JMIR Public Health and Surveillance

Impact Factor (2023): 3.5
Volume 8 (2022), Issue 9 ISSN 2369-2960 Editor in Chief: Travis Sanchez, PhD, MPH

Contents

Original Papers

Original Paper

Monitoring the Nutrient Composition of Food Prepared Out-of-Home in the United Kingdom: Database Development and Case Study

Yuru Huang[1], MHS; Thomas Burgoine[1], PhD; Michael Essman[1], PhD; Dolly R Z Theis[1], MPhil; Tom R P Bishop[1], MA, MEng; Jean Adams[1], PhD

Medical Research Council (MRC) Epidemiology Unit, University of Cambridge, Cambridge, United Kingdom

Corresponding Author:
Yuru Huang, MHS
Medical Research Council (MRC) Epidemiology Unit
University of Cambridge
Box 285 Institute of Metabolic Science
Cambridge Biomedical Campus
Cambridge, CB2 0QQ
United Kingdom
Phone: 44 01223 330315
Email: Yuru.Huang@mrc-epid.cam.ac.uk

Abstract

Background: Hand transcribing nutrient composition data from websites requires extensive human resources and is prone to error. As a result, there are limited nutrient composition data on food prepared out of the home in the United Kingdom. Such data are crucial for understanding and monitoring the out-of-home food environment, which aids policy making. Automated data collection from publicly available sources offers a potential low-resource solution to address this gap.

Objective: In this paper, we describe the first UK longitudinal nutritional database of food prepared out of the home, MenuTracker. As large chains will be required to display calorie information on their UK menus from April 2022, we also aimed to identify which chains reported their nutritional information online in November 2021. In a case study to demonstrate the utility of MenuTracker, we estimated the proportions of menu items exceeding recommended energy and nutrient intake (eg, >600 kcal per meal).

Methods: We have collated nutrient composition data of menu items sold by large chain restaurants quarterly since March 2021. Large chains were defined as those with 250 employees or more (those covered by the new calorie labeling policy) or belonging to the top 100 restaurants based on sales volume. We developed scripts in Python to automate the data collection process from business websites. Various techniques were used to harvest web data and extract data from nutritional tables in PDF format.

Results: Automated Python programs reduced approximately 85% of manual work, totaling 500 hours saved for each wave of data collection. As of January 2022, MenuTracker has 76,405 records from 88 large out-of-home food chains at 4 different time points (ie, March, June, September, and December) in 2021. In constructing the database, we found that one-quarter (24.5%, 256/1043) of large chains, which are likely to be subject to the United Kingdom's calorie menu labeling regulations, provided their nutritional information online in November 2021. Across these chains, 24.7% (16,391/66,295) of menu items exceeded the UK government's recommendation of a maximum of 600 kcal for a *single meal*. Comparable figures were 46.4% (29,411/63,416) for saturated fat, 34.7% (21,964/63,388) for total fat, 17.6% (11,260/64,051) for carbohydrates, 17.8% (11,434/64,059) for sugar, and 35.2% (22,588/64,086) for salt. Furthermore, 0.7% to 7.1% of the menu items exceeded the maximum *daily* recommended intake for these nutrients.

Conclusions: MenuTracker is a valuable resource that harnesses the power of data science techniques to use publicly available data online. Researchers, policy makers, and consumers can use MenuTracker to understand and assess foods available from out-of-home food outlets. The methods used in development are available online and can be used to establish similar databases elsewhere.

(JMIR Public Health Surveill 2022;8(9):e39033) doi:10.2196/39033

KEYWORDS

nutritional database; web scraping; food prepared out of the home; out-of-home; data science; chains

Introduction

The consumption of food prepared out of the home is increasing worldwide. Eating outside of the home accounts for over half of food expenditures in the United States [1], 34% in Spain [2], and 27% in New Zealand [3]. In the United Kingdom, the percentage of total food expenditure outside the home was 28% in 2018/2019 [4]. In addition to dining out of the home, the rapid expansion of online delivery services also facilitates the consumption of food prepared outside the home. In an international study, 15% of respondents reported online delivery use [5]. The frequent consumption of food prepared out of the home is a public health concern as it is typically high in energy, salt, saturated fat, and sugar [6-11]. Frequent consumption of these foods has been associated with a higher BMI and an elevated risk of cardiovascular diseases [12,13].

The increasing frequency of eating out of the home makes food prepared by these chains an important avenue for improving population dietary quality. Internationally, decision makers are developing policies that promote healthier out-of-home options. The goal is to improve the out-of-home food environment and ensure that "the healthy choice is the easy choice" [14,15]. In the United Kingdom, for example, the government introduced the mandatory calorie menu labeling policy as part of a wider obesity strategy [16-18]. It requires large out-of-home food chains with 250 employees or more to add calorie labeling to menus for most of the food they sell starting from April 6, 2022 [17,19]. The effects of the policy may be not only to help consumers make informed choices but also to incentivize out-of-home chains to reformulate or provide healthier offerings [20].

Despite progress in policy, there are limited data on the nutrient composition of food prepared out of the home. As a recent World Health Organization report highlighted, a lack of quality data hinders the monitoring of the out-of-home food environment, creating barriers and challenges for policy development and evaluation [2]. With respect to the UK calorie labeling policy, a longitudinal nutritional database of food prepared out of the home is needed to investigate the direct effects of this policy on menus (eg, healthier menu options) and the overall effect on population dietary intake. In addition to aiding policy evaluations, nutrient composition data for out-of-home foods might also improve nutrient intake estimation in epidemiology studies by incorporating brand-specific information that is currently rarely included [21].

Many restaurants post nutritional information of their menu items online, and this information can be valuable for research. In the United States, a longitudinal restaurant nutritional database, MenuStat, was established in 2013 using information sourced from restaurant websites [22]. It has proven to be a valuable resource for researchers to advance the understanding of the restaurant food environment [23,24], assess changes in restaurant foods over time [25-28], and evaluate the potential impact of the calorie menu labeling policy in the United States [29]. Elsewhere, similar nutritional data for food and drinks prepared out of the home have been collected in New Zealand [30], Australia [31], and Canada [32]. However, to the best of our knowledge, these databases' nutritional data were manually collected by researchers, and perhaps as a result, they have not been updated regularly, if at all. Manual collection of restaurant nutritional data by hand requires extensive human resources and is prone to error.

Web scraping, or automated data extraction from websites, provides an efficient, reliable, and flexible alternative to hand transcribing website data [33,34]. In the United Kingdom, web scraping has been used to establish a longitudinal nutritional database—foodDB—of packaged foods sold in large supermarkets [33]. Yet nutritional data are still limited for food prepared out of the home, largely due to the heterogeneity of how and what nutritional information is presented on out-of-home chains' websites.

This study presents MenuTracker, the first longitudinal nutritional database, updated quarterly, of food prepared by large out-of-home food chains in the United Kingdom. In the future, we will use this database to describe and characterize changes in the nutrient content of out-of-home foods over time, to evaluate the effect of the calorie menu labeling policy, and potentially to improve nutrient intake estimation in nutritional epidemiology studies. In this paper, we aim to describe MenuTracker and its data collection methods, identify gaps in the presentation of nutritional information online, and demonstrate an example application of MenuTracker in food and nutrition research.

Methods

Overview

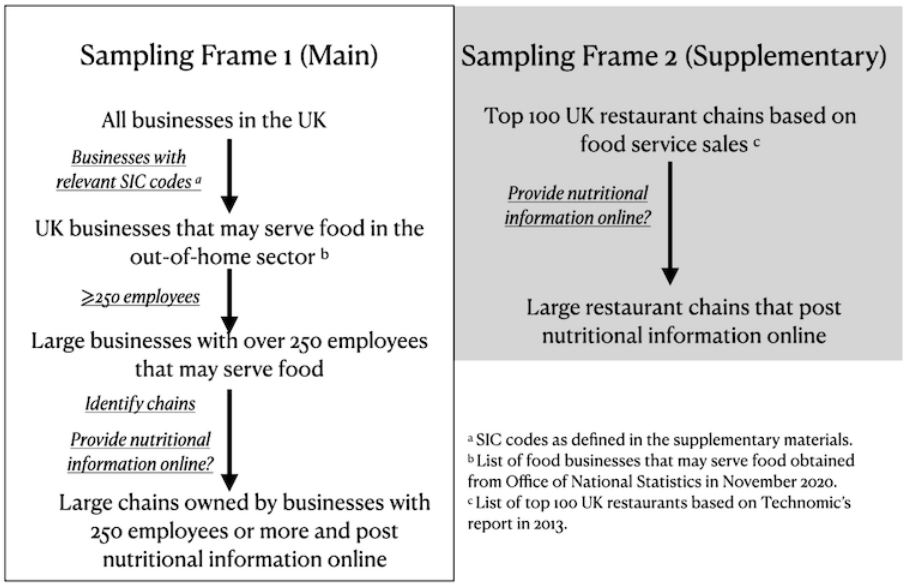
We have collated data on the nutritional composition of menu items sold by large UK food businesses (likely to be subject to the calorie labeling policy) from their websites quarterly since March 2021. We automated this data collection using web scraping techniques and PDF extraction tools. In the example application of this data, we examined the proportion of menu items exceeding recommended energy and nutrient intake values for the UK population.

Out-of-Home Food Chain Inclusion Criteria

Out-of-home food chains were defined as any chain where food or drink is prepared for immediate consumption by the person who buys it [18]. Figure 1 shows inclusion criteria for out-of-home food chains in MenuTracker. Essentially, two sampling frames were used for MenuTracker. Sampling frame one—the primary sampling frame—was a list of businesses that were potentially relevant (ie, may serve food) to the UK calorie menu labeling policy. In this study, we use the term business to refer to the parent company and chain to refer to the brands belonging to the businesses. We obtained the list from the Office for National Statistics (ONS) in October 2020. This list contained all businesses with Standard Industrial Classification

(SIC) codes that indicated they might serve food (eg, "SIC 47.11: Retail sale in non-specialized stores with food, beverages or tobacco predominating") and their employee numbers. A full list of included SIC codes included can be found in Multimedia Appendix 1. We then filtered to businesses with 250 employees or more and reviewed them to determine which of those provided nutritional information online. If there were multiple chains under one business, each chain was reviewed to determine the availability of online nutritional information. For example, under the business Mitchells & Butlers, there were more than 10 different chains, including Sizzling Pubs, Vintage Inns, Harvester, Ember Inns, and Toby Carvery. Each chain was reviewed and included if the chain provided nutritional information online. Sampling frame two—the supplementary sampling frame—contained the top 100 UK restaurants based on sales volume. Sales data were provided by Technomic, a market research company specializing in the food service industry, in 2013 [35]. This list of the top 100 food businesses supplemented our primary sampling frame to capture all large food businesses in the United Kingdom that may be eligible for calorie labeling. Each of these listed businesses were reviewed to determine whether they provided nutritional information online and would thus be included. Both lists are reviewed annually to check for changes in chains that provide nutritional information online.

Figure 1. Out-of-home food chains inclusion criteria. SIC: Standard Industrial Classification.



Menu and Menu Item Inclusion Criteria

All out-of-home menu items with online nutritional information were included in the data collection. In this paper, we used "nutritional information" to refer to "energy and nutritional information." Menu item data were collected as they appeared on websites. We collected the out-of-home food chain name, menu item name, menu section, item description, serving size, and nutritional information. Additionally, ingredient statements, allergens, and dietary information (eg, vegetarian) were extracted if available on the same page or in the same PDF document.

When menus differed between locations (eg, Weatherspoon had different food menus at different locations), the first listed location in London was selected to represent the out-of-home chain. If the chain did not have a presence in London, a random location was selected. The same location for the chain was used in different data collection waves. When an out-of-home chain had different menus (eg, "Core" or "Delivery"), the main menu (eg, "Core Menu" or "Main Menu") was used. The children's menu and relevant promotional menus were also included in addition to the main menu, where available. If the nutritional document was last updated more than 3 years ago, it was deemed invalid. Only 1 restaurant was excluded due to this criterion.

Menu items of different sizes and beverages with different customization options were also included. For example, beverages with multiple choices of milk (eg, oat milk, soy milk, or whole milk) were entered as individual records, as well as pizza of different sizes (eg, individual, medium, large, or XXL). However, highly customizable menu items such as *building your own burritos* can lead to a large number of possible combinations. We collected the default customizations for these. If there were no default customizations, meal components for each menu item were collected and assigned an item ID for future linkage.

Data Collection

Before MenuTracker data collection, we collected four waves of data in a pilot study, which has been described in detail elsewhere [36,37]. Using the sampling frames described above, we used automated Python scripts to collect data for MenuTracker proper beginning March 2021. The codebase was developed from October 2020 to February 2021. Included chains presented their nutritional information directly on web pages or in separate downloadable PDF files. Despite the variations in how this information could be shown on web pages (eg, some were presented on individual item pages, while others were presented as nutrition tables separate from the item page), the

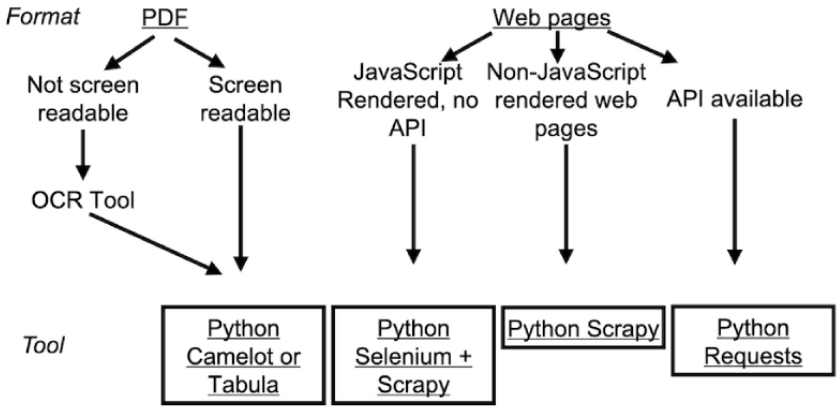
web scraping fundamentals were the same. Hence, we can describe the web scraping method for all "web pages," irrespective of how the nutritional information was presented (Figure 2).

For nutritional information presented in non–screen-readable PDF format, we first used the "Scan & OCR" tool in Adobe Acrobat to convert the PDF to be screen readable. We then used the Python packages Tabula or Camelot to extract data tables from PDFs. Both packages are designed to enable table extraction from PDFs, with Camelot allowing for more user customization and Tabula providing a more stable user interface to select table boundaries. The choice of package depended on the quality of output from these packages. In both packages, two table parsing methods, "stream" and "lattice," were available. The "stream" parsing method estimates the number of columns based on row ranges and table areas. It performed better for PDF tables without clear boundaries and lines. The "lattice" parsing method defines tables based on table lines. It performed better for PDF tables with clear line segments. For each chain, we randomly selected one menu item in our data extract and compared to data on the website to ensure accuracy. We also checked all outliers for energy and nutrient values (eg, top/bottom 5%) in extracted data against websites.

Different data scraping methods were used to harvest nutritional information from web pages. For simple, non–JavaScript-rendered web pages, we used the Scrapy framework in Python to extract data. Scrapy is a powerful and highly customizable web scraping framework. However, the Scrapy framework alone cannot gather data from websites rendered by JavaScript. We used Selenium WebDriver within the Scrapy framework in these instances. There were a few websites where nutritional data were loaded through application programming interface (API) requests. An API is built for information retrieval, enabling data transmission between software. For example, when a web page is being loaded, the web server requests data from the company's database/server through the API. For websites where nutritional information was loaded through an API (identified through the inspection of developer tools in Chrome), we used the Python Request library to pull data directly.

We abided by the UK ONS's safe web scraping policy to minimize the burden of our extraction on web site owners [38]. Additionally, we worked within what is allowed from a copyright perspective. The UK government's guidance on copyright outlines several exceptions, including limited use of copyright works for noncommercial research studies [39]. Scripts were checked quarterly and updated to accommodate any changes in web site structures that may have occurred since the previous scrape. The most up-to-date scripts are available publicly on GitHub [40].

Figure 2. Restaurant Nutritional information formats and data collection tools. API: application programming interface; OCR: optical character recognition.



Data Cleaning and Standardization

To address inconsistency in the portion size information for pizzas prepared by large pizza chains such as Pizza Hut and Papa John's, we calculated the energy and nutrient values for 3 slices of pizza if an item was described as "large," "family," "for sharing," or "medium," and the whole pizza if an item was described as "small" or "individual." This was in accordance with the way Domino's—the leading pizza chain in the United Kingdom—presented the nutritional information on its pizzas. We also saved the original energy and nutrient values for pizza items.

The field names were also standardized for each out-of-home food chain. For example, "sugar," "sugar content," and "sugars" were all standardized as "sugar." Operators in the nutrition values were also removed before converting to numeric values. As an example, "<0.05" was replaced with "0.05" for conservative estimates. Nutrition values with "-" or blanks were set to missing. All verbatim texts (including operators) were stored in each restaurant's data collection folder.

After standardization and data cleaning, we compiled all data into one master file for each quarterly data collection.

Energy and Daily Nutrient Intake Values

For our example application of MenuTracker data, we estimated the proportion of menu items exceeding the United Kingdom's per meal recommendations and daily reference intakes in 2021. The daily reference intakes for an adult are 2000 kcal for energy, less than 70 g for total fat, less than 20 g for saturated fat, 260 g for carbohydrates, 90 g for total sugars, and less than 6 g for salt [41]. The reference intake values for energy and nutrients are calculated based on an average female with an average

amount of physical exercise. The UK government recommends adults also consume no more than 600 kcal for lunch or dinner [42]. Although there are no specific *per meal* recommendations for other nutrients, it has been suggested that any meal components should not exceed 30% of daily reference intake, in line with the UK government guidelines [43]. As such, we set the *per meal* recommendations for total fat, saturated fat, carbohydrates, sugars, and salt at 30% of the daily reference intakes, proportional to the energy recommendations. We used all MenuTracker records collected in 2021 for this analysis.

Results

Availability of Calorie Information Among Large Out-of-Home Food Chains

A total of 1043 businesses with 250 employees or more were identified in October 2020. This represents a likely overestimate of the number of businesses potentially eligible for the UK calorie labeling policy (eg, not all "historical sites and buildings and similar visitor attractions" in SIC 91.03 served food). Among these 1043 businesses, 256 (24.5%) presented nutritional information for their menu items (food prepared out of the home) available online in November 2021. Companies operating as franchisees of other businesses (n=196) typically serve the same menu as provided by the franchisor. As such, data on franchisees were not collected unless the main chain had not been captured (n=3; eg, Taco Bell UK).

In total, 82 unique chains provided nutritional information online using the main sampling frame in March 2021. The supplementary sampling frame added 3 additional food chain brands (ie, Papa John's, PAUL, and Ben & Jerry's). This gave a total of 85 unique chains.

Data Collection Automation

In our pilot study, it was estimated to have taken one researcher 36 working days to collect and transcribe data from 42 out-of-home food chains in 2018 [44]. Using automated programs written in Python, we were able to collect data from around 85 food chains in about 10 working days. This is an 85% reduction in hours, totaling approximately 500 hours, compared to the manual transcription in 2018.

Descriptive Statistics

As shown in Table 1, a total of 85, 83, 79, and 81 out-of-home food chain brands were included in MenuTracker across March, June, September, and December 2021, respectively. A list of all included chains as of March 2021 can be found in Multimedia Appendix 2. The number of included food chains varied as some out-of-home food chains stopped providing nutritional information online while others started during 2021. Some chains did not provide nutritional information for every menu item listed. Among menu items with calorie information (86.1-87.6% of all items identified), information on fat, saturated fat, carbohydrates, sugars and salt were available for the majority (94.6-97.5%). However, only 36.7%-42.4% of items with calorie information had associated serving size information and around half of them provided fibre information.

Table 1. MenuTracker 2021 data summary statistics.

	March 2021	June 2021	September 2021	December 2021
Out-of-home chains, n	85	83	79	81
Menu items, n	18,005	19,310	19,392	19,698
Item-level availability				
Energy, n	15,766	16,678	16,882	16,969
Fat, n (%)[a]	15,244 (96.7)	15,785 (94.6)	16,069 (95.2)	16,290 (96.0)
Saturated fat, n (%)[a]	15,261 (96.8)	15,774 (94.6)	16,028 (94.9)	16,353 (96.4)
Carbohydrates, n (%)[a]	15,183 (96.3)	16,021 (96.1)	16,308 (96.6)	16,539 (97.5)
Sugars, n (%)[a]	15,233 (96.6)	16,028 (96.1)	16,279 (96.4)	16,519 (97.3)
Protein, n (%)[a]	15,160 (96.2)	15,777 (94.6)	16,078 (95.2)	16,194 (95.4)
Salt, n (%)[a]	15,179 (96.3)	16,009 (96.0)	16,357 (96.9)	16,541 (97.5)
Fiber, n (%)[a]	8167 (51.8)	8367 (50.2)	8750 (51.8)	8229 (48.5)
Serving weight, n (%)[a,b]	6348 (40.3)	6721 (40.3)	7153 (42.4)	6235 (36.7)

[a]Percentage per items that provided calorie information.

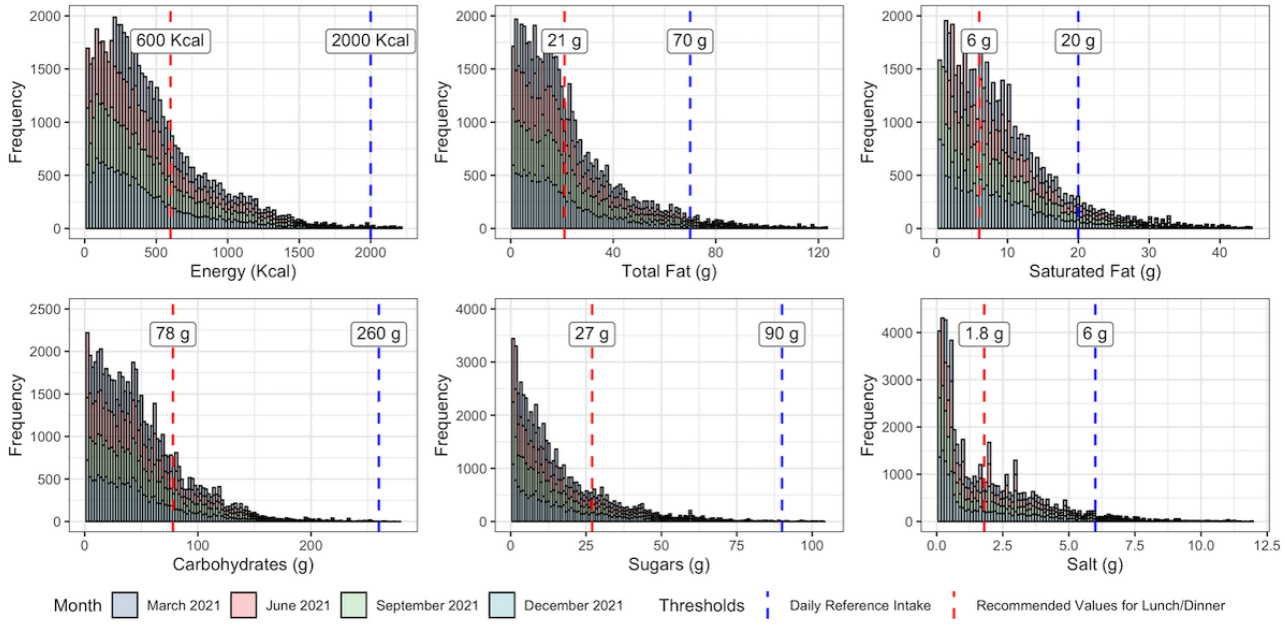
[b]Items that provided serving size information directly or if the information can be calculated through nutrient per serving and nutrient density.

Proportion of Menu Items Exceeding Per Meal and Daily Reference Intake

As shown in Figure 3, the largest proportion of any nutrient exceeding allowances was saturated fat, where 46.4% (29,411/63,416) of menu items exceeded per meal recommendations and 7.1% (4523/63,416) exceeded the daily reference intake. Proportions of menu items exceeding the per meal recommendation for salt, total fat, energy, sugar, carbohydrates were 35.2% (22,588/64,086), 34.7%

(21,964/63,388), 24.7% (16,391/66,295), 17.8% (11,434/64,059), and 17.6% (11,260/64,051), respectively. Comparable figures for proportions of items exceeding daily reference intakes were 0.7% (497/66,295) for energy, 3.6% (2258/63,388) for total fat, 0.1% (75/64,051) for carbohydrates, 0.4% (245/64,059) for sugar, and 4.2% (2722/64,086) for salt. Detailed proportions by different data collection waves can be found in Multimedia Appendix 3.

Figure 3. Menu item energy and nutrient distributions.



Discussion

Summary of Findings

In this study, we described MenuTracker, the first longitudinal nutritional database of food prepared by out-of-home food chains in the United Kingdom who provided this information online. As of December 2021, MenuTracker includes 76,405 menu item records from over 80 large out-of-home food chains, collected across 4 time points. The database is semiautomated and time stamped. In constructing the database, we found that less than one-quarter of businesses that might be subject to a calorie menu labeling policy in the United Kingdom had presented menu item energy information on their websites as of November 2021. Across chains that provided nutritional information online, a large proportion of menu items did not have associated serving weight or nutrient density information. Using MenuTracker data, we found that more than one-third of items available in the database were high in saturated fat, total fat, or salt, and one-fourth were high in energy in comparison to recommendations for a main meal.

Interpretation of Findings

Uniqueness of the Database

MenuTracker data represents a valuable resource for dietary public health and nutrition research, as well as policy making in this space. As it is regularly updated and data are time stamped, it allows researchers and policy makers to track nutritional composition changes in the UK out-of-home food environment over time. The nutritional data contained are harvested directly from the official websites of food chains in a systematic fashion, ensuring accuracy. Among other potential future applications, MenuTracker will enable the evaluation of calorie menu labeling, assessments of the out-of-home food environment, and refinement of nutrient estimation in nutritional epidemiologic studies. In the UK policy context, few obesity policies have been proposed with an evaluation plan and MenuTracker data may facilitate policy evaluations [45].

Manual collection of nutritional composition data is labor-intensive. In 2018 and 2019, researchers in our group hand transcribed MenuTracker data annually. The automation codebase reduced 85% of manual work hours, allowing us to continue collecting MenuTracker data every quarter. We currently have resources to continue to collect MenuTracker data at least until spring 2023.

Out-of-Home Food Chains Nutrient Reporting in the United Kingdom

In the United Kingdom, starting from April 6, 2022, the mandatory calorie labeling policy is now in effect [19]. In this study, we found less than one-quarter of potentially eligible businesses posted calorie information on their websites in November 2021. This is broadly consistent with a previous UK study in 2018 where only 17% of large chains were found to provide calorie labeling in store [46]. However, our calculated percentage could be an underestimation, as some of these out-of-home businesses could be exempt from the calorie menu labeling policy (eg, seasonal items only), or they may not serve food at all. Nonetheless, our results highlight a gap in nutrient reporting for out-of-home chains before the regulations came into effect, which may indicate the industry's reluctance to present this information voluntarily.

Notably, for out-of-home food chains that post nutritional composition online, nutrient density (eg, kcal per 100 g)

information (or serving weight, which would permit its calculation) was missing for around 60% of items. As most voluntary reduction programs (eg, for salt and sugar) set targets based on nutrient density, this may prohibit monitoring and evaluation of these initiatives [47,48]. Moreover, this information is critical for identifying menu item reformulations—a key potential impact of menu labeling regulation—as any overall change in nutrient content may be caused by reformulation or change in serving size. Without serving size indicators, these possibilities cannot be distinguished. Mandating the declaration of serving sizes (alongside calorie information) could enable more comprehensive evaluations of interventions targeting the out-of-home food retail sector.

Example Use Cases of MenuTracker

In this study, we demonstrated an example application of MenuTracker data. We used MenuTracker data to estimate the proportions of menu items excessively high in energy and nutrient content. The proportion of menu items exceeding the per meal energy recommendation was broadly similar to that previously reported in the United Kingdom [6,7]. Our results also draw attention to other nutrients high in out-of-home food, such as saturated fat and salt. Our data reaffirms that in 2021 food prepared by large out-of-home chains were high in energy and nutrients such as sugars (for which intake should be limited).

In a recent paper, we demonstrated the feasibility of using MenuTracker to monitor changes in the nutritional composition of food prepared out of the home over time [36]. Elsewhere, we used US MenuStat (equivalent to MenuTracker) data to draw international comparisons in the nutritional composition of food away from home [49]. These applications of MenuTracker demonstrate its power as a research tool.

Limitations and Future Directions

While most of the MenuTracker data collection has been automated, manual review and modification of code are still needed at each wave of data collection. This need arises due to two main issues: challenges of data extraction from PDF documents and changing web site structures. PDF conversion tools are imperfect and unable to correctly identify table boundaries at times, which necessitates the manual checking of results. Web site structures and design are also subject to change, which requires the updating of paths for certain elements or the rewriting of scripts. Currently, we monitor the changes in web site structures each quarter to ensure the codebase works properly for each data wave. However, as more chains start providing nutritional information on their web pages, along with advances in PDF conversion tools, full automation might be achievable in the future. Alternatively, as the calorie labeling regulations extend to third-party delivery platforms such as Just Eat and Deliveroo, it would be less resource-intensive to collect calorie information from these delivery platforms, compared to 80 individual websites. In the future, we may transition to obtaining calorie information in this way. However, at this time, information on other nutrients remains unavailable on these platforms, meaning that such a transition would lead to loss of the breadth of information.

MenuTracker itself is not without limitations. MenuTracker focuses on large out-of-home food chains and does not include energy and nutritional information from smaller chains or independent businesses. However, the UK Government estimates that these large chains make up 50% of all out-of-home food and drink sales [18]. Additionally, MenuTracker focuses on online menus from chains' official web sites, which may differ from the physical menus in-store or on delivery platforms. This could be important, as throughout the COVID-19 pandemic, use of online food delivery services has increased worldwide [50,51]. Future research is needed to understand potential differences between online menus from chains' official web sites, in-store menus, and menus on delivery service web sites. Moreover, chains and menu items with online nutritional information may also have different characteristics compared with those without. Future research could also explore what types of chains and menu items are more likely to have the full energy and nutritional information.

Another limitation relates to our sampling frame. The list of food businesses we obtained was for October 2020, and businesses are likely to have both been added to and dropped from the list since then. To mitigate this concern, we will review this list annually. However, there remains the possibility that new large businesses have not been subsequently included in MenuTracker. Moreover, the fact that only one-quarter of businesses potentially eligible for the calorie labeling policy provided nutritional information online undermines the market coverage of MenuTracker. As the calorie labeling policy is now in effect, MenuTracker will be expanded to include new out-of-home food chains that start providing relevant information—although this may well be limited to energy information only. Moreover, MenuTracker relies on self-reporting of nutritional information by chains, which may not be entirely accurate. However, we believe that these large chains have the incentive to provide accurate nutritional information. Lastly, the nutritional information presented on chain web sites may be outdated. It was difficult for us to determine when and how nutritional information was obtained for each chain if no time stamp was provided.

In addition to the inclusion of new businesses and potentially obtaining data from delivery platforms, MenuTracker will benefit in the future from the development of machine learning models for the categorization of menu items and automated linkage. This will allow category-specific (eg, food and drinks) tracking of energy and nutrient content over time while saving hundreds of hours of manual labeling for each data wave. We piloted a record linkage process in our recent paper, which was used to track energy and nutrient changes for the same set of menu items over time [36]. We plan to refine this technique and implement record linkage in existing and future MenuTracker data.

Conclusions

Using data science techniques, we established MenuTracker, a valuable database for researchers and policy makers to understand and assess foods available from large chains in the United Kingdom who provide this information online. In constructing the database, we found less than one-quarter of

chains potentially eligible for the calorie labeling policy provided nutritional information online, and serving size information was missing for a large proportion of menu items in 2021. This may present challenges for monitoring the out-of-home food environment. This study also adds to the growing body of evidence suggesting that foods prepared out of the home are high in saturated fat, total fat, and salt in the United Kingdom. The methods used in development are available online and can be used to establish similar databases elsewhere.

Acknowledgments

This paper was funded by a UK Research and Innovation grant MC_UU_00006/7. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) license to any author-accepted manuscript version that arises. YH is supported through a Gates Cambridge Scholarship. DRZT is supported by a PhD studentship awarded by the National Institute for Health Research, School for Public Health Research (grant PD-SPH-2015-10025). No funders had any role in the study design; collection, analysis, and interpretation of data; the writing of the manuscript; or the decision to submit the manuscript for publication.

Data Availability

The codebase is publicly available on GitHub [40]. All analysis codes are also available upon request. The anonymized data set used in the analyses are available on request. Use of our data is only permitted for noncommercial purposes.

Authors' Contributions

YH and JA conceptualized the study. YH developed the database, conducted the formal analysis, and wrote the original draft. ME contributed to the study sample frame review. JA and TB supervised study. TB, ME, DRZT, TRPB, and JA reviewed and edited the paper. TRPB reviewed the code. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Standard Industrial Classification codes used for selecting businesses that may serve food.
[DOCX File , 21 KB - publichealth_v8i9e39033_app1.docx]

Multimedia Appendix 2
Chains included in MenuTracker, March 2021.
[DOCX File , 17 KB - publichealth_v8i9e39033_app2.docx]

Multimedia Appendix 3
Proportion of menu items exceeding per meal and daily reference intake by data collection wave.
[DOCX File , 15 KB - publichealth_v8i9e39033_app3.docx]

References

1. Saksena MJ, Okrent AM, Anekwe TD, Cho C, Dicken C, Effland A, et al. America's eating habits: food away from home. Econ Inf Bull 2018:172.
2. The out-of-home food sector – exponential growth in an unregulated market. World Health Organization. 2021. URL: https://www.euro.who.int/en/health-topics/disease-prevention/nutrition/news/news/2021/9/the-out-of-home-food-sector-exponential-growth-in-an-unregulated-market [accessed 2021-12-09]
3. Kiwis growing taste for takeaways and eating out. Stats NZ. 2020 Aug 13. URL: https://www.stats.govt.nz/news/kiwis-growing-taste-for-takeaways-and-eating-out [accessed 2021-06-18]
4. National statistics: Family Food 2018/19. GOV.UK. 2020. URL: https://www.gov.uk/government/statistics/family-food-201819/family-food-201819 [accessed 2021-09-24]
5. Keeble M, Adams J, Sacks G, Vanderlee L, White CM, Hammond D, et al. Use of online food delivery services to order food prepared away-from-home and associated sociodemographic characteristics: a cross-sectional, multi-country analysis. Int J Environ Res Public Health 2020 Jul 17;17(14):5190 [FREE Full text] [doi: 10.3390/ijerph17145190] [Medline: 32709148]
6. Robinson E, Jones A, Whitelock V, Mead BR, Haynes A. (Over)eating out at major UK restaurant chains: observational study of energy content of main meals. BMJ 2018 Dec 12;363:k4982 [FREE Full text] [doi: 10.1136/bmj.k4982] [Medline: 30541906]

7. Muc M, Jones A, Roberts C, Sheen F, Haynes A, Robinson E. A bit or a lot on the side? Observational study of the energy content of starters, sides and desserts in major UK restaurant chains. BMJ Open 2019 Oct 07;9(10):e029679 [FREE Full text] [doi: 10.1136/bmjopen-2019-029679] [Medline: 31594875]
8. Jaworowska A, Blackham T, Stevenson L. Nutritional composition of takeaway meals served by independent small outlets. Proc Nutr Soc 2011 Oct 14;70(OCE4):E166. [doi: 10.1017/s0029665111002175]
9. Jaworowska A, Blackham TM, Long R, Taylor C, Ashton M, Stevenson L, et al. Nutritional composition of takeaway food in the UK. Nutr Food Sci 2014 Sep 2;44(5):414-430. [doi: 10.1108/nfs-08-2013-0093]
10. Jaworowska A, Blackham T, Davies IG, Stevenson L. Nutritional challenges and health implications of takeaway and fast food. Nutr Rev 2013 May;71(5):310-318. [doi: 10.1111/nure.12031] [Medline: 23590707]
11. Jaworowska A, Blackham T, Stevenson L, Davies IG. Determination of salt content in hot takeaway meals in the United Kingdom. Appetite 2012 Oct;59(2):517-522. [doi: 10.1016/j.appet.2012.06.018] [Medline: 22772043]
12. Du Y, Rong S, Sun Y, Liu B, Wu Y, Snetselaar L, et al. Association between frequency of eating away-from-home meals and risk of all-cause and cause-specific mortality. J Acad Nutr Diet 2021 Sep;121(9):1741-1749.e1. [doi: 10.1016/j.jand.2021.01.012] [Medline: 33775622]
13. Braithwaite I, Stewart AW, Hancox RJ, Beasley R, Murphy R, Mitchell EA, ISAAC Phase Three Study Group. Fast-food consumption and body mass index in children and adolescents: an international cross-sectional study. BMJ Open 2014 Dec 08;4(12):e005813 [FREE Full text] [doi: 10.1136/bmjopen-2014-005813] [Medline: 25488096]
14. Public Health England. Encouraging healthier 'out of home' food provision. GOV.UK. 2017. URL: https://www.gov.uk/government/publications/encouraging-healthier-out-of-home-food-provision [accessed 2021-02-15]
15. Ashe M, Graff S, Spector C. Changing places: policies to make a healthy choice the easy choice. Public Health 2011 Dec;125(12):889-895. [doi: 10.1016/j.puhe.2011.04.010] [Medline: 21917279]
16. Department of Health and Social Care, Chuchill J. Calorie labelling on menus to be introduced in cafes, restaurants and takeaways. GOV.UK. 2021. URL: https://www.gov.uk/government/news/calorie-labelling-on-menus-to-be-introduced-in-cafes-restaurants-and-takeaways [accessed 2021-06-02]
17. Department of Health and Social Care. Calorie labelling in the out of home sector: implementation guidance. GOV.UK. 2021. URL: https://www.gov.uk/government/publications/calorie-labelling-in-the-out-of-home-sector/calorie-labelling-in-the-out-of-home-sector-implementation-guidance [accessed 2021-01-06]
18. Department of Health and Social Care. Calorie labelling for food and drink served outside of the home. GOV.UK. 2018. URL: https://www.gov.uk/government/consultations/calorie-labelling-for-food-and-drink-served-outside-of-the-home [accessed 2021-02-15]
19. Department of Health and Social Care. New calorie labelling rules come into force to improve nation's health. GOV.UK. 2022. URL: https://www.gov.uk/government/news/new-calorie-labelling-rules-come-into-force-to-improve-nations-health [accessed 2022-04-15]
20. Robinson E, Marty L, Jones A, White M, Smith R, Adams J. Will calorie labels for food and drink served outside the home improve public health? BMJ 2021 Jan 20;372:n40. [doi: 10.1136/bmj.n40] [Medline: 33472836]
21. Learn more about food composition databases for 24-hour dietary recalls and food records. Dietary Assessment Primer. URL: https://dietassessmentprimer.cancer.gov/learn/recall-record.html [accessed 2022-03-17]
22. Niederman S, Leonard E, Clapp J. Restaurant nutrition reporting and impact on surveillance. J Food Composition Analysis 2017 Dec;64:73-77. [doi: 10.1016/j.jfca.2017.04.011]
23. Alexander E, Rutkow L, Gudzune KA, Cohen JE, McGinty EE. Healthiness of US chain restaurant meals in 2017. J Acad Nutr Diet 2020 Aug;120(8):1359-1367. [doi: 10.1016/j.jand.2020.01.006] [Medline: 32169296]
24. Jarlenski M, Wolfson J, Bleich S. Macronutrient composition of menu offerings in fast food restaurants in the U.S. Am J Prev Med 2016 Oct;51(4):e91-e97. [doi: 10.1016/j.amepre.2016.03.023] [Medline: 27180027]
25. Wolfson JA, Moran AJ, Jarlenski MP, Bleich SN. Trends in sodium content of menu items in large chain restaurants in the U.S. Am J Prev Med 2018 Jan;54(1):28-36. [doi: 10.1016/j.amepre.2017.08.018] [Medline: 29056370]
26. Moran AJ, Block JP, Goshev SG, Bleich SN, Roberto CA. Trends in nutrient content of children's menu items in U.S. chain restaurants. Am J Prev Med 2017 Mar;52(3):284-291 [FREE Full text] [doi: 10.1016/j.amepre.2016.11.007] [Medline: 28089130]
27. Bleich SN, Wolfson JA, Jarlenski MP. Calorie changes in large chain restaurants: declines in new menu items but room for improvement. Am J Prev Med 2016 Jan;50(1):e1-e8 [FREE Full text] [doi: 10.1016/j.amepre.2015.05.007] [Medline: 26163168]
28. Bleich SN, Soto MJ, Dunn CG, Moran AJ, Block JP. Calorie and nutrient trends in large U.S. chain restaurants, 2012-2018. PLoS One 2020;15(2):e0228891 [FREE Full text] [doi: 10.1371/journal.pone.0228891] [Medline: 32040526]
29. Petimar J, Zhang F, Cleveland LP, Simon D, Gortmaker SL, Polacsek M, et al. Estimating the effect of calorie menu labeling on calories purchased in a large restaurant franchise in the southern United States: quasi-experimental study. BMJ 2019 Oct 30;367:l5837 [FREE Full text] [doi: 10.1136/bmj.l5837] [Medline: 31666218]
30. Eyles H, Jiang Y, Blakely T, Neal B, Crowley J, Cleghorn C, et al. Five year trends in the serve size, energy, and sodium contents of New Zealand fast foods: 2012 to 2016. Nutr J 2018 Jul 09;17(1):65 [FREE Full text] [doi: 10.1186/s12937-018-0373-7] [Medline: 29983114]

31. Wellard-Cole L, Goldsbury D, Havill M, Hughes C, Watson WL, Dunford EK, et al. Monitoring the changes to the nutrient composition of fast foods following the introduction of menu labelling in New South Wales, Australia: an observational study. Public Health Nutr 2018 Apr;21(6):1194-1199. [doi: 10.1017/S1368980017003706] [Medline: 29262878]
32. The Canadian Food Supply. L'Abbe Lab. 2022. URL: https://labbelab.utoronto.ca/projects/the-canadian-food-supply/ [accessed 2022-03-17]
33. Harrington RA, Adhikari V, Rayner M, Scarborough P. Nutrient composition databases in the age of big data: foodDB, a comprehensive, real-time database infrastructure. BMJ Open 2019 Jun 27;9(6):e026652 [FREE Full text] [doi: 10.1136/bmjopen-2018-026652] [Medline: 31253615]
34. Hillen J. Web scraping for food price research. Br Food J 2019 Nov 12;121(12):3350-3361. [doi: 10.1108/bfj-02-2019-0081]
35. Top 100 U.K. chain restaurant report. Technomic. 2014. URL: https://www.technomic.com/available-studies/industry-reports [accessed 2020-12-02]
36. Huang Y, Theis D, Burgoine T, Adams J. Trends in energy and nutrient content of menu items served by large UK chain restaurants from 2018 to 2020: an observational study. BMJ Open 2021 Dec 30;11(12):e054804. [doi: 10.1136/bmjopen-2021-054804]
37. Theis DRZ, Adams J. Differences in energy and nutritional content of menu items served by popular UK chain restaurants with versus without voluntary menu labelling: a cross-sectional study. PLoS One 2019;14(10):e0222773 [FREE Full text] [doi: 10.1371/journal.pone.0222773] [Medline: 31618202]
38. Web scraping policy. Office for National Statistics. 2022 Aug 24. URL: https://www.ons.gov.uk/aboutus/transparencyandgovernance/datastrategy/datapolicies/webscrapingpolicy [accessed 2022-08-24]
39. Intellectual Property Office. Exceptions to copyright. GOV.UK. 2021. URL: https://www.gov.uk/guidance/exceptions-to-copyright [accessed 2022-03-17]
40. Huang Y. MenuTracker. GitHub. 2022. URL: https://github.com/YuruHuang/MenuTracker [accessed 2022-03-17]
41. Food labels. NHS. 2020. URL: https://www.nhs.uk/live-well/eat-well/food-guidelines-and-food-labels/how-to-read-food-labels/ [accessed 2022-08-24]
42. Public Health England. Plans to cut excess calorie consumption unveiled. GOV.UK. 2018. URL: https://www.gov.uk/government/news/plans-to-cut-excess-calorie-consumption-unveiled [accessed 2021-12-16]
43. Saunders P, Saunders A, Middleton J. Living in a 'fat swamp': exposure to multiple sources of accessible, cheap, energy-dense fast foods in a deprived community. Br J Nutr 2015 Jun 14;113(11):1828-1834. [doi: 10.1017/S0007114515001063] [Medline: 25885785]
44. Dolly RZT. MenuStat UK: Establishment of a Database of The Nutritional Content of Food and Drink Served by UK Chain Restaurants and Cross-Sectional Analysis of Energy Content. Cambridge: The University of Cambridge; 2018.
45. Theis DRZ, White M. Is obesity policy in England fit for purpose? Analysis of government strategies and policies, 1992-2020. Milbank Q 2021 Mar;99(1):126-170 [FREE Full text] [doi: 10.1111/1468-0009.12498] [Medline: 33464689]
46. Robinson E, Burton S, Gough T, Jones A, Haynes A. Point of choice kilocalorie labelling in the UK eating out of home sector: a descriptive study of major chains. BMC Public Health 2019 May 28;19(1):649 [FREE Full text] [doi: 10.1186/s12889-019-7017-5] [Medline: 31138179]
47. Public Health England. Salt reduction: targets for 2024. GOV.UK. 2020. URL: https://www.gov.uk/government/publications/salt-reduction-targets-for-2024 [accessed 2021-08-10]
48. Public Health England. Sugar reduction and wider reformulation. GOV.UK. 2018. URL: https://www.gov.uk/government/collections/sugar-reduction [accessed 2022-08-24]
49. Huang Y, Burgoine T, Theis DR, Adams J. Differences in energy and nutrient content of menu items served by large chain restaurants in the USA and the UK in 2018. Public Health Nutr 2022 Jun 01:1-9. [doi: 10.1017/S1368980022001379] [Medline: 35642073]
50. Online food delivery. Statista. 2022. URL: https://www.statista.com/outlook/dmo/eservices/online-food-delivery/worldwide#global-comparison [accessed 2022-07-04]
51. Ellison B, McFadden B, Rickard B, Wilson N. Examining food purchase behavior and food values during the COVID-19 pandemic. Appl Econ Perspect Policy 2020 Nov 04;43(1):58-72. [doi: 10.1002/aepp.13118]

Abbreviations

API: application programming interface
ONS: Office for National Statistics
SIC: Standard Industrial Classification

Edited by Y Khader; submitted 05.05.22; peer-reviewed by S Hua, Y Yang; comments to author 27.06.22; revised version received 22.07.22; accepted 29.07.22; published 08.09.22.

Please cite as:
Huang Y, Burgoine T, Essman M, Theis DRZ, Bishop TRP, Adams J
Monitoring the Nutrient Composition of Food Prepared Out-of-Home in the United Kingdom: Database Development and Case Study
JMIR Public Health Surveill 2022;8(9):e39033
URL: https://publichealth.jmir.org/2022/9/e39033
doi:10.2196/39033
PMID:36074559

Original Paper

A Standard-Based Citywide Health Information Exchange for Public Health in Response to COVID-19: Development Study

Bala Hota[1], MD, MPH; Paul Casey[2], MD; Anne F McIntyre[3], PhD, MPH; Jawad Khan[4], BS; Shafiq Rab[4], MD, MPH; Aneesh Chopra[5], BA, MPP; Omar Lateef[2], DO; Jennifer E Layden[3], MD, PhD

[1]Tendo Systems, Inc, Hinsdale, IL, United States

[2]Rush University Medical Center, Chicago, IL, United States

[3]Centers for Disease Control and Prevention, Atlanta, GA, United States

[4]Wellforce, Burlington, MA, United States

[5]CareJourney, Washington, DC, United States

Corresponding Author:
Bala Hota, MD, MPH
Tendo Systems, Inc
5617 S Oak St
Hinsdale, IL, 60521
United States
Phone: 1 708 362 3767
Email: bala.hota@gmail.com

Abstract

Background: Disease surveillance is a critical function of public health, provides essential information about the disease burden and the clinical and epidemiologic parameters of disease, and is an important element of effective and timely case and contact tracing. The COVID-19 pandemic demonstrates the essential role of disease surveillance in preserving public health. In theory, the standard data formats and exchange methods provided by electronic health record (EHR) meaningful use should enable rapid health care data exchange in the setting of disruptive health care events, such as a pandemic. In reality, access to data remains challenging and, even if available, often lacks conformity to regulated standards.

Objective: We sought to use regulated interoperability standards already in production to generate awareness of regional bed capacity and enhance the capture of epidemiological risk factors and clinical variables among patients tested for SARS-CoV-2. We described the technical and operational components, governance model, and timelines required to implement the public health order that mandated electronic reporting of data from EHRs among hospitals in the Chicago jurisdiction. We also evaluated the data sources, infrastructure requirements, and the completeness of data supplied to the platform and the capacity to link these sources.

