Int J Public Health (2014) 59:841–849
DOI 10.1007/s00038-014-0581-7

ORIGINAL ARTICLE

Demand-based web surveillance of sexually transmitted infections in Russia

Alexander Domnich · Eva K. Arbuzova · Alessio Signori · Daniela Amicizia · Donatella Panatto · Roberto Gasparini

Received: 16 December 2013 / Revised: 27 May 2014 / Accepted: 18 June 2014 / Published online: 11 July 2014

Abstract

Objectives To investigate the possibility of using HIV- and syphilis-related web queries to predict incident diagnosis rates of sexually transmitted infections in Russia.

Methods The regional volume of HIV/syphilis queries, normalized to the total number of queries submitted to the most popular search engine, was used to predict the notification rates of HIV/syphilis in each region by applying both global non-spatial and spatial statistics.

Results Nationwide, both search volumes and regional HIV/syphilis diagnosis rates were positively spatially autocorrelated, indicating a clustered pattern of spatial distribution. A high positive correlation between notification rates and search volume was observed. Compared with linear models, spatially explicit geographically weighted models adjusted for broadband Internet diffusion proved superior in predicting the regional level of the HIV/syphilis epidemic on the basis of their search volume.

Conclusions Timeliness, easy availability, low cost, and transparency make HIV- and syphilis-related web queries a promising addition to traditional methods of disease surveillance in Russia. Geographically weighted regression provides useful insights, as it is able to capture the spatial heterogeneity of the relationship between search volume and disease incidence.

Keywords Disease surveillance · HIV · Search engine · Search volume · Infodemiology

Electronic supplementary material The online version of this article (doi:10.1007/s00038-014-0581-7) contains supplementary material, which is available to authorized users.

A. Domnich (✉) · A. Signori · D. Amicizia · D. Panatto · R. Gasparini
Department of Health Sciences, University of Genoa, Genoa, Italy
e-mail: alexander.domnich@gmail.com

E. K. Arbuzova
State Budgetary Healthcare Institution “Specialized Clinical Hospital for Infectious Diseases”, Healthcare Department of Krasnodar Region, Krasnodar, Russia

Introduction

Human Immunodeficiency Virus (HIV) and other Sexually Transmitted Infections (STIs) are a major public health challenge in Russia (Moran and Jordaan 2007; Zhan et al. 2011). The first case of HIV in the former Soviet Union was documented in 1987. Since then, the number of newly diagnosed cases has increased explosively each year (Dehne et al. 1999); while less than a thousand prevalent cases were reported in 1994, people living with HIV (PLWH) now number about 700,000. The geographical distribution of HIV has also changed over the years. In the late 1990s, one of the most seriously affected regions was Kaliningrad oblast in the most western part of Russia, while today’s highly affected regions are situated mainly in the Ural, Siberian, and Volga federal districts (Federal AIDS centre 2013). The heterogeneous regional prevalence of HIV in some way reflects the circumstances of post-Soviet transition, and has proved to be associated with gross regional product, urbanization, population mobility, social dislocation, and drug use (Moran and Jordaan 2007).

Notification rates (NRs) of some STIs, unlike HIV infection, have diminished since the late 1990s. For instance, the NR of syphilis was 277.3 per 100,000 inhabitants in 1997, while in 2012 it was just 33.0 (Russian Ministry of Health and Social Development 2013; Riedner et al. 2000). However, although the NR of syphilis has declined significantly, it remains among the highest in Europe (Centralized Information System for Infectious Diseases 2013). There is also a disproportionately large difference (up to 45-fold) in syphilis NRs among single Russian regions. Indeed, in 2012 the lowest NR was recorded in the Republic of Dagestan in the Northern Caucasus (4.4 per 100,000 inhabitants), while that of the Republic of Tuva in southern Siberia reached nearly 200 per 100,000 inhabitants (Russian Ministry of Health and Social Development 2013).

Analyzing regional patterns and understanding discrepancies in the epidemic features of HIV and other STIs contribute to the planning of surveillance of these infections (Webster 2005). Geographical stratification for infectious disease surveillance is of particular relevance in Russia, the world's largest country, which consists of 83 Federal Subjects (henceforth referred to as regions) grouped into eight federal districts, namely: Central, Northwestern, Southern, North Caucasian, Volga, Ural, Siberian, and Far Eastern.

HIV surveillance in Russia is currently based on the mandatory testing of blood donors and the registration of HIV-infected persons and AIDS cases. The weakness of this system is its scant linkage with STI, drug use and tuberculosis surveillance, conjointly with inadequate socio-epidemiological monitoring; its advantages include a well-organized reporting mechanism from the periphery to the center (World Health Organization 2011a), which yelds official statistics (Russian Ministry of Health and Social Development 2013; Federal AIDS centre 2013). However, reporting official national or regional statistics on disease occurrence usually requires some time. In view of the growing concerns about pandemics, emerging infectious diseases, and bioterrorism, there is a great need for a sensitive and robust real-time surveillance system (Carneiro and Mylonakis 2009).

Nowadays, the Internet is a very important tool for laypeople, as more than half of Internet users, at least in industrialized societies, search the Web for health-related information (Atkinson et al. 2009). It has been shown that common search engines, which are usually the starting-points of searches for specific topics, are useful tools for mapping and predicting the spatio-temporal incidence of several infectious diseases. Indeed, queries submitted to common search engines by laypeople have proved to be highly correlated with official surveillance data on influenza and influenza-like illness, chickenpox, Lyme disease, tuberculosis, dengue, and HIV (Bernardo et al. 2013). Such novel methods have already been applied to the Russian setting (Zheluk et al. 2012, 2013).

