Green River
167 Main Street, Suite 103
Brattleboro, VT 05301
802.257.0641 | www.greenriver.com

White Paper

Protecting privacy in the neighborhood-level release of health information: An algorithm for publishing localized health data in compliance with the Health Insurance Portability and Accountability Act of 1996

May 2022

Michael Knapp, Ph.D., M.E.M.
Chief Executive Officer, Green River Data Analysis, Inc.

Todd Blackman, M.S., B.S.
Software Engineer, Green River Data Analysis, Inc.

Matthew Muspratt, J.D., M.E.M.

Acknowledgements

Green River's work on the software described in this white paper would not have been possible without the support of the Delaware Department of Health and Social Services (DHSS), for whom Green River developed *My Healthy Community*, DHSS's online public health data portal.[1] Many of the descriptions, examples, and figures presented in this paper relate to or draw from *My Healthy Community*, which was the first application of this software to data made publicly available. The authors are grateful to their DHSS colleagues for their collaboration and especially thank state epidemiologist Dr. Tabatha Offutt-Powell and Division of Public Health Associate Deputy Director Cassandra Codes-Johnson, M.P.A. for their work spearheading the *My Healthy Community* project.

Abstract

Public health officials seek to provide citizens—and public sector colleagues—with localized data about disease and health (*e.g.* case rates, clusters, and trends). But publishing health data at sub-state levels such as county and ZIP code risks a precision that could expose individuals' confidential health information and violate the Health Insurance Portability and Accountability Act of 1996 (HIPAA). This white paper describes "HIPAA checker" software developed by Green River to detect and suppress non-HIPAA-compliant data at highly granular resolutions, including

[1] Delaware Department of Health and Social Services, *My Healthy Community*, https://myhealthycommunity.dhss.delaware.gov/locations/state.

as low as the Census block group level, a geographic division comprising as few as 600 people. Employing a HIPAA-satisfying (k, p)-anonymity test to detect problematic data, including computationally complex implied disclosures in the extensive stratifications and spatial and temporal resolutions that make granular data useful, our software is also capable of providing end users with a "next best option" if the indicator they request has been suppressed for the desired geography and time period. The software and its algorithm therefore offer a powerful solution for disclosing local health information while protecting confidentiality—examples of which are presented throughout the paper.

Introduction

One of the ongoing problems facing public health officials is how to communicate detailed, actionable information about disease and health while protecting the confidentiality of citizens. Community organizations, the media, and the general public benefit greatly from a highly localized understanding of case rates, clusters, trends, and vulnerabilities—the widespread dissemination of small area data and maps describing the COVID-19 pandemic provides a recent, high profile example—but information that is too precise about *what*, *where*, and *when* risks exposing *who*. In other words, it could expose an individual's confidential health information.

In the United States, most use and disclosure of health information must employ de-identification methods that meet a standard under the Health Insurance Portability and Accountability Act of 1996 (HIPAA).[2] While enforcement was relaxed for the COVID-19 pandemic,[3] HIPAA's Privacy Rule establishes the general, baseline conditions under which health data may be disclosed under non-pandemic circumstances.[4] Compliance with the Privacy Rule can be straightforward via a "safe harbor" provision that deems data sufficiently de-identified if 18 specific identifiers are removed (*e.g.* names, sub-state geopolitical subdivisions, birth dates, various phone, account, and record numbers, *etc*.), but this method frequently undermines the usefulness of the data by stripping it of the local precision needed to make the information actionable. Knowing the overall statewide rates of preterm births and neonatal abstinence syndrome, for example, is not nearly as useful for maternal health advocates as knowing in precisely which neighborhoods and at what ages and for which ethnicities the rates are highest.

The challenge for those publishing localized public health information, therefore, is meeting the Privacy Rule's alternative test, which considers "health information" to be "not individually identifiable… only if… the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information."[5] This means, for example,

---

[2] Per HIPAA, 45 CFR § 160.103 - Definitions, "health information" is "any information, including genetic information, whether oral or recorded in any form or medium, that: (1) Is created or received by a health care provider, health plan, public health authority, employer, life insurer, school or university, or health care clearinghouse; and (2) Relates to the past, present, or future physical or mental health or condition of an individual; the provision of health care to an individual; or the past, present, or future payment for the provision of health care to an individual."

[3] See, e.g., U.S. Department of Health and Human Services (HHS), Notification of Enforcement Discretion under HIPAA to Allow Uses and Disclosures of Protected Health Information by Business Associates for Public Health and Health Oversight Activities in Response to COVID-19, 85 Fed. Reg. 19,392 (Apr. 7, 2020), https://www.govinfo.gov/content/pkg/FR-2020-04-07/pdf/2020-07268.pdf.

[4] 45 C.F.R. § 164.514(b) *Implementation specifications: Requirements for de-identification of protected health information*.

[5] 45 C.F.R. § 164.514(b) *Implementation specifications: Requirements for de-identification of protected health information.*

publishing ZIP code-level or Census block-group level rates for a particular disease is entirely permissible provided no one can reverse engineer the data and determine the identity of the individuals who had the disease.

As it turns out, though, such reverse engineering is not only "quick" to expose individuals—Harvard University's Laranya Sweeney has shown that ZIP code, gender, and birthday is enough information to uniquely identify nearly 90 percent of Americans[6]—but the innumerable combinations in which datasets can be sliced (by age, sex, race/ethnicity, geography, time period, *etc*.) makes ascertaining the HIPAA compliance of a particular stratification at a particular spatial and temporal resolution a substantially complex mathematical problem. It may be useful to present the public with a longitudinal comparison of preterm birth rates among teenagers living in urban, suburban, and rural postal codes, but the existence of datasets of intersecting geographic division and/or time period (*e.g.*, postal code and school district; annual and five-year rates) and the arithmetical interplay of stratifications (*e.g.*, implied disclosure: a count for one subgroup can provide information about the count of another subgroup) greatly complicate the HIPAA compliance of publishing those rates, a question that requires sophisticated computational investigation to untangle.

Below, we—developers at Green River, an impact-focused software and analytics firm headquartered in Brattleboro, Vermont—describe a HIPAA protection algorithm designed to detect and suppress non-compliant data at sub-state (and state) level resolutions, including as low as the Census block group level, a small geographic division comprising as few as 600 people. Importantly, the software, which currently supports the Delaware Department of Health and Social Services's (DHSS) public-facing health data platform, *My Healthy Community*,[7] operates as more than a "red flag" warning signal. For any given problematic statistic, the software is also designed to present the "next best option" by seeking out an alternative "version" of the statistic that is HIPAA-compliant by virtue of either its longer temporal or broader geographic resolution. (As explained below, the process is akin—but not identical—to aggregating data across longer time periods or greater geographic areas until the statistic meets an HIPAA-acceptable threshold.) The software is therefore especially useful in contexts where, like Delaware, population totals are relatively small and geopolitical subdivisions few (Delaware is divided into only three counties), potentially obviating the need for HIPAA waivers where publication of granular public health informatics is desired. Indeed, *My Healthy Community* is the rare public platform capable of presenting COVID-19 information and numerous other health indicators in compliance with HIPAA's Privacy Rule at resolutions as low as the census block group level.

