverage or representativeness of monitored population [3–7]. In our experience, the effective use of such approaches in GP based surveillance is difficult as it is not possible to have a stable roster of GPs who want to participate in surveillance [8]. The interest of a priori computations to identify an optimal set of providers is further reduced if candidates can refuse participation. Furthermore, turnover of participants, which is common [1], may jeopardize carefully arranged providers sets. Interestingly, we found that in France, the maximum coverage algorithm proposed by Polgreen et al. [3] ended up selecting places with slightly below average GP density, leading to upward biased incidences. It may be possible to alter this coverage algorithm so that the only GPs proposed for inclusion are those practicing in areas with average GP density, at the price of reducing coverage. Unexpectedly, the self-selection of GPs in the Sentinelles network led to an average GP density in the monitored regions that was close to the average national GP density. This suggests that participation to surveillance is indeed independent from local GP density, a feature that is required in the estimators described above.

When a priori optimal selection of data providers for surveillance is not carried out, post-hoc stratification can address the issue. This requires identifying characteristics associated with the estimated quantity, as local GP density with the number of cases reported by SGPs. At worst, calibration with an irrelevant characteristic increases variance but not bias [28]. A feature of practical importance is that local GP density is in most situations easy to obtain, when other candidate characteristics to reduce bias, like consultation numbers [19], is more difficult to obtain. Additionally, GP density is easily recomputed when a GP joins or stops surveillance. Post-stratification may furthermore be used in all kinds of GP selection schemes, including the algorithmic approaches described above.

We used a simulation approach to investigate the impact of spatial sampling of providers on incidence estimates. The model for ILI simulations [16] was previously shown to reproduce the major characteristics of flu epidemics. The various observation networks that we compared corresponded with plausible choices, for example large cities (similar to the 122 cities mortality system of CDC [29]) or maximum coverage. That there is bias associated with these choices is important to stress, as these may be considered in setting up surveillance networks. As with all surveillance network data, empirical validation is difficult as the real incidence is unknown. The performance of estimators is often assessed by bias^2 + variance. Since improved estimates do not increase bias, reducing variance provides good evidence of improved performance in our situation.

**Conclusions**

Bias in incidence estimated through epidemiological data from voluntary surveillance networks could be reduced by using sampling weights based on local GP density. It can contribute to improving disease monitoring when the selection of participants is not controllable.

![Fig. 5 Root mean square error of incidence estimates based on estimators accounting for local GP density compared to real incidences in the various simulated networks](image)

---

**Souty and Boelle BMC Medical Research Methodology (2016) 16:156**

Page 7 of 8
Additional file

Additional file 1: “Computational details”. Detailed computations of incidence estimators. (DOCX 39 kb)

Abbreviations
Cl: Confidence interval; GP: General practitioner; ILI: Influenza-like illness; LAU: Local administrative units; NUTS: Nomenclature of territorial units for statistics; RMSE: Root mean squared error; SGP: Sentinel general practitioner

Acknowledgments
We thank all participating general practitioners of the Sentinelles network.

Funding
This work was funded by Université Pierre et Marie Curie (France).

Availability of data and materials
The dataset of incidence estimated by the French Sentinelles network are publicly available in the repository: http://www.sentientweb.fr/?page=databases.

Authors’ contributions
CS and PYB concept and performed the analyses. CS and PYB drafted and revised the manuscript. Both authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
The protocol was conducted in agreement with the Helsinki declaration. Authorization was obtained from the French Data Protection Agency (CNIL, registration number #471393).

Author details
1 Sorbonne Universités, UPMC Univ Paris 06, INSERM, Institut Pierre Louis d’épidémiologie et de Santé Publique (IPLESP UMRS 1136), F-75012 Paris, France. 2 Département de santé publique, AP-HP, Hôpital Saint-Antoine, F-75012 Paris, France.

Received: 2 June 2016 Accepted: 7 November 2016

Published online: 15 November 2016

References