Ondrikova et al. BMC Public Health (2021) 21:1245
https://doi.org/10.1186/s12889-021-11317-3

BMC Public Health

RESEARCH Open Access

Understanding norovirus reporting patterns in England: a mixed model approach

N. Ondrikova[1,2*], H. E. Clough[1,3], N. A. Cunliffe[1,4], M. Iturriza-Gomara[1,5], R. Vivancos[3,4,6] and J. P. Harris[1,4]

Abstract

Background: Norovirus has a higher level of under-reporting in England compared to other intestinal infectious agents such as *Campylobacter* or *Salmonella*, despite being recognised as the most common cause of gastroenteritis globally. In England, this under-reporting is a consequence of the frequently mild/self-limiting nature of the disease, combined with the passive surveillance system for infectious diseases reporting. We investigated heterogeneity in passive surveillance system in order to improve understanding of differences in reporting and laboratory testing practices of norovirus in England.

Methods: The reporting patterns of norovirus relating to age and geographical region of England were investigated using a multivariate negative binomial model. Multiple model formulations were compared, and the best performing model was determined by proper scoring rules based on one-week-ahead predictions. The reporting patterns are represented by epidemic and endemic random intercepts; values close to one and less than one imply a lower number of reports than expected in the given region and age-group.

Results: The best performing model highlighted atypically large and small amounts of reporting by comparison with the average in England. Endemic random intercept varied from the lowest in East Midlands in those in the under 5 year age-group (0.36, CI 0.18–0.72) to the highest in the same age group in South West (3.00, CI 1.68–5.35) and Yorkshire & the Humber (2.93, CI 1.74–4.94). Reporting by age groups showed the highest variability in young children.

Conclusion: We identified substantial variability in reporting patterns of norovirus by age and by region of England. Our findings highlight the importance of considering uncertainty in the design of forecasting tools for norovirus, and to inform the development of more targeted risk management approaches for norovirus disease.

Keywords: Norovirus, HHH4, Underestimation, Public health surveillance, Mixed-effects, Negative binomial

Background

Norovirus is recognised as the most common cause of diarrhoeal disease globally [1] but has the highest levels of under-reporting compared to other intestinal infectious agents such as *Campylobacter* or *Salmonella* [2]. In England, the level of this under-reporting is a consequence of the nature of the disease and the surveillance system for infectious diseases. The illness is characterised by a sudden onset of symptoms and is generally self-limiting, lasting around two to 3 days in otherwise healthy individuals, and most people recover without contacting medical services. However, some individuals suffer more severe disease outcomes in [3, 4]. Norovirus is not a notifiable disease in England. However, there is a statutory duty on the providers of diagnostic laboratory services to report to Public Health England (PHE) isolates of an infectious agent within 7 days [5] and norovirus is often reported this way. Reporting of outbreaks in care homes and in health care settings is encouraged by the regulator (Care Quality Commission) in England, but it remains voluntary. All

* Correspondence: nikola.ondrikova@liverpool.ac.uk
[1]Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Liverpool, UK
[2]Institute for Risk & Uncertainty, University of Liverpool, Liverpool, UK
Full list of author information is available at the end of the article

routine laboratory reports of norovirus are reported to the national laboratory surveillance system – Second-Generation Surveillance System (SGSS) [6].

Norovirus places a considerable burden on health care services in England both financial, with the cost estimated at between £63 and £106 million annually [7], and in terms of capacity adding significantly to the annual "winter pressures". To reduce this burden, PHE recommends that affected individuals stay at home until symptoms have resolved. While not seeking medical attention prevents norovirus from spreading, it limits the chances of isolated cases to feature in a national surveillance system [8]. Consequently, the surveillance system might be prone to represent cases associated with outbreaks particularly from semi-closed settings such as care homes and hospitals rather than those in the community [9].

A better understanding of the heterogeneity in passive surveillance system improves understanding of differences in reporting and laboratory testing practices across geographic regions and age groups. Statistical modelling approaches can help quantify this heterogeneity. Count data are often modelled using a Poisson distribution: however biological data commonly exhibit greater variation than the Poisson model can accommodate (called "over-dispersion"). The negative binomial distribution is a viable alternative. To account for heterogeneity due to individual or regional differences, random effects are commonly used. However, the use of random effects brings challenges when it comes to identifying the best model formulation. Typically, Akaike (AIC) or Bayesian information criterion (BIC) is used. When the model contains random effects the definition of the AIC and BIC is not straightforward [10], and the use of proper scoring rules is one approach to overcome these difficulties [11]. This approach has been described elsewhere [12]. Briefly, the approach assesses predictive distribution based on predictions from the proposed model.

We investigated the reporting patterns of norovirus relating to geographical region and age in England. A simpler analysis of earlier data which motivated the current study is reported in [13]. First, an age-stratified multivariate discrete spatio-temporal model was fitted. Then, we incorporated random effects into four model formulations which were evaluated against the simpler model and each other. This led to the selection of the best performing model based on one-week-ahead predictions. Finally, we used the best performing model to highlight regions with atypically large or small amounts of reporting by comparison with the average in England. This information is highly relevant for public health policy and planning but also for any research using routine data.