Methods: Following a public health order mandating data submission by all acute care hospitals in Chicago, we developed the technical infrastructure to combine multiple data feeds from those EHR systems—a regional data hub to enhance public health surveillance. A cloud-based environment was created that received ELR, consolidated clinical data architecture, and bed capacity data feeds from sites. Data governance was planned from the project initiation to aid in consensus and principles for data use. We measured the completeness of each feed and the match rate between feeds.

Results: Data from 88,906 persons from CCDA records among 14 facilities and 408,741 persons from ELR records among 88 facilities were submitted. Most (n=448,380, 90.1%) records could be matched between CCDA and ELR feeds. Data fields absent from ELR feeds included travel histories, clinical symptoms, and comorbidities. Less than 5% of CCDA data fields were empty. Merging CCDA with ELR data improved race, ethnicity, comorbidity, and hospitalization information data availability.

Conclusions: We described the development of a citywide public health data hub for the surveillance of SARS-CoV-2 infection. We were able to assess the completeness of existing ELR feeds, augment those feeds with CCDA documents, establish secure transfer methods for data exchange, develop a cloud-based architecture to enable secure data storage and analytics, and produce dashboards for monitoring of capacity and the disease burden. We consider this public health and clinical data registry as an informative example of the power of common standards across EHRs and a potential template for future use of standards to improve public health surveillance.

(JMIR Public Health Surveill 2022;8(9):e35973) doi:10.2196/35973

KEYWORDS

public health; informatics; surveillance; disease surveillance; epidemiology; health data; electronic health record; data hub; acute care hospital; COVID-19; pandemic; data governance

Introduction

Since the emergence of SARS-CoV-2, the virus that causes COVID-19, in Wuhan, China [1], a global pandemic was declared in 2020 [2], and widespread and sustained transmission was observed across the United States. As of March 23, 2022, there were 79,621,004 cases and 971,422 deaths in the United States [3].

Disease surveillance is a critical function of public health in the United States. It provides essential information about the disease burden and the clinical and epidemiologic parameters of disease and is an important element to conduct effective and timely case investigations. In addition to individual and aggregated patient data, the pandemic has required careful monitoring of health care capacity and utilization to ensure clinical care needs are met, especially in times of surges of cases that have strained capacity; ongoing surveillance of case counts can aid this need to be met

Support for the public health functions of the surveillance and epidemiology of diseases has been embedded in key national informatics initiatives in the United States for nearly 2 decades through federal programs and mandates. These efforts have included syndromic surveillance [4], electronic laboratory reporting (ELR) [5] in the meaningful use program [6] (the program in which health systems were empowered to implement electronic health records [EHRs] through multiple federal incentives), and the growth of the National Healthcare Safety Network (NHSN) [7]. These programs created linkages between hospitals, commercial laboratories, and public health across the United States that collect and organize data, often through EHR and order workflows in order to improve the timeliness and completeness of reporting.

In theory, the standard data formats and exchange methods provided by the meaningful use program should enable rapid health care data exchange in the setting of disruptive health care events, such as a pandemic. In reality, access to data remains challenging and, even if available, often lacks conformity to regulated standards [8]. The current COVID-19 pandemic revealed gaps in data liquidity (ie, data entered into a system at 1 point should be usable at other points downstream in the system) and difficulty in quickly gathering information by key stakeholders, such as policy makers and public health authorities [9].

In the early phase of the pandemic, the Chicago Department of Public Health (CDPH) and health systems in Chicago tried to address 2 major challenges: first, the ability to efficiently submit necessary clinical data elements for SARS-CoV-2–tested patients, and second, the ability to capture aggregated capacity data for resource planning in an administratively efficient manner. Despite significant EHR investments among the city's hospitals and health systems, the inability of EHR systems to automate delivery of important data elements to public health surveillance systems meant that providers and health systems had to manually enter data into the public health reporting system. However, the high volume of patients and significant work demands on health systems limited timely and complete manual data entry. As the pandemic unfolded, multiple agencies requested bed and surge capacity information, including the NHSN, the Federal Emergency Management Agency (FEMA), the National Guard, and the Illinois Department of Public Health (IDPH), all with slightly varying data element definitions (Multimedia Appendix 1). Locally, an important aspect of capturing the resource capacity data was to monitor the surge capacity and assist with coordination of resources. The multiple reporting requirements, varying definitions, and limited mechanisms for automated, real-time submission of key resource metrics, such as bed capacities, raised concern about the ability to locally monitor the resource capacity across our systems.

In response to these challenges, the CDPH issued a public health order requiring electronic data sharing and partnered with the Rush University Medical Center to leverage existing health information technology (HIT) infrastructure for COVID-19 to develop a platform for data exchange. In this paper, we describe the technical and operational components, governance model, and timelines required to implement the public health order that mandated electronic reporting of data from EHRs among hospitals in the Chicago jurisdiction. We also evaluate the data sources, infrastructure requirements, and the completeness of data supplied to the platform and the capacity to link these sources. As an example of clinically relevant fields of interest for reporting, we compared available fields in data feeds to the *Human Infection with 2019 Novel Coronavirus Case Report* (also referred to as the COVID-19 Persons Under Investigation [PUI] Form) [10]. Finally, we reflect on success factors that enabled the rapid implementation of data sharing in the region.

Methods

Setting

This project was conducted by the CPDH in partnership with the Rush University Medical Center, which was made a third-party agent of the CDPH to develop and support the analytics and provide the infrastructure to support the data collection.

Public Health Notice

On April 6, 2020, the CDPH issued public health order 2020-4 requiring hospitals in Chicago to share EHR data with the CDPH [9] for all patients tested for SARS-CoV-2. The order outlined a constrained set of data to be submitted for all SARS-CoV-2–tested patients. This order was disseminated through the CDPH's clinical Health Alert Network (HAN), posted on the department's website, and shared with city hospital leadership on calls. The CDPH constituted a governance

committee comprising medical directors and informaticists from hospital systems in Chicago, Illinois.

Data Feeds

ELR feeds were accessed from the Illinois National Electronic Disease Surveillance System (I-NEDSS) to provide baseline information on laboratory-confirmed cases in the city. As a result of meaningful use mandates, each positive test result for COVID-19 obtained from diagnostic laboratories and present in EHRs was being sent to I-NEDSS. These feeds contained records of patient demographics, test name, results, and dates of service and were being submitted by 88 facilities. To meet public health order 2020-4, Chicago hospitals were provided with multiple mechanisms to submit consolidated clinical data architecture (CCDA) records for SARS-CoV-2–tested patients. This included (1) a report via a secure mailbox that used the DIRECT protocol [11], a recognized data standard by the Office of the National Coordinator for Health Information Technology (ONC) for the 1-way transmission of EHRs to a centralized instance of the Epic EHR [12] for the city, or (2) a report directly to the CDPH's instance of the Microsoft Azure cloud [13] via DIRECT or an application programming interface (API), which could receive and accept the CCDA records. In either case, the CCDAs were parsed into a database within a dedicated tenant in Azure for analytics. Additionally, a third data set of NHSN patient safety and hospital capacity was included, where hospitals were asked to either enter into a Research Electronic Data Capture (REDCap) database or send electronically to the Azure tenant. All data feeds were operational data (ie, used for purposes of public health reporting or obtained from electronic records used in patient care) and contained protected health information (PHI).

Technical Evaluation

At the project start, we developed the requirements of a solution to collect data from sites and produced the required analytics. At the start of this project, the accepted method for COVID-19 case-related data to be submitted to health departments was the Person Under Investigation (PUI) surveillance form. These forms were available as paper forms or via survey instruments hosted on a RedCap survey tool by the IDPH. Entry was time-consuming and often incomplete due to clinical burdens. Responsibility for form completion rested with infection control practitioners or clinical staff and was considered neither timely nor complete due to competing tasks for these individuals. We evaluated the gap between the existing COVID-19 PUI form fields and the electronic data elements available in federal standard–based data feeds and developed a crosswalk of reporting requirements to ensure that the data set could function as a reporting gateway for sites and reduce the burden of reporting. Feeds evaluated were ELR, CCDA, and Fast Healthcare Interoperability Resource (FHIR, pronounced as *fire*) fields. Missingness and usefulness were evaluated among CCDA and ELR feeds. Missingness refers to whether data are present in the field. Usefulness refers to clean and complete information in the data field. Data were labeled not useful if any of the following were present in their respective fields: "unknown" in race, ethnicity, or other string fields; the presence of PO boxes, unknown, homeless, or not applicable (N/A) for an address; the absence of a telephone number, an implausible number (eg, 111-1111 or 999-999-9999), or not enough numbers for the phone number; and less than 5 digits or 99999, 00000, or text (eg, "UUUUU") for zip codes. Records were deduplicated using name and date of birth. The record match rate between CCDA and ELR data feeds was assessed: a deterministic match process using an exact match of characters in 12 different combinations ("keys") of last name, first name, and date of birth was implemented, which has been shown to have efficacy in matching using surveillance registries [14]. We did not attempt to resolve close matches. For the 3 fields demonstrating the most missing or low-quality data (ie, race, ethnicity, and telephone number), we examined the additional completeness to ELR feeds by augmenting with CCDA data; this was accomplished by using complete and useful data when ELR feeds were missing for an individual person.

Ethics

This investigation was part of the ongoing public health response to COVID-19. This activity was reviewed by the Centers for Disease Control and Prevention (CDC) and was conducted consistent with applicable federal law and CDC policy (see, eg, 45 C.F.R. part 46.102(l)(2); 21 C.F.R. part 56; 42 U.S.C. §241(d); 5 U.S.C. §552a; 44 U.S.C. §3501 et seq.).

Results

State Surveillance System Baseline Reporting

In Chicago, a significant proportion of reported cases of SARS-CoV-2 infections are reported through ELR. As of June 30, 2020, ELR alone provided 73.7% of cases, while ELR combined with other modalities (eg, submission of a case report from a hospital or health care provider to I-NEDSS) accounted for 94% of reported cases. ELR data reported key fields requested in the COVID-19 PUI form (Table 1) but not all; data fields routinely absent from ELR feeds included travel histories, clinical symptoms, and comorbidities.

Table 1. Crosswalk table to compare coverage of Human Infection with 2019 Novel Coronavirus Case Report form fields and ELR[a], the CCDA[b], and FHIR[c].

CDC[d] PUI[e] form field	Covered in CCDA	Covered in ELR	Covered in FHIR	Covered in other data sources
What is the current status of this person?				
PUI: testing pending[f]	Yes (lab test and result information)	N/A[g]	Yes	N/A
PUI: tested negative[f]	Yes (lab test and result information)	N/A	Yes	N/A
Presumptive case (positive local test): confirmatory testing pending	N/A	N/A	N/A	N/A
Presumptive case (positive local test): confirmatory tested negative	N/A	N/A	N/A	N/A
Laboratory-confirmed case	Yes	Yes	Yes	N/A
Report date of PUI to CDC	N/A	N/A	N/A	N/A
Report date of case to CDC	N/A	N/A	N/A	N/A
County of residence	Yes	Yes	Yes	N/A
State of residence	Yes	Yes	Yes	N/A
Ethnicity	Yes	Yes	Yes	N/A
Race	Yes	Yes	Yes	N/A
Sex	Yes	Yes	Yes	N/A
Date of birth	Yes	Yes	Yes	N/A
Age	Yes	Yes	Yes	N/A
Was the patient hospitalized? Date?	Yes	N/A	Yes	ADT[h] or Census data
Was the patient admitted to the ICU[i]?	N/A	N/A	Yes	ADT or Census data
Did the patient receive mechanical ventilation (MV) or intubation? Days of MV?	N/A	N/A	N/A	Custom report
Did the patient receive extracorporeal membrane oxygenation (ECMO)?	N/A	N/A	N/A	Custom report
Did the patient die as a result of this illness? Date?	Yes	N/A	Yes	ADT
Date of first positive specimen collection	Yes	Yes	Yes	N/A
Did the patient develop pneumonia?	Yes	N/A	Yes	N/A
Did the patient have acute respiratory distress syndrome?	Yes	N/A	Yes	N/A
Did the patient have another diagnosis/etiology for their illness?	N/A	N/A	N/A	N/A
Did the patient have an abnormal chest X-ray?	N/A	N/A	Yes[f]	N/A
Symptoms present during course of illness: (symptomatic/asymptomatic/unknown)	N/A	N/A	N/A	N/A
Symptom onset date	N/A	N/A	N/A	N/A
Symptom resolution date	N/A	N/A	N/A	N/A
Is the patient a health care worker in the United States?	N/A	N/A	N/A	N/A
Does the patient have a history of being in a health care facility (as a patient worker or visitor) in China?	N/A	N/A	N/A	N/A
In the 14 days prior to illness onset, did the patient have any of the following exposures (check all that apply)?				
Travel to Wuhan	N/A	N/A	N/A	N/A
Travel to Hubei	N/A	N/A	N/A	N/A
Travel to mainland China/other non-US country	N/A	N/A	N/A	N/A

CDC[d] PUI[e] form field	Covered in CCDA	Covered in ELR	Covered in FHIR	Covered in other data sources
Community contact with another lab-confirmed COVID-19 case	N/A	N/A	N/A	N/A
Any health care contact with another lab-confirmed COVID-19 case (patient/visitor/health care worker [HCW])	N/A	N/A	N/A	N/A
Exposure to a cluster of patients with severe acute lower respiratory distress of unknown etiology	N/A	N/A	N/A	N/A
Household contact with another lab- confirmed COVID-19 case	N/A	N/A	N/A	N/A
Animal exposure	N/A	N/A	N/A	N/A
If the patient had contact with another COVID-19 case, was this person a US case?	N/A	N/A	N/A	N/A
Under what process was the PUI or case first identified (check all that apply)?				
Clinical evaluation leading to PUI determination	N/A	N/A	Yes[f]	N/A
Contact tracing of the patient	N/A	N/A	N/A	N/A
Routine surveillance	N/A	N/A	N/A	N/A
Epidemic Information Exchange (EpiX) notification of travelers, if checked	N/A	N/A	N/A	N/A
Unknown	N/A	N/A	N/A	N/A
Other (specify)	N/A	N/A	N/A	N/A
Symptoms				
Fever >100.4°F (38°C)	N/A	N/A	Yes[f]	N/A
Subjective fever (felt feverish)	N/A	N/A	Yes[f]	N/A
Chills	N/A	N/A	Yes[f]	N/A
Muscle aches (myalgia)	N/A	N/A	Yes[f]	N/A
Runny nose (rhinorrhea)	N/A	N/A	Yes[f]	N/A
Sore throat	N/A	N/A	Yes[f]	N/A
Cough (new onset or worsening of chronic cough)	N/A	N/A	Yes[f]	N/A
Shortness of breath (dyspnea)	N/A	N/A	Yes[f]	N/A
Nausea or vomiting	N/A	N/A	Yes[f]	N/A
Headache	N/A	N/A	Yes[f]	N/A
Abdominal pain	N/A	N/A	Yes[f]	N/A
Diarrhea (≥3 loose/looser-than-normal stools/24-hour period)	N/A	N/A	Yes[f]	N/A
Other	N/A	N/A		N/A
Pre-existing medical conditions				
Chronic lung disease (asthma/emphysema/chronic obstructive pulmonary disease [COPD])	Yes	N/A	Yes	N/A
Diabetes mellitus	Yes	N/A	Yes	N/A
Cardiovascular disease	Yes	N/A	Yes	N/A
Chronic renal disease	Yes	N/A	Yes	N/A
Chronic liver disease	Yes	N/A	Yes	N/A
Immunocompromised condition	Yes	N/A	Yes	N/A
Neurologic/neurodevelopmental intellectual disability	Yes	N/A	Yes	N/A

CDC[d] PUI[e] form field	Covered in CCDA	Covered in ELR	Covered in FHIR	Covered in other data sources
Other chronic diseases	Yes	N/A	Yes	N/A
If female, currently pregnant	N/A	N/A	N/A	N/A
Current smoker	Yes	N/A	Yes	N/A
Former smoker	Yes	N/A	Yes	N/A
Respiratory diagnostic testing test (respiratory virus testing panel information)	Yes	N/A	Yes	N/A
Specimens for COVID-19 testing				
Nasopharyngeal swab/oropharyngeal swab/sputum/other (specify)	N/A	N/A	N/A	N/A

[a]ELR: electronic laboratory reporting.

[b]CCDA: consolidated clinical document architecture.

[c]FHIR: Fast Healthcare Interoperability Resources.

[d]CDC: Centers for Disease Control and Prevention.

[e]PUI: Person Under Investigation.

[f]If notes are shared through FHIR.

[g]N/A: not applicable.

[h]ADT: admission, discharge, and transfer.

[i]ICU: intensive care unit.

Response to the Public Health Notice

On April 6, 2020, Public Health Order 2020-4 was shared via the HAN in Chicago with all eligible institutions (ie, health systems within the Chicago City borders). The order mandated the sharing with the CDPH of 3 main data types: (1) ELR feeds of SARS-CoV-2–tested individuals, which were an existing state mandate; (2) CCDA records from hospitals for SARS-CoV-2–tested patients; and (3) NHSN capacity module reporting, which was asked to be sent centrally to the CDPH. These data were requested to be sent at a minimum once per day by 10:00 a.m. US Central Time. Sites also provided contact information for key Rush University Medical Center personnel who were leading the implementation. A series of calls with hospital technical staff were conducted by the Rush University Medical Center chief information officer to introduce the project, review the rationale, and describe technical approaches.

An Azure-hosted and isolated environment was established, with 5 individual modalities for connectivity, all feeding into a centralized data hub from more than 40 organizations and hundreds of thousands of transactions per week. Over the next 30 days, all sites were approached to initiate data sharing; a CDPH data governance committee comprising chief medical officers and chief medical informatics officers from select institutions was created through which issues could be discussed and additional roadmaps could be generated; collaboration with Epic and Cerner EHR developers was established and mechanisms for enterprise scale sharing created; and data were sent centrally to the CDPH Azure instance.

Technical Architecture

An overview of the technical architecture of the project is shown in Figure 1 and was designed to maximize security and privacy of data, keeping the CDPH at the center of data use. At a high level, because of the tools from meaningful use adoption, connections existed between stakeholders in the system, which could support secure file sharing with the ability to choose records based on criteria. These tools included (1) standard-based representation of clinical data (eg, CCDA), (2) secure methods of data transport both within and external to EHR systems (eg, CareEverywhere within Epic, DIRECT mailboxes, and API-based authenticated pathways), and (3) existing implementation of complex public health rules within EHRs to identify cases and submit to public health (eg, ELR). Limited mapping of semantic content was required because data shared between health systems and public health used CCDA and Health Level Seven International (HL7) meaningful use standards, with content mapped to standard vocabularies before submission. Vocabularies used were HL7 race, gender, and ethnicity categories; *International Classification of Diseases, Tenth Revision* (ICD-10) and Current Procedural Terminology (CPT) codes for diagnoses and procedures; and Logical Observation Identifiers Names and Codes (LOINC) for lab test names. The cloud-based environment was Health Insurance Portability and Accountability Act (HIPAA) certified, and data were encrypted at rest and in transit. DIRECT mailboxes leveraged certificate-based encryption, and API pathways used hypertext transfer protocol secure (https).

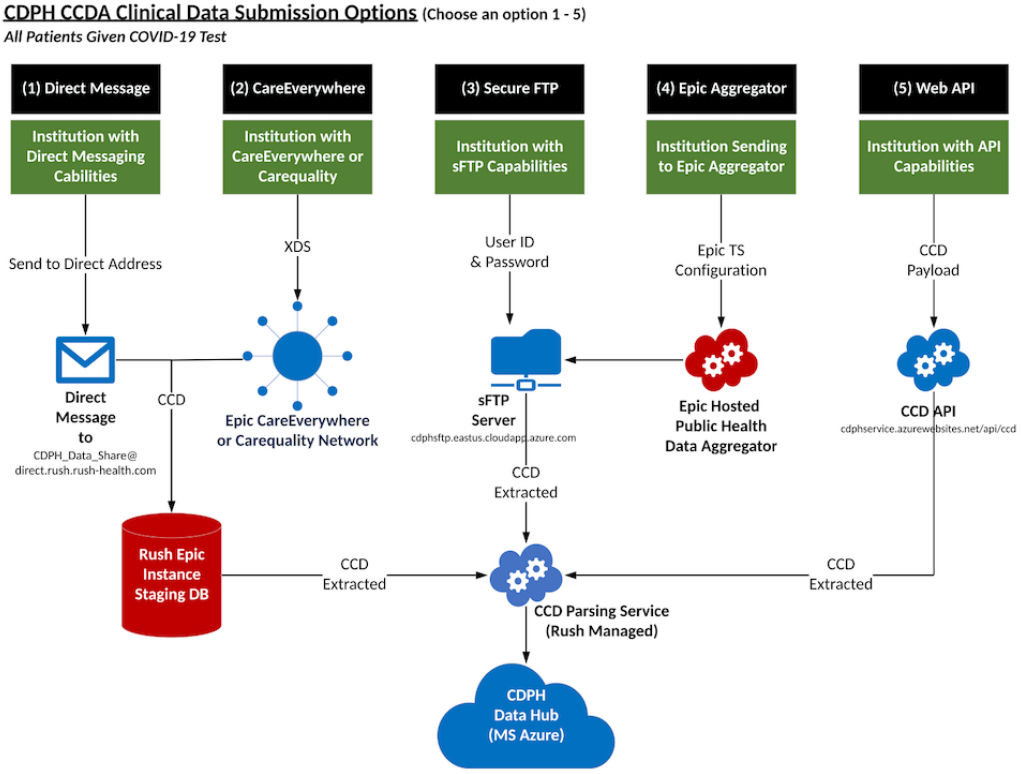
A cloud-based environment was created that was totally isolated from the Rush University Medical Center EHR instance and patient records. This environment was built to support over 40 organizations within the city of Chicago and designed to scale across public health departments.

ELR data feeds were the most straightforward to use in the model, as existing connections between hospital systems were present for communicable disease reporting. Hospitals were required to implement new logic at the outset of SARS-CoV-2 infections in Chicago to identify and report lab-identified cases of COVID-19 to the CDPH and tested patients as those are

PUIs. ELR feeds are submitted to the state public health agency, which makes these available to the CDPH and local health departments.

To isolate data, the Rush University Medical Center created an isolated Azure Data Repository, including Microsoft Azure SQL Warehouse, and a CosmosDB for survey forms data was created. We found that not all cross-enterprise document sharing (XDS) and DIRECT messages could avoid our EHR instance, so we needed to identify a way to enforce separation of data. We addressed this by pulling data from the Epic staging area. In addition, infrastructure components were created that included an XDS service server, DIRECT message communication, a continuity of care document (CCD) to the FHIR service, and integration with Epic via a community health aggregator. Google Apigee handled the API layer, and services were handled behind Apigee for token control. Data collection via manual entries was handled via REDCap forms with integration via the API into the Azure environment.

Figure 1. High-level architecture of the CDPH data hub CCDA submission options. API: application programming interface; CCD: continuity of care document; CCDA: consolidated clinical data architecture; CDPH: Chicago Department of Public Health; sFTP: secure file transfer protocol; TS: technical services; XDS: cross-enterprise document sharing.



Governance

Data governance was planned from the project beginning to aid in consensus and principles for data use. Although the local health department, with its public health orders, was a necessary recipient and data user, participants recognized the value of a larger sharing initiative, plus site participation to engage on use cases and mechanisms to leverage the information. The governance committee comprised the chief medical officer (CMO), the chief medical informatics officer (CMIO), or the technical lead from each of the 12 sites. These leaders also brought content and guidance back to site participants and sought to bridge varying degrees of internal technical capabilities among systems. The committee met weekly and helped to build trust among participating sites. General principles were modeled after rules implemented for use of Centers for Medicare & Medicaid Services (CMS) data [15] and were established among sites through this committee. These principles were:

- Openness: promoting and facilitating the open sharing of knowledge about COVID-19 data
- Communication: promoting partnerships across the region to eliminate duplication of effort, a source of truth for regional data that may enable reducing administrative burden, and a valuable regional and national resource
- Accountability: ensuring compliance with approved data management principles and policies and understanding the objectives of current and future strategic or programmatic initiatives and how they impact, or are impacted by, existing data management principles and policies and current privacy and security protocols

Reporting of Bed, Supply, and Clinical Capacity

Metrics mandated for reporting to multiple agencies and groups for Chicago hospitals at the time of the hub creation are shown in Multimedia Appendix 1. In this inventory, over 100 measures to 4 systems were required: the NHSN, EMResource, FEMA, and the Illinois National Guard. The systems measure bed usage, emergency department (ED) usage, ventilator usage, supply usage and need, and laboratory testing. Of note, 57 different bed usage measures alone exist among the 4 systems. Although metrics shown had similar definitions, these still require separate administrative efforts for the data collection and reporting.As

of July 31, 2020, 14 hospitals in Chicago were reporting data to the hub. For bed capacity reporting, 7 were reporting NHSN data through manual data submission, 2 were reporting through electronic queries from their EHRs with electronic submission to the hub, and 14 were submitting to EMResource.

Completeness of Reporting via ELR and the CCDA

Data from 86,499 persons from CCDA records among 14 facilities and 408,741 persons from ELR records among 88 facilities were submitted, representing records meeting criteria to be reported under the public health order. Table 2 shows the volume and completeness of data feeds related to COVID-19, as obtained from CCDA and ELR feeds. Patients with records in these feeds were those diagnosed through reverse transcription–polymerase chain reaction (RT-PCR) testing with a Chicago address through July 31, 2020. For those individuals with more than 1 test reported, data were deduplicated. Among individuals with CCDA records submitted, 11,491 (13.3%) had a positive test compared to 53,968 (13.2%) among ELR feeds.

We examined CCDA and ELR data fields for completeness defined as a populated (ie, nonmissing) data field and usefulness defined as clean, complete information in a data field. CCDA data provided an improvement in the quality of data available for surveillance. ELR feeds had gaps in the usability or quality of race and ethnicity data (race: n=382,097, 93.5%, nonmissing and n=215,273, 52.7%, useful; ethnicity: n=333,122, 81.5%, nonmissing and n=165,715, 49.7%, useful). The CCDA was highly complete with <5% missing information in data fields for all records types except for patient phone numbers. In addition, 99.2% of CCDA data was nonmissing for both race (n=85,794) and ethnicity (n=85,799), and 82.5% of CCDA data was useful for race (n=71,345) and 79.2% for ethnicity (n=68,507). The CCDA, although covering fewer records, also had information related to encounters and hospitalization, and the presence of comorbidities.

CCDA and ELR data feeds were matched by name and date of birth among 90.6% (n=78,378) of patients in the CCDA field. With matching, some improvement in data completeness for the 3 most incomplete fields was noted for ELR data: race, ethnicity, and telephone number. Of the 78,378 matched CCDA and ELR feeds, ELR race data alone improved from 79.4% to 88.5% (n=62,232-69,365) useful data with the CCDA, while ELR ethnicity data alone improved from 58.2% to 86.7% (n=45,616-67,954) with the CCDA. Telephone number data were 78.6% (n=321,121) complete in ELR; combining the CCDA and ELR improved completeness to 80.0% (n=326,993). In addition, for the matched set, complete hospitalization and comorbidity information was present.

For presentation, data were displayed on a dashboard available for CDPH analysts, via the Microsoft Azure Power BI platform, and are shown in Figure 2. Data from the dashboard were shared to contributing hospitals over a business intelligence portal hosted by the Rush University Medical Center and via email of bed capacity reports and analytic descriptions of case counts by subgroup. Bed capacity reports aligned with bed types listed in Multimedia Appendix 1: critical care versus general medical, and overall capacity versus COVID-19 utilization.

Table 2. Completeness of data submitted via the CCDA[a] and ELR[b].

Data field	CCDA data (N=86,499), n (%)	ELR data (N=408,741), n (%)
Lab-confirmed SARS-CoV-2	11,491 (13.3)	53,968 (13.2)
Facility name/reporting lab		
Nonmissing	86,499 (100.0)	408,737 (100.0)
Useful	86,499 (100.0)	408,463 (99.9)
Patient first name		
Nonmissing	86,499 (100.0)	408,732 (100.0)
Useful	86,499 (100.0)	408,717 (100.0)
Patient last name		
Nonmissing	86,499 (100.0)	408,732 (100.0)
Useful	86,497 (100.0)	408,718 (100.0)
Patient date of birth		
Nonmissing	86,489 (100.0)	408,270 (99.9)
Useful	86,480 (100.0)	407,730 (99.9)
Patient sex (male/female/unknown)		
Nonmissing	86,416 (99.9)	408,540 (100.0)
Useful	86,405 (99.9)	398,590 (97.5)
Patient race		
Nonmissing	85,794 (99.2)	382,097 (93.5)
Useful	71,345 (82.5)	215,273 (52.7)
Patient ethnicity		
Nonmissing	85,799 (99.2)	333,122 (81.5)
Useful	68,507 (79.2)	165,715 (49.7)
Patient address		
Nonmissing	86,498 (100.0)	385,073 (94.2)
Useful	85,471 (98.8)	384,000 (93.9)
Patient city		
Nonmissing	86,499 (100.0)	408,741 (100.0)
Useful	86,499 (100.0)	408,741 (100.0)
Patient zip code		
Nonmissing	86,377 (99.9)	408,026 (99.8)
Useful	86,375 (99.9)	407,918 (99.8)
Patient home or cell phone[c]		
Nonmissing	20,712 (23.9)	321,121 (78.6)
Useful	20,712 (23.9)	319,974 (78.3)
Test name (open text field)		
Nonmissing	86,499 (100.0)	408,694 (100.0)
Useful	86,499 (100.0)	408,694 (100.0)
Logical Observation Identifiers Names and Codes (LOINC)		
Nonmissing	0	408,741 (100.0)
Useful	0	408,727 (100.0)
Test results (raw feed/open text field)		

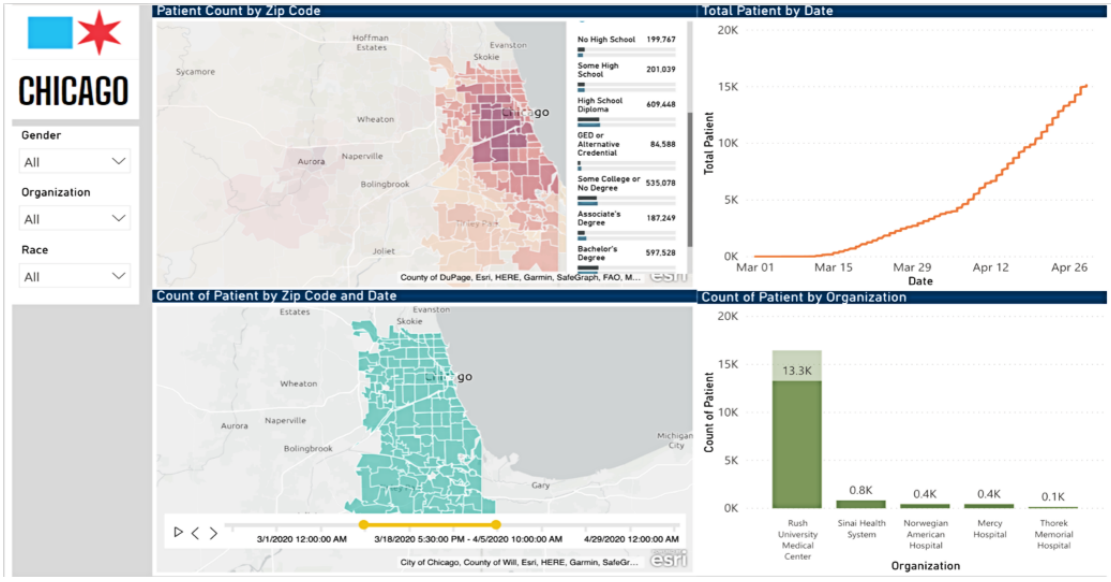
Data field	CCDA data (N=86,499), n (%)	ELR data (N=408,741), n (%)
Nonmissing	55,783 (64.5)	405,650 (99.2)
Useful	55,235 (62.1)	396,110 (96.9)
Test results (interpreted from open text field[c])		
Nonmissing	86,499 (100.0)	408,741 (100.0)
Useful	86,499 (100.0)	408,741 (100.0)
Test date		
Nonmissing	83,999 (97.1)	408,046 (99.8)
Useful	83,999 (97.1)	408,046 (99.8)
Hospitalization (yes/no)[d]		
Nonmissing	86,499 (100.0)	0
Useful	86,499 (100.0)	0
Comorbidities		
Nonmissing	86,499 (100.0)	0
Useful	86,499 (100.0)	0

[a]CCDA: consolidated clinical document architecture.

[b]ELR: electronic laboratory reporting.

[c]CCDA completeness represents at least 1 phone number from either the home or cell data fields; the ELR feed has 1 phone field, so home and fields cell are not differentiated. “Nonmissing” refers to a populated data field. “Useful” refers to clean, complete information in a data field. Data were labeled not useful if any of the following were present in their respective fields: “unknown” in race, ethnicity, address, or other string fields; for address, the presence of PO boxes, unknown, homeless, or N/A; for phone, an implausible number (eg, 111-1111 or 999-999-9999), or less than 10 numbers; and for zip code, less than 5 digits or 99999, 00000, or letters (eg, “UUUUU”).

Figure 2. Epidemiologic dashboards for assessment of outbreak, CDPH data hub. CDPH: Chicago Department of Public Health.



Discussion

Principal Findings

In this report, we described the development of a citywide public health data hub for the surveillance of SARS-CoV-2 infection in Chicago, Illinois. We were able to assess the completeness of existing ELR feeds, augment these feeds with CCDA documents, establish secure transfer methods for data exchange, develop a cloud-based architecture to enable secure data storage and analytics, and produce meaningful dashboards for the monitoring of capacity and disease burden.

An underlying need in public health that drove our work was an aim to improve the automation, completeness, and usefulness of data submitted to public health agencies. The work builds on the known utility of ELR with improved data quality. ELR, or the submission electronically of laboratory tests to a public health department through implementation of business logic for detection, has been found in multiple studies to improve the timeliness and completeness of reporting [16-20] at potentially

lower costs [21]. A review prior to widespread electronic reporting use found that despite legal mandates for reporting, passive surveillance yielded completeness rates of 23%-81% for communicable diseases with higher rates for active surveillance [22] and timeliness of reporting between 10 and 13 days after laboratory result dates [23]. ELR systems have improved the reporting of data to public health for surveillance, with the volume and timeliness of reporting improving 2.3-4.4-fold and 3.8-7.9 days earlier, respectively [24]. ELR has been a major advance in that it can improve the completeness of reporting over what is found through passive surveillance [21,25].

ELR data have been hampered by ongoing issues with completeness. In prior reports, ELR data have been found to vary in their completeness: the completeness of fields reported via ELR within basic HL7 v2.x messages ranges from 38% (race) to 98% (date of birth) [25]. To increase completeness, improvements have been proposed: (1) increase in mandatory fields in ELR HL7 v2.x messages [24]; (2) augmenting of ELR feeds with data from a health information exchange, which improves completeness for race to 60% [25]; and (3) electronic case report forms that are completed through either automated data capture or manual completion [26]. Significant limitations in case reporting have been identified during the COVID-19 pandemic, including limited data on key variables such as age, race/ethnicity, hospitalization, and intensive care unit (ICU) status [27].

We also found that ELR data do not provide all the information needed for adequate case investigation. Demographic and risk factor information may not be complete in the HL7 feeds for ELR, and case report forms continue to play a critical role in the work of public health practice. Additionally, comorbid conditions, a significant predictor of disease outcome, are not captured. We found that CCDA data have a broader set of clinical fields and have the advantage of providing valuable comorbidity information. Although only small improvements in completeness were achieved, a high match rate to ELR data makes the CCDA a compelling addition to ELR to improve the analytic power of public health data sets. The CCDA had some fields that remained incomplete, indicating that data capture and sharing at the source remain crucial issues for use of these data.

Initiatives to standardize and automate case report form completion have been developed [28] and piloted [29], which have shown promise at reducing the time to complete reporting. Similar to our results, others have found that health information exchanges show value in prepopulating key elements for reporting through automated matching and searches in the patient record [30]. The use of FHIR [31] may provide an additional path for automated public health case reporting and reducing the administrative burden through API-based connections between public health and EHR systems. An example workflow could be the submission of case data via traditional ELR methods to public health agencies, followed by a “pull” of information from EHR systems by public health via FHIR API calls to complete a record. When combined with an ELR-based trigger for a case (eg, sexually transmitted infection cases), an app that executes FHIR-based queries could complete an electronic case report form in 85% of cases [26]. Additionally, all the key components of FHIR-based workflows for public health reporting are often in place [32]. In the recent past, alignment on the US Core Data for Interoperability (USCDI), with use of FHIR standards, has created a baseline for fields, vocabularies, and content that may enhance existing mandates from meaningful use. Our technical architecture supports the use of mandated as well as available data to create a unified public health data set in the data hub.

A feature of our solution is that it supports the central role of local health departments in data aggregation and reporting. An important component of the public health response in many communities is “home rule” for public health agencies [33] or local jurisdiction and control of policy and approach for local health departments. Home rule laws empower local governments to address public health issues and fill gaps in the patchwork of the national and state-based public health response [33]. In the current pandemic, robust local responses that can enable targeted interventions and planning can allow more sophisticated preparedness planning, pandemic control, and epidemiologic analysis.

For the most efficient data exchange, standards for the structure of data sharing and the semantic representation of information are critical. In this context, the technical and nontechnical handshakes and handoffs related to data are key factors in successful programs. In this setting, technical handshakes are the trust relationships between systems to enable data sharing: the ability to use both authenticated API-based transfers and DIRECT mailbox shares accelerated time to implementation for the project. Technical handoffs were the ability to have seamless data parsing because of robust standards implemented via meaningful use. Given the greater coverage of fields in the COVID-19 PUI form by CCDA files, the ability to leverage the CCDA to increase the completeness of overall COVID-19 PUI reporting is a sign of the value of federal standards for clinical data interchange.

Of more importance were the nontechnical handshakes (ie, relationship building and the development of consensus among institutions to enable sharing of data) and handoffs (ie, the partnerships between public and private entities). A data governance committee was essential to promote trust and enabled the scaling of the program to new data sets and deeper information within sets. At a time of a surge in COVID-19 cases, a private and academic partner (Rush University Medical Center) with the technical capacity was able to rapidly implement a solution. Three implications emerge from the system developed in Chicago. First, relationships and collaborations were critical in the setting of the pandemic to ensure success. Second, the role of public health in driving adoption through the use of mandates was also critical. Finally, the existence of standards and API-based data exchange accelerated adoption in the region.

Limitations

Our efforts were subject to several limitations. First, the solution that was implemented was used in a single public health jurisdiction and was not deployed to multiple locations. We believe that the use of file types that are widely available through

federal mandates (CCDA and ELR data) suggests that our approach could be scalable to multiple health departments, but further investigation is required. An additional limitation was the use of a public health mandate to encourage engagement and participation. Without a requirement for data sharing, lower rates of data sharing likely would have occurred. Finally, although we made significant process in our effort at regional data exchange for public health purposes, much work remains nationally to facilitate scalable data sharing. To avoid the challenges faced in this pandemic with data liquidity, more work is needed for automation of data collection and networks of "on-the-ready" data sharing built outside of pandemics.

Conclusion

We consider this public health and clinical data hub to be an informative example of how common standards across electronic records can be used to create a more complete surveillance record for public health. This report may be a potential template for future extension of the use of standards to improve public health surveillance. Through merging of data, small improvements in completeness were achieved, particularly for comorbidity and hospitalization information for COVID-19 surveillance. A reduction in the administrative burden in reporting remains a goal but will require more broad changes to the US reporting infrastructure.

Acknowledgments

We acknowledge the cooperation of hospital representatives in the city of Chicago. Illinois, who participated in this initiative, collaborated in data governance, and shared their data.

Authors' Contributions

The findings and conclusions in this report are those of the author(s) and do not necessarily represent the official position of the Centers for Disease Control and Prevention (CDC).

Conflicts of Interest

AC serves as the President of CareJourney.

Multimedia Appendix 1
List of measures and agencies with mandated reporting in April 2020 for COVID-19 in Chicago, Illinois.
[DOCX File , 33 KB - publichealth_v8i9e35973_app1.docx]

References

1. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med 2020 Feb 20;382(8):727-733 [FREE Full text] [doi: 10.1056/nejmoa2001017]
2. World Health Organization. WHO Director-General's Opening Remarks at the Media Briefing on COVID-19 - 11 March 2020. 2020. URL: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 [accessed 2020-04-27]
3. Centers for Disease Control and Prevention. COVID Data Tracker. 2022. URL: https://covid.cdc.gov/covid-data-tracker/#county-view?list_select_state=all_states&list_select_county=all_counties&data-type=Cases [accessed 2022-03-23]
4. Centers for Disease Control and Prevention. Syndromic Surveillance (SS): Meaningful Use. 2020. URL: https://www.cdc.gov/ehrmeaningfuluse/Syndromic.html [accessed 2020-04-25]
5. Centers for Disease Control and Prevention. Electronic Laboratory Reporting (ELR): Meaningful Use. 2020. URL: https://www.cdc.gov/ehrmeaningfuluse/elr.html [accessed 2020-04-25]
6. Centers for Disease Control and Prevention. Guides: Meaningful Use. 2020. URL: https://www.cdc.gov/ehrmeaningfuluse/guides.html [accessed 2020-04-25]
7. Arnold K, Thompson N. Building data quality and confidence in data reported to the National Healthcare Safety Network. Infect Control Hosp Epidemiol 2012 May;33(5):446-448 [FREE Full text] [doi: 10.1086/665311] [Medline: 22476269]
8. Holmgren A, Apathy N, Adler-Milstein J. Barriers to hospital electronic public health reporting and implications for the COVID-19 pandemic. J Am Med Inform Assoc 2020 Aug 01;27(8):1306-1309 [FREE Full text] [doi: 10.1093/jamia/ocaa112] [Medline: 32442266]
9. Strengthening the Public Health Infrastructure: The Role of Data in Controlling the Spread of COVID-19. 2020. URL: https://opcast.org/OPCAST_Public_Health_Data_Report_07-28-20.pdf [accessed 2020-04-25]
10. Centers for Disease Control and Prevention. Human Infection with 2019 Novel Coronavirus Case Report Form. 2020. URL: https://www.cdc.gov/coronavirus/2019-ncov/downloads/pui-form.pdf [accessed 2020-04-25]
11. The Direct Project. 2020. URL: https://www.healthit.gov/sites/default/files/hie-interoperability/the-direct-project-data-sheet.pdf [accessed 2020-04-25]
12. Epic...with the Patient at the Heart. 2020. URL: https://www.epic.com/ [accessed 2020-04-25]
13. Microsoft. Cloud Computing Services: Microsoft Azure. URL: https://azure.microsoft.com/en-us/ [accessed 2020-04-25]

14. Drobnik A, Pinchoff J, Bushnell G, Ly S, Yuan J, Varma J. Matching HIV, tuberculosis, viral hepatitis, and sexually transmitted diseases surveillance data, 2000-2010: identification of infectious disease syndemics in New York City. JPHMP 2000;20(5):506-512. [doi: 10.1097/phh.0b013e3182a95607]
15. CMS Information Systems Security and Privacy Policy. 2019. URL: https://www.cms.gov/Research-Statistics-Data-and-Systems/CMS-Information-Technology/InformationSecurity/Downloads/CMS-IS2P2.pdf [accessed 2020-04-25]
16. Centers for Disease Control and Prevention (CDC). Automated detection and reporting of notifiable diseases using electronic medical records versus passive surveillance–Massachusetts, June 2006-July 2007. Morb Mortal Wkly Rep 2008 Apr;57(14):373-376.
17. Centers for Disease Control and Prevention (CDC). Potential effects of electronic laboratory reporting on improving timeliness of infectious disease notification–Florida, 2002-2006. Morb Mortal Wkly Rep 2008 Dec;57(49):1325-1328.
18. Centers for Disease Control and Prevention (CDC). Effect of electronic laboratory reporting on the burden of Lyme disease surveillance–New Jersey, 2001-2006. Morb Mortal Wkly Rep 2008 Jan;57(2):42-45.
19. Nguyen TQ, Thorpe L, Makki HA, Mostashari F. Benefits and barriers to electronic laboratory results reporting for notifiable diseases: the New York City Department of Health and Mental Hygiene experience. Am J Public Health 2007 Apr;97(Supplement_1):S142-S145. [doi: 10.2105/ajph.2006.098996]
20. Swaan C, van den Broek A, Kretzschmar M, Richardus JH. Timeliness of notification systems for infectious diseases: a systematic literature review. PLoS One 2018 Jun 14;13(6):e0198845 [FREE Full text] [doi: 10.1371/journal.pone.0198845] [Medline: 29902216]
21. Samoff E, DiBiase L, Fangman MT, Fleischauer AT, Waller AE, MacDonald PD. We can have it all: improved surveillance outcomes and decreased personnel costs associated with electronic reportable disease surveillance, North Carolina, 2010. Am J Public Health 2013 Dec;103(12):2292-2297. [doi: 10.2105/ajph.2013.301353]
22. Doyle T, Glynn M, Groseclose S. Completeness of notifiable infectious disease reporting in the United States: an analytical literature review. Am J Epidemiol 2002 May 01;155(9):866-874. [doi: 10.1093/aje/155.9.866] [Medline: 11978592]
23. Jajosky RA, Groseclose SL. Evaluation of reporting timeliness of public health surveillance systems for infectious diseases. BMC Public Health 2004 Jul 26;4(1):29 [FREE Full text] [doi: 10.1186/1471-2458-4-29] [Medline: 15274746]
24. Rajeev D, Staes C, Evans R, Mottice S, Rolfs R, Samore M. Development of an electronic public health case report using HL7 v2.5 to meet public health needs. JAMIA 2010;17(1):34-41. [doi: 10.1197/jamia.m3299]
25. Dixon BE, McGowan JJ, Grannis SJ. Electronic laboratory data quality and the value of a health information exchange to support public health reporting processes. 2011 Presented at: AMIA Annual Symposium Proceedings; October 22-26, 2011; Washington, DC p. 322-330.
26. Dixon B, Taylor D, Choi M, Riley M, Schneider T, Duke J. Integration of FHIR to facilitate electronic case reporting: results from a pilot study. Stud Health Technol Inform 2019 Aug;264:940-944.
27. CDC COVID-19 Response Team. Severe outcomes among patients with coronavirus disease 2019 (COVID-19) - United States, February 12-March 16, 2020. Morb Mortal Wkly Rep 2020 Mar 27;69(12):343-346 [FREE Full text] [doi: 10.15585/mmwr.mm6912e2] [Medline: 32214079]
28. Mac Kenzie WR, Davidson A, Wiesenthal A, Engel J, Turner K, Conn L, et al. The promise of electronic case reporting. Public Health Rep 2016 Nov;131(6):742-746 [FREE Full text] [doi: 10.1177/0033354916670871] [Medline: 28123218]
29. Whipple A, Jackson J, Ridderhoff J, Nakashima AK. Piloting electronic case reporting for improved surveillance of sexually transmitted diseases in Utah. Online J Public Health Inform 2019 Sep 20;11(2):e7 [FREE Full text] [doi: 10.5210/ojphi.v11i2.9733] [Medline: 31632601]
30. Painter I, Revere D, Gibson PJ, Baseman J. Leveraging public health's participation in a health information exchange to improve communicable disease reporting. Online J Public Health Inform 2017 Sep 08;9(2):e186 [FREE Full text] [doi: 10.5210/ojphi.v9i2.8001] [Medline: 29026452]
31. Index - FHIR v4.0.1. URL: https://www.hl7.org/fhir/ [accessed 2020-05-02]
32. Mishra N, Duke J, Lenert L, Karki S. Public health reporting and outbreak response: synergies with evolving clinical standards for interoperability. J Am Med Inform Assoc 2020 Jul 01;27(7):1136-1138 [FREE Full text] [doi: 10.1093/jamia/ocaa059] [Medline: 32692844]
33. McCarty KL, Nelson GD, Hodge JG, Gebbie KM. Major components and themes of local public health laws in select U.S. jurisdictions. Public Health Rep 2009 Aug 03;124(3):458-462 [FREE Full text] [doi: 10.1177/003335490912400317] [Medline: 19445424]

Abbreviations

API: application programming interface
CCDA: consolidated clinical data architecture
CDC: Centers for Disease Control and Prevention
CDPH: Chicago Department of Public Health
ED: emergency department

EHR: electronic health record
ELR: electronic laboratory reporting
FEMA: Federal Emergency Management Agency
FHIR: Fast Healthcare Interoperability Resources
HAN: Health Alert Network
HL7: Health Level Seven International
ICU: intensive care unit
IDPH: Illinois Department of Public Health
I-NEDSS: Illinois National Electronic Disease Surveillance System
NHSN: National Healthcare Safety Network
PUI: Person Under Investigation
REDCap: Research Electronic Data Capture
XDS: cross-enterprise document sharing

Edited by T Sanchez, A Mavragani, G Eysenbach; submitted 24.12.21; peer-reviewed by C El-Hayek, H Schwermer, T Karen; comments to author 25.01.22; revised version received 27.03.22; accepted 07.05.22; published 27.09.22.

