Northwestern Medicine

Robert H. Lurie Comprehensive Cancer Center of Northwestern University

Protecting Privacy, Amplifying Insights: De-identifying Clinical Trial Data through Geographic Aggregation with R gatpkg

Daniel Antonio, MPH

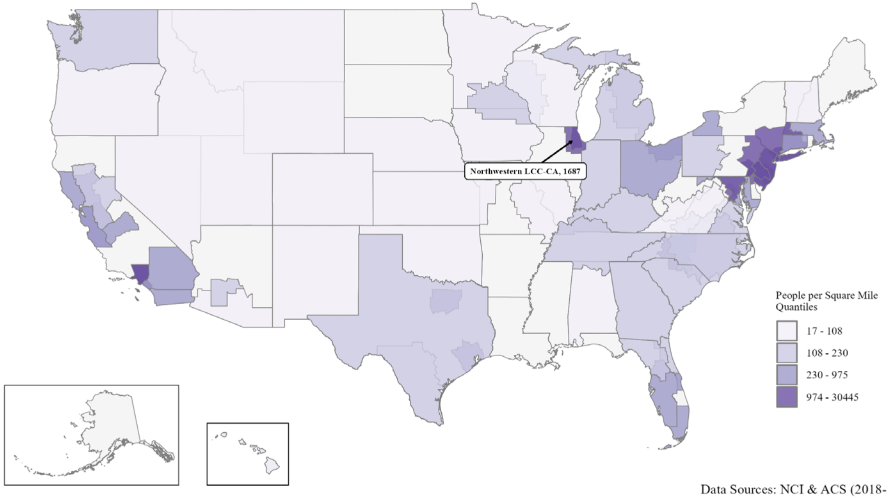
Statistical Analyst

Background

Catchment Area

- Nine county catchment area
- 11th in population density
- 17th in total population
- Total population – 8.7 million
  - 66% of the state population
- 90% of LCC patients come from catchment area

NCI-Designated Cancer Centers Catchment Areas by Population Density

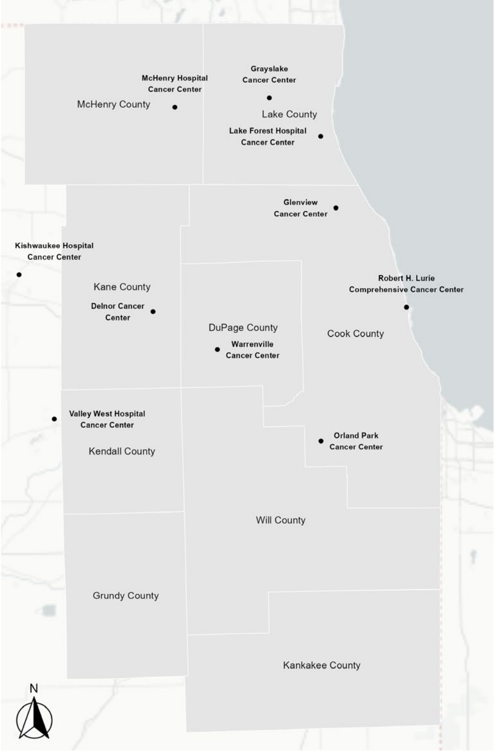


Data Sources: NCI & ACS (2018-2022)

Background

Making it Happen

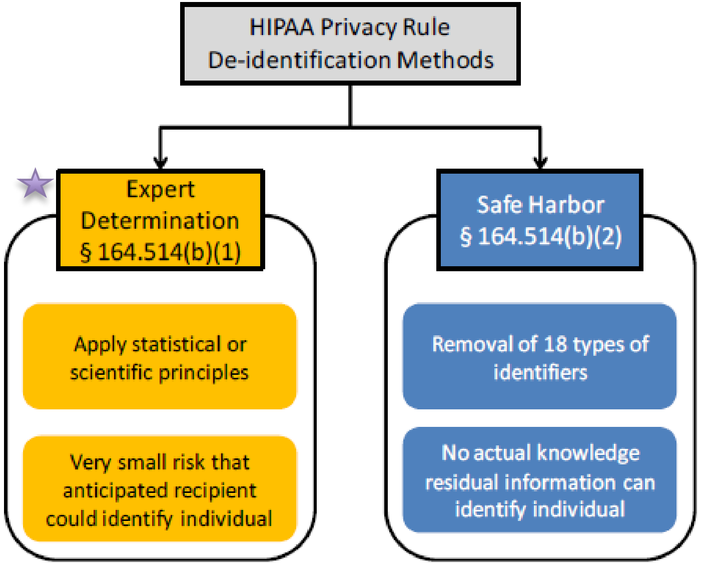
- **Goal:** Map Northwestern Medicine cancer clinical trial data to identify low enrollment areas in our catchment area
- **Issue:** Data must be de-identified before use
- **Solution:** De-identify data with the R Geographic Aggregation Tool (GAT)