The absolute number of web users in Russia has increased very rapidly, becoming the highest in Europe (Zheluk et al. 2012); however, Internet Penetration (IP) is still only about 50 % nationwide. A particular feature of Internet use in Russia is the enormous difference in Internet access among single regions. For instance, in the two federal cities of Moscow and Saint-Petersburg, web penetration is above 70 %, while in some Republics of the Northern Caucasus it is about 5 % (Russian Federation Federal State Statistics Service 2012). This picture illustrates the phenomenon of the digital divide, i.e. the inequality in access, use and knowledge of Information and Communication Technology (ICT) among different groups of people. Diverse Internet access has been described among people of different socioeconomic status, age-class, gender and geographical locations (Chen and Wellman 2004). The tumultuous growth of ICT has led to a gradual shift of the public focus on the digital divide towards the concept of inequalities in Broadband Internet Penetration (BIP), which allows users to send and receive enormous quantities of variegated data (Prieger 2007).

As research has increasingly suggested the usefulness of analyzing the web search behavior regarding notifiable diseases, we investigated whether HIV- and syphilis-related queries submitted to a search engine could provide additional near real-time information on the regional incidence of HIV/syphilis, taking into account the particular features of ICT development in contemporary Russia.

Methods

Data source

Annual numbers of newly diagnosed HIV and syphilis (all forms) infections (NR) as of 2011 and 2012 in each region were gathered from data supplied by the Russian Ministry of Health and Social Development (2013).

Regional levels of IP were taken from the Russian Federation Federal State Statistics Service (2012) data gathered from a survey on household budgets, while regional fixed BIP (web access with a download data transfer rate of 256 kbit/s or higher) was determined on the basis of data supplied by the Russian Ministry of Communications and Mass Media (2012), as of 2011 and 2012.

Analysis of search trends was performed through the Yandex© search engine (www.yandex.ru), which is the most used site nationwide, handling 60 % of all search traffic in Russia (Zheluk et al. 2012). The site provides a publically available search pattern tool, Yandex Wordstat©, which assesses the search volume of a query (i.e. a request for information) in a desired time interval within 2 years and can display detailed regional search patterns (Yandex Wordstat© 2013). While other similar tools may not show queries if their search volume is insufficient (Zheluk et al. 2013; Zhou and Shen 2010), Yandex

Wordstat© is able to display every single query. Moreover, it provides summary statistics of all queries containing a given term. For example, submission of the term "HIV" will display all phrases containing this term, such as "HIV symptoms", "risk of HIV", "HIV test", etc. (Yandex Wordstat© 2013). Yandex Wordstat© has been validated as a tool for measuring online behavior in Russia (Zheluk et al. 2012, 2013).

Regional search volumes of the terms "HIV" and "syphilis" (in Russian, *ВИЧ* and *сифилис*, respectively) in each single region were retrieved and analyzed by two native Russian speakers (AD and EA). The search volume of the term "AIDS" (in Russian, *СПИД*) was not considered for two reasons. First, previous research (Zheluk et al. 2013) has found a higher correlation coefficient between regional HIV prevalence and "HIV" searches than with "AIDS" searches. This has been explained by the misclassification, for purely linguistic reasons, of searches unrelated to the disease (the English word "speed" is often transliterated into Russian as "*спид*"—identical to the acronym that means "AIDS"—which produces numerous queries, such as "*Нид фор спид*" ["Need for speed"—a popular computer game] or "*Спид тест*" ["Speed test"—online test to measure Internet speed]). Second, the World AIDS Day on 1st December attracts worldwide attention, causing the number of searches unconnected with the occurrence of the disease to soar around this date (Zhou and Shen 2010). By contrast, we did not find a significant increase in "HIV" searches in November/December of the two consecutive years.

To offset the possible effect of increases in search volume and make regional search volumes comparable, queries were normalized (Polgreen et al. 2008) to obtain the Demand Prevalence Rate (DPR) (Eysenbach 2009); this is the ratio between the absolute number of "HIV"/"syphilis" queries and the total number of queries submitted to the search engine from this region in a given period of time, and was expressed per 100,000 queries. At the time of extraction, data on "HIV"/"syphilis" searches were available from July 2011 to July 2013. Therefore, the DPR of "HIV"/"syphilis" in 2011 considered the absolute number of "HIV"/"syphilis" queries in relation to the total number of queries submitted from July to December 2011; similarly, the 2012 DPR of "HIV"/"syphilis" considered those submitted from July to December 2012. All statistics provided by Yandex Wordstat© are presented in a rigorously anonymous aggregated form; no information on the user's identity or Internet protocol address is provided.

Statistical analysis

Spearman's correlation coefficient was used to measure the correlation between the DPRs and NRs. Bootstrapping of 9,999 rounds was used to compute 95 % confidence intervals (CIs) of Spearman correlation coefficients. Global Moran's *I* was used to measure overall clustering of DPRs and NRs. Positive *I* values indicate strong geographic patterns of spatial clustering, negative values show clustering of dissimilar values, while values close to zero indicate complete spatial randomness. The interpretation of Moran's *I* is similar to that of Pearson's correlation, as it ranges from −1 to 1. To measure local spatial association, local Moran's *I* cluster maps were subsequently created to provide information on the location of clusters and types of spatial association. Specifically, positive spatial autocorrelation signifies the presence of spatial clusters; positive association occurs when a region with high NR or DPR is surrounded by other regions with high values (high–high, "hot spots") or when a region with low NR or DPR borders on regions which also display low values (low–low, "cold spots"). By contrast, negative spatial autocorrelation indicates spatial outliers, i.e. when a region with low NR or DPR is surrounded by regions with high values (low–high) or vice versa (high–low). Owing to the presence of an exclave (region of Kaliningrad) and Sakhalin island, and the fact that eastern regions are much larger than western ones, for Moran's *I* statistics *k* nearest neighbor spatial weights matrix was employed (Anselin 2002); a *k* value of 4 was used as it has been shown that *k* values of 4–6 are optimal (Duncan et al. 2012). The significance of Moran's *I* was checked by means of a pseudo *p* value calculated through a Monte Carlo randomization test with 9,999 permutations.

