Moriña *et al. BMC Medical Research Methodology* (2021) 21:6
https://doi.org/10.1186/s12874-020-01188-4

BMC Medical Research Methodology

RESEARCH ARTICLE **Open Access**

Quantifying the under-reporting of uncorrelated longitudal data: the genital warts example

David Moriña[1,2*], Amanda Fernández-Fontelo[3], Alejandra Cabaña[4], Pedro Puig[4], Laura Monfil[5], Maria Brotons[5] and Mireia Diaz[5]

Abstract

Background: Genital warts are a common and highly contagious sexually transmitted disease. They have a large economic burden and affect several aspects of quality of life. Incidence data underestimate the real occurrence of genital warts because this infection is often under-reported, mostly due to their specific characteristics such as the asymptomatic course.

Methods: Genital warts cases for the analysis were obtained from the Catalan public health system database (SIDIAP) for the period 2009-2016. People under 15 and over 94 years old were excluded from the analysis as the incidence of genital warts in this population is negligible. This work introduces a time series model based on a mixture of two distributions, capable of detecting the presence of under-reporting in the data. In order to identify potential differences in the magnitude of the under-reporting issue depending on sex and age, these covariates were included in the model.

Results: This work shows that only about 80% in average of genital warts incidence in Catalunya in the period 2009-2016 was registered, although the frequency of under-reporting has been decreasing over the study period. It can also be seen that this issue has a deeper impact on women over 30 years old.

Conclusions: Although this study shows that the quality of the registered data has improved over the considered period of time, the Catalan public health system is underestimating genital warts real burden in almost 10,000 cases, around 23% of the registered cases. The total annual cost is underestimated in about 10 million Euros respect the 54 million Euros annually devoted to genital warts in Catalunya, representing 0.4% of the total budget.

Keywords: Genital warts, Estimation, HPV, Under-reporting, Time series

Background

Health information systems are essential to ensure the safety and quality of health care and improve adherence to clinical practice guidelines, but they are also a very powerful tool concerning resources management and control, decision making, and effective and efficient planning of prevention and control interventions [1, 2]. However, the incompleteness and inaccuracy of the information is common in this type of registries and can lead to problems at a clinical level, but also at a population level such as the underestimation of some diseases. In Catalunya (Spain), the Information System for Research in Primary Care (SIDIAP) was launched in 2010 with the integration of data from the clinical work station of primary care (ECAP) of the Catalan Health Institute (ICS), which

*Correspondence: dmorina@ub.edu
[1] Department of Econometrics, Statistics and Applied Economics, Riskcenter-IREA, Universitat de Barcelona (UB), Avinguda Diagonal, 690, 08034 Barcelona, Spain
[2] Centre de Recerca Matemàtica, Universitat Autònoma de Barcelona (UAB), Edifici C, Campus de Bellaterra, 08193 Cerdanyola del Vallès, Spain
Full list of author information is available at the end of the article

started in 1998, and other complementary sources [3]. The ICS is the main provider of health services in Catalunya and manages 283 out of 370 Primary Care Teams with a catchment of 5,564,292 people, approximately 74% of the Catalan population (http://ics.gencat.cat/es/lics/). Nevertheless, it is reasonable to assume that the incidence of genital warts (GW) will be very similar among the Catalan population not covered by ICS. In the particular case of sexually transmitted diseases, it is even more important to have reliable information due to their remarkable morbidity, and therefore, the importance of controlling trends over time and priority setting (see [4] for a comprehensive discussion focused on developing countries). GW are a common and highly contagious sexually transmitted disease in Catalunya (in 2016 the incidence was about 107 cases per 100,000 women and 139 cases per 100,000 men[5]) caused by a subset of HPV types, with the most common being genotypes 6 and 11. They are usually benign, or non-cancerous, skin growths that develop on the genital area. However, they have an important negative impact on the health service and the individual, in addition to have a large economic burden and affect several aspects of quality of life [6–8]. A higher risk of CIN2+ lesions in women following a GW diagnose has been reported in a comprehensive recent study, even more than four years after the GW diagnose [9]. It is well known that incidence data underestimate, to some degree, the real occurrence of genital warts because this infection is often under-reported, mostly due to their specific characteristics such as the asymptomatic course of the disease [10]. This issue might be even more severe in specific vulnerable populations as imprisoned women [11]. Further, the SIDIAP database only includes data from the public healthcare sector and around 28% of the general population in Catalunya have a double health insurance coverage, public and private, so this fact can also explain why GW incidence rates are underestimated [12], although this source of under-reporting cannot be detected by the proposed model as we only have data from the public health system. There has been a growing interest in the past recent years to deal with data that are only partially registered or under-reported in the biomedical literature [13–18]. Most of these previous works deal with discrete-valued time series, whereas this paper is focused on the incidence of a disease, which should be treated as a continuous-valued time series. Therefore, the aim of this work is to quantify the under-reporting of genital warts cases in Catalunya and the reconstruction of the actual incidence in the period 2009-2016 on the basis of the mixture model described in the next Section.