This white paper continues with a review of HIPAA's Privacy Rule parameters; a description of the statistical processes that Green River's algorithm employs to solve privacy complications and satisfy the Privacy Rule,

[6] Sweeney, L., "Simple Demographics Often Identify People Uniquely." Carnegie Mellon University, Data Privacy Working Paper 3, Pittsburgh 2000. https://dataprivacylab.org/projects/identifiability/paper1.pdf.

[7] Delaware Department of Health and Social Services, *My Healthy Community*, https://myhealthycommunity.dhss.delaware.gov/locations/state.

supplemented with examples of the algorithm in use on the *My Healthy Community* platform; and a discussion of the implications of such software for effective public health communication.

Meeting HIPAA's Privacy Rule with (*k*, *p*) anonymity

Section 45 C.F.R. § 164.514(b) of HIPAA outlines two means of determining whether "health information is not individually identifiable health information."[8] The "safe harbor" process requires removal of 18 direct identifiers, which as discussed above leaves little leeway for publishing localized rate information, in particular the proscription against "geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes."[9]

(For clarity and completeness, we note that, as with DHSS's *My Healthy Community* portal, the intended application of our software is to present frequencies, tables, charts, and other statistical summaries—not to present record-level data or any information displaying direct identifiers.)

The alternative approach, 45 C.F.R. § 164.514(b)(1), to which Green River adheres in its software, states that:

> …health information is not individually identifiable health information only if:
>
> > (1) A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable:
> >
> > > (i) Applying such principles and methods, determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and
> > >
> > > (ii) Documents the methods and results of the analysis that justify such determination;[10]

[8] 45 C.F.R. § 164.514(b) *Implementation specifications: Requirements for de-identification of protected health information*.

[9] The geographic limitation under HIPAA's "safe harbor" guidelines does allow exceptions for the initial three digits of a ZIP code if a population threshold of 20,000 is met. That, we contend, still falls short of the granularity that is often desirable in presenting population health data.

[10] 45 C.F.R. § 164.514(b) *Implementation specifications: Requirements for de-identification of protected health information.*

In other words, a dataset, report, chart, or similar presentation derived from individually identifiable health information[11]—such as those employed prolifically and granularly on DHSS's *My Healthy Community* platform—is deemed de-identified under HIPAA if there is a very small risk that a person whose information contributes to that presentation could be identified.

To meet this standard, Green River's software relies on the (k, p)-anonymity technique, which, in general and applied here, requires that an indirect identifier (*i.e.* a non-unique attribute such as sex or age) or any combination of indirect identifiers be common to at least k individuals in the population under examination ($k > 1$), except for at most a fraction of p of the population members ($p < 1$). Such a rule, usually with very small p, ensures multiple indistinguishable individuals are behind any presented rate.

In the case of the software underpinning *My Health Community* and its various frequency reports—rate tables, charts, trendlines, maps, *etc*.—our algorithm will suppress (*i.e.* not present) counts or rates when fewer than 11 individuals contribute. This satisfies (k, p) anonymity with $k = 11$, $p = 0$. In 2020, qualified statistical consultants retained by Green River assessed and certified the adequacy of our HIPAA processes, affirming that:

> [the (k, p)-anonymity technique establishes] a risk measure comparable to the risk involved in the HIPAA Safe Harbor Rule [and the] size limit of eleven (11) patients per cell has been used in similar contexts by government agencies such as the Centers for Medicare and Medicaid Studies. Similar methods of statistical disclosure limitation are discussed in Working Paper 22 which has been explicitly cited by the Department of Health and Human Services as a suitable reference for HIPAA compliance.[12]

Additionally, our software requires frequency reports to describe populations of at least 500 individuals, a "denominator rule" that at once provides a further anonymity safeguard and helps to uniformly stabilize rates. (We also establish a numerator count ceiling to protect anonymity in areas of high frequency. That is, reporting a 98 percent incidence rate (490/500) for a particular geography would effectively disclose that (nearly) all members have the attribute.)

Scanning datasets for thresholds of 11 and 500 is, of course, an undemanding software task, but as mentioned above, the nature of public health data—the extensive stratifications and spatial and temporal resolutions that make it

---

[11] HIPAA also employs the term "protected health information," which is defined at 45 CFR § 160.103 - Definitions and includes, with exceptions, "individually identifiable health information… that is (i) Transmitted by electronic media; (ii) Maintained in electronic media; or (iii) Transmitted or maintained in any other form or medium."

[12] Scheuren, F. and Baier, P. (2020), *HIPAA Certification for Green River Data Analysis, LLC*., citing U.S. Federal Committee on Statistical Methodology, Statistical Policy Working Paper 22 (second version 2005), Report on Statistical Disclosure Limitation Methodology (2005), https://www.hhs.gov/sites/default/files/spwp22.pdf. On file with the authors.

useful—greatly complicates the picture, rendering these thresholds on their own an insufficient test for HIPAA compliance. The additional analysis required to catch nuanced (but very real) disclosure risks comprises the "black box" of our software, and is discussed below.

Disclosure by subtraction

The inadequacy of cell count thresholds alone for assessing HIPAA compliance can be demonstrated with a simple example. Suppose a particular geography *D* is divided into three subregions, *a*, *b*, and *c* (as the state of Delaware is into three counties), with counts of a particular illness distributed as follows:

Region	Case count
a	15
b	15
c	10
D (=a+b+c)	40

Publishing a frequency report for macro region *D* would be permissible under our (*k*, *p*)-anonymity test (where $k = 11$) because $40 \geq 11$. The same would *seem* to be true for subregions *a* and *b* as well since their counts also exceed our cell threshold ($15 \geq 11$). But the failure of region *c* to satisfy the test ($10 < 11$) in fact necessitates the additional suppression of either subregion *a* or *b* because the publication of counts for both *a* and *b* alongside a known total count (*i.e.*, *D*'s count of 40) reveals *c*'s count is less than 11, a violation of our (*k*, *p*)-anonymity requirement.

Such disclosure by subtraction, or "implied disclosure," can occur along any dimension, not just spatially. Our software, for instance, would suppress sex-stratified cancer rates for a given geography and time period if the male rate were based on a count of less than 11 cases, no matter the female count—subtracting stratifications from overall values can potentially reveal counts (or populations) that are too small. Thus, similarly, we cannot report that among a total of 15 cases, 12 concern patients older than 18 years of age, because doing so implicitly discloses that 3 cases concern minors, even though our report may not explicitly feature an "under 18" age group.

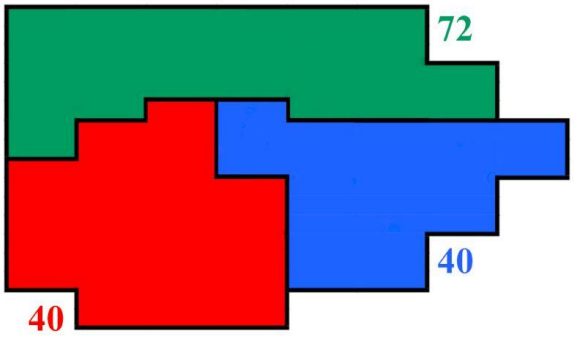
As we will discuss below, our algorithm handles such disclosures by, first, suppressing the offending rate and then, second, presenting a permissible rate comprised of counts and/or populations representing longer periods of time or larger geographic units. Before discussing the algorithmic details of this process, however—and highlighting

examples from *My Healthy Community*—we first note other types of inadvertent disclosure that our software must guard against.