Methods

Data

All diagnostic laboratories in England report data to the SGSS. The process of data validation and management is described elsewhere, e.g. [6]. We obtained the weekly numbers of laboratory confirmed norovirus cases between week 27, 2014 (June) and week 26, 2019 (July) for nine regions of England from SGSS stratified by age. The period did not coincide with an emergence of a new strain; the norovirus Sydney2012 strain was dominant throughout the study [14].

The Office for National Statistics (ONS) provides age-stratified population estimates for England. We obtained regional population data stratified by six age groups: 0–4, 5–14, 15–24, 25–44, 45–64, 65+. Social contacts matrix for these age groups is based on physical as well as non-physical contacts of UK subset of the POLYMOD study [15].

We also obtained the numbers of primary schools and nurseries [16] and hospitals [17] in each region, since norovirus is known to cause outbreaks in closed and semi-closed environments. These regional counts were normalised to range from 0 to 1 and thus providing regional proportions.

The R code to import and prepare data for modelling is provided in public GitHub repository (see Availability of data and materials).

Statistical modelling

The multivariate time-series modelling framework allows for additive decomposition of aggregated time series into endemic and epidemic components representing origins of an infection spread [11, 18, 19]. The initial formulation described in Held et al. [18] was later extended to consider social contacts within a population [19], and heterogeneity was often found in count data [11]. We adopted these methods to investigate reporting patterns of norovirus in England, but the terminology is inherited.

The endemic component represents sporadic cases: in other words, the number of reported norovirus cases that would be expected in specific regions and age-groups in the absence of outbreaks. The epidemic component conceptually represents reported cases emerging from outbreaks. The spatial and temporal spread is specified in terms of power-law distance decay. One of the power-law decay benefits is the relaxation of a simple assumption that the epidemic can only spread to a directly neighbouring region, i.e. a region with a shared border. Moreover, when distance decay depends on the population, it can describe temporal (i.e. within-region) spread, often referred to as a gravity model, e.g. [20].

Weekly count of norovirus cases $(t = 1, \ldots, 52)$ per age group $(g = 1, \ldots, 6)$ per region of England $(r = 1, \ldots, 9)$ is

denoted by Y_{grt}. Then, conditional upon previous observations, Y_{grt} is assumed to follow a negative binomial distribution with mean

$$\mu_{grt} = pop_{gr}\nu_{grt} + \phi_{grt}\sum_{g'r'}\left\lfloor C_{g'g}W_{r'r}\right\rfloor Y_{g',r',t-1} \quad (1)$$

The population fraction pop_{gr} based on mid-2016 OSN estimates is used as an offset for the endemic component ν_{grt}. Since we were interested in reflecting the differences in population size across age groups and geographic units rather than temporal variation in the population per se, temporal variation in population data was not explicitly modelled. The row-normalised product of POLYMOD contact matrix $C_{g'g}$ and spatial weights $W_{r'r}$ summed over the age group g' and region r' forms the epidemic component $\phi_{grt.}$ In other words, the product determines how the counts from the previous period affect the current mean in the age group and region [19]. Mathematical expression of ν_{grt} and ϕ_{grt} in fixed-effect formulations from the previous equation are

$$\begin{aligned}\log(\nu_{grt}) = {} & \alpha_0^{(\nu)} + \alpha_g^{(\nu)} + \alpha_1 x_{1r}^{(\nu)} + \alpha_2 x_{2r}^{(\nu)} \\ & + \alpha_3 x_{3r}^{(\nu)} + \beta_t + \sum_{s=2}^{s}\{\gamma_s^{(\nu)}\sin(\omega_s t) \\ & + \delta_s^{(\nu)}\cos(\omega_s t)\}\end{aligned} \quad (2)$$

$$\begin{aligned}\log(\phi_{grt}) = {} & \alpha_0^{(\phi)} + \alpha_g^{(\phi)} + \tau log(pop_{gr}) \\ & + \sum_{s=0}^{s}\{\gamma_s^{(\phi)}\sin(\omega_s t) \\ & + \delta_s^{(\phi)}\cos(\omega_s t)\}\end{aligned} \quad (3)$$

We included age-specific fixed effects α_g in both components to account for age-related susceptibility, linear trend β_t and S number of harmonic waves in curly brackets where the γ_s and δ_s signify seasonal parameters and $\omega_s = 2\pi s/52$ represents Fourier frequencies for weekly data. Additionally, logged regional proportions of number of primary schools $\alpha_1 x_{1r}^{(\nu)}$, nurseries $\alpha_2 x_{2r}^{(\nu)}$ and hospitals $\alpha_3 x_{3r}^{(\nu)}$ were included as endemic covariates. For the rest of the model formulations, random effects $b_{gr}^{(.)}$ are added to capture any remaining heterogeneity; $b_{gr}^{(\nu)} \sim N(0,\sigma_\nu^2), b_{gr}^{(\phi)} \sim N(0,\sigma_\phi^2)$. These can be correlated or uncorrelated.

As previously mentioned, comparing models employing random effects can be challenging. As recommended by Paul and Held in [11], we use strictly proper scoring rules, namely ranked probability score (RPS) and logarithmic score (logS). The earlier is less sensitive to extreme values, whereas the latter will more strictly penalise them. In other words, logS is more sensitive to a misprediction in outbreak period than RPS. Both scores were calculated from one-week-ahead predictions based on unseen data from norovirus season 2018–2019. The train and test partitions are illustrated in Additional file 1. Model selection is then based on the scores and permutation tests, determining whether one score is significantly better than the other.