McHenry Hospital Cancer Center
Grayslake Cancer Center
McHenry County
Lake County
Lake Forest Hospital Cancer Center
Glenview Cancer Center
Kishwaukee Hospital Cancer Center
Kane County
Delnor Cancer Center
Robert H. Lurie Comprehensive Cancer Center
DuPage County
Warrenville Cancer Center
Cook County
Valley West Hospital Cancer Center
Kendall County
Orland Park Cancer Center
Will County
Grundy County
Kankakee County
N

Background

De-Identification: Expert Determination

- **Techniques:** suppression, generalization, & randomization
- **Goal:** transform the data to have a very small probability of identifying individual subjects using external data
- Risk Assessment conducted after data transformation ensures low risk



HIPAA Privacy Rule De-identification Methods

Expert Determination § 164.514(b)(1)

Apply statistical or scientific principles

Very small risk that anticipated recipient could identify individual

Safe Harbor § 164.514(b)(2)

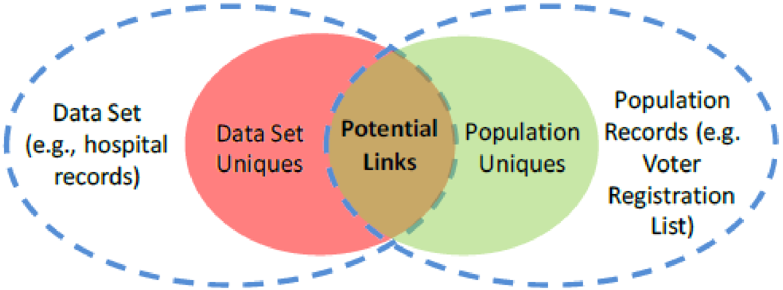
Removal of 18 types of identifiers

No actual knowledge residual information can identify individual

Background

De-Identification: Risk Assessment

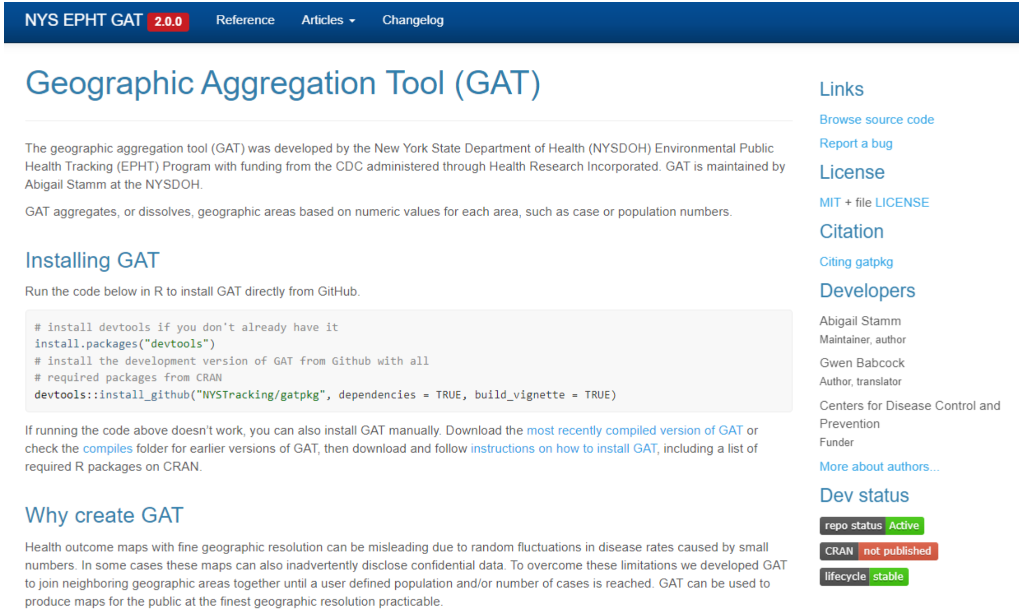
- Evaluates probably of identifying a subject using an external dataset
- Example
  - Hospital records and voter registration lists could have potential links
  - An individual could be matched using unique links in both datasets
- Different probability thresholds exists 1 in 5 (P=0.20) to 1 in 20 (P=0.05)



