Los Alamos NATIONAL LABORATORY EST.1943

LA-UR-17-21702 (Accepted Manuscript)

Epidemiological Data Challenges: Planning for a More Robust Future Through Data Standards

Fairchild, Geoffrey
Tasseff, Byron Alexander
Khalsa, Hari Singh
Generous, Eric Nicholas Alexandre
Daughton, Ashlynn Rae
Velappan, Nileena
Priedhorsky, Reid
Deshpande, Alina

Provided by the author(s) and the Los Alamos National Laboratory (2019-02-20).

To be published in: Frontiers in Public Health

DOI to publisher's version: 10.3389/fpubh.2018.00336

Permalink to record: http://permalink.lanl.gov/object/view?what=info:lanl-repo/lareport/LA-UR-17-21702

Disclaimer:
Approved for public release. Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the Los Alamos National Security, LLC for the National Nuclear Security Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

RESEARCH LIBRARY

frontiers in Public Health

PERSPECTIVE
published: 23 November 2018
doi: 10.3389/fpubh.2018.00336

Epidemiological Data Challenges: Planning for a More Robust Future Through Data Standards

Geoffrey Fairchild[1], Byron Tasseff[1], Hari Khalsa[1], Nicholas Generous[1], Ashlynn R. Daughton[1], Nileena Velappan[2], Reid Priedhorsky[3] and Alina Deshpande[2]*

[1] Analytics, Intelligence, and Technology Division, Los Alamos National Laboratory, Los Alamos, NM, United States, [2] Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM, United States, [3] High Performance Computing Division, Los Alamos National Laboratory, Los Alamos, NM, United States

Accessible epidemiological data are of great value for emergency preparedness and response, understanding disease progression through a population, and building statistical and mechanistic disease models that enable forecasting. The status quo, however, renders acquiring and using such data difficult in practice. In many cases, a primary way of obtaining epidemiological data is through the internet, but the methods by which the data are presented to the public often differ drastically among institutions. As a result, there is a strong need for better data sharing practices. This paper identifies, in detail and with examples, the three key challenges one encounters when attempting to acquire and use epidemiological data: (1) *interfaces,* (2) *data formatting,* and (3) *reporting*. These challenges are used to provide suggestions and guidance for improvement as these systems evolve in the future. If these suggested data and interface recommendations were adhered to, epidemiological and public health analysis, modeling, and informatics work would be significantly streamlined, which can in turn yield better public health decision-making capabilities.

Keywords: data, computational epidemiology, public health, disease modeling, informatics, disease surveillance

OPEN ACCESS

Edited by:
Jimmy Thomas Efird, University of Newcastle, Australia

Reviewed by:
Vimal kishor Singh, Amity University Gurgaon, India
Marco Brandizi, Rothamsted Research (BBSRC), United Kingdom
Reza M. Salek, International Agency For Research On Cancer (IARC), France

****Correspondence:***
Geoffrey Fairchild
gfairchild@lanl.gov

Specialty section:
This article was submitted to Epidemiology, a section of the journal Frontiers in Public Health

Received: *17 July 2018*
Accepted: *01 November 2018*
Published: *23 November 2018*

Citation:
Fairchild G, Tasseff B, Khalsa H, Generous N, Daughton AR, Velappan N, Priedhorsky R and Deshpande A (2018) Epidemiological Data Challenges: Planning for a More Robust Future Through Data Standards. Front. Public Health 6:336. doi: 10.3389/fpubh.2018.00336

1. INTRODUCTION

At the heart of disease surveillance and modeling are epidemiological data. These data are generally presented as a time series of cases, T, for a geographic region, G, and for a demographic, D. The type of cases presented may vary depending on the context. For example, T may be a time series of confirmed or suspected cases, or it might be hospitalizations or deaths; in some circumstances, it may be a summation of some combination of these (e.g., confirmed + suspected cases). G is most commonly a political boundary; it might be a country, state/province, county/district, city, or sub-city region, such as a postal code or United States (U.S.) Census Bureau census tract. Depending on the context, D may simply be the the entire population of G, or it might be stratified by age, sex, race, education, or other relevant factors.

Epidemiological data have a variety of uses. From a public health perspective, they can be used to gain an understanding of population-level disease progression. This understanding can in turn be used to aid in decision-making and allocation of resources. Recent outbreaks like Ebola and Zika have demonstrated the value of accessible epidemiological data for emergency preparedness and the need for better data sharing (1). These data may influence vaccine distribution (2), and hospitals can anticipate surge capacity during an outbreak, allowing them to obtain extra temporary help if necessary (3, 4).

From a modeler's perspective, high quality reference data (also commonly referred to as ground truth data) are needed to enable prediction and forecasting (5). These data can be used to parameterize compartmental models (6) as well as stochastic agent-based models [e.g., (7–11)]. They can also be used to train and validate machine learning and statistical models [e.g., (12–19)].

The internet has become the predominant way to publish, share, and collect epidemiological data. While data standards exist for observational studies (20) and clinical research (21), for example, no such standards exist for the publication of the kind of public health-related epidemiological data described above. Despite the strong need to share and consume data, there are many legal, technical, political, and cultural challenges in implementing a standardized epidemiological data framework (22, 23). As a result, the methods by which data are presented to the public often differ significantly among data-sharing institutions (e.g., public health departments, ministries of health, data collection or aggregation services). Moreover, these problems are not unique to epidemiological data; the issues described in this paper are common across many different disciplines.

First, epidemiological data on the internet are presented to the user through a variety of *interfaces*. These interfaces vary widely not only in their appearance but also in their functionality. Some data are openly available through clear modern web interfaces, complete with well-documented programmer-friendly application programming interfaces (APIs), while others are displayed as static web pages that require error-prone and brittle web scraping. Still others are offered as machine-readable documents [e.g., comma-separate values (CSV), Microsoft Excel, Extensible Markup Language (XML), Adobe PDF]. Finally, some necessitate contacting a human, who then prepares and sends the requested data manually.

Second, there are many *data formats*. Data containers [e.g., CSV, JavaScript Object Notation (JSON)] and element formats (e.g., timestamp format, location name format) may differ. Character encodings (24) (e.g., ASCII, UTF-8) and line endings (25) (e.g., `\r\n`, `\n`) may also differ. Compounding these issues, formats can change over time (e.g., renaming or reordering spreadsheet columns). More broadly, these challenges are closely tied to schema, data model, and vocabulary standardization.