Please cite as:
Hota B, Casey P, McIntyre AF, Khan J, Rab S, Chopra A, Lateef O, Layden JE
A Standard-Based Citywide Health Information Exchange for Public Health in Response to COVID-19: Development Study
JMIR Public Health Surveill 2022;8(9):e35973
URL: https://publichealth.jmir.org/2022/9/e35973
doi:10.2196/35973
PMID:35544440

Original Paper

The Relationships Between Social Media and Human Papillomavirus Awareness and Knowledge: Cross-sectional Study

Soojung Jo[1], RN, PhD; Keenan A Pituch[2], PhD; Nancy Howe[2], MSc

[1]School of Nursing, Purdue University, West Lafayette, IN, United States

[2]Edson College of Nursing and Health Innovation, Arizona State University, Phoenix, AZ, United States

Corresponding Author:
Soojung Jo, RN, PhD
School of Nursing
Purdue University
502 N University Street
West Lafayette, IN, 47907
United States
Phone: 1 7654942225
Email: soojungj@purdue.edu

Abstract

Background: Human papillomavirus (HPV) is the most common sexually transmitted infection. HPV can infect both females and males, and it can cause many cancers, including anal, cervical, vaginal, vulvar, and penile cancers. HPV vaccination rates are lower than vaccination rates within other national vaccination programs, despite its importance. Research literature indicates that people obtain health-related information from internet sources and social media; however, the association between such health-seeking behavior on social media and HPV-related behaviors has not been consistently demonstrated in the literature.

Objective: This study aims to examine the association between social media usage and HPV knowledge and HPV awareness.

Methods: This study analyzed public health data collected through the Health Information National Trends Survey (HINTS) conducted by the US National Cancer Institute. The analysis used data collected in 2020; in total, 2948 responses were included in the analysis. Six HPV-related questions were used to identify HPV awareness, HPV vaccine awareness, and HPV knowledge about HPV-related cancers. Four questions about social media usage and one question about online health information–seeking behavior were used to analyze the associations between social media usage and HPV-related behaviors. Initially, six logistic regressions were conducted using replicate weights. Based on the results, significant factors were included in a second set of regression analyses that also included demographic variables.

Results: About half of the respondents were aware of HPV (68.40%), the HPV vaccine (64.04%), and the relationship between HPV and cervical cancer (48.00%). However, fewer respondents were knowledgeable about the relationships between HPV and penile cancer (19.18%), anal cancer (18.33%), and oral cancer (19.86%). Although social media usage is associated with HPV awareness, HPV vaccine awareness, and knowledge of cervical cancer, these associations were not significant after adjusting for demographic variables. Those less likely to report HPV awareness and knowledge included older participants, males, those with a household income of less than US $20,000, those with a formal education equal to or less than high school, or those who resided in a household where adults are not fluent in English.

Conclusions: After adjusting for demographic variables, social media use was not related to HPV knowledge and awareness, and survey respondents were generally not aware that HPV can lead to specific types of cancer, other than cervical cancer. These results suggest that perhaps a lack of high-quality information on social media may impede HPV awareness and knowledge. Efforts to educate the public about HPV via social media might be improved by using techniques like storytelling or infographics, especially targeting vulnerable populations, such as older participants, males, those with low incomes, those with less formal education, or those who reside in the United States but are not fluent in English.

(JMIR Public Health Surveill 2022;8(9):e37274) doi:10.2196/37274

KEYWORDS

papillomavirus infections; vaccination; social media; health promotion; public reporting of health care data; human papillomavirus

Introduction

Human papillomavirus (HPV) is a common sexually transmitted infection [1]. HPV infection is associated with many different cancers, including anal, cervical, vaginal, vulvar, and penile cancers [2], and can infect both females and males [3]. About 39,221 cancers annually can be caused by HPV [4]. HPV infection is the cause of nearly all cervical cancers, about 90% of all anal cancers, about 75% of all vaginal cancers, about 70% of all vulvar cancers, and about 60% of all penile cancers [2]. Recent studies have shown strong relationships among HPV infection, tobacco use, and oropharynx cancers [5], which make up about 70% of all HPV-related cancers in the United States [2]. According to the American Cancer Society, it is estimated that 1,918,030 new cancers are diagnosed and 609,360 deaths are attributed to cancer annually. In 2022, 14,100 new cases of cervical cancer were diagnosed and 4280 deaths were attributed to cervical cancer [6].

Detection of HPV infection is difficult since most HPV infections are typically asymptomatic. HPV infection is easily preventable by vaccination. HPV vaccination can prevent up to 93% of cervical cancer occurrences [7]. Despite the importance of HPV vaccination, the rate of HPV vaccination in the United States remains low. Among adolescents aged 13 to 17 years, HPV vaccination completion in the United States was estimated at 54.2% in 2019 and 58.6% in 2020, a rate that is much lower than the vaccination rate for measles, mumps, and rubella or for tetanus, diphtheria, and pertussis, which all have vaccination rates of more than 90% [8].

HPV vaccination among adolescents is highly associated with parental desire and recommendation rather than a health care provider's recommendation [9,10]. Literature has shown that HPV vaccination intention among parents for their children was associated with HPV awareness and HPV knowledge [11,12]. Despite the importance of HPV knowledge, research indicates that people have little knowledge about HPV [13]. Health care providers' recommendations or information on HPV is needed to increase HPV awareness and knowledge. However, research indicates that people do not obtain information about HPV and related cancers from health care providers. Instead, HPV information is gathered from different internet sites and social media [12,14].

The internet enables people to search online for health-related information. A literature review reported that people who seek health-related information for specific diseases or for public health concerns felt emotionally supported by peer interactions that occur through social media [15]. Even though health issues are private, people who trust social media interactions are more likely to disclose their health issues over social media [16], leading to greater engagement with others about health issues and greater sharing of health-related concerns.

In terms of cancer prevention behavior, some research has shown the positive impact of social media [17,18]. People who use social media are more likely to have heard of hepatitis C virus and HPV [17]. A systematic review indicated that individuals who engage in social media have higher HPV awareness [18]. However, the results of the impact of social media were inconsistent. Social media engagement was not associated with HPV vaccine uptake [18]. While interventions using social media could improve HPV knowledge [18], it is difficult to assess whether there is an association between social media and HPV knowledge. The quality of the information available through social media mattered as well. Research showed positive encouragement for HPV vaccination on Twitter [19]. However, research that analyzed websites that mentioned HPV discovered that antivaccine content was greater than provaccine content: 50.7% of content was antivaccine and only 37.4% of content was provaccine [20]. Similarly, the majority of YouTube videos on HPV were antivaccine videos (57%), whereas only 31% were provaccine videos [21].

Antivaccine messages about HPV can negatively impact HPV awareness and HPV knowledge, and unvaccinated people are likely to contribute to an infectious disease outbreak. For example, 49 out of 110 (45%) patients infected in the California measles outbreak of 2014-2015 were unvaccinated, and 28 of them intentionally did not get vaccinated because of their health beliefs [22]. Social media may contribute to an antivaccine sentiment, as misinformation can be quickly and easily disseminated, increasing exposure to misinformation [23]. Moreover, social media communities could further spread misinformation by interacting with other community members and reinforce the users' incorrect beliefs. In this way, the impact of misinformation could be much greater in a small, robust antivaccination community than the impact of accurate, more traditional forms of online information, such as certified web pages. Despite the important role of social media, no research has evaluated the influence of social media usage on HPV knowledge. Therefore, this study aims to examine the association between social media usage and HPV knowledge and HPV awareness.

Methods

Recruitment

In this study, a secondary data analysis was performed using public health data collected through the Health Information National Trends Survey (HINTS) conducted by the US National Cancer Institute. HINTS is a national survey that collects data on cancer and health-related knowledge and attitudes about cancer and preventive behaviors. This study is based on the most recent data, which was collected from the 2020 HINTS 5, Cycle 4 survey. Data sampling was based on random sampling using addresses, and the data were grouped into two stratifications: high concentrations of minority populations and low concentrations of minority populations. The survey was sent to the selected population by mail, and survey respondents returned their surveys by mail. Data collection occurred from February 24, 2020, through June 15, 2020. The survey was sent to 15,347 households by grouping two sampling strata: high concentrations of minority populations (n=11,050) and low concentrations of minority populations (n=4300). The response rate was 37%, with a total of 3865 responses collected. In this study, responses with missing data were excluded, resulting in a total of 2948 responses included in the final analysis.

Ethical Considerations

The data are publicly available on the HINTS website, enabling anyone to use the data without requiring specific approval from an institutional review board.

Measurements

Six HPV-related questions were used as dependent variables. HPV awareness was evaluated by the question, "Have you ever heard of HPV? HPV stands for Human Papillomavirus. It is not HCV, HIV, HSV, or herpes." HPV vaccine awareness was evaluated by the question, "A vaccine to prevent HPV infection is available and is called the HPV shot, cervical cancer vaccine, GARDASIL. Before today, have you ever heard of the cervical cancer vaccine or HPV shot?" Data show that cervical cancer vaccination is more commonly advertised than HPV vaccination [24] and that more people are aware of cervical and HPV vaccination than are aware of HPV infection [10]. Hence, we analyzed each of two HPV and HPV vaccine awareness variables.

Four questions targeted HPV knowledge by asking about four HPV-related cancers: "Do you think HPV can cause a) cervical, b) penile, c) anal, or d) oral cancer?" Previous studies have reported that there is a lack of knowledge about the relationship between HPV and the cancers that occur in males, including penile and anal cancers. Only 36.1% of respondents in prior studies were aware that HPV can cause noncervical cancers, whereas 79.6% were aware that HPV can cause cervical cancer [25]. Therefore, this study analyzed each of the four cancers as dependent variables instead of summing a single question as a continuous variable. HPV knowledge and HPV vaccine awareness were coded as "yes" or "no." HPV knowledge was screened by the HPV awareness question with three possible responses: "yes," "no," and "not sure." In this study, respondents who answered either "no" or "not sure" about HPV knowledge were both coded as "no" with respect to the HPV knowledge questions.

One extended question addressed four different methods for using social media. These four methods of using the internet were defined in a single question as follows: "Sometimes people use the Internet to connect with other people online through social networks like Facebook or Twitter. This is often called 'social media.' In the last 12 months, have you used the Internet for any of the following reasons? a) To visit a social networking site, such as Facebook or LinkedIn. b) To share health information on social networking sites, such as Facebook or Twitter. c) To participate in an online forum or support group for people with a similar health or medical issue. d) To watch a health-related video on YouTube." Each of the four possibilities was considered as one variable. All four variables were coded as either "yes" or "no." In addition to the social media questions, online health information–seeking behavior was measured by asking, "In the past 12 months, have you used a computer, smartphone, or other electronic means to look for health or medical information for yourself?" Answers were coded as "yes" or "no."

Control variables included demographic characteristics, English proficiency, and descriptions of online health information–seeking behaviors. Demographic characteristics included age, biological sex, household income ("<US $20,000," "$20,000 to <$35,000," "$35,000 to <$50,000," "$50,000 to <$75,000," or "≥$75,000"), educational level ("equal to or less than high school," "post–high school training to college graduate," or "postgraduate"), employment ("employed" or "other"), and marital status ("single, never been married" or "other"). The HINTS survey gathered data on linguistically isolated strata by identifying that 30% of households do not include members older than 14 years who speak English well. English proficiency is a significant factor associated with HPV knowledge, especially among immigrant populations [26]. Therefore, English proficiency was included as a control variable and was coded as "yes" or "no."

Statistical Analysis

Initially, six logistic regressions were used to examine the relationship between social media usage and HPV-related behaviors: HPV awareness, HPV vaccine awareness, and four HPV knowledge variables. Based on the results of the analysis, associations between three HPV-related variables and the significant predictors were further analyzed by adding demographic characteristics as control variables. Replicate weights using the jackknife replication method were used to estimate the sampling variability among the population estimates. To strike a reasonable balance between type I and type II error rates, we used an α of .01 when testing each regression coefficient and obtained the corresponding 99% CI for each odds ratio (OR). All analyses were conducted using SAS software (version 9.4; SAS Institute Inc).

Results

Demographic Characteristics

Table 1 shows the demographic characteristics of the sample. The mean age was 46.70 (SE 0.33) years. About half of the respondents were female (50.47%—weighted proportions are reported throughout the paper) and earned US $75,000 or more in household income (44.82%). Most respondents had post–high school or some college training or were college graduates (60.01%), followed by those with an educational level equal to or less than high school (27.61%) and postgraduates (12.38%). About one-third were employed (36.87%). There were 32.25% respondents who were single and had never been married. In total, 7.31% respondents indicated that their households did not include adults who are fluent in English.

Table 1. Demographic characteristics of the sample.

Characteristic	Sample (N=2948), n (%)	Weighted sample, n	Weighted proportion, % (SE)
Sex			
Female	1691 (57.36)	104,697,568	50.47 (0.56)
Male	1257 (42.64)	102,732,540	49.53 (0.56)
Household income (US $)			
<20,000	441 (14.96)	27,599,477	13.31 (0.97)
20,000 to <35,000	372 (12.62)	22,584,273	10.89 (0.81)
35,000 to <50,000	375 (12.72)	24,560,620	11.84 (0.90)
50,000 to <75,000	535 (18.15)	39,722,602	19.15 (1.63)
≥75,000	1225 (41.55)	92,963,137	44.82 (1.73)
Educational level			
Equal to or less than high school	662 (22.46)	57,275,441	27.61 (1.04)
Post–high school training, some college, or college graduate	1696 (57.53)	124,479,334	60.01 (1.23)
Postgraduate	590 (20.01)	25,675,333	12.38 (0.75)
Employment			
Employed	1333 (45.22)	76,489,111	36.87 (1.27)
Other[a]	1615 (54.78)	130,940,997	63.13 (1.27)
Marital status			
Single, never been married	528 (17.91)	66,886,401	32.25 (0.51)
Other[b]	2420 (82.09)	140,543,707	67.75 (0.51)
English proficiency			
Good	2689 (91.21)	192,276,056	92.69 (0.57)
Not good	259 (8.79)	15,154,053	7.31 (0.57)
Online health information seeking			
Yes	2204 (74.76)	156,874,229	75.63 (1.34)
No	744 (25.24)	50,555,880	24.37 (1.34)
Visited a social networking site, such as Facebook or LinkedIn			
Yes	2096 (71.10)	160,485,232	77.37 (1.17)
No	852 (28.90)	46,944,877	22.63 (1.17)
Shared health information on social networking sites, such as Facebook or Twitter			
Yes	416 (14.11)	31,485,688	15.18 (1.05)
No	2532 (85.89)	175,944,421	84.82 (1.05)
Participated in an online forum or support group for people with a similar health or medical issue			
Yes	268 (9.09)	20,857,999	10.06 (0.78)
No	2680 (90.91)	186,572,110	89.94 (0.78)
Watched a health-related video on YouTube			
Yes	1156 (39.21)	87,106,159	41.99 (1.40)
No	1792 (60.79)	120,323,949	58.01 (1.40)
HPV[c] awareness			
Yes	1953 (66.25)	141,878,406	68.40 (1.60)
No	995 (33.75)	65,551,702	31.60 (1.60)
HPV vaccine awareness			

Characteristic	Sample (N=2948), n (%)	Weighted sample, n	Weighted proportion, % (SE)
Yes	1855 (62.92)	132,835,764	64.04 (1.45)
No	1093 (37.08)	74,594,344	35.96 (1.45)
HPV knowledge (can cause cervical cancer)			
Yes	1412 (47.90)	99,569,406	48.00 (1.55)
No	1536 (52.10)	107,860,702	52.00 (1.55)
HPV knowledge (can cause penile cancer)			
Yes	554 (18.79)	39,791,723	19.18 (1.26)
No	2394 (81.21)	167,638,385	80.82 (1.26)
HPV knowledge (can cause anal cancer)			
Yes	531 (18.01)	38,012,201	18.33 (1.34)
No	2417 (81.99)	169,417,907	81.67 (1.34)
HPV knowledge (can cause oral cancer)			
Yes	566 (19.20)	41,193,057	19.86 (1.31)
No	2382 (80.80)	166,237,051	80.14 (1.31)

[a]Unemployed, homemaker, student, retired, disabled, or other response.

[b]Married, living as married, or living with a romantic partner; divorced; widowed; or separated.

[c]HPV: human papillomavirus.

With respect to social media usage–related variables, about 77.37% of respondents have visited a social networking site, 15.18% have shared health information on social networking sites, 10.06% have participated in an online forum or support group for people with similar health or medical issues, and 41.99% have watched a health-related video on YouTube. About 75.63% of respondents indicated that they seek health information online.

More than half of the respondents were aware of HPV (68.40%), and a similar number were aware of the HPV vaccine (64.04%). Less than half of the respondents knew that HPV could cause cervical cancer (48.00%). However, far fewer respondents were knowledgeable about the relationships between HPV and penile cancer (19.18%), HPV and anal cancer (18.33%), and HPV and oral cancer (19.86%).

Relationship Between Social Media Usage and HPV-Related Behaviors

Table 2 shows the results of six logistic regressions assessing the relationship between social media usage and HPV-related behaviors. In general, seeking health information online and having visited a social networking site were associated with HPV-related behaviors. Specifically, people who sought health information online were more likely to be aware of HPV (OR 2.25, 99% CI 1.49-3.40), the HPV vaccine (OR 1.85, 99% CI 1.27-2.70), and the relationship between HPV and cervical cancer (OR 2.73, 99% CI 1.69-4.42). Likewise, people who visited a social networking site were more likely to be aware of HPV (OR 2.10, 99% CI 1.28-3.43), the HPV vaccine (OR 2.10, 99% CI 1.34-3.30), and the relationship between HPV and cervical cancer (OR 1.94, 99% CI 1.17-3.22). Moreover, people who have participated in an online forum or support group for people with similar health or medical issues had higher HPV awareness (OR 2.35, 99% CI 1.05-5.26). However, having shared health information on social networking sites and having watched a health-related video on YouTube were not significant factors. Moreover, none of the social media variables were significantly related to the knowledge that HPV causes penile cancer, anal cancer, or oral cancer.

Table 2. Relationship between social media usage and HPV-related behaviors.

Predictor	HPV[a] awareness		HPV vaccine awareness		HPV knowledge (can cause cervical cancer)		HPV knowledge (can cause penile cancer)		HPV knowledge (can cause anal cancer)		HPV knowledge (can cause oral cancer)	
	OR[b] (99% CI)	*P* value	OR (99% CI)	*P* value	OR (99% CI)	*P* value	OR (99% CI)	*P* value	OR (99% CI)	*P* value	OR (99% CI)	*P* value
Online health information seeking	2.25 (1.49-3.40)	<.001	1.85 (1.27-2.70)	<.001	2.73 (1.69-4.42)	<.001	1.37 (0.64-2.91)	.27	1.40 (0.64-3.05)	.25	1.58 (0.71-3.51)	.13
Visited a social networking site, such as Facebook or LinkedIn	2.10 (1.28-3.43)	<.001	2.10 (1.34-3.30)	<.001	1.94 (1.17-3.22)	<.001	1.58 (0.82-3.03)	.07	1.22 (0.62-2.41)	.44	1.22 (0.68-2.21)	.37
Shared health information on social networking sites, such as Facebook or Twitter	1.02 (0.58-1.78)	.94	1.50 (0.92-2.45)	.03	1.17 (0.75-1.83)	.35	1.22 (0.73-2.05)	.31	1.43 (0.83-2.47)	.09	0.95 (0.54-1.68)	.82
Participated in an online forum or support group for people with a similar health or medical issue	2.35 (1.05-5.26)	.006	2.02 (0.97-4.20)	.01	1.62 (0.92-2.85)	.03	1.38 (0.74-2.59)	.17	1.10 (0.59-2.05)	.69	1.04 (0.50-2.17)	.88
Watched a health-related video on YouTube	1.10 (0.71-1.71)	.57	1.20 (0.81-1.78)	.22	1.05 (0.68-1.61)	.78	0.99 (0.58-1.67)	.94	1.06 (0.64-1.76)	.78	1.09 (0.71-1.67)	.59

[a]HPV: human papillomavirus.

[b]OR: odds ratio.

Adjusted Associations Between Social Media Usage and HPV Outcomes

Based on the previous results, demographic variables, along with the three significant social media variables, were included in regression models for HPV awareness, HPV vaccine awareness, and knowledge that HPV can cause cervical cancer. Table 3 shows the results of these logistic regressions. Unlike the previous results from Table 2, social media usage was not significantly associated with any of the HPV variables. However, seeking health information online was marginally associated with HPV awareness (OR 1.53, 99% CI 0.99-2.39; *P*=.01). Also, knowledge of the relationship between HPV and cervical cancer (OR 1.65, 99% CI 1.00-2.74; *P*=.01) and having visited a social networking site were marginally related to HPV vaccine awareness (OR 1.62, 99% CI 0.99-2.66; *P*=.01).

Among the demographic variables, age (OR 0.97, 99% CI 0.96-0.99), sex (OR 0.47, 99% CI 0.29-0.76), income, educational level, and English proficiency were significantly associated with HPV outcomes. Older people and males were less likely to be aware of the HPV vaccine. Individuals with a household income greater than or equal to US $75,000 were more likely to be aware of the HPV vaccine compared to individuals with a household income less than US $20,000 (OR 1.97, 99% CI 1.06-3.68). Respondents who are college graduates were more likely to be aware of HPV (OR 1.79, 99% CI 1.11-2.88) and the HPV vaccine (OR 2.30, 99% CI 1.43-3.72) as well as to know about relationships between HPV and cervical cancer (OR 2.97, 99% CI 1.96-4.49) compared to respondents whose educational level did not exceed high school. Similarly, respondents who indicated that they are postgraduates were more likely to be aware of HPV (OR 2.67, 99% CI 1.21-5.91), the HPV vaccine (OR 2.76, 99% CI 1.46-5.21), and the relationship between HPV and cervical cancer (OR 5.98, 99% CI 3.40-10.50) compared to respondents whose educational level did not exceed high school. Respondents whose households included adults who are fluent in English were more likely to be aware of the HPV vaccine (OR 2.12, 99% CI 1.04-4.34).

Table 3. Associations between social media usage and HPV awareness, HPV vaccine awareness, and HPV knowledge.

Predictor	HPV[a] awareness		HPV vaccine awareness		HPV knowledge (can cause cervical cancer)	
	OR[b] (99% CI)	*P* value	OR (99% CI)	*P* value	OR (99% CI)	*P* value
Age	0.97 (0.96-0.99)	<.001	0.98 (0.96-1.00)	.002	0.98 (0.96-0.99)	<.001
Gender: male (reference: female)	0.47 (0.29-0.76)	<.001	0.30 (0.20-0.44)	<.001	0.44 (0.28-0.69)	<.001
Income (US $)						
20,000 to <35,000 (reference: <20,000)	0.95 (0.48-1.87)	.83	0.91 (0.47-1.76)	.71	0.67 (0.31-1.47)	.18
35,000 to <50,000	1.12 (0.53-2.37)	.68	1.07 (0.58-1.95)	.78	0.83 (0.39-1.78)	.52
50,000 to <75,000	0.96 (0.42-2.16)	.88	0.73 (0.42-1.29)	.14	0.77 (0.37-1.61)	.34
≥75,000	1.71 (0.75-3.90)	.09	1.97 (1.06-3.68)	.005	1.38 (0.62-3.05)	.29
Educational level						
Post–high school training, some college, or college graduate (reference: equal to or less than high school)	1.79 (1.11-2.88)	.002	2.30 (1.43-3.72)	<.001	2.97 (1.96-4.49)	<.001
Postgraduate	2.67 (1.21-5.91)	.002	2.76 (1.46-5.21)	<.001	5.98 (3.40-10.50)	<.001
Employed (reference: other)	0.96 (0.60-1.53)	.81	1.06 (0.67-1.69)	.72	1.05 (0.67-1.63)	.78
Single, never been married (reference: other)	1.10 (0.63-1.91)	.65	1.15 (0.67-1.99)	.49	1.19 (0.70-2.03)	.39
English proficiency	1.39 (0.72-2.68)	.19	2.12 (1.04-4.34)	.007	1.12 (0.60-2.10)	.63
Online health information seeking	1.53 (0.99-2.39)	.01	1.15 (0.74-1.79)	.40	1.65 (1.00-2.74)	.01
Visited a social networking site, such as Facebook or LinkedIn	1.44 (0.89-2.32)	.05	1.62 (0.99-2.66)	.01	1.45 (0.84-2.50)	.07
Participated in an online forum or support group for people with a similar health or medical issue	2.14 (0.95-4.82)	.02	2.11 (0.97-4.61)	.01	1.42 (0.81-2.50)	.10

[a]HPV: human papillomavirus.

[b]OR: odds ratio.

Discussion

Principal Findings

This study analyzed the relationship between social media usage and awareness of HPV, the HPV vaccine, and HPV-related knowledge about cervical, anal, penile, and oral cancers. Although social media usage is associated with HPV awareness and knowledge, these associations were not significant after adjusting for demographic variables and were only marginally related to HPV-related behaviors. Meanwhile, the demographic variables age, sex, educational level, income, and English proficiency were significantly associated with HPV-related behaviors.

The nonsignificant associations between social media usage and HPV awareness, HPV vaccine awareness, and knowledge related to cervical cancer might be related to the quality of information on social media. Earlier research that analyzed websites that mention HPV reported that only 4.81% of those websites included information that HPV can cause cervical cancer [20]. In addition, research that specifically analyzed Twitter postings reported that most tweets about HPV were written by nonprofessionals. Twitter tweets about HPV more often contained links to layperson blogs compared to links to professional information or websites [19]. Thus, our results showing that respondents who used social media did not possess more HPV knowledge than respondents who did not use social media may reflect the poor quality of information posted on social media and on some websites and blogs that are linked to poor-quality social media posts. Moreover, searching for health information on social media may be triggered by the needs of people who have special health concerns or health issues [15]. Using social media for personal health knowledge could not be addressed directly with respect to HPV. People may search for other topics or general health concerns when they use social media [27]. Since HPV infection does not result in any symptoms and can cause cancer multiple years after infection, people might not search specifically for information about HPV. People are less likely to search for information about HPV unless someone actually recommends that they should research HPV.

Another explanation could be the characteristics of social media for the information exchange perspectives. Using social media could limit the information that circulates within the community and could lead to a lack of knowledge [23]. People obtain health information by interacting with peers inside social media or inside specific, smaller, and more robust communities found on social media [15]. Information provided by social media could be reinforced by exposure based on the number of users and networks, so it could increase the proliferation of misinformation and could reinforce the incorrect beliefs of the

viewers [23]. If there is no one to correct misinformation, people using social media may have difficulty discerning correct from incorrect information. Unfortunately, experts or government agencies are often unable to correct misinformation. Compelling personal stories that contain misinformation can be especially difficult to correct, further impeding the promotion of accurate health information on social media.

Our results also show that watching health-related videos on YouTube is not associated with either HPV awareness or HPV knowledge. We suggest that there may be three possible explanations for this unexpected result. First, the majority of YouTube videos about HPV were videos that contained an antivaccine bias: 57% of YouTube videos presented an antivaccine philosophy compared to only 31% of YouTube videos that promoted the health benefits of HPV vaccination [21]. Second, most of the viewers' top comments about HPV-related videos highlighted potential negative side effects of vaccination and supported conspiracy theories about recommendations for HPV vaccination [21]. This combination of inaccurate and biased information and the prominence of negative viewer comments about HPV vaccination might explain why watching health-related videos on YouTube is not associated with respondents having more HPV knowledge compared to respondents who do not watch YouTube for health-related information. Finally, a third explanation is related to the way in which viewers find information on YouTube. YouTube provides personalized videos based on the viewer's history of watching [28]. YouTube algorithms employ user-provided performance, watch history, and recognition of the specific videos that users watch to suggest additional videos to users. The nonsignificant association between HPV awareness and watching YouTube videos might be a consequence of how users find information on YouTube. People who are unaware of HPV may be less likely to enter HPV-related keywords and, therefore, their searches would be less likely to trigger YouTube algorithms to suggest videos about HPV. The HINTS questions did not ask respondents about the specific health issues that they researched online or through social media. Further research is suggested to investigate the causal relationships.

Literature indicates that the HPV vaccine is often described as a cervical cancer vaccine [24] and that respondents report greater awareness of a cervical cancer vaccine instead of its product name, GARDASIL 9, or the HPV vaccine itself [10]. There is less knowledge that HPV infection can cause cancers other than cervical cancer, including penile and oral cancers [13]. This suggests that respondents who searched online for HPV-related health information were already more knowledgeable about HPV than the respondents who did not search online for HPV-related health information. Perhaps people who have heard of HPV search online for the detailed information, and this might explain the marginal effect of seeking health information online and visiting social networking sites regarding the HPV vaccine and knowledge that cervical cancer is caused by HPV. In addition, our research supports the findings of previous literature showing that knowledge of the relationship between HPV infection and penile, oral, and anal cancers is very low: fewer than 20% of our sample knew about these relationships [13]. This low number may indicate that even people who have some knowledge about HPV and who search online for more information still lack information about HPV infection and its relationship to multiple cancers. Greater efforts are needed to inform people that HPV can cause a variety of cancers, and that HPV vaccination is an effective method of preventing these cancers.

Other control variables, such as age, income, occupational status, and English proficiency, were associated with HPV vaccination or HPV vaccination intention. The results of this study confirmed previous findings about factors associated with HPV vaccination [11,29,30].

Limitations

Although this study indicated that social media usage has a significant role in HPV-related behaviors, this study has some limitations. First, this study was based on secondary data from a public survey. The authors did not develop a survey that focused on specific knowledge about HPV, and the questions about social media usage did not directly address HPV. This limits generalization of our findings about the association between HPV behaviors and social media usage. Second, this study did not examine HPV vaccination intention or vaccine uptake. Although HPV vaccination intention may be highly related to HPV vaccine uptake, some people who initiate HPV vaccination do not complete the entire series of two or three shots required for effective HPV vaccination [31]. Additional research is needed to determine the percentage of people who have HPV vaccination intention and fully complete HPV vaccination. A third limitation of this study is that the respondents were mainly adults. HPV vaccination is suggested for youth and young adults aged 11 to 26 years, [32] who, as a group, are more likely to access and use social media compared to older adults. Further studies that target this group are recommended.

Conclusions and Implications

Previous research has revealed the significance of HPV awareness, but it has not addressed HPV vaccine awareness and HPV-related knowledge. This study provided further evidence of the nonsignificant relationship between social media usage and HPV-related behaviors. Earlier research has shown that the majority of HPV-related videos and most of the top viewer comments on YouTube reflect antivaccination bias [21]. Our results suggest that there is a lack of high-quality, accurate information on social media. Unlike traditional media, it is hard for health care professionals to intervene through social media. Rather than rely on individual health care workers, government-level policies or efforts are needed to provide accurate information and promote HPV vaccination. Information about HPV has to be accurate and easy for nonprofessionals to understand. A previous study that analyzed Instagram posts about HPV reported that personal stories were the prevalent source of antivaccine postings [33]. Efforts to use storytelling on social media could be one approach to persuade the public. In addition, information that is well suited to social media (eg, infographics) could also increase knowledge about HPV vaccination. An intervention study that used infographics on social media reported that infographics were able to reduce misperceptions about COVID-19 [34]. Government agencies

that wish to inform the public with accurate information about HPV should develop communication methods that are appropriate to be shared within social media, based on the public's level of understanding, and should make strong efforts to disseminate this information within social media.

Authors' Contributions

SJ was responsible for the study design, data analysis, and writing the manuscript. KAP was responsible for data analysis and writing the manuscript. NH was responsible for writing the manuscript.

Conflicts of Interest

None declared.

References

1. Satterwhite CL, Torrone E, Meites E, Dunne EF, Mahajan R, Ocfemia MCB, et al. Sexually transmitted infections among US women and men: Prevalence and incidence estimates, 2008. Sex Transm Dis 2013 Mar;40(3):187-193. [doi: 10.1097/OLQ.0b013e318286bb53] [Medline: 23403598]
2. Cancers associated with human papillomavirus (HPV). Centers for Disease Control and Prevention. 2021. URL: https://www.cdc.gov/cancer/hpv/basic_info/cancers.htm [accessed 2022-09-08]
3. Human papillomavirus (HPV) vaccines. National Cancer Institute. 2021. URL: https://www.cancer.gov/about-cancer/causes-prevention/risk/infectious-agents/hpv-vaccine-fact-sheet [accessed 2022-09-08]
4. Saraiya M, Unger E, Thompson T, Lynch CF, Hernandez BY, Lyu CW, HPV Typing of Cancers Workgroup. US assessment of HPV types in cancers: Implications for current and 9-valent HPV vaccines. J Natl Cancer Inst 2015 Jun;107(6):djv086 [FREE Full text] [doi: 10.1093/jnci/djv086] [Medline: 25925419]
5. Chaturvedi AK, Engels EA, Pfeiffer RM, Hernandez BY, Xiao W, Kim E, et al. Human papillomavirus and rising oropharyngeal cancer incidence in the United States. J Clin Oncol 2011 Nov 10;29(32):4294-4301 [FREE Full text] [doi: 10.1200/JCO.2011.36.4596] [Medline: 21969503]
6. Cancer Statistics Center, American Cancer Society. 2021. URL: https://cancerstatisticscenter.cancer.org/#!/ [accessed 2022-01-21]
7. Brotherton JM, Tabrizi SN, Phillips S, Pyman J, Cornall AM, Lambie N, et al. Looking beyond human papillomavirus (HPV) genotype 16 and 18: Defining HPV genotype distribution in cervical cancers in Australia prior to vaccination. Int J Cancer 2017 Oct 15;141(8):1576-1584 [FREE Full text] [doi: 10.1002/ijc.30871] [Medline: 28677147]
8. Pingali C, Yankey D, Elam-Evans L, Markowitz LE, Williams CL, Fredua B, et al. National, regional, state, and selected local area vaccination coverage among adolescents aged 13-17 years - United States, 2020. MMWR Morb Mortal Wkly Rep 2021 Sep 03;70(35):1183-1190 [FREE Full text] [doi: 10.15585/mmwr.mm7035a1] [Medline: 34473682]
9. Radisic G, Chapman J, Flight I, Wilson C. Factors associated with parents' attitudes to the HPV vaccination of their adolescent sons: A systematic review. Prev Med 2017 Feb;95:26-37. [doi: 10.1016/j.ypmed.2016.11.019] [Medline: 27932052]
10. Kim M, Lee H, Kiang P, Aronowitz T, Sheldon LK, Shi L, et al. HPV vaccination and Korean American college women: Cultural factors, knowledge, and attitudes in cervical cancer prevention. J Community Health 2019 Aug;44(4):646-655 [FREE Full text] [doi: 10.1007/s10900-019-00634-9] [Medline: 30863974]
11. Jo S, Han S, Walters C. Factors associated with the HPV vaccination among Korean Americans and Koreans: A systematic review. Int J Environ Res Public Health 2021 Dec 21;19(1):51 [FREE Full text] [doi: 10.3390/ijerph19010051] [Medline: 35010311]
12. López N, Salamanca de la Cueva I, Vergés E, Suárez Vicent E, Sánchez A, López AB, et al. Factors influencing HPV knowledge and vaccine acceptability in parents of adolescent children: Results from a survey-based study (KAPPAS study). Hum Vaccin Immunother 2022 Dec 31;18(1):2024065 [FREE Full text] [doi: 10.1080/21645515.2021.2024065] [Medline: 35103571]
13. Wigfall L, Sherman L, Garney W, Patterson M, Montiel Ishino FA, Vadaparampil S. Are health care providers making the most of patient encounters to promote HPV vaccination among cigarette smokers? Patient Educ Couns 2020 Jan;103(1):180-188 [FREE Full text] [doi: 10.1016/j.pec.2019.07.026] [Medline: 31383561]
14. Gor BJ, Chilton JA, Camingue PT, Hajek RA. Young Asian Americans' knowledge and perceptions of cervical cancer and the human papillomavirus. J Immigr Minor Health 2011 Feb;13(1):81-86 [FREE Full text] [doi: 10.1007/s10903-010-9343-7] [Medline: 20414727]
15. Zhao Y, Zhang J. Consumer health information seeking in social media: A literature review. Health Info Libr J 2017 Dec;34(4):268-283 [FREE Full text] [doi: 10.1111/hir.12192] [Medline: 29045011]
16. Lin WY, Zhang X, Song H, Omori K. Health information seeking in the Web 2.0 age: Trust in social media, uncertainty reduction, and self-disclosure. Comput Human Behav 2016 Mar;56:289-294. [doi: 10.1016/j.chb.2015.11.055]

17. Qin L, Zhang X, Wu A, Miser JS, Liu YL, Hsu JC, et al. Association between social media use and cancer screening awareness and behavior for people without a cancer diagnosis: Matched cohort study. J Med Internet Res 2021 Aug 27;23(8):e26395 [FREE Full text] [doi: 10.2196/26395] [Medline: 34448708]
18. Ortiz RR, Smith A, Coyne-Beasley T. A systematic literature review to examine the potential for social media to impact HPV vaccine uptake and awareness, knowledge, and attitudes about HPV and HPV vaccination. Hum Vaccin Immunother 2019;15(7-8):1465-1475 [FREE Full text] [doi: 10.1080/21645515.2019.1581543] [Medline: 30779682]
19. Keim-Malpass J, Mitchell EM, Sun E, Kennedy C. Using Twitter to understand public perceptions regarding the #HPV vaccine: Opportunities for public health nurses to engage in social marketing. Public Health Nurs 2017 Jul;34(4):316-323. [doi: 10.1111/phn.12318] [Medline: 28261846]
20. Okuhara T, Ishikawa H, Okada M, Kato M, Kiuchi T. Contents of Japanese pro- and anti-HPV vaccination websites: A text mining analysis. Patient Educ Couns 2018 Mar;101(3):406-413. [doi: 10.1016/j.pec.2017.09.014] [Medline: 29031425]
21. Ekram S, Debiec KE, Pumper MA, Moreno MA. Content and commentary: HPV vaccine and YouTube. J Pediatr Adolesc Gynecol 2019 Apr;32(2):153-157. [doi: 10.1016/j.jpag.2018.11.001] [Medline: 30445163]
22. Zipprich J, Winter K, Hacker J, Xia D, Watt J, Harriman K. Measles outbreak--California, December 2014-February 2015. MMWR Morb Mortal Wkly Rep 2015 Feb 20;64(6):153-154 [FREE Full text] [Medline: 25695321]
23. Smith N, Graham T. Mapping the anti-vaccination movement on Facebook. Inf Commun Soc 2019;22(9):1310-1327.
24. Gollust S, LoRusso S, Nagler R, Fowler E. Understanding the role of the news media in HPV vaccine uptake in the United States: Synthesis and commentary. Hum Vaccin Immunother 2016 Jun 02;12(6):1430-1434 [FREE Full text] [doi: 10.1080/21645515.2015.1109169] [Medline: 26554612]
25. Thompson EL, Wheldon CW, Rosen BL, Maness SB, Kasting ML, Massey PM. Awareness and knowledge of HPV and HPV vaccination among adults ages 27-45 years. Vaccine 2020 Mar 30;38(15):3143-3148. [doi: 10.1016/j.vaccine.2020.01.053] [Medline: 32029321]
26. Garcini LM, Murray KE, Barnack-Tavlaris JL, Zhou AQ, Malcarne VL, Klonoff EA. Awareness and knowledge of human papillomavirus (HPV) among ethnically diverse women varying in generation status. J Immigr Minor Health 2015 Feb;17(1):29-36. [doi: 10.1007/s10903-013-9913-6] [Medline: 24052478]
27. Lin LY, Sidani JE, Shensa A, Radovic A, Miller E, Colditz JB, et al. Association between social media use and depression among U.S. young adults. Depress Anxiety 2016 Apr;33(4):323-331 [FREE Full text] [doi: 10.1002/da.22466] [Medline: 26783723]
28. Cooper P. How the YouTube algorithm works in 2022: The complete guide. Hootsuite. 2021 Jun 21. URL: https://blog.hootsuite.com/how-the-youtube-algorithm-works/ [accessed 2022-09-08]
29. Guo Y, Bowling J. Human papillomavirus (HPV) vaccination initiation and completion among adult males in the United States. J Am Board Fam Med 2020;33(4):592-599 [FREE Full text] [doi: 10.3122/jabfm.2020.04.190464] [Medline: 32675270]
30. Lee HY, Luo Y, Neese J, Daniel C, Hahm HC. The role of English proficiency in HPV and HPV vaccine awareness: A cross-sectional study across race/ethnicity. Asian Pac J Cancer Prev 2021 Feb 01;22(2):349-357 [FREE Full text] [doi: 10.31557/apjcp.2021.22.2.349]
31. Jeudin P, Liveright E, Del Carmen MG, Perkins RB. Race, ethnicity, and income factors impacting human papillomavirus vaccination rates. Clin Ther 2014 Jan 01;36(1):24-37. [doi: 10.1016/j.clinthera.2013.11.001] [Medline: 24417783]
32. HPV vaccination recommendations. Centers for Disease Control and Prevention. 2021. URL: https://www.cdc.gov/vaccines/vpd/hpv/hcp/recommendations.html [accessed 2022-01-08]
33. Massey P, Kearney M, Hauer M, Selvan P, Koku E, Leader A. Dimensions of misinformation about the HPV vaccine on Instagram: Content and network analysis of social media characteristics. J Med Internet Res 2020 Dec 03;22(12):e21451 [FREE Full text] [doi: 10.2196/21451] [Medline: 33270038]
34. Vraga E, Bode L. Addressing COVID-19 misinformation on social media preemptively and responsively. Emerg Infect Dis 2021 Feb;27(2):396-403 [FREE Full text] [doi: 10.3201/eid2702.203139] [Medline: 33395379]

Abbreviations

HINTS: Health Information National Trends Survey
HPV: human papillomavirus
OR: odds ratio

Edited by Y Khader; submitted 14.02.22; peer-reviewed by M Lotto, E Said-Hung; comments to author 24.04.22; revised version received 12.07.22; accepted 29.07.22; published 20.09.22.

Please cite as:
Jo S, Pituch KA, Howe N
The Relationships Between Social Media and Human Papillomavirus Awareness and Knowledge: Cross-sectional Study
JMIR Public Health Surveill 2022;8(9):e37274
URL: https://publichealth.jmir.org/2022/9/e37274
doi:10.2196/37274
PMID:36125858

Original Paper

Effectiveness of Cash Transfer Delivered Along With Combination HIV Prevention Interventions in Reducing the Risky Sexual Behavior of Adolescent Girls and Young Women in Tanzania: Cluster Randomized Controlled Trial

Evodius Kuringe[1], MD; Alice Christensen[2], MSN; Jacqueline Materu[1], MSc; Mary Drake[2], MPH; Esther Majani[2], MPH; Caterina Casalini[2], MD; Deusdedit Mjungu[2], MD; Gaspar Mbita[2], MSc; Esther Kalage[2], MPH; Albert Komba[2], MPH; Daniel Nyato[1], MA; Soori Nnko[1], PhD; Amani Shao[1], PhD; John Changalucha[1], MSc; Mwita Wambura[1], PhD

[1]Department of Sexual and Reproductive Health, National Institute for Medical Research, Mwanza, United Republic of Tanzania

[2]Sauti Project, Jhpiego (an affiliate of John Hopkins University), Dar-es-Salaam, United Republic of Tanzania

Corresponding Author:
Evodius Kuringe, MD
Department of Sexual and Reproductive Health
National Institute for Medical Research
Isamilo Road, Box 1462
Mwanza, 33104
United Republic of Tanzania
Phone: 255 0282500399
Fax: 255 282500399
Email: evokur@gmail.com

Abstract

Background: Poverty and social inequality exacerbate HIV risk among adolescent girls and young women (AGYW) in sub-Saharan Africa. Cash transfers can influence the structural determinants of health, thereby reducing HIV risk.

Objective: This study assessed the effectiveness of cash transfer delivered along with combination HIV prevention (CHP) interventions in reducing the risky sexual behavior of AGYW in Tanzania. The incidence of herpes simplex virus type 2 (HSV-2) infection was used as a proxy for sexual risk behavior.

Methods: A cluster randomized controlled trial was conducted in 15 matched pairs of communities (1:1 intervention to control) across 3 strata (urban, rural high-risk, and rural low-risk populations) of the Shinyanga Region, Tanzania. The target population was out-of-school AGYW aged 15-23 years who had completed 10-hour sessions of social and behavior change communication. Eligible communities were randomly assigned to receive CHP along with cash transfer quarterly (intervention group) or solely CHP interventions (control group) with no masking. Study recruitment and baseline survey were conducted between October 30, 2017 and December 1, 2017. Participants completed an audio computer-assisted self-interview, HIV counselling and testing, and HSV-2 testing at baseline and during follow-up visits at 6, 12, and 18 months after the baseline survey. A Cox proportional hazards model with random effects specified at the level of clusters (shared frailty) adjusted for matching pairs and other baseline imbalances was fitted to assess the effects of cash transfer on the incidence of HSV-2 infection (primary outcome). Secondary outcomes included HIV prevalence at follow-up, self-reported intergenerational sex, and self-reported compensated sex. All secondary outcomes were measured at each study visit.

Results: Of the 3026 AGYW enrolled in the trial (1482 in the intervention and 1544 in the control), 2720 AGYW (1373 in the intervention and 1347 in the control) were included in the final analysis. Overall, HSV-2 incidence was not significantly different at all follow-up points between the study arms in the adjusted analysis (hazard ratio 0.96, 95% CI 0.67-1.38; *P*=.83). However, HSV-2 incidence was significantly lower in the rural low-risk populations who received the cash transfer intervention (hazard ratio 0.45, 95% CI 0.29-0.71; *P*=.001), adjusted for potential confounders.

Conclusions: Although this trial showed no significant impact of the cash transfer intervention on HSV-2 incidence among AGYW overall, the intervention significantly reduced HSV-2 incidence among AGYW in rural low-risk communities. Factors such as lesser poverty and more asset ownership in urban and rural high-risk communities may have undermined the impact of cash transfer.

Trial Registration: ClinicalTrials.gov NCT03597243; https://clinicaltrials.gov/show/NCT03597243

(JMIR Public Health Surveill 2022;8(9):e30372) doi:10.2196/30372

KEYWORDS

adolescent; female; HIV infections/epidemiology; HIV infections/prevention and control; herpes simplex virus type 2; incidence; motivation; Tanzania

Introduction

Adolescent girls and young women (AGYW) aged 15-24 years in eastern and southern regions of Africa have high rates of HIV infection [1-4]. In Tanzania, the HIV incidence was 0.14% (95% CI 0%-0.31%) and 0% (95% CI 0%-0.23%) in 2016 among AGYW and adolescent boys and young men, respectively [5]. Targeted HIV prevention efforts among AGYW are crucial to realize an HIV-free generation [6]. Interactions among biological [7-9], behavioral [10-15], and structural determinants of health [10,16] have been associated with higher HIV infection risks in AGYW, with poverty [10,16] and gender inequality [10,17] exacerbating their vulnerability. AGYW from low-income households are likely to engage in transactional sex [12,15] to cater to their daily subsistence and are less likely to negotiate safer sexual practices [9,10,18]. This highlights the need for incorporating structural interventions in combination with biomedical and behavioral approaches for impactful HIV prevention among AGYW [19-21]. Cash transfer as a structural intervention has shown promising results, with some studies demonstrating its impact on the prevention of sexually transmitted infections and delaying marriage and childbearing [22-25]. Of the 8 cash transfer studies conducted in sub-Saharan Africa with sexually transmitted infections as an outcome, 4 reported a significant reduction; cash transfer provided to in-school AGYW aged 13-22 years and their parents reduced HIV prevalence by 64% (95% CI 9%-86%) and herpes simplex virus type 2 (HSV-2) prevalence by 76% (95% CI 35%-91%) [22]. Short-term financial incentives to engage in safe sex in Lesotho reduced HIV incidence among males and females aged 18-30 years by 25% (95% CI 3%-42%) [25]. Similarly, another financial incentive to engage in safe sex in Eswatini, which was conditional on staying sexually transmitted infection–negative, reduced HIV incidence among schoolgirls by 23% (95% CI 1%-40%) compared to those not eligible for educational cash transfer [23,26]. A South African study [27,28] showed reduced HSV-2 incidence by 30% (95% CI 14%-43%) among school-based girls and boys but had too few new infections for HIV incidence analysis. Other studies [29,30] reported that cash transfer had no effect on HIV prevalence in South Africa among in-school adolescent girls (adjusted odds ratio [OR] 1.17, 95% CI 0.80-1.72) [29] and among men and women in Malawi conditional on remaining HIV-negative (β=.001, robust SE=.005) [30]. Cash transfer studies among in-school adolescent orphans showed no effect on HIV infection (OR 1.15, 95% CI 0.47-2.79) and HSV-2 infection (OR 1.46, 95% CI 0.50-4.26) in Zimbabwe [31] and on HIV infection (adjusted OR 0.72, 95% CI 0.15-3.42) and HSV-2 infection (OR 0.98, 95% CI 0.54-4.26) in Kenya [32]. However, none of these studies in sub-Saharan Africa were conducted among out-of-school AGYW aged 15-23 years to ascertain the combined effect of cash transfer and combination HIV prevention (CHP) services.