R Geographic Aggregation Tool (GAT)

An easy solution using generalization

- Addresses small case counts and confidential data by merging geographies based on user-defined requirements
- High Customization:
  - Exclusions
  - Min & Max
  - Merging Boundaries
  - Merging Algorithms



Geographic Aggregation Tool (GAT)

The geographic aggregation tool (GAT) was developed by the New York State Department of Health (NYSDOH) Environmental Public Health Tracking (EPHT) Program with funding from the CDC administered through Health Research Incorporated. GAT is maintained by Abigail Stamm at the NYSDOH.

GAT aggregates, or dissolves, geographic areas based on numeric values for each area, such as case or population numbers.

Installing GAT

Run the code below in R to install GAT directly from GitHub.

```
# install devtools if you don't already have it
install.packages("devtools")
# install the development version of GAT from Github with all
# required packages from CRAN
devtools::install_github("NYSTracking/gatpkg", dependencies = TRUE, build_vignette = TRUE)
```

If running the code above doesn't work, you can also install GAT manually. Download the most recently compiled version of GAT or check the compiles folder for earlier versions of GAT, then download and follow instructions on how to install GAT, including a list of required R packages on CRAN.

Why create GAT

Health outcome maps with fine geographic resolution can be misleading due to random fluctuations in disease rates caused by small numbers. In some cases these maps can also inadvertently disclose confidential data. To overcome these limitations we developed GAT to join neighboring geographic areas together until a user defined population and/or number of cases is reached. GAT can be used to produce maps for the public at the finest geographic resolution practicable.

Links

Browse source code

Report a bug

License

MIT + file LICENSE

Citation

Citing gatpkg

Developers

Abigail Stamm
Maintainer, author

Gwen Babcock
Author, translator

Centers for Disease Control and Prevention
Funder

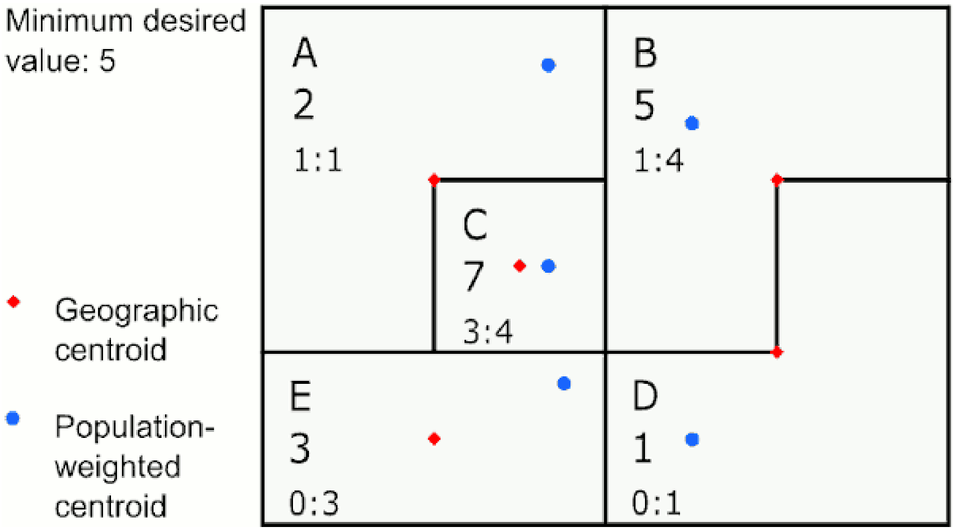
More about authors...