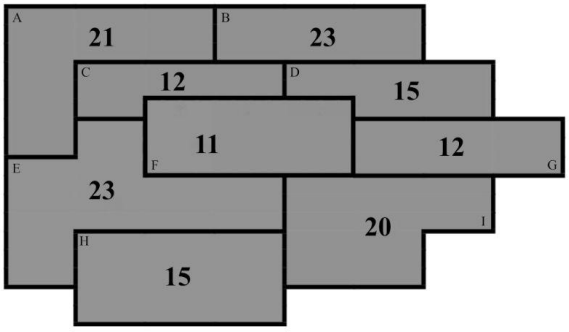
The sliver problem

Smaller geographies do not always fit perfectly into larger geographies. While our example above imagines the tidy subdivision of a state into three distinct, non-intersecting counties, the geopolitical spatial extents of importance to public health may not subsume each other so conveniently. ZIP codes cross county lines (and even, in a few cases, state lines); Census block groups may not coincide with ZIP code boundaries; cities can span multiple counties; and so on. The existence of these overlapping and hierarchically "incompatible" geographies greatly complicates the assessment of disclosure risks because they generate "slivers" of areas where counts and/or populations fall short of our (k, p)-anonymity parameters.

As an example, consider the intersections of a region divided into three Census block groups—*r*, *g*, and *b*, shaded red, green, and blue, respectively—that encompass nine relatively smaller neighborhoods, including one that happens to span all three block groups (Neighborhood F, outlined with a dashed border in the third diagram below).[13] ("Neighborhood," a granular division attractive to public health advocates due to its granularity and the public's familiarity with its extent, is defined on *My Healthy Community* according to data available from Zillow.) The diagrams and table below indicate hypothetical case counts in each of these subdivisions:

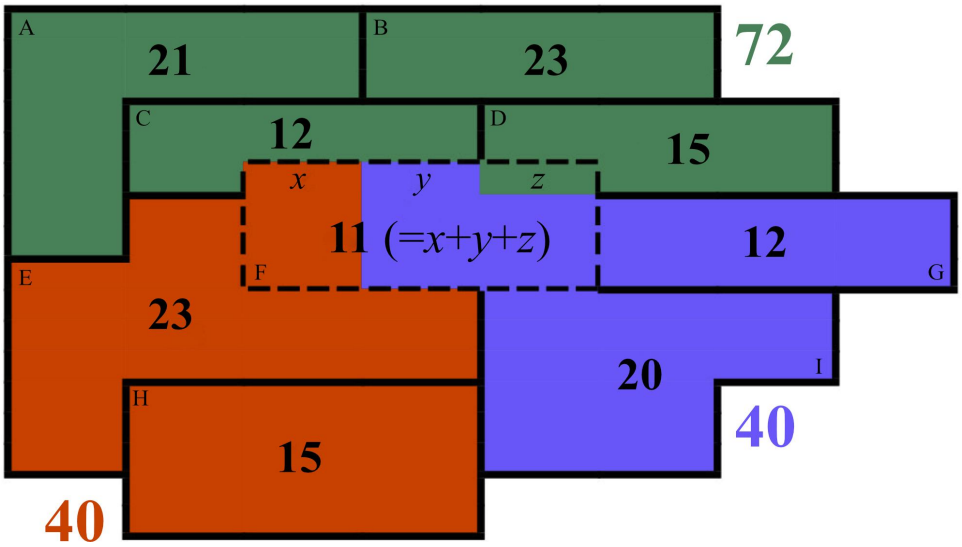


Case counts for Census block groups *r*, *g*, and *b*



Case counts for Neighborhoods A through I

[13] This example and its accompanying diagrams borrow from and lightly modify material in Scheuren, F. and Baier, P. (2020), *HIPAA Certification for Green River Data Analysis, LLC.*, on file with the authors.



Case counts by Census block group and Neighborhood

Region	Case count
Census block group *g*	72
Census block group *r*	40
Census block group *b*	40
Total	**152**
Neighborhood A	21
Neighborhood B	23
Neighborhood C	12
Neighborhood D	15
Neighborhood E	23
Neighborhood F	11
Neighborhood G	12
Neighborhood H	15
Neighborhood I	20
Total	**152**

All Census block groups and all neighborhoods satisfy our (k, p)-anonymity threshold of 11 cases, but we nonetheless are prohibited from publishing these figures. The reason is slivers x, y, and z, which together comprise neighborhood F but fall in different Census block groups. While neighborhood F itself contains an acceptable count of 11 cases, the intersecting nature of blocks groups and neighborhoods provides means to calculate counts for x, y, and z:

For g, $72 = 21 + 23 + 12 + 15 + z$, and therefore $z = 1$ case
For r, $40 = 23 + 15 + x$, and therefore $x = 2$ cases
For b, $40 = 20 + 12 + y$, and therefore $y = 8$ cases

In short, by publishing the full set of seemingly acceptable counts, we create three stratifications of neighborhood F and inadvertently, or implicitly, "publish" their case counts (1, 2, and 8)—which are all inadmissible under our (k, p)-anonymity test.

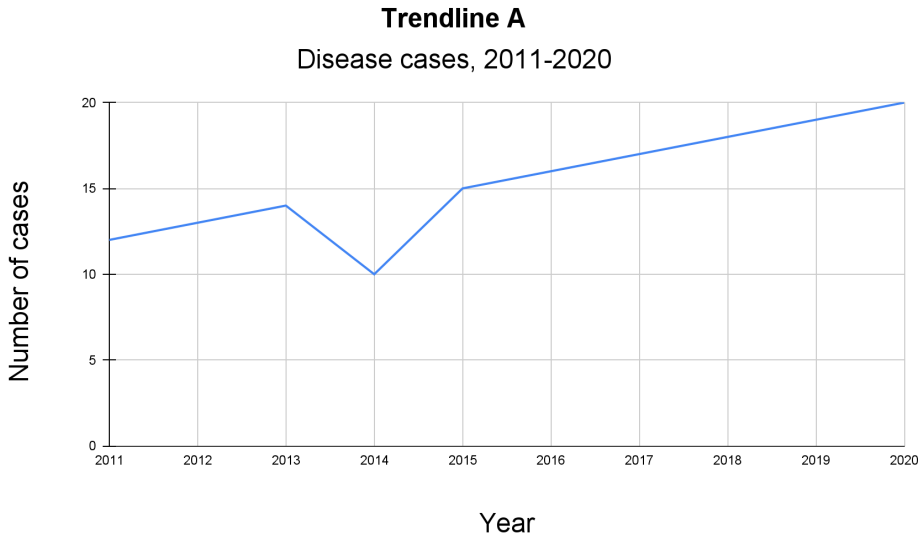
As with the disclosures by subtraction discussed above, the sliver problem requires intensive computational investigation, especially with the breadth of geopolitical divisions, time periods, stratifications, and indicators that public health advocates need at their disposal. Importantly—and to borrow from the above example—one group of advocates may desire to publish a particular case rate at Census block group level, while another group may have good reason to publish the same rate at neighborhood granularity. A robust informatics tool enables this to the greatest extent possible, minimizing suppression and presenting a "next best option" only when necessary. We will turn to our algorithm's "next best option" process after reviewing one more example of inadvertent disclosure.