The analyses were conducted using R software [21]. Social contacts matrix was obtained from the R package 'socialMixer' [15], and modelling tasks were performed with R packages 'surveillance' [22] and 'hhh4contacts' [19] (Eqs. 1). The penalised log-likelihood in the hhh4 function is penalised using the quasi-Newton algorithm by default. In the case of the mixed-effects models, Nelder-Mead penalisation was preferred to maximise the marginal likelihood concerning the variance parameters [11]. The regression parameters were optimised on the log-scale.

Results

Initially, we compared fixed-effect models with one (S = 1), two (S = 2) and three (S = 3) seasonal waves in endemic component to determine the baseline model formulation (see Additional file 2). The predictive performance was not significantly different between these models and so two seasonal waves were selected as the baseline.

Table 1 shows that all the models were well-calibrated, as suggested by $p > .05$. A model is well-calibrated when its predictive distribution covers the observed value; for example, when the prediction for week 15 in South West England is 200 cases and the upper bound of the predictive distribution is 190 cases, miscalibration is suspected. Models including random effects with harmonic waves in the epidemic component achieved the lowest scores.

The two best performing models, B2 and C2, were selected for permutation test-based comparison. The comparison of error scores showed no difference between models in both, RPS ($p = 0.408$) and logS ($p = 0.067$). The B2 model showed lower score but as the models were not significantly different in the predictive performance the second-best model could have been selected as well.

The point estimates, confidence intervals (CI) and standard errors of the best performing negative binomial regression model with uncorrelated random effects and epidemic seasonal component are reported in Table 2. Considering fixed age-group coefficients, the higher epidemic intercept for the 65+ group (1.796, CI 1.085–2.971) compared to the other groups and the endemic intercept in the same group suggests there is a bias towards reporting of outbreak-generated cases from care homes and hospitals. Generally, the models without epidemic seasonality (B1, C1) suggest that 87% of the reported cases originate from outbreaks, and 13% are endemic in nature. However, models considering seasonal waves (B2, C2) in epidemic component showed

Table 1 Performance evaluation of selected models based on proper scoring rules

Model	RPS		logS	
Endemic seasonality (S = 2) + linear trend + covariates	**Mean score**	**Calibration test (*p*-Value)**	**Mean score**	**Calibration test (*p*-Value)**
No random effects:				
A1 epidemic (S = 0)	0.938	0.385	1.480	0.270
A2 epidemic (S = 1)	0.934	0.251	1.480	0.177
Uncorrelated random effects:				
B1 epidemic (S = 0)	0.890	0.888	1.430	0.499
B2 epidemic (S = 1)	0.884	0.514	1.430	0.245
Correlated random effects:				
C1 epidemic (S = 0)	0.890	0.810	1.430	0.448
C2 epidemic (S = 1)	0.884	0.452	1.430	0.211

Table 2 Coefficient estimates from the best performing model (endemic seasonality (S2) + epidemic seasonality (S1) + trend + uncorrelated random effects)

	Estimates	CI 2.5%	CI 97.5%	Std. Error
Epidemic Component:				
Age [05–14]	0.041	0.024	0.072	0.012
Age [15–24]	0.064	0.040	0.101	0.015
Age [25–44]	0.044	0.024	0.080	0.014
Age [45–64]	0.114	0.063	0.206	0.034
Age [65+	1.796	1.085	2.971	0.461
Population Size	1.431	1.036	1.978	0.165
Sine (2* π *t/52)	0.929	0.871	0.992	0.032
Cosine (2* π *t/52)	0.744	0.697	0.794	0.019
Random Intercept	7.548	1.434	39.731	6.396
Endemic Component:				
Age [05–14]	0.180	0.092	0.350	0.061
Age [15–24]	0.115	0.058	0.227	0.040
Age [25–44]	0.136	0.071	0.263	0.046
Age [45–64]	0.154	0.080	0.297	0.052
Age [65+	0.355	0.181	0.695	0.122
Primary Schools (%)	0.869	0.657	1.149	0.142
Nurseries (%)	1.113	0.866	1.431	0.128
Hospitals (%)	0.997	0.898	1.107	0.053
Sine (2* π *t/52)	1.112	0.991	1.248	0.071
Cosine (2* π *t/52)	1.142	1.000	1.303	0.060
Sine (4* π *t/52)	0.850	0.781	0.925	0.038
Cosine (4* π *t/52)	0.830	0.771	0.894	0.019
Random Intercept	89.232	48.502	164.162	27.754
Spatial weights (d)	3.617	3.313	3.948	0.162
Overdispersion	1.337	1.305	1.370	0.012

that the proportion of outbreak-related reporting is lower in summer (57%). The relative contribution of endemic and epidemic components per region and age group is illustrated in Figs. 1 and 2 respectively. Figure 2 also shows high levels of within-group spread in small children and the elderly.

Norovirus reporting patterns

Reporting patterns were described in terms of endemic and epidemic random intercepts (RI) per age group and region (Fig. 3); values close to one and less than one imply lower number of reports than expected given the linear trend, harmonic waves, population structure, number of primary schools, nurseries and hospitals. Endemic random intercept varied from the lowest in East Midlands (UKF) in those in the under 5 year age-group (0.36, CI 0.18–0.72) to the highest in the same age group in South West (UKK)(3.00, CI 1.68–5.35) and Yorkshire & the Humber (UKE) (2.93, CI 1.74–4.94). Overall, regions displayed in purple consistently across the age groups (North West – UKD, West Midlands – UKG) are the most likely suspects for underestimation of norovirus burden (Fig. 3). In contrast, regions such as Yorkshire & the Humber and South West (UKK) are displayed in shades of green and blue.