To discover whether the regional search volume could predict the NR of HIV/syphilis, we used linear models, unadjusted and adjusted for an indicator of Internet diffusion. During the analytical phase, the Shapiro–Wilk test was used to test normality of the dependent variable. As distributions of NRs were slightly skewed, a natural logarithm transformation was used to produce near-normal distributions; the Shapiro–Wilk test was then repeated to confirm data normality. In sum, linear models were set to $\ln(\mathrm{NR}) = b_0 + b_{\mathrm{DPR}} \cdot x_{\mathrm{DPR}} + \varepsilon$ and $\ln(\mathrm{NR}) = b_0 + b_{\mathrm{DPR}} \cdot x_{\mathrm{DPR}} + b_{\mathrm{INTERNET}} \cdot x_{\mathrm{INTERNET}} + \varepsilon$, where ln(NR) is the log-transformed dependent variable of HIV/syphilis NR, b_0 is intercept, $b_{\mathrm{DPR},}$ and b_{INTERNET} are regression coefficients of "HIV"/"syphilis" DPR and indicator of Internet diffusion (IP or BIP), respectively, x_{DPR} and x_{INTERNET} are independent variables of "HIV"/"syphilis" DPR and the indicator of Internet diffusion, respectively, and ε is random error.

The application of linear models to spatial data may raise some important concerns, as such models assume spatial immobility, i.e. any association among variables is constant over space; this may potentially veil important local variations in the association between dependent and independent variables. Geographically Weighted Regression (GWR) (Fotheringham et al. 2002) has been specifically designed to overcome spatial heterogeneity; a modification of traditional regression, it performs, by moving a spatial kernel across the study area, many local regressions, each of which is influenced by the neighboring data. As the spatial distribution of single regions was inhomogeneous in the present study, an adaptive bisquare spatial kernel with a bandwidth determined by minimizing the corrected Akaike Information Criterion (AICc) was used. The GWR models used the same dataset as the linear models and were set to $\ln(\text{NR}) = b_0(u_\text{i},v_\text{i}) + b_\text{DPR}(u_\text{i},v_\text{i})\cdot x_\text{iDPR} + \varepsilon_\text{i}$ and $\ln(\text{NR}) = b_0(u_\text{i},v_\text{i}) + b_\text{DPR}(u_\text{i},v_\text{i})\cdot x_\text{iDPR} + b_\text{INTERNET}(u_\text{i},v_\text{i})\cdot x_\text{iINTERNET} + \varepsilon_\text{i}$, where i refers to a location in which the GWR model is calibrated and (u_i,v_i) refers to the spatial coordinates of the location i.

To compare the performance of the models AICc minimization criterion was used. Furthermore, the spatial fit of the models was examined using global Moran's I of the models' residuals. Collinearity of all models adjusted for Internet diffusion was checked; this is an important issue, especially in GWR modeling, because its effects may be more pronounced when smaller samples are used to calibrate single local regressions, as some locations may exhibit collinearity while others do not. Collinearity was measured by means of the Variance Inflation Factor (VIF); local VIFs may be used for collinearity diagnostics in GWR models (Brunsdon et al. 2012). VIF values >4 were considered critical (Miles and Shevlin 2001).

The best-performing models were subsequently externally validated by testing parameter estimates obtained from 2011 data against a dataset from 2012 and assessing the Spearman correlation coefficients between predicted and reported NRs. Prediction accuracy was also measured by means of percentage mean absolute error (%MAE) and as the fraction of predicted values within 20 % of the reported values.

The level of significance adopted was two-tailed $P < 0.05$. All data were analyzed by means of the *R* stats package, version 3.0.1 (R Development Core Team 2012).

Results

Nationwide, the DPRs of "HIV" and "syphilis" increased from 7.05 per 100,000 queries in 2011 to 7.57 in 2012 and from 2.76 per 100,000 queries in 2011 to 3.11 in 2012, respectively.

The global autocorrelation analysis of annual regional DPRs of "HIV" revealed significant ($P < 0.001$) Moran's I values, indicating a clustered pattern of spatial distribution (0.34 in 2011 and 0.24 in 2012). A similarly significant ($P < 0.001$) spatial pattern was observed among NRs of HIV (0.45 in 2011 and 0.40 in 2012). Although highly significant, the overall tendency was moderate. Likewise, a clustered pattern of spatial distribution was documented both among DPRs of "syphilis" ($I = 0.59$, $P < 0.001$ in 2011 and $I = 0.56$, $P < 0.001$ in 2012) and among NRs of syphilis ($I = 0.60$, $P < 0.001$ in 2011 and $I = 0.49$, $P < 0.001$ in 2012).

Fig. 1 **a** Local Moran's *I* cluster map of natural logs of HIV ▶ notification rates in Russia in 2011 (**a**) and 2012 (**b**) and "HIV" demand prevalence rates in 2011 (**c**) and 2012 (**d**) *DPR* demand prevalence rate, *NR* notification rate. **b** Local Moran's *I* cluster map of natural logs of syphilis notification rates in Russia in 2011 (**a**) and 2012 (**b**) and "Syphilis" demand prevalence rates in 2011 (**c**) and 2012 (**d**). *DPR* demand prevalence rate, *NR* notification rate

Figure 1a, b shows regions with significant local Moran statistics for both NRs and DPRs; it is noteworthy that local Moran's I_i of NRs and DPRs are very similar. No statistically significant negative spatial autocorrelation (outliers) was detected. Significantly high DPR of "HIV" and HIV NR were observed in most regions of the Ural Federal district and in some Volga and Siberian regions, while high DPR of "syphilis" and syphilis NR were concentrated in the Siberian Federal district and in some Far Eastern regions. Low DPRs and NRs of both HIV and syphilis were mainly located in the North Caucasus and adjacent regions.

Overall, a high positive Spearman correlation between NR of HIV and the corresponding DPR was observed [0.73 (95 % CI: 0.59–0.84) in 2011, $P < 0.001$ and 0.73 (95 % CI: 0.57–0.84) in 2012, $P < 0.001$]. The Spearman correlation between the NR of syphilis and the DPR of "syphilis" was even higher in both years [0.82 (95 % CI: 0.69–0.90) in 2011, $P < 0.001$ and 0.79 (95 % CI: 0.70–0.86) in 2012, $P < 0.001$].