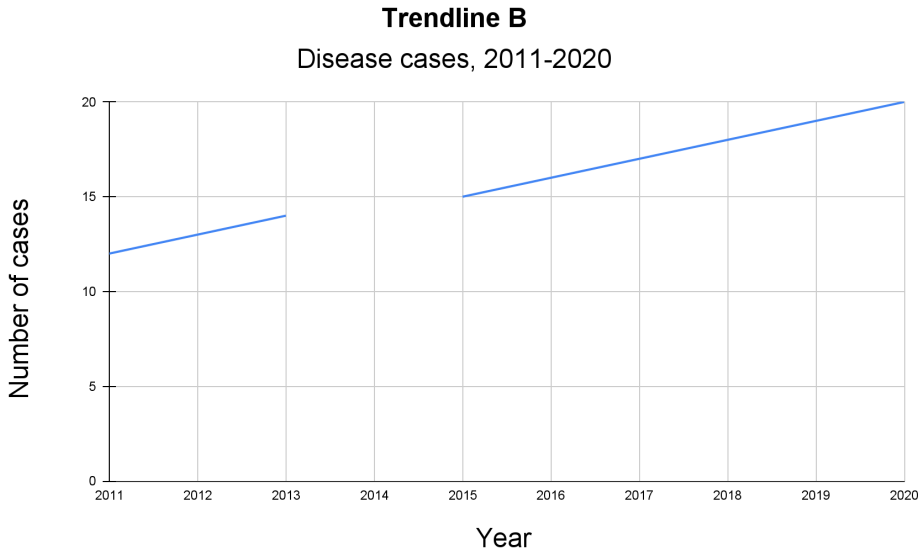
Trendlines and "problem years"

Trendlines are an effective data visualization for public health professionals and the general public alike. They provide, at just a quick glance, a good sense of whether an indicator over time is increasing or decreasing, stable or fluctuating. Once again, however, a brief example demonstrates scenarios when seemingly HIPAA-safe case counts must be suppressed—another "red flag" that our software is designed to detect and suppress.

Consider the following hypothetical data and two trendline depictions for prevalence of a disease in a particular region:

Year	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**
Cases	12	13	14	10	15	16	17	18	19	20





The "problem year" is 2014, when case counts dip to 10, an impermissibly low total under our (k, p)-anonymity test over an otherwise upward trending decade. Presenting a chart like Trendline A, therefore, is clearly not an option for a public-facing public health portal like *My Healthy Community*. But neither is Trendline B. Even though Trendline B omits the data of offending year 2014, the omission itself signals that a suppression issue is at play. (Recall that HIPAA's Privacy Rule guards against the "combination" of health data "with other reasonably available information," which could be expected to include supporting notes such as suppression thresholds.)

In such circumstances and in this example, therefore, our software would determine that not only is the presentation of cases for the year 2014 a Privacy Rule violation, but so is the presentation of a 10-year trendline (or data table)

spanning 2011 to 2020. For a given geography, when one time period in a sequence of such time periods must be suppressed, our software suppresses all statistics for the variable in question at the same temporal and spatial resolution.

Delivering HIPAA-compliant data to end users

While Green River's "HIPAA checker" software can serve as a standalone system for verifying the HIPAA-compliance of any given statistic (*i.e.*, by detecting (k, p)-anonymity test violations as outlined above), thereby enabling the presentation of granular, sub-state level data, it offers additional utility by identifying a "next best option" when a requested statistic is suppressed.

In practice, therefore, in the context of a public-facing portal such as DHSS's *My Healthy Community*, the application of our software involves:

1. deciding the universe of potential statistics to present at all temporal and geographic resolutions (*e.g.*, monthly, annual, and five-year rates for n number of indicators, including stratifications, at state, county, ZIP code, Census tract, and other geographic divisions);
2. running the Green River "HIPAA checker" software against those statistics to redact non-compliant data;
3. publishing the resulting HIPAA-compliant statistics such that the end user may search and investigate the vetted data (*e.g.* on a platform like *My Healthy Community*);
4. for any statistic requested by the end user, presenting that statistic or, if it has been suppressed during step 2, presenting that statistic at an incrementally longer temporal and/or incrementally broader geographic resolution than requested (and which is HIPAA-compliant).

The algorithm governing step 4 is, in essence, a search function, since—due to the suppressions in step 2—the final set of statistics available to the end user have all been confirmed HIPAA-compliant. For example, if the annual rate of preterm births in a particular ZIP code has been suppressed for failing our (k, p)-anonymity test, the available HIPAA-compliant set might nonetheless include *five-year* preterm birth rates for that ZIP code and/or annual preterm birth rates for *the county* encompassing the ZIP code, among other resolutions. Our algorithm is designed to return the best of these alternatives when an end user requests the offending statistic.

Given the primacy of location in the public health work our software supports, the algorithm prioritizes an end user's geographic selection in its search for alternatives. That is, it first holds the end user's desired geography unchanged while iteratively testing the availability of the statistic at the next shortest time interval. If rates at all temporal resolutions for a selected geography have been suppressed, the algorithm returns its search to the requested time interval, but this time tests the availability of the statistic at the next smallest geographic division encompassing

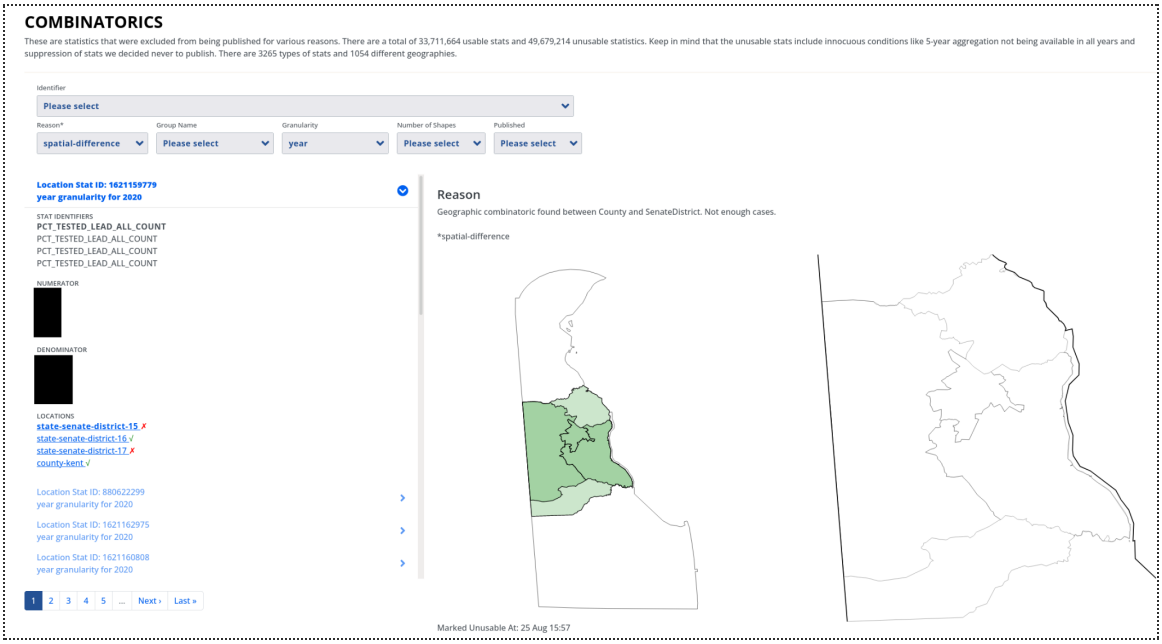
the desired geography. Through this hill climbing technique, it arrives at a "next best option," which is as close as possible in geographic and temporal scope as the suppressed statistic originally requested by the end user.