Finally, there are differences among institutions in their *reporting* habits; even within a single institution, there are often reporting nuances among diseases. For example, one context may be reported monthly (e.g., Q fever in Australia), while another context is reported weekly (e.g., influenza in the U.S.) or even more finely (e.g., 2014 West African Ebola outbreak). Furthermore, what is meant by "weekly" in one context may be different than another context (e.g., CDC epi weeks vs. irregular reporting intervals in Poland, as described later).

Together, these challenges make large-scale public health data analysis and modeling significantly more difficult and time-consuming. Gathering, cleaning, and eliciting relevant data often require more time than the actual analysis itself. This paper discusses these three key technical challenges involving public health-related epidemiological data, in detail and with examples that were identified through detailed analysis of data deposition practices around the globe. Building from this analysis, we offer a framework of best practices comprised of modern standards that should be adhered to when releasing epidemiological data to the public. Such a framework will enable a more robust future for accurate and high-confidence epidemiological data and analysis.

2. DISCUSSION

2.1. Interface Challenges

The *interface* is the mechanism by which data are presented to a user for consumption.

Epidemiological data repositories implementing current best practices provide an interactive web-based searching and filtering interface that enables users to easily export desired data in a variety of formats. These are generally accompanied by an API that allows users to programmatically acquire desired data. For example, if one wants to download the latest influenza surveillance data weekly, instead of manually navigating an interactive web interface each week to export the data, the process could be automated by writing code that interacts with the API. Such an interface provides the simplest and most powerful method of data acquisition. Examples of this type of interface are the U.S. Centers for Disease Control and Prevention (CDC)[1] and the World Health Organization (WHO) Global Health Observatory (GHO)[2].

While an interactive web-based interface coupled with an API is a best practice, it can be complex and expensive to implement. Many public health departments are under resource constraints and depend on older websites that tend to release data in one of two ways: 1) data are uploaded in some common format (e.g., CSV, Microsoft Excel, PDF) or 2) data are displayed in Hypertext Markup Language (HTML) tables. An example of the first is seen via Israel's Ministry of Health website, where data are provided weekly in Microsoft Excel formats (26). An example of the second is seen via Australia's Department of Health website, where data are provided within simply-formatted HTML tables (27).

Data uploaded in a common format can often be automatically downloaded and processed, and HTML tables can generally be automatically scraped and processed. While HTML scraping is often straightforward, there are some instances where it can be quite difficult. One example of a difficult-to-scrape data source is the Robert Koch Institute SurvStat 2.0 website (28). Although the service is capable of providing epidemiological data at superior spatial and temporal resolutions (county- and week-level, respectively), the interface is not easily amenable to scraping. First, the HTTP requests formed by the ASP.NET application cannot be easily reverse-engineered; this necessitates the use of browser-automation software like Selenium[3], which enables automating website user interaction, such as mouse clicks and keyboard presses, for data scraping. Second, the selection of new filters, attributes, and display options results in a

[1] https://data.cdc.gov/
[2] http://www.who.int/gho/en/
[3] http://www.seleniumhq.org/

newly-refreshed page for each change; because many options are required to obtain each desired dataset, scraping can take a long time.

Additionally, while there may be no technical barriers to downloading or scraping data, there may be barriers relating to a website's terms of service (TOS). In some instances, the TOS may prevent users from scraping or downloading data en masse; this is sometimes done to prevent unreasonable load on the website, for example. Ignoring the TOS raises ethical issues that are often overlooked in research; after all, the goals of most epidemiological researchers are benevolent, and the data are public and usually funded by taxpayers. Ignoring a website's TOS could also raise logistical issues related to publishing and institutional review board (IRB) approval.

A concern underlying all scraping efforts is that data scraping scripts are brittle. Web scraping relies on patterns in the HTML/CSS source code of a website. If an institution modifies its layouts, even slightly, scrapers may exhibit unexpected behavior.

In some cases, a human must be contacted directly, who then prepares and sends the requested data. However, these manually requested and prepared data are often saddled with many restrictions. For example, when one of the authors contacted a ministry of health for more detailed epidemiological data, the data were offered with a five-page data request form that significantly restricted use and sharing of the data. Furthermore, it stated that it would take "up to 3 months" to be released because of the review and approval from the various data owners (local, state, and territory health departments). These types of restrictions and hurdles to data access prevent the development and adoption of advanced analytics.

Finally, finding epidemiological data interfaces or data within an interface is often a time-consuming and error-prone task. For example, the Zika virus epidemic has resulted in increased global attention for Brazil, but it has not resulted in a single easy-to-understand machine-readable interface (29). Until just recently, Brazil's Ministry of Health maintained two separate lists of mosquito-borne illness epidemiological bulletins (30, 31). Although these lists pointed to the exact same bulletins, (31) is consistently more up-to-date than (30) (see **Figures 1**, **2**). Having multiple interfaces increases the likelihood of human error when collecting epidemiological data. For instance, if one assumes that there is only one official source for Zika, the most current information may be overlooked.

2.2. Data Format Challenges

The *data format* specifies how the data are read and written. There are two layers: 1) the *data container* and 2) the *element format*. The data container specifies how individual elements should be agglomerated; CSV is an example of a data container. The element format specifies how each individual element should be arranged; the ISO 8601 date and time specification is an example of an element formatting standard.

Data format challenges often provide the biggest obstacles that users must overcome. In order for an analyst to use data from multiple sources, they must first be merged. In practice, however, data from one institution are seldom available in a format that can be directly compared to data from another institution.

2.2.1. Data Containers

First, data container issues must be addressed. For example, CSV files are among the simplest file types to parse; they are plain text files with a simple structure (i.e., columns are separated by a comma, rows are separated by a newline). **Figure 3** demonstrates how epidemiological data might be provided in a CSV file. Any spreadsheet software can open CSV files natively, and most programming languages require no third-party libraries to read and write CSV files.