On simple linear regression, the independent variables "HIV" and "syphilis" DPRs were positively associated with the corresponding NRs. When linear models were adjusted for the indicators of regional Internet diffusion, estimates of "HIV"/"syphilis" DPR did not change significantly. On the other hand, as indicated by both adjR^2 and AICc, models predicting HIV NR adjusted for Internet diffusion performed better, while linear models predicting syphilis NR performed almost identically. The additional analysis of spatial autocorrelation of residuals showed a moderate positive spatial autocorrelation of all linear models, suggesting a violation of the standard assumption of homoscedasticity. Adjusted models were not subject to bias derived from collinearity, as indicated by the VIF (Table 1).

In comparison with linear models, GWR had greater explanatory power. The GWR results revealed considerable spatial heterogeneity in parameter estimates from different regional models. As shown by the AICc, GWR models predicting both HIV and syphilis NRs and adjusted

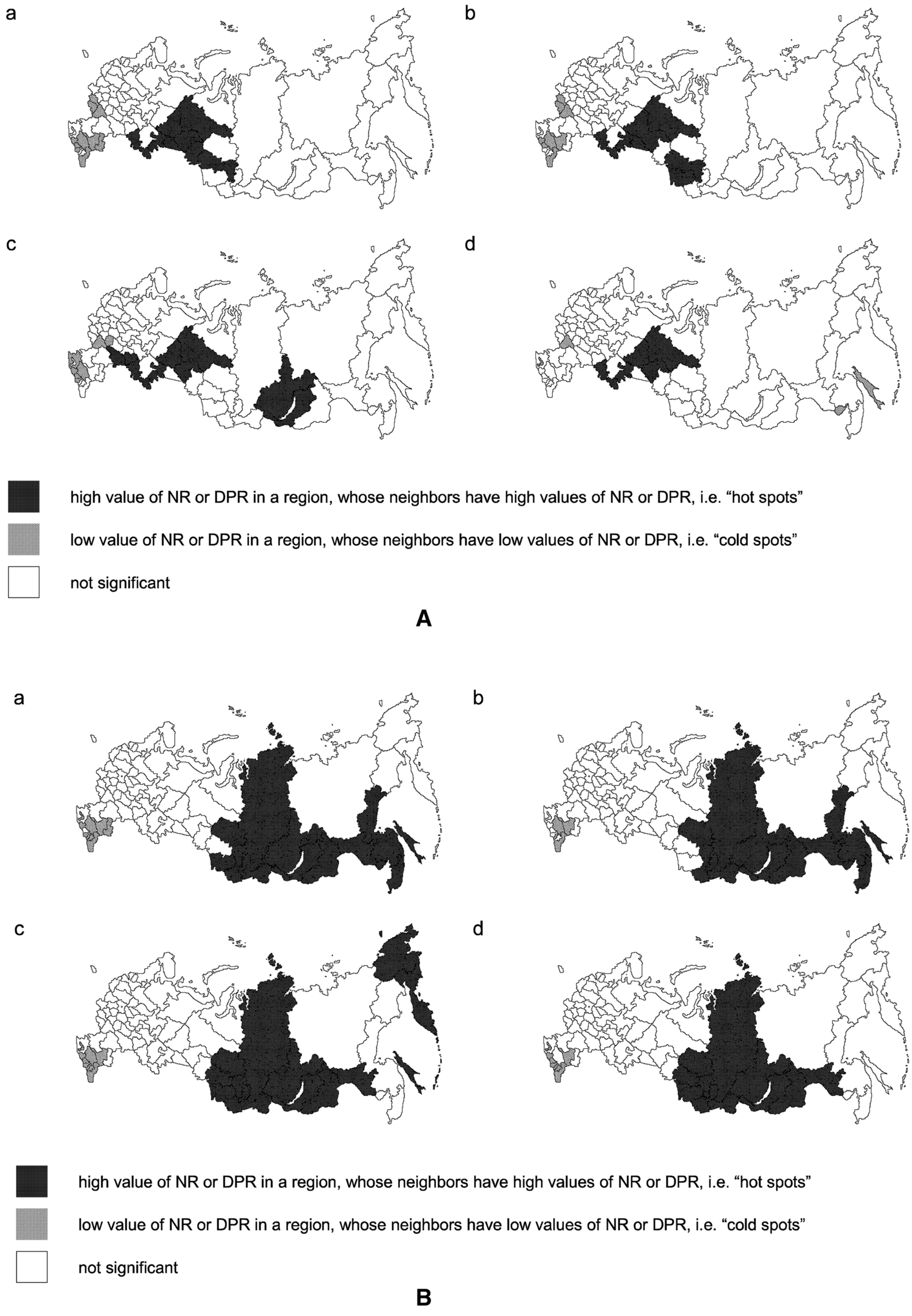
a
b
c
d
high value of NR or DPR in a region, whose neighbors have high values of NR or DPR, i.e. “hot spots”
low value of NR or DPR in a region, whose neighbors have low values of NR or DPR, i.e. “cold spots”
not significant
A
a
b
c
d
high value of NR or DPR in a region, whose neighbors have high values of NR or DPR, i.e. “hot spots”
low value of NR or DPR in a region, whose neighbors have low values of NR or DPR, i.e. “cold spots”
not significant
B

Table 1 Linear models to predict regional HIV and syphilis notification rates in Russia in 2011