It is worth noting that the algorithm described above resembles in outcome, if not process, aggregation techniques frequently employed to present sparse or otherwise constrained data in public health and other fields. Though our software is in fact used to load portals like *My Healthy Community* with only HIPAA-compliant data—and then algorithmically present the statistic closest in geography and time period to the one requested—from an end user's perspective our software appears to be aggregating data over time and/or space when confronted with a suppressed indicator. Viewed this way, upon detecting problematic data at a given temporal resolution in a given geography, our software *appears* to, first, aggregate counts across a longer time period (*e.g.* per year instead of per month), holding the geography unchanged, and, second, if the data still fail to satisfy the Privacy Rule, aggregate counts over a larger geographic area (*e.g.* county instead of ZIP code).

Also worth noting is that a third method of aggregation would be available for some indicators upon "maxing out" temporal and geographic extents: aggregation by attribute (*e.g.*, presenting counts or rates for "all skin cancer types" instead of only "basal cell carcinoma" or only "melanoma"). In the current application of our software, this possibility is accommodated in step 1 above, when software engineers (and other project participants) decide on statistics and stratifications. (*i.e.* It is not an automated process in the software.)

Software interface and end user experience

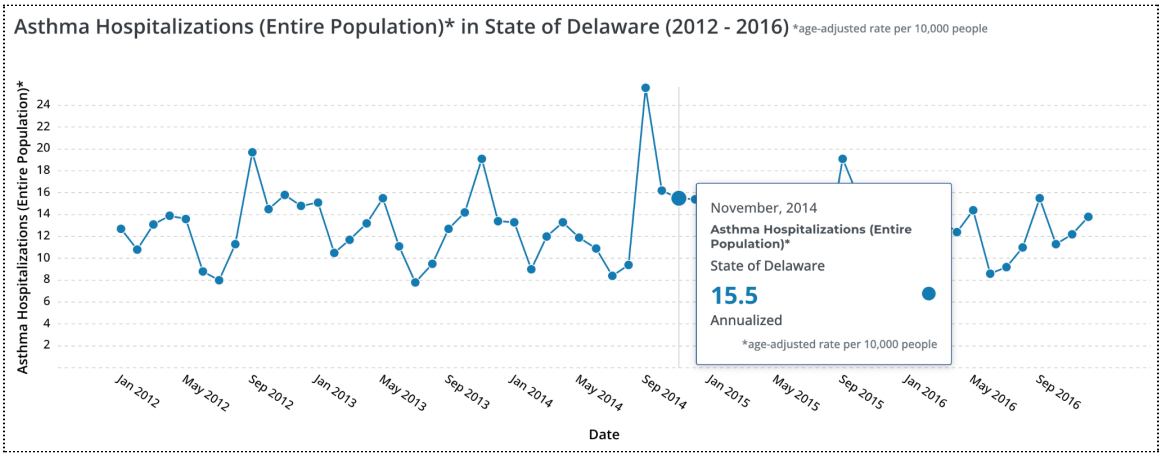
Our software is experienced in two ways by users: the "HIPAA checker" interface (employed by users preparing public health data for presentation) and outputs of end user requests for statistics (*e.g.* a portal like *My Healthy Community*). An example of the former, the software's interface as encountered during step 2 above (*i.e.* the results of the software's computations for detecting non-HIPAA-compliant data), is below:



Software results detecting a disclosure problem in lead testing data for Delaware

Turning to the end user's experience, and using the *My Healthy Community* portal for examples, we can examine the public-facing results of the software's work in a presentation of asthma hospitalization rates for various Delaware geographies.[14]

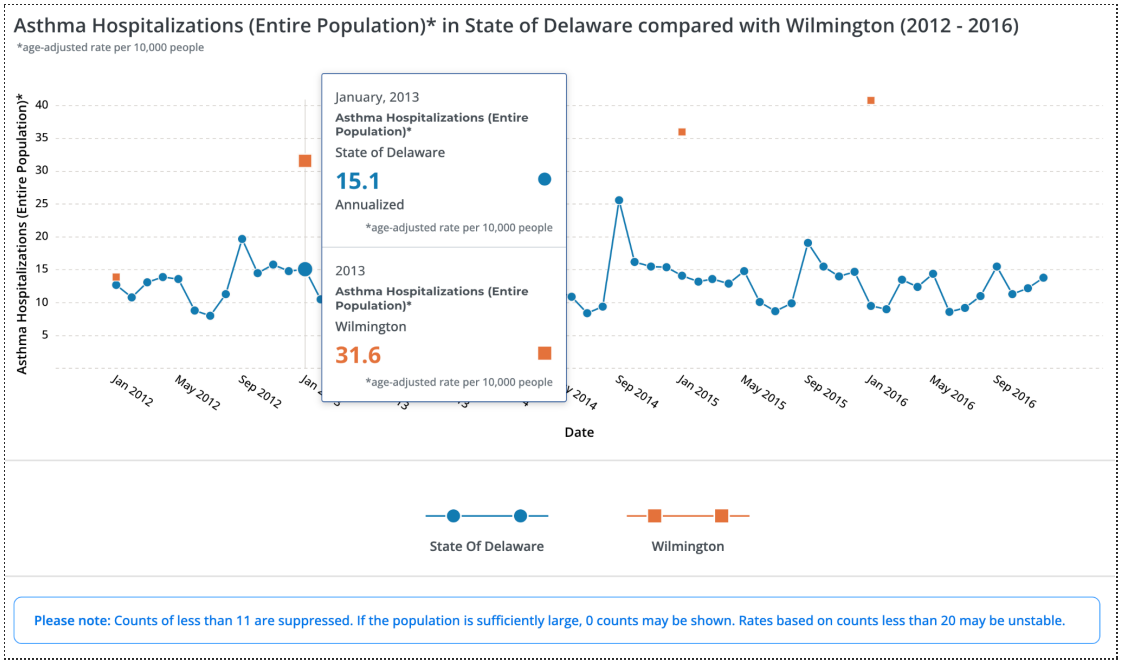
To start, below is the platform's rendering of data for the entire state from 2012-2016, which the software determined is HIPAA-compatible at a monthly temporal resolution:



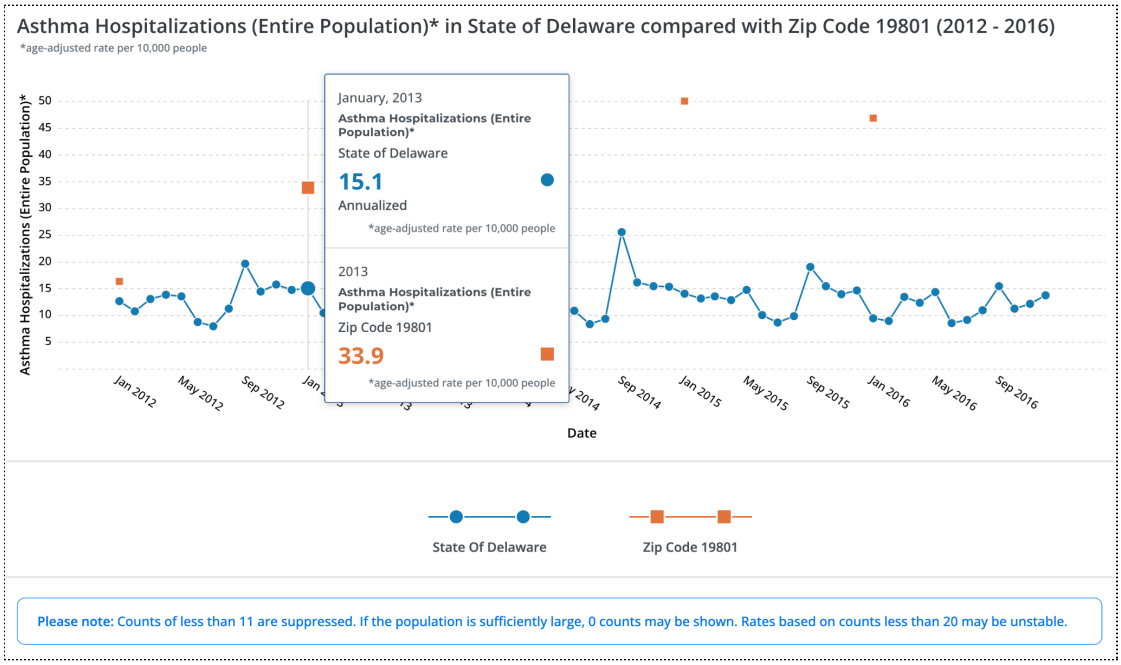
Asthma hospitalization rates, Delaware, monthly, 2012-2016

[14] Delaware Department of Health and Social Services, *My Healthy Community*, https://myhealthycommunity.dhss.delaware.gov/locations/state.

My Healthy Community can also display data for the city of Wilmington—Delaware's largest—as well as ZIP code 19801, whose spatial extent includes a portion of the city. In both cases, however, the software detected a HIPAA compliance problem. A request for asthma hospitalizations rates at these geographic granularities, therefore, returned annual instead of monthly figures. Below, the platform reports a January 2013 rate for Delaware but only a 2013 (annual) rate for Wilmington and postal code 19801:

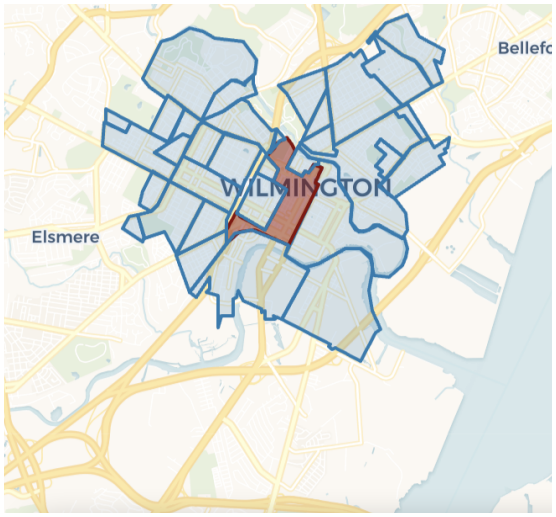


Asthma hospitalization rates, Delaware (monthly) and Wilmington (annual), 2012-2016

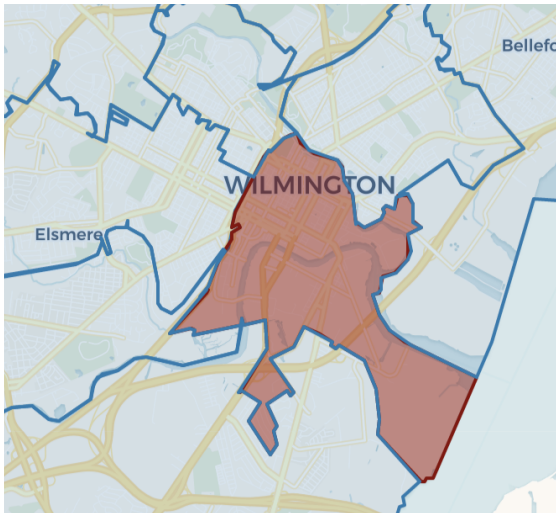


Asthma hospitalization rates, Delaware (monthly) and ZIP code 19801 (annual), 2012-2016

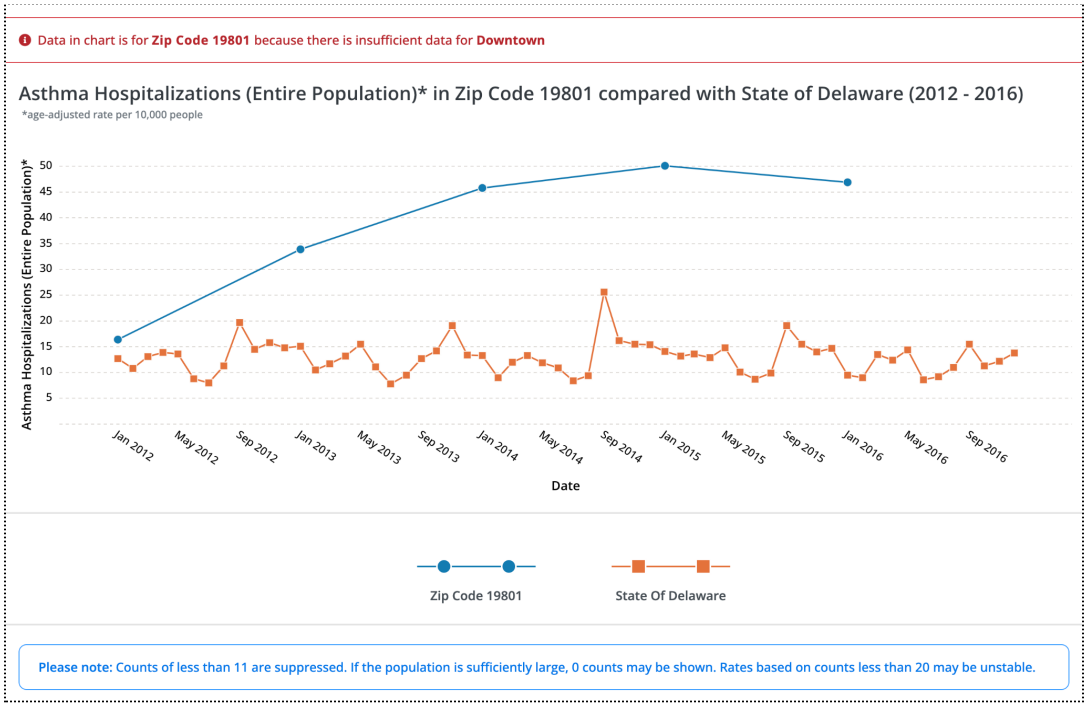
Investigating asthma hospitalization rates even more granularly reveals our algorithm's second approach towards the "next best option." A *My Healthy Community* search for rates within the Wilmington neighborhood known as Downtown in fact yields the ZIP 19801 data along with a red-lettered banner that Downtown's data is "insufficient," which informs the user that a search for the next best geographic division for the statistic has occurred. Note the spatial extents of the geographies and the resulting chart:



"Downtown" neighborhood of Wilmington



ZIP code 19801

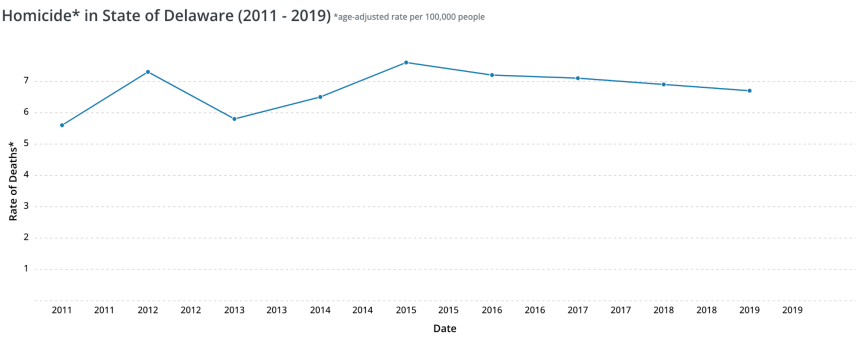


Asthma hospitalization rates, Delaware (monthly) and ZIP code 19801 (annual), 2012-2016, upon portal search for state- and neighborhood-level data

(Again, note that an end user not privy to how our software is applied to *My Healthy Community* might experience the above results as aggregations. A search for asthma hospitalization rates in Wilmington and ZIP code 19801 yielded data "aggregated" to an annual figure, instead of the monthly rates available for Delaware as a whole. And a search for asthma hospitalization rates in the Downtown neighborhood yielded data "aggregated" to an annual figure and a rate for the broader, "aggregated" geography of postal code 19801.)

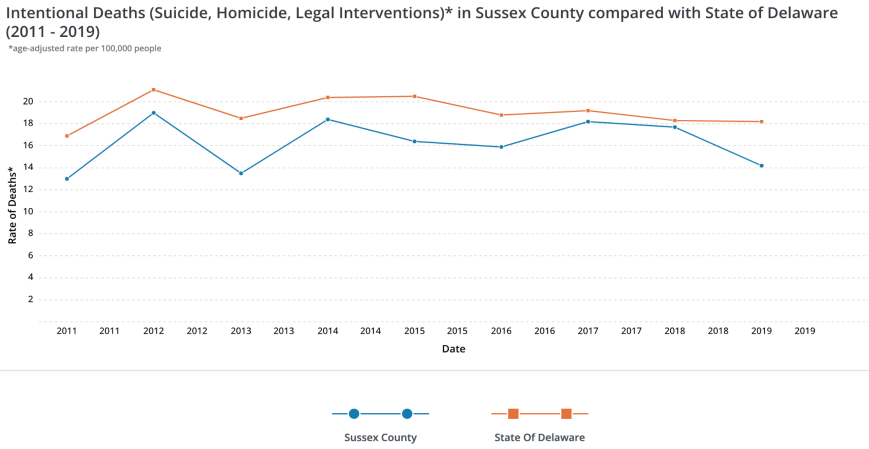
The complexity with which the algorithm seeks optimal geographic alternatives is evidenced by the diversity of "higher" geographies at its disposal when related indicators are investigated. For instance, the charts below all report that the small, southwestern Delaware town of Laurel (2017 population: 4,147) experienced insufficient homicides, intentional deaths, and accidental deaths to display annual rates. For each indicator, however, the algorithm has identified a different geographic subdivision as the next best option (*i.e.* state, county, and state senate district):

Data in chart is for **State of Delaware** because there is insufficient data for **Laurel**



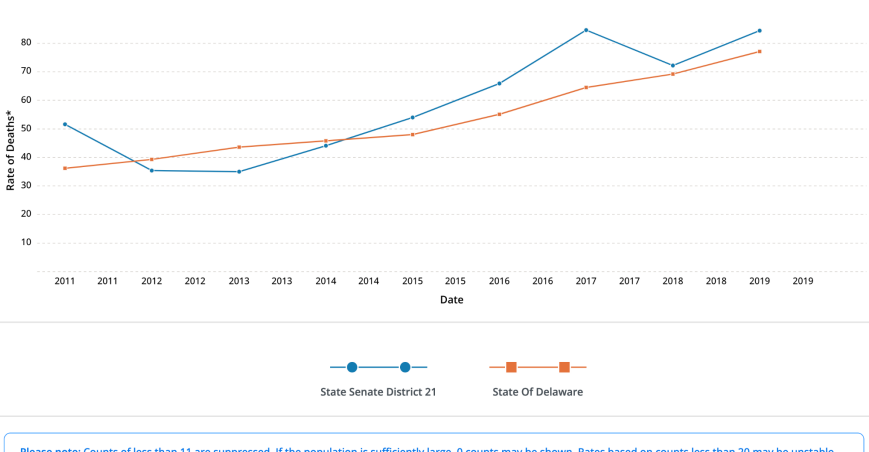
Please note: Counts of less than 11 are suppressed. If the population is sufficiently large, 0 counts may be shown. Rates based on counts less than 20 may be unstable.

Data in chart is for **Sussex County** because there is insufficient data for **Laurel**



Please note: Counts of less than 11 are suppressed. If the population is sufficiently large, 0 counts may be shown. Rates based on counts less than 20 may be unstable.

Data in chart is for **State Senate District 21** because there is insufficient data for **Laurel**



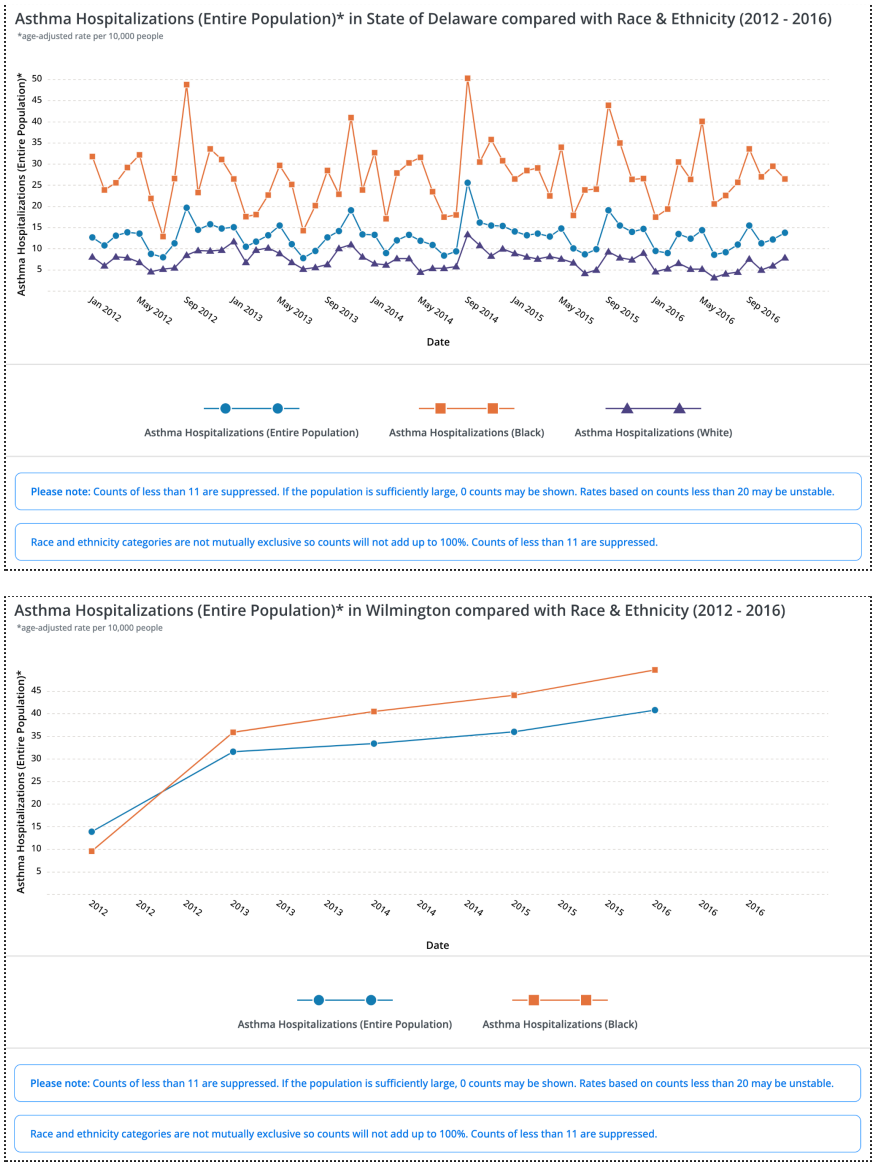
Please note: Counts of less than 11 are suppressed. If the population is sufficiently large, 0 counts may be shown. Rates based on counts less than 20 may be unstable.

