Denominator Issues for Personally Generated Data in Population Health Monitoring

Rumi Chunara, PhD,1,2 Lauren E. Wisk, PhD,3,4 Elissa R. Weitzman, ScD, MSc3,4,5

INTRODUCTION

Widespread use of Internet and mobile technologies provides opportunities to gather health-related information to complement data generated through traditional healthcare and public health systems. These personally generated data (PGD) are increasingly viewed as informative of the patient experience of conditions, symptoms, treatments, and side effects.1 Behavior, sentiment, and disease patterns can be discerned from mining unstructured PGD in text, image, or metadata form, and from analyzing PGD collected via structured, opt-in, and web-enabled platforms and devices, including wearables.2,3 Models that employ PGD from distributed cohorts are being used increasingly to measure public health outcomes; moreover, PGD collection forms the centerpiece of important new federal investments into personalized medicine that seek to energize vast cohorts in donating data via apps and devices.4 PGD offer the opportunity to inform gap areas of health research through high-resolution views into spatial, temporal, or demographic features. However, when PGD are used to answer epidemiologic questions, it is not always clear what constitutes the population at risk (PAR); when PAR is unobservable, different proxy metrics are often used.9 The exact nature of what might constitute a useful and valid choice for defining PAR depends on the research focus and is a source of difficulty.5,9 All analyses drawing on chart reviews, electronic medical records, and telephone surveys must address the extent to which the sample represents the source population (i.e., generalizability). Even when samples are drawn directly from the PAR, the dynamic nature of human populations and subtleties of selection, differential non-response, and attrition may introduce variability and impact validity of inferences. Anticipation and mitigation of these issues are vital for investigations drawing on traditional healthcare-based data and PGD. As with RCTs, where findings may not generalize with the imposition of strict eligibility criteria and participant self-selection into research,10 PGD may not fully represent patient or community populations. Moreover, once individuals contribute data, bias may be introduced through imperfect methods of observation and engagement. Researchers wishing to use PGD will need to acknowledge and, when possible, measure how characteristics of a sample differ from the PAR (i.e., issues relevant to external validity). Three challenges to using PGD stand out and are elaborated here—sample representativeness and selection bias, poorly characterized reference populations, and spatiotemporally inconsistent denominators (Table 1 and Appendix Table 1, available online).

Challenges in specifying the population at risk

Individuals considered capable of acquiring a disease are considered the PAR (Figure 1). Accurate knowledge of the PAR is a fundamental requirement in preventive medicine and public health, and is necessary for measuring effects and gauging impact. Challenges in assessing PAR have been deliberated even for traditional healthcare-based data sources;6 when PAR is unobservable, different proxy metrics are often used.7 The exact nature of what might constitute a useful and valid choice for defining PAR depends on the research focus and is a source of difficulty.5,9 All analyses drawing on chart reviews, electronic medical records, and telephone surveys must address the extent to which the sample represents the source population (i.e., generalizability). Even when samples are drawn directly from the PAR, the dynamic nature of human populations and subtleties of selection, differential non-response, and attrition may introduce variability and impact validity of inferences. Anticipation and mitigation of these issues are vital for investigations drawing on traditional healthcare-based data and PGD. As with RCTs, where findings may not generalize with the imposition of strict eligibility criteria and participant self-selection into research,10 PGD may not fully represent patient or community populations. Moreover, once individuals contribute data, bias may be introduced through imperfect methods of observation and engagement. Researchers wishing to use PGD will need to acknowledge and, when possible, measure how characteristics of a sample differ from the PAR (i.e., issues relevant to external validity). Three challenges to using PGD stand out and are elaborated here—sample representativeness and selection bias, poorly characterized reference populations, and spatiotemporally inconsistent denominators (Table 1 and Appendix Table 1, available online).
Sample Representativeness and Selection Bias

Selection biases arise when the sample does not accurately represent an underlying population. They are among the most common validity threats when making inferences (Appendix Table 1, available online). In healthcare-based studies, different types of settings, recruitment procedures, and inclusion and exclusion criteria must be carefully considered to ensure valid estimates and extrapolation of findings. For example, access bias was found to affect the Behavioral Risk Factor Surveillance System, a U.S. survey that collects data via random-digit-dial telephone interviews. However, using only data from landline-based sampling excluded important segments of the U.S. population, and resulted in measurable biases for many key health indicators. Consequently, the Behavioral Risk Factor Surveillance System shifted their sampling approach to additionally incorporate cell phones, yielding more valid, reliable, and representative measures.