Disease	Model	Parameter							
		Intercept, (SE) [P]	DPR, estimate (SE) [P]	IP, estimate (SE) [P]	BIP, estimate (SE) [P]	adjR^2	AICc	VIF	Moran's *I* of residuals (*P*)
HIV	1[a]	1.15 (0.24) [<0.001]	0.31 (0.04) [<0.001]	–	–	0.49	168.69	–	0.21 (0.002)
	2[b]	0.32 (0.29) [0.27]	0.29 (0.03) [<0.001]	2.10 (0.50) [<0.001]	–	0.58	154.46	1.04	0.16 (0.013)
	3[c]	0.65 (0.24) [0.009]	0.29 (0.03) [<0.001]	–	6.14 (1.37) [<0.001]	0.59	152.39	1.03	0.11 (0.047)
Syphilis	1[d]	2.10 (0.15) [<0.001]	0.49 (0.05) [<0.001]	–	–	0.53	96.89	–	0.27 (<0.001)
	2[e]	2.12 (0.21) [<0.001]	0.50 (0.05) [<0.001]	−0.05 (0.36) [0.88]	–	0.53	99.07	1.03	0.27 (<0.001)
	3[f]	2.05 (0.19) [<0.001]	0.50 (0.05) [<0.001]	–	0.37 (0.98) [0.71]	0.53	98.95	1.01	0.27 (<0.001)

AICc corrected Akaike information criterion, *BIP* broadband Internet penetration, *DPR* demand prevalence rate, *HIV* human immunodeficiency virus, *IP* internet penetration, *NR* notification rate, *SE* standard error, *VIF* variance inflation factor

[a] $\ln(\text{NR of HIV in 2011}) = b_0 + b_{(\text{DPR of ``HIV'' in 2011})} \cdot x_{(\text{DPR of ``HIV'' in 2011})} + \varepsilon$

[b] $\ln(\text{NR of HIV in 2011}) = b_0 + b_{(\text{DPR of ``HIV'' in 2011})} \cdot x_{(\text{DPR of ``HIV'' in 2011})} + b_{(\text{IP in 2011})} \cdot x_{(\text{IP in 2011})} + \varepsilon$

[c] $\ln(\text{NR of HIV in 2011}) = b_0 + b_{(\text{DPR of ``HIV'' in 2011})} \cdot x_{(\text{DPR of ``HIV'' in 2011})} + b_{(\text{BIP in 2011})} \cdot x_{(\text{BIP in 2011})} + \varepsilon$

[d] $\ln(\text{NR of syphilis in 2011}) = b_0 + b_{(\text{DPR of ``syphilis'' in 2011})} \cdot x_{(\text{DPR of ``syphilis'' in 2011})} + \varepsilon$

[e] $\ln(\text{NR of syphilis in 2011}) = b_0 + b_{(\text{DPR of ``syphilis'' in 2011})} \cdot x_{(\text{DPR of ``syphilis'' in 2011})} + b_{(\text{IP in 2011})} \cdot x_{(\text{IP in 2011})} + \varepsilon$

[f] $\ln(\text{NR of syphilis in 2011}) = b_0 + b_{(\text{DPR of ``syphilis'' in 2011})} \cdot x_{(\text{DPR of ``syphilis'' in 2011})} + b_{(\text{BIP in 2011})} \cdot x_{(\text{BIP in 2011})} + \varepsilon$

for BIP performed best. The residuals of all models predicting HIV NR outcome showed no spatial autocorrelation. Models predicting syphilis NR were slightly spatially autocorrelated; however, in comparison with linear models, the spatial autocorrelation of GWR residuals diminished about two fold. No signs of collinearity of the adjusted GWR models were detected (Table 2, Online Resource).

When parameter estimates based on the results of the 2011 GWR model predicting HIV NR adjusted for BIP were applied to 2012 data, a high positive correlation between predicted and reported HIV NRs was observed. More than a quarter of predicted NRs were within 20 % of reported values. Analogous results were documented for the 2011 GWR model predicting syphilis NR and adjusted for BIP (Table 3).

Discussion

Our results demonstrate that HIV- and syphilis-related queries submitted to a search engine reflect regional HIV and syphilis epidemiology; they are therefore a potentially valuable additional tool for disease surveillance. Apart from the fact that Yandex© is the most used search engine in Russia, Zheluk et al. (2012) have noted that Yandex Wordstat© is able to provide more detailed and precise subnational data on search patterns than Google Trends©. According to the Federal AIDS centre (2013) HIV epidemiological investigation and reporting in some Russian regions are insufficient in terms of completeness and timeliness; this undoubtedly affects down-to-earth appraisal of the epidemic process at both regional and national levels. Although web query-based surveillance is unlikely to replace traditional STI surveillance, it can be viewed as an opportunity to overcome at least some limitations of this latter. However, the peculiarities of STI epidemics in Russia (Moran and Jordaan 2007; Regushevskaya et al. 2010) and of the Russian language Internet require the development of a specific country-oriented online surveillance system.

To date, few studies have examined the spatial association between search volumes and the incidence of notifiable diseases. In the present study, we found a high correlation between the search term "HIV" and HIV NR, which is consistent with previous research (Jena et al. 2013; Zheluk et al. 2013). In particular, the correlation between "HIV" and incident HIV diagnoses in our study was comparable to that found by Jena et al. (2013) in a correlation study between "HIV" queries submitted to Google© and HIV incidence in single American states. Another recently published paper (Zheluk et al. 2013) found a high Spearman correlation coefficient (0.88) between "HIV" queries submitted to Yandex© in 2011 and HIV prevalent diagnosis rates in Russian regions. However, correlation coefficients in our study were lower, probably for the following reasons. First, different denominators were used to express search volume. In their work, Zheluk et al. (2013) used regional per-capita HIV-related search volume, while in the present study search volume was expressed as a proportion of "HIV" queries to the total number of queries submitted from a region. We chose this denominator to account for a 1 year increase in

Table 2 Geographically weighted regressions to predict regional HIV and syphilis notification rates in Russia in 2011

