



IRB Wizard - USA
User Manual for Web Version

Document Information

Document Title: IRB Wizard - USA
Original Document Date: 9 March 2011
Document Version: Version 2
Copyright: Privacy Analytics Inc.
Author: Lisa Gaudette, Khaled El Emam, and Gunes Koru
More Information: www.privacyanalytics.ca

Table of Contents

1	INTRODUCTION	2
2	PARAMETERS	3
3	INTERPRETING RESULTS	5
4	REFERENCES	6
5	CONTACT INFORMATION.....	7

1 Introduction

The IRB Wizard has been created to allow data custodians in the US evaluate re-identification risk for their data sets. The model underlying IRB Wizard has been constructed by analyzing US Census data.

We called the tool an “IRB” Wizard because the basic use case is for an institutional review board that needs to make a re-identification risk decision based on information provided in a protocol. Under that scenario the IRB does not have actual data because the data has not been collected yet, but they still need to make a re-identification risk decision. In that case the intention is to collect data that is de-identified or that will be de-identified at the earliest opportunity after collection. Therefore, we expect IRBs and organizations that have similar use cases to be primary users of this tool.

The purpose of this tool is to estimate the percentage of individuals who will be unique or nearly unique based on several key demographic and geographic variables, in order to evaluate the privacy risks involved with a particular study even before any data is collected. This tool can be used to accept or reject a proposal, to help determine the level of detail that can be collected, and to help determine the level of controls needed when sharing data. Individuals who are unique or nearly unique based on basic demographic data can often be re-identified easily by combining the basic data with public sources of information such as a voters registration lists, or can potentially be re-identified by neighbors or colleagues who have background information about them.

For full year, 5 year ranges, and 10 year ranges, the exact values are computed from the 2000 census data. For Year/Month, the values are estimated based on combining census data with the most detailed birth data available, which varies by year. This provides a more precise measurement than other sources which assume a uniform distribution of births.

There has been some work on estimating re-identification risk using publicly available census data [1-3]. That previous work assumed that births are uniformly distributed throughout the year. We made a more specific assumption in our analysis in that we used vital statistics data to determine the actual distribution of births going back to the early 1900's, and used that information to compute re-identification risk. In theory this approach should produce more accurate results. More comparative data comparing the two approaches, for those interested, will be made available soon.

2 Parameters

1. **Selected Area:** The area of interest; Options include the entire US, each state, the District of Columbia, and Puerto Rico. These can be selected from the map or the drop down menu. Clicking on the background of the map will select the entire country.
2. **Zip Code Digits:** The number of zip code digits that will be collected. From 1 to 5.
3. **Age Range:** The age detail that will be recorded. Currently available options are Month/Year, Year, 5 year range, and 10 year range.
4. **Group Size:** A group is made up of the people who are identical based on the other attributes; People in groups of less than or equal to the group size are considered to be at risk. Group size should be set higher for data that is more sensitive, or that will be released to a wider audience. The group size represents the concept of “k-anonymity” in the computational disclosure control literature. From 1 to 25.
5. **Gender:** Whether or not Gender will be collected.
6. **At Risk:** The estimated percentage of residents of the selected geographic area who are in a group of less than or equal to Group Size based on the other attributes.

The IRB Wizard is a work-in-progress and we do expect these variables and options to expand over time in future releases of the tool.

3 Interpreting Results

One way to interpret the results is to start off with the HIPAA Privacy Rule Safe Harbor items. If we have the three digit ZIP code, year of birth, and gender then the overall US percentage of individuals who are unique is zero. This risk changes mildly as the group size increases, but the change does vary state by state. Therefore, the re-identification risk from a HIPAA Safe Harbor data set (using only the demographics) is low. The risk would still be quite low even if we had the 5 digit ZIP code and the year of birth (around 0.175% for the whole of the country).

On the other hand if we change the geography to a 5 digit ZIP code and make the age in terms of month and year, then the overall percentage of the US population that is unique is approximately 4.2% (example shown in the figure), and it is close to 9.5% if the group size is 5.

To define what acceptable risk is, two parameters need to be specified. The first is the group size, and the second is the percentage of individuals in groups that are smaller than the threshold group size. For example, a (4,1) set of parameters means that individuals whose demographics put them in groups of size 4 or smaller (or groups smaller than 5) are considered at a high risk of re-identification. If more than 1% of the population is at high risk then this would be considered unacceptable.

Historically, different acceptable thresholds have been used to define an acceptable percentage of people at risk and acceptable group sizes. A justifiable default would probably be (4,5). But this would need to be adjusted up or down depending on the sensitivity of the data and the trust one has in the data recipient and the strength of their security and privacy practices.

4 References

1. Sweeney L. Uniqueness of Simple Demographics in the US Population. 2000; Carnegie Mellon University, Laboratory for International Data Privacy.
2. Golle P. Revisiting the uniqueness of simple demographics in the US population. Workshop on Privacy in the Electronic Society. 2006.
3. Benitez K, Malin B. Evaluating re-identification risks with respect to the HIPAA privacy rule. *Journal of the American Medical Informatics Association*, 2010; 17(2):169-177.

5 Contact Information

For more information contact us at:

Privacy Analytics Inc.
800 King Edward Ave. Suite 3042
Ottawa, Ontario K1N 6N5
Canada

Tel: +1 613.369-4313

Fax: +1 613 369 4312

Email: info@privacyanalytics.ca

www.privacyanalytics.ca