Methods

Population and incidence estimation

The study population included all residents in Catalunya assigned to an ICS primary care center (74% of the Catalan population). Monthly GW incident cases for the analysis were obtained from the SIDIAP database for the period 2009-2016. Episodes of GW were classified as incident if they were preceded by at least 12-month period without any episode. People under 15 and over 94 years old were excluded from the analysis as the incidence of GW in this population is negligible (averages of 0.24 cases and 0.22 x 100,000 individuals over the period of study respectively).

Model

Consider X_t the series of real GW incidence, where $t = 1, 2, \ldots$ is the time, following a normal distribution with mean μ and variance σ^2. In our setting, this process cannot be directly observed, and all we can see is a part of it, expressed as

$$Y_t = \begin{cases} X_t & \text{with probability } 1 - \omega_t \\ q \cdot X_t & \text{with probability } \omega_t \end{cases} \tag{1}$$

The series Y_t represents the registered values corresponding to GW incidence in the part of Catalunya covered by ICS. According to Eq. (1), the registered observations series Y_t is a mixture of two normally distributed random variables $Y_t = Y_{1t}$ with probability $(1 - \omega_t)$ and $Y_t = Y_{2t}$ with probability ω_t, where Y_{1t} coincides with the unobserved process X_t and Y_{2t} is a normal random variable with mean $q \cdot \mu$ and variance $q^2 \cdot \sigma^2$. The parameter ω_t is modeled as $logit(\omega_t) = \alpha_0 + \alpha_1 \cdot t$ and can be interpreted as the frequency of under-reporting at a time t, while q can be interpreted as the intensity of such under-reporting, both taking values between 0 and 1. When $q = 0$ the observed incidence is $Y_t = 0$ and when $q = 1$ there is no under-reporting. A value of ω_t equal to 0 indicates that the observed value at time t is not under-reported, and a value of ω_t equal to 1 means that under-reporting is for sure happening. In order to detect potential differences in GW incidence depending on sex (men and women) and age (16-29 and 30-94), these covariates were included in the model, so the mean of the observed process Y_{1t} was modeled as $\mu_{1,t} = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot a + \beta_3 \cdot s + \beta_4 \cdot a * s$ (where a is the age, s is the sex and $a * s$ is the interaction between age and sex). The average of the second component Y_{2t} can be recovered as $\mu_{2,t} = q \cdot (\beta_0 + \beta_1 \cdot t + \beta_2 \cdot a + \beta_3 \cdot s + \beta_4 \cdot a * s)$. After fitting the previous model and performing residuals examination, a seasonal behavior with period 3 months was observed. Hence the model was updated by including the following trigonometric function to reflect this periodic behavior: $f(t) = \beta_5 \cdot sin\left(\frac{2 \cdot \pi \cdot t}{3}\right) + \beta_6 \cdot cos\left(\frac{2 \cdot \pi \cdot t}{3}\right)$ on the terms $\mu_{1,t}$ and $\mu_{2,t}$. Other similar models were considered and the best fitting one according to the validation process described in the next Section was chosen. In particular, as coefficients β_1 and β_6 are not significant, models without linear trend and with only one periodicity term were considered but the resulting validations were not satisfactory. Therefore, the final expressions were $\mu_{1,t} = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot a + \beta_3 \cdot s + \beta_4 \cdot a * s + \beta_5 \cdot sin\left(\frac{2 \cdot \pi \cdot t}{3}\right) + \beta_6 \cdot cos\left(\frac{2 \cdot \pi \cdot t}{3}\right)$ and $\mu_{2,t} = q \cdot \left(\beta_0 + \beta_1 \cdot t + \beta_2 \cdot a + \beta_3 \cdot s + \beta_4 \cdot a * s + \beta_5 \cdot sin\left(\frac{2 \cdot \pi \cdot t}{3}\right) + \beta_6 \cdot cos\left(\frac{2 \cdot \pi \cdot t}{3}\right)\right)$. The estimates and their associated standard errors were obtained by maximizing the log-likelihood function described in Eq. (2) and from its Hessian matrix respectively, using the *nlm* procedure in R [19].

$$l(Y, \theta) = \sum_{t=1}^{n} \log \left((1 - \omega_t) \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(y_t - \mu_{1,t})^2}{2\sigma^2}} + \omega_t \frac{1}{\sqrt{2\pi} q\sigma} e^{\frac{(y_t - \mu_{2,t})^2}{2q^2\sigma^2}} \right), \tag{2}$$

where $Y = y_1, \ldots, y_n$ is the observed series, $\theta = (\alpha_0, \alpha_1, \gamma, \beta_0, \ldots, \beta_6, \sigma)$, $\omega_t = \frac{e^{\alpha_0 + \alpha_1 t}}{1 + e^{\alpha_0 + \alpha_1 t}}$, $q = \frac{e^{\gamma}}{1 + e^{\gamma}}$ and $\mu_{1,t}$ and $\mu_{2,t}$ are as defined before.