Many types of selection bias also apply to PGD (Appendix Table 1, available online). Access bias, for example, is a concern when disparities in Internet access affect representativeness of PGD. Efforts to close the Internet access gap have succeeded, with only 15% of the U.S. population offline in the past year, yet variability and inequalities persist. For instance, among teens, Internet platform preferences vary by household income. Among adults, health literacy is highly correlated with health information seeking on the Internet and self-rated health. Hence, individuals who passively source or actively contribute data via online platforms may vary by income and be more literate and healthy than the general population. Access differences can also reflect group preferences, norms, technology diffusion, and sharing patterns. For example, in the U.S., 40% of African Americans aged 18–29 years who use the Internet say that they use Twitter compared with 28% among their white counterparts. Across multiple influenza-focused participatory surveillance systems, women have higher participation levels than men, with peak levels found among those aged 30–60 years. As well, in participatory diabetes surveillance within an online diabetes community, early adopters of an app that enabled PGD sharing for population health monitoring reported greater diabetes control than did later adopters, and preference toward openness in data sharing was greater among individuals with better diabetes control.

Because technology diffusion, preferences, and norms vary by age, income, and gender, observations gleaned passively from processing of Internet search or microblogging data (as done for sentiment analysis or outbreak detection) may be affected by differential selection stemming from the composition of groups using one versus another platform or tool. Bias may also be present in active surveillance, which relies on the planned donation of information by persons using Internet or mobile tools. For example, in participatory influenza reporting systems, individuals tend to sign up more when their symptoms are worse, potentially skewing incidence estimates. Technology standards, protections on data use, or controls over secondary data use can also affect representativeness. To address these issues, investigators might preselect a subsample of users that represents the PAR using probability sampling, conduct sensitivity analyses to ascertain how sample composition affects observations, or integrate data gleaned from multiple platforms (Table 1).
Poorly Characterized Reference Populations

Even if recruitment, access, and other factors are accounted for, there still can be multiple PAR options. Complexity when choosing a reference population using healthcare-based data has been discussed when assessing cohort patterns for cerebral palsy (CP), for example. For a congenital disorder such as CP, in which newborns and infants might logically be considered to constitute the PAR, the birth cohort from which a case arises is generally used as the denominator. However, poorly defined or tracked birth cohorts may necessitate the use of other denominators, such as all children aged <2 years residing in a particular area. Yet, if families of a child with CP move and cluster around health centers that are best suited for caring for children with CP, prevalence estimates might skew higher; alternatively, such health centers might be more adept at screening and diagnosis, leading to a diagnostic access bias. Ultimately, statistics calculated with different denominators will answer slightly different questions and researchers must select the most relevant estimate for their objectives.

Similarly, when using PGD, there will be multiple options for specifying the reference population. For example, investigators may elect to use all search queries, all geotagged queries, or all queries by week, and these elections may yield different estimates. Accordingly, researchers will need to select and justify their choice of a reference population based on their aims. Investigators can evaluate their choice of denominator by comparing estimates that reflect different underlying PARs or they can benchmark PGD measures against criterion standards (assessing concurrent validity), as has been done for

<table>
<thead>
<tr>
<th>PGD challenge type</th>
<th>Illustrative case(s)</th>
<th>Corrective approaches (via design or analytic strategy)</th>
<th>Practical examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample representativeness and selection bias</td>
<td>Use of social media data for health monitoring where data are collected via keyword search</td>
<td>• Select out a subsample that represents the PAR using probability sampling or other matching methods</td>
<td>• Propensity score matching used to evaluate effect of information on Twitter users.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Conduct sensitivity analyses varying denominator</td>
<td>• Assess effect of participatory symptom contribution frequency on population burden estimates.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Control for potential confounders at appropriate scale</td>
<td>• SES controlled for at ZIP code when assessing relationship between online check-ins and obesity levels.</td>
</tr>
<tr>
<td>Poorly characterized reference populations</td>
<td>Use of anonymized data as indicator of health phenomenon that do not include demographic or sampling details, such as internet search data from Google Trends</td>
<td>• Focus on concurrent validation against criterion standards, with the understanding that criteria are rarely available to compare against</td>
<td>• Assess flu incidence from internet sources versus CDC data.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Utilize multiple datasets (including a criterion standard, when available) to establish concurrent validity</td>
<td>• Combine multiple flu proxy data to improve incidence estimates.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Use Bayesian evidence synthesis to improve model performance</td>
<td>• Estimate burden of influenza from hospital-reported case data.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Supervised learning and some ground truth data to perform inference of latent characteristics</td>
<td>• Infer demographics of Twitter users.</td>
</tr>
<tr>
<td>Spatiotemporally inconsistent denominator</td>
<td>Data donation of steps or symptoms by members of an opt-in distributed, online health community</td>
<td>• Engage a cohort for cross-sectional or prospective evaluation where spatial and temporal frames can be identified</td>
<td>• Estimate influenza burden via examining contribution of participatory reports.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Partner with PGD sources/platforms to elucidate denominator features</td>
<td>• Obtain geographic information on members of an online diabetes social network and compare to geographic patterns of app uptake to test for bias.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Imputation for missing or outlier data</td>
<td>• Remove spurious spikes in influenza data.</td>
</tr>
</tbody>
</table>

CDC, Centers for Disease Control and Prevention; PAR, population at risk; PGD, personally generated data.
influenza surveillance when anonymous Internet search activity data are modeled against healthcare measures. Yet, criterion standards are not always available, particularly as PGD are often used to fill existing data gaps. Other approaches, including triangulating PGD, may improve estimates. For example, investigators might compare influenza symptom data derived from Twitter and a participatory surveillance system, where both sources derive from the same geographic area and time period. Participatory methods, even where extrapolation to PAR and formulation of a denominator is not feasible, may be informative too, where information about a health phenomenon is otherwise missing. Such numerator-only estimates gleaned through voluntary reporting may shed light on emerging phenomenon without a practical reference population for benchmarking change (Table 1). In this case, transparency around measurement and caveats on inference are especially important.