We examined the synergetic effect of cash transfer and CHP interventions among out-of-school AGYW in the Shinyanga Region, Tanzania, where the Sauti (meaning *Voices in Kiswahili)* project provided CHP interventions to AGYW [33]. We had 2 aims for this study. First, we sought to assess the synergetic effect of cash transfer and CHP interventions on HSV-2 incidence. Second, we sought to examine the effect of cash transfer along with CHP on AGYW's sexual behavior. We hypothesized that cash transfer along with CHP interventions would be associated with a reduced risky sexual behavior compared with CHP alone. HSV-2 incidence was chosen as a proxy measure for HIV incidence, as the study would not be powered to detect a difference in HIV incidence in a relatively low incidence setting like Tanzania. Other measures of risky sexual behavior such as self-reported behavioral indicators are affected by bias and low validity compared to biomarkers such as HIV and HSV-2 infection [34,35]. HSV-2 is a sexually transmitted infection like HIV, more prevalent in Tanzania, and has been used as a marker for risky sexual behavior in other similar studies conducted in sub-Saharan Africa [29,36]. Our findings contribute to the body of evidence on the effectiveness of cash transfer in the reduction of risky sexual behavior among AGYW in Tanzania.

Methods

Study Design and Setting

This study was a 2-arm cluster-randomized controlled trial with 1:1 allocation ratio, implemented among out-of-school AGYW in Tanzania by the Sauti project—a project implementing community-based interventions under the DREAMS (Determined, Resilient, Empowered, AIDS-free, Mentored, and Safe) initiative in Tanzania [4,33]. The Sauti project collaborated with the Government of Tanzania and civil society organizations (CSOs) to provide community-based CHP interventions to AGYW in selected regions of Tanzania, which is the largest country in East Africa and is hierarchically subdivided into regions, districts, divisions, wards, and villages in rural settings and into neighborhoods *(mtaa)* in urban settings. A description of the methods used in this trial has been previously published [33]. Briefly, this study used 15 matched clusters randomly allocated to intervention (cash transfer plus CHP) or control (CHP only) arms. We randomly assigned clusters rather than individuals because cash transfer could be shared among family members even when allocated in different arms (risk of contamination: AGYW are likely to share the cash given to the participant in the intervention arm with relatives allocated to the control arm, thus reducing/diluting the impact of the study)

[37,38]. In addition, a 5-km buffer zone was maintained between clusters to minimize the dilution of the intervention.

Ethics Approval

Ethics approval for this study was obtained from the Johns Hopkins University School of Public Health Research Ethics Committee (00007976) and the Tanzanian Medical Research Coordinating Committee of the National Institute for Medical Research (NIMR/HQ/R.8a/VolIX/2287). This trial was registered at ClinicalTrials.gov (NCT03597243) and was conducted and reported following the Consolidated Standards of Reporting Trials (CONSORT) guidelines for cluster randomized trials [39].

Selection of Participants

This study enrolled participants from a pool of potential DREAMS beneficiaries in the Shinyanga Region. The project conducted a household survey among AGYW in the identified villages to determine AGYW who were out-of-school and developed a pool of potential cash transfer program beneficiaries. Potential beneficiaries were then invited to attend 10 sessions of social and behavior change communication (SBCC) and other project interventions. The SBCC sessions were peer-led group sessions designed to address significant determinants of HIV risk, gender, and reproductive health. During the SBCC sessions, data on all villages receiving CHP were generated for study eligibility assessment and randomization. In the identified study clusters (intervention and control), potential participants were also given information about the study during group meetings for SBCC. After 10 hours of SBCC sessions, AGYW were informed about study enrolment, recruitment dates, and locations. AGYW were eligible for the study if they were aged 15-23 years; out-of-school (defined as either never enrolled or have dropped out of school for at least a month at the time of study enrolment) as documented by the household survey; residents of the village of recruitment; had completed 10 hours of SBCC training; willing to take part in the study, including testing for HIV and receiving results; and willing to participate in the cash transfer program (applicable to the intervention arm). Each study participant provided written informed consent or assent. Consent and assent forms were administered in Kiswahili. AGYW who tested positive for HIV or HSV-2 infection at baseline were enrolled in the study to avoid inadvertent disclosure of their serostatus. However, those with a positive HSV-2 status at baseline were excluded from the main analysis.

Randomization and Masking

Villages were eligible for the study if they were receiving other Sauti project interventions except for cash transfer, identified as potentially eligible for cash transfer program, and having between 110 and 150 AGYW aged 15-23 years who were out of school. All eligible clusters were matched into pairs by location (rural vs urban areas) and the presence or absence of HIV high-risk areas, generating 3 strata: rural clusters in the high-risk area, urban clusters, and rural clusters in the low-risk area. Matching was conducted to minimize the between-community variance in HSV-2 incidence within the matched clusters. One cluster was randomly selected from each matched pair to receive the intervention package, and the other was automatically assigned to the control arm. No blinding was performed in this study.

Study Implementation

Participants in the intervention arm received CHP and unconditional cash transfer in quarterly instalments of 70,000 Tanzania shillings (~US $31) for 18 months through mobile money on a project-provided cellular phone. Participants in the control arm received CHP only and cellular phones provided by the project. The aim was to make sure that the 2 arms were comparable except for the intervention. The interventions are described elsewhere [33]. In short, all study participants received Sauti's core package of interventions, including risk reduction counselling, HIV testing services, condom use skills and provision, family planning counselling and service provision, sexually transmitted infection screening and treatment, gender-based violence interventions (escorted referrals and the desk for social, legal, and medical services provided to survivors of gender-based violence), tuberculosis and alcohol and drug abuse screening, and referral to services. The other features were SBCC training sessions and economic empowerment community banking also called as the Women Organizing Resources Together plus (WORTH+) intervention [33]. The cash transfer program was implemented in the intervention arm only. The WORTH+ intervention consisted of financial literacy training that aimed to build microbusiness development skills and community banking.

At baseline, following consent and enrolment into the study, AGYW completed audio computer-assisted self-interview (ACASI), which collected data on demographic information, factors related to HIV vulnerabilities, family planning, sexual risk behavior, and gender-based violence. Sexual behavior data collected included compensated sex (sexual encounters motivated by exchange for money, material support, or other benefits) and intergenerational sex (a sexual partnership between AGYW and a man 10 or more years older). Further, data on sex work defined as having negotiated payment for sex and transactional sex defined as initiating a sexual relationship with an expectation to receive money or gifts were collected. After data collection using ACASI, trained government health care workers offered HIV pretest and posttest counselling to participants. Blood was drawn for HIV and HSV-2 testing. HIV testing was done following the National Guidelines for the Management of HIV and AIDS [40]. All study procedures were conducted in a confidential environment in preidentified venues in the respective communities. Participants were seen every 6 months for study activities (6, 12, and 18 months), while the Sauti program provided CHP interventions. Each study visit included ACASI, HIV pretest and posttest counselling, and HSV-2 testing (if negative at the previous visit). Blood specimens were taken and transported to the local health facility laboratory for serum separation and temporarily stored at –20 °C before transportation to the National Institute for Medical Research (Mwanza laboratory) for HSV-2 testing.

HIV testing was conducted using 2 HIV rapid tests following Tanzania's national HIV testing and counselling guidelines. Participants who were HIV-positive received escorted referrals

to care and treatment centers. HSV-2 testing was conducted using the HSV-2 IgG enzyme-linked immunosorbent assay (Kalon Biological Ltd). Participants with positive HSV-2 test results were given posttest counselling and referral for treatment where required. At each follow-up visit, nurse counsellors assessed the social harm events by actively interrogating the study participant. If social harm was reported, it was recorded and graded per Sauti project safety guidelines, which defined and outlined procedures for reporting and management. All social harms were reported to the study steering committee, and the Sauti project initiated investigations and responses as appropriate. It was anticipated that any harm to AGYW owing to study participation would be minimal. The primary outcome was HSV-2 incidence, while secondary outcomes included HIV prevalence at follow-up, self-reported intergenerational sex, and self-reported compensated sex. All secondary outcomes were measured at each study visit.

Statistical Analysis

All calculations for sample size were conducted using methods for matched cluster-randomized trials [41]. A sample of 14 paired clusters (28 clusters) with 70 participants per cluster was estimated to achieve over 80% power of detecting a 35% reduction in HSV-2 incidence in the intervention arm at the end of 18 months. The within-pair coefficient of variation between clusters was assumed to be 0.25 [42], and the significance level of the test was .05. It was estimated that, at baseline, HSV-2 prevalence would be 20% among AGYW aged 15-23 years [43], the attrition rate would be 10%, and the nonresponse rate would be 18% over 18 months. Thus, the sample size (70 AGYW per cluster) was increased to 104 AGYW per cluster (1560 per arm) and paired clusters increased to 15 (30 clusters). The HSV-2 incidence estimate was based on a sample size of approximately 1575 person-years per cluster by month 18. All analyses performed were prespecified. The primary analysis was intention-to-treat and based on individual-level data because the study had a sufficient number of clusters per arm, and the cluster size was anticipated to differ considerably at follow-up [41].

Descriptive analysis was done using standard methods for the analysis of a pair-matched cluster randomized trial with a small number of clusters [44]. Baseline data were used to assess balance across study arms in sociodemographic and other key characteristics associated with an outcome. Since covariates such as age, marital status, whether AGYW had emotional/psychological support, whether AGWY had debt at the time of the survey, and whether AGYW lacked food in the past 4 weeks were imbalanced across the study arms, they were adjusted for in the analysis.

A Cox proportional hazards model with random effects specified at the level of clusters (shared frailty), adjusted for matching pairs and other baseline imbalances, was fitted to assess the effects of cash transfer on the incidence of HSV-2 infection. The significance of the intervention was assessed using a value of .05 (2-sided) after verification of the validity of the proportional hazards assumption. Secondary analysis of the effect of cash transfer on HIV prevalence at follow-up was estimated using a log-binomial model adjusting for matched pairs, age, and other variables with baseline imbalance between the arms and adjusting for standard errors for clustering at the village level. Intergenerational sex, compensated sex, transactional sex, and other behavioral outcomes were compared between study arms by using generalized estimating equations with identity logit, binomial distribution, and robust variance to account for repeated measures on each participant. The generalized estimating equations regression models were also fitted to assess for interaction between the trial arm and strata (rural cluster in the high-risk area, urban cluster, rural cluster in the low-risk area).

Results

Sociodemographic Characteristics of the Participants

Study recruitment took place between October 30, 2017 and December 1, 2017. Of the 3105 participants screened for eligibility (Figure 1), 3071 were eligible and 3055 consented/assented to study participation. Of these, 3026 (1482 in the intervention and 1544 in the control) participants completed baseline survey procedures and 2720 (1373 in the intervention and 1347 in the control) attended at least one follow-up study visit. Of those with follow-up data, 865 (443 in the intervention and 422 in the control) were infected with HSV-2 at baseline and, therefore, excluded from the HSV-2 longitudinal data analysis. Therefore, 1855 (930 in the intervention and 925 in the control) were included in the longitudinal analysis of the primary outcome contributing to 2524.2 personal years of observation (PYO) (1293.9 PYO in the intervention and 1230.3 PYO in the control) and had a retention of 61.3% (1855/3026; 930/1482, 62.7% in the intervention and 925/1544, 59.9% in the control). At baseline, the median age of the participants was 20 (IQR 18-22) years, and the intervention and control groups were similar for key sexual behavior variables, HIV prevalence, and HSV-2 prevalence. However, there were imbalances in the sociodemographic variables such as age, marital status, debt, and going to bed hungry (Table 1). The imbalance between the arms was adjusted for in the impact analysis.

Figure 1. Trial profile for a cluster randomized trial among adolescent girls and young women in Shinyanga Region, Tanzania, in October 2017 to July 2019. AGYW: adolescent girls and young women; HSV-2: herpes simplex virus type 2.

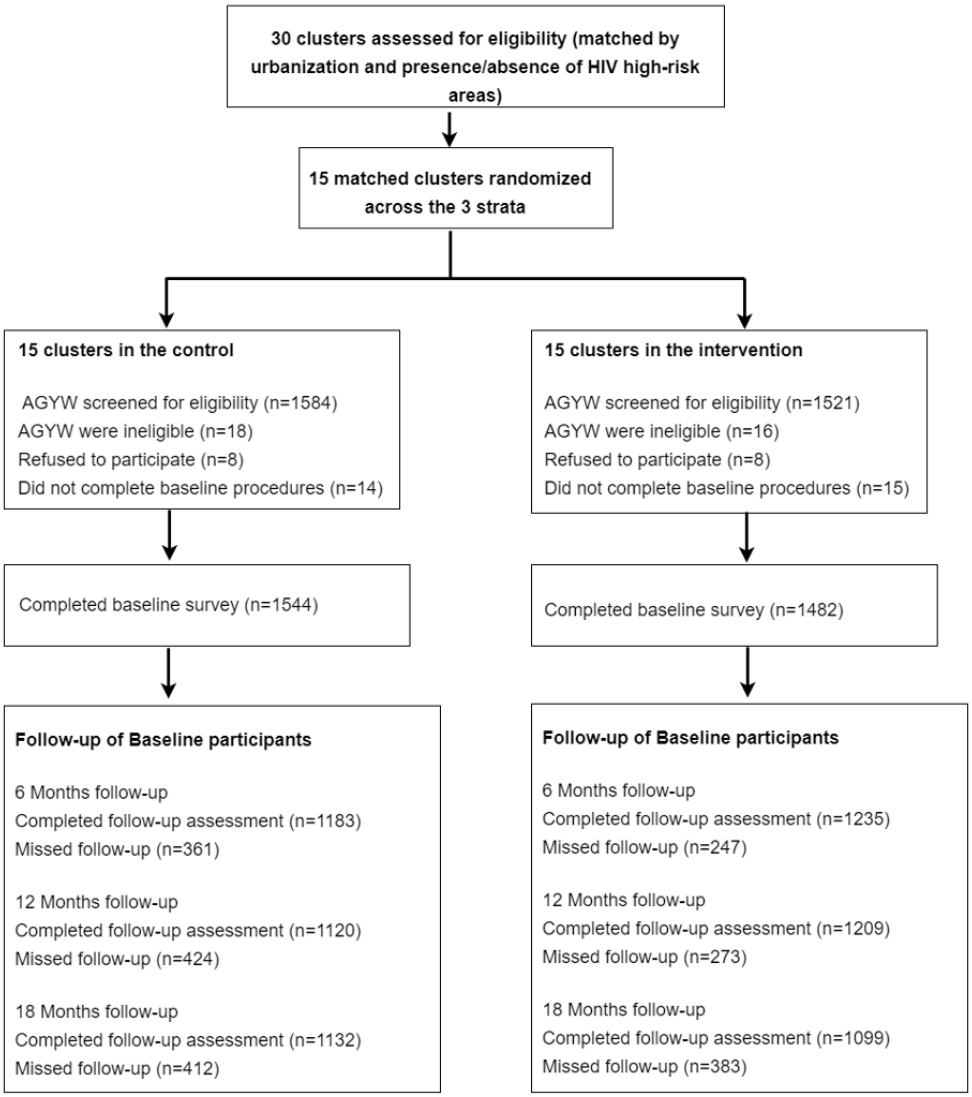


Table 1. Baseline sociodemographic characteristics of the study participants in a cluster randomized trial among adolescent girls and young women in the Shinyanga Region, Tanzania, in October 2017 to July 2019.

Characteristics	Total (N=3026)	Intervention group (n=1482)	Control group (n=1544)
Age (years), median (IQR)	20 (18-22)	20 (18-22)	19 (17-22)
Marital status, n (%)			
Single	1323 (43.7)	561 (37.9)	762 (49.4)
Married	1514 (50)	828 (55.9)	686 (44.4)
Separated, divorced, or widowed	189 (6.3)	93 (6.3)	96 (6.2)
Educational status, n (%)			
No formal/not completed primary school	856 (28.3)	402 (27.1)	454 (29.4)
Completed primary school	1609 (53.2)	790 (53.3)	819 (53)
Complete/incomplete secondary school	561 (18.5)	290 (19.6)	271 (17.6)
Had emotional/psychological support, n (%)	2387 (78.9)	1200 (81)	1187 (76.9)
Had debt at the time of the survey, n (%)	1119 (37)	616 (41.6)	503 (32.6)
Lacked food (went to bed hungry past 4 weeks), n (%)	702 (23.2)	309 (20.9)	393 (25.5)
HIV-positive status[a], n (%)	109 (3.6)	45 (3.1)	64 (4.2)
Herpes simplex virus type 2–positive status[b], n (%)	956 (32)	478 (32.8)	478 (31.2)
Reported sex work (past 6 months)[c], n (%)	387 (17)	183 (16.7)	204 (17.1)
Reported transactional sex (6 months)[c], n (%)	694 (30.4)	316 (28.9)	378 (31.8)
Reported compensated sex (6 months)[c], n (%)	792 (34.7)	365 (33.4)	427 (35.9)
Reported intergenerational sex (6 months)[c,d], n (%)	283 (12.6)	127 (11.9)	156 (13.3)
Reported concurrent partnerships (6 months)[c], n (%)	300 (13.1)	138 (12.6)	162 (13.6)
Reported condom use (nonmarital partner, 6 months)[c], n (%)	3 (0.1)	1 (0.1)	2 (0.2)
Reported sexual partner violence (6 months)[c], n (%)	817 (35.8)	377 (34.5)	440 (37)

[a]26 participants (18 in intervention and 8 in control) had either missing data or indeterminate results.

[b]36 participants (23 in intervention and 13 in control) had missing data.

[c]Restricted to those who were sexually active (n=2283), defined by self-reported vaginal or anal sex history (1093 in intervention and 1190 in control).

[d]48 participants (23 in intervention and 25 in control) had missing data.

Study Implementation

Of the 1373 AGYW in the intervention arm included in the longitudinal data analysis, 465 (33.9%), 373 (27.2%), and 414 (30.2%) received cash transfer 3 times or less, 4 times, and over 4 times, respectively, while 121 (8.8%) did not disclose the number of times they received cash transfer. HSV-2 incidence was 5.7/100 PYO (ie, 5.7 girls become HSV-2 positive for 100 years of observation/57 seroconverts for every 1000 years), 9.1/100 PYO, 9.6/100 PYO, and 4.8/100 PYO among AGYW who received cash transfer 3 times or less, 4 times, over 4 times, and among those with missing information on the number of times they received cash transfer, respectively.

Intervention Impact on HSV-2 Incidence and Other Outcomes

Two hundred incident HSV-2 infections were diagnosed in the study (98 in the control and 102 in the intervention), resulting in an annual incidence of 7.9/100 (95% CI 6.9/100-9.1/100) PYO. There was no significant difference in HSV-2 incidence between the study arms (adjusted hazard ratio 0.96, 95% CI 0.67-1.38; *P*=.83; Table 2), although location significantly modified the effect of cash transfer on HSV-2 incidence (effect modification *P*<.001). In urban and rural areas at high risk of HIV infection, cash transfer was associated with a nonsignificant increased hazard ratio adjusted for baseline confounders (Multimedia Appendix 1 and Multimedia Appendix 2). However, in rural areas with a low risk of HIV infection, the adjusted hazard ratio for incident HSV-2 infection in the cash transfer group was 0.45 (95% CI 0.29-0.71; *P*=.001; Multimedia Appendix 3).

There was no significant difference by study arm on HIV prevalence at follow-up and behavioral risk factors except for transactional sex and condom use with nonmarital sexual partners. The cash transfer intervention reduced the proportion of AGYW reporting transactional sex (adjusted OR 0.84, 0.73-0.96; *P*=.01), increased their savings (adjusted OR 1.87, 1.69-2.08; *P*<.001), and increased their utilization of community-based biomedical services (adjusted OR 2.10,

1.95-2.26; *P*<.001). There was no significant difference between the study groups in the proportion reporting compensated and intergenerational sex, sex work, and more than one sexual partner in the last 12 months. However, urbanization modified the effect of cash transfer intervention on savings (effect modification *P*<.001), utilization of biomedical services (*P*<.001), and sex work (*P*=.02). In urban areas, cash transfer was associated with a significant increase in reporting of sex work among those receiving the intervention compared to that in the control group adjusted for baseline confounders, although this was not the case in rural areas at high and low risks for HIV infection.

The 5 major items that AGYW spent their quarterly cash transfer on were starting a business (651/1373, 47.4%), toiletries (221/1373, 16.1%), savings (211/1373, 15.4%), food (143/1373, 10.4%), supporting dependents (114/1373, 8.3%), and other uses (33/1373, 2.4%). Only 3 social harm events were reported during the study—3 in the intervention and 0 in the control clusters—and were associated with minor teasing that cash transfer and phones provided to facilitate cash transfer may be associated with Freemasonry.

Table 2. Effect of cash transfer on primary and secondary outcomes in a cluster randomized trial among adolescent girls and young women in Shinyanga Region, Tanzania, in October 2017 to July 2019.

	Overall		Urban stratum		Rural high-risk stratum		Rural low-risk stratum	
	Intervention group	Control group	Intervention group	Control group	Intervention group	Control group	Intervention group	Control group
HSV-2 new cases out of total at risk, n/N (%)	102/930 (11)	98/925 (10.6)	29/310 (9.4)	17/327 (5.2)	45/322 (14)	21/261 (8)	28/298 (9.4)	60/337 (17.8)
Time of follow-up (years)	1293.9	1230.3	426.0	438.9	450.0	346.3	417.9	445.1
HSV-2 incidence (per 100 personal years of observation)	7.9	8.0	6.8	3.9	10.0	6.1	6.7	13.5
Hazard ratio[a]	0.96 (0.67-1.38)		1.55 (0.84-2.84)		1.60 (0.95-2.71)		0.45 (0.29-0.71)	
Hazard ratio *P* value	.83		.16		.08		.001	
P value for interaction	<.001		__[b]	—	—	—	—	—
New HIV cases out of total at risk, n/N (%)	11/1429 (0.8)	9/1478 (0.6)	—	—	—	—	—	—
Time of follow-up (years)	1954.7	1811.8	—	—	—	—	—	—
HIV incidence (per 1000 personal years of observation)	5.6	5.0	—	—	—	—	—	—
Hazard ratio[a]	0.78 (0.40-1.53)		—	—	—	—	—	—
Hazard ratio *P* value	.47		—	—	—	—	—	—
HIV prevalence (follow-up), n/N (%)	58/1466 (4)	74/1537 (4.8)	—	—	—	—	—	—
Risk ratio[a]	0.75 (0.52-1.08)		—	—	—	—	—	—
Risk ratio *P* value	.13		—	—	—	—	—	—
Compensated sex, n/N (%)	552/1373 (40.2)	580/1347 (43.1)	N/A[c]	N/A	N/A	N/A	N/A	N/A
Odds ratio[a]	0.91 (0.80-1.04)		N/A	N/A	N/A	N/A	N/A	N/A
Odds ratio *P* value	.15		N/A	N/A	N/A	N/A	N/A	N/A
P value for interaction	.46		—	—	—	—	—	—
Intergenerational sex, n/N (%)	374/1373 (27.2)	362/1347 (26.9)	N/A	N/A	N/A	N/A	N/A	N/A
Odds ratio[a]	0.91 (0.77-1.07)		N/A	N/A	N/A	N/A	N/A	N/A
Odds ratio *P* value	.25		N/A	N/A	N/A	N/A	N/A	N/A
P value for interaction	.88		—	—	—	—	—	—
Transactional sex, n/N (%)	486/1373 (35.4)	533/1347 (39.6)	N/A	N/A	N/A	N/A	N/A	N/A
Odds ratio[a]	0.84 (0.73-0.96)		N/A	N/A	N/A	N/A	N/A	N/A
Odds ratio *P* value	.01		N/A	N/A	N/A	N/A	N/A	N/A
P value for interaction	.50		—	—	—	—	—	—
Condom use (nonmarital)[d], n/N (%)	867/1284 (67.5)	738/1205 (61.2)	N/A	N/A	N/A	N/A	N/A	N/A
Odds ratio[a]	1.28 (1.16-1.42)		N/A	N/A	N/A	N/A	N/A	N/A
Odds ratio *P* value	<.001		N/A	N/A	N/A	N/A	N/A	N/A
P value for interaction	.22		—	—	—	—	—	—

	Overall		Urban stratum		Rural high-risk stratum		Rural low-risk stratum	
	Intervention group	Control group	Intervention group	Control group	Intervention group	Control group	Intervention group	Control group
>1 sexual partner (12 months), n/N (%)	211/1373 (15.4)	207/1347 (15.4)	N/A	N/A	N/A	N/A	N/A	N/A
Odds ratio[a]	0.97 (0.79-1.20)		N/A	N/A	N/A	N/A	N/A	N/A
Odds ratio *P* value	.80		N/A	N/A	N/A	N/A	N/A	N/A
P value for interaction	.12		—	—	—	—	—	—
Savings, n/N (%)	1275/1373 (92.9)	1140/1347 (84.6)	402/458 (87.8)	379/437 (86.7)	462/480 (96.3)	338/421 (80.3)	411/435 (94.5)	423/489 (86.5)
Odds ratio[a]	1.87 (1.69-2.08)		0.92 (0.76-1.11)		3.13 (2.63-3.72)		2.28 (1.91-2.71)	
Odds ratio *P* value	<.001		.40		<.001		<.001	
P value for interaction	<.001		—	—	—	—	—	—
Used biomedical services, n/N (%)	1161/1373 (84.6)	820/1347 (60.9)	393/458 (85.8)	225/437 (51.5)	362/480 (75.4)	275/421 (65.3)	406/435 (93.3)	320/489 (65.4)
Odds ratio[a]	2.10 (1.95-2.26)		2.63 (2.28-3.03)		1.35 (1.18-1.54)		2.42 (2.15-2.72)	
Odds ratio *P* value	<.001		<.001		<.001		<.001	
P value for interaction	<.001		—	—	—	—	—	—
Sexual partner violence, n/N (%)	711/1373 (51.8)	685/1347 (50.9)	N/A	N/A	N/A	N/A	N/A	N/A
Odds ratio[a]	0.94 (0.83-1.06)		N/A	N/A	N/A	N/A	N/A	N/A
Odds ratio *P* value	.29		N/A	N/A	N/A	N/A	N/A	N/A
P value for interaction	.83		—	—	—	—	—	—
Sex work, n/N (%)	310/1373 (22.6)	315/1347 (23.4)	87/458 (19)	58/437 (13.3)	114/480 (23.8)	121/421 (28.7)	109/435 (25.1)	136/489 (27.8)
Odds ratio[a]	1.06 (0.89-1.26)		1.66 (1.16-2.36)		0.90 (0.68-1.19)		0.92 (0.69-1.22)	
Odds ratio *P* value	.53		.005		.47		.56	
P value for interaction	.02		—	—	—	—	—	—

[a]Adjusted for matching pairs and variables that were significantly different between the arms at baseline.

[b]Not available.

[c]N/A: not applicable.

[d]Restricted to adolescent girls and young women who had follow-up data and reported sexual activity in the last 6 months.

Discussion

Overall, this study observed no significant effect of the quarterly cash transfer on HSV-2 incidence among out-of-school AGYW after 18 months of intervention. However, cash transfer was associated with a reduced incidence of HSV-2 infection in rural communities at low risk of HIV infections but not in urban and rural communities with high risk of HIV infections. This difference may be because of factors in urban and rural high-risk communities that may affect the effectiveness of cash transfer as a structural intervention for HIV infection. These factors may include less poverty, high mobility, and migration in the urban areas and rural high-risk communities where there are mines compared to rural low-risk areas. For instance, almost 81% of the low-income population in Tanzania reside in rural areas, depending on subsistence agriculture for their livelihood [45]. In 2015, 20.8% of the rural households were clustered in the lowest-income quartile against 3.7% of the urban households [46]. In this study, only 12% (11% urban, 9% rural high-risk villages vs 15% rural low-risk villages; *P*<.001) of the AGYW were living in households supported by the Government of Tanzania social action fund, which targets the lowest-income and vulnerable households [47]. Previous studies have shown that cash transfer has a significant impact on low-income communities as compared to that on high-income communities [48,49]. Since AGYW in rural areas have less access to jobs and other income-generating activities, they may engage in transactional sex to fulfil basic needs [50,51]. It is therefore likely that cash transfer was associated with a reduced incidence of HSV-2 infection in rural low-risk communities because a high proportion of AGYW in these communities were able to meet their basic needs through cash transfer and thus reduce risky sexual behavior [29]. Moreover, 10.6% (107/1014) of AGYW in urban areas, 10.6% (108/1021) of AGYW in rural high-risk areas, and 6.6% (65/991) of AGYW in rural low-risk

areas who participated in the baseline survey had migrated to other distant areas and were not seen in the subsequent follow-up rounds even after extensive tracing. Evidence suggests that mobile populations and migrants tend to be more vulnerable to sexually transmitted infections than nonmigrating populations [52,53]. Mining areas and towns surrounding mines attract a large number of male miners, thereby affecting the age-specific sex ratio in the area [54], altering the number of available sexual partners [55] and sexual networks [54,56] and placing AGYW in an environment conducive to practice transactional sex [56]. In this study, the cash transfer intervention was associated with higher reporting of sex work in urban areas and higher HSV-2 incidence rates among AGYW with more exposure due to higher cash transfer.

Urban and rural communities with many small-scale mining activities that put them at high risk of HIV infection have better asset ownership than rural communities [46,57]. In control communities, the incidence of HSV-2 was 13.5/100 PYO in the rural low-risk stratum, 6.1/100 PYO in the rural high-risk stratum, and 3.9/100 PYO in the urban areas. The high incidence in rural remote villages may be because of small densely connected sexual networks, which have been shown to be highly effective in spreading viral sexually transmitted infections [58,59]. Owing to the high prevalence of HSV-2 at baseline in these communities, AGYW selecting new sexual partners in rural areas are more likely to choose a partner who is HSV-2 positive because of their remoteness and small population size [58,59], unlike in rural high-risk and urban areas where there are large-scale migrations and mobility. Finally, AGYW in remote rural villages may also have limited access to health care, especially the management of viral sexually transmitted infections, which may elevate their risk of HSV-2 infection [60].

The cash transfer intervention did not reduce HIV incidence in the intervention arm compared to that in the control arm. Most of the evidences related to the impact of cash transfer on HIV prevention among AGYW in sub-Saharan Africa have been reported in studies conducted among adolescent schoolgirls and mostly in rural areas [22]. Keeping young girls in school is associated with a reduced risk of HIV infection [29]. Our study was conducted among out-of-school AGYW who were economically vulnerable and therefore more likely to be at high risk of HIV infections compared to economically nonvulnerable AGYW. The economic vulnerability of AGYW who dropped out of school may be caused by factors that led them to leave school, such as lack of financial support, loss of parents, sick parents, pregnancy, or early marriage [61]. Thus, the cash amount provided through the cash transfer intervention may not have been sufficient to bring the AGYW out of the risk behavior.

The baseline findings from this study showed HSV-2 prevalence of 32% (956/2990), which is comparable to that reported in earlier studies conducted in similar settings [43,62,63]. HSV-2 infection increases the risk of acquiring HIV infection up to 5-fold [64]. The high baseline HSV-2 prevalence is probably because of several large-scale and small-scale gold mining activities with a large number of men attracting many economically vulnerable young women, creating a niche for transactional sex and sexual mixing where the partnership is formed between partners with different HIV risk profiles [65]. The high HSV-2 prevalence observed in this study indicates the need for continued targeted prevention efforts among AGYW to saturate the region with interventions to reduce new infections. Our study contributes to the literature on cash transfer, as it reports the synergetic effect of cash transfer as a social protection scheme along with CHP among out-of-school girls where HIV infection is the highest. To our knowledge, only the adolescent girls' initiative study in Kenya has been implemented in this group in sub-Saharan Africa [66]. Previous cash transfer interventions have produced mixed results, with some demonstrating impact on preventing sexually transmitted infections among school girls by delaying sexual activity [22,24,25]. Our study shows that cash transfer programs targeting the low-income and highly vulnerable populations in rural areas are more likely to reduce risky sexual behavior among AGYW. However, more studies are needed to evaluate the effect of cash transfer among out-of-school AGYW aged 15-24 years in different settings and the different amounts of cash transfer because of high HIV infection rates in this subpopulation.

This study has several strengths. First, we used ACASI to collect sexual behavior data and other sensitive data. Studies comparing face-to-face interviews with ACASI have reported that respondents are more likely to be open and honest when using ACASI in reporting sensitive information [67,68]. Second, to assess the effect of cash transfer, our study collected longitudinal data on biomedical, behavioral, and structural interventions on a large sample of AGYW in urban and rural areas who were out of school and therefore vulnerable to HIV infections. Third, the CHP intervention package was developed after broad consultation and engagement with local leaders; the Ministry of Health, Community Development, Gender, Elderly and Children; Government of Tanzania social action fund; National AIDS Control Program; Tanzania Commission for HIV/AIDS; mobile communication companies; and CSOs serving AGYW and AGYW representatives among other stakeholders. These dialogues led to developing a tailored and prioritized CHP intervention package to AGYW offered by the Sauti project, including the cash transfer amount and payment modality. The involvement of CSOs serving AGYW and AGYW representatives was crucial in advising on the content of the study materials and data collection techniques such as ACASI. CSOs were crucial in the recruitment of AGYW for CHP interventions alongside the recruitment of study participants and monitoring of study activities and tracing of study participants. CSOs were also involved in the dissemination of the study findings at the community and regional levels. A limitation of this study is that the clusters were selected only from the Shinyanga Region, as the prevalence of HIV in this region was higher than the national average, and thus, these findings may not be generalizable to other study regions with lower HIV prevalence in Tanzania.

In conclusion, this trial showed no significant impact of the cash transfer intervention on HSV-2 incidence among AGYW in Tanzania. Although the intervention appears to have reduced HSV-2 incidence among AGYW in rural low-risk communities,

this effect was not observed in urban high-risk communities. Factors such as less poverty and more asset ownership in urban and rural high-risk (mining) communities may have undermined the effect of cash transfer.

Acknowledgments

This study was funded by US President's Emergency Plan for AIDS Relief through the US Agency for International Development under Cooperative Agreement (AID-621-A-15-00003). We are grateful to the participants, data collection team, Sauti project, civil society organization's implementing services, and health authorities in Tanzania for their dedication to this study.

Authors' Contributions

E Kuringe, AC, MD, EM, CC, DM, AK, DN, SN, AS, JC, and MW contributed to the study design. E Kuringe, JM, EM, CC, DM, GM, E Kalage, AK, DN, SN, AS, JC, and MW oversaw the implementation of the intervention and control strategies. DN, GM, JM, E Kuringe, SN, AS, E Kalage, EM, and MW managed the data collection. JM, GM, and MW performed the data analysis. E Kuringe, AC, and MW wrote the first draft of the manuscript. All authors contributed to the interpretation of the results and critical revision of the manuscript for important intellectual content.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Kaplan-Meier curve for herpes simplex virus type 2 seroconversion in urban areas in a cluster randomized trial among adolescent girls and young women in Shinyanga region, Tanzania, in October 2017-July 2019.
[PNG File , 20 KB - publichealth_v8i9e30372_app1.png]

Multimedia Appendix 2
Kaplan-Meier curve for herpes simplex virus type 2 seroconversion in rural areas at high HIV risk in a cluster randomized trial among adolescent girls and young women in Shinyanga region, Tanzania, in October 2017-July 2019.
[PNG File , 20 KB - publichealth_v8i9e30372_app2.png]

Multimedia Appendix 3
Kaplan-Meier curve for herpes simplex virus type 2 seroconversion in rural areas at low HIV risk in a cluster randomized trial among adolescent girls and young women in Shinyanga region, Tanzania, in October 2017-July 2019.
[PNG File , 21 KB - publichealth_v8i9e30372_app3.png]

Multimedia Appendix 4
CONSORT EHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1265 KB - publichealth_v8i9e30372_app4.pdf]

References

1. Gouws E, Stanecki KA, Lyerla R, Ghys PD. The epidemiology of HIV infection among young people aged 15-24 years in southern Africa. AIDS 2008 Dec;22 Suppl 4:S5-16. [doi: 10.1097/01.aids.0000341773.86500.9d] [Medline: 19033755]
2. Govender K, Masebo WGB, Nyamaruze P, Cowden RG, Schunter BT, Bains A. HIV Prevention in Adolescents and Young People in the Eastern and Southern African Region: A Review of Key Challenges Impeding Actions for an Effective Response. Open AIDS J 2018;12:68 [FREE Full text] [doi: 10.2174/1874613601812010068] [Medline: 30197723]
3. Dellar RC, Dlamini S, Karim QA. Adolescent girls and young women: key populations for HIV epidemic control. J Int AIDS Soc 2015;18(2 Suppl 1):19408 [FREE Full text] [doi: 10.7448/IAS.18.2.19408] [Medline: 25724504]
4. Saul J, Bachman G, Allen S, Toiv NF, Cooney C, Beamon T. The DREAMS core package of interventions: A comprehensive approach to preventing HIV among adolescent girls and young women. PLoS One 2018;13(12):e0208167 [FREE Full text] [doi: 10.1371/journal.pone.0208167] [Medline: 30532210]
5. Tanzania HIV impact survey (THIS) 2016-2017: final report. Tanzania Commission for AIDS. 2018. URL: https://phia.icap.columbia.edu/wp-content/uploads/2020/02/FINAL_THIS-2016-2017_Final-Report__06.21.19_for-web_TS.pdf [accessed 2022-08-03]
6. The gap report. The Joint United Nations Programme on HIV/AIDS. 2014. URL: https://www.unaids.org/sites/default/files/media_asset/UNAIDS_Gap_report_en.pdf [accessed 2022-09-03]
7. Yi TJ, Shannon B, Prodger J, McKinnon L, Kaul R. Genital immunology and HIV susceptibility in young women. Am J Reprod Immunol 2013 Mar;69 Suppl 1:74-79. [doi: 10.1111/aji.12035] [Medline: 23157424]

8. Hwang LY, Scott ME, Ma Y, Moscicki A. Higher levels of cervicovaginal inflammatory and regulatory cytokines and chemokines in healthy young women with immature cervical epithelium. J Reprod Immunol 2011 Jan;88(1):66-71 [FREE Full text] [doi: 10.1016/j.jri.2010.07.008] [Medline: 21051089]
9. Abdool Karim Q, Sibeko S, Baxter C. Preventing HIV infection in women: a global health imperative. Clin Infect Dis 2010 May 15;50 Suppl 3:S122-S129 [FREE Full text] [doi: 10.1086/651483] [Medline: 20397940]
10. Mabaso M, Sokhela Z, Mohlabane N, Chibi B, Zuma K, Simbayi L. Determinants of HIV infection among adolescent girls and young women aged 15-24 years in South Africa: a 2012 population-based national household survey. BMC Public Health 2018 Jan 26;18(1):183 [FREE Full text] [doi: 10.1186/s12889-018-5051-3] [Medline: 29373958]
11. Stoner MCD, Nguyen N, Kilburn K, Gómez-Olivé FX, Edwards JK, Selin A, et al. Age-disparate partnerships and incident HIV infection in adolescent girls and young women in rural South Africa. AIDS 2019 Jan 27;33(1):83-91 [FREE Full text] [doi: 10.1097/QAD.0000000000002037] [Medline: 30289813]
12. Swartzendruber A, Zenilman JM, Niccolai LM, Kershaw TS, Brown JL, Diclemente RJ, et al. It takes 2: partner attributes associated with sexually transmitted infections among adolescents. Sex Transm Dis 2013 May;40(5):372-378 [FREE Full text] [doi: 10.1097/OLQ.0b013e318283d2c9] [Medline: 23588126]
13. Ritchwood TD, Hughes JP, Jennings L, MacPhail C, Williamson B, Selin A, et al. Characteristics of Age-Discordant Partnerships Associated With HIV Risk Among Young South African Women (HPTN 068). J Acquir Immune Defic Syndr 2016 Aug 01;72(4):423-429 [FREE Full text] [doi: 10.1097/QAI.0000000000000988] [Medline: 26977748]
14. Gottert A, Pulerwitz J, Siu G, Katahoire A, Okal J, Ayebare F, et al. Male partners of young women in Uganda: Understanding their relationships and use of HIV testing. PLoS One 2018;13(8):e0200920 [FREE Full text] [doi: 10.1371/journal.pone.0200920] [Medline: 30096147]
15. Becker ML, Bhattacharjee P, Blanchard JF, Cheuk E, Isac S, Musyoki HK, et al. Vulnerabilities at First Sex and Their Association With Lifetime Gender-Based Violence and HIV Prevalence Among Adolescent Girls and Young Women Engaged in Sex Work, Transactional Sex, and Casual Sex in Kenya. J Acquir Immune Defic Syndr 2018 Nov 01;79(3):296-304 [FREE Full text] [doi: 10.1097/QAI.0000000000001826] [Medline: 30113403]
16. Butts SA, Parmley LE, Alcaide ML, Rodriguez VJ, Kayukwa A, Chitalu N, et al. Let us fight and support one another: adolescent girls and young women on contributors and solutions to HIV risk in Zambia. Int J Womens Health 2017;9:727-737 [FREE Full text] [doi: 10.2147/IJWH.S142232] [Medline: 29033613]
17. Sherwood J, Sharp A, Cooper B, Roose-Snyder B, Blumenthal S. HIV/AIDS National Strategic Plans of Sub-Saharan African countries: an analysis for gender equality and sex-disaggregated HIV targets. Health Policy Plan 2017 Dec 01;32(10):1361-1367 [FREE Full text] [doi: 10.1093/heapol/czx101] [Medline: 28973358]
18. Economic strengthening for female sex workers: A review of the literature. fhi360. 2014. URL: https://www.fhi360.org/sites/default/files/media/documents/Economic_Strengthening_for_Female_Sex_Workers.pdf [accessed 2022-09-03]
19. Kim J, Ferrari G, Abramsky T, Watts C, Hargreaves J, Morison L, et al. Assessing the incremental effects of combining economic and health interventions: the IMAGE study in South Africa. Bull World Health Organ 2009 Nov;87(11):824-832 [FREE Full text] [doi: 10.2471/blt.08.056580] [Medline: 20072767]
20. Sumartojo E, Doll L, Holtgrave D, Gayle H, Merson M. Enriching the mix: incorporating structural factors into HIV prevention. AIDS 2000 Jun;14 Suppl 1:S1-S2. [doi: 10.1097/00002030-200006001-00001] [Medline: 10981468]
21. Gibbs A, Willan S, Misselhorn A, Mangoma J. Combined structural interventions for gender equality and livelihood security: a critical review of the evidence from southern and eastern Africa and the implications for young people. J Int AIDS Soc 2012 Jun 14;15 Suppl 1:1-10 [FREE Full text] [doi: 10.7448/IAS.15.3.17362] [Medline: 22713350]
22. Baird SJ, Garfein RS, McIntosh CT, Ozler B. Effect of a cash transfer programme for schooling on prevalence of HIV and herpes simplex type 2 in Malawi: a cluster randomised trial. Lancet 2012 Apr 07;379(9823):1320-1329. [doi: 10.1016/S0140-6736(11)61709-1] [Medline: 22341825]
23. Gorgens M, Longosz AF, Ketende S, Nkambule M, Dlamini T, Mabuza M, et al. Evaluating the effectiveness of incentives to improve HIV prevention outcomes for young females in Eswatini: Sitakhela Likusasa impact evaluation protocol and baseline results. BMC Public Health 2020 Oct 22;20(1):1591 [FREE Full text] [doi: 10.1186/s12889-020-09680-8] [Medline: 33092558]
24. de Walque D, Dow WH, Nathan R, Abdul R, Abilahi F, Gong E, et al. Incentivising safe sex: a randomised trial of conditional cash transfers for HIV and sexually transmitted infection prevention in rural Tanzania. BMJ Open 2012;2:e000747 [FREE Full text] [doi: 10.1136/bmjopen-2011-000747] [Medline: 22318666]
25. Björkman-Nyqvist M, Corno L, Walque DD, Svensson J. P4.120 Evaluating the impact of short term financial incentives on HIV and STI incidence among youth in Lesotho: A randomised trial. In: Sex Transm Infect. 2013 Presented at: Abstract TUPDC0106. 7th IAS Conference on HIV Pathogenesis, Treatment and Prevention; July 13; Kuala Lumpur, Malaysia. [doi: 10.1136/sextrans-2013-051184.1017]
26. Gorgens M, Ketende S, Tsododo V, Heard W, Mabuza M, Longosz A. Results of a cluster randomized control trial (cRCT) of financial incentives for HIV prevention among adolescent girls and young women (AGYW) in Eswatini. 2019 Presented at: IAS Conference on HIV Science; July 23; Mexico City URL: https://programme.ias2019.org/Abstract/Abstract/4943

27. Abdool KQ, Leask K, Kharsany A, Humphries H, Ntombela F, Samsunder N. Impact of conditional cash incentives on HSV-2 and HIV prevention in rural South African high school students: results of the CAPRISA 007 cluster randomized controlled trial. Journal of the International AIDS Society 2015;18:77-89. [doi: 10.1007/978-3-319-47518-9_6]
28. Abdool KQ. Impact of conditional cash incentives on HSV-2 and HIV prevention in rural South African high school students: results of the CAPRISA 007 cluster randomized controlled trial. Journal of the International AIDS Society 2015:18. [doi: 10.7448/IAS.18.5.20547]
29. Pettifor A, MacPhail C, Hughes JP, Selin A, Wang J, Gómez-Olivé FX, et al. The effect of a conditional cash transfer on HIV incidence in young women in rural South Africa (HPTN 068): a phase 3, randomised controlled trial. Lancet Glob Health 2016 Dec;4(12):e978-e988 [FREE Full text] [doi: 10.1016/S2214-109X(16)30253-4] [Medline: 27815148]
30. Kohler H, Thornton R. Conditional Cash Transfers and HIV/AIDS Prevention: Unconditionally Promising? World Bank Econ Rev 2012 Jun 01;26(2):165-190 [FREE Full text] [doi: 10.1093/wber/lhr041] [Medline: 24319306]
31. Hallfors DD, Cho H, Rusakaniko S, Mapfumo J, Iritani B, Zhang L, et al. The impact of school subsidies on HIV-related outcomes among adolescent female orphans. J Adolesc Health 2015 Jan;56(1):79-84 [FREE Full text] [doi: 10.1016/j.jadohealth.2014.09.004] [Medline: 25530603]
32. Cho H, Mbai I, Luseno WK, Hobbs M, Halpern C, Hallfors DD. School Support as Structural HIV Prevention for Adolescent Orphans in Western Kenya. J Adolesc Health 2018 Jan;62(1):44-51 [FREE Full text] [doi: 10.1016/j.jadohealth.2017.07.015] [Medline: 29107569]
33. Wambura M, Drake M, Kuringe E, Majani E, Nyato D, Casalini C, et al. Cash Transfer to Adolescent Girls and Young Women to Reduce Sexual Risk Behavior (CARE): Protocol for a Cluster Randomized Controlled Trial. JMIR Res Protoc 2019 Dec 20;8(12):e14696 [FREE Full text] [doi: 10.2196/14696] [Medline: 31859686]
34. Plummer ML, Ross DA, Wight D, Changalucha J, Mshana G, Wamoyi J, et al. "A bit more truthful": the validity of adolescent sexual behaviour data collected in rural northern Tanzania using five methods. Sex Transm Infect 2004 Dec;80 Suppl 2:ii49-ii56 [FREE Full text] [doi: 10.1136/sti.2004.011924] [Medline: 15572640]
35. Palen L, Smith EA, Caldwell LL, Flisher AJ, Wegner L, Vergnani T. Inconsistent reports of sexual intercourse among South African high school students. J Adolesc Health 2008 Mar;42(3):221-227 [FREE Full text] [doi: 10.1016/j.jadohealth.2007.08.024] [Medline: 18295129]
36. Hayes RJ, Donnell D, Floyd S, Mandla N, Bwalya J, Sabapathy K, HPTN 071 (PopART) Study Team. Effect of Universal Testing and Treatment on HIV Incidence - HPTN 071 (PopART). N Engl J Med 2019 Jul 18;381(3):207-218 [FREE Full text] [doi: 10.1056/NEJMoa1814556] [Medline: 31314965]
37. Torgerson DJ. Contamination in trials: is cluster randomisation the answer? BMJ 2001 Mar 10;322(7282):355-357 [FREE Full text] [doi: 10.1136/bmj.322.7282.355] [Medline: 11159665]
38. DiGuiseppi C, Coupland C. The design and use of cluster randomised controlled trials in evaluating injury prevention interventions: part 1. Rationale, design and informed consent. Inj Prev 2010 Mar;16(1):61-67. [doi: 10.1136/ip.2009.023119] [Medline: 20179039]
39. Campbell MK, Piaggio G, Elbourne DR, Altman DG, CONSORT Group. Consort 2010 statement: extension to cluster randomised trials. BMJ 2012 Sep 04;345:e5661. [doi: 10.1136/bmj.e5661] [Medline: 22951546]
40. NACP. National comprehensive guidelines for HIV testing services in Tanzania. The United Republic of Tanzania Ministry of Health National Aids Control Programme. 2019. URL: https://nacp.go.tz/download/national-comprehensive-guidelines-on-hiv-testing-services/ [accessed 2022-09-03]
41. Hayes RJ, Moulton LH. Cluster Randomised Trials. New York: Chapman and Hall/CRC; 2017.
42. Grosskurth H, Mosha F, Todd J, Senkoro K, Newell J, Klokke A, et al. A community trial of the impact of improved sexually transmitted disease treatment on the HIV epidemic in rural Tanzania: 2. Baseline survey results. AIDS 1995 Aug;9(8):927-934. [doi: 10.1097/00002030-199508000-00015] [Medline: 7576329]
43. Kapiga SH, Ewings FM, Ao T, Chilongani J, Mongi A, Baisley K, et al. The epidemiology of HIV and HSV-2 infections among women participating in microbicide and vaccine feasibility studies in Northern Tanzania. PLoS One 2013;8(7):e68825 [FREE Full text] [doi: 10.1371/journal.pone.0068825] [Medline: 23874780]
44. Leyrat C, Morgan KE, Leurent B, Kahan BC. Cluster randomized trials with a small number of clusters: which analyses should be used? Int J Epidemiol 2018 Feb 01;47(1):321-331. [doi: 10.1093/ije/dyx169] [Medline: 29025158]
45. Tanzania mainland household budget survey 2017-18: key indicators report. Tanzania National Bureau of Statistics. 2019. URL: https://www.nbs.go.tz/index.php/en/census-surveys/poverty-indicators-statistics/household-budget-survey-hbs/653-household-budget-survey-2017-18-tanzania-mainland-final-report [accessed 2022-06-21]
46. Maliti E. Inequality in Education and Wealth in Tanzania: A 25-Year Perspective. Soc Indic Res 2018 Jan 29;145(3):901-921. [doi: 10.1007/s11205-018-1838-y]
47. Rosas N, Zaldivar S, Granata M, Lertsuridej G, Wilson N, Chuwa A. Evaluating Tanzania's productive social safety net: findings from the midline survey. Technical Report 2016 Nov:21 [FREE Full text] [doi: 10.13140/RG.2.2.12257.97120]
48. Kakwani N, Soares F, Son H. Cash transfers for school age children in African countries: simulation of impacts on poverty and school attendance. Development Policy Review 2006:5 [FREE Full text] [doi: 10.1111/j.1467-7679.2006.00347.x]

