CITY OF MILWAUKEE
HEALTH DEPARTMENT

JUSTIFICATION REPORT: De-identification Methods for Milwaukee's COVID-19 Data Set

OVERVIEW

This report aims to discuss the importance of de-identification methods in the creation of a public COVID-19 data set for the city of Milwaukee. As the data set contains sensitive information about confirmed infections and the vaccination status of individuals, protecting privacy and ensuring data security is essential. To protect the information of community members in our data set, identifying and utilizing de-identification methods plays a crucial role in reducing disclosure risk of protected health information.

PROBLEM AND POSSIBLE SOLUTIONS

Data collected from database systems like the Wisconsin Electronic Surveillance System (WEDSS) and Wisconsin Immunization Registry (WIR) contain identifiable information such as names, addresses, and other health data. Publicly releasing this data set can therefore present a risk to individual privacy. To address this, de-identification methods must be used to minimize the possibility of re-identification and protect sensitive information. Two common methods for de-identification are expert determination and the "safe harbor" method, both of which aim to meet de-identification and HIPAA standards to protect individuals' rights.

Expert determination involves a qualified expert, with expertise in protected health information or security, to make informed decisions regarding the de-identification process. This method allows for tailored approaches specific to the project's needs, ensuring transparency while maintaining the privacy regulations. On the other hand, the "safe harbor" method follows determined guidelines for removing the specific identifiers, which can reduce the chances of errors, but limits flexibility.

CONSIDERATIONS AND OUR SOLUTION

The expert determination method has greater flexibility in decisions of publishing data while ensuring HIPAA and privacy compliance, but this can be a complex process. De-identification and risk assessment requires thorough documentation and still involves some uncertainty. On the other hand, the "safe harbor" method provides standardized guidelines which can minimize errors but lacks with reduced flexibility.

For the public COVID-19 data set in Milwaukee, we have chosen the expert determination method. This approach allows us to specify the de-identification process to meet standard requirements and disclose data we deem essential for informing future analysis. With the expertise of multiple qualified experts at the Milwaukee Health Department, this can enhance the accuracy of the de-identification process, ensuring maximum protection of the individual and their private information.

SUPPRESSION AND MASKING

Certain information that must not be made public can simply be omitted. Certain identifiers with personal detail including names, address, date of birth, client or patient identification numbers, and date of death will be removed completely. However, it's worth noting that one of our purposes is to connect individual cases with their vaccination status. We need to retain certain identifiers to link the two, which will be first name, last name, and date of birth. This was temporary and only remained for internal use during creation and is also eliminated after coding was completed, which occurs prior to the data set publishment. Finally, randomized identification numbers were produced for every confirmed case and client.

GENERALIZATION AND AGGREGATION

Another technique, known as generalization, is the process of organizing the data in an abstract view that introduces ambiguity and reduces risk of identification. This often includes grouping specified identifiers. To show the data in specific geographical locations, which may be either by census tract or ZIP code, an expert input and final decision must be made. Census tract data may be more useful in showing detail, however, this can be problematic. Because using census tract data largely depends on population densities, which vary widely across Milwaukee census tracts, ZIP code is preferred and is used in this public data set despite losing some specificity.

Other individual characteristics can present prominent identifiers and must be grouped accordingly while also maintaining valuable detail. To allow for ease in analysis, we remained consistent with the American Community Survey (ACS) age estimates for the following groups:

Under 5 years	**20 - 24 years**	**55 - 64 years**
5 - 9 years	**25 - 34 years**	**65 - 74 years**
10 - 14 years	**35 - 44 years**	**75 - 84 years**
15 - 19 years	**45 - 54 years**	**85 years +**

Gender is categorized based on sex of the individual, which includes male or female, ethnicity identifies if individual is of Hispanic or Latino descent, and race is grouped into:

American Indian / Alaska Native

Asian / Pacific Islander

Black / African American

White

Multiple Races

Other

Lastly, specimen and vaccination collection dates are grouped into weekly reports. Weekly intervals provide more detail tracking in trends of overall infection and changes in virus strain waves.

LIMITATIONS

The Wisconsin Electronic Disease Surveillance System (WEDSS) and Wisconsin Immunization Registry (WIR) contain all data necessary for valuable analysis. However, the COVID-19 pandemic proved a challenging time, and many errors are included in these databases. Especially in the early phases of the pandemic, standard procedures were not consistent. Firstly, WIR demographic characteristics were not required to be documented and multiple race categories could not be retained in the system for individual records which resulted in imperfect or incomplete information. Additionally, many individuals who performed data collection were not properly trained, flooded with patients, often fatigued, and may not have taken much detail during documentation. This lack of consistence over time also applies to other data where, for example, vaccine doses were reported as ordinal positions in vaccine types. Their position meanings have changed over time as each vaccine type has had changes in vaccine recommendations within their respective series and/or boosters. To gain that consistency, the vaccine doses for individuals are only reported as the number of vaccines they have received overall.

CONCLUSION

Using the expert determination method, various techniques can be applied, including suppression and generalization. Identifiers such as names and addresses were removed (the suppression technique), while critical data like age were grouped to track infections and vaccinations effectively (generalization). By doing so, individuals within the data set cannot be identified or linked in combination with other data sets to possibly identify them. Protecting privacy in data collection and sharing is highly important. By using the expert determination method and included techniques, we can effectively provide security assurance for the public COVID-19 data set of Milwaukee. It ensures compliance with privacy regulations and reduces the risk of compromising protected health information. Ultimately, the city of Milwaukee can confidently utilize this de-identified data for public health initiatives or education purposes.