In order to get proper initial values for the maximization routine, an Expectation-Maximization (EM) algorithm for mixtures of linear regressions was used, through the R package *mixtools* [20]. The estimates provided by the EM algorithm could have been used directly, but although this methodology is widely used when dealing with mixtures of distributions, it is unable to produce standard errors directly [21], and this is an important drawback in our context and in many other situations. If the main focus was not on quantifying the under-reporting issue, an alternative approach to analyze these data might be a hierarchical generalized linear model with random effects [22], implemented in the R package *HGLMM* [23]. By means of this methodology the most likely unobserved real GW incidence process is reconstructed based on the classification (underreported or not underreported) given by the posterior probabilities for the observations, provided by the output of the *mixtools* procedure, and on the estimates of the parameters. All the R code used to fit the models and to obtain the reported results and figures is available as Supplementary material.

Validation

The goodness of fit of the proposed mixture approach can be assessed by means of the Akaike's Information Criterion (AIC) compared to a single normal model. In this case, this measure favors the proposed model (AIC: 1717.9) in front of the single normal model (AIC: 1826.1). The model has been validated by analyzing its residuals. Figure 1 shows that they behave like white noise as expected and that there are no significant auto-correlations that should be accounted for. The residuals r_t have been estimated as

$$\hat{r}_t = Y_t - \left(\hat{\omega}_t \cdot \hat{q} \cdot \left(\hat{\beta}_0 + \hat{\beta}_1 \cdot t + \hat{\beta}_2 \cdot a + \hat{\beta}_3 \cdot s + \hat{\beta}_4 \cdot a * s + \hat{\beta}_5 \cdot sin\left(\frac{2 \cdot \pi \cdot t}{3}\right) + \hat{\beta}_6 \cdot cos\left(\frac{2 \cdot \pi \cdot t}{3}\right)\right) + (1 - \hat{\omega}_t) \cdot \left(\hat{\beta}_0 + \hat{\beta}_1 \cdot t + \hat{\beta}_2 \cdot a + \hat{\beta}_3 \cdot s + \hat{\beta}_4 \cdot a * s + \hat{\beta}_5 \cdot sin\left(\frac{2 \cdot \pi \cdot t}{3}\right) + \hat{\beta}_6 \cdot cos\left(\frac{2 \cdot \pi \cdot t}{3}\right)\right)\right) \tag{3}$$