Also, regions varied in age-related reporting patterns. For example, the lowest endemic RI for North East (UKC) was identified in those in the 15–24 age group (0.43, CI 0.18–1.04), for East Midlands (UKF) it was the young children (0.36, CI 0.18–0.72) and for South East (UKJ) it was school age children (0.44, CI 0.23–0.81). Some combinations of age groups and regions showed wide confidence interval ranges spanning from below one to over one pointing towards high levels of uncertainty, e.g. South West (UKK) in elderly (1.48, CI 0.80–2.74).

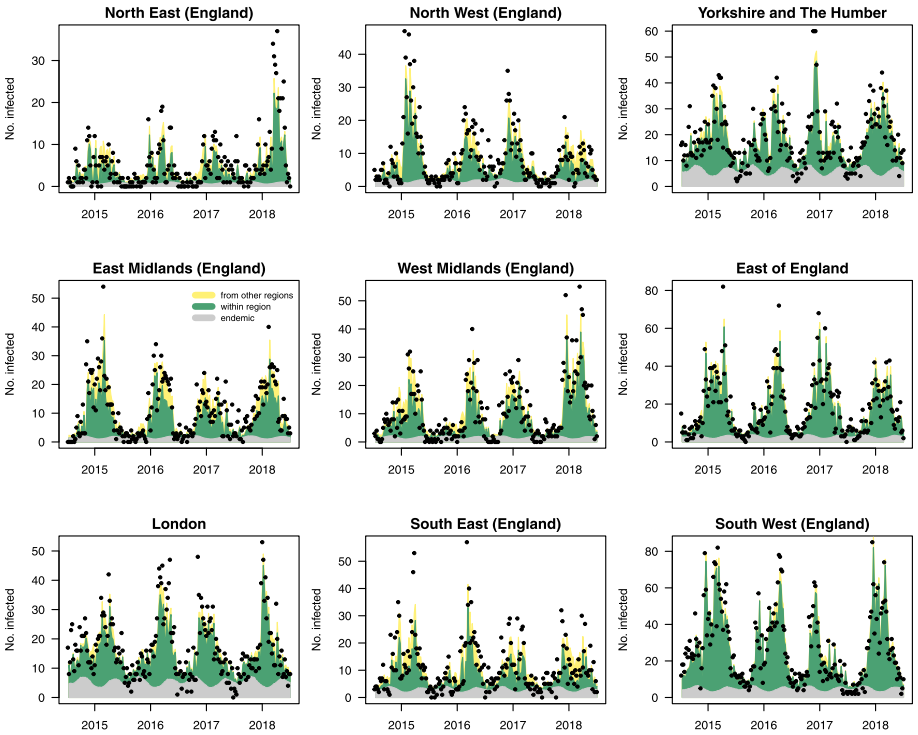

Fig. 1 Regional model fit and the relative contribution of model components. Epidemic (within + between) and endemic components per region, aggregated over age groups from the best performing model (B2). The dots represent the actual reported norovirus cases

As judged by variance, the regional differences in the epidemic RI were less pronounced ($Var(b_{gr}^{(\phi)}) = 0.129$) compare to endemic RI ($Var(b_{gr}^{(\nu)}) = 0.450$). As in the endemic random intercepts, the under 5 years age group showed the highest variation across regions with the highest epidemic RI in the South West (2.49, CI 1.83–3.38; < 5 yrs) and the lowest in the West Midlands (0.44, CI 0.29–0.66; < 5 yrs). All regions were closer to the expected incidence, except South East. In the South East, across all age groups, the epidemic RI was lower than the endemic with the lowest epidemic RI in elderly (0.71, CI 0.52–0.97). North East showed a slightly different pattern in epidemic and endemic RI. The model identified an unexpectedly low number of reports in epidemic RI but only in the age-groups from 5 to 64 years. In terms of endemic RI, young children were the only group to reach a value above one. Further details are provided in the Additional files 3 and 4. The patterns of heterogeneity were stable across models in both components.

Discussion

We aimed to describe norovirus reporting patterns in England using mixed-effect modelling. We started by describing a relationship between age-stratified weekly reports of norovirus by region and a set of predictors including seasonal waves, relevant regional covariates, spatial relationship between regions, within region norovirus activity and average contact between the age-groups. After determining the best model, we analysed endemic and epidemic random intercepts. We found that reporting practices vary greatly across regions and subpopulations, and that the seasonal changes in reporting related to differences between outbreaks and sporadic cases.