A conceptually similar file type to CSV is Microsoft Excel's XLSX. XLSX is a spreadsheet format developed by Microsoft and is a part of the Office Open XML (OOXML) specification. OOXML is a complex specification comprised of zipped XML files and other embedded data (e.g., images) (32). This format is common among public health practitioners due to the ubiquity of Microsoft Excel. For the programmer, however, this format presents a variety of challenges not present with CSV files. Due to the file type's complexity, a third-party library will be necessary in virtually all circumstances for reading/writing XLSX files (e.g., xlrd[4] and xlwt[5] for Python, Apache POI[6] for Java). Depending on the maturity of the library used, formulas, pivot tables, and other complex features should be handled with varying degrees of trust.

JSON is another common data container used on the internet (see **Figure 4** for an example). For instance, JSON data are commonly returned when querying an API endpoint. JSON is easy and fast to use; many programming languages offer built-in JSON read/write support (e.g., Python and Java). Additionally, similar to CSV, JSON is a plain text format that is human-readable. Unlike CSV, however, JSON is not limited to tabular data. JSON can represent more complex relationships between data and is conceptually more similar to XML. Due to its ubiquity and structure, a number of application-specific JSON standards are available. For example, GeoJSON (33) and TopoJSON (34) enable sharing geographic data. In 2016, Finnie et al. proposed EpiJSON, which offers a standardized way to encode epidemiological data (35). Although EpiJSON shows promise, it is young and has yet to be broadly embraced. To make adoption simpler, open-source EpiJSON libraries could be developed for common programming languages; currently, the only such library exists for the programming language R (36), but additional libraries should be developed for Python, Java, and other languages commonly used for epidemiological data analysis.

PDF files provide a number of unique challenges in addition to complexity. Extraction of data is the biggest challenge, as epidemiological information is often provided in mixed formats: textual (e.g., paragraphs of descriptive text in a report), graphical (e.g., bar and line charts), and tabular. Simply extracting the text of a PDF correctly and in the right order can prove to be a non-trivial challenge. Named-entity recognition and extraction, a natural language processing task, can be used to elicit case counts from unstructured text (19), but this supervised machine learning task requires knowledge of the language in which the

[4]https://github.com/python-excel/xlrd
[5]https://github.com/python-excel/xlwt
[6]https://poi.apache.org/

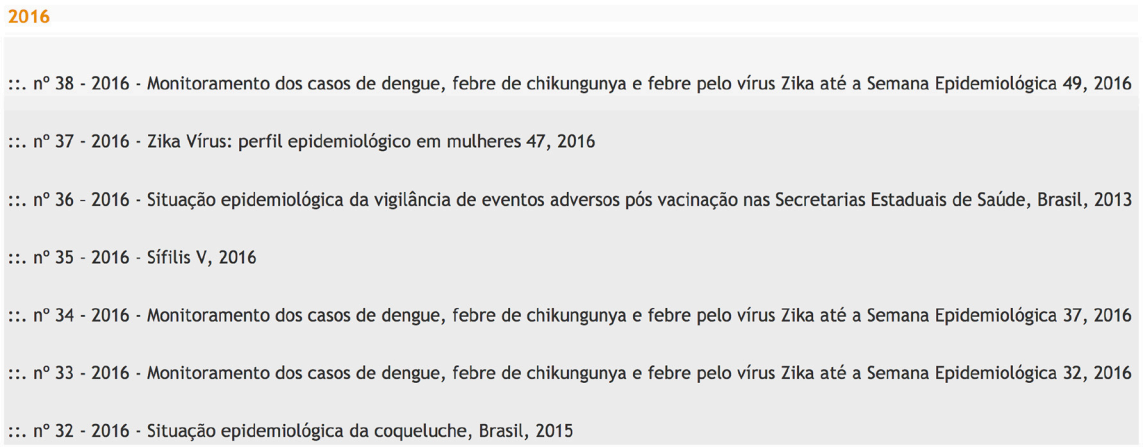

FIGURE 1 | Screenshot showing part of the mosquito-borne illness epidemiological bulletin list available at (31). This is the most current and complete list, with data available through the 38th week of 2016.

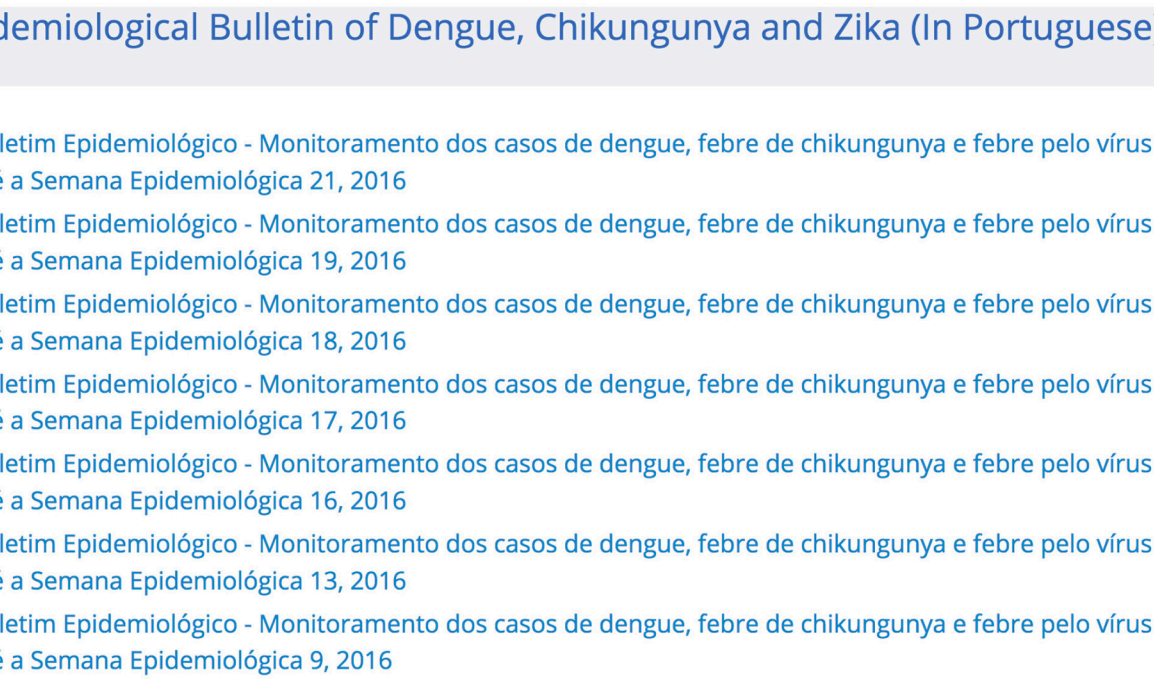

FIGURE 2 | Screenshot showing part of the mosquito-borne illness epidemiological bulletin list available at (30). This list only goes through week 21 of 2016 and is missing a number of weeks when compared to the list in **Figure 1**. This screenshot was taken at the same time as the one in **Figure 1**.