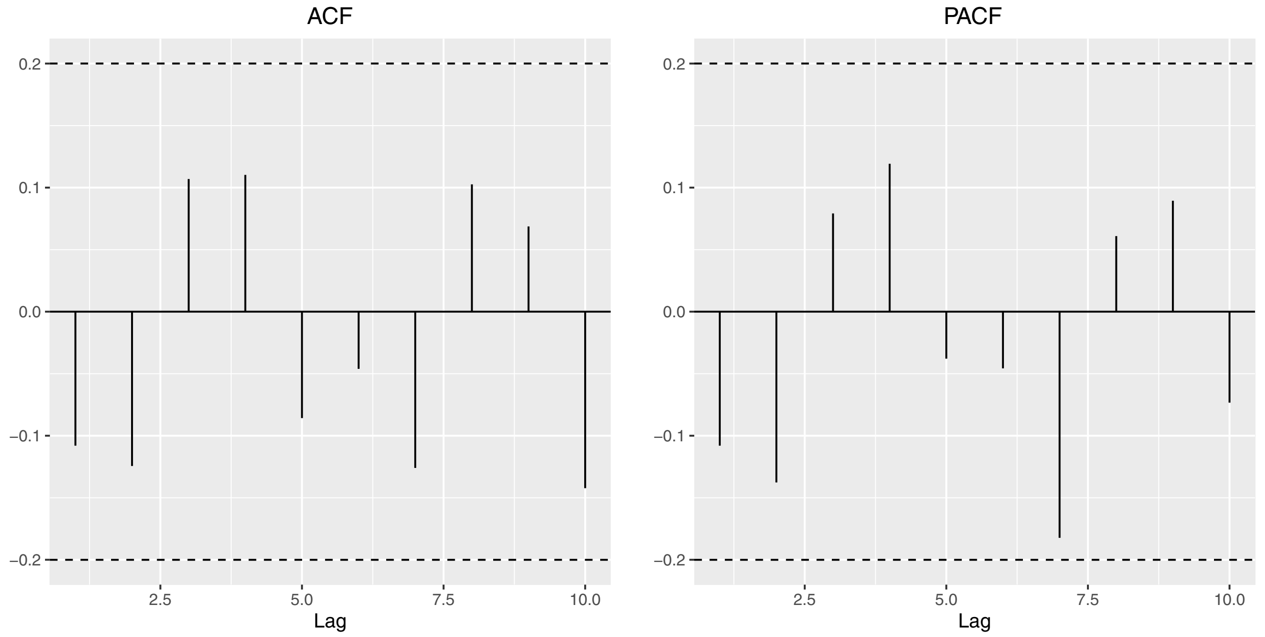

Fig. 1 Auto-correlations and partial auto-correlations of the model residuals

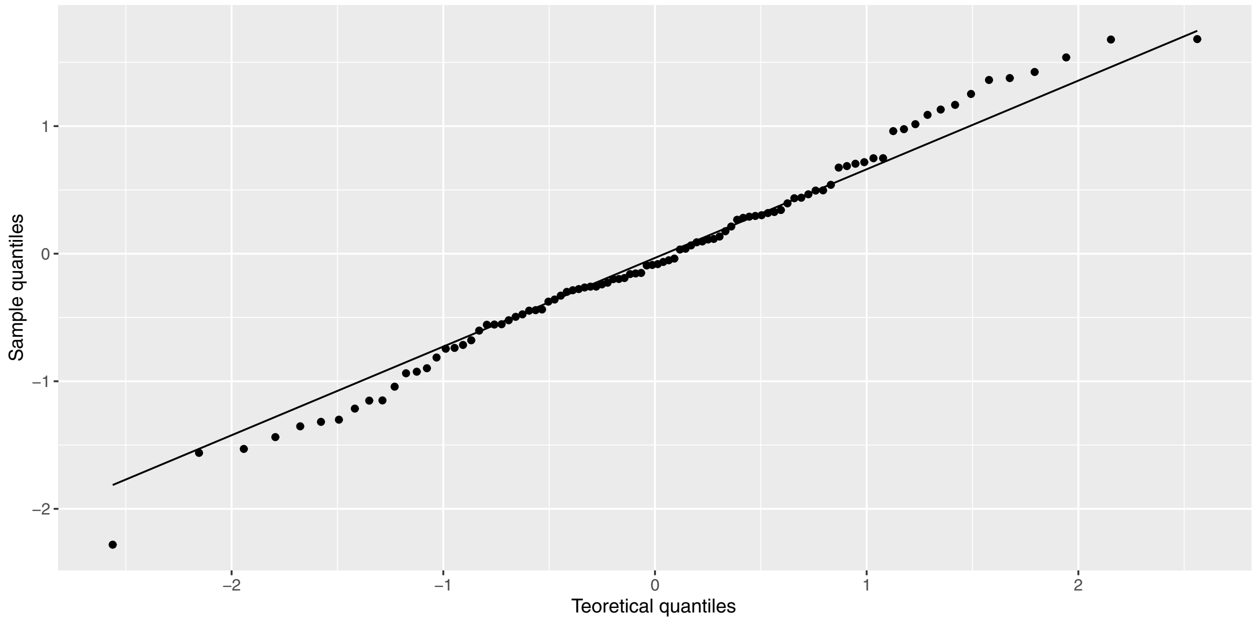

Fig. 2 Residuals qqplot

where Y_t is the total observed GW incidence at time t, and the letters with a hat $\hat{()}$ represent the estimated parameters.

If we were dealing with counts (number of cases) instead of incidence, the underlying distribution might be a Poisson, although the monthly number of GW cases is large enough to be approximated by a normal distribution. Additionally, the assumption that the underlying distributions of the two processes are Gaussian seems reasonable considering the qqplot of the residuals shown in Fig. 2.

Results

Our analysis estimates that, globally, only around 80% of actual GW incidence was registered in the SIDIAP database in the period 2009-2016. For women over 30 years old, the monthly average registered incidence is 3.9 cases per 100,000 women, while the estimated monthly incidence is 4.9 cases per 100,000 women, 24.9% higher. On males over 30 years old, the registered series has a monthly average of 5.9 cases per 100,000 men for 7.1 cases per 100,000 men on the reconstructed series, 21.8% higher. Regarding males under 30 years old, the reconstructed series is 13.3% higher (monthly averages of 18.4 and 20.8 cases per 100,000 men for the registered and reconstructed processes respectively). For women under 30 years old, the monthly average registered incidence of GW in Catalunya is 19.0 per 100,000 women, while the reconstructed hidden process has an average of 23.0 cases per 100,000 women, about 21.0% larger. This information is summarized in Table 1 and described in more detail in the Supplementary material (Table S1).