Context

Our analysis identified geographic areas where reporting of norovirus was lower than expected, given the age structure of the population, social contacts between groups and covariates such as number of primary schools, nurseries and hospitals. As explained by Gibbons et al. [23], there are two main reasons for disease burden underestimation: 1) Under-ascertainment and 2) Under-reporting. The former occurs when community cases do not seek healthcare, and the latter when cases presenting to healthcare do not reach the surveillance system due to failure to diagnose or report a pathogen correctly. Reporting patterns described in this study capture these instances with random effects. The endemic random effects in some regions were low for all the age groups (North West and West Midlands) suggesting that under-ascertainment or under-reporting is more likely

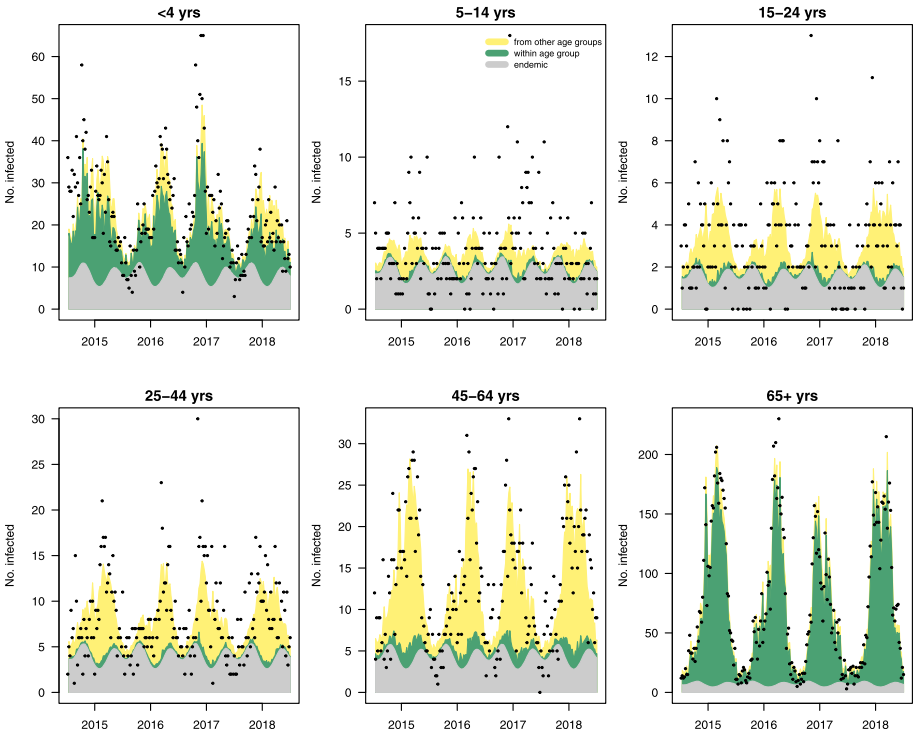

Fig. 2 Age-group-based model fit and the relative contribution of model components. Epidemic (within + between) and endemic components per age group, aggregated over regions from the best performing model (B2). The dots represent the actual reported norovirus cases

in these areas. Since norovirus is not a notifiable disease, this may lead to the perception that it is a low priority pathogen. The perception together with unavailability of a specific norovirus treatment could play a critical role when a clinician decides on whether to request a sample. For a reference laboratory reporting of an identified pathogen is mandatory through legislation [5]; therefore, here, the variability may be explained by differences in testing practices rather than reporting per se. Besides, the speed of recovery can be another factor as most of the people will recover between 12 and 72 h [3] and so they may not have contact with medical services to provide a sample during the period of illness. In contrast, for some regions, random effects were atypically high. This may in part reflect regions and laboratories that have historically functioned as sentinel surveillance centres for norovirus (for example Avon in South West as described in [24, 25]) or are very proactive in reporting (for example Yorkshire & the Humber), and to a degree may respond to particular research interests among virologists or infectious diseases specialists in the region.

Furthermore, our data suggest that the regional reporting variation was most pronounced in the subpopulation of young children, who had the largest difference in reporting between the most passive and active regions in both outbreaks and sporadic cases. Thus, in some areas, norovirus in younger children could be more likely to be underestimated compared to other groups. These results agree with previous studies indicating that norovirus in children is underdiagnosed in England [26]. However, further research is needed to clarify the extent of the issue compared to other subpopulations.

The model suggests that cases from outbreaks are more likely to be reported; disproportionally higher elderly populations were shown to be associated with an increased epidemic incidence of reported norovirus infection, with a weaker association identified in the endemic sub-model. Also, the distribution of epidemic random intercepts was narrower compare to the endemic suggesting that reporting practices are relatively similar across regions and age-groups when it comes to outbreak-generated cases. The variation we see is likely related to the number of samples collected per outbreak. For example, a study investigating care homes outbreaks in North West points out that even though at least six samples are recommended, the median is only three [27]. Most of the reported cases of norovirus are epidemic in nature (86%). These findings strongly support the hypothesis that cases of norovirus from outbreaks in nursing homes and hospitals are more likely to appear in national statistics.