49. Olken B, Onishi J, Wong S. Indonesia's PNPM Generasi program: final impact evaluation report. World Bank. 2011. URL: https://openknowledge.worldbank.org/bitstream/handle/10986/21595/691420REVISED00aluation0Report02011.pdf?sequence=1&isAllowed=y [accessed 2022-09-03]
50. Wamoyi J, Fenwick A, Urassa M, Zaba B, Stones W. "Women's bodies are shops": beliefs about transactional sex and implications for understanding gender power and HIV prevention in Tanzania. Arch Sex Behav 2011 Mar;40(1):5-15. [doi: 10.1007/s10508-010-9646-8] [Medline: 20652390]
51. Wamoyi J, Stobeanau K, Bobrova N, Abramsky T, Watts C. Transactional sex and risk for HIV infection in sub-Saharan Africa: a systematic review and meta-analysis. J Int AIDS Soc 2016;19(1):20992 [FREE Full text] [doi: 10.7448/IAS.19.1.20992] [Medline: 27809960]
52. Camlin CS, Cassels S, Seeley J. Bringing population mobility into focus to achieve HIV prevention goals. J Int AIDS Soc 2018 Jul;21 Suppl 4:e25136 [FREE Full text] [doi: 10.1002/jia2.25136] [Medline: 30027588]
53. Camlin CS, Charlebois ED. Mobility and its Effects on HIV Acquisition and Treatment Engagement: Recent Theoretical and Empirical Advances. Curr HIV/AIDS Rep 2019 Aug;16(4):314-323 [FREE Full text] [doi: 10.1007/s11904-019-00457-2] [Medline: 31256348]
54. Quinn TC. Population migration and the spread of types 1 and 2 human immunodeficiency viruses. Proc Natl Acad Sci U S A 1994 Mar 29;91(7):2407-2414 [FREE Full text] [doi: 10.1073/pnas.91.7.2407] [Medline: 8146131]
55. Decosas J, Pedneault V. Women and AIDS in Africa: demographic implications for health promotion. Health Policy Plan 1992;7(3):227-233. [doi: 10.1093/heapol/7.3.227]
56. Decosas J, Kane F, Anarfi JK, Sodji KD, Wagner HU. Migration and AIDS. Lancet 1995 Sep 23;346(8978):826-828. [doi: 10.1016/s0140-6736(95)91631-8] [Medline: 7674750]
57. Tanzania mainland poverty assessment. World Bank. 2019. URL: https://openknowledge.worldbank.org/bitstream/handle/10986/33031/Executive-Summary.pdf?sequence=6&isAllowed=y [accessed 2022-06-21]
58. Carnegie NB, Morris M. Size matters: concurrency and the epidemic potential of HIV in small networks. PLoS One 2012;7(8):e43048 [FREE Full text] [doi: 10.1371/journal.pone.0043048] [Medline: 22937011]
59. Hazel A, Holland Jones J. Remoteness influences access to sexual partners and drives patterns of viral sexually transmitted infection prevalence among nomadic pastoralists. PLoS One 2018;13(1):e0191168 [FREE Full text] [doi: 10.1371/journal.pone.0191168] [Medline: 29385170]
60. WHO guidelines for the treatment of genital herpes simplex virus. World Health Organization. 2016. URL: https://www.who.int/publications-detail-redirect/9789241549987 [accessed 2022-06-21]
61. Global initiative on out-of-school children: Tanzania country study. UNICEF. 2018. URL: https://www.unicef.org/tanzania/media/596/file/Tanzania-2018-Global-Initiative-Out-of-School-Children-Country-Report.pdf [accessed 2022-06-21]
62. Watson-Jones D, Weiss HA, Rusizoka M, Baisley K, Mugeye K, Changalucha J, et al. Risk factors for herpes simplex virus type 2 and HIV among women at high risk in northwestern Tanzania: preparing for an HSV-2 intervention trial. J Acquir Immune Defic Syndr 2007 Dec 15;46(5):631-642 [FREE Full text] [doi: 10.1097/QAI.0b013e31815b2d9c] [Medline: 18043318]
63. Tassiopoulos KK, Seage G, Sam N, Kiwelu I, Shao J, Ao TTH, et al. Predictors of herpes simplex virus type 2 prevalence and incidence among bar and hotel workers in Moshi, Tanzania. J Infect Dis 2007 Mar 15;195(4):493-501. [doi: 10.1086/510537] [Medline: 17230408]
64. Looker KJ, Elmes JAR, Gottlieb SL, Schiffer JT, Vickerman P, Turner KME, et al. Effect of HSV-2 infection on subsequent HIV acquisition: an updated systematic review and meta-analysis. Lancet Infect Dis 2017 Dec;17(12):1303-1316 [FREE Full text] [doi: 10.1016/S1473-3099(17)30405-X] [Medline: 28843576]
65. Gregson S, Nyamukapa CA, Garnett GP, Mason PR, Zhuwau T, Caraël M, et al. Sexual mixing patterns and sex-differentials in teenage exposure to HIV infection in rural Zimbabwe. Lancet 2002 Jun 01;359(9321):1896-1903. [doi: 10.1016/S0140-6736(02)08780-9] [Medline: 12057552]
66. Stoner MCD, Kilburn K, Godfrey-Faussett P, Ghys P, Pettifor AE. Cash transfers for HIV prevention: A systematic review. PLoS Med 2021 Nov;18(11):e1003866 [FREE Full text] [doi: 10.1371/journal.pmed.1003866] [Medline: 34843468]
67. Estes LJ, Lloyd LE, Teti M, Raja S, Bowleg L, Allgood KL, et al. Perceptions of audio computer-assisted self-interviewing (ACASI) among women in an HIV-positive prevention program. PLoS One 2010 Mar 10;5(2):e9149 [FREE Full text] [doi: 10.1371/journal.pone.0009149] [Medline: 20161771]
68. Morrison-Beedy D, Carey MP, Tu X. Accuracy of audio computer-assisted self-interviewing (ACASI) and self-administered questionnaires for the assessment of sexual behavior. AIDS Behav 2006 Sep;10(5):541-552 [FREE Full text] [doi: 10.1007/s10461-006-9081-y] [Medline: 16721506]

Abbreviations

ACASI: audio computer-assisted self-interview
AGYW: adolescent girls and young women
CHP: combination HIV prevention
CONSORT: Consolidated Standards of Reporting Trials

CSO: civil society organization
DREAMS: Determined, Resilient, Empowered, AIDS-free, Mentored, and Safe
HSV-2: herpes simplex virus type 2
OR: odds ratio
PYO: personal years of observation
SBCC: social and behavior change communication
WORTH+: Women Organizing Resources Together plus

Edited by T Sanchez, A Mavragani; submitted 12.05.21; peer-reviewed by AS Ibrahim, K Waters, A Chwałczyńska; comments to author 23.01.22; revised version received 19.05.22; accepted 02.08.22; published 19.09.22.

Please cite as:
Kuringe E, Christensen A, Materu J, Drake M, Majani E, Casalini C, Mjungu D, Mbita G, Kalage E, Komba A, Nyato D, Nnko S, Shao A, Changalucha J, Wambura M
Effectiveness of Cash Transfer Delivered Along With Combination HIV Prevention Interventions in Reducing the Risky Sexual Behavior of Adolescent Girls and Young Women in Tanzania: Cluster Randomized Controlled Trial
JMIR Public Health Surveill 2022;8(9):e30372
URL: https://publichealth.jmir.org/2022/9/e30372
doi:10.2196/30372
PMID:36121686

Original Paper

Spatiotemporal Analysis of Online Purchase of HIV Self-testing Kits in China, 2015-2017: Longitudinal Observational Study

Yi Lv[1*], PhD; Qiyu Zhu[1*], MPH; Chengdong Xu[2*], PhD; Guanbin Zhang[3,4*], PhD; Yan Jiang[5], MD, PhD; Mengjie Han[1], MPH; Cong Jin[1], PhD

[1]National Center for AIDS/STD Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing, China

[2]State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China

[3]National Engineering Research Center for Beijing Biochip Technology, Beijing, China

[4]Department of Infectious Diseases, 404 Hospital of Mianyang, Sichuan, China

[5]Chinese Association of STD and AIDS Prevention and Control, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Cong Jin, PhD
National Center for AIDS/STD Control and Prevention
Chinese Center for Disease Control and Prevention
155 Changbai Road
Changping District
Beijing, 102206
China
Phone: 86 10 58900995
Email: jinc@chinaaids.cn

Abstract

Background: Since the introduction of HIV self-testing by UNAIDS in 2014, the practice has been extensively implemented around the world. HIV self-testing (HIVST) was developed in China around 2015, and the online purchase of HIVST kits through e-commerce platforms has since become the most important delivery method for self-testing, with advantages such as user-friendliness, speed, and better privacy protection.

Objective: Understanding the spatiotemporal characteristics of online HIVST kit purchasing behavior and identifying potential impacting factors will help promote the HIV self-testing strategy.

Methods: The online retail data of HIVST kits from the 2 largest e-commerce platforms in China from 2015 to 2017 were collected for this study. The Bayesian spatiotemporal hierarchical model was used to investigate the spatiotemporal characteristics of online purchased HIVST kits. Ordinary least squares regression was used to identify potential factors associated with online purchase, including GDP per capita, population density, road density, HIV screening laboratory density, and newly diagnosed HIV/AIDS cases per 100,000 persons. The *q* statistics calculated by Geodetector were used to determine the interactive effect of every 2 factors on the online purchase.

Results: The online purchase of HIVST kits increased rapidly in China from 2015 to 2017, with annual peak sales in May and December. Five economically superior regions in China, Pearl River Delta, Yangtze River Delta, Chengdu and surrounding areas, Beijing and Tianjin areas, and Shandong Peninsula, showed a comparatively higher spatial preference for online purchased HIVST kits. The GDP per capita (P<.001) and the rate of newly diagnosed HIV/AIDS cases per 100,000 persons (P<.001) were identified as 2 factors positively associated with online purchase. Among the factors we investigated in this study, 2 factors associated with online purchase, GDP per capita and the rate of newly diagnosed HIV/AIDS cases per 100,000 persons, also displayed the strongest interactive effect, with a *q* value of 0.66.

Conclusions: Individuals in better-off areas are more inclined to purchase HIVST kits online. In addition to economic status, the severity of the HIV epidemic is also a factor influencing the online purchase of HIVST kits.

(JMIR Public Health Surveill 2022;8(9):e37922) doi:10.2196/37922

KEYWORDS

spatiotemporal; characteristics; online; purchase; HIV; self-testing; e-commerce; economic status; HIV epidemic; China

Introduction

To help end the AIDS epidemic, the Joint United Nations Program on HIV/AIDS (UNAIDS) proposed the 90-90-90 target in 2014 [1]. This ambitious but achievable target of 90% of people living with HIV being aware of their status [1] and that HIV testing services (HTS) play important roles and serve as the gateway to treatment, prevention, and care [2]. HIV self-testing (HIVST), defined as a type of testing strategy in which sample collection, testing, and interpretation are all performed by individuals who wish to learn about their HIV status on their own in a private environment, is known as a confidential way of HIV testing [3]. The interest in the use of HIV rapid diagnostic tests (RDTs) for self-testing has increased worldwide since 2015 [2]. As a private and convenient method, HIVST has been widely accepted by diverse populations, including those who might not otherwise be tested due to stigma and confidentiality concerns [4-6]. HIVST is especially cost-effective in resource-limited areas and has proven to be an effective way to expand HIV testing service [7]. As of June 2020, 41 countries around the world had implemented HIVST. In addition, 45 countries had allowed HIV self-testing policies [8].

The provider-initiated HIV testing and counseling (PITC) and voluntary counseling and testing (VCT) have been routinely provided in China. Since the World Health Organization advocated HIV self-testing in 2015 [2], a series of national policies on HIVST were implemented in China. Because HIVST kits belong to the third category of medical devices in China, under the regulations of the National Medical Products Administration, only a small number of pharmacies that had the business and sales qualification of the third category of medical devices were qualified to sell HIVST kits. The requirements of online HIVST retail are also very strict. To sell HIVST kits online, retailers need the business and sales qualification of the third category of medical devices and the business qualification of internet medical devices [9]. However, the logistics in China developed fast, and online retail could meet the purchasing needs of a wider area. Therefore, online purchase of HIVST kits through e-commerce platforms has become the most important form of delivery for private self-testing in China, with the advantages of being user-friendly, fast, and privacy protection. Purchasing HIVST kits online met the needs of certain populations who were reluctant to seek HIV testing services at health facilities. There are 3 types of HIVST kits sold online in China, including fingertip blood test kits, oral mucosal transudate test kits, and urine test kits [10]. The HIVST kits sold online range in price from 20 to 200 RMB (US $2.93-$29.34), with an average price of about 50 RMB (US $7.34) [11,12].

Since the introduction and promotion of HIVST in China, it has been gradually accepted and welcomed by users. It was reported that about 220 HIVST kits were sold per hour in 2017 by an online pharmaceutical store in China [13]. However, the characteristics of online HIVST purchasing behaviors have not been systematically studied. In this study, we analyzed the online HIVST kit sales data from 2015 to 2017 from JD and Taobao, the top 2 e-commerce platforms in China, to reveal the spatiotemporal characteristics of online HIVST kit purchases. Further, we investigated potential factors associated with online purchase of HIVST kits and roughly assessed the contribution of online acquisition of HIVST to HIV diagnosis. Our findings will inform the behavioral profile of those who purchase of HIVST kits online, which can further promote self-testing.

Methods

Data Source

JD and Taobao are two of the largest and most popular e-commerce platforms in China. The online retail data of HIVST kits at city level from 2015 to 2017 were collected from JD and Taobao, the largest retailer of HIVST kits, which sold 70% of HIVST kits. The online retail data we collected for this study had the city location of purchasers but did not include customer shipping addresses. City-level socioeconomic and demographic data were extracted from the Statistical Yearbook of each province or the official website of the local statistics department, including population, gross domestic product (GDP), and gross domestic product per capita (GDP per capita). Geographic data were downloaded from the National Catalogue Service for Geographic Information [14]. The data on newly diagnosed HIV cases at city level were collected from the provincial Centers for Disease Control and Prevention and the data center of China Public Health Science [15]. The data on HIV screening laboratories were obtained from the National HIV/AIDS Laboratory Management Information System of China.

Spatiotemporal Distribution of the Online Purchase of HIVST Kits

The spatial and temporal characteristics of the online purchase of HIVST kits were investigated using the Bayesian spatiotemporal hierarchical model. To eliminate the influence of population size and better reflect the behavioral characteristics of people in the city purchasing HIVST kits online, the purchase rate of HIVST kits per capita was calculated by dividing the amount of HIVST kits sold in a city by the number of people living in that city. Poisson and log link regression functions were used to perform analyses as follows:

In Poisson likelihood function (equation 1), y_{it} is the amount of online HIVST kit sales in place i at time t, n_{it} is the number of total populations in place i at time t, and r_{it} is the rate of online HIVST kit sales per capita in place i at time t.

In the log link regression function (equation 2) that describes the rate of online HIVST kit sales per capita, α is the overall average for the entire study period and S_i is the overall spatial trend. Additionally, t^* is the median time of the study period,

($b_0t^* + v_t$) and $b_{1i}t^*$ specify the overall and local temporal trend at place i, respectively. ε_{it} is the term for random Gaussian noise.

Factors Associated With the Online Purchase Rate of HIVST Kits Per Capita

To investigate factors associated with the behavior of online purchase of HIVST kits, the ordinary least squares regression (OLSR) was performed using data in 2017 to analyze the correlation between the online purchase rate of HIVST kits per capita and 5 potential factors: (1) population density, (2) GDP per capita, (3) road density, (4) HIV/AIDS screening laboratory density, and (5) rate of newly diagnosed HIV/AIDS cases per 100,000 persons. We also used the statistical tool Geodetector [16-18] to quantify the interactive effect of every 2 factors on influencing the online purchase rate of HIVST kits per capita with a *q* statistic value. The *q* statistic value ranged between 0 and 1, and a higher *q* statistic value means stronger influencing power.

Contribution of HIVST Through Online Purchase to HIV Diagnosis

The contribution of HIVST through online purchase to HIV diagnosis was estimated using the city-level data in 2017 with the following formula: [(Amount of HIVST kits purchased online in a city * HIV prevalence) / newly diagnosed HIV cases in a city] * 100%. Since city-level HIV prevalence was not available for this study, we used the national HIV prevalence of 0.09% instead [19]. Cities with more than 300 newly diagnosed HIV/AIDS cases in 2017 were included in the analysis.

Data Management and Statistical Analysis

ArcGIS software (Esri) was used to calculate population density ($1/km^2$), road density (km/km^2), and HIV screening laboratory density ($1/km^2$). The Bayesian hierarchical model was conducted with WinBUGS (University of Cambridge). OLSR was performed using GeoDa version 1.12.1.139 software. All geographic maps were created with ArcGIS 10.2 software.

Ethics Approval

The study was approved by the ethics review committees of the National Center for AIDS/STD Control and Prevention, Chinese Center for Disease Control and Prevention (X131022302), and all procedures were performed in accordance with the relevant guidelines and regulations.

Results

Temporal Trend and Geographic Distribution of Online HIVST Kit Sales

Between 2015 and 2017, a total of 1,482,773 HIVST kits were sold online. Online sales of HIVST kits showed an evident upward temporal trend and undulation over the seasonal variations, with annual peak sales occurring in May and December (Figure 1A). Among all provinces, Guangdong, Sichuan, Beijing, Jiangsu, Shandong, and Zhejiang ranked the top 6 in online purchases of HIVST kits in 2017, with sales of more than 60,000 kits in each of the provinces (Figure 1B).

Geographic distribution analysis showed that, in 2015, only 7 major cities (Beijing, Chengdu, Chongqing, Shanghai, Shenzhen, Guangzhou, and Wuhan) purchased more than 1500 HIVST kits, and in 2017, this number had rapidly increased to 132 cities (Figure 2 A-C). Spatial distribution of online HIVST kit sales was uneven across the country (Figure 2C). If we draw a Heihe-Tengchong Line, also known as the Hu Line since 1934 [20], which marks a striking difference in the distribution of population in China, we can see that most HIVST kits were sold online to cities located southeast of the Heihe-Tengchong Line.

To exclude the impact of population density on the online sales of HIVST kits, we analyzed the geographic distribution of the online purchase rate of HIVST kits per capita (Figure 3 A-C). Urumchi, Lhasa, Hohhot, Sining, and Lanchow (as shown in blue spots) and other cities located northwest of the Heihe-Tengchong Line showed similar online purchase rate of HIVST kits per capita in 2017 compared with cities located southwest of the Heihe-Tengchong Line (Figure 3C). These results suggested that although the online purchase amount of HIVST kits in a city appeared to be associated with the number of people in that city, the willingness to purchase HIVST kits online is uniform across the country and independent of population size.

Figure 1. Online purchase of HIVST kits in China 2015-2017: (A) monthly online purchase amount of HIVST kits and (B) online purchase amount of HIVST kits in provinces. HIVST: HIV self-testing.

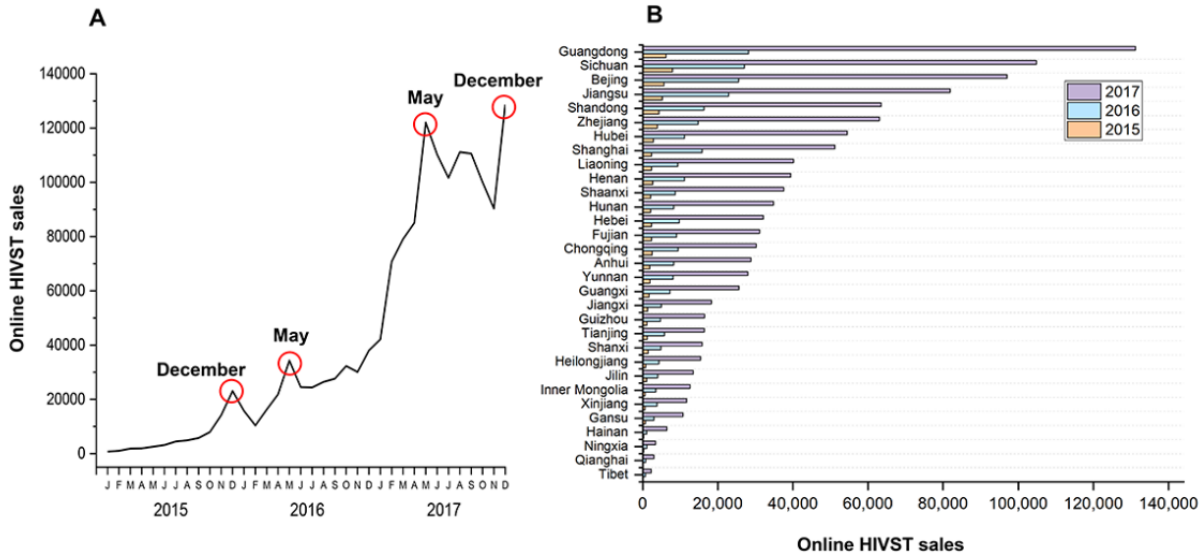


Figure 2. Geographic distribution of online sale amount of HIVST kits in China 2015-2017: (A) 2015 map, (B) 2016 map, and (C) 2017 map. Most HIVST kits were sold online to cities located southeast of the Heihe-Tengchong Line (as shown the red line), which marked a striking difference in the distribution of population in China. Color coded from pale orange to dark orange, and dark orange indicates high level. HIVST: HIV self-testing.

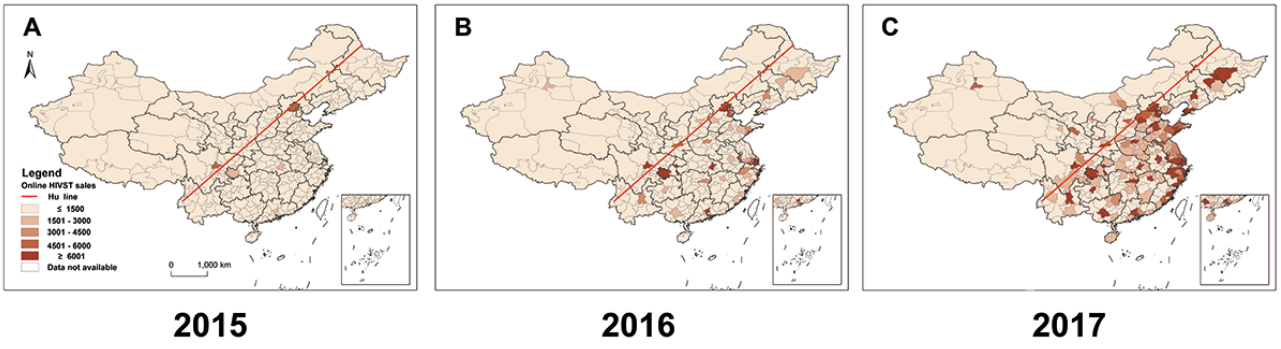
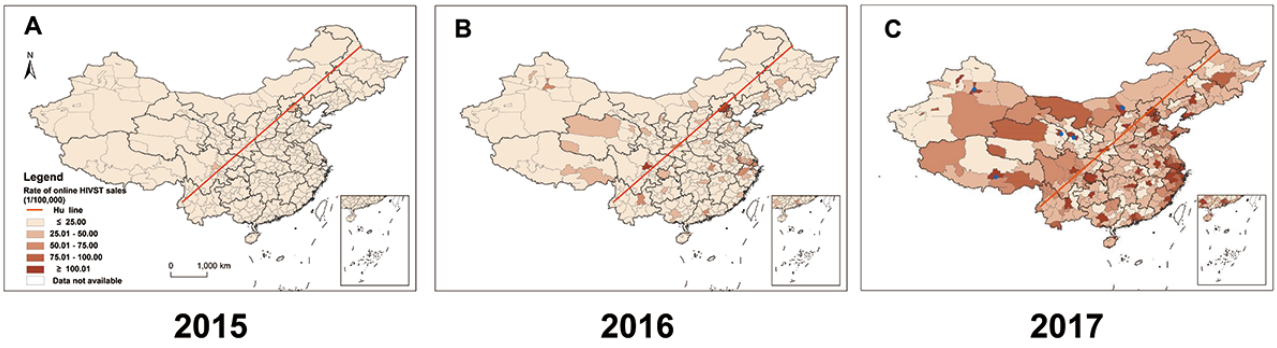


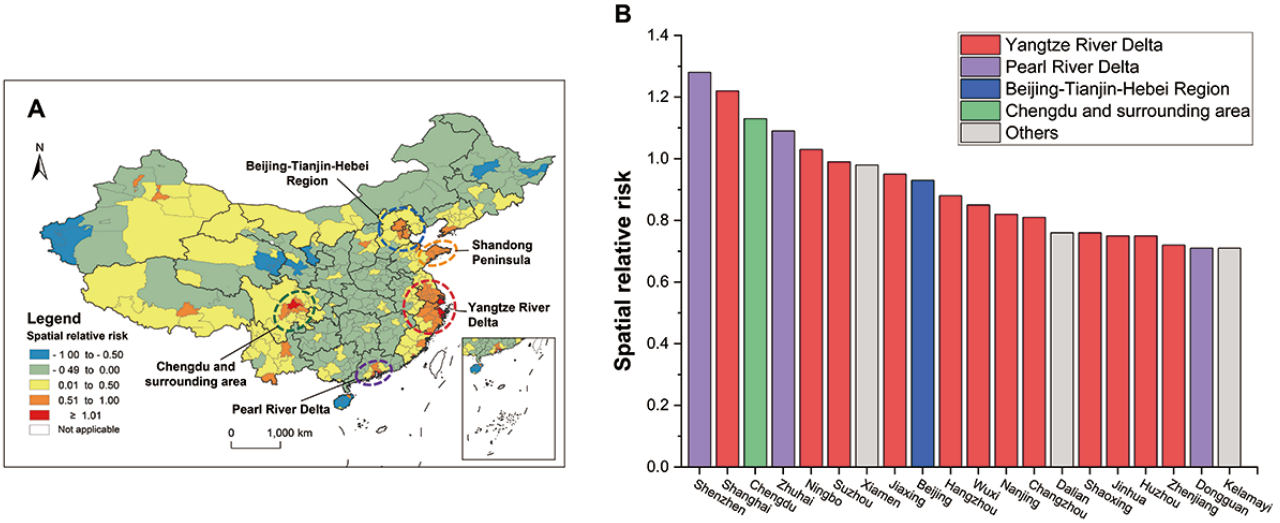
Figure 3. Geographic distribution of the online purchase rate of HIVST kits per capita in China 2015-2017: (A) 2015 map, (B) 2016 map, and (C) 2017 map. To exclude the impact of population density on the online sales of HIVST kits, we analyzed the geographic distribution of the online purchase rate of HIVST kits per capita. Urumchi, Lhasa, Hohhot, Sining and Lanchow (as shown in blue spots) and other cities located northwest of the Heihe-Tengchong Line showed similar online purchase rate of HIVST kits per capita in 2017 compared with cities located southwest of the Heihe-Tengchong Line (as shown the red line). Color coded from pale orange to dark orange, and dark orange indicates high level. HIVST: HIV self-testing.



Spatial Distribution Characteristics of the Online Purchase Rate of HIVST Kits per 100,000 Persons

According to the results of spatial distribution analysis by the Bayesian hierarchical model, the online HIVST sales were unevenly distributed (Figure 4A). There were 5 regions that showed a comparatively higher spatial preference for purchasing HIVST kits online, including the Pearl River Delta, Yangtze River Delta, Chengdu and surrounding areas, Beijing and Tianjin areas, and Shandong Peninsula. These 5 regions also had a higher economic status than other areas in China [21], and the top 20 cities showing a high spatial preference for online purchasing of HIVST kits were mostly from these 5 regions (Figure 4B).

Figure 4. Spatial pattern of online purchase of HIV self-testing (HIVST) kits: (A) spatial pattern of online purchase preference of HIVST kits 2015-2017 and (B) the top 20 cities showed high spatial preference for online purchasing HIVST kits.



Factors Associated With the Online Purchase Rate of HIVST Kits per 100,000 Persons

We further analyzed potential factors associated with the online purchase rate of HIVST kits per 100,000 persons, including population density, GDP per capita, road density, HIV/AIDS screening laboratory density, and the rate of newly diagnosed HIV/AIDS cases per capita. OLSR analysis showed that GDP per capita and the rate of newly diagnosed HIV/AIDS cases per 100,000 persons were positively associated with the online purchase rate of HIVST kits per 100,000 persons (Table 1). Furthermore, according to the results of the Geodetector analysis (Figure 5), GDP per capita and the rate of newly diagnosed HIV/AIDS cases per 100,000 persons exerted the strongest interactive effect on the online purchase rate of HIVST kits per 100,000 persons with a *q* value of 0.66, which means that 66% heterogeneity of the rate of online HIVST kit sales per 100,000 persons can be explained by these 2 factors.

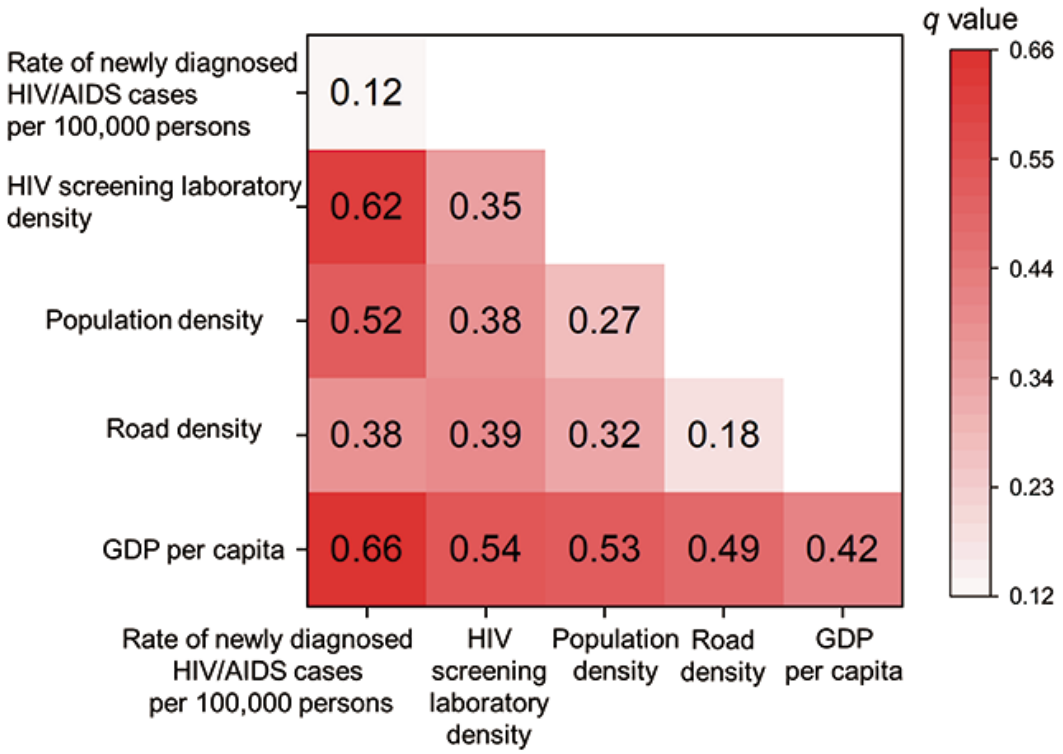
Table 1. Potential factors associated with the online purchase rates of HIV self-testing (HIVST) kits per 100,000 persons using ordinary least squares regression.

	Coefficient	*P* value
GDP[a] per capita	0.001	<.001
Road density	21.65	.60
Population density	–0.004	.79
HIV screening laboratory density	4.26	.07
Rate of Newly diagnosed HIV/AIDS cases per 100,000 persons	0.99	<.001
R^2	0.48	<.001
AIC[b]	3578.22	<.001

[a]GDP: gross domestic product.

[b]AIC: Akaike information criterion.

Figure 5. Interactive effects of potential factors on the online purchase of HIV self-testing (HIVST) kits. The darkness of color represents the value of q. GDP: gross domestic product.



Contribution of Online Purchasing of HIVST Kits to the Detection of HIV Cases

We roughly estimated the contribution of online purchasing of HIVST kits to the detection of HIV-infected individuals in 2017 and found that among the top 20 cities with the highest contribution, the contribution ratio ranged from 1.28% to 3.51% (Table 2). Among the top 20 cities with comparatively higher contributions to the detection of HIV-infected individuals, 12 cities are located in the 4 major economic regions in China as shown in Figure 4A (Pearl River Delta, Yangtze River Delta, Beijing-Tianjin-Hebei Region, and Shandong Peninsula), 7 cities are provincial capitals, and 1 city is a municipality under independent planning status. When further looking at the online HIVST kit purchase amounts and the proportion of newly diagnosed HIV/AIDS cases in these 20 cities, 14 of the 20 cities with the highest contribution to the detection of HIV-infected individuals also entered the top 20 in the proportion of online HIVST kit purchase amounts, although their ranks in the proportion of newly diagnosed HIV/AIDS cases were not high. It is worth noting that among the top 20 cities with comparatively higher contributions to the detection of HIV-infected individuals, 14 cities were also identified in the top 20 cities with a high spatial preference for online purchasing HIVST kits (Figure 4B). These findings suggested that the HIV self-testing through online access contributed to the detection of HIV-infected individuals in a city; however, it was noted that the contribution ratio was not high.

Table 2. The top 20 cities with comparatively higher contributions to detection of HIV-infected individuals.

City	Contribution ratio (%)	Rank	Proportion of online HIVST[a] purchase amount (%)	Rank	Proportion of newly diagnosed HIV/AIDS cases (%)	Rank	Location
Wuhan	3.51	1	3.45	5	0.63	38	Provincial capitals
Shanghai	2.69	2	4.57	4	1.11	11	Yangtze River Delta
Beijing	2.49	3	8.66	1	2.24	4	Beijing-Tianjin-Hebei Region
Guangzhou	2.38	4	4.80	3	1.30	10	Pearl River Delta
Nanjing	2.33	5	1.63	12	0.45	65	Yangtze River Delta
Hefei	2.27	6	0.95	18	0.27	96	Yangtze River Delta
Xi'an	2.24	7	2.53	8	0.73	28	Provincial capitals
Qingdao	1.99	8	0.94	19	0.31	88	Shandong Peninsula
Zhengzhou	1.94	9	1.20	15	0.40	71	Provincial capitals
Suzhou	1.89	10	1.75	11	0.60	42	Yangtze River Delta
Shenyang	1.87	11	1.84	9	0.63	39	Provincial capitals
Tianjin	1.84	12	1.47	14	0.52	60	Beijing-Tianjin-Hebei Region
Jinan	1.71	13	0.72	26	0.27	93	Provincial capitals
Hangzhou	1.56	14	1.83	10	0.76	24	Yangtze River Delta
Xiamen	1.48	15	0.62	32	0.27	95	Others
Changzhou	1.43	16	0.54	36	0.24	113	Yangtze River Delta
Wuxi	1.41	17	0.78	24	0.36	83	Yangtze River Delta
Changsha	1.29	18	1.50	13	0.75	26	Provincial capitals
Shijiazhuang	1.29	19	0.61	34	0.30	90	Beijing-Tianjin-Hebei Region
Taiyuan	1.28	20	0.48	39	0.24	112	Provincial capitals

[a]HIVST: HIV self-testing.

Discussion

Principal Findings

Our temporal analysis found that online HIVST kit sales in China showed an obvious undulation within a year, with significant peaks in May and December. This could be due to some special events occurring during certain time periods. In China, a week-long holiday usually begins on May 1, International Workers' Day, during which individuals are more likely to engage in risky sexual behaviors [22,23], increasing the demand for HIVST kits. We also observed an uptick in online HIVST kit sales around World AIDS Day on December 1 [24], likely due to the high level of advocacy and campaigns to promote HIV testing during this period [25]. It is worth noting that, by the time the State Council of China released the Thirteenth Five-Year Plan (2017-2022) in January 2017 announcing that China would promote HIVST by selling HIVST kits in pharmacies and online, the online HIVST kit sales had tripled, representing a high demand and acceptance of HIV self-testing among people in need. Therefore, it is of great significance to investigate the behaviors of those who purchase HIVST kits online to better promote the HIV self-testing strategy in China.

The spatial Bayesian analyses in this study found that cities located in economically developed regions have a relatively high spatial preference for the online purchase of HIVST kits. Furthermore, regression analysis identified GDP per capita and the rate of newly diagnosed HIV/AIDS cases per 100,000 persons as 2 factors associated with the online purchase rate of HIVST kits per 100,000 persons. Therefore, our findings suggested that when HIVST was launched in China from 2015 to 2017, online purchase of HIVST was more acceptable to high-risk individuals with good financial statuses, which was consistent with our previous survey results on a small group of people who purchased HIVST online [26]. The Geodetector analysis also identified GDP per capita and the rate of newly diagnosed HIV/AIDS cases per 100,000 persons as 2 factors having the strongest interaction with the online purchase rate of HIVST kits per 100,000 persons, which echoes a previous report that income status was positively related to high-risk behaviors among men who have sex with men (MSM) in China [27].

Another interesting finding is that online HIVST sales exhibited a long-tail effect. For the traditional HIV testing services provided by medical institutions through PITC and VCT, 80% of HIV screening tests were reported to have originated from 17.3% of the laboratories [28], which showed a phenomenon of the Pareto principle (80-20 phenomenon) [29]. The Pareto

principle states that approximately 80% of the effects come from 20% of the causes and is commonly used in the field of traditional place-based sales channels. In our study, when looking at the proportion of online HIVST sales at the city-level in 2017, 4 first-tier cities and 15 new first-tier cities with 16.2% of the HIV screening laboratories accounted for only 48.52% of online HIVST sales. This long-tail effect suggested that the testing delivery strategy through online purchase of HIVST kits had a more generalized effect than traditional place-based HIV testing. The long-tail effect has recently been widely used to improve public health [30], library service [31], and industrial parks [32].

In this study, we roughly estimated the contribution of online purchase of HIVST kits to detect HIV/AIDS cases in 2017 and found that the contribution rate was not high from 2015 to 2017. The highest contribution rate was only 3.5% in Wuhan city. We suspected this is because from 2015 to 2017 when HIVST was initially introduced in China and gradually implemented nationwide, self-testing was unfamiliar to people in need who had previously sought HIV testing at local hospitals, maternal centers, and public health laboratories. In addition, when estimating the contribution of online HIVST kits to newly diagnosed HIV/AIDS cases, we used national HIV prevalence (0.09%) in the general population which could underestimate the contribution ratio since HIV risk behaviors are higher among individuals who purchased HIVST kits online. A study conducted in MSM across China shows that internet-based self-test behavior was more common among first-time testers, many of whom reported a higher risk of sexual behaviors [33].

In addition, we found that the rate of newly diagnosed HIV/AIDS cases per 100,000 persons is a factor associated with online purchasing of HIVST kits, raising another important question as to whether the HIV epidemic affects the purchase of HIVST kits or the purchase of HIVST kits can indicate or predict the occurrence of the epidemic. Regardless, more research is needed to clarify these issues.

Limitations

The rapid growth in sales of online HIVST kits from 2015 to 2017 indicates that this new HIV testing service is generally accepted by those in need. However, it is noteworthy that the causes of the increasing trend in online purchase of HIVST kits are heterogeneous. Specifically, due to the lack of data on online shopping activity, our study was unable to assess whether the increased online purchase of HIVST kits in a region is associated with the high online purchase engagement among local residents. To address this concern, future research should analyze in-depth the contribution of web development to the online purchase of HIVST kits. In addition, some information about HIVST kits sold by the 2 e-commerce platforms was missing, such as price, types of products, and retailers. These factors could also potentially impact customer behavior in online purchasing of HIVST kits.

Conclusion

This study found that the online purchasing of HIVST kits has been generally accepted by those in need and were preferred by individuals in regions with a good economy. In addition to economic status, a higher rate of newly diagnosed HIV/AIDS cases per 100,000 persons is also associated with online purchasing of HIVST kits.

Acknowledgments

We gratefully acknowledge and appreciate the contribution of Haolan Sun from the University of Queensland to manuscript polishing and improvement. This study was supported by grant 2018ZX10732101-001-010 from the China National Science and Technology Major Projects in Infectious Disease from Ministry of Science and Technology and grant D171100006717001 from the Science and Technology Program of Beijing, China. The funder had no role in the study design, data collection and analysis, decision to publish, and preparation of the manuscript.

Authors' Contributions

YL was responsible for investigation and methodology. QZ and CX performed data analysis. QZ and CX were responsible for software. CX was responsible for visualization. GZ, MH, and CJ were responsible for conceptualization. YJ was responsible for resources and supervision. YL, QZ, and CJ wrote the original draft. GZ, MH, and CJ edited the manuscript. MH reviewed the manuscript. Authors CJ (jinc@chinaaids.cn) and MH (mjhan@chinaaids.cn) are co-corresponding authors for this article.

Conflicts of Interest

None declared.

References

1. An ambitious treatment target to help end the AIDS epidemic. Geneva: Joint United Nations Programme on HIV and AIDS; 2014. URL: https://www.unaids.org/sites/default/files/media_asset/90-90-90_en.pdf [accessed 2022-08-16]
2. UNITAID. Landscape for HIV rapid diagnostic tests for HIV self-testing. Geneva: World Health Organization; 2015. URL: https://www.aidsdatahub.org/sites/default/files/resource/landscape-hiv-rapid-diagnostic-tests-hiv-self-testing.pdf [accessed 2022-08-16]

3. Guidelines on HIV self-testing and partner notification: supplement to consolidated guidelines on HIV testing services. Geneva: World Health Organization; 2016. URL: https://apps.who.int/iris/bitstream/handle/10665/251655/9789241549868-eng.pdf?sequence=1&isAllowed=y [accessed 2022-03-16]
4. Figueroa C, Johnson C, Verster A, Baggaley R. Attitudes and acceptability on HIV self-testing among key populations: a literature review. AIDS Behav 2015 Nov;19(11):1949-1965 [FREE Full text] [doi: 10.1007/s10461-015-1097-8] [Medline: 26054390]
5. Johnson C, Baggaley R, Forsythe S, van Rooyen H, Ford N, Napierala MS, et al. Realizing the potential for HIV self-testing. AIDS Behav 2014 Jul;18 Suppl 4:S391-S395. [doi: 10.1007/s10461-014-0832-x] [Medline: 24986599]
6. Pant PN, Sharma J, Shivkumar S, Pillay S, Vadnais C, Joseph L, et al. Supervised and unsupervised self-testing for HIV in high- and low-risk populations: a systematic review. PLoS Med 2013;10(4):e1001414 [FREE Full text] [doi: 10.1371/journal.pmed.1001414] [Medline: 23565066]
7. Cambiano V, Ford D, Mabugu T, Napierala Mavedzenge S, Miners A, Mugurungi O, et al. Assessment of the potential impact and cost-effectiveness of self-testing for HIV in low-income countries. J Infect Dis 2015 Aug 15;212(4):570-577 [FREE Full text] [doi: 10.1093/infdis/jiv040] [Medline: 25767214]
8. HIV self-testing research and policy hub. URL: https://hivst.fjelltopp.org/policy [accessed 2022-03-16]
9. Shan D, Xiong R, Shi Y, Li J, Pan L, Xiao D, et al. A survey on use of online sold fingertip blood HIV rapid test kits for self-testing. Dis Surveil 2020;35(9):862-865. [doi: 10.3784/j.issn.1003-9961.2020.09.020]
10. Shan D, Xiong R, Xiao D, Li J, Shi Y, Pan L, et al. A survey on current situation of HIV self-testing kits use. Chin J AIDS STD 2021;27(2):166-169. [doi: 10.13419/j.cnki.aids.2021.02.13]
11. Ma Z, Chen B, Chang H, Pei L, Xing W. Application of self-testing in detection of human immunodeficiency virus antibody. Chin J AIDS STD 2022;23(12):1169-1172. [doi: 10.13419/j.cnki.aids.2017.12.28]
12. Ren Y, Ma Z, Xu B, Wang J, Pei L, Jiang Y, et al. Physical performance and analytical sensitivity of rapid diagnostic test reagents sold online to screen HIV infection. Chin J AIDS STD 2020;26(3):243-246. [doi: 10.13419/j.cnki.aids.2020.03.05]
13. Tang W, Wu D. Opportunities and challenges for HIV self-testing in China. Lancet HIV 2018 Nov;5(11):e611-e612. [doi: 10.1016/S2352-3018(18)30244-3] [Medline: 30213725]
14. National Catalogue Service For Geographic Information. National Geomatics Center of China. URL: https://www.webmap.cn [accessed 2022-03-16]
15. Data center of China Public Health Science. Chinese Center for Disease Control and Prevention. URL: https://www.phsciencedata.cn [accessed 2022-03-17]
16. Wang JF, Li XH, Christakos G, Liao YL, Zhang T, Gu X, et al. Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun region, China. International Journal of Geographical Information Science 2010;24(1):107-127. [doi: 10.1080/13658810802443457]
17. Wang JF, Zhang TL, Fu BJ. A measure of spatial stratified heterogeneity. Ecological Indicators 2016;67:250-256. [doi: 10.1016/j.ecolind.2016.02.052]
18. Wang J, Chengdong XU. Geodetector:Principle and prospective. Acta Geographica Sinica 2017;72(1):116-134. [doi: 10.11821/dlxb201701010]
19. The prevalence of AIDS in China is about nine out of ten thousand. National Health Commission of the People's Republic of China. URL: http://www.gov.cn/xinwen/2018-11/23/content_5342852.htm [accessed 2022-08-16]
20. She W. Hu Huanyong: father of China's population geography. China Popul Today 1998 Aug;15(4):20. [Medline: 12294257]
21. Report on the integration of urban agglomerations in China. Beijing: China Development Research Foundation URL: https://cdrf.org.cn/jjh/pdf/zhongguochengshiqunfazhanbaogao.pdf [accessed 2022-04-26]
22. Calafat A, Hughes K, Blay N, Bellis MA, Mendes F, Juan M, et al. Sexual harassment among young tourists visiting Mediterranean resorts. Arch Sex Behav 2013 May;42(4):603-613. [doi: 10.1007/s10508-012-9979-6] [Medline: 22733155]
23. Downing J, Hughes K, Bellis MA, Calafat A, Juan M, Blay N. Factors associated with risky sexual behaviour: a comparison of British, Spanish and German holidaymakers to the Balearics. Eur J Public Health 2011 Jun;21(3):275-281. [doi: 10.1093/eurpub/ckq021] [Medline: 20231212]
24. World AIDS Day. Wikipedia. URL: https://en.volupedia.org/wiki/World_AIDS_Day [accessed 2022-08-16]
25. Campaigns to promote HIV testing around World AIDS Day on December 1. URL: http://news.china.com.cn/node_7253302.htm [accessed 2022-08-16]
26. Lv Y, He X, Ma J, Yao J, Xing W, Liu P, et al. Characteristics of population who purchase HIV rapid test kit on internet for self-testing. Chin J AIDS STD 2019;25(6):559-561. [doi: 10.13419/j.cnki.aids.2019.06.04]
27. Shi T, Zhang B, Li X, Xu J, Wang N, Zhou S, et al. [Study on the comparison of high risk behaviors related to AIDS among different status of income in men who have had sex with men]. Chin J Epidemiol 2008 May;29(5):426-429. [Medline: 18956671]
28. Zhang J, Ding X, Zhou X, Chen W, Yao J, Guo Z, et al. Performance of HIV detection in Zhejiang province in China: the Pareto principle at work. J Clin Lab Anal 2021 Jun;35(6):e23794 [FREE Full text] [doi: 10.1002/jcla.23794] [Medline: 33942384]
29. Erridge P. The Pareto principle. Br Dent J 2006 Oct 07;201(7):419. [doi: 10.1038/sj.bdj.4814131] [Medline: 17031332]

30. Kreuter MW, Hovmand P, Pfeiffer DJ, Fairchild M, Rath S, Golla B, et al. The "long tail" and public health: new thinking for addressing health disparities. Am J Public Health 2014 Dec;104(12):2271-2278. [doi: 10.2105/AJPH.2014.302039] [Medline: 25322308]
31. Jia Y, Yao J, Li Y. Research on library long-tail server based on K-means algorithms. AIP Conf Proc 2019:1. [doi: 10.1063/1.5090733]
32. Hou D, Li G, Chen D, Zhu B, Hu S. Evaluation and analysis on the green development of China's industrial parks using the long-tail effect model. J Environ Manage 2019 Oct 15;248:109288. [doi: 10.1016/j.jenvman.2019.109288] [Medline: 31382195]
33. Jin X, Xu J, Smith MK, Xiao D, Rapheal ER, Xiu X, et al. An internet-based self-testing model (Easy Test): cross-sectional survey targeting men who have sex with men who never tested for HIV in 14 provinces of China. J Med Internet Res 2019 May 15;21(5):e11854 [FREE Full text] [doi: 10.2196/11854] [Medline: 31094339]

Abbreviations

GDP: gross domestic product
HIVST: HIV self-testing
HTS: HIV testing service
MSM: men who have sex with men
OLSR: ordinary least squares regression
PITC: provider-initiated HIV testing and counseling
RDT: rapid diagnostic test
UNAIDS: Joint United Nations Program on HIV/AIDS
VCT: voluntary counseling and testing

Edited by Y Khader; submitted 27.04.22; peer-reviewed by B Chakalov, ER Khalilian, J Xu; comments to author 22.06.22; revised version received 13.07.22; accepted 21.07.22; published 27.09.22.