Table 2 shows the estimated effect of the age and sex over the under-reporting issue. In particular, it can be seen that the GW incidence is higher among younger populations and men. It can also be noticed that a significant interaction between sex and age group is found, which can be interpreted as a distinguishable impact of sex on GW incidence depending on the age group.

Table 1 Registered and estimated GW monthly average incidence (number of cases x 100,000 individuals) in the period 2009-2016

Sex	Age	Incidence (registered)	Incidence (estimated)	Difference (%)
	15-29	19.0	23.0	21.0%
Females	30-94	3.9	4.9	24.9%
	Average	6.8	8.4	23.2%
	15-29	18.4	20.8	13.3%
Males	30-94	5.9	7.1	21.8%
	Average	8.3	9.8	18.3%
Global		7.6	9.1	19.9%

Table 2 Parameter estimates

Covariate	Parameter	Estimate (95% CI)
	α_0	2.99 (1.77; 4.20)
t	α_1	-4.31 (-6.53; -2.09)
	β_0	13.76 (7.11; 20.40)
t	β_1	0.36 (-12.75; 13.46)
age	β_2	-13.53 (-14.13; -12.92)
sex	β_3	-1.60 (-2.24; -0.95)
age ∗ *sex*	β_4	3.25 (2.44; 4.06)
	β_5	4.16 (0.44; 7.88)
	β_6	0.52 (-5.59; 6.64)
	q	0.75 (0.72; 0.77)

Figure 3 shows the registered (solid black line) and reconstructed unobserved (dashed red line) processes for each of the considered sub-populations. Although this figure shows increasing trends for all series, they are not well explained by coefficient β_1, which is not significantly different from zero. Increasing trends are mainly explained by the significant coefficient α_1, which leads to a decreasing frequency of under-reporting ω_t.

The under-reporting frequency is about 95% in 2009 (ω_1) and around 21% in 2016 (ω_{96}). This is measured by parameter α_1 in model (1), and should not be confused to overall under-reporting of the data, as its intensity (measured by parameter q in the model) also plays a crucial role. For instance, all observations in a certain period of time could be slightly under-reported ($\omega = 1$, q near to 1), resulting in small differences between registered and estimated values or just a few observations might be under-reported (ω near to zero) but with a high intensity (q near to zero), potentially resulting on large differences between registered and estimated values. Table 3 shows the total number of GW cases registered in the SIDIAP in the period of study, the reconstructed values according to these registered cases and the projection over the whole Catalan population, assuming that the incidence on the area outside ICS coverage is the same.

Discussion

The results of this work show that in relative terms, the under-reporting issue has a deeper impact on people over 30 years old (where GW incidence is lower), especially among women. Nonetheless, the relative difference between registered and estimated annual averages range