document is published, as well as epidemiological subject matter expertise. Graphical data are intended for the human eye. While graphical data can potentially be digitized using software like WebPlotDigitizer[7], this cannot always be reliably automated. Even tabular data, which visually appear structured, are typically difficult to extract due to the variety of ways a table can be presented in a PDF document (37).

Furthermore, PDF files need not even contain text. In a number of circumstances, the PDF files that institutions provide simply contain scanned images of documents. The resulting PDF simply contains the image, rather than the raw text that comprised the original document. For example, many of the weekly reports available through the Department of Health website for the Philippines (38) are PDFs of scanned documents [e.g., (39, 40)]. Text can potentially be elicited with optical character recognition (OCR) software, but the quality of the resulting textual data will

[7]http://arohatgi.info/WebPlotDigitizer/

```
date,location,cases
2013-11-05,United States,4
2013-11-05,Germany,8
2013-11-11,South Africa,9
2013-11-12,Japan,6
```

FIGURE 3 | Sample epidemiological case count data in CSV format. CSV files are plain text files that allow tabular data to be laid out as rows separated by newlines and columns separated by commas. This time series does not contain real data and only exists for demonstration purposes.

```
[
  {
    "date":"2013-11-05",
    "locations":{
      "United States":4,
      "Germany":8
    }
  },
  {
    "date":"2013-11-11",
    "locations":{
      "South Africa":9
    }
  },
  {
    "date":"2013-11-12",
    "locations":{
      "Japan":6
    }
  }
]
```

FIGURE 4 | Sample epidemiological case count data in a simple JSON format. Compared to CSV (demonstrated in **Figure 3**), JSON contains more structure that can more rigorously specify data relationships (including hierarchical relationships). Note that this is *not* EpiJSON; EpiJSON can be quite verbose (due to, for example, metadata specifications and GeoJSON-specified locations), and the authors felt a complete EpiJSON example would take up an unreasonable amount of space in this paper. As in **Figure 3**, this time series does not contain real data and only exists for demonstration purposes.

vary significantly depending on the quality of the scanned images.

Finally, one must be aware of the character encoding when reading text. Since, at the basic level, computers represent all data using binary bits, there must be some binary representation of each character or symbol in an alphabet or language; the character encoding specifies how the raw bits stored in a file should be converted to readable text and vice versa (24). While there are a number of possible encodings, ASCII, ISO-8859-1, and UTF-8 are among the most common encodings encountered in practice. In 2012, UTF-8 surpassed 60% adoption across the web (41) and is currently approaching the 90% mark (42). Encoding differences are important; for example, while reading ASCII text as UTF-8 yields correct results, the converse does not.

2.2.2. Element Format

2.2.2.1. Date and Time

Beyond data container challenges, there are a number of element formatting differences that must be addressed. First and foremost are date and time formatting discrepancies. While the ISO 8601 date/time standard has existed since 1988, it is often bypassed in favor of locale-dependent formats. For example, much of Europe follows the `day-month-year` convention, the U.S. follows the `month-day-year` convention, and China follows the `year-month-day` convention. Depending on the locale, `03-09-2005` may refer to March 9, 2005 or September 3, 2005. Additionally, not all locales use the Gregorian calendar. Thailand, for example, uses the Buddhist calendar. The current year, as represented by the Gregorian calendar, is 2018; a Thai timestamp would instead specify 2561. Finally, some locales use 24-h time, while others use 12-h time.

In addition to these timestamp parsing differences, there are significant *implicit* timestamp differences that must be understood. To understand these, one must first recognize that a timestamp on a typical disease curve usually implicitly refers to an interval of time (i.e., actual event-level epidemiological data are rare). To illustrate this, consider the time series in **Table 1**. Each timestamp can be interpreted using one of three possible *interval types*:

Leading: The timestamp starts the interval, and the interval ends the "instant" before the next specified timestamp. **Table 2** shows how the time series in **Table 1** would be transformed to an *interval series* with an interval type of leading.

Trailing exclusive: The timestamp ends the interval but is not included in the interval; **Table 3** demonstrates this transformation.

Trailing inclusive: The timestamp ends the interval and is included in the interval; **Table 4** shows this transformation.

Note that we do not currently feel it is necessary to include a leading exclusive option; leading will always be inclusive.

As an example, the CDC standardizes reporting dates in the U.S. using the notion of an "MMWR week" or "epi week" (43). MMWR weeks always begin on Sundays and end on Saturdays. Weeks can be numbered 1–53, and, as a result, many institutions choose to report them as such (e.g., the interval `[2016-05-15 00:00, 2016-05-22 00:00)` is reported as "2016, week 20"). However, while most U.S.-based health departments respect the weekly MMWR aggregation standard, many continue to report timestamps based on the MMWR week concept. For example, **Figure 5** shows how Texas identifies its weekly influenza surveillance PDF reports by trailing inclusive timestamps rather than by MMWR week.

Outside of the U.S., a variety of reporting date standards exist. In Japan, for example, the epi week starts on Monday and ends on Sunday (45). In Poland, the reporting is even more different. Poland reports influenza cases four "weeks" a month, regardless

TABLE 1 | Sample historical weekly epidemiological time series consisting of timestamps and case counts.

Timestamp	Cases
2014-08-07 00:00	2
2014-08-14 00:00	5
2014-08-21 00:00	4

TABLE 2 | Explicit transformation of **Table 1** into a *leading* interval series.

Interval start	Interval end	Cases
2014-08-07 00:00	2014-08-14 00:00	2
2014-08-14 00:00	2014-08-21 00:00	5
2014-08-21 00:00	2014-08-28 00:00	4

The interval start and end are inclusive and exclusive, respectively.