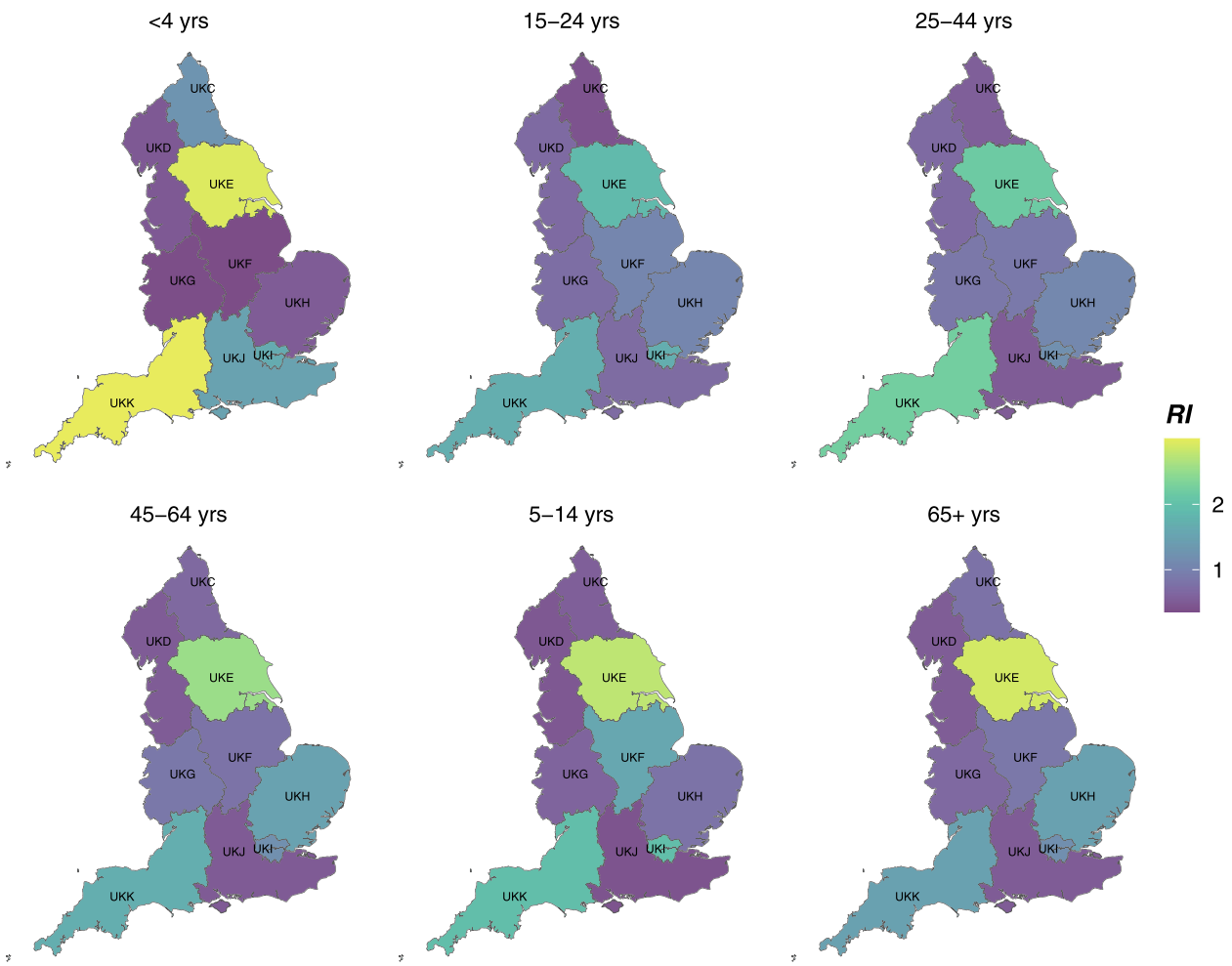

Fig. 3 Age group- and region-specific endemic random intercepts. RI < 1 suggests lower number of norovirus reports than expected. The Nomenclature of Territorial Units for Statistics (NUTS) Level 1 codes can be mapped to regional names as follows: UKC - North East, UKD – North West, UKE – Yorkshire & Humber, UKF – East Midlands, UKG – West Midlands, UKH – East of England, UKI – London, UKJ – South East, UKK – South West

What this study adds

This study adapted HHH4 analytic pipeline [11, 18, 19] to analyse norovirus reporting practices across regions and selected age groups in England. One of the benefits of this method was that it allows borrowing strength between age-groups as opposed to formulating a separate model for each group [19], some of which have low numbers of reported cases. This means that the same procedure can be followed by regional PHE units using higher spatial granularity. Additionally, we pointed out some of the biases in the otherwise stable and consistent national surveillance system, such as the tendency to see outbreak-related rather than sporadic cases in the national statistics and regional variation. Furthermore, we described the reporting patterns of norovirus, which can be of use to public health policymakers. Models facilitate focusing upon regions and subpopulations in which under-reporting may be most pronounced and have the power to highlight whether the reporting for different subpopulations within a particular region is low. Variability in reporting emphasises the importance of considering uncertainty as it has implications for decisions regarding the development of more targeted norovirus risk management (e.g. vaccine), and overall, these insights are relevant to potential norovirus forecasting efforts which are likely to follow the path of seasonal influenza, e.g. [28]. In light of this, we point out that multivariate approaches have clear benefits over separate age- or region-based models as they allow for spatial relationships. However, at the regional geographic level of granularity, the spatial effects are relatively small and so modelling every region of England on its own when age-stratification is unnecessary or unavailable could yield valid predictions as well.

Limitations

A weakness of this analysis is that the level of geographical resolution in the data is coarse, and this limits the depth of inference. Despite the inclusion of regional factors (number of hospitals, schools and nurseries) and population structure, random effects may have captured some residual spatial variation that could potentially be explained by other factors such as genotype. However, given that routine analysis did not detect any changes or shift in the main genotypes circulating at this time [29], our analysis is unlikely to have been affected by such change. Also, our approach was not able to differentiate between under-ascertainment and under-reporting. Moreover, the POLYMOD study took place in 2005/2006, and the contact patterns in the study period could be different during the studied period. Despite these limitations, we have applied the methods to consistently collected data with the best resources available.