Examples of aggregation to various geographies

Further examples of suppression in practice

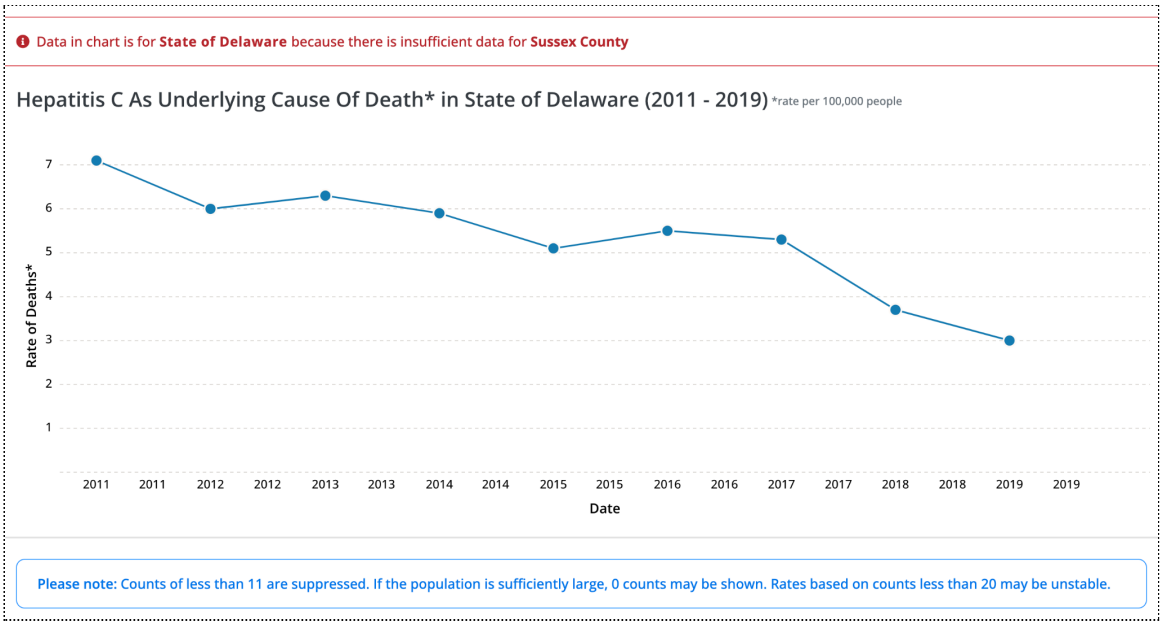
As alluded to earlier, one of the benefits of Green River's software is the flexibility it affords investigation into disease—it ensures that the "baby is not thrown out with the bathwater," that rates of a particular resolution are not wholesale suppressed because of a confidentiality concern at a particular stratification or resolution.

For instance, returning to asthma hospitalization rates, the software permits investigation into race and ethnicity stratifications at the Delaware-wide level even though it suppresses one variable (white) when displaying the same rates at a city granularity (Wilmington).



Suppression of one variable ("Asthma Hospitalizations (White)") upon changing indicator search geography from Delaware to Wilmington

Though not readily apparent on its face, an example of a trendline "problem year" occurs when investigating deaths due to the hepatitis C virus at the county level. A search for Sussex County data instead yields a trendline for the entire state of Delaware accompanied by the red-lettered warning of insufficient data:



Replacement of suppressed county-level county-level trendline with state-level data

Additionally, apart from the numerous statistics available via tables, charts, and the like on DHSS's *My Healthy Community*, the principles described herein are also applied to the portal's animated heat map visualizations. For further details on that application, please see our Green River white paper on the topic, *Better Maps: A methodology for creating granular—yet anonymized—topologies of disease*, available at https://www.greenriver.com/blog/posts/56.

Discussion

As the public has learned over the course of the COVID-19 pandemic experience, *local* health information is *useful* health information. Citizens and organizations need granular information in order to act locally towards improved health outcomes. Given the (necessary and desirable) privacy guardrails of the HIPAA Privacy Rule, we require a way for health departments to disclose local health information while protecting confidentiality. The "HIPAA checker" algorithm developed by Green River in collaboration with DHSS is valuable precisely for public health communication.

Tangential to the sheer informative benefits of local health data are several other advantages. For one, the ability to publish HIPAA-compliant information according to a multitude of geographic extents aligns with how decisions are

made. Governors might need—and wish to present—state-level data; county officials will want county-level information; local legislators, from state congresspersons to mayors, want to know and talk about yet other slices, from ZIP code to Census block group, from school district to neighborhood, across cities and towns (as do nonprofits and community organizations). To present information at multiple non-hierarchical and "incompatible" geopolitical divisions, software that can detect and suppress non-HIPAA-compliant data is required.

On the statistical front, the capability to present data locally and independent of geopolitical boundaries—or at least at multiple, varied, and overlapping geographies—helps guard against the modifiable areal unit problem (MAUP). The MAUP, briefly, describes a bias in data calculated and presented according to geopolitical unit—patterns and trends appearing in such data are in actuality determined by the geopolitical unit chosen for data aggregation (*e.g.* ZIP code, Census tract, county, *etc*.) because those geopolitical boundary lines are dictating the groupings of reported events (rate numerators) and populations (rate denominators).[15] Multiple granular lenses into rates—as our software enables—helps expose MAUPs by offering several slices of the same data and less "room" for MAUP error.

Finally, Green River's software is an answer to the dilemma presented at the outset of this white paper—it represents a resolution to the tension between confidentiality and disclosure of information vital to the public. Public health advocates must work with and communicate specific details of disease, but in the name of protecting individuals' personal health information there must be a tradeoff, and a tradeoff that is optimized. Ideally, the public is privy to no more health information than is necessary—and no less—and "confidentiality" serves as neither an excuse for inaction nor a diversion from obligation. The HIPAA checker algorithm achieves an optimum balance by disclosing the most specific information possible while still protecting confidentiality.

[15] For further discussion of the modifiable areal unit problem and how Green River's public health heat map tool evades the bias, see our white paper: Knapp, M., Blackman, T., and Muspratt, M. (2021), *Better Maps: A methodology for creating granular—yet anonymized—topologies of disease*, https://www.greenriver.com/blog/posts/56.