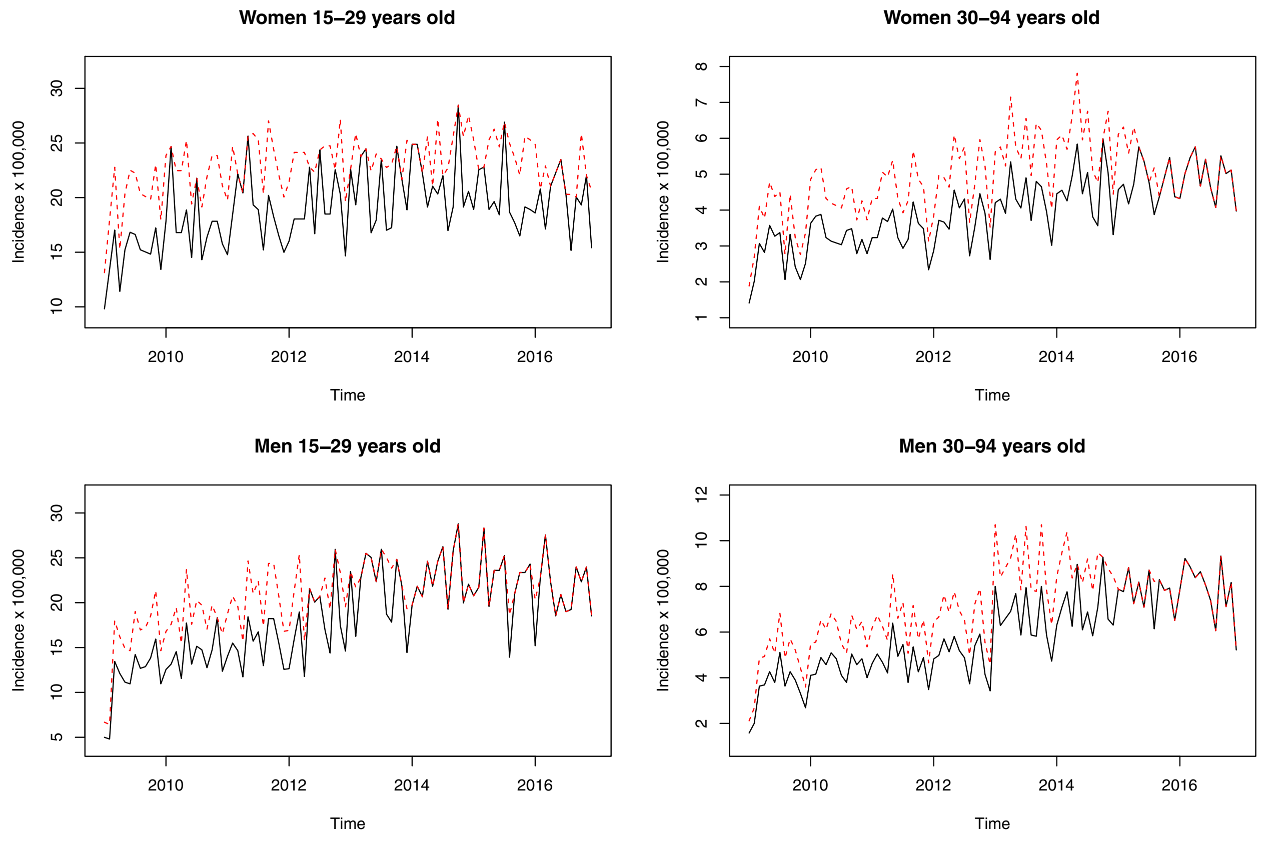

Fig. 3 Registered (solid black line) and estimated underlying series (dashed red line) for each of the considered sub-populations

Table 3 Registered, estimated and projected number of GW cases in Catalunya

Sex	Age	SIDIAP (registered)	SIDIAP (estimated)	Catalunya (registered projection)	Catalunya (estimated projection)
	15-29	8,051	9,769	10,280	12,460
Females	30-94	7,625	9,520	9,062	11,337
	Total	15,676	19,289	19,342	23,797
	15-29	7,967	9,097	10,166	23,797
Males	30-94	10,774	13,842	12,914	16,598
	Total	18,741	22,939	23,080	28,182
Global		34,417	42,228	42,422	51,979

between 13.3% and 24.9%. It is also remarkable that the quality of SIDIAP register regarding GW in Catalunya has been significantly improving during the study period, as the frequency of under-reported observations has been decreasing over time. Facing under-reported information from public health registers is very common in many situations, especially regarding potentially asymptomatic diseases like GW. The proposed methodology considers the potential under-reporting in continuous time series data in a very flexible way, estimating its frequency and intensity, and it is general enough to be appropriate in a wide range of real situations in the public health context. Additionally, the most likely non-observed process can be reconstructed on the basis of estimated posterior probabilities. Moreover, the GW data show that these models can deal with time-dependent under-reporting parameters, seasonal behavior, trends and also incorporate the effect of other factors by including covariates.

One of the potential limitations of this study is that the database used included data from the public healthcare setting and not from the private sector. In Catalunya, it is estimated that 33% of women and 25% of men aged 15 to 44 years have a double health insurance coverage (i.e. the public health insurance and a private insurance plan) [12], so the rates estimated in our study are likely still underestimating the real incidence of GW. One of its strengths is that the same methodology (possibly with minor model modifications) could be used to analyze the frequency and intensity of potential under-reporting issues for any condition or setting in the absence of temporal dependence among the observations.

Conclusions

The GW incidence registered in SIDIAP is underestimating the real burden in almost 10,000 cases in Catalunya, around 23% of the registered cases. The annual per person cost of GW was around 1000 Euros [8], so the potential total annual cost is underestimated in at least about 10 million Euros respect the 54 million Euros devoted to GW in Catalunya annually, representing 0.4% of the total budget of the Catalan Government intended for health, although about 2.8 million Euros would correspond to private insurances. It is, therefore, clear that knowing the true burden of GW at the general population level is important for health policy makers, especially after the introduction of prophylactic vaccines against HPV in many countries, as it plays a crucial role in developing and evaluating prevention strategies [24, 25]. This work presents a methodology that opens a wide field for future research lines. In particular, if temporal correlations are found in the data, an appropriate model should take this structure into account, similarly to [13, 18].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-020-01188-4.

Additional file 1: Additional tables and R code to reproduce the analyses.

Abbreviations

AIC: Akaike's information criterion; CI: Confidence interval; ECAP: e-Centre d'atenció primària (e-Primary attention center); EM: Expectation-maximization; ICS: Institut català de la salut (Catalan health institute); GW: Genital warts; SIDIAP: Sistema d'informació per al desenvolupament de la investigació en atenció primària (Information system for the development of primary attention research)

Acknowledgements

We acknowledge the SIDIAP, with special thanks to Maria Aragon for her help in data collection. We thank CERCA Programme / Generalitat de Catalunya for institutional support.