Conclusion

Our findings contribute to the understanding of norovirus reporting patterns in England and provide a basis for future norovirus forecasting endeavours. There is inherent uncertainty in the routinely collected surveillance data, which needs to be recognised and methods of analysis adjusted accordingly. Regional differences were anticipated as attitudes towards the importance of norovirus surveillance as well as testing practices vary across reference laboratories and hospitals. Our analysis highlighted regions in which sporadic cases may be underestimated. Understanding the biases in surveillance data and sources of variation is crucial, especially as disease forecasting tools are increasingly developed and applied. In this context, multivariate approaches should be favoured over separate age- or region-based models. Besides forecasting, future research could enhance understanding of why under-reporting takes place, and so inform targeted norovirus risk management strategies.

Abbreviations

PHE: Public Health England; AIC: Akaike information criterion; BIC: Bayesian information criterion; SGSS: Second Generation Surveillance System; ONS: Office for National Statistics; RI: Random intercept; NUTS: Nomenclature of Territorial Units for Statistics

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12889-021-11317-3.

Additional file 1. Number of confirmed norovirus cases in England (2014/15–2018/19). Colours mark the partition of the data into training (2014 w27–2018 w26) and test periods (2018 w27–2019 w26).

Additional file 2. Modelling results to determine the number of seasonal waves. Comparison of fixed-effect models with one (S = 1), two (S = 2) and three (S = 3) seasonal waves in endemic component to determine the baseline model formulation.

Additional file 3. Epidemic and endemic random intercepts by region and age group with confidence intervals.

Additional file 4. Age group- and region-specific epidemic random intercepts.

Acknowledgements

Nikola Ondrikova would like to acknowledge the gracious support of this work through the EPSRC and ESRC Centre for Doctoral Training on Quantification and Management of Risk Uncertainty in Complex Systems & Environments Grant No. (EP/L015927/1). Also, without surveillance data from Public Health England, none of this work would be possible.
Helen Clough, Nigel Cunliffe and Roberto Vivancos are affiliated to the National Institute for Health Research (NIHR) Health Protection Research Unit in Gastrointestinal Infections at University of Liverpool, in partnership with Public Health England, in collaboration with University of Warwick. The views expressed are those of the author(s) and not necessarily those of the NIHR, the Department of Health and Social Care or Public Health England.
Finally, we would like to thank two antonymous reviewers for their useful comments.

Authors' contributions

HEC had an initial idea for the study that was further developed by NO under the supervision of HEC, MIG and JPH. NO performed the analysis with HEC providing advice on statistical modelling. NO wrote the manuscript and prepared the supporting materials. HEC, NAC, MIG, RV, and JPH were involved in discussing and interpreting the results and editing and reviewing the manuscript. All authors have given final approval of the version of the manuscript submitted for publication.

Funding

EPSRC and ESRC Centre for Doctoral Training on Quantification and Management of Risk Uncertainty in Complex Systems & Environments Grant No. (EP/L015927/1). The funder had no role in the design of the study, data analysis, interpretation of the results and in writing the manuscript.

Availability of data and materials

The dataset analysed during the current study is available from the corresponding author on reasonable request. A synthetic version of the dataset and R codes supporting the conclusions of this article are available in the GitHub repository, [https://doi.org/10.5281/zenodo.4464124].

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no conflict of interest.

Author details

[1]Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Liverpool, UK. [2]Institute for Risk & Uncertainty, University of Liverpool, Liverpool, UK. [3]NIHR Health Protection Unit in Gastrointestinal Infections, University of Liverpool, Liverpool, UK. [4]Public Health England, Liverpool, UK. [5]Centre for Vaccine Innovation and Access, PATH, Geneva, Switzerland. [6]NIHR Health Protection Unit in Emerging and Zoonotic Infections, University of Liverpool, Liverpool, UK.

Received: 25 January 2021 Accepted: 18 June 2021
Published online: 28 June 2021

References

1. Belliot G, Lopman BA, Ambert-Balay K, Pothier P. The burden of norovirus gastroenteritis: an important foodborne and healthcare-related infection. Clin Microbiol Infect. 2014;20(8):724–30. https://doi.org/10.1111/1469-0691.12722.