TABLE 3 | Explicit transformation of **Table 1** into a *trailing exclusive* interval series.

Interval start	Interval end	Cases
2014-07-31 00:00	2014-08-07 00:00	2
2014-08-07 00:00	2014-08-14 00:00	5
2014-08-14 00:00	2014-08-21 00:00	4

The interval start and end are inclusive and exclusive, respectively.

TABLE 4 | Explicit transformation of **Table 1** into a *trailing inclusive* interval series.

Interval start	Interval end	Cases
2014-08-01 00:00	2014-08-08 00:00	2
2014-08-08 00:00	2014-08-15 00:00	5
2014-08-15 00:00	2014-08-22 00:00	4

The interval start and end are inclusive and exclusive, respectively.

of the length of the month. As a result, the intervals between reports are not regular. For example, the four influenza reports in May 2016 are shown in **Table 5**.

One remaining concern related to date and time is time zone. With the increasing use of internet data streams in disease forecasting and surveillance, it is important to be able to precisely place reference epidemiological data since associated internet data streams might be timestamped down to the second. Many data sources fail to report a time zone, so local time is often assumed. An incorrect time zone may impact analysis of high resolution data. For example, norovirus data are sometimes provided hourly [e.g., (50, 51)], and time zone errors could have a potentially drastic negative effect on model results or analysis.

2.2.2.2. Geography

Political boundaries and names must be carefully managed. Subtle differences in names (e.g., Zurich vs. Zürich) may lead to incorrect results during an analysis. The ISO 3166 standard defines country and principle subdivision (e.g., state or province) names, but it does not handle finer-than-subdivision regions, such as counties, districts, or cities.

Moreover, political boundaries (and thus populations and demographics) change over time. For example, South Sudan's split from Sudan in 2011 decreased Sudan's population by more than ten million people and dramatically changed its political boundary. Computing the historical attack rate for a disease (e.g., influenza incidence per 100,000 people), for instance, must take into account these changes.

2.3. Data Reporting Challenges

Beyond interface and data format challenges, there are challenges that lie within the bureaucratic *reporting* process for an epidemiological institution. Modern disease surveillance systems rely on complex reporting hierarchies; raw data are initially captured at each provider, who then anonymizes and aggregates data as necessary before sending it to the next level in the hierarchy (perhaps a local or state public health department) (52). This hierarchy can have many levels. Even in many of the most developed regions of the world, much of this process continues to be done by hand, although the push to electronic medical records is gaining traction. As a result, most disease surveillance systems across the world experience reporting lags of at least one to two weeks.

This reporting lag can, in some cases, affect both an intuitive understanding of the situation as well as computational forecasting models. In an effort to combat surveillance system reporting lag, a number of attempts have been made to "fill in" the gaps using internet data [e.g., (13–15, 17)], but these studies require moderate to high levels of internet usage in the locales of interest, which are often not guaranteed.

Another issue is heterogeneous case definitions across jurisdictions. Many times, the case definitions used in epidemiological data are not clearly defined, and it is often difficult to navigate websites to identify the definitions. For example, many of the influenza surveillance systems in Europe use common, but not identical, case definitions (53). Contextual differences in case definitions for Ebola (54) could make interpreting data for the 2014 West African Ebola outbreak difficult.

One must also be concerned about language issues. Data are often provided in the native language of the region of the world in which they originate. For example, Thailand's Bureau of Epidemiology website (55) is natively displayed in Thai but also offers an English version (56). While online language translation services do exist (e.g., Google Translate[8]), these are not always reliable, and they cannot easily translate text in images (e.g., a website header comprised of images). To assist with language issues, formal disease- and epidemiological-focused ontologies [e.g., (57)] can help; translations, abbreviations, and alternate names can be encoded in an ontology to help automatically map different records to the same concept.

Furthermore, even within a single institution, there are often reporting nuances among diseases. For example, one context may be reported monthly, while another context may be reported

[8]https://translate.google.com/

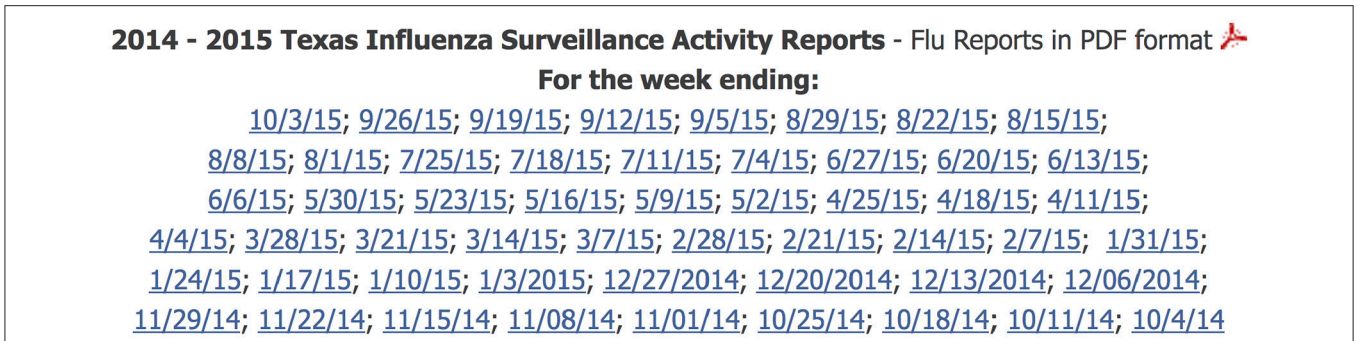

FIGURE 5 | Screenshot taken from Texas' Department of State Health Services 2014–2015 weekly influenza reports web page (44). Texas identifies its weekly influenza surveillance PDF reports by trailing inclusive timestamps rather than by MMWR week (e.g., "10/3/15" instead of "2015, week 39"). Interestingly, much of the data in each PDF uses MMWR week numbers rather than timestamps.

TABLE 5 | Influenza reporting intervals in Poland in May 2016.