Disease	Model	Parameter							
		Intercept (range)	DPR, estimate (range)	IP, estimate (range)	BIP, estimate (range)	adjR^2	AICc	Local VIF (range)	Moran's *I* of residuals (*P*)
HIV	1[a]	0.87–1.90	0.21–0.38	–	–	0.58	158.46	–	0.03 (0.28)
	2[b]	−0.05–1.15	0.27–0.33	0.85–2.96	–	0.60	152.76	1.00–1.10	0.07 (0.13)
	3[c]	−0.01–1.58	0.24–0.35	–	0.06–8.29	0.64	146.71	1.00–1.24	0.02 (0.29)
Syphilis	1[d]	0.89–3.41	0.15–0.95	–	–	0.68	71.66	–	0.12 (0.033)
	2[e]	1.61–4.46	0.22–0.71	−2.52–0.26	–	0.66	71.96	1.00–1.46	0.15 (0.016)
	3[f]	1.56–4.45	0.09–0.70	–	−7.49–0.60	0.67	71.22	1.01–1.48	0.13 (0.030)

AICc corrected Akaike information criterion, *BIP* broadband internet penetration, *DPR* demand prevalence rate, *HIV* human immunodeficiency virus, *IP* internet penetration, *NR* notification rate, *VIF* variance inflation factor

[a] $\ln(\text{NR of HIV in 2011}) = b_0(u_i,v_i) + b_{(\text{DPR of “HIV” in 2011})}(u_i,v_i)\cdot x_{i(\text{DPR of “HIV” in 2011})} + \varepsilon_i$

[b] $\ln(\text{NR of HIV in 2011}) = b_0(u_i,v_i) + b_{(\text{DPR of “HIV” in 2011})}(u_i,v_i)\cdot x_{i(\text{DPR of “HIV” in 2011})} + b_{(\text{IP in 2011})}(\mathrm{u_i},\mathrm{v_i})\cdot x_{i(\text{IP in 2011})} + \varepsilon_i$

[c] $\ln(\text{NR of HIV in 2011}) = b_0(u_i,v_i) + b_{(\text{DPR of “HIV” in 2011})}(u_i,v_i)\cdot x_{i(\text{DPR of “HIV” in 2011})} + b_{(\text{BIP in 2011})}(u_i,v_i)\cdot x_{i(\text{BIP in 2011})} + \varepsilon_i$

[d] $\ln(\text{NR of syphilis in 2011}) = b_0(u_i,v_i) + b_{(\text{DPR of “syphilis” in 2011})}(u_i,v_i)\cdot x_{i(\text{DPR of “syphilis” in 2011})} + \varepsilon_i$

[e] $\ln(\text{NR of syphilis in 2011}) = b_0(u_i,v_i) + b_{(\text{DPR of “syphilis” in 2011})}(u_i,v_i)\cdot x_{i(\text{DPR of “syphilis” in 2011})} + b_{(\text{IP in 2011})}(u_i,v_i)\cdot x_{i(\text{IP in 2011})} + \varepsilon_i$

[f] $\ln(\text{NR of syphilis in 2011}) = b_0(u_i,v_i) + b_{(\text{DPR of “syphilis” in 2011})}(u_i,v_i)\cdot x_{i(\text{DPR of “syphilis” in 2011})} + b_{(\text{BIP in 2011})}(u_i,v_i)\cdot x_{i(\text{BIP in 2011})} + \varepsilon_i$

Table 3 Temporal external validation of regression estimates obtained from the 2011 geographically weighted models to predict regional HIV and syphilis notification rates in Russia and applied to 2012 data

Outcome	Correlation between predicted and reported NRs, *ρ* (95 % CI) [*P*]	%MAE	Predicted values within 20 % of reported values, *N*/total (%)
NRs of HIV in 2012[a]	0.84 (0.74–0.89) [<0.001]	53.64	21/83 (25.30)
NRs of syphilis in 2012[b]	0.76 (0.64–0.84) [<0.001]	59.50	24/83 (28.92)

BIP broadband internet penetration, *CI* confidence interval, *DPR* demand prevalence rate, *HIV* human immunodeficiency virus, *MAE* mean absolute error, *NR* notification rate

[a] $\text{NR of HIV in 2012} = \exp[b_0(u_i,v_i) \text{ in 2011} + b_{(\text{DPR of “HIV” in 2011})}(u_i,v_i)\cdot x_{i(\text{DPR of “HIV” in 2012})} + b_{(\text{BIP in 2011})}(u_i,v_i)\cdot x_{i(\text{BIP in 2012})}]$

[b] $\text{NR of syphilis in 2012} = \exp[b_0(u_i,v_i) \text{ in 2011} + b_{(\text{DPR of “syphilis” in 2011})}(u_i,v_i)\cdot x_{i(\text{DPR of “syphilis” in 2012})} + b_{(\text{BIP in 2011})}(u_i,v_i)\cdot x_{i(\text{BIP in 2012})}]$

total search volume, to make regional “HIV” searches comparable and to align our methodology and results to the growing amount of research that uses normalized search volume data from Google trends©. Second, the outcome of the present study was the incident diagnosis rate of HIV, while the study by Zheluk et al. (2013) used the prevalence rate as the study outcome. An enormous increase in the use of ICT in Russia over the last few years may have motivated people infected years ago to try new forms of communication, which could produce a higher correlation between search volume and prevalent HIV diagnoses. Indeed, Reeves (2000, 2001) has noted that making social connections is the second most frequent use of Internet (after seeking HIV/AIDS-related information) among people coping with HIV/AIDS.

With regard to syphilis, the nationwide DPR of “syphilis” was approximately 13 % higher in 2012 than in 2011. Similar observations have been documented in Germany, where the normalized search volume of the term “syphilis” has increased since mid-2010; almost in parallel with Google© searches for “syphilis”, an increase in syphilis NRs has also been observed (Noll-Hussong 2012). By contrast, the NRs of syphilis in Russia have declined: from 37.6 in 2011 to 33.0 per 100,000 inhabitants in 2012 (Russian Ministry of Health and Social Development 2013). However, this decline, which started in 1997, should be interpreted with caution. Indeed, patients wishing to avoid the stigma associated with the public healthcare sector are increasingly using private laboratories and anonymous outpatient clinics, which may not report STI cases to the authorities. Moreover, antibiotics have become available over the counter for self-treatment (Koniuchova et al. 2013; Regushevskaya et al. 2010; Riedner et al. 2000). Therefore, some degree of underestimation is highly likely (Netesov and Conrad 2001). In sum, real-time online surveillance offers several advantages over traditional reporting: it yields estimates much more rapidly, is easily available, fully automatic, easy to adapt to different needs, and economical in terms of both implementation and maintenance. It therefore constitutes a useful supplementary tool for health policy planning (Hulth and Rydevik 2011; Reis and Brownstein 2010; Zhou and Shen 2010).