2. Tam CC, Rodrigues LC, Viviani L, Dodds JP, Evans MR, Hunter PR, et al. Longitudinal study of infectious intestinal disease in the UK (IID2 study): incidence in the community and presenting to general practice. Gut. 2012; 61(1):69–77. https://doi.org/10.1136/gut.2011.238386.
3. Estes MK, Prasad BV, Atmar RL. Noroviruses everywhere: has something changed? Curr Opin Infect Dis. 2006;19(5):467–74. https://doi.org/10.1097/01.QCO.0000244053.69253.3D.
4. Harris JP, Edmunds WJ, Pebody RG, Brown DW, Lopman BA. Deaths from norovirus among the elderly, England and Wales - volume 14, number 10—October 2008 - emerging infectious diseases journal - CDC. Emerg Infect Dis. 2008;14(10):1546–52. https://doi.org/10.3201/EID1410.080188.
5. HEALTH PROTECTION (NOTIFICATION) REGULATIONS 2010. https://www.legislation.gov.uk/uksi/2010/659/made.
6. Public Health England. Second Generation Surveillance System (SGSS). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/926838/PHE_Laboratory_reporting_guidelines_October-2020-v3.pdf.
7. Tam CC, O'Brien SJ. Economic cost of campylobacter, Norovirus and Rotavirus Disease in the United Kingdom. PLoS One. 2016;11(2):e0138526. https://doi.org/10.1371/JOURNAL.PONE.0138526.
8. Harris JP. Norovirus Surveillance: An Epidemiological Perspective. J Infect Dis. 2016;213(suppl_1):S8–11. https://doi.org/10.1093/INFDIS/JIV452.
9. Harris JP, Iturriza-Gomara M, O'Brien SJ. Estimating disability-adjusted life years (DALYs) in community cases of norovirus in England. Viruses. 2019; 11(2):184. https://doi.org/10.3390/v11020184.
10. Vaida F, Blanchard S. Conditional Akaike information for mixed-effects models. Biometrika; 2005.
11. Paul M, Held L. Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts. Stat Med. 2011; 30(10):1118–36. http://www.ncbi.nlm.nih.gov/pubmed/21484849. https://doi.org/10.1002/sim.4177.
12. Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc. 2007;102(477):359–78. https://doi.org/10.1198/016214506000001437.
13. Clough HE, Hardstaff J, Harris JP, O'Brien SJ. Challenges in understanding the spatio-temporal epidemiology of norovirus infection in England using routine public health surveillance data. Conference of the Royal Statistical Society. 2016.
14. Allen DJ, Adams NL, Aladin F, Harris JP, Brown DWG. Emergence of the GII-4 Norovirus Sydney2012 Strain in England, Winter 2012–2013. PLoS One. 2014; 9(2):e88978. https://doi.org/10.1371/JOURNAL.PONE.0088978.
15. Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Med. 2008;5(3):e74. https://doi.org/10.1371/journal.pmed.0050074.
16. Department for Education. Schools, pupils and their characteristics: January 2016. 2016. https://www.gov.uk/government/statistics/schools-pupils-and-their-characteristics-january-2016. Accessed 23 Jun 2020.
17. NHS Choices. Hospitals. 2015. https://data.gov.uk/dataset/f4420d1c-043a-42bc-afbc-4c0f7d3f1620/hospitals. Accessed 23 Jun 2020.
18. Held L, Höhle M, Hofmann M. A statistical framework for the analysis of multivariate infectious disease surveillance counts. 2005;5:187–99. https://doi.org/10.1191/1471082X05ST098OA.
19. Meyer S, Held L. Incorporating social contact data in spatio-temporal models for infectious disease spread. Biostatistics. 2017;18(2):338–51. https://doi.org/10.1093/biostatistics/kxw051.
20. Xia YÂ, ONÂ B, BTÂ G. Measles Metapopulation Dynamics: A Gravity Model for Epidemiological Coupling and Dynamics. 2004;164:267–81. https://doi.org/10.1086/422341.
21. R Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2020.
22. Meyer S, Held L, Höhle M. Spatio-Temporal Analysis of Epidemic Phenomena Using the R Package surveillance. J Stat Softw. 2017;77:1–55. https://doi.org/10.18637/jss.v077.i11.
23. Gibbons CL, Mangen M-JJ, Plass D, Havelaar AH, Brooke RJ, Kramarz P, et al. Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods. BMC Public Health. 2014;14:1–17. https://doi.org/10.1186/1471-2458-14-147.
24. Lopman BA, Reacher MH, Vipond IB, Sarangi J, Brown DWG. Clinical manifestation of norovirus gastroenteritis in health care settings. Clin Infect Dis. 2004;39(3):318–24. http://www.ncbi.nlm.nih.gov/pubmed/15306997. https://doi.org/10.1086/421948.
25. Lopman BA, Reacher MH, Vipond IB, Hill D, Perry C, Halladay T, et al. Epidemiology and cost of nosocomial gastroenteritis, Avon, England, 2002-2003. Emerg Infect Dis. 2004;10(10):1827–34. http://www.ncbi.nlm.nih.gov/pubmed/15504271. https://doi.org/10.3201/eid1010.030941.
26. Chamberland RR, Burnham CA, Storch GA, Jackups R, Doern CD. Prevalence and seasonal distribution of norovirus detection in stools submitted from pediatric patients for enteric pathogen testing. J Pediatric Infect Dis Soc. 2015;4(3):264-6. https://doi.org/10.1093/jpids/piu040.
27. Inns T, Wilson D, Manley P, Harris JP, O'Brien SJ, Vivancos R. What proportion of care home outbreaks are caused by norovirus? An analysis of viral causes of gastroenteritis outbreaks in care homes, north East England, 2016-2018. BMC Infect Dis. 2019;20(1):1–8. https://doi.org/10.1186/s12879-019-4726-4.
28. Lutz CS, Huynh MP, Schroeder M, Anyatonwu S, Dahlgren FS, Danyluk G, et al. Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples. BMC Public Health. 2019;19: 1–12. https://doi.org/10.1186/S12889-019-7966-8.
29. Public Health England. National norovirus and rotavirus bulletin: management information. 2020. https://www.gov.uk/government/statistical-data-sets/national-norovirus-and-rotavirus-bulletin-management-information.

Publisher's Note