Please cite as:
Lv Y, Zhu Q, Xu C, Zhang G, Jiang Y, Han M, Jin C
Spatiotemporal Analysis of Online Purchase of HIV Self-testing Kits in China, 2015-2017: Longitudinal Observational Study
JMIR Public Health Surveill 2022;8(9):e37922
URL: https://publichealth.jmir.org/2022/9/e37922
doi:10.2196/37922
PMID:35918844

Original Paper

Privacy of Study Participants in Open-access Health and Demographic Surveillance System Data: Requirements Analysis for Data Anonymization

Matthias Templ[1*], PhD; Chifundo Kanjala[2*], PhD; Inken Siems[3*], MA

[1]Institute of Data Analysis and Process Design, Zurich University of Applied Sciences, Winterthur, Switzerland

[2]Department of Population Health, London School of Hygiene and Tropical Medicine, Lilongwe, Malawi

[3]Economics and Social Statistics, University of Trier, Trier, Germany

[*]all authors contributed equally

Corresponding Author:
Matthias Templ, PhD
Institute of Data Analysis and Process Design
Zurich University of Applied Sciences
Rosenstrasse 3
Winterthur, 8404
Switzerland
Phone: 41 793221578
Email: matthias.templ@zhaw.ch

Abstract

Background: Data anonymization and sharing have become popular topics for individuals, organizations, and countries worldwide. Open-access sharing of anonymized data containing sensitive information about individuals makes the most sense whenever the utility of the data can be preserved and the risk of disclosure can be kept below acceptable levels. In this case, researchers can use the data without access restrictions and limitations.

Objective: This study aimed to highlight the requirements and possible solutions for sharing health surveillance event history data. The challenges lie in the anonymization of multiple event dates and time-varying variables.

Methods: A sequential approach that adds noise to event dates is proposed. This approach maintains the event order and preserves the average time between events. In addition, a nosy neighbor distance-based matching approach to estimate the risk is proposed. Regarding the key variables that change over time, such as educational level or occupation, we make 2 proposals: one based on limiting the intermediate statuses of the individual and the other to achieve k-anonymity in subsets of the data. The proposed approaches were applied to the Karonga health and demographic surveillance system (HDSS) core residency data set, which contains longitudinal data from 1995 to the end of 2016 and includes 280,381 events with time-varying socioeconomic variables and demographic information.

Results: An anonymized version of the event history data, including longitudinal information on individuals over time, with high data utility, was created.

Conclusions: The proposed anonymization of event history data comprising static and time-varying variables applied to HDSS data led to acceptable disclosure risk, preserved utility, and being sharable as public use data. It was found that high utility was achieved, even with the highest level of noise added to the core event dates. The details are important to ensure consistency or credibility. Importantly, the sequential noise addition approach presented in this study does not only maintain the event order recorded in the original data but also maintains the time between events. We proposed an approach that preserves the data utility well but limits the number of response categories for the time-varying variables. Furthermore, using distance-based neighborhood matching, we simulated an attack under a nosy neighbor situation and by using a worst-case scenario where attackers have full information on the original data. We showed that the disclosure risk is very low, even when assuming that the attacker's database and information are optimal. The HDSS and medical science research communities in low- and middle-income country settings will be the primary beneficiaries of the results and methods presented in this paper; however, the results will be useful for anyone working on anonymizing longitudinal event history data with time-varying variables for the purposes of sharing.

(JMIR Public Health Surveill 2022;8(9):e34472) doi:10.2196/34472

KEYWORDS

longitudinal data and event history data; low- and middle-income countries; LMIC; anonymization; health and demographic surveillance system

Introduction

Background

Although health research data sharing has many benefits and great value [1,2], one of the main concerns is maintaining the privacy of study participants. The rationale for both data sharing and privacy is widely recognized. In the field of medical science research, the issue of privacy is central to good ethical practice. Anonymization of data provides an opportunity to mitigate this tension between sharing data and preserving the privacy of those whose data are shared. However, it is often unclear how data can be shared without unduly compromising the privacy of the individuals included in a data set.

A fundamental issue with personal data disclosure is whether an attacker can learn anything about an individual if the data or analysis results are provided or predictions are made. On the one hand, one can ask whether an attacker can successfully match individuals with the data at their disposal. In addition, are attackers' efforts (and related costs) higher than the benefits of disclosing information? On the other hand, the needs of the users of data are of high utility, allowing for high-quality analysis. Data providers are interested in providing such information without disclosing the identities of the individuals in the data.

Similar to all other areas of health research, longitudinal population studies in low- and middle-income countries (LMIC), such as health and demographic surveillance system (HDSS) [3], face the challenge of finding the right balance between data sharing and privacy protection.

The HDSS must take a position that allows the sharing required by research funders and journal publishers [2,4] while minimizing the risk of compromising the privacy of individuals who make their data available for research.

However, the important issue of health data privacy has not been adequately explored in LMIC in general and HDSSs in particular. HDSSs currently share data in most cases without anonymizing them beyond masking direct identifiers [5]. There is a possibility that attackers may use indirect identifiers such as education level, sex, and age—in cases where these are shared [6]—to identify participants and, consequently, their health status, which they did not intend to share beyond the boundaries of the research in which they participated. The extent of such risks has not been fully explored in the HDSS data sets, and consequently, no measures have been taken to mitigate these risks; that is, to the best of our knowledge, this has not been addressed in the literature on health, statistics, and privacy.

Note that for some selected data sets and general anonymization problems, the World Bank Group, PARIS21 and Organization for Economic Cooperation and Development, and the International Household Survey Network supported the development of the anonymization software sdcMicro [7], and they all recommend it [8]. sdcMicro is actively used in many organizations, ranging from statistical offices [9] and social and political science [10] to the United Nations High Commissioner for Refugees [11] and health [12-14]. However, there is a need to justify the use of this software for the specific needs arising from longitudinal population health data in LMIC.

Longitudinal data include records of different attributes of the same participants observed and measured at multiple points in time. Existing theories and software are suitable only for anonymizing and assessing the disclosure risk of cross-sectional data. An extension of this theory is needed to quantify and control the disclosure risk for longitudinal data.

Karonga HDSS

An HDSS is a combination of field and computing procedures for collecting demographic, health risk, and exposure and outcome data from a defined population within a defined geographical area on a longitudinal basis [3,15]. HDSSs are set up to monitor open or dynamic population cohorts, building longitudinal databases of this population over time [15]. A substantial body of literature has considered various HDSS aspects, including the rationale for their establishment in LMIC [3,15], the definition of core HDSS concepts and processes [5,16], and the reference data model [17] among many others. The data set used for illustration is from an HDSS in Malawi, the Karonga HDSS. This HDSS has been described in detail elsewhere [18]. Briefly, its surveillance site is in northern rural Malawi and has been in operation from its initial census in 2002 to 2004. The Karonga HDSS contains longitudinally linked health data from the study population.

The Karonga HDSS is part of a collaborative research program under the Malawi Epidemiology and Intervention Research Unit [19].

HDSS Core Residency Data

The generic data set structure on which we based this data anonymization requirements analysis is in the core residency data format. This standard data set is widely used in HDSS for data sharing and analysis [19]. An extended version of this data set is comprehensive enough to cover the considerations that need to be made in anonymizing HDSS event history data. This data set essentially comprises the core HDSS events for each individual under surveillance and attributes relating to the individual and to the core events. The events occur in a particular order that defines entry or exit from the study population. The first event for any individual is one of the following: a baseline census enumeration, a birth, or an in-migration. The last event is one of the following: an out-migration, a death, or the end of observation (censoring). The intervening events observed for any individual need to be logical; for example, an individual born within the surveillance area cannot have in-migration as the next event. The core events change the residency status of an individual and, thus, the name of the data set, core residency data [20].

The basic form of the core residency data includes the following variables: an individual identifier, date of birth, sex, core event, and event date. This form contains all the data on the numerators and person-years of surveillance (exposure) required to calculate the demographic rates for the HDSS population and perform event history analyses.

This basic form can be extended to capture other observations made within the HDSS population. These may include disaggregation of the migration events by distinguishing between migration within the surveillance area (internal) and migration to or from outside the area (external), as well as the inclusion of attributes that change over time, such as education level, occupation, and specific disease status (eg, HIV and tuberculosis).

To elaborate on the anonymization requirements, we distinguish between three variable groupings that can go into these HDSS core residency data:

1. Static variables: These are variables in which the observations on individuals do not change over time, such as sex and date of birth.
2. Status (time-varying) variables: These are variables in which the observations on individuals change over time, such as occupation or education level.
3. Core events variables: These are the variables in which the observations are specific to the event. The observed event and the event date fall into this category.

Our approach investigates the requirements for anonymizing variables falling into these 3 groups.

Karonga Residency Data

The variables in this data set largely overlap with those found in the publicly available Karonga HDSS core residency data set on the iSHARE data repository [21]. The extended version used in this study has status variables on occupation and education level, in addition to those found in the Karonga core residency file.

This data set contains information recorded from October 1995 to the end of 2016, comprising 14 variables, 280,381 rows (events), and 72,935 individuals ever observed since the HDSS's inception.

The main variables of the data set for this work are as follows:

- Static variables: sex
- Status variables: occupation with categories not working, student, unskilled manual, farmer, fisherman, skilled manual, nonmanual, small trader or business, unskilled manual, skilled manual, nonmanual, and professional; and education with categories none, 1 to 3 years primary, 4 to 7 years primary, primary completed, Junior Certificate of Education completed, Malawi School Certificate of Education completed, and tertiary
- Core event variables: event code with dates on the baseline, date of birth, in-migration, out-migration, and date of death
- Household ID, mother's ID, father's ID, and polygamy ID

Objective

To contribute toward filling this gap, we propose a set of requirements for anonymizing the HDSS longitudinal data. Our proposal customizes and applies traditional methods that work on the premise of keeping the data quality as high as possible while slightly altering the data until the disclosure risk is below a fixed threshold. The main contributions of this study are as follows:

- We define anonymization requirements peculiar to longitudinal event history data.
- We propose steps to take to meet these requirements, including assessing and controlling for disclosure risk for the static and time-varying variables and core event dates.
- We implement the proposed steps and show the results.
- We place our proposal within the larger context of data anonymization approaches, outlining how our method of choice contrasts with the alternatives within the LMIC HDSS context.

Methods

In this section, we outline the methods and procedures for anonymizing HDSS core residency data.

Different Concepts for Different Needs

Our approach of keeping data quality as high as possible by modifying data slightly until the disclosure risk is below a certain threshold does not stand alone but rather is part of a broader ecosystem of data anonymization methods. We briefly review this ecosystem and emphasize that the choice of anonymization approaches depends heavily on the needs of the user group and the cost of implementing the solution. We briefly outline 4 important anonymization concepts before discussing their applicability for sharing HDSS data. They are listed in ascending order of data analysis potential as follows: privacy-preserving computation, synthetic data, secure laboratories, and the approach used in this study (anonymized individual-level data using methods of statistical disclosure control [SDC]). With privacy-preserving computation, data remain on the data owner's side. This can be extended to a secure multiparty computation with multiple clients (data holders). Two popular privacy-preserving computation methods are differential privacy [22] and federated learning with Private Aggregation of Teacher Ensembles [23]. However, there are several limitations, as highlighted in the studies by Domingo-Ferrer et al [24], Francis et al [25], and Bambauer et al [26]. Furthermore, the user must trust the predictions without evaluating the model and the data behind the model. Another way of providing anonymized data is by generating synthetic data that exhibit the same characteristics as the original data [27], usually using machine learning and statistical modeling methods. Synthetic data typically have very low disclosure but have also relatively low data utility when the original data possess complex structures [6]. Synthetic data can also be used in remote execution environments, whereby registered researchers work on the synthetic data to develop an analysis code, and the staff of the data holder finally runs the code on the original data. The final analysis output is checked for privacy

by laboratory staff as this checking can hardly be fully automated [28-30].

Difficulties in Using Alternative Concepts

For HDSS data, using privacy-preserving computation would mean first setting up a framework to compute privacy and, for known users (test data), providing a predictive value for a meaningful piece of information (eg, the date of migration or the health status of a person) based on a machine learning prediction approach. It is evident that these approaches have some difficulties in providing good predictions for complex longitudinal data sets. Privacy-preserving computational approaches are also not sustainable options for health and survival data for LMIC because of the high cost and the users' need for detailed data, instead of simply receiving predictions for sensitive information or working with aggregated data. Synthetic, close-to-reality data have the potential of being a viable approach; however, the complexity of longitudinal event history data from HDSS makes it difficult to model and represent all relationships and logical conditions adequately. Remote access to secure laboratories offers the advantage of working on real data but can only provide access to a small number of trusted researchers and requires permanent staff to perform output checks to keep the software on the servers up to date and the server and access secure.

Methods for SDC

For these reasons, methods of SDC are the most suitable. The core concept of SDC comprises transforming data in such a way as to reduce the reidentification risks of the persons represented in the data. More precisely, the aim of SDC is to reduce the risk to a level below a predefined threshold on the one hand and to maintain the data quality and analysis potential and research questions on the other. This is a complex task that requires the application and development of complex methods and, in our particular case, the understanding of specific health population data sets.

Data Release Types: Public Use Versus Scientific Use Files

In line with lowering the barriers to data access, as encouraged by funders [2], and in the interest of implementing sustainable data sharing models, open data through the sharing of the so-called public use files [31] would be a typical mechanism for sharing HDSS data. Public use files require that a potential user agrees to the terms of use and then get access to the data without seeking approval from the data custodians. A reason for this is the resource-efficient publication and distribution of data. Once distributed, there is no need for further labor-intensive steps, as is the case with remote execution and remote access solutions. The next level up would be the scientific use files [31]. This requires a potential user to go through a review process by a data access team to confirm that they are a bona fide researcher from a reputable institution. This sharing demands that the custodians set aside staff time to review data access applications, prepare the data for sharing, customize the shared data to suit the request, and communicate and supervise the researchers. These demands of staff time are suboptimal as they will take staff away from their daily work and are rarely sufficiently funded in LMIC medical science research projects.

Pseudoanonymization

In pseudoanonymization, a string—the exact name of a person or any other direct identification feature (eg, social security number)—is replaced by a pseudonym, usually a 256-bit hash code produced by a cryptography hash function from a salted string [32,33]. The pseudoanonymization of the HDSS core residency data on the iSHARE data repository is performed in a simplified manner. An ascending ID is assigned per person instead of listing their names or identifiers used in the dynamic HDSS databases. Note that as more data with complex interrelationships are shared through platforms such as the Implementation Network for Sharing Population Information from Research Entities (INSPIRE) data, more elaborate pseudoanonymization will become necessary. However, pseudoanonymization does not solve the data protection problem as it only prevents attacks on direct identifiers.

Identifying Key Variables—the Disclosure Scenario

The key question here is what information does an attacker have access to that they could match with the data to be released to identify individuals? Before the key variables (also often called quasi-identifiers) are identified, a check is made to see what other existing data a potential attacker could access and use to link to the current data and identify individuals. This is called the (archive) disclosure scenario [34]. Existing data may include census, voters' roll, population surveys, or administrative data held by government departments and national statistical offices. In most LMIC, not many data sets are available for broad access, and hence, this should not be a major problem.

The biggest challenge may be that an attacker has additional knowledge of some information pertaining to an individual in the data being released. This is often called the nosy neighbor scenario in the literature [34]. An attacker can potentially use this information to identify individuals.

In general, defining these scenarios requires input from subject matter experts who work with the data being released and who are also aware of other common data.

Anonymization Methods for Static and Status Variables

Traditional anonymization of population data uses the concept of uniqueness. By combining several variables (quasi-identifiers from the *Identifying Key Variables—the Disclosure Scenario* section), an individual can be uniquely identified in the data. A key is unique if its frequency is 1, and thus, only one person has the combination of characteristics defined by the key. For example, the key postcode *8404*, citizenship *Austria*, sex *male*, and age *45* are unique in a demographic population data set of Switzerland. A commonly used concept for measuring uniqueness and "almost uniques" is k-anonymity. A data set is k-anonymous if each key (ie, combination of key variables) belongs to at least *k* observations. An approach that also evaluates subsets of key variables is called the special uniques detection algorithm [35,36]. This approach allows for a more

detailed analysis and evaluation of uniques in subsets of key variables.

To achieve k-anonymity and low special uniques detection algorithm scores, the first step typically involves use case–specific recoding of the categorical key variables into broader categories [6]. With recoding, the risk can be significantly reduced. If some individuals still have an increased risk and further recoding would lead to an excessive loss of quality of data, local suppression is typically considered next [6]. This suppresses certain values to guarantee, for example, k-anonymity. The aim is to find specific patterns in categorical key variables and replace these patterns with missing values. (heuristic) optimization methods must be applied to find a minimal suppression pattern [7].

If the number of categorical key variables is large or many of these variables have many categories, the number of keys in a data set is large, and many keys will be unique. In this case, recoding and local suppression would significantly change the data to achieve, for example, k-anonymity. Applying the postrandomization method (PRAM) [37] to a subset of key variables would be a good alternative to recoding and suppressing all key variables. In the PRAM, values are exchanged between the categories of a variable with certain transition probabilities. An attacker can never be sure whether a value is true or has been swapped.

Handling Static and Status Variables With Varying Status of a Person Over Time

Cross-sectional data sets typically contain observations for a single time point, and the application of anonymization methods is generally straightforward (eg, using the guidelines presented by Templ et al [6]).

In the following paragraphs, the extension to longitudinal information, in particular to status variables (eg, *occupation* or *education*), for which the observed values (can) change over time, is discussed. Table 1 shows the problem of using a toy data set with 2 individuals in a simplified manner. It can be easily seen that for person 1, both educational level and occupational have improved over time. When only the baseline status in 2010 is considered, both individuals share the same level of education and occupation category; thus, they are not unique in the data set. If only 2015 were considered, the 2 individuals would not be unique. If only the latest status of a person is considered, both individuals would be unique in this toy data set, considering the key variables of occupation and education level. Moreover, if each status is reported each year, the 2 individuals would also be unique.

A number of alternative representations could be used to anonymize the status variables, each of which has its own advantages and disadvantages.

If only the initial status of a person is reported, the variable would no longer be considered a status variable that changes over time, which simplifies anonymization. The disadvantage is that we can no longer see the progress, for example, in the person's occupational and educational level over time.

If only the first and last statuses of a person in a record are reported, all events in between must either be deleted or replaced by the first stage or the last status.

Another very strict alternative would be to delete the link of a person from one year to the other; that is, for each person, another ID is provided from one year to another. However, this makes a longitudinal analysis difficult; thus, the data utility would suffer significantly.

Postrandomization could be an option, although the order and consistency of educational and occupational levels are either lost or biased to higher levels. For example, it makes no sense to lower a person's education level over time; therefore, with realistic swapping probabilities in the PRAM, the education level would randomly increase but never decrease.

Another approach would be to apply traditional anonymization methods to patterns or subsets of the data, whereby individuals with the same pattern of event occurrence are considered as a subset to be anonymized. For example, the 2 individuals in Table 1 do not have the same pattern as they have a different number of events. This approach leads to a potentially large oversuppression but reduces the disclosure risk heavily. Studies aimed at analyzing the education and occupation of individuals over time might be possible, especially when data analysts impute the suppressed information.

Before deciding on one of these or even other alternative approaches, one has to think about the disclosure scenario. How likely is it that an attacker can merge their database with the anonymized data set provided to match and identify individuals? How likely is a nosy neighbor scenario and to what extent?

For an archive scenario, the following assumptions regarding the attacker's knowledge are made:

- Only the last status of education of a person is known to the attacker, assuming that the attacker's database is more or less an up-to-date archive containing the current educational level of a person used for matching. Here, it is neglected that the attacker has access to the historical sociodemographic status data of individuals.
- Only the last occupational status is known by an attacker, provided that the attacker's database is more or less an up-to-date archive containing the current profession of a person used for matching.
- The attacker has knowledge of the static variables of sex and birth date.
- The attacker does not know the reason for in- and out-migration but knows the birth date, the start date, and the stop date.

For a nosy neighbor scenario, the following assumptions about the attacker's knowledge are made:

- The (changing status) of the education of a person is known to the attacker over time, assuming that the attacker has individual knowledge of the historical development of the educational and occupational levels of a few individuals.
- The attacker has knowledge of the static variables of sex and birth date.

- The attacker may know the reason for in- and out-migration for certain individuals and the corresponding event time, and they may have knowledge about the birth date of certain individuals.

As the data go public as an open-access data set, a nosy neighbor scenario is possible and, thus, in focus. Therefore, we use the approach in which only the first and last observed statuses of a person are reported. This is a solution in which the change in a person's status is reported without their intermediate improvements, whereas local suppression results in a low number of suppressions as not all stages are reported.

Table 1. Toy data set supporting a simple explanation to the problem to deal with time-varying information on status variables.

Person ID	(Event) year	Occupation	Education level
1	2010	2	2
1	2011	2	2
1	2012	3	2
1	2013	3	2
1	2014	3	2
1	2015	3	3
1	2016	4	3
2	2010	2	2
2	2015	3	3
2	2016	3	3

Handling Event History Dates

General Considerations

To prevent (exact) record linkage and closest distance–based neighborhood matching, we suggest adding random noise to the event dates. An adequate obvious choice is to add approximately 100 days randomly. This prevents an attacker from successfully applying record linkage and is likely to prevent distance-based matching.

However, care must be taken to ensure that the order of events is maintained. For example, if a person has a birth date of May 15, 2009, and we hypothetically assume that this person out-migrated on June 5, 2009, in-migrated on July 6, and died on August 1, 2009, then a random noise of +40 or –40 to +60 or –60 days will completely upset the event order.

Thus, we need to modify the event data by adding or subtracting a sufficient number of days so that the individual cannot be identified, although the data utility and event order of the data are retained. More specifically, the addition of noise must be performed with the following constraints: (1) the order of events must be maintained; (2) the time span between events should remain the same as much as possible, naturally fulfilled by adding noise; (3) attacks with record linkage should not be successful; and (4) the number of events per person should remain unchanged.

This leads to a sequential approach that adds noise for each person, event by event, under certain restrictions, explained in more detail in the following paragraphs. Of course, the main parameter—the level of noise—must be determined on a use case and data set–specific basis.

Add Noise to One Event Date

For simplicity, equation 1 shows the case for 3 events, whereby noise is added for 1 person for event 2. Figure 1 shows this case with 3 event dates t_1, t_2, and t_3, and the time span between events 1 and 2 ($\Delta_{2;1}$) and events 2 and 3 ($\Delta_{3;2}$).

It should be noted that extension to any number of events per person is possible and straightforward to implement, although the notation becomes more complicated.

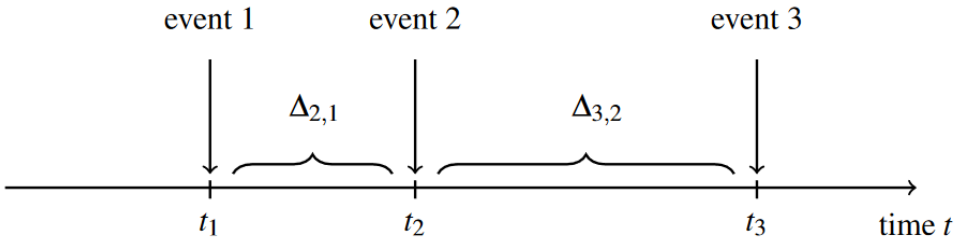
With s, a Bernoulli random values$\in \{-1, 1\}$ with P=.50 for random addition or subtraction of the event date, and $u \sim U[_{min}; _{max}]$, which controls the number of noise (in days), a new (anonymized) event date t_2^* is calculated using the following:

$$t_2^* = t_2 + u \cdot s \text{ , if } \Delta_{2,1} >_{max} \wedge \Delta_{3,2} >_{max} t_2^* = t_2 + u \text{ , if } \Delta_{2,1} \leq_{max} \wedge \Delta_{3,2} >_{max} t_2^* = t_2 - u \text{ , if } \Delta_{2,1} >_{max} \wedge \Delta_{3,2} \leq_{max} t_2^* = t_2 - u - (\Delta_{2,1} - 1) \text{ , if } \Delta_{2,1} >_{max} \wedge \Delta_{3,2} \leq_{max} \wedge \min(\Delta_{2,1}, \Delta_{3,2}) = \Delta_{2,1}$$

This ensures that the event order is preserved for t_1, t_2, and t_3. Except for the first case, restrictions were applied as the distance between event data was smaller than the specified minimum noise range.

An alternative noise addition method is to draw $u \sim N(\mu, \sigma^2)$ and round it to the next integer value.

Figure 1. Schematic overview of 3 event history dates for one person and corresponding time span between the events.



Add Noise Sequentially Event by Event

The extension of equation 1 to all events of a person is achieved by the sequential application of noise to each event of a person. First, all recorded data of one person are stored, and the number of events of this individual, as well as the distance between all events, are recorded. For the first event, date t_1 noise is either randomly subtracted or added; more precisely, it is subtracted without any restrictions and added less than the distance to the second event. Subsequently, for all other events recorded, in an additional loop considering one event date at the time, noise is added, as described above (equation 1) according to a predefined noise level (see *Disclosure Risk* and *Data Utility* section for further discussion on the level of noise). Therefore, first, for t_1, noise is added leading to t_1^*, and then, noise is added to t_2, considering possible restrictions from t_3 and t_1^* to not change the event order. Subsequently, noise is added to t_3 considering t_2^* and t_4,..., until the last event date. Using this sequential approach, preservation of the event order is guaranteed.

Restrictions may occur if 3 consecutive events are very close to each other. If the maximal noise of the respective noise level is larger than the difference between t_2^* and t_3 and t_3 and t_4, it proceeds as follows. If the minimum of the event difference $\min(\Delta_{2;1};\Delta_{3;2})$ is larger than the predefined minimum noise, then take minimum=minimum noise and maximum noise=$\Delta_{2;1}$ and $\Delta_{3;2}$, respectively, and sample at random. If the minimum of event difference $\min(\Delta_{2;1};\Delta_{3;2})$ is smaller than the minimum noise, then sample from a univariate distribution $U(0; \Delta_{2;1})$; same with $\Delta_{3;2}$ in the respective sampling direction as maximum or minimum noise. In the case of normal distribution while (noise $< \Delta_{2;1} \wedge$ noise $> \Delta_{3;2}$), draw a new value from $N(\mu=0; \sigma=50)$ until a valid noise is obtained.

Furthermore, we would like to briefly point out that it is necessary to consider the special data structure. It has already been mentioned that the event history dates cannot ideally be represented in columns, as there are different numbers of events and different events per person. Therefore, a separate row for each event in the data set is used to store the event code and date for a person; that is, individuals are represented in multiple rows. If a person was born within the observation period, he or she has an additional entry as an event in addition to the actual date of birth. Thus, if no birth date is registered under event dates, as the individual was born before data collection, then only one number is randomly added to the date of birth of a person in all rows of this person. If birth is also represented as event date information, the same noise (used to noise the event date on birth) has to be taken as for the column holding the birth date of the person; that is, the information on birth date and the event birth date is linked and must be considered adequately and consistently.

In the *Results* section, the noise level chosen for the HDSS core data set is presented, and further insights into the choice of noise level are provided.

Putting It All Together

The event data are particularly important as they are numerical information that can be used for record linkage if the attacker has a database of exact event data. However, an attacker might only know the year of birth and death and then use this information for matching. In addition to the event history dates, variables with varying statuses over time must also be considered. Therefore, the changes in education and occupational levels are limited by indicating only the first and last status (Textbox 1).

For certain studies, for example, on fertility by educational level, the full history of event dates and changes in the educational level is needed. This is also true for various studies on the occupational level of individuals over time (eg, answering the question of whether well-educated individuals change their occupational levels quicker). In this case, the entire history of event data might be needed, and the previous procedure has to be adapted, in this case, for example, by anonymizing the patterns, as outlined previously.

Textbox 1. Steps of putting it together.

Step 1

- Add random noise to event dates for each person sequentially, as described in the *Handling Event History Dates* section. This prevents record linkage and nearest-neighbor matching with an external database containing exact event dates and preserves the order of events.

Step 2

- Aggregate data (ie, from long to wide representation, where each line represents a person) so that each row contains the information of a person for the static variables (such as sex and birth date), first and latest education, and first and latest occupation and build new variables containing the year of birth, year of death, and number of events of a person.

Step 3

- Perform k-anonymity using local suppression using the implemented methods in sdcMicro [7] using the variables mentioned in step 2 to avoid uniques and prevent successful matching. If the year of the earliest or latest event or the year of birth is suppressed, the noised year and noised event date should also be suppressed. It should be noted that this was hardly the case as the importance was set such that the year of birth, year of death, and number of events of a person are the most important variables; thus, the suppression algorithm uses the remaining variables to make local suppressions.

Step 4

- Disaggregate the anonymized aggregated data (from wide to long representation, where each line represents an event). The data set now includes only the anonymized information on sex and the earliest and latest occupational and educational codes of a person.

Estimation of the Disclosure Risk

The theory for estimating disclosure risk in a cross-sectional data set is well implemented, for example, in the R package sdcMicro [6,7]. In fact, for survey sample data, the approach of Franconi et al [38] or, for example, Skinner et al [27,39] can be used, or, for population data, the concepts of k-anonymity [40,41] or sample uniqueness [35,36]. We introduce an extension of this theory that provides a practical tool for quantifying disclosure risk for event history data.

Typically in anonymization, methods differ when continuous or categorical information is anonymized [6]. In addition, we distinguish between 2 scenarios—the matching of event dates (continuous measurements) and an attack on categorical key variables.

Event data are considered continuous measurements as there are multiple records for each person on a time scale.

As k-anonymity is already ensured (step 3) and population data are used, there is no need to quantify the disclosure risk for categorical key variables.

For continuous event dates, a neighborhood distance-based approach is proposed. Neighborhood matching, as introduced here and further introduced and applied in the *Results* section, assumes that the attacker has a database with exact event dates, which represents a worst-case scenario. For each individual in the anonymized data set, the nearest 3 individuals in the original nonanonymized data are determined by using Euclidean distances between event dates in the original and anonymized files. This is performed with replacement, meaning that the nearest neighbors are available to match for another individual in the data set. In case 1 of the 3 nearest neighbors is the correct match, we identify this observation to be of high risk. The number of risky observations is reported. The *Results* section shows the specific settings for our application.

Results

Anonymization of the Karonga HDSS Core Residency Data Set

First, it should be noted that the data set obviously cannot be spread into columns of events as migration and other event codes have possibly >1 entry, and the number of events differs between individuals. This makes it difficult to anonymize the data as the individuals have different events and different numbers of the same events at different times.

The key (identifying) variables are listed in Table 2.

Experiments with the HDSS core residency data set have shown that an additional identifying variable, the ID of the mother of a child, ID of the father and of the household, and the reason for in-migration and out-migration (reasons are marriage, divorce, start or end of work or education, and others) could potentially enlarge possible matches to approximately 10% of the original possible matches or individuals. Polygamy identifiers are not considered in this study. The usual approach for handling cluster information (eg, persons in households) for risk estimation of (enlarged) risk is, for example, described in Templ et al [6] and implemented in sdcMicro under the term of hierarchical risk estimation. However, as no further household information is available in this data set, this approach can be neglected. This is because household information can be used to identify individuals more easily; however, such additional household information is not available in our data set.

Other socioeconomic or sensible variables (eg, health status) were not included in the open-access data set.

Table 2. Key (identifying) variables of the health and demographic surveillance system core residency data set.

Key variable	Kind
Biological sex	Static variable
Year of birth	Static variable
Year of death	Static variable
Exact event date	Core event date[a]
Education	Status variable
Occupation	Status variable
Number of events per person	Static variable

[a]Contains dates at which the observed core events occurred (birth, death, in-migration, or out-migration).

Anonymization of Event Dates (Details Related to Step 1)

According to the random principle, a drawn number of days is randomly added to or subtracted from the event dates of birth, death, in-migration, and out-migration (equation 1; *Add Noise to One Event Date* section).

Four levels of noise were considered. In 3 scenarios, integer numbers (noise in days denoted by ε) for each event of a person (with E being the number of events of a person) were drawn with equal probability from the following intervals—depending on the noise level. In addition, a fourth scenario with normally distributed random noise is considered:

1. Noise level 1: $\varepsilon_{min}=46$; $\varepsilon_{max}=62$
2. Noise level 2: $\varepsilon_{min}=76$; $\varepsilon_{max}=93$
3. Noise level 3: $\varepsilon_{min}=106$; $\varepsilon_{max}=124$
4. Noise level 4: $u \sim N(\mu=0; \sigma=50)$

As described previously, random noise is added sequentially to the birth date, in-migration and out-migration dates, and death date to prevent record linkage and nearest-neighbor matching, with an external database containing exact event dates and information on sex, number of events, year of birth, year of death, occupational status, and educational level.

Anonymization of Static and Status Key Variables (Details to Steps 2 to 3)

To prevent successful matching, we achieved 3-anonymity through global recoding and local suppression using the heuristic implemented in the R package sdcMicro [6,7].

New variables are built for the year of birth, year of death, and year of the first change of educational and occupational status and used as key variables along with the sex of a person and the number of events of a person. Intermediate changes in educational and occupational levels are dropped. K-anonymity is then achieved by local suppression using the implemented methods in sdcMicro [7]. If the year of the latest event or the year of birth is suppressed, the noised year and noised event date are also suppressed. The number of events and the year of birth and death are set to the highest importance so that the implemented (weighted) local suppression algorithm in Templ et al [7] likely does not include missing values in these variables. Note that one suppression in a variable with high importance would increase the loss (function) in utility for >1 suppression in a variable with low importance (see Templ et al [7] for details).

After event date anonymization and status variable anonymization, the data are again matched to transform them into their original shape.

Disclosure Risk

To assess whether a data set was successfully anonymized, we quantified the disclosure risk. It must be reported only for event dates as, for the categorical key variables, k-anonymity is achieved, which satisfies our need to prevent successful matching.

The disclosure risk is calculated by matching each individual of the raw data set with the 3 nearest neighbors of the anonymized data with replacement using distance-based matching. In addition, an individual is matched with individuals who are born, died, or migrated within plus minus the same year as the true match, respectively, having the same (final) education, the same (final) occupation, and the same sex. If an individual has a missing value for one of these variables because of local suppression, that person is still considered a possible match if the rest of the variables meet the requirement.

If the match is correct, we assume that the attack was successful, and an individual can be reidentified. This means that if a person is in 3 of the nearest distances, we consider it unsafe. False-positive matches are not taken into account.

Table 3 reports the absolute and relative disclosure risk (in percentage) of the anonymized Karonga data set for all 4 scenarios, considering only individuals as possible matches who were born or had died or migrated in the range of +1 or –1 year of the date of birth, death, or migration, respectively, of the real match. We can observe that the risk is very low and that an attacker can hardly reidentify individuals. Note that the disclosure risk is already based on a worst-case scenario with 3 neighbors and by assuming the attacker uses the original nonanonymized data for matching. The low risk can also be explained by the fact that we choose ε_{min} to be relatively large; for example, for noise level 1 it is 46, meaning that for each event, the date is changed within at least 46 days. However, for death and birth, the risk increases as death is more unique than any of the other variables. The highest risk is connected with normal noise.

The computation time for neighborhood-based risk measurement, as proposed here, is high, and an implementation that uses parallel computing is preferable. Currently, the anonymization runs for 4 hours on a single-core Intel(R) Core i7-6700HQ central processing unit (CPU) with 2.60 GHz, and 8 days are spent for the risk assessment on all 4 noise levels on the HDSS core residency data set using 32 CPUs, Intel Xeon(R) Gold 5218 CPU with 2.30 GHz.

Table 3. Counts on successfully matched individuals and relative disclosure risk (in percentage; number of risky individuals divided by the number of individuals times 100) of the anonymized Karonga data set for all 4 levels of noises based on the matching scenario.

Scenario	Birth (number of successful matches)	Death (number of successful matches)	IMG[a] (number of successful matches)	OMG[b] (number of successful matches)
Absolute risk				
U(46;62)	1669	177	220	394
U(76;93)	1452	154	222	388
U(106;124)	1271	151	178	383
N(μ=0; σ=50)	1513	619	197	242
Relative risk (%)				
U(46;62)	2.3	5.0	0.5	0.8
U(76;93)	2.0	4.3	0.5	0.8
U(106;124)	1.7	4.2	0.4	0.8
N(μ=0; σ=50)	2.1	17.3	0.4	0.5

[a]IMG: in-migration.

[b]OMG: out-migration.

Utility

Utility measures specialized in a particular field should always be preferred to general measures ([42]; eg, as implemented in sdcMicro). To check the data utility after anonymization, visual comparisons of the original nonanonymized and anonymized data sets, as well as chi-square tests comparing contingency tables obtained from original and anonymized data, are shown.

Figure 2 shows the distribution of the date of birth from the original data and the noised data sets. The original data show a heaping in 1925, 1937, and 1945, which is still visible in the modified versions of the data set. This is not surprising as the noise was not too large.

The 2 midyear population pyramids for 2005 and 2015 are depicted in Figure 3. We distinguish between the population pyramids for the original nonanonymized data and anonymized data with noise levels of 1 to 4. Almost no differences were observed.

We do not explicitly show further graphs on the distribution of the date of death, in-migration, and out-migration, as the results are very similar to the previous figures; that is, there are no significant differences in the distributions.

Table 4 shows summary statistics of the time span between in-migration and subsequent out-migration of individuals. It shows only minimal differences; that is, all statistics are well preserved. The best results are obtained with noise scenario 4 (normal distributed noise). The results for out- to in-migration are comparable, except for the time between out- to in-migration. This can be shown in more details by a visualization.

Figure 4 visualizes this time span between in-migration and subsequent out-migration, as well as between out-migration and in-migration by box plots. The x-axis is presented on a $\log_{10}$ scale to better see minimal differences in the distribution of the time span between the original nonanonymized data and the anonymized data (almost no differences can be seen in the original scale). Almost no differences were found in the time span for in-migration to out-migration.

For the number of days between the out-migration and in-migration of a person, the worst results were obtained by scenario 4 (normal distributed noise). The reason for this difference between in- and out-migration is that people tend to return after out-migration much earlier than they leave the place after in-migration. Normal noise tends to increase the number of days of consecutive events if the events are close together.

Table 5 presents the results of the statistical test. The cross-tabulation for age class×event code×sex×event time category (2000-2004, 2005-2009, 2010-2014, and 2015-2020) was calculated from the original nonanonymized data and for the anonymized data. The corresponding cell counts were compared with each other by using a chi-square test. The results of the chi-square tests (Table 5) showed that the null hypothesis of equality of anonymized and original data can never be rejected.

Naturally, the differences between original and anonymization increase with an increasing level of noise, as can be seen in all the presented tables and visualizations of data utility. The best utility was achieved by adding normal noise (Table 5). However, even with noise level 3, the structure is well preserved, and the data utility is very high for all 4 noise levels investigated.

For the anonymization of the status variables on education and occupation, including sex, number of events of a person, year of birth, and year of death, a few values were suppressed to

achieve 3-anonymity (Table 6). The highest number of suppressions is present in variable end education (last educational status of a person), with approximately 0.64% (3735/583,480) suppression. Overall, 0.14% (808/583,480) of values were suppressed.

For the static and status variables, one of the most important information might be the last status of occupation and education. Figure 5 shows the frequencies of the corresponding contingency tables. The differences were minimal and not detectable by visual comparison. This is even more true for the other tabulations.

Figure 2. Distribution of the date of birth of the original data set and for the anonymized data set according to noise levels 1, 2, 3, and 4.

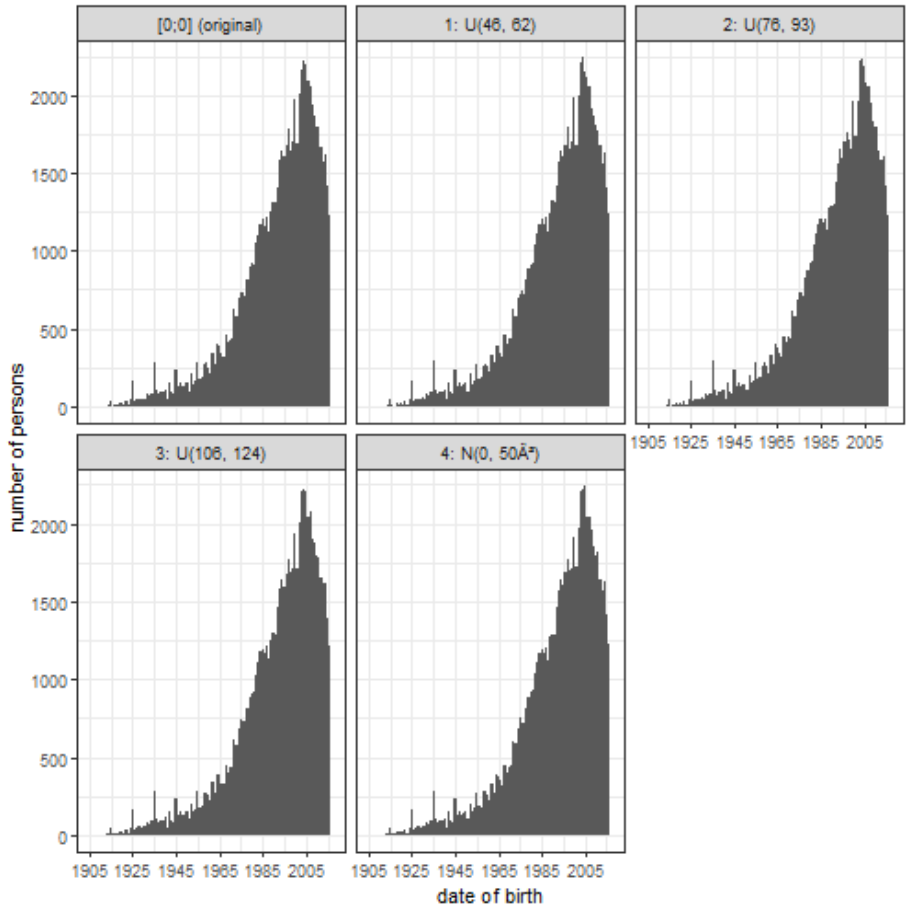


Figure 3. Population pyramids for 2005 and 2015 midyear population and age structure of the original and anonymized data according to noise levels 1, 2, 3, and 4 for men (left bars) and women (right bars).

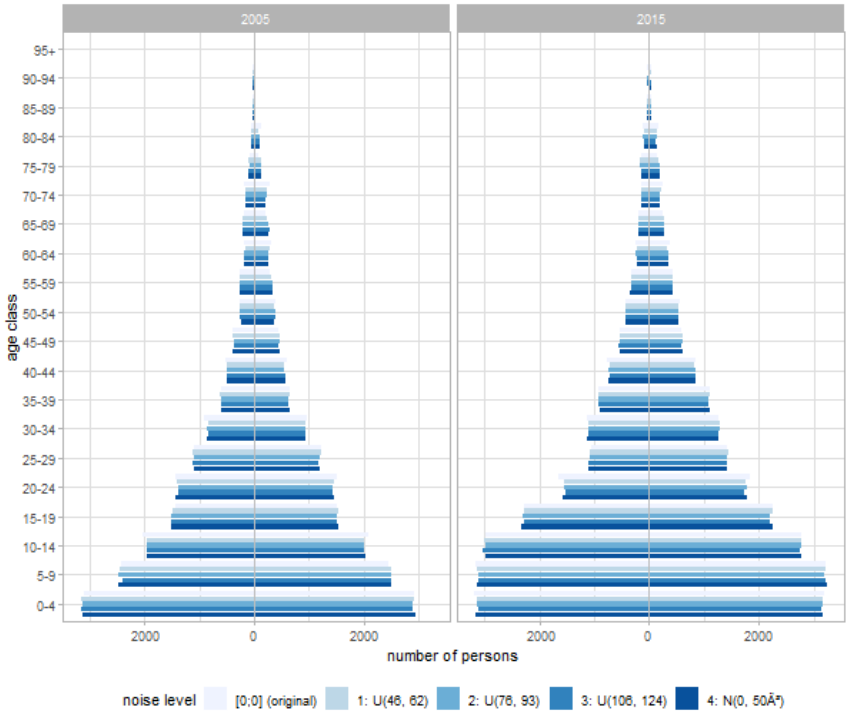


Table 4. Summary statistics for the number of days between in-migration and subsequent out-migration of a person for noise levels 1 to 4.

Scenario	Values (minimum-maximum)	Values, mean (SD)	<100 days (%)
(0;0) (original)	(0-5909)	862.05 (714)	2.2
U(46;62)	(0-5805)	846.67 (716)	3.4
U(76;93)	(0-5832)	839.25 (717)	4.4
U(106;124)	(0-5906)	831.30 (720)	5.5
N(μ=0; σ=50)	(0-5859)	862.58 (716)	2.9

Figure 4. Time span (in log10 scale) between in-migration and subsequent out-migration and out-migration to subsequent in-migration of the original data set and for the anonymized data sets by noise levels 1, 2, 3, and 4. Regarding in-migration to out-migration and out-migration to in-migration only individuals who in- or out-migrate, respectively, are considered.

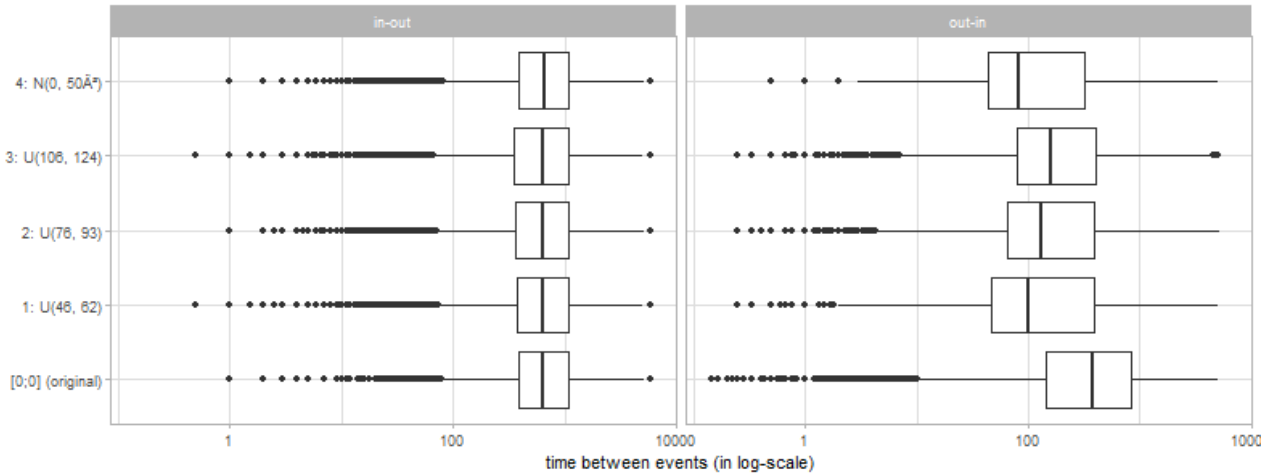


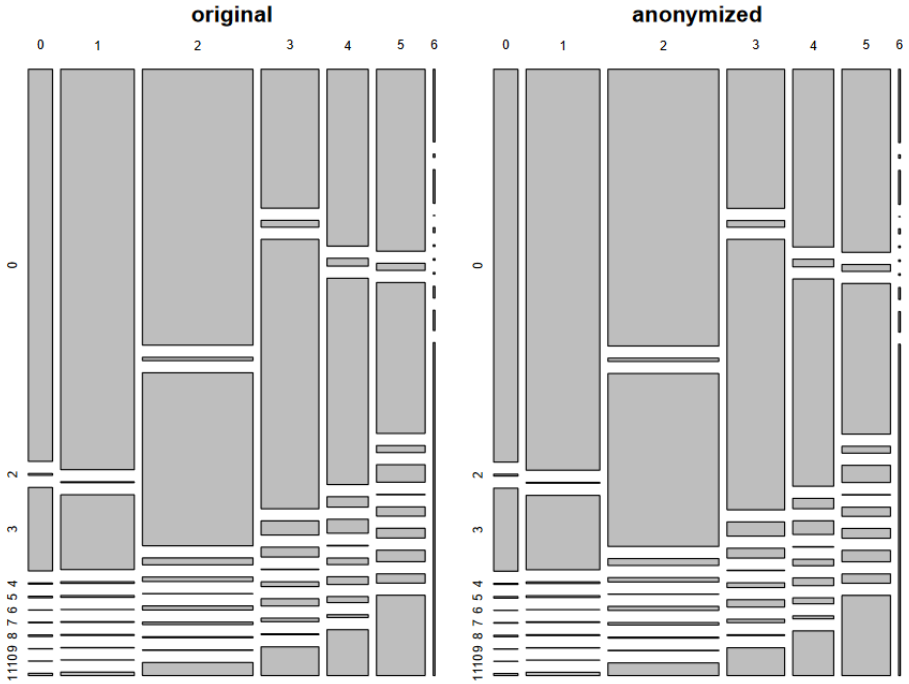
Table 5. Comparison of 4-dimensional contingency tables of the anonymized and original data using a chi-square test.