We found that adjustment for an indicator of Internet access improved the performance of models predicting HIV NRs. We believe that the level of ICT diffusion should be considered not only in studies on spatial correlation between official surveillance statistics and data from search engines, but also in studies on temporal trends, given the continuous growth of Internet use, especially in developing countries. Moreover, the GWR models adjusted for BIP performed slightly better than the GWR models adjusted for IP; the reasons for this are probably multiple. For instance, people might make greater use of the high-speed Internet for STI-related purposes, or the local digital divide in terms of BIP may be more pronounced than that of total IP. This finding is in line with previous research: Atkinson et al. (2009) documented that people with cable, satellite or DSL (digital subscriber line) connections were significantly more likely to look for health-related topics than those with a dial-up connection. Regional socioeconomic inequalities within Russia may also play a role; indeed, in their principal component analysis of HIV prevalence and "HIV" search volume, Zheluk et al. (2013) found a strong relationship between search volumes and broadband access prices in different macro-regions. Moreover, in Russia in 2011/12, about 80 % of newly diagnosed HIV infections were registered among 20–40-year-olds (Federal AIDS centre 2013) and the age group of young adults has the greatest broadband access at home (Amstadter et al. 2009). Finally, it could be due to methodological considerations, as IP level is an estimate obtained from a representative sample during a survey on household budgets, while BIP is based on the exact number of subscriptions (Russian Federal State Statistics Service 2012; Russian Ministry of Communications and Mass Media 2012).

The present study points out the utility of applying spatial analysis to demand-based web surveillance of diseases. Specifically, we noted a certain spatial variability in HIV- and syphilis-related search volumes, which reflects the geographical pattern of HIV/syphilis epidemiology across the whole country, as was shown by Moran's *I* statistics; this justifies the use of spatially explicit modeling. Indeed, GWR was able to capture spatial heterogeneity, thereby providing useful insights into the direction and magnitude of the relationship between online search behavior and NRs at the regional level. This capacity to deal with geographical non-uniformity constituted the superiority of GWR over the global linear models.

This study has some limitations. First of all, although Yandex© is the most widely used search engine nationwide (Zheluk et al. 2013), its use may vary among regions; models should therefore be corrected for this variable. However, to our knowledge, there are no publically available data on the market shares of search engines in single regions. To overcome this limitation, we expressed the search volume as a proportion of the total number of queries, rather than per capita or per web user search volume. Second, as Yandex Wordstat© displays aggregated monthly data for the last 2 years, this paper is based on 2-year data; further research over a longer period is needed. Third, this study used the relatively simple entry terms "HIV" and "syphilis", as we aimed at establishing a quick and simple additional method of STI surveillance in Russia; this simplistic approach, however, may be subject to bias due to media activity. Indeed, according to Eysenbach (2009), in some circumstances data on search volumes may not take users' intentions into account (e.g. searching for symptoms by a subject without these symptoms). To increase the accuracy of estimates, analysis of both search and click data has been proposed (for instance, a person initially searches for "syphilis" by using a search engine and then clicks to a specific site on syphilis treatment). Alternatively, aggregated data on more specific search terms and their synonyms—defined as information "concept" (Eysenbach 2009)—such as "positive HIV test" could be used. However, the variety of such specific search terms and their spellings is huge. Moreover, search behavior and queries submitted to the search engine could be affected by regional, ethnic, or socioeconomic differences, as is the case of a multicultural Russia. This shortcoming may be partially mitigated, as all possible phrases containing the terms "HIV" or "syphilis" are counted.

In view of the rapid development of Internet in contemporary Russia, there is a need to fill some "key" gaps in our knowledge of the use of Internet for both generic health- and STI-related purposes among the general population, people at greater risk of STIs and PLWH. It may also be useful to investigate online search patterns related to major STI risk factors and behaviors. Such "behavioral online HIV/STIs surveillance" would be more in line with the second-generation surveillance for HIV/AIDS promoted by the World Health Organization (2011b). For example, online surveys among Men who have Sex with Men (MSM) may help to estimate the incidence of HIV and other STIs in this community. Marcus et al. (2013) found a strong correlation between online self-reported HIV diagnoses and notification rates of HIV among MSM in 38 countries, including Russia, and plausible underreporting or a high level of misclassification of MSM in other transmission categories, probably due to the stigmatization of homosexuality (Marcus et al. 2013).

Conflict of interest The authors declare that they have no conflict of interest.

References

Amstadter AB, Broman-Fulks J, Zinzow H, Ruggiero KJ, Cercone J (2009) Internet-based interventions for traumatic stress-related mental health problems: a review and suggestion for future research. Clin Psychol Rev 29:410–420

Anselin L (2002) Under the hood issues in the specification and interpretation of spatial regression models. Agric Econ 27:247–267

Atkinson NL, Saperstein SL, Pleis J (2009) Using the internet for health-related activities: findings from a national probability sample. J Med Internet Res 11:e4

Bernardo TM, Rajic A, Young I, Robiadek K, Pham MT, Funk JA (2013) Scoping review on search queries and social media for disease surveillance: a chronology of innovation. J Med Internet Res 15:e147

Brunsdon C, Charlton M, Harris P (2012) Living with collinearity in local regression models. In: Proceedings of the 10th international symposium on spatial accuracy assessment in natural resources and environmental sciences. http://www.spatial-accuracy.org/BrunsdonAccuracy2012. Accessed 29 November 2013