Interval start	Interval end	Duration (days)	Source
2016-05-01	2016-05-07	7	(46)
2016-05-08	2016-05-15	8	(47)
2016-05-16	2016-05-22	7	(48)
2016-05-23	2016-05-31	9	(49)

Instead of reporting data at regular intervals (e.g., every 7 days), Poland reports data four "weeks" a month, regardless of the length of the month. This yields irregular interval durations. Here, the interval start and end are inclusive.

weekly or daily. Some contexts may not be regularly reported; irregular reporting can lead to questions like, "Is the value for a missing timestamp zero or unknown?"

Finally, case count data are often retroactively updated as new data are made available. In other words, historical data are not fixed the first time they are published. For example, a case count data point published today may be updated next week or the following week, as new data appear. This problem, often called "backfill," is due to the number and variety of members that comprise the complex reporting hierarchy that modern disease surveillance systems rely on; if a surveillance member's computer system goes down temporarily, for example, it may not be able to submit its data until the following week. Backfill can in some cases drastically affect analyses, so analysts and modelers must be aware of this potential issue (58, 59).

3. CONCLUSIONS

We have identified three key challenges involving epidemiological data: 1) interface challenges, 2) data format challenges, and 3) data reporting challenges. Each of these challenges can be addressed to simplify the efforts of analysts and modelers. Here, we propose a framework of best practices comprised of modern standards that should be adhered to when releasing epidemiological data to the public:

1. Present the user with an interactive web interface to search and filter data. This interface should allow users to export data in common open formats (e.g., JSON using the EpiJSON standard, CSV).
2. Provide a web-based API to allow automated data retrieval.
3. Always use ISO 8601 dates, times, timestamps, and durations. Timestamps should either explicitly provide the local time zone or be adjusted for UTC, as specified by the ISO 8601 standard.
4. When providing time series, clearly define the interval type so that timestamps can be interpreted properly.
5. When possible, use ISO 3166 location names.
6. Ensure all data are encoded using UTF-8.
7. Ensure website can be run through an online language translation service (e.g., do not place important text in an image).
8. When reporting case counts, the case definitions should be made explicit and clear.
9. Clearly distinguish between unknown and zero values.

These suggestions are not prescribing a single format or process; instead, these items provide a means for clearly defining and presenting epidemiological data to the public.

We implore the members of the global public health community to work together to create and follow standards for publishing data. Many institutions attempt to publish similar types of data using similar interfaces. In general, a user selects locations, diseases, optional time periods, and optional demographics in order to retrieve the desired data. Because many analysts and modelers have similar data desires, we feel this provides an opportunity for a generic shared epidemiological data access platform. Currently used by the CDC, one possibility might be Socrata[9], a platform that allows governments to share data openly. Socrata provides not only a modern interactive web interface but also an extensive API. Another option may be for public health institutions to collaboratively develop a free and open source solution that each could use. Such a platform may be more easily implemented by resource-constrained public health departments that have neither the time nor the money to develop their own solutions. Additionally, as web standards

[9]https://socrata.com/

evolve, a shared data access platform could be updated in order to propagate these changes to each institution.

If a standard platform could be employed by institutions worldwide, then one could envision a future where global data could be easily collected without the challenges we currently face. This would in turn streamline epidemiological and public health analysis, modeling, and informatics, resulting in better public health decision-making capabilities.

Additionally, while this paper focuses on the public health and epidemiological communities, many of the challenges and solutions discussed here are not unique to them. Many of these same challenges are present whenever data of the same type are published globally by separate institutions that do not have an a priori agreed-upon set of standards. For example, weather and economic data have many of the same features as epidemiological data (e.g., locations, time intervals) and should also adhere to the ISO 8601 and 3166 standards, be encoded in UTF-8, and clearly distinguish between unknown and zero values.

Finally, it is important to recognize that this paper focuses on capable public health institutions with enough funding to collect and disseminate their epidemiological data. It should be noted, however, that a number of regions worldwide do not meet this criterion and are struggling even to monitor and care for their constituent populations, let alone publish reliable data. Unfortunately, it is precisely in these underserved regions that the public health community often desires data. Until worldwide public health infrastructure improves significantly, the suggestions here will remain peripheral for many; thus, the problems and suggested solutions put forth in this paper are likely only to be relevant to well-funded institutions. While the solutions presented in this paper may not be as effective at present time due to the lack of coverage, we offer them in preparation for the expanding global coverage that is continuously occurring, and it is only a matter of time until 100% global internet coverage becomes a reality.

AUTHOR CONTRIBUTIONS

GF and BT compiled the initial manuscript. All authors identified and described challenges and standards, and provided critical revisions of the manuscript.

FUNDING

This work was supported by the Defense Threat Reduction Agency's Joint Science and Technology Office for Chemical and Biological Defense under project numbers CB3656 and CB10007.

ACKNOWLEDGMENTS

This work was supported by the U.S. Department of Energy through the Los Alamos National Laboratory. Los Alamos National Laboratory is operated by Triad National Security, LLC, for the National Nuclear Security Administration of U.S. Department of Energy (Contract No. 89233218NCA000001).