Authors' contributions

DM, AFF, AC and PP developed the statistical model. DM implemented it in R software. LM, MB and MD helped providing an epidemiological context for the methodology in the framework of genital warts and enriched the discussion focusing on the relevance on public health practice. All authors have read and approved the manuscript.

Funding

David Moriña acknowledges financial support from the Spanish Ministry of Economy and Competitiveness, through the María de Maeztu Programme for Units of Excellence in R&D (MDM-2014-0445) and Fundación Santander Universidades. This work has partial funding promoted by the Department of Health of the Generalitat de Catalunya for the execution of the project "Monitorización y evaluación del impacto de la introducción de nuevas estrategias preventivas del cáncer de cuello de útero en Catalunya" (reference 0599S/7613/2010). This work was partially funded by the Instituto de Salud Carlos III-ISCIII (Spanish Government) through the projects PIE16/00049, PI16/01254, PI16/01056, PI19/01118, Río Hortega CM15/00061 (Co-funded by FEDER funds / European Regional Development Fund. ERDF, a way to build

Europe), Agència de Gestió d'Ajuts Universitaris i de Recerca (2017SGR1718) and by grant RTI2018-096072-B-I00 from the Spanish Ministry of Science, Innovation and Universities. The funding sources had no role in the data collection, analysis or interpretation of the results.

Availability of data and materials

R codes used in the described analyses are available as Supplementary material.

Ethics approval and consent to participate

This study was approved by the Clinical Research Ethics Committee and the Institutional Review Board of the University Institute for Primary Care Research (IDIAP) Jordi Gol (P15/106).

Consent for publication

Not applicable.

Competing interests

The Cancer Epidemiology Research Programme, where LM, MB and MD are affiliated to, has received institutional sponsorship for grants from Merck and GlaxoSmithKline.

Author details

[1]Department of Econometrics, Statistics and Applied Economics, Riskcenter-IREA, Universitat de Barcelona (UB), Avinguda Diagonal, 690, 08034 Barcelona, Spain. [2]Centre de Recerca Matemàtica, Universitat Autònoma de Barcelona (UAB), Edifici C, Campus de Bellaterra, 08193 Cerdanyola del Vallès, Spain. [3]Chair of Statistics, School of Business and Economics, Humboldt-Universität zu Berlin, Berlin, Germany. [4]Barcelona Graduate School of Mathematics (BGSMath), Departament de Matemàtiques, Universitat Autònoma de Barcelona (UAB), Edifici C, Campus de Bellaterra, 08193 Cerdanyola del Vallès, Spain. [5]Unit of Infections and Cancer - Information and Interventions (UNIC - I&I), Cancer Epidemiology Research Program (CERP), Catalan Institute of Oncology (ICO)-IDIBELL, L'Hospitalet de Llobregat, Spain.

Received: 2 July 2020 Accepted: 4 December 2020
Published online: 06 January 2021