Carneiro HA, Mylonakis E (2009) Google trends: a web-based tool for real-time surveillance of disease outbreaks. Clin Infect Dis 49:1557–1564

Centralized Information System for Infectious Diseases (2013) Syphilis, incidence rate. http://data.euro.who.int/cisid/?TabID=326108. Accessed 29 November 2013

Chen W, Wellman B (2004) The global digital divide—within and between countries. It Soc 1:39–45

Dehne KL, Khodakevich L, Hamers FF, Schwartländer B (1999) The HIV/AIDS epidemic in eastern Europe: recent patterns and trends and their implications for policy-making. AIDS 13:741–749

Development Core Team R (2012) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Duncan DT, Castro MC, Gortmaker SL, Aldstadt J, Melly SJ, Bennett GG (2012) Racial differences in the built environment–body mass index relationship? A geospatial analysis of adolescents in urban neighborhoods. Int J Health Geogr 11:11

Eysenbach G (2009) Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. J Med Internet Res 11:e11

Federal AIDS centre. Statistics. http://www.hivrussia.ru/stat/index.shtml. 28 November 2013

Fotheringham S, Brunsdon C, Charlton M (2002) Geographically weighted regression: the analysis of spatially varying relationships. Wiley, Chichester

Hulth A, Rydevik G (2011) Web query-based surveillance in Sweden during the influenza A(H1N1)2009 pandemic, April 2009 to February 2010. Euro Surveill 16:pii: 19856

Jena AB, Karaca-Mandic P, Weaver L, Seabury SA (2013) Predicting new diagnoses of HIV infection using internet search engine data. Clin Infect Dis 56:1352–1353

Koniuchova K, Rozova M, Savicheva A, Sokolovskiy E, Domeika M, Unemo M (2013) Epidemiology of sexually transmitted infections in Tver. Sex Transm Infect 89(Suppl 1):A158

Marcus U, Hickson F, Weatherburn P, Schmidt AJ (2013) Estimating the size of the MSM populations for 38 European countries by calculating the survey-surveillance discrepancies (SSD) between self-reported new HIV diagnoses from the European MSM internet survey (EMIS) and surveillance-reported HIV diagnoses among MSM in 2009. BMC Public Health 13:919

Miles J, Shevlin M (2001) Applying regression & correlation: a guide for students and researchers. SAGE Publications Ltd, London

Moran D, Jordaan JA (2007) HIV/AIDS in Russia: determinants of regional prevalence. Int J Health Geogr 6:22

Netesov SV, Conrad JL (2001) Emerging infectious diseases in Russia, 1990-1999. Emerg Infect Dis 7:1–5

Noll-Hussong M (2012) Syphilis and the internet. Euro Surveill 17:pii:20249

Polgreen PM, Chen Y, Pennock DM, Nelson FD (2008) Using internet searches for influenza surveillance. Clin Infect Dis 47:1443–1448

Prieger JE (2007) The supply side of the digital divide: is there equal availability in the broadband Internet access market? Econ Inq 41:346–363

Reeves PM (2000) Coping in cyberspace: the impact of Internet use on the ability of HIV-positive individuals to deal with their illness. J Health Commun 5(Suppl):47–59

Reeves PM (2001) How individuals coping with HIV/AIDS use the Internet. Health Educ Res 16:709–719

Regushevskaya E, Dubikaytis T, Laanpere M, Nikula M, Kuznetsova O, Karro H, Haavio-Mannila E, Hemminki E (2010) The determinants of sexually transmitted infections among reproductive age women in St. Petersburg, Estonia and Finland. Int J Public Health 55:581–589

Reis BY, Brownstein JS (2010) Measuring the impact of health policies using Internet search patterns: the case of abortion. BMC Public Health 10:514

Riedner G, Dehne KL, Gromyko A (2000) Recent declines in reported syphilis rates in eastern Europe and central Asia: are the epidemics over? Sex Transm Infect 76:363–365

Russian Federation Federal State Statistics Service (2012) Households with a personal computer and Internet access, as a percentage of the total number of households of the correspondent federal subject in 2011. http://www.gks.ru/bgd/regl/b12_14p/IssWWW.exe/Stg/d02/20-07.htm. Accessed 28 November 2013

Russian Ministry of Communications and Mass Media (2012) Industry statistics. http://minsvyaz.ru/ru/directions/stat/stat/. Accessed 28 November 2013

Russian Ministry of Health and Social Development (2013) Statistics. http://www.rosminzdrav.ru/docs/mzsr/stat/47. Accessed 28 November 2013

Webster P (2005) Russia underestimates HIV/AIDS incidence. CMAJ 172:985

World Health Organization (WHO) (2011a) Introduction of Second-generation HIV Surveillance Guidelines in some Newly Independent States of Eastern Europe. Report on a WHO meeting. http://www.euro.who.int/__data/assets/pdf_file/0007/120301/E74470.pdf. Accessed 28 November 2013

World Health Organization (WHO) (2011b) UNAIDS. Guidelines on surveillance among populations most at risk for HIV, WHO, Geneva

Yandex Wordstat© (2013) Search terms statistics. http://wordstat.yandex.ru/. Accessed 28 November 2013

Zhan W, Krasnoselskikh TV, Niccolai LM, Golovanov S, Kozlov AP, Abdala N (2011) Concurrent sexual partnerships and sexually transmitted diseases in Russia. Sex Transm Dis 38:543–547

Zheluk A, Gillespie JA, Quinn C (2012) Searching for truth: internet search patterns as a method of investigating online responses to a Russian illicit drug policy debate. J Med Internet Res 14:e165

Zheluk A, Quinn C, Hercz D, Gillespie JA (2013) Internet search patterns of human immunodeficiency virus and the digital divide in the Russian Federation: infoveillance study. J Med Internet Res 15:e256

Zhou XC, Shen HB (2010) Notifiable infectious disease surveillance with data collected by search engine. J Zhejiang Univ-Sci C 11:241–248