REFERENCES

1. Chretien JP, Rivers CM, Johansson MA. Make data sharing routine to prepare for public health emergencies. *PLoS Med*. (2016) 13:e1002109. doi: 10.1371/journal.pmed.1002109
2. U S Department of Health and Human Services, U S Department of Homeland Security. *Guidance on Allocating and Targeting Pandemic Influenza Vaccine* (2008). Available online at: https://asprtracie.hhs.gov/technical-resources/resource/2846/guidance-on-allocating-and-targeting-pandemic-influenza-vaccine
3. Nap RE, Andriessen MPHM, Meessen NEL, van der Werf TS. Pandemic influenza and hospital resources. *Emerg Infect Dis.* (2007) 13:1714–9. doi: 10.3201/eid1311.070103
4. Hota S, Fried E, Burry L, Stewart TE, Christian MD. Preparing your intensive care unit for the second wave of H1N1 and future surges. *Crit Care Med*. (2010) 38:e110–9. doi: 10.1097/CCM.0b013e3181c66940
5. Moran KR, Fairchild G, Generous N, Hickmann KS, Osthus D, Priedhorsky R, et al. Epidemic forecasting is Messier than weather forecasting: the role of human behavior and internet data streams in epidemic forecast. *J Infect Dis*. (2016) 214(Suppl. 4):S404–8. doi: 10.1093/infdis/jiw375
6. Hethcote HW. The Mathematics of infectious diseases. *SIAM Rev*. (2000) 42:599–653. doi: 10.1137/S0036144500371907
7. Eubank S, Guclu H, Kumar VSA, Marathe MV, Srinivasan A, Toroczkai Z, et al. Modelling disease outbreaks in realistic urban social networks. *Nature* (2004) 429:180–4. doi: 10.1038/nature02541
8. Bisset KR, Chen J, Feng X, Kumar VSA, Marathe MV. EpiFast: a fast algorithm for large scale realistic epidemic simulations on distributed memory systems. in *Proceedings of the 23rd International Conference on Supercomputing*. New York, NY: ACM Press (2009). p. 430–439.
9. Chao DL, Halstead SB, Halloran ME, Longini, Jr IM. Controlling dengue with vaccines in Thailand. *PLoS Negl Trop Dis.* (2012) 6:e1876. doi: 10.1371/journal.pntd.0001876
10. Grefenstette JJ, Brown ST, Rosenfeld R, DePasse J, Stone NT, Cooley PC, et al. FRED (A Framework for Reconstructing Epidemic Dynamics): an open-source software system for modeling infectious diseases and control strategies using census-based populations. *BMC Public Health* (2013) 13:940. doi: 10.1186/1471-2458-13-940
11. McMahon BH, Manore CA, Hyman JM, LaBute MX, Fair JM. Coupling vector-host dynamics with weather geography and mitigation measures to model Rift Valley fever in Africa. *Math Model Nat Phenom.* (2014) 9:161–77. doi: 10.1051/mmnp/20149211
12. Viboud C, Boëlle PY, Carrat F, Valleron AJ, Flahault A. Prediction of the spread of influenza epidemics by the method of analogues. *Am J Epidemiol.* (2003) 158:996–1006. doi: 10.1093/aje/kwg239
13. Polgreen PM, Chen Y, Pennock DM, Nelson FD. Using internet searches for influenza surveillance. *Clin Infect Dis.* (2008) 47:1443–8. doi: 10.1086/593098
14. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. *Nature* (2009) 457:1012–4. doi: 10.1038/nature07634
15. Signorini A, Segre AM, Polgreen PM. The use of Twitter to track levels of disease activity and public concern in the U.S. during the Influenza A H1N1 pandemic. *PLoS ONE* (2011) 6:e19467. doi: 10.1371/journal.pone.0019467
16. Shaman J, Karspeck A, Yang W, Tamerius J, Lipsitch M. Real-time influenza forecasts during the 2012–2013 season. *Nat Commun.* (2013) 4:1–10. doi: 10.1038/nmcomms3837
17. Generous N, Fairchild G, Deshpande A, Del Valle SY, Priedhorsky R. Global disease monitoring and forecasting with wikipedia. *PLoS Comput Biol.* (2014) 10:e1003892. doi: 10.1371/journal.pcbi.1003892.
18. Hickmann KS, Fairchild G, Priedhorsky R, Generous N, Hyman JM, Deshpande A, et al. Forecasting the 2013-2014 influenza