References

1. Groseclose SL, Buckeridge DL. Public Health Surveillance Systems: Recent Advances in Their Use and Evaluation. Ann Rev Public Health. 2017;38(1): 57–79. https://doi.org/10.1146/annurev-publhealth-031816-044348.
2. Ford MA, Spicer CM. Monitoring HIV Care in the United States: Indicators and Data Systems; 2012. http://www.nap.edu/catalog.php?record_id=13225. Accessed 12 Apr 2019.
3. SIDIAP, Information System for Research in Primary Care [Internet]. [cited 2019 Mar 29]. https://www.sidiap.org/.
4. McCormack D, Koons K. Sexually Transmitted Infections. W.B. Saunders. 2019. https://doi.org/10.1016/j.emc.2019.07.009.
5. Brotons M, Monfil L, Roura E, Duarte-Salles T, Casabona J, Urbiztondo L, Cabezas C, Bosch FX, de Sanjosé S, Bruni L. Impact of a single-cohort HPV vaccination strategy with quadrivalent vaccine in northeast Spain: Population-based analysis of genital warts in men and women. In: EUROGIN; 2018. https://www.eurogin.com/content/dam/Informa/eurogin/previous/Abstracts-Eurogin-2018.pdf.
6. Woodhall SC, Jit M, Soldan K, Kinghorn G, Gilson R, Nathan M, Ross JD, Lacey CJN, study group Q. The impact of genital warts: loss of quality of life and cost of treatment in eight sexual health clinics in the UK. Sex Transm Infect. 2011;87(6):458–63. https://doi.org/10.1136/sextrans-2011-050073.
7. Sénécal M, Brisson M, Maunsell E, Ferenczy A, Franco EL, Ratnam S, Coutlée F, Palefsky JM, Mansi JA. Loss of quality of life associated with genital warts: baseline analyses from a prospective study,. Sex Transm Infect. 2011;87(3):209–15. https://doi.org/10.1136/sti.2009.039982.
8. Castellsagué X, Cohet C, Puig-Tintoré LM, Acebes LO, Salinas J, Martin MS, Breitscheidel L, Rémy V. Epidemiology and cost of treatment of genital warts in Spain. Eur J Public Health. 2009;19(1):106–10. https://doi.org/10.1093/eurpub/ckn127.
9. Blomberg M, Dehlendorff C, Kjaer SK. Risk of CIN2+ following a diagnosis of genital warts: A nationwide cohort study. Sex Transm Infect. 2019;95(8): 614–8. https://doi.org/10.1136/sextrans-2019-054008.
10. Hsueh P-R. Human papillomavirus, genital warts, and vaccines. J Microbiol Immunol Infect Wei mian yu gan ran za zhi. 2009;42(2):101–6.
11. Escobar N, Plugge E. Prevalence of human papillomavirus infection, cervical intraepithelial neoplasia and cervical cancer in imprisoned women worldwide: A systematic review and meta-analysis. BMJ Publ Group. 2020. https://doi.org/10.1136/jech-2019-212557.
12. Enquesta de salut de Catalunya (ESCA 2017) [Internet]. Departament de Salut. 2017 [cited 2019 Mar 31]. https://salutweb.gencat.cat/web/.content/_departament/estadistiques-sanitaries/enquestes/Enquesta-de-salut-de-Catalunya/Resultats-de-lenquesta-de-salut-de-Catalunya/documents/els40prals_2017_web.xlsx.
13. Fernández-Fontelo A, Cabaña A, Puig P, Moriña D. Under-reported data analysis with INAR-hidden Markov chains. Stat Med. 2016;35(26):4875–90. https://doi.org/10.1002/sim.7026.
14. Bernard H, Werber D, Höhle M. Estimating the under-reporting of norovirus illness in Germany utilizing enhanced awareness of diarrhoea during a large outbreak of Shiga toxin-producing E. coli O104:H4 in 2011–a time series analysis. BMC Infect Dis. 2014;14:116. https://doi.org/10.1186/1471-2334-14-116.
15. Alfonso JH, Løvseth EK, Samant Y, Holm J.-Ø.. Work-related skin diseases in Norway may be underreported: data from 2000 to 2013. Contact Dermatitis. 2015;72(6):409–12. https://doi.org/10.1111/cod.12355.
16. Rosenman KD, Kalush A, Reilly MJ, Gardiner JC, Reeves M, Luo Z. How much work-related injury and illness is missed by the current national surveillance system? J Occup Environ Med. 2006;48(4):357–65. https://doi.org/10.1097/01.jom.0000205864.81970.63.
17. Arendt S, Rajagopal L, Strohbehn C, Stokes N, Meyer J, Mandernach S. Reporting of foodborne illness by U.S. consumers and healthcare professionals,. Int J Env Res Public Health. 2013;10(8):3684–714. https://doi.org/10.3390/ijerph10083684.
18. Fernández-Fontelo A, Cabaña A, Joe H, Puig P, Moriña D. Untangling serially dependent underreported count data for gender-based violence. Stat Med. 2019;38(22):4404–22. https://doi.org/10.1002/sim.8306.
19. R Core Team. R: A Language and Environment for Statistical Computing. 2019. https://www.r-project.org/.
20. Benaglia T, Chauveau D, Hunter DR, Young D. mixtools : An R Package for Analyzing Finite Mixture Models. J Stat Softw. 2009;32(6):1–29. https://doi.org/10.18637/jss.v032.i06.
21. Jamshidian M, Jennrich RI. Standard errors for EM estimation. J R Stat Soc Ser B (Stat Methodol). 2000;62(2):257–70. https://doi.org/10.1111/1467-9868.00230.
22. Lee Y, Nelder JA. Double Hierarchical Generalized Linear Models. J R Stat Soc Ser B (Methodol). 1996;58(4):619–78.
23. Molas M, Lesaffre E. Hierarchical generalized linear models: The R package HGLMMM. J Stat Softw. 2011;39(13):1–20. https://doi.org/10.18637/jss.v039.i13.
24. Kjær S, Tran T, Sparen P, Tryggvadottir L, Munk C, Dasbach E, Liaw K, Nygård J, Nygård M. The Burden of Genital Warts: A Study of Nearly 70,000 Women from the General Female Population in the 4 Nordic Countries. J Infect Dis. 2008;196(10):1447–54. https://doi.org/10.1086/522863.
25. Kostaras D, Karampli E, Athanasakis K. Vaccination against HPV virus: a systematic review of economic evaluation studies for developed countries: Taylor and Francis Ltd; 2019. https://doi.org/10.1080/14737167.2019.1555039.

Publisher's Note