Statistics	U(46;62)	U(76;93)	U(106;124)	$N(\mu=0; \sigma=50)$
Test statistic	46.08	73.58	121.39	37.52
Critical value	237.24	237.24	237.24	237.24
P value	.99	.99	.99	.99

Table 6. Percentage of suppressions per variable and total number of suppressions per variable.

Suppression	Sex	Base education	Base occupation	End education	End occupation	Number of events	Year of birth	Year of death
Suppressions (%)	0.03	0.22	0.07	0.64	0.13	0.02	0	0
Total suppressions	23	160	53	465	94	13	0	0

Figure 5. Relative frequencies of the latest educational and latest occupational status of individuals for the original and the anonymized data set.



Discussion

Principal Findings

Providing open data (public use files) is a typical mechanism for HDSS data sharing, which is consistent with the funders' [2] call for lowering barriers to data access and in the interest of implementing sustainable data sharing models. However, more stringent anonymization is required than that for access-restricted and contracted files used for scientific purposes.

Anonymizing HDSS data is challenging, and no easy-to-apply solutions are available. The details matter to ensure consistency or credibility, and context knowledge is key for successful implementation. The presented approach is novel in several respects. This is the first time that a systematic approach has been adopted to determine the anonymization requirements for residency data from LMIC HDSS studies or for any other longitudinal data generated in these settings. Previously, anonymization of HDSS data was performed on an ad hoc basis. We grouped the variables into static, status (time-varying), and core event–specific variables and tackled the anonymization relating to the variables in each of these groupings.

We achieved an anonymized data set with very low disclosure risk and high utility, ready for sharing as a public use data file.

Using distance-based neighborhood matching, we simulated an attack under a nosy neighbor situation and using the worst-case scenario, where attackers have full information on the original data. We showed that the risk of disclosure is very low, even when assuming the worst-case scenario.

We explicitly defined a procedure for anonymizing core event dates as a major part of the HDSS event history data anonymization. Different levels of noise addition to the event history dates were evaluated for disclosure risk and data utility. It was found that high utility was maintained, even with the highest level of noise. The basic properties of the event data such as order, time span, and number of events were preserved compared with the original data. As can be seen from the

application and anonymization of event history dates, it is likely that the noise level and the loss of data utility will balance each other. Thus, a medium level of noise may be recommended to preserve the properties and usefulness of the data. In addition, the preservation of the time intervals between events is important for the successful implementation of this anonymization method. If the interval is too small, the added noise will is also automatically reduced by the algorithm.

Furthermore, our work explores the extent to which methods or tools such as sdcMicro can be used and for which aspects of longitudinal data. The guides for these tools focus on cross-sectional data and thus do not naturally lend themselves to the anonymization of multiple records per individual, which is the case in the Karonga HDSS core residency data that we used. In this regard, we transformed the time-varying variables of education level and occupation, year of death, year of birth, and the number of events for an individual before feeding them into the sdcMicro R package. The transformation involved limiting the number of transitions an individual had in the time-varying variables over time. This strategy preserves the data utility well, albeit providing fewer details than the original data.

The HDSS and medical science research communities in LMIC settings will be the primary beneficiaries of the results and methods presented in this paper; however, the results will be useful for anyone working on anonymizing longitudinal data sets, possibly including time-varying information and event history data with time-varying variables for purposes of sharing. If more sensitive variables such as medical conditions are added, l-diversity should also be checked. Alternatively, the PRAM [37] should be applied to medical conditions.

Future Work

The proposed approach of combining the range of values for the status variables into a baseline value and a final value may not be optimal for some analyses. This is one of the realities of data anonymization; it almost always results in data of lower utility than the original data. Further work is required to explore alternative handling of the status variables to determine the optimal handling of the transitions in the time-varying variables.

The disclosure risk is calculated based on 3 nearest-neighbor distance-based matchings. This matching strategy is already quite complex, with some constraints described previously, as well as dealing with missing values. However, other matching strategies might be possible, and specialized record linkage software [43] might also be considered.

Further work is also required to determine the right amount of offset for the core event dates. To determine this, it might be important to gather data from the participants to estimate what it would take to sufficiently offset the dates so that the potential nosy neighbors are unable to make guesses even in cases where events such as in-migration are rare.

Of course, not all data sets might have exactly the same structure as the HDSS residency data set used here. Other longitudinal data sets from HDSS settings, such as those generated from the observation of tuberculosis episodes or sexual partnership episodes, may contain features not fully catered for by our approach here. These issues need to be explored further.

Acknowledgments

The work of CK and MT was supported by a start-up grant from Network for the promotion of Institutional Health Partnerships, Switzerland. An interview about this grant and further details about the project can be found in German, English, and French [44]. The authors would especially like to thank Dörte Petit and Judith Safford from the University of Bern for their support on this project.

Malawi Epidemiology and Intervention Research Unit (MEIRU) and Zurich University of Applied Sciences (ZHAW) contributed in kind for some of CK's and MT's time on this project to enable them to fully explore the research collaboration and the methods used for anonymization.

The authors' gratitude also goes to the study participants and the iSHARE team for providing a platform through which health and demographic surveillance system data can be shared.

Conflicts of Interest

None declared.

References

1. Pisani E, Aaby P, Breugelmans JG, Carr D, Groves T, Helinski M, et al. Beyond open data: realising the health benefits of sharing data. BMJ 2016 Oct 10;355:i5295 [FREE Full text] [doi: 10.1136/bmj.i5295] [Medline: 27758792]
2. Walport M, Brest P. Sharing research data to improve public health. Lancet 2011 Feb 12;377(9765):537-539. [doi: 10.1016/S0140-6736(10)62234-9] [Medline: 21216456]
3. Sankoh O, Byass P. The INDEPTH Network: filling vital gaps in global epidemiology. Int J Epidemiol 2012 Jun;41(3):579-588 [FREE Full text] [doi: 10.1093/ije/dys081] [Medline: 22798690]
4. Federer LM, Belter CW, Joubert DJ, Livinski A, Lu YL, Snyders LN, et al. Data sharing in PLOS ONE: an analysis of data availability statements. PLoS One 2018 May 2;13(5):e0194768 [FREE Full text] [doi: 10.1371/journal.pone.0194768] [Medline: 29719004]

5. Herbst K, Juvekar S, Bhattacharjee T, Bangha M, Patharia N, Tei T, et al. The INDEPTH data repository: an international resource for longitudinal population and health data from health and demographic surveillance systems. J Empir Res Hum Res Ethics 2015 Jul;10(3):324-333 [FREE Full text] [doi: 10.1177/1556264615594600] [Medline: 26297754]
6. Templ M. Statistical Disclosure Control for Microdata: Methods and Applications in R. Cham, Switzerland: Springer; 2017.
7. Templ M, Kowarik A, Meindl B. Statistical disclosure control for micro-data using the R package sdcMicro. J Stat Soft 2015;67(4):1-36. [doi: 10.18637/jss.v067.i04]
8. Statistical Disclosure Control (sdcMicro). International Household Survey Network. URL: http://www.ihsn.org/software/disclosure-control-toolbox, [accessed 2022-02-22]
9. Templ M, Todorov V. The software environment R for official statistics and survey methodology. Aust J Stat 2016 Feb 29;45(1):97-124. [doi: 10.17713/ajs.v45i1.100]
10. Milliff A. Data security in practitioner-academic partnerships: an agenda for improvement. SSRN J 2020 Sep 16. [doi: 10.2139/ssrn.3693330]
11. Statistical Disclosure Control. The Centre for Humanitarian Data. 2019. URL: https://centre.humdata.org/guidance-note-statistical-disclosure-control/ [accessed 2021-10-20]
12. Hummerl M. Data-intensive computing with genomic data. BiobankCloud. 2013. URL: https://cordis.europa.eu/docs/projects/cnect/1/317871/080/deliverables/001-D52.pdf [accessed 2022-08-01]
13. Song X, Waitman LR, Hu Y, Luo B, Li F, Liu M. The impact of medical big data anonymization on early acute kidney injury risk prediction. AMIA Jt Summits Transl Sci Proc 2020 May 30;2020:617-625 [FREE Full text] [Medline: 32477684]
14. COVID-19 Case Privacy Review. GitHub. URL: https://github.com/CDCgov/covid_case_privacy_review/ [accessed 2021-10-20]
15. INDEPTH Network. Population and Health in Developing Countries: Population, Health, and Survival at INDEPTH Sites. Ottawa, ON, Canada: International Development Research Centre; 2002.
16. Ye Y, Wamukoya M, Ezeh A, Emina JB, Sankoh O. Health and demographic surveillance systems: a step towards full civil registration and vital statistics system in sub-Sahara Africa? BMC Public Health 2012 Sep 05;12:741 [FREE Full text] [doi: 10.1186/1471-2458-12-741] [Medline: 22950896]
17. Benzler BJ, Herbst K, MacLeod B. A data model for demographic surveillance systems. INDEPTH Network. 1998. URL: http://www.indepth-network.org/Resource%20Kit/INDEPTH%20DSS%20Resource%20Kit/LinkedDocuments/HRS2%20DSS%20Reference%20Data%20Model%20Paper.pdf [accessed 2021-10-20]
18. Crampin AC, Dube A, Mboma S, Price A, Chihana M, Jahn A, et al. Profile: the Karonga health and demographic surveillance system. Int J Epidemiol 2012 Jun;41(3):676-685 [FREE Full text] [doi: 10.1093/ije/dys088] [Medline: 22729235]
19. Crampin AC, Kayuni N, Amberbir A, Musicha C, Koole O, Tafatatha T, et al. Hypertension and diabetes in Africa: design and implementation of a large population-based study of burden and risk factors in rural and urban Malawi. Emerg Themes Epidemiol 2016 Feb 1;13:3 [FREE Full text] [doi: 10.1186/s12982-015-0039-2] [Medline: 26839575]
20. Bocquier P, Ginsburg C, Herbst K, Sankoh O, Collinson MA. A training manual for event history data management using Health and Demographic Surveillance System data. BMC Res Notes 2017 Jun 26;10(1):224 [FREE Full text] [doi: 10.1186/s13104-017-2541-9] [Medline: 28651610]
21. Dube A, Crampin AC. Malawi - Karonga HDSS INDEPTH Core Dataset 2003-2017 (Release 2019). INDEPTH Network Data Repository. 2019. URL: https://datacompass.lshtm.ac.uk/id/eprint/1738/ [accessed 2021-10-20]
22. Dwork C. Differential privacy: a survey of results. In: Proceedings of the 5th International Conference on the Theory and Applications of Models of Computation. 2008 Presented at: TAMC '08; April 25-29, 2008; Xi'an, China p. 1-19. [doi: 10.1007/978-3-540-79228-4_1]
23. Abadi M, Erlingsson Ú, Goodfellow I, McMahan HB, Mironov I, Papernot N, et al. On the protection of private information in machine learning systems: two recent approches. In: Proceedings of the IEEE 30th Computer Security Foundations Symposium. 2017 Presented at: CSF '17; August 21-25, 2017; Santa Barbara, CA, USA p. 1-6. [doi: 10.1109/csf.2017.10]
24. Domingo-Ferrer J, Sánchez D, Blanco-Justicia A. The limits of differential privacy (and its misuse in data release and machine learning). Commun ACM 2021 Jul;64(7):33-35 [FREE Full text] [doi: 10.1145/3433638]
25. Francis P. Dear differential privacy, put up or shut up. The Max Planck Institute for Software Systems. 2020 Jan 9. URL: http://www.mpi-sws.org/tr/2020-005.pdf [accessed 2021-10-20]
26. Bambauer J, Muralidhar K, Sarathy R. Fool's gold: an illustrated critique of differential privacy. Vanderbilt J Entertain Technol Law 2020;16(4):701-755.
27. Skinner CJ, Holmes DJ. Estimating the re-identification risk per record in microdata. J Off Stat 1998;14(4):361-372.
28. Hochguertel T, Weiss E. De facto anonymity in results. FDZ-Arbeitspapier Nr. 2012. URL: https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2011/50_Hochguertel-Weiss.pdf [accessed 2021-10-20]
29. Bond S, Brandt M, de Wolf PP. Guidelines for the checking of output based on microdata research. Data without Boundaries. 2013. URL: https://ec.europa.eu/eurostat/cros/system/files/dwb_standalone-document_output-checking-guidelines.pdf [accessed 2021-10-20]
30. Griffiths E, Greci C, Kotrotsios Y, Parker S, Scott J, Welpton R, et al. Handbook on Statistical Disclosure Control for Outputs. figshare. 2019. URL: https://figshare.com/articles/book/SDC_Handbook/9958520/1 [accessed 2021-10-20]

31. Dupriez O, Boyko E. Dissemination of microdata files: principles procedures and practices. International Household Survey Network. 2010 Aug. URL: http://www.ihsn.org/sites/default/files/resources/IHSN-WP005.pdf [accessed 2021-10-20]
32. Borde DS, Hebare PA, Dhanedhar PD. Overview of Web password hashing using salt techiques. Int Res J Eng Technol 2017 Nov;4(11):152-154.
33. Sauermann S, Kanjala C, Templ M, Austin CC, RDA COVID-19 WG. Preservation of individuals' privacy in shared COVID-19 related data. SSRN J 2020 Jul 17. [doi: 10.2139/ssrn.3648430]
34. Hundelpool A, Domingo-Ferrer J, Franconi L, Giessing S, Nordholt ES, Spicer K, et al. Statistical Disclosure Control. Hoboken, NJ, USA: Wiley; 2012.
35. Manning AM, Haglin DJ, Keane JA. A recursive search algorithm for statistical disclosure assessment. Data Min Knowl Disc 2007 Jul 10;16(2):165-196. [doi: 10.1007/s10618-007-0078-6]
36. Manning AM, Haglin DJ. A new algorithm for finding minimal sample uniques for use in statistical disclosure assessment. In: Proceedings of the 5th IEEE International Conference on Data Mining. 2005 Presented at: ICDM '05; November 27-30, 2005; Houston, TX, USA p. 290-297 URL: http://dblp.uni-trier.de/db/conf/icdm/icdm2005.html%5C#ManningH05 [doi: 10.1109/icdm.2005.10]
37. Gouweleeuw JM, Kooiman P, Willenborg LC, de Wolf PP. Post randomisation for statistical disclosure control: theory and implementation. J Off Stat 1998;14(4):463-478.
38. Franconi L, Polettini S. Individual risk estimation in μ-argus: a review. In: Proceedings of the CASC Project International Workshop on the Privacy in Statistical Databases. 2004 Presented at: PSD '04; June 9-11, 2004; Barcelona, Spain p. 262-272. [doi: 10.1007/978-3-540-25955-8_20]
39. Skinner C, Shlomo N. Assessing identification risk in survey microdata using log-linear models. J Am Stat Assoc 2008 Sep;103(483):989-1001. [doi: 10.1198/016214507000001328]
40. Samarati P, Sweeney L. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Electronic Privacy Information Center. 1998. URL: https://epic.org/wp-content/uploads/privacy/reidentification/Samarati_Sweeney_paper.pdf [accessed 2021-10-20]
41. Samarati P. Protecting respondents identities in microdata release. IEEE Trans Knowl Data Eng 2001;13(6):1010-1027. [doi: 10.1109/69.971193]
42. Templ M. Quality indicators for statistical disclosure methods: a case study on the structure of earnings survey. J Off Stat 2015 Dec 16;31(4):737-761. [doi: 10.1515/jos-2015-0043]
43. Sariyar M, Borg A. The RecordLinkage package: detecting errors in data. R J 2010;2(2):61-67. [doi: 10.32614/rj-2010-017]
44. Wurz J. A partnership building on health research data from Malawi. Esther Switzerland. 2021 Jun 1. URL: https://www.esther-switzerland.ch/a-partnership-building-on-health-research-data-from-malawi/ [accessed 2022-08-02]

Abbreviations

CPU: central processing unit
HDSS: health and demographic surveillance system
INSPIRE: Implementation Network for Sharing Population Information from Research Entities
LMIC: low- and middle-income countries
PRAM: postrandomization method
SDC: statistical disclosure control

Edited by H Bradley; submitted 25.10.21; peer-reviewed by M Sariyar, K Herbst; comments to author 23.02.22; revised version received 19.04.22; accepted 10.05.22; published 02.09.22.

Please cite as:
Templ M, Kanjala C, Siems I
Privacy of Study Participants in Open-access Health and Demographic Surveillance System Data: Requirements Analysis for Data Anonymization
JMIR Public Health Surveill 2022;8(9):e34472
URL: https://publichealth.jmir.org/2022/9/e34472
doi:10.2196/34472
PMID:36053573

Original Paper

The Impact of Nonrandom Missingness in Surveillance Data for Population-Level Summaries: Simulation Study

Paul Samuel Weiss[1*], MS; Lance Allyn Waller[1*], PhD

Rollins School of Public Health, Emory University, Atlanta, GA, United States

[*]all authors contributed equally

Corresponding Author:
Paul Samuel Weiss, MS
Rollins School of Public Health
Emory University
1518 Clifton Rd NE
Room 308
Atlanta, GA, 30322-4201
United States
Phone: 1 404 712 9641
Email: paul.weiss@emory.edu

Abstract

Background: Surveillance data are essential public health resources for guiding policy and allocation of human and capital resources. These data often consist of large collections of information based on nonrandom sample designs. Population estimates based on such data may be impacted by the underlying sample distribution compared to the true population of interest. In this study, we simulate a population of interest and allow response rates to vary in nonrandom ways to illustrate and measure the effect this has on population-based estimates of an important public health policy outcome.

Objective: The aim of this study was to illustrate the effect of nonrandom missingness on population-based survey sample estimation.

Methods: We simulated a population of respondents answering a survey question about their satisfaction with their community's policy regarding vaccination mandates for government personnel. We allowed response rates to differ between the generally satisfied and dissatisfied and considered the effect of common efforts to control for potential bias such as sampling weights, sample size inflation, and hypothesis tests for determining missingness at random. We compared these conditions via mean squared errors and sampling variability to characterize the bias in estimation arising under these different approaches.

Results: Sample estimates present clear and quantifiable bias, even in the most favorable response profile. On a 5-point Likert scale, nonrandom missingness resulted in errors averaging to almost a full point away from the truth. Efforts to mitigate bias through sample size inflation and sampling weights have negligible effects on the overall results. Additionally, hypothesis testing for departures from random missingness rarely detect the nonrandom missingness across the widest range of response profiles considered.

Conclusions: Our results suggest that assuming surveillance data are missing at random during analysis could provide estimates that are widely different from what we might see in the whole population. Policy decisions based on such potentially biased estimates could be devastating in terms of community disengagement and health disparities. Alternative approaches to analysis that move away from broad generalization of a mismeasured population at risk are necessary to identify the marginalized groups, where overall response may be very different from those observed in measured respondents.

(JMIR Public Health Surveill 2022;8(9):e37887) doi:10.2196/37887

KEYWORDS

surveillance; estimation; missing data; population-level estimates; health policy; public health policy; estimates; data; policy decision; bias; response rate

Introduction

The emergence of COVID-19 in 2019 has given rise to numerous challenges in global health. Many of those challenges have been easily observable and measurable. The intervening months produced countless publications on social distancing and vaccination measures and their resulting effects on the spread of the infection. Even now, epidemiological papers provide current updates on the disease's differential impact in high-risk populations compared to susceptible people whose risk may not be as high. Most of these analyses were conducted quickly, using available but incomplete data to provide rapid assessments. A challenge that has not been explored in as much detail is how the analysis of incomplete data without proper adjustments may be producing biased results that can lead to detrimental effects as we try to measure knowledge, attitudes, and behaviors related to various aspects of COVID-19.

Public health surveillance data are useful for noninvasively monitoring community health [1]. In some cases, these data are collected as part of an ongoing protocol with defined data elements and quality checks [eg, 11]. Increasingly, however, public health surveillance systems seek to draw conclusions and understanding from a broader collection of data available from administrative, commercial, or other sources [eg, 8-10].

Public health surveillance can be used to address a host of epidemiological questions at a micro level, drilling down to community clusters to identify the who, where, and when of disease concentration. A problem arises when the analyst tries to scale the analysis to the macro level when a nonrandom sample of individuals is used to try to draw inference to a population that the data cannot and do not accurately represent [2-5]. Brick [6] presents a number of potential solutions for reducing nonresponse bias, but these solutions tend to focus on improving response rates as well as statistical adjustment methods for reducing bias in data collections where nonresponse has occurred. In this paper, we quantify and illustrate the range and magnitude of problems encountered when we tried to infer the underlying global properties from an incompletely measured sample where the missingness of the data varied from random to nonrandom. In practice, analysts often turn to sampling weights [6] to control and reduce potential impacts of bias due to nonresponse [2]. In this study, we also examine when and if the use of sampling weights achieves this desired goal in public health surveillance and determine when and if such a strategy makes sense when considering data from a nonrandom microlevel sample for making macrolevel decisions.

Many statistical methods for dealing with missing data require that the data be missing at random (MAR). Investigators turn to methods like those presented in Cohen and Cohen [7], Simonoff [8], or Little and Rubin [9], applying statistical tests to their data to see if they meet this requirement, but these approaches may not provide sufficient rigor for identifying the underlying missingness mechanism, especially if the missingness mechanism is not associated with the auxiliary variables used in the testing [eg, 10]. These approaches are based on a null hypothesis that the data are MAR, and a failure to reject does not provide proof that the null is true. Such approaches also focus on missingness due to the variables involved in the testing and may not have strong statistical power to detect nonrandom missingness due to other reasons [7-9].

An additional approach favored by investigators interested in surveillance involves expanding the sample size through the addition of observations, widening eligibility criteria, or adding additional questions onto an existing large-scale questionnaire [eg, 8,11]. In the case of public-use data sets and surveillance systems, there is often an abundance of observations available for analysis. Extremely large sample sizes are considered to be rich data sources and provide an excellent opportunity to "find something." Nonprobability samples designed to maximize the number of respondents may present analysts with a wealth of data, but the impact of nonrandom missingness may limit the value of inference drawn from such studies. Although numerous examples of "spam the list" samples and imperfect censuses exist in the literature, we prefer to focus on the statistical impact of such methods rather than calling out our colleagues and peers in this paper for using such methods [eg, 9,11].

Applications of public health surveillance often focus on the data at hand rather than general principles of analytic performance in the presence of nonrandom missingness. In the sections below, we use simulation to explore and illustrate the impact of nonrandom missingness on a single survey item. Our approach allows us to investigate and quantify the error in the estimation of a mean when the randomness of the missingness varies from semicomplete to not complete at all. We also provide an illustration of how increasing the sample size impacts an estimator when the data are not MAR. Finally, we present the results of Cohen and Cohen's approach [7] for missingness at random for all of our results to assess the performance of this diagnostic approach in identifying when it may be unsafe to assume missingness at random in a given public health surveillance data set. Although it may be well known that, in theory, nonrandom missingness can influence statistical inference, our example provides an illustration of the nature and magnitude of this influence in a simple but realistic setting and in a simple tool for exploration and discovery by readers, students, and researchers.

Methods

Overview

A more detailed description of our methods could be seen in Multimedia Appendix 1. Briefly, we present a simulated example of item missingness using a Likert-scale outcome with 5 levels, similar to the kinds of questions often collected in public health surveys. To provide a frame of reference, we consider the outcome to be the answer to the question "how satisfied are you with your community's efforts to mandate vaccination for local government employees and public servants?" and simulate answers ranging from 1 to 5, one being very dissatisfied and 5 being very satisfied. The simulation uses a discrete random number generator to generate a large (N=100,000) population of potential respondents, where the response pattern is allowed to vary. We present some simulations where an individual's probability of response is

generally uniform across the values, some skewed toward the more satisfied and some skewed toward the less satisfied.

We induce missingness in the data via a uniform random value for each respondent. In our simulation, we compare the effects of data missing completely at random (MCAR) to not missing at random (NMAR) data, where the missingness is not random. We define mechanism as the reason for the data's missingness, as per the study by Little and Rubin [10]. When the mechanism is completely independent of the survey, then the data are MCAR. When the mechanism is directly associated with the missingness, then the data are NMAR. In the case where the mechanism can be identified and shown to be independent of the data of interest, then the data are MAR. Identifying the mechanism of missingness may be easier to do in the case of item missing data, where nonresponse of certain survey items may be analyzed using completeness in other items. In the case of unit nonresponse, it may be impossible to truly identify the missingness mechanism, as all information on nonrespondents is unavailable. When a mechanism is identified, it may be possible to control for it using multivariable modeling approaches. In this study, we simulate MCAR and NMAR data for a single survey item. For each simulated observation in the population, we also have complete data for race and sex. These demographic items provide auxiliary variables for Cohen and Cohen's approach [7]. We implement this approach to investigate the test's ability to effectively detect the NMAR mechanism.

Our simulation replicates 1000 random samples of our overall population and assigns observed values in the sample. Sampling weights [6] are introduced to allow the missing observations to be represented by complete observations.

We quantify the effect of missingness and weighting with the mean squared error (MSE) [11]. The MSE summarizes how far away an estimator is from the truth (on average) and summarizes two components of estimation performance: sampling variability (or sampling error) and bias. A full discussion of the MSE may be found in the Multimedia Appendix 1. Our simulation replicates samples and produces estimator variability, allowing us to estimate sampling variance as a summary of variation in estimation error from sample to sample. The square root of the difference gives us a simulation-based estimate of the estimator's bias. In the event of rounding leading to negative values of $bias^2$, we assign the observed bias a value of zero. In our simulation, the bias describes how far away, on average, our sample estimator is away from the true population mean satisfaction, rating in points on a 5-point Likert scale.

We present summary results for the following three population conditions:

- Uniform response across categories (ie, no response is more likely than other).
- Generally satisfied respondents in the population (ie, two satisfied responses are more likely than unsatisfied responses).
- Generally dissatisfied respondents in the population (ie, two dissatisfied responses are more likely than satisfied responses).

Under these conditions, we presented a constant response rate of 90% for the generally satisfied respondents (response of three or higher on the question) and allowed the missingness to vary from 10% to 90% for the dissatisfied respondents to explore the impact of nonrandom missingness. We also compared results for two sample sizes (800 and 8000) to see how this affects the estimators' behavior. A sample of 800 was chosen for a margin of error of approximately 3.5% for estimating the percentage of those satisfied with the community's vaccine mandates for government employees and civil servants. The sample size of 8000 was arbitrarily chosen as an inflation by a factor of 10 without specific statistical justification. The simulation was written in SAS 9.4 (Cary, NC). Refer to Multimedia Appendix 2 for the full program.

Ethical Considerations

No human subjects were involved in this simulation, so no institutional review board approval was necessary.

Results

Uniform Response Pattern

We used a uniform response pattern to describe a community without a particularly strong opinion about their government's efforts toward a vaccine mandate. Our response rates are assigned using a hypothetical convention that people who are generally supportive of public health practices will be inclined to respond to the survey and share their positive opinions, whereas people who are unhappy with the current state of affairs will decline (at a range of levels) to talk about their concerns with a stranger. We hold the response rate constant at 90% for the satisfied members of the community, suggesting their willingness to participate in the survey. We consider scenarios wherein the response rate in the dissatisfied group becomes progressively worse to measure impacts of this differential response on sampling variability, MSE, and bias, in terms of points on a 5-point Likert scale. We also report results after calculating weighted means in an attempt to adjust for nonresponse for samples from this community.

The first row in Figure 1 compares the performance of the estimator as the nonresponse rate becomes increasingly worse in the dissatisfied group. When response rates are similar between those who are generally satisfied and those who are generally dissatisfied, we see little evidence of bias; as the gap widens in response disparity, we can see a clear upward trend in MSE. Sampling variability appears to be relatively unaffected, but the sharp increase in bias indicates that although our estimator has considerable precision, our intervals are not likely to contain the true satisfaction rating of our full population. At its worst, the estimated policy satisfaction rating is off almost an entire scale point compared to the population truth. The first columns of Table 1 show how often Cohen and Cohen's approach [7] correctly identifies a departure from missingness at random. We see approximately 5% of the samples presenting with an association between one of the demographic variables and missingness, but we rarely see evidence to indicate missingness not at random using this approach, suggesting low statistical power to detect nonrandom missingness in our setting. In addition, Table 1 also reveals that the bias appears to be

considerable even when adjusting for nonrandom missingness using traditional adjustment weights. That indicates the use of sampling weights will not remove the underlying problem.

Interestingly, our results remained consistent when we expanded the sample size (Table 2). Increasing the sample size does not appear to reduce the bias in the estimator nor does it seem to have an impact on its overall variability. The inflated sample size neither reduced nor inflated the inherent bias of the estimator and had no apparent impact on the power of the Cohen and Cohens' approach to detect departures from missingness at random. Since MSE is a linear combination of variance and $bias^2$, we see no change in these quantities when the sample size increases. The sampling variance is improving, but negligibly so when compared to the impact the bias has on the quality of the estimator. The missing data contribute to a heavily biased satisfaction estimate, so the MSE or the average distance of our sample estimates from the true mean, is driven by the Bias component. The sample means vary very little between replicates, whereas they vary greatly from the true mean of the population.

Figure 1. Mean squared error (MSE), sampling variance, and bias by sample size and response pattern.

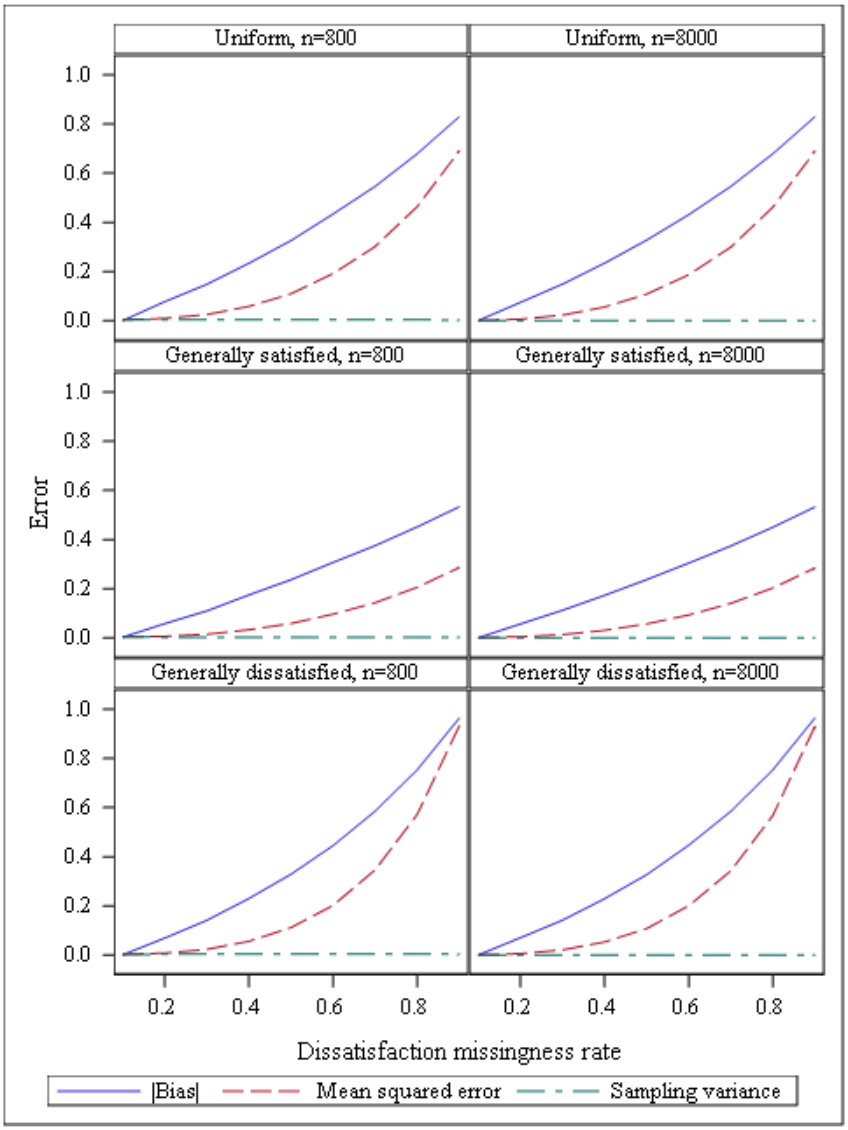


Table 1. Number of samples out of 1000, where Cohen and Cohen's approach [7] identifies nonrandom missingness using sex and race based on a sample size of 800.

Dissatisfied nonresponse rate	Uniform, n			Generally satisfied, n			Generally dissatisfied, n		
	Race	Sex	Both	Race	Sex	Both	Race	Sex	Both
10%	50	43	3	41	54	0	49	55	3
20%	46	52	4	52	53	5	59	49	2
30%	62	46	3	55	55	4	52	57	4
40%	45	48	3	54	69	3	57	61	6
50%	42	48	0	52	41	3	37	47	1
60%	51	37	1	43	40	1	46	59	1
70%	55	42	5	46	59	2	56	52	3
80%	53	47	1	50	63	4	51	61	2
90%	49	38	3	70	53	2	57	57	3

Table 2. Number of samples out of 1000, where Cohen and Cohen's approach [7] identifies nonrandom missingness using sex and race based on a sample size of 8000.

Dissatisfied nonresponse rate	Uniform, n			Generally satisfied, n			Generally dissatisfied, n		
	Race	Sex	Both	Race	Sex	Both	Race	Sex	Both
10%	34	38	1	43	44	1	43	52	2
20%	36	50	2	34	37	3	32	39	1
30%	35	43	2	39	37	3	36	43	2
40%	42	40	1	46	43	2	37	52	6
50%	34	49	0	45	58	3	42	51	1
60%	46	43	4	40	57	3	50	37	1
70%	38	50	2	53	44	5	48	36	0
80%	42	29	1	51	43	2	49	50	2
90%	29	31	2	60	60	3	46	51	1

Generally Satisfied Response Pattern

When the simulated respondents were generally satisfied, we observed less overall missingness in the data, even as the nonresponse rate for dissatisfied respondents increased. The second row in Figure 1 shows the estimator's behavior under a favorable response profile. In this population, we see that bias is considerably reduced because our sample is more representative of a truly more favorable population. We see that sampling variability is comparable between response profiles because the underlying sampling distribution of the estimator has not changed, so the variation of the estimates from sample to sample is unaffected. However, since these sample estimators are closer to the truth, we also see an arrested increase in the MSE and bias even as the dissatisfied response rate falls. The simulation also reveals that increasing the sample size had little impact on the bias in either direction.

Generally Dissatisfied Response Pattern

The third row of Figure 1 illustrates a missingness pattern where a large part of the population is both disenfranchised and disinclined to participate in the survey. In this scenario, the respondents present a much different population estimate than what is actually true. As with the other scenarios, there is little sample-to-sample variability. In the generally satisfied population, this posed a different kind of problem, as the respondents who were less likely to respond comprised a smaller part of the population as a whole. In the generally dissatisfied population, however, the respondent-based estimate was far removed from the population's truth; the resulting naive confidence intervals have no reliable coverage, while providing the appearance of high precision, suggesting a mostly satisfied population, even after adjusting with sampling weights. The estimator becomes biased much more quickly in the generally dissatisfied population, where nonresponse rates of 40%-50% result in the same apparent bias as much higher nonresponse rates in the two other populations considered. As in the previous case, the simulation results reveal that increasing the sample size does not appear to make a significant difference in this effect, and Cohen and Cohens' approach does not reliably result in a detection of the missingness' departure from randomness.

Discussion

Principal Findings

Our simulations illustrate the impact nonrandom missing data can have on population-based estimates even when analyzing a fairly simple survey sample. What we present in our examples indicate that basic diagnostic tests of missingness at random or the use of sampling weights do not automatically control for such biases and are not simple guarantees or workarounds to improve the quality of estimates.

Statistical discussions of missingness tend to focus on reducing nonresponse in the survey implementation [6] or fixing the data in the analysis [10]. These methods can be elegant and applicable to data collected under a specified design. Under MCAR missingness, a sample is simply reduced but not in a manner that generates bias. Under NMAR missingness, however, the "true" observed sample is a combination of the design (with known probabilities of selection) and the missingness pattern (typically with unknown probability of observation).

In surveillance data, particularly in a public health crisis, where data are needed quickly, existing surveys often are repurposed for additional data collection, or analysts include convenient data of unknown (if any) design. In this repurposed use (eg, through the addition of COVID-19 questions to ongoing surveys), we may well expect new (and unknown) patterns of missingness. Adjusting for design alone (via the design-based weights based on designed probability of selection but not necessarily probability of response) can adapt estimates for the anticipated design; however as seen above, important impacts of new causes of missingness will be missed. Specifically, the examples in our study illustrate how dissonance between sampling weights (adjusting for the probability of "selection") and patterns of missingness (which changes the probability of "response") can result in bias. Such impacts can be mitigated if the missingness occurs in subpopulations with low weights (as in our generally satisfied population example) but can be inflated if the missingness occurs in subpopulations receiving high sampling weights (as in our generally dissatisfied population example). Unless we know both the probability of selection and the probability of response, we cannot see the full picture and cannot adjust estimates appropriately with traditional reweighting methods.

As our simple example illustrates, the application of design weights should not be considered as a panacea for the challenges of extending survey designs for surveillance purposes. A closer look at Figure 1 reveals evidence as to why caution is necessary when applying weights in practice, particularly in settings where probabilities of response are unknown. In our simulation examples, while the MSE and bias trend upward as the dissatisfied response rate decreases, the sampling variability remains constant. The sampling variability is the essential statistic for producing confidence intervals and evaluating hypothesis tests—two broad statistical applications of inferential methods. We can see that confidence intervals produced from surveillance data may have the desired width determined by the sample size calculation, but the bias (due to nonrandom missingness) will result in a precise interval around the wrong number, potentially leading to very poor decisions, policies, and their consequences. Since sampling variability does not fully account for deviation from the truth the way the MSE does, in practice, we may never truly know how far away our sample's estimate is from the true but unknown population value. If we assume that the missingness is completely at random and produce biased estimates, our reported estimates may (and most likely will) lead to incorrect decisions with potentially long-standing public health implications.

A larger problem comes from the intention to use surveillance data to make global statements about a community. Extrapolation is often mentioned as a concern in modeling but rarely translated to estimators drawn from a nonrandom sample inferring parameters of a larger population. Our simulator shows that as the response rate becomes increasingly worse in a population subgroup, the sample's effectiveness at representing the larger community abates, in many cases rather drastically. Using data from a sample with unknown probabilities of observation, particularly survey data where the data may not be MAR, is a clear example of extrapolation. Ultimately, a failure to adequately represent a marginalized population may lead to political and social unrest. Policy decisions based on such data could result in creating or widening disparities already detrimental to social justice and health equity outcomes.

The simulations we showed in our study illustrate that estimates from survey samples have the potential to be heavily biased when extended beyond their design, especially in the presence of differential missingness due to imbalanced probability of response. Although the potential of such bias is known in theory, our simulations provide a basic but practical illustration of the potential magnitude of the problem. We note that these simulations represent a simplified (but perhaps not rare) illustration of the problem; the direction and magnitude of the bias can and likely will vary considerably as the relationship missingness shares with the survey changes. We contend that surveys with missing data will rarely (if ever) be random to some extent in surveillance settings and recommend considerable caution when applying survey weights based on sampling plans alone without consideration of potential differential missingness. In particular, we recommend that thoughtful summaries of potential biases accompany analyses and interpretation, especially those drawing from multiple available data sources. We recommend that, rather than using the survey data to look upward to the community, analysts should be encouraged to consider looking down to the observed population, instead.

Although it makes analytical sense to assume the data are MAR, this decision may come with a sizable cost. If we assume missingness at random in error, we reach conclusions that are far from the truth and could lead to devastating social consequences. If we assume that the missingness is not random in error, we make more cautious conclusions and open avenues to better identify and understand potentially underserved sections of our population of interest. Erring on the side of nonrandom missingness leads to a more socially responsible analysis of all the available information.

One limitation of our study is that we applied simple random sampling to simulate the survey experience where most surveillance data sets are multistage cluster designs. We note that a more complex design typically would likely lead to an inflation in the sampling variability but would not reduce the mean squared error or bias inherent in the differential response. In our examples, we also only considered three response patterns and arbitrarily assigned the population responses based on our own characterization of stronger satisfaction and dissatisfaction propensity. Our simulator is available to readers (See Multimedia Appendix 2) and is easily reprogrammed for more complex population response profiles. The simulator can also be modified to measure other kinds of response types (eg, continuous or binary) in addition to our Likert scale example. We limited our analysis to a basic survey design and response because the setting clearly illustrates our point and mirrors a very common setting when analyzing surveillance data.

Conclusions

Surveillance is an essential component of public health practice. Surveillance data allow us to produce useful descriptive measures to characterize the movement of disease through a population at risk. The current pandemic has produced vast amounts of data, much of which come from nonrandom samples or are drawn from surveys where data missingness patterns can obscure original sampling plans to the point that traditional sampling weights alone cannot provide proper adjustments to estimates. Our examples suggest an opportunity to develop new methods that move away from classical design-only approaches and move toward methods that explore designs for data collection and adjustment for patterns in data completeness that let us use the information more effectively for making better public health decisions for the entire population.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full description of methods.
[DOCX File , 16 KB - publichealth_v8i9e37887_app1.docx]

Multimedia Appendix 2
SAS Simulation Macro.
[DOCX File , 21 KB - publichealth_v8i9e37887_app2.docx]

References

1. Declich S, Carter AO. Public health surveillance: historical origins, methods and evaluation. Bull World Health Organ 1994;72(2):285-304 [FREE Full text] [Medline: 8205649]
2. Rader B, White LF, Burns MR, Chen J, Brilliant J, Cohen J, et al. Mask-wearing and control of SARS-CoV-2 transmission in the USA: a cross-sectional study. The Lancet Digital Health 2021 Mar;3(3):e148-e157. [doi: 10.1016/s2589-7500(20)30293-4]
3. Zhang L, Zhu S, Yao H, Li M, Si G, Tan X. Study on factors of people's wearing masks based on two online surveys: cross-sectional evidence from China. Int J Environ Res Public Health 2021 Mar 26;18(7):3447 [FREE Full text] [doi: 10.3390/ijerph18073447] [Medline: 33810355]
4. Gazmararian J, Weingart R, Campbell K, Cronin T, Ashta J. Impact of COVID-19 pandemic on the mental health of students from 2 semi-rural high schools in Georgia. J Sch Health 2021 May 12;91(5):356-369 [FREE Full text] [doi: 10.1111/josh.13007] [Medline: 33843084]
5. Basta NE, Sohel N, Sulis G, Wolfson C, Maimon G, Griffith LE, for the Canadian Longitudinal Study on Aging (CLSA) Research Team. Factors associated with willingness to receive a COVID-19 vaccine among 23,819 adults aged 50 years or older: an analysis of the Canadian longitudinal study on aging. Am J Epidemiol 2022 May 20;191(6):987-998 [FREE Full text] [doi: 10.1093/aje/kwac029] [Medline: 35166332]
6. Brick J. Unit nonresponse and weighting adjustments: a critical review. J Off Stat 2013;29(3):329-353. [doi: 10.2478/jos-2013-0026]
7. Cohen J, Cohen P, West SG, Aiken LS. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Hillside, NJ: Routledge; 1983.
8. Simonoff JS. Regression diagnostics to detect nonrandom missingness in linear regression. Technometrics 1988 May;30(2):205-214. [doi: 10.1080/00401706.1988.10488368]
9. Little RJA. A test of missing completely at random for multivariate data with missing values. JASA 1988 Dec;83(404):1198-1202. [doi: 10.1080/01621459.1988.10478722]
10. Little R. In: Rubin D, editor. Statistical Analysis with Missing Data, Third Edition. Hoboken, NJ: John Wiley and Sons; 2019.
11. Casella G, Berger RL. Statistical inference. Biometrics 1993 Mar;49(1):320. [doi: 10.2307/2532634]

Abbreviations

MSE: mean squared error
MAR: missing at random
MCAR: missing completely at random
NMAR: not missing at random

Edited by Y Khader; submitted 10.03.22; peer-reviewed by K Cummins, M Raimi; comments to author 03.04.22; revised version received 20.05.22; accepted 05.08.22; published 09.09.22.

Please cite as:
Weiss PS, Waller LA
The Impact of Nonrandom Missingness in Surveillance Data for Population-Level Summaries: Simulation Study
JMIR Public Health Surveill 2022;8(9):e37887
URL: https://publichealth.jmir.org/2022/9/e37887
doi:10.2196/37887
PMID:36083618

Original Paper

Psychometric Properties of the COVID-19 Pandemic Fatigue Scale: Cross-sectional Online Survey Study

Carmen Rodriguez-Blazquez[1], PhD; Maria Romay-Barja[2], PhD; Maria Falcon[3], PhD; Alba Ayala[4], MSc; Maria João Forjaz[1], PhD

[1]National Epidemiology Center, Carlos III Health Institute, Madrid, Spain

[2]Nacional Center of Tropical Diseases, Carlos III Health Institute, Madrid, Spain

[3]School of Medicine, University of Murcia, Murcia, Spain

[4]Department of Statistics, Carlos III University, Getafe, Spain

Corresponding Author:
Maria Falcon, PhD
School of Medicine
University of Murcia
Campus de Espinardo
Murcia, 31000
Spain
Phone: 34 868887583
Email: falcon@um.es

Abstract

Background: Pandemic fatigue is defined as feelings of demotivation to follow preventive measures against COVID-19, together with decreased trust in government and frequency of information-seeking behaviors.

Objective: This study aims to analyze the psychometric properties of the COVID-19–specific pandemic fatigue scale according to classical test theory (CTT) and Rasch model approaches in the general Spanish population.

Methods: This was a cross-sectional study in a representative sample of 1018 adults who completed an online survey in November 2020 in the framework of the COVID-19 Snapshot Monitoring (COSMO)-Spain project. The assessments included the 6-item COVID-19 Pandemic Fatigue Scale (CPFS) and other COVID-19–related variables: COVID-19 infection, adherence to preventive behaviors, information-seeking behavior, self-efficacy, worry, and cognitive and affective risk perception. Data quality, acceptability, reliability, and validity were analyzed according to CTT, and the fit to the Rasch model, unidimensionality, appropriateness of the response scale, item local independency, reliability (person-separation index [PSI]), and item-person distribution were also calculated.

Results: The mean CPFS score was 17.06 (SD 5.04, range 6-30), with higher scores for women, younger participants, participants who never seek information on COVID-19, those who think they would contract a mild disease in case of infection, those with higher level of worry about coronavirus/COVID-19, and those who felt depressed or felt the coronavirus/COVID-19 is spreading slowly (all $P<.01$). The Cronbach alpha for the CPFS was 0.74. In the confirmatory factor analysis, one factor was identified (root mean square error of approximation [RMSEA]=.02; comparative fit index [CFI]=.99; $\chi^2_5=8.06$, $P=.15$). The CPFS showed good fit to the Rasch model ($\chi^2_{24}=42.025$, $P=.01$, PSI=.642), unidimensionality (binomial 95% CI –.005 to .045), and item local independency.

Conclusions: Our results suggest that the CPFS has moderate reliability and internal consistency and it is composed of a single dimension. It is a useful tool to ascertain the level of pandemic fatigue in the general population, which may help to guide the communication and information strategies to face the COVID-19 pandemic.