Season Using wikipedia. *PLoS Comput Biol.* (2015) 11:e1004239. doi: 10.1371/journal.pcbi.1004239
19. Fairchild G, De Silva L, Del Valle SY, Segre AM. Eliciting disease data from wikipedia articles. in *Ninth International AAAI Conference on Weblogs and Social Media - Wikipedia Workshop*. Oxford, UK (2015). p. 26–33.
20. STROBE Initiative. *STROBE Statement* (2018). Available online at: https://www.strobe-statement.org/ (Accessed October 01,2018).
21. CDISC. *CDISC | Strength Through Collaboration* (2018). Available online at: https://www.cdisc.org/ (Accessed October 01,2018).
22. Pisani E, AbouZahr C. Sharing health data: good intentions are not enough. *Bull. World Health Organ.* (2010) 88:462–6. doi: 10.2471/BLT.09.074393
23. Sane J, Edelstein M. *Overcoming Barriers to Data Sharing in Public Health: A Global Perspective.* London: Chatham House: The Royal Institute of International Affairs (2015). Available online at: https://www.chathamhouse.org/publication/overcoming-barriers-data-sharing-public-health-global-perspective
24. Zentgraf DC. *What Every Programmer Absolutely, Positively Needs to Know About Encodings and Character Sets to Work With Text* (2015). Available online at: http://kunststube.net/encoding/ (Accessed August 23, 2016).
25. Atwood J. *The Great Newline Schism* (2010). Available online at: https://blog.codinghorror.com/the-great-newline-schism/ (Accessed October 01, 2016).
26. Israeli Ministry of Health. *Epidemiological Reports Archive* (2016). Available online at: http://www.health.gov.il/UnitsOffice/HD/PH/epidemiology/Pages/epidemiology_report.aspx?WPID=WPQ7&PN=1 (Accessed September 04, 2016).
27. Australian Government Department of Health. *National Notifiable Disease Surveillance System* (2016). Available online at: http://www9.health.gov.au/cda/source/rpt_1_sel.cfm (Accessed September 04, 2016).
28. Robert Koch Institute. *SurvStat@RKI 2.0* (2016). Available online at: https://survstat.rki.de (Accessed September 04, 2016).
29. Coelho FC, Codeço CT, Cruz OG, Camargo S, Bliman PA. Epidemiological data accessibility in Brazil. *Lancet Infect Dis.* (2016) 16:524–5. doi: 10.1016/S1473-3099(16)30007-X
30. Federal Government of Brazil. *Epidemiological Situation* (2017). Available online at: http://www.combateaedes.saude.gov.br/en/epidemiological-situation (Accessed January 19, 2017).
31. Secretaria de Vigilância em Saúde – Ministério da Saúde. *Boletim Epidemiológico* (2017). Available online at: http://portalsaude.saude.gov.br/index.php/o-ministerio/principal/leia-mais-o-ministerio/197-secretaria-svs/11955-boletins-epidemiologicos-arquivos (Accessed January 19, 2017).
32. Apache OpenOffice. *Office Open XML* (2016). Available online at: https://wiki.openoffice.org/wiki/Office_Open_XML (Accessed August 18, 2016).
33. GeoJSON. *GeoJSON* (2016). Available online at: http://geojson.org/ (Accessed August 22, 2016).
34. TopoJSON. *TopoJSON* (2016). Available online at: https://github.com/topojson/topojson (Accessed August 22, 2016).
35. Finnie TJR, South A, Bento A, Sherrard-Smith E, Jombart T. EpiJSON: a unified data-format for epidemiology. *Epidemics* (2016) 15:20–6. doi: 10.1016/j.epidem.2015.12.002
36. Finnie T. *Repijson* (2015). Available online at: https://github.com/Hackout2/repijson (Accessed October 23, 2018).
37. Khusro S, Latif A, Ullah I. On methods and tools of table detection, extraction and annotation in PDF documents. *J Inform Sci.* (2015) 41:41–57. doi: 10.1177/0165551514551903
38. Republic of the Philippines Department of Health. *Statistics*. Department of Health (2016). Available online at: http://www.doh.gov.ph/statistics (Accessed December 13, 2016).
39. Republic of the Philippines Department of Health. *2015 Dengue Morbidity Week 26* (2016). Available online at: http://www.doh.gov.ph/sites/default/files/statistics/dengueMw26.compressed.pdf (Accessed December 13, 2016).
40. Republic of the Philippines Department of Health. *2016 Diphtheria Morbidity Week 12* (2016). Available online at: http://www.doh.gov.ph/sites/default/files/statistics/DIPHTHERIA%20MW12.pdf (Accessed December 13, 2016).
41. Davis M. *Unicode Over 60 Percent of the Web* (2016). Avialable online at: https://googleblog.blogspot.com/2012/02/unicode-over-60-percent-of-web.html (Accessed August 23, 2016).
42. W3Techs. *Usage of Character Encodings for Websites*. (2016). Available online at: https://w3techs.com/technologies/overview/character_encoding/all (Accessed August 23, 2016).
43. Centers for Disease Control and Prevention. *MMWR Weeks* (2016). Available online at: https://wwwn.cdc.gov/nndss/document/MMWR_week_overview.pdf (Accessed August 22, 2016).
44. Texas Department of State Health Services. *2014–2015 Texas Influenza Surveillance Activity Report* (2016). Available online at: http://www.dshs.texas.gov/idcu/disease/influenza/surveillance/2015/ (Accessed December 20, 2016).
45. National Institute of Infectious Diseases, Japan. *Weeks Ending Log* (2016). Available online at: http://www.nih.go.jp/niid/en/calendar-e.html (Accessed Decemeber 13, 2016).
46. National Institute of Public Health, Poland. *Influenza and influenza-like illness in Poland, 5A* (2016). Available online at: http://wwwold.pzh.gov.pl/oldpage/epimeld/grypa/2016/I_16_05B.pdf (Accessed Decemeber 13, 2016).
47. National Institute of Public Health, Poland. *Influenza and influenza-Like Illness in Poland, 5B* (2016). Available online at: http://wwwold.pzh.gov.pl/oldpage/epimeld/grypa/2016/I_16_05A.pdf (Accessed Decemeber 13, 2016).
48. National Institute of Public Health, Poland. *Influenza and influenza-like illness in Poland, 5C* (2016) Available online at: http://wwwold.pzh.gov.pl/oldpage/epimeld/grypa/2016/I_16_05C.pdf (Accessed Decemeber 13, 2016).
49. National Institute of Public Health, Poland. *Influenza and influenza-like illness in Poland, 5D* (2016). Available online at: http://wwwold.pzh.gov.pl/oldpage/epimeld/grypa/2016/I_16_05D.pdf (Accessed Decemeber 13, 2016).
50. Guzman-Herrador B, Heier BT, Osborg EJ, Nguyen VH, Vold L. Outbreak of norovirus infection in a hotel in Oslo, Norway, January 2011. *Euro Surveill.* (2011) 16:19928. Available online at: https://www.eurosurveillance.org/content/10.2807/ese.16.30.19928-en
51. Mayet A, Andréo V, Bédubourg G, Victorion S, Plantec JY, Soullié B, et al. Food-borne outbreak of norovirus infection in a French military parachuting unit, April 2011. *Euro Surveill.* (2011) 16:19930. Available online at: https://www.eurosurveillance.org/content/10.2807/ese.16.30.19930-en
52. Fairchild G. *Improving Disease Surveillance: Sentinel Surveillance Network Design and Novel Uses of Wikipedia*. Ph.D. Dissertation. University of Iowa (2014).
53. Aguilera JF, Paget WJ, Mosnier A, Heijnen ML, Uphoff H, van der Velden J, et al. Heterogeneous case definitions used for the surveillance of influenza in Europe. *Eur J Epidemiol.* (2003) 18:751–4. doi: 10.1023/A:1025337616327.
54. World Health Organization. *Case Definition Recommendations for Ebola or Marburg Virus Diseases* (2017). Available online at: http://www.who.int/csr/resources/publications/ebola/ebola-case-definition-contact-en.pdf (Accessed January 10, 2017).
55. Thailand Bureau of Epidemiology. *Thailand Bureau of Epidemiology* (2016). Available online at: http://203.157.15.110/boe/ (Accessed August 23, 2016).
56. Thailand Bureau of Epidemiology. *Thailand Bureau of Epidemiology* (2016). Available online at: http://203.157.15.110/boeeng/ (Accessed August 23, 2016).
57. Schriml L. *Disease Ontology - Institute for Genome Sciences @ University of Maryland* (2011). Available online at: http://disease-ontology.org/ (Accessed October 01, 2018).
58. Farrow DC, Brooks LC, Hyun S, Tibshirani RJ, Burke DS, Rosenfeld R. A human judgment approach to epidemiological forecasting. *PLoS Comput Biol.* (2017) 13:e1005248. doi: 10.1371/journal.pcbi.1005248
59. Osthus D, Gattiker J, Priedhorsky R, Del Valle SY. Dynamic Bayesian influenza forecasting in the United States with hierarchical discrepancy. *arXiv:1708.09481* arXiv pre-print (2017). doi: 10.1214/18-BA1117

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Fairchild, Tasseff, Khalsa, Generous, Daughton, Velappan, Priedhorsky and Deshpande. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.