**Spatiotemporally Inconsistent Denominator**

In healthcare-based data, prevalence incidence bias, diagnostic vogue bias, and noncontemporaneous control bias are all known contributors to a denominator that varies over time or by place (Appendix Table 1, available online). In traditional longitudinal cohort studies, inclusion time is demarcated by directly observed benchmarks (e.g., date of diagnosis or treatment). For PGD, the absence of structured reporting formats or observation periods and the influence of exogenous factors, such as tool availability, create subtle spatiotemporal issues. Fixing a known and appropriate spatiotemporal data frame is vital because the underlying, relevant PAR could be dynamic, which affects interpretability, reproducibility, and prediction. For example, inclusion time is conditioned on availability of search engines, operating systems, and technology standards. Shifts in app or device popularity and privacy policies may affect data availability and reporting. Views into these factors may be obscured when data are proprietary, resulting in transparency concerns, such as those observed in pharmaceutical trials. Although some statistics on Internet and mobile tool use exist, they are not necessarily reproducible or consistent over time, providing static views into dynamic patterns.

Other exogenous factors can also cause spatiotemporal shifts. News regarding an outbreak can generate “spurious spikes” attributed to increased discussion/awareness (rather than experience of actual symptoms) in incidence time series. This can affect the validity and stability of disease incidence estimates generated from close monitoring of Internet search activity patterns, a common indicator of infectious disease activity. These spikes may not reproducibly occur in every situation; monitoring of malaria trends via Google search queries in Thailand did not reveal “spurious spikes”—possibly due to low media activity in the area. Variations in annual influenza-related search incidence estimates drawn from Google search queries reflect the lack of structure in PGD and underscore the importance of understanding how the dynamic nature of population reporting behaviors may affect the denominator. As well, PGD may be influenced by the condition of participants, analogous to the healthy participant effect. For example, associations between self-report of health status and willingness to share personal health information in studies via data sharing from web-based sources have been noted.

Temporally inconsistent denominators can be addressed by utilizing data from only individuals with acceptable participation levels over a particular time period, as done to understand incidence from participatory influenza reports. However, efforts must be made to ensure unmeasured confounding is not introduced. Spurious spikes have been distinguished from true increases in illness burden by identifying the effects of media or other exogenous factors unrelated to actual disease dynamics. Comparing with reasonable growth rates of disease allows investigators to distinguish excess variation over time. Methods to reduce these spikes can be developed after or simultaneous to data collection, as has been done in monitoring dengue trends from Internet search query data and influenza-like illness trends from participatory reports.

**CONCLUSIONS**

Challenges common to both healthcare-based data and PGD pertain to knowledge about PAR and specification of a denominator. All surveillance data have challenges that may lead to samples that do not represent the population from which they are drawn. Excitement over the potential for PGD to accelerate knowledge and inform prevention can obscure the significance of these challenges when using PGD—simply put, extra care is needed to define PAR and a denominator. The commercial nature of many novel data types amplifies these challenges. Ambitious new initiatives such as NIH’s call for creation of a vast, volunteer cohort to donate data to advance precision medicine will require exquisite attention to how PGD are invoked, sampled, compiled, and used. Similar attention is vital to optimizing use of PGD distilled from web platforms, search activity, and other sources. Overall, in both healthcare-based data and PGD, there are no “one-size-fits-all” solutions. For PGD as with other forms of data, concerns about measurement, inference, and extrapolation are mitigated where careful specification of PAR is evident, limitations addressed forthrightly and through confirmatory investigations as feasible.
ACKNOWLEDGMENTS

This work is supported by grant R21 AA023901-01 from the National Institutes of Alcohol Abuse and Alcoholism at NIH (ERW and RC), and grant IIS-1343968 from the National Science Foundation (RC). Rumi Chunara, PhD and Elissa R. Weitzman, ScD, MSc, conceived the paper and Rumi Chunara, PhD, Elissa Weitzman, ScD, MSc, and Lauren Wisk, PhD, wrote the paper. The authors thank Melanie Kenney for help with assembling relevant literature.

No financial disclosures were reported by the authors of this paper.

SUPPLEMENTAL MATERIAL

Supplemental materials associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.amepre.2016.10.038.

REFERENCES