(JMIR Public Health Surveill 2022;8(9):e34675) doi:10.2196/34675

KEYWORDS

COVID-19; pandemic fatigue; psychometric properties; Rasch analysis; validation; online survey; pandemic; fatigue; mental health; information seeking; health information

Introduction

The world has become familiar with the term "pandemic fatigue" in the context of COVID-19 [1-3]. This term has been used to describe different phenomena related to psychological distress and demotivation to follow preventive measures, as well as decreased trust in the government and frequency of information-seeking behaviors [3].

In 2020, the World Health Organization (WHO) defined pandemic fatigue as a "demotivation to follow recommended protective behaviors, emerging gradually over time and affected by a number of emotions, experiences and perceptions," proposing a framework on how to "maintain and reinvigorate" people's motivation to comply with COVID-19 response policies [4]. The WHO proposes that pandemic fatigue is expressed through an increasing number of people not sufficiently following or accepting recommendations and restrictions or decreasing their effort to keep themselves informed about the pandemic [4].

The restrictions adopted by the authorities to tackle this public health crisis have saved many lives but have also affected the mental and physical well-being of the population, social cohesion, economic stability, and community resilience [5]. After more than 2 years of restrictions, fatigue is an expected and natural response [4], and several authors have already measured how support of and compliance with nonpharmaceutical interventions (NPIs) have decreased in Spain [6] and worldwide as the pandemic evolves [7-12].

In Spain, compliance with health authorities' recommendations has been high [13], and the vaccination campaign has been a success [14,15]. Even so, compliance with NPIs remains important and will be in the future, as long as new variants continue to be a threat [16]. On the other hand, social, psychological, and economic consequences of the pandemic will continue over time, so the availability of a tool that measures pandemic fatigue is crucial.

Another challenge that public health authorities face is to keep the population informed in the context of a health crisis of unpredictable duration. Disinterest and information fatigue might be an obstacle for adherence to NPIs to combat the pandemic [17]. In Spain, there is a national strategy to improve health communication and address pandemic fatigue [18]. Public health communication strategies should focus on raising awareness in the event of future outbreaks and new restrictions [17].

Lilleholt et al [19], in 2021, conceptualized pandemic fatigue to represent a general demotivation toward following COVID-19–related health protective behaviors and staying informed about the development of the pandemic. The authors developed and validated the COVID-19 Pandemic Fatigue Scale (CPFS) with the aim of identifying who experiences it, analyzing related emotions and perceptions, and shedding light on the relationship between pandemic fatigue and health protective behaviors.

The aim of this study was to assess the psychometric properties of the Spanish version of the CPFS to measure pandemic fatigue in the Spanish general population, administered online, using 2 complementary methodological approaches: classic test theory (CTT) and the Rasch model. In addition to reliability and internal validity, Rasch analysis provides unique information such as differential item functioning for population groups and adequacy of the response scale. Finally, scales that fit the Rasch model provide results in a linear scale.

Methods

Design and Procedures

This is a cross-sectional, observational, nationwide study with survey data collected using an online questionnaire. This study is part of a larger project, the COVID-19 Snapshot Monitoring (COSMO)-SPAIN project [20], based on the COSMO tool developed by the WHO Europe Regional Office [6,21], with the aim of monitoring the knowledge, attitudes, compliance with the preventive measures, and risk perception of the Spanish population toward the COVID-19 pandemic, as well as informing COVID-19 outbreak response measures, including policies, interventions, and communications. More details can be found in the protocol of the COSMO-Spain study [22].

A nationally representative sample of 1018 subjects living in Spain was recruited. The sample was stratified to match the Spanish general population in terms of age, education, gender, and area of residence. A research company invited the potential participants who fit the inclusion criteria (both sexes, aged 18 years or older, and being able to answer an online questionnaire) and carried out the survey. The research market company has a panel of 157,535 members from the Spanish population. They contacted panel members who fit the inclusion criteria by email; 2655 invitations were sent, 1777 members participated in the survey (response rate 67%), and 1020 complete questionnaires were obtained. The data were collected between November 24, 2020, and November 27, 2020, at the end of the "second pandemic wave" in Spain. During that period, 60,462 cases of COVID-19 were detected, with a cumulative incidence of 128.6 over 14 days [23]. Mobility restrictions and capacity limitations in commercial establishments were present in different Spanish regions.

Ethical Review

The Ethics Committee of Carlos III Health Institute (CEI PI 59-2020-v2) approved the study protocol. The survey was anonymous, and the research company provided data with no identifying information to the researchers. Participants were informed on the purpose and characteristics of the study and provided informed consent by clicking a box.

Variables

Online Survey

The online survey included questions about participants' sociodemographic characteristics: sex (male, female), age, education (highest level of education attained: incomplete primary or less, primary, secondary, high school, and university), area of residence (village, 2000 to 50,000; town, 50,000 to 400,000; city, >400,000), and employment situation (working, student, domestic care, retired/pensioner, long-term unemployed,

unemployed, or Spanish temporary employment regulation due to COVID-19).

CPFS

The CPFS is a self-reported questionnaire based on the original version by Lilleholt et al [19]. It asks about demotivation toward COVID-19–related health-protective behaviors and staying informed about the development of the pandemic. The CPFS includes 6 items rated from 1 (strongly disagree) to 5 (strongly agree): "I am tired of all the COVID-19 discussions in TV shows, newspapers and radio programs, etc.," "I feel strained from following all of the behavioral regulations and recommendations around COVID-19," "I am sick of hearing about COVID-19," "I am tired of restraining myself to save those who are most vulnerable to COVID-19," "when friends or family members talk about COVID-19, I try to change the subject because I do not want to talk about it anymore," and "I am losing my spirit to fight against COVID-19." The total CPFS score is obtained by summing the items, with a maximum of 30 points that is indicative of a higher degree of COVID-19 pandemic fatigue. In the original study, it reached Cronbach α values of 0.83 and 0.87 (Danish and German studies, respectively) [19].

Other Variables

Other variables were included, as described in the following paragraphs (see also Multimedia Appendix 1).

COVID-19 infection was ascertained using the question "To your knowledge, are you, or have you been, infected with COVID-19?", with yes/no response options.

Adherence to preventive behaviors was assessed by asking how frequently respondents carried out a list of 12 measures to prevent infection from coronavirus/COVID-19, with 1 (never) to 5 (always) as scoring options. The listed behaviors included use of face masks: (1) using face masks following the recommendations and (2) wearing face masks in the presence of relatives and friends. It also included questions on hygienic behavior: (3) ventilating closed spaces; (4) using hydro alcoholic gel or disinfectants; (5) disinfecting surfaces; (6) washing hands; and (7) avoiding touching eyes, nose, and mouth with unwashed hands. Finally, it included physical distancing: (8) avoiding public transportation, (9) ensuring physical distancing, (10) avoiding social or family events, (11) not visiting relatives and friends if they are in quarantine, and (12) avoiding crowded spaces. The total number of preventive behaviors was calculated for each participant, computing the scores 4 and 5 as a positive response (=1) and summing them to obtain a score ranging from 0 to 12. The same procedure was applied to calculate the score for each type of preventive behavior.

Information-seeking behavior was assessed by asking respondents about the frequency of searching for information on coronavirus/COVID-19, answered on a scale from 1 (never) to 5 (several times a day).

Perceived self-efficacy was surveyed using the question "avoiding an infection with coronavirus/COVID-19 in the current situation is...?", with a response scale from 1 (very difficult) to 5 (very easy). This question, addressing self-assessed COVID-19 self-protection and avoidance ability, has been adapted from a previous study [24] by the original authors of the COSMO survey [21].

Level of worry about the coronavirus/COVID-19 in general was collected using a response scale from 1 (do not worry at all) to 5 (worry a lot).

Cognitive risk perception was measured using a question on perceived probability of getting infected with coronavirus/COVID-19, answered from 1 (very unlikely) to 5 (very likely), and a question on how severe would contracting the coronavirus/COVID-19 be for you, answered on a scale from 1 (not severe) to 5 (very severe) [19]. Both items were multiplied to obtain the value of the cognitive risk perception, ranging from 1 to 25, with higher scores indicative of higher risk perception.

Affective risk perception was collected using "the coronavirus/COVID-19 to me feels...," including 3 items: speed of propagation, with a response scale ranging from 1 (spreading slowly) and 5 (spreading fast); fear, scored from 1 (not fear-inducing) to 5 (fear-inducing); and mood, with a scale from 1 (it does not affect my mood) to 5 (makes me feel depressed) [19]. The responses to these questions were summed, obtaining a score that ranged from 3 to 15, with higher values indicating higher affective risk perception.

All items were originally in English and were translated by professional translators, reviewed and slightly modified by the COSMO-Spain team to adapt them to the Spanish context.

Data Analysis

Variables were summarized using descriptive statistics, including central tendency and dispersion measures (mean, median, and SD) and frequency and percentages, depending on their format.

Since the total CPFS score fit a normal distribution (Shapiro-Wilk test, P=.26), parametric statistics were used. According to the CTT [25], the following psychometric properties of the CPFS were analyzed: data quality and acceptability, structural validity, hypotheses testing (construct validity), and internal consistency.

Data quality and acceptability were computed by the mean, median, SD, and range of the observed versus theoretical values; skewness (criterion: −1 to +1); floor and ceiling effects (criterion: ≤15%) of the CPFS items; and total score [26].

For structural validity, exploratory (EFA) and confirmatory factor analyses (CFA) were used. For EFA, a principal component analysis with varimax rotation was applied. CFA used maximum likelihood estimations. A root mean squared error of approximation (RMSEA) ≤0.06 and comparative fit index (CFI) >0.9 indicated a good fit to the model [27].

Hypotheses testing (construct validity) included convergent and discriminative validity. Convergent validity was analyzed using Pearson correlation coefficients to ascertain the association of pandemic fatigue with related continuous variables: age, number and type of protective behaviors, and cognitive and affective risk perception. Following the literature [4,19], moderate-to-high correlation coefficients (r≥0.30 and r≥0.60) [28] between CPFS

and these variables were hypothesized. Regarding discriminative (known groups) validity, mean differences in total CPFS score in the sample grouped by relevant variables were calculated, using ANOVA and Student *t* tests. The following hypotheses were established, according to the literature [4,12]: a higher CPFS score would be reached in younger participants; those with lower education levels, risk perception, perceived severity, self-efficacy, level of worry, or information-seeking behavior; and those with higher levels of depression.

Internal consistency was examined by computing the Cronbach α coefficient (criterion ≥0.70), item total corrected correlations (standard *r* ≥0.40), interitem correlations, and the item homogeneity index (criterion >.30) [29].

The Rasch model, one of the most used applications of item response theory, was also applied to complete the information on the measurement properties of the CPFS provided by the CTT. According to the Rasch model, the answer to a certain item is a function between the person's ability (level of pandemic fatigue) and the item's difficulty (level of construct represented by that item), expressed in logits [30]. The following measurement properties were assessed: fit to the Rasch model, unidimensionality, appropriateness of the response scale, item local independency, reliability (person-separation index [PSI]), and item-person distribution. There are excellent tutorials and examples explaining the Rasch analysis process [31,32].

Since small deviations from the Rasch model are signaled as statistically significant when using large sample sizes, resulting in unnecessary model modifications, a random sample of 300 was drawn. This sample size allows for stable estimates regardless of targeting [33]. Fit to the Rasch model was considered when there was a nonsignificant chi-square test using Bonferroni correction for number of items (*P*>.008) [31]. Also, fit residuals were expected to be within the interval of –2.5 to +2.5 and item and person estimates to follow a normal distribution with a mean of 0 and SD of 1. Modifications were performed iteratively until model fit is achieved. PSI measures reliability and is interpreted similarly to Cronbach alpha. Threshold is the point of equal answer probability between 2 adjacent response categories. In case of disordered thresholds, adjacent response categories were collapsed. Unidimensionality was checked using a principal component analysis of residuals and then comparing person estimates with a binomial test; a lower bound of the 95% CI should be ≤.05 [34,35]. Local item independency, or the degree to which 1 item response does not lead to the response to another item, was analyzed though the correlation matrix of the residuals [36]. Differential item functioning (DIF) occurs when, for the same construct level, 2 or more sample groups answer in a statistically different way [37]. DIF was inspected through ANOVA by the following groups: age (groups defined by the median: ≤46 years; >46 years), gender, and education level (low: up to 14 years old; medium: secondary or professional training; high: university). Finally, the person-item threshold distribution was visually inspected.

CTT analysis was performed using SPSS 27.0 (IBM Corp, Armonk, NY) and Rasch analysis using RUMM2030 statistical software.

Results

The sample was formed by the same number of men and women, with a mean age 46.1 (SD 14.2, range 18-85) years (Table 1). Most participants were working (577/1018, 56.7%), 27.7% (282/1018) of them in a setting with a moderate risk of infection.

Table 1. Sociodemographic characteristics of the sample (n=1018) in the COVID-19 Snapshot Monitoring (COSMO)-Spain study, November 2020.

Variables	Results, n (%)
Sex	
Women	509 (50.0)
Men	509 (50.0)
Age groups (years)	
18-29	177 (17.4)
30-44	301 (29.6)
45-60	336 (33.0)
≥61	204 (20.0)
Education level	
Incomplete primary or less	31 (3.0)
Primary	240 (23.6)
Secondary	308 (30.3)
University	439 (43.1)
Employment	
Working	577 (56.7)
Student	41 (4.0)
Homemaker	32 (3.1)
Retired/pensioner	186 (18.3)
Long-term unemployed	100 (9.8)
Unemployed or ERTE[a]	82 (8.1)
Type of work	
With high risk of contagion	101 (9.9)
With moderate risk of contagion	282 (27.7)
No risk	69 (6.8)
Telework	102 (10.0)
Health care staff	23 (2.3)

[a]ERTE: Spanish Temporary Employment Regulation due to COVID-19; in Spanish, “expediente de regulación temporal de empleo.”

Psychometric Properties According to CTT

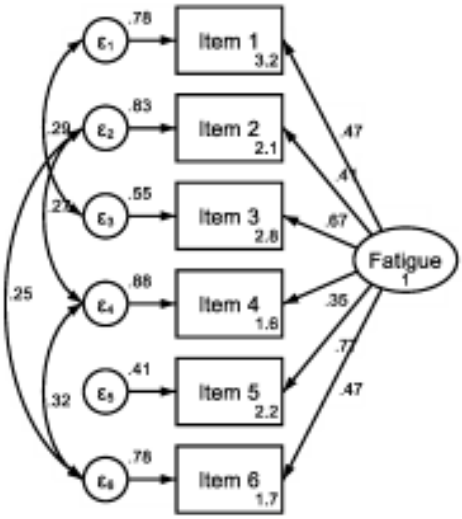
Table 2 shows the data quality and acceptability analysis of the CPFS. The mean total CPFS score was 17.06 (median 17.0, SD 5.04, range 6-30). Skewness of the total CPFS score was .13. All items reached the expected score range (1-5), and most of them showed a floor effect, especially items 6 and 4. Ceiling effect was marked in items 1 and 3. The total PFS score did not present floor or ceiling effects.

Table 2. Data quality and acceptability of the COVID-19 Pandemic Fatigue Scale (CPFS) in the COVID-19 Snapshot Monitoring (COSMO)-Spain study, November 2020.

CPFS item	Mean (SD)	Median	Minimum to maximum	Floor effect, %	Ceiling effect, %
1. I am tired of all the COVID-19 discussions in TV shows, newspapers and radio programs, etc.	3.88 (1.23)	4.00	1-5	7.3	43.3
2. I feel strained from following all of the behavioral regulations and recommendations around COVID-19	2.77 (1.32)	3.00	1-5	24.2	12.5
3. I am sick of hearing about COVID-19	3.52 (1.28)	4.00	1-5	9.8	30.3
4. I am tired of restraining myself to save those who are most vulnerable to COVID-19	2.07 (1.27)	2.00	1-5	48.3	7.2
5. When friends or family members talk about COVID-19, I try to change the subject because I do not want to talk about it anymore	2.81 (1.30)	3.00	1-5	21.6	13.5
6. I am losing my spirit to fight against COVID-19	2.01 (1.20)	2.00	1-5	49.7	5.1
CPFS Total	17.06 (5.04)	17.00	6-30	1.9	1.4

Regarding structural validity, EFA identified 2 factors with a correlation coefficient of .70, explaining 62.6% of variance. Figure 1 shows the path diagram of the CPFS using CFA. The 1-factor model obtained an RMSEA of .02 and CFI of .99 (χ^2_5=8.06, *P*=.15).

Figure 1. Path diagram of the COVID-19 Pandemic Fatigue Scale (CPFS) model in the COVID-19 Snapshot Monitoring (COSMO)-Spain study, November 2020. CPFS items: 1. I am tired of all the COVID-19 discussions in TV shows, newspapers and radio programs, etc.; 2. I feel strained from following all of the behavioral regulations and recommendations around COVID-19; 3. I am sick of hearing about COVID-19; 4. I am tired of restraining myself to save those who are most vulnerable to COVID-19; 5. When friends or family members talk about COVID-19, I try to change the subject because I do not want to talk about it anymore; 6. I am losing my spirit to fight against COVID-19.



The construct validity results of the CPFS appear in Table 3. CPFS scores were significantly higher for women, younger participants, participants who never seek information on COVID-19, those who think they would contract a mild disease in case of infection, those with lower level of worry about coronavirus/COVID-19, those who felt depressed, or participants who felt the coronavirus/COVID-19 was spreading slowly. The CPFS correlated with age (*r*=–0.20; *P*<.001), number of preventive measures (*r*=–0.16; *P*<.001), use of face masks and physical distancing (*r*=–0.12; *P*<.001), hygienic behavior (*r*=–0.13; *P*<.001), and affective risk perception (*r*=0.07; *P*=.03).

Internal consistency statistics are displayed in Table 4. The Cronbach α was .74, item homogeneity was .33, and item-total corrected correlation ranged from *r*=0.42 (item 1) to *r*=0.56 (item 3). Intercorrelation between items ranged from *r*=0.17 (item 1 with items 4 and 6) to *r*=0.51 (item 1 with items 3 and 5).

Table 3. Construct validity of the COVID-19 Pandemic Fatigue Scale (CPFS) in the COVID-19 Snapshot Monitoring (COSMO)-Spain study, November 2020.

Variables	Number of participants, n	CPFS scores, mean (SD)	*P* value[a]	Correlation with total CPFS[b]	*P* value
Sex					
Men	509	16.59 (4.75)	.003	__[c]	—
Women	509	17.53 (5.28)		—	—
Age (years)[d]	—	—	—	–0.20	<.001
Age group (years)					
18-29	177	18.90 (5.07)	<.001	—	—
30-44	301	17.32 (4.93)		—	—
45-60	336	16.51 (5.21)		—	—
≥61	204	15.99 (4.42)		—	—
Education level					
Primary or less	78	17.67 (5.53)	.47	—	—
Secondary	501	16.92 (4.97)		—	—
University	439	17.10 (5.04)		—	—
COVID-19 infection (past or present)					
Yes	69	17.45 (4.72)	.40	—	—
No	949	17.03 (5.07)		—	—
Number of preventive behaviors[d]	—	—	—	–0.16	<.001
Type of preventive behaviors					
Use of face masks[d]	—	—	—	–0.12	<.001
Hygienic behavior[d]	—	—	—	–0.13	<.001
Physical distancing[d]	—	—	—	–0.12	<.001
Information-seeking behavior					
Never (1-2)	270	18.53 (5.08)	<.001	—	—
Occasionally (3)	402	16.67 (4.45)		—	—
Several times a day (4-5)	346	16.35 (5.43)		—	—
Perceived self-efficacy: for me, avoiding an infection with coronavirus/COVID-19 is...?					
Very difficult/difficult (1-2)	246	17.52 (5.22)	.19	—	—
Neutral (3)	528	17.01 (4.72)		—	—
Easy/very easy (4-5)	244	16.70 (5.50)		—	—
General level of worry					
Not worry at all (1-2)	116	18.81 (5.56)	<.001	—	—
Moderate (3)	301	17.55 (4.66)		—	—
Worry a lot (4-5)	601	16.47 (5.02)		—	—
Cognitive risk perception[d]	—	—	—	–0.01	.85
Perceived probability of getting infected with coronavirus/COVID-19					
Very unlikely/unlikely (1-2)	268	17.24 (5.29)	.03	—	—
Neutral (3)	487	16.65 (4.63)		—	—
Likely/very likely (4-5)	263	17.62 (5.46)		—	—
How severe would contracting the coronavirus/ COVID-19 be for you					

Variables	Number of participants, n	CPFS scores, mean (SD)	*P* value[a]	Correlation with total CPFS[b]	*P* value
Very mild/mild (1-2)	176	17.59 (5.67)	.03	—	—
Normal (3)	480	17.28 (4.50)		—	—
Severe/very severe (4-5)	362	16.51 (5.35)		—	—
Affective risk perception[d]	—	—	—	0.07	.03
The coronavirus/COVID-19 to me feels…					
It is spreading slowly (1-2)	31	19.58 (6.37)	.01	—	—
Neutral (3)	185	17.36 (4.55)		—	—
It is spreading fast (4-5)	802	16.89 (5.07)		—	—
The coronavirus/ COVID-19 to me feels…					
Not fear-inducing (1-2)	265	17.02 (5.40)	.96	—	—
Neutral (3)	344	17.12 (4.61)		—	—
Fear-inducing (4-5)	409	17.03 (5.16)		—	—
The coronavirus/ COVID-19 to me feels…					
It does not affect my mood (1-2)	275	15.97 (4.94)	<.001	—	—
Neutral (3)	319	16.76 (4.60)		—	—
Depressed (4-5)	424	17.99 (5.26)		—	—

[a]Student *t* and ANOVA tests with Bonferroni correction.

[b]Pearson correlation coefficients for continuous variables.

[c]Not applicable.

[d]Continuous variable.

Table 4. Internal consistency of the COVID-19 Pandemic Fatigue Scale (CPFS) in the COVID-19 Snapshot Monitoring (COSMO)-Spain study, November 2020.

CPFS item	ITCC[a]	Cronbach α after item deletion	Intercorrelations					
			Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1. I am tired of all the COVID-19 discussions in TV shows, newspapers and radio programs, etc.	.42	.73	—[b]	0.21	0.51	0.17	0.37	0.17
2. I feel strained from following all of the behavioral regulations and recommendations around COVID-19	.46	.72	0.21	—	0.29	0.37	0.30	0.39
3. I am sick of hearing about COVID-19	.56	.69	0.51	0.29	—	0.23	0.51	0.31
4. I am tired of restraining myself to save those who are most vulnerable to COVID-19	.42	.72	0.17	0.37	0.23	—	0.26	0.43
5. When friends or family members talk about COVID-19, I try to change the subject because I do not want to talk about it anymore	.54	.69	0.37	0.30	0.51	0.26	—	0.37
6. I am losing my spirit to fight against COVID-19	.50	.70	0.17	0.39	0.31	0.43	0.37	—

[a]ITCC: item total corrected correlation.

[b]Not applicable.

Psychometric Properties According to the Rasch Model

The Rasch analysis showed that all items displayed disordered thresholds. After reducing the response options to scales with 2 to 4 points, according to the item, data showed a good fit to the Rasch model (χ^2_{24}=42.025; *P*=.01; PSI=.642; Table 5), unidimensionality (binomial 95% CI: –005 to .045), and item local independency.

Table 5. Goodness of fit to the Rasch Model of the COVID-19 Pandemic Fatigue Scale (CPFS) in the COVID-19 Snapshot Monitoring (COSMO)-Spain study, November 2020.

Attribute	Criteria	CPFS
Item fit residual		
Mean	0	–.189
SD	1	.852
Person fit residual		
Mean	0	–.285
SD	1	.736
Item-trait, χ^2 (df)	Low	42.025 (24)
Interaction *P* value	NS[a]	.0128
PSI[b]	>0.70	.642
Unidimensionality		
Independent *t* tests	<5%	2.00%
95% CI binomial	*[c]	.042-.091

[a]NS: nonsignificant.

[b]PSI: Personal Separation Index

[c]Lower bound should be ≤.05.

Table 6 presents the fit at the item level. Item 1 ("I am tired of all the COVID-19 discussions in TV shows, newspapers and radio programs, etc.") showed DIF by age, with older adults overestimating pandemic fatigue (Figure 2). No DIF was observed by sex or education level. The person-item threshold distribution was close to normality, with no floor or ceiling effects and item threshold locations ranging from –2 to 2 logits. There was a lack of items representing persons with lower and higher pandemic fatigue levels (Figure 3).

Table 6. Individual item fit of the COVID-19 Pandemic Fatigue Scale (CPFS) in the COVID-19 Snapshot Monitoring (COSMO)-Spain study, November 2020.

Item	Location	Standard error	Fit residual	χ^2_4	*P* value
1. I am tired of all the COVID-19 discussions in TV shows, newspapers and radio programs, etc.	–1.596	.099	1.177	8.672	.07
2. I feel strained from following all of the behavioral regulations and recommendations around COVID-19	0.289	.074	0.029	5.762	.22
3. I am sick of hearing about COVID-19	–1.346	.097	–1.179	10.429	.03
4. I am tired of restraining myself to save those who are most vulnerable to COVID-19	1.777	.183	–0.766	7.832	.10
5. When friends or family members talk about COVID-19, I try to change the subject because I do not want to talk about it anymore	–0.220	.090	0.251	3.163	.53
6. I am losing my spirit to fight against COVID-19	1.095	.088	–0.647	6.167	.19

Figure 2. Differential item functioning for item 1, by age groups defined by the median (≤46 years, >46 years) in the COVID-19 Snapshot Monitoring (COSMO)-Spain study, November 2020.

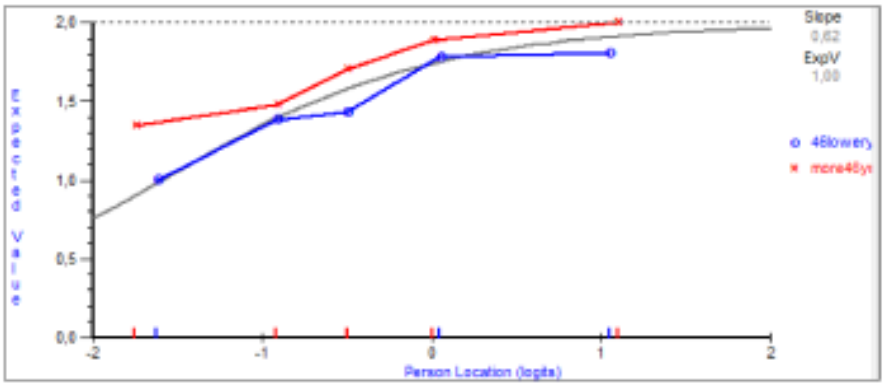
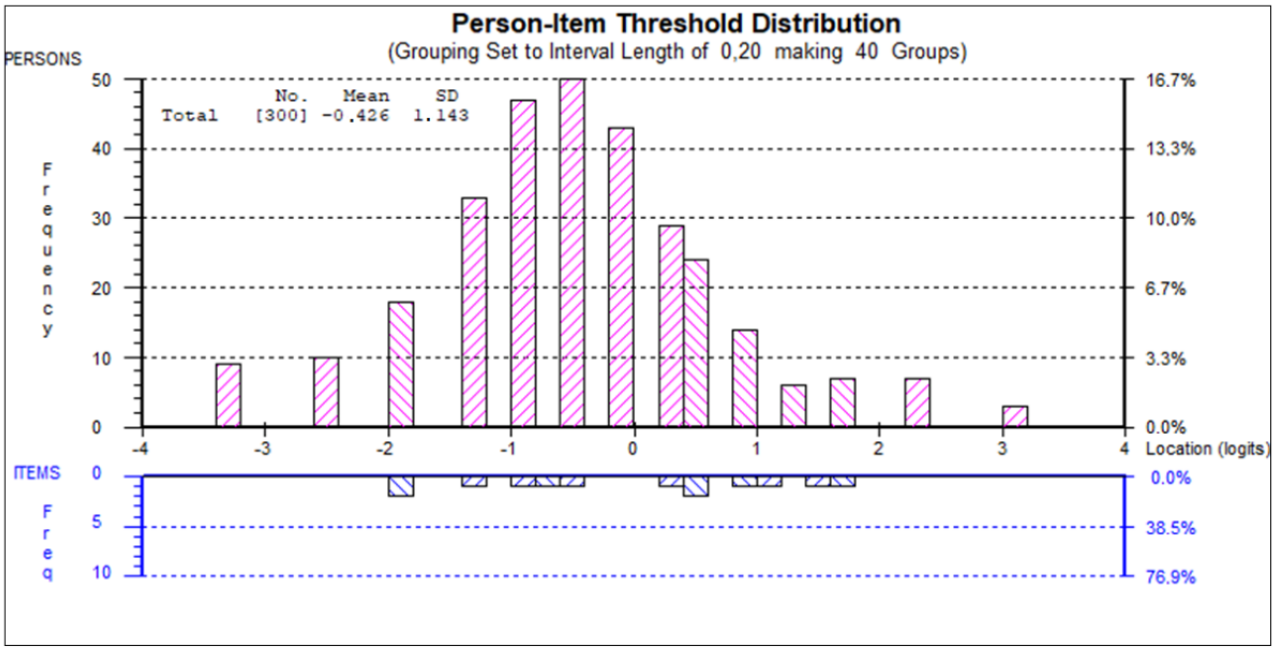


Figure 3. Person-item threshold distribution of the COVID-19 Pandemic Fatigue Scale (CPFS) in the COVID-19 Snapshot Monitoring (COSMO)-Spain study, November 2020. Note: the distribution of persons (top part) and item thresholds (bottom part) locations are shown on the same logit scale. Persons with higher fatigue level and more difficult thresholds are represented on the right.



Discussion

Pandemic fatigue is an important barrier to implementing NPIs, and monitoring its population levels requires the availability of valid and reliable measures [4]. This is the first validation study of the Spanish version of the CPFS scale in a representative sample of the Spanish general population, as part of an international behavioral insights initiative, the COSMO study [6,21]. The use of 2 complementary methodological approaches, the CTT and Rasch model, provides a robust testing of the psychometric properties of the scale.

Principal Findings

Results indicate that the Spanish population reported a moderate-to-high level of pandemic fatigue during the second pandemic wave (mean 17.06 points on a 24-point scale). Our data are in accordance with those of another Spanish study that reported a moderate level of pandemic fatigue (around 3 on a 1-4 scale) in the same period, using a different scale [38].

The CPFS displayed moderate reliability and internal consistency, allowing for group comparisons. Both CFA and Rasch analysis supported that the scale measures a single construct, indicating that the items may be summed to provide a meaningful total score. In addition, Rasch analysis allows converting the raw scores into a true interval scale, supporting the calculation of change scores and use of parametric statistics. The Rasch analysis results also indicate that respondents were not able to distinguish between all levels of the 5-point response scale. If these results are confirmed in further studies, the response scale may be simplified. However, there is no need to change the way the scale is administered, only how it is coded.

Items 1 (I am tired of all the COVID-19 discussions in TV shows, newspapers and radio programs, etc.) and 3 (I am sick of hearing about COVID-19) had a ceiling effect, indicating that a high percentage of respondents scored the highest level. This is consistent with the findings from the Rasch analysis, where these items presented the lowest locations, meaning that people with low levels of pandemic fatigue will easily endorse

items 1 and 3. On the contrary, items 4 and 6 are endorsed by people with higher levels of pandemic fatigue. The item hierarchy indicates that, when people start feeling some pandemic fatigue, they will first feel tired of COVID-19 discussions in the media (item 1). Only respondents with very high levels of pandemic fatigue will acknowledge that they are tired of restraining themselves to save those who are most vulnerable to COVID-19 (item 6). These results support the content validity of the scale.

One item presented a bias by age, with older adults overestimating pandemic fatigue scores in the same construct level. If further research confirms these results, separate item 1 locations may be calculated for each age groups. In the meanwhile, differences by age should be interpreted cautiously.

Comparison With Prior Work

The known-groups validity results showed that young people reported higher levels of pandemic fatigue. This may be explained by a higher impact of NPI on their social lives, which, in Spain, takes place mostly outside of home. Moreover, young people suffer from asymptomatic or mild disease if infected; they are less likely to adhere to preventive measures [8] and have a decreased risk perception [39]. As a result, they play a crucial role in the increase of incidence rates in several countries, such as in Spain during the summer of 2021 [40]. Younger age was also found to be associated with greater risk of decreasing or diminished interest and avoidance of news about COVID-19 [41]. Therefore, it is necessary to develop campaigns and information strategies specifically addressed to this group of population to overcome these difficulties.

The absence of DIF by sex indicates that the observed significant sex differences are not due to an item bias. We found higher pandemic fatigue scores in women than men, consistent with a study reporting that women are less likely to sustain long-term confinement [42]. However, another study reported higher levels of pandemic fatigue in Spanish men than in women [38]. These discrepancies might be due to how pandemic fatigue was measured in different studies.

In general, our hypotheses about convergent and discriminative validity were supported by the results. As explained in the previous pararaphs, women and young people showed higher levels of pandemic fatigue. Other variables associated with pandemic fatigue were the number and type of preventive behaviors, although with low correlation coefficients. Although pandemic fatigue impacts the adherence to protective behaviors, as stated in the definition by the WHO [4], in Spain, the levels of compliance with the main protective measures (use of face masks, washing hands, and social distance) were very high [13], and the use of face masks was compulsory at the time of data gathering. These results, similar to those in other countries [2], could be an explanation for the low correlation of the CPFS with preventive measures.

Decreased information-seeking behavior is another consequence of pandemic fatigue, and, as our results suggest, people who never or almost never look for information on COVID-19 scored significantly higher on the CPFS. Related to this, items 1 and 3 of the CPFS, which enquire about “information fatigue,” reached the highest mean scores. However, it is difficult to judge if the pandemic fatigue caused the decrease in information-seeking behavior. As hypothesized, people who reported higher levels of pandemic fatigue were those with lower levels of concern, who perceived they are unlikely to be infected, who believed the disease was spreading slowly, who thought they would experience mild disease if infected, or those with depression. Although no causal inferences can be inferred here, other studies have found that less fear of COVID-19 predicted diminished interest in or avoidance of COVID-19 news [41], which is part of the pandemic fatigue definition. In addition, information avoidance predicted a reluctance to engage in COVID-19 preventive behaviors in China [43]. Information avoidance was found to be related to more negative attitudes toward information searching, negative affective responses to risk, and perceived information overload [44].

Limitations and Strengths

This study has some limitations. First, we used a cross-sectional design, which provides data from the specific time of an evolving pandemic. Second, data were collected using an online survey, which might not reach minority, hard-to-reach population groups. However, the representative sample provides strength for the external validity of the study.

This study presents information on the measurement properties of the CPFS to measure pandemic fatigue in a valid and reliable way. Study strengths include the use of a representative population sample and both CTT and Rasch model methods. Results indicate that the CPFS is useful to monitor the level of population pandemic fatigue from the perspective of individuals. Being formed by only 6 items, the questionnaire is quick to apply, while providing good-quality measurement data for group comparisons. The use of the CPFS could help identify groups of people at risk of higher pandemic fatigue and the design of adequate intervention programs and information campaigns addressed at them.

Conclusions

In conclusion, the Spanish version of the CPFS is a promising questionnaire to measure pandemic fatigue at the population level, an important public health implication. Its strengths include that it is a brief, unidimensional scale, with a reliability level that allows for group comparisons, absence of bias by gender or education level, and satisfactory validity. As weaknesses and room for improvement, the reliability of the CPFS is not suitable for comparisons at the individual level; 1 item presented bias by age, and there was a lower than expected association with some behavioral aspects.

Further research is needed to test DIF by country. This is very important considering that the CPFS is being used in the WHO behavioral insight survey, in which more than 30 countries are participating. In addition, information on the scale’s sensitivity to change will be very useful to monitor changes due to pandemic progression and public health interventions. In addition, studies of the associated factors to pandemic fatigue in Spain and other countries, measured through the CPFS scale, would be very useful to design public health interventions to prevent and ease pandemic fatigue. Still, our study suggests that

younger adults, women, and people with lower risk perception are more susceptible to presenting with higher levels of fatigue. Communication strategies targeted at these groups will likely have a positive impact on lowering pandemic fatigue [18] and, consequently, increase adherence to protection measures.

Acknowledgments

The research was funded by the Carlos III Health Institute. The funder had no role in the study design; collection, analysis, and interpretation of data; writing of the paper; or decision to submit for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Description of the variables.
[DOCX File , 13 KB - publichealth_v8i9e34675_app1.docx]

References

1. Michie S, West R, Harvey N. The concept of "fatigue" in tackling covid-19. BMJ 2020 Nov 02;371:m4171. [doi: 10.1136/bmj.m4171] [Medline: 33139254]
2. Reicher S, Drury J. Pandemic fatigue? How adherence to covid-19 regulations has been misrepresented and why it matters. BMJ 2021 Jan 18;372:n137. [doi: 10.1136/bmj.n137] [Medline: 33461963]
3. Meichtry S, Sugden J, Barnett A. Pandemic Fatigue Is Real—And It's Spreading. The Wall Street Journal. 2020 Oct 26. URL: https://www.wsj.com/articles/pandemic-fatigue-is-realand-its-spreading-11603704601?reflink=desktopwebshare_permalink [accessed 2022-08-08]
4. Regional Office for Europe. Pandemic fatigue: reinvigorating the public to prevent COVID-19: policy considerations for Member States in the WHO European Region. World Health Organization. Regional Office for Europe. 2020. URL: https://apps.who.int/iris/handle/10665/335820 [accessed 2022-08-08]
5. Mukhtar S. Psychological health during the coronavirus disease 2019 pandemic outbreak. Int J Soc Psychiatry 2020 Aug 21;66(5):512-516 [FREE Full text] [doi: 10.1177/0020764020925835] [Medline: 32434402]
6. Rodríguez-Blázquez C, Romay-Barja M, Falcón M, Ayala A, Forjaz MJ. The COSMO-Spain survey: three first rounds of the WHO behavioral insights tool. Front Public Health 2021 May 31;9:678926 [FREE Full text] [doi: 10.3389/fpubh.2021.678926] [Medline: 34136459]
7. Petherick A, Goldszmidt R, Andrade EB, Furst R, Hale T, Pott A, et al. A worldwide assessment of changes in adherence to COVID-19 protective behaviours and hypothesized pandemic fatigue. Nat Hum Behav 2021 Sep;5(9):1145-1160. [doi: 10.1038/s41562-021-01181-x] [Medline: 34345009]
8. Wright L, Fancourt D. Do predictors of adherence to pandemic guidelines change over time? A panel study of 22,000 UK adults during the COVID-19 pandemic. Prev Med 2021 Dec;153:106713 [FREE Full text] [doi: 10.1016/j.ypmed.2021.106713] [Medline: 34242662]
9. Franzen A, Wöhner F. Fatigue during the COVID-19 pandemic: Evidence of social distancing adherence from a panel study of young adults in Switzerland. PLoS One 2021 Dec 10;16(12):e0261276 [FREE Full text] [doi: 10.1371/journal.pone.0261276] [Medline: 34890414]
10. Łaszewska A, Helter T, Simon J. Perceptions of Covid-19 lockdowns and related public health measures in Austria: a longitudinal online survey. BMC Public Health 2021 Aug 04;21(1):1502 [FREE Full text] [doi: 10.1186/s12889-021-11476-3] [Medline: 34344343]
11. Haktanir A, Can N, Seki T, Kurnaz MF, Dilmaç B. Do we experience pandemic fatigue? current state, predictors, and prevention. Curr Psychol 2021 Oct 20:1-12 [FREE Full text] [doi: 10.1007/s12144-021-02397-w] [Medline: 34690475]
12. MacIntyre CR, Nguyen P, Chughtai AA, Trent M, Gerber B, Steinhofel K, et al. Mask use, risk-mitigation behaviours and pandemic fatigue during the COVID-19 pandemic in five cities in Australia, the UK and USA: A cross-sectional survey. Int J Infect Dis 2021 May;106:199-207 [FREE Full text] [doi: 10.1016/j.ijid.2021.03.056] [Medline: 33771668]
13. Beca-Martínez MT, Romay-Barja M, Falcón-Romero M, Rodríguez-Blázquez C, Benito-Llanes A, Forjaz MJ. Compliance with the main preventive measures of COVID-19 in Spain: The role of knowledge, attitudes, practices, and risk perception. Transbound Emerg Dis 2022 Jul 10;69(4):e871-e882 [FREE Full text] [doi: 10.1111/tbed.14364] [Medline: 34730277]
14. Barandalla I, Alvarez C, Barreiro P, de Mendoza C, González-Crespo R, Soriano V. Impact of scaling up SARS-CoV-2 vaccination on COVID-19 hospitalizations in Spain. Int J Infect Dis 2021 Nov;112:81-88 [FREE Full text] [doi: 10.1016/j.ijid.2021.09.022] [Medline: 34536609]

15. Amiel S. How struggling Spain became one of Europe's vaccination champions. EuroNews. 2021 Mar 09. URL: https://www.euronews.com/my-europe/2021/09/03/how-struggling-spain-became-one-of-europe-s-vaccination-champions [accessed 2022-08-08]
16. de Bruin M, Suk J, Baggio M, Blomquist S, Falcon M, Forjaz M, et al. Behavioural insights and the evolving COVID-19 pandemic. Euro Surveill 2022 May;27(18):2100615 [FREE Full text] [doi: 10.2807/1560-7917.ES.2022.27.18.2100615] [Medline: 35514309]
17. Link E, Rosset M, Freytag A. Patterns of online information seeking and avoidance about SARS-CoV-2 and COVID-19. EJHC 2022 Apr 21;3(1):53-75. [doi: 10.47368/ejhc.2022.103]
18. Recomendaciones sobre Estrategias Comunicativas frente a la Fatiga Pandémica. Consejo Interterritorial del Sistema Nacional de Salud. 2021 Feb 03. URL: https://www.sanidad.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/documentos/Recomendaciones_estrategias_comunicacion_fatiga_pandemica.pdf [accessed 2022-08-08]
19. Lilleholt L, Zettler I, Betsch C, Böhm R. Pandemic fatigue: measurement, correlates, and consequences. PsyArXiv. Preprint posted online on December 20, 2020 [FREE Full text] [doi: 10.31234/osf.io/2xvbr]
20. Monitorización del comportamiento y las actitudes de la población relacionadas con la COVID-19 en España (COSMO-SPAIN). COSMO-Spain. URL: https://portalcne.isciii.es/cosmo-spain/ [accessed 2022-08-08]
21. Betsch C, Wieler LH, Habersaat K. Monitoring behavioural insights related to COVID-19. The Lancet 2020 Apr;395(10232):1255-1256. [doi: 10.1016/s0140-6736(20)30729-7]
22. Forjaz M, Romay Barja M, Falcón Romero M, Rodriguez-Blazquez C. Spain COVID-19 Snapshot MOnitoring (COSMO Spain): Monitoring knowledge, risk perceptions, preventive behaviours, and public trust in the current coronavirus outbreak in Spain. PsychArchives 2021:1 [FREE Full text] [doi: 10.23668/psycharchives.4877]
23. Equipo C. Informe nº 54. Situación de COVID-19 en España. Casos diagnosticados a partir 10 de mayo. RENAVE. CNE. CNM (ISCIII). Instituto de Salud Carlos III. 2020 Nov 25. URL: https://www.isciii.es/QueHacemos/Servicios/VigilanciaSaludPublicaRENAVE/EnfermedadesTransmisibles/Documents/INFORMES/Informes%20COVID-19/Informe%20COVID-19.%20N%c2%ba%2054_25%20de%20noviembre%20de%202020.pdf [accessed 2022-08-08]
24. Renner B, Schwarzer R. The motivation to eat a healthy diet: How intenders and nonintenders differ in terms of risk perception, outcome expectancies, self-efficacy, and nutrition behavior. Polish Psychological Bulletin 2005;36(1):15 [FREE Full text]
25. Aaronson N, Alonso J, Burnam A, Lohr KN, Patrick DL, Perrin E, et al. Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res 2002 May;11(3):193-205. [doi: 10.1023/a:1015291021312] [Medline: 12074258]
26. McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? Qual Life Res 1995 Aug;4(4):293-307. [doi: 10.1007/bf01593882]
27. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal 1999 Jan;6(1):1-55. [doi: 10.1080/10705519909540118]
28. Fisk JD, Brown MG, Sketris IS, Metz LM, Murray TJ, Stadnyk KJ. A comparison of health utility measures for the evaluation of multiple sclerosis treatments. J Neurol Neurosurg Psychiatry 2005 Jan 01;76(1):58-63 [FREE Full text] [doi: 10.1136/jnnp.2003.017897] [Medline: 15607996]
29. Jenkinson C, Fitzpatrick R. Cross-cultural evaluation of the short form 8-item Parkinson's Disease Questionnaire (PDQ-8): results from America, Canada, Japan, Italy and Spain. Parkinsonism Relat Disord 2007 Feb;13(1):22-28. [doi: 10.1016/j.parkreldis.2006.06.006] [Medline: 16931104]
30. Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. San Diego, CA: Mesa Press; 1993:773-702.
31. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Rheum 2007 Dec 15;57(8):1358-1362 [FREE Full text] [doi: 10.1002/art.23108] [Medline: 18050173]
32. Pallant JF, Tennant A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). Br J Clin Psychol 2007 Mar 24;46(Pt 1):1-18. [doi: 10.1348/014466506x96931] [Medline: 17472198]
33. Linacre J. Sample size and item calibration or person measure stability. Rasch Measurement Transactions 1994;7(4):328 [FREE Full text]
34. Smith EJ. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. J Appl Meas 2002;3(2):205-231. [Medline: 12011501]
35. Tennant A, Pallant JF. Unidimensionality Matters! (A Tale of Two Smiths?). Rasch Measurement Transactions 2006;20(1):1048-1051 [FREE Full text]
36. Baghaei P. Local dependency and Rasch measures. Rasch Measurement Transactions 2007;21(3):1105-1106 [FREE Full text]
37. Tennant A, Pallant J. DIF matters: A practical approach to test if Differential Item Functioning makes a difference. Rasch Measurement Transactions 2007;20(4):1082-1084 [FREE Full text]

38. Cuadrado E, Maldonado MA, Tabernero C, Arenas A, Castillo-Mayén R, Luque B. Construction and validation of a brief pandemic fatigue scale in the context of the coronavirus-19 public health crisis. Int J Public Health 2021 Aug 30;66:1604260 [FREE Full text] [doi: 10.3389/ijph.2021.1604260] [Medline: 34566554]
39. Maytin L, Maytin J, Agarwal P, Krenitsky A, Krenitsky J, Epstein RS. Attitudes and perceptions toward COVID-19 digital surveillance: survey of young adults in the United States. JMIR Form Res 2021 Jan 08;5(1):e23000 [FREE Full text] [doi: 10.2196/23000] [Medline: 33347420]
40. Spain's COVID-19 incidence rate rises, but officials see signs of hope. Reuters. 2021 Jul 26. URL: https://www.reuters.com/world/europe/spains-covid-19-incidence-rate-rises-officials-see-signs-hope-2021-07-26/ [accessed 2022-08-08]
41. Buneviciene I, Bunevicius R, Bagdonas S, Bunevicius A. COVID-19 media fatigue: predictors of decreasing interest and avoidance of COVID-19-related news. Public Health 2021 Jul;196:124-128 [FREE Full text] [doi: 10.1016/j.puhe.2021.05.024] [Medline: 34192604]
42. Martinez-Garcia M, Rabasa A, Barber X, Polotskaya K, Roomp K, Oliver N. Key factors affecting people's unwillingness to be confined during the COVID-19 pandemic in Spain: a large-scale population study. Sci Rep 2021 Sep 20;11(1):18626 [FREE Full text] [doi: 10.1038/s41598-021-97645-1] [Medline: 34545107]
43. Song S, Yao X, Wen N. What motivates Chinese consumers to avoid information about the COVID-19 pandemic?: The perspective of the stimulus-organism-response model. Inf Process Manag 2021 Jan;58(1):102407 [FREE Full text] [doi: 10.1016/j.ipm.2020.102407] [Medline: 33041437]
44. Link E. Information avoidance during health crises: Predictors of avoiding information about the COVID-19 pandemic among german news consumers. Inf Process Manag 2021 Nov;58(6):102714 [FREE Full text] [doi: 10.1016/j.ipm.2021.102714] [Medline: 34539039]

Abbreviations

CFA: confirmatory factor analyses
CFI: comparative fit index
COSMO: COVID-19 Snapshot Monitoring
CPFS: COVID-19 Pandemic Fatigue Scale
CTT: classical test theory
DIF: differential item functioning
EFA: exploratory factor analyses
NPI: nonpharmaceutical intervention
PSI: person-separation index
RMSEA: root mean square error of approximation
WHO: World Health Organization

Edited by T Sanchez, A Mavragani; submitted 03.11.21; peer-reviewed by S Song, J Voss; comments to author 30.04.22; revised version received 18.05.22; accepted 28.06.22; published 08.09.22.

Please cite as:
Rodriguez-Blazquez C, Romay-Barja M, Falcon M, Ayala A, Forjaz MJ
Psychometric Properties of the COVID-19 Pandemic Fatigue Scale: Cross-sectional Online Survey Study
JMIR Public Health Surveill 2022;8(9):e34675
URL: https://publichealth.jmir.org/2022/9/e34675
doi:10.2196/34675
PMID:35785547

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org