Dev status

Geographic Aggregation Process

Merging to neighbor with least value

- Step 1. Select the smallest number (D) and merge with the neighbor with the smallest value (E)
- Step 2. Repeat until the minimum value is achieved
- **Pros:** largest number of areas (more granular)
- **Cons:** could produce “weird snaky shapes and possibly donuts”

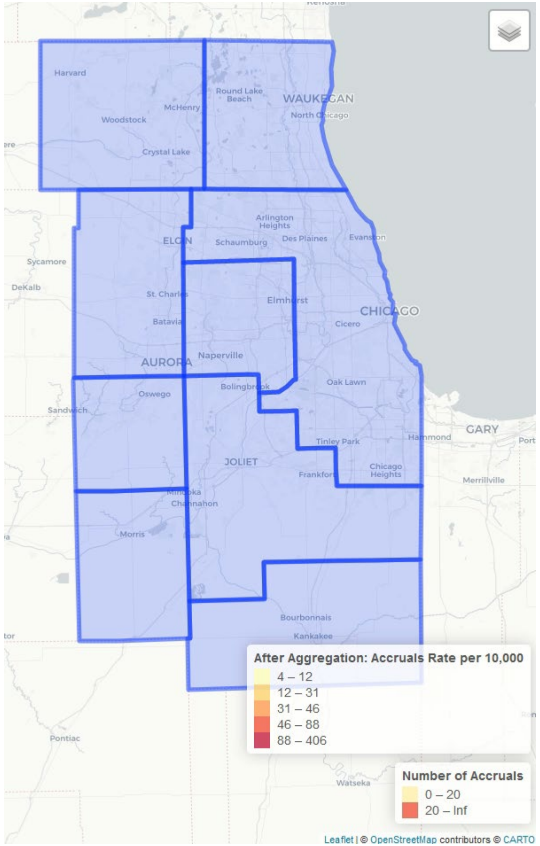


Minimum desired value: 5

De-identification

Our workflow

- Geocoded clinical trial patients from 2018-2023 (DeGAUSS) from all northwestern facilities
- Count the number of accruals per census tract
- Deploy GAT to merge census tracts to meet a minimum value of 20 accruals (P=0.05)
  - Using the least value algorithm
- Exclude tracts with no accruals
- Calculate accrual rates per 10,000
- Conduct risk assessment



Findings

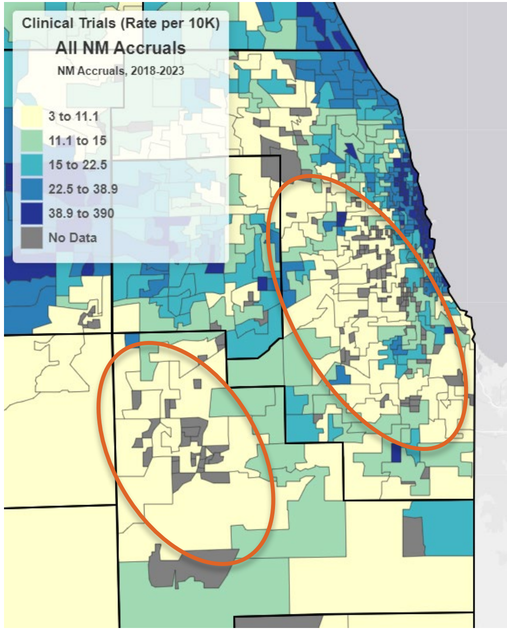
Does it meet HIPAA Standards?

	Before Aggregation, N = 2,100	After Aggregation, N = 695
Number of Accruals	5 (3, 9)	26 (23, 31)
Less than 20	1,793 (85%)	3 (0.4%)
More than 20	103 (5%)	478 (69%)
Zero	214 (10%)	214 (31%)
Re-identification probability	1/1 = 1.0	1/20 = 0.05

Findings

What can we learn?

	Bottom Quantile	Top Quantile
Distance from NM Facility	10 Miles	3 Miles
Uninsured	9%	3%
Household Income	82K	128K
Above HS Education	61%	91%
Minority Population	44%	16%
History of Cancer Diagnosis	5%	5%



COE Activities

Guiding outreach and research

- Disseminating findings and data to key stakeholders
  - Community Advisory Board, Clinical Trials Office, Leadership, Institutes
- Integrating data within Cancer InFocus applications and other reporting systems
- Targeting specific communities with education and outreach
- Conducting further research
  - Target low enrollment areas/groups
  - Explore other determinants of clinical trial participation
  - Explore alignment between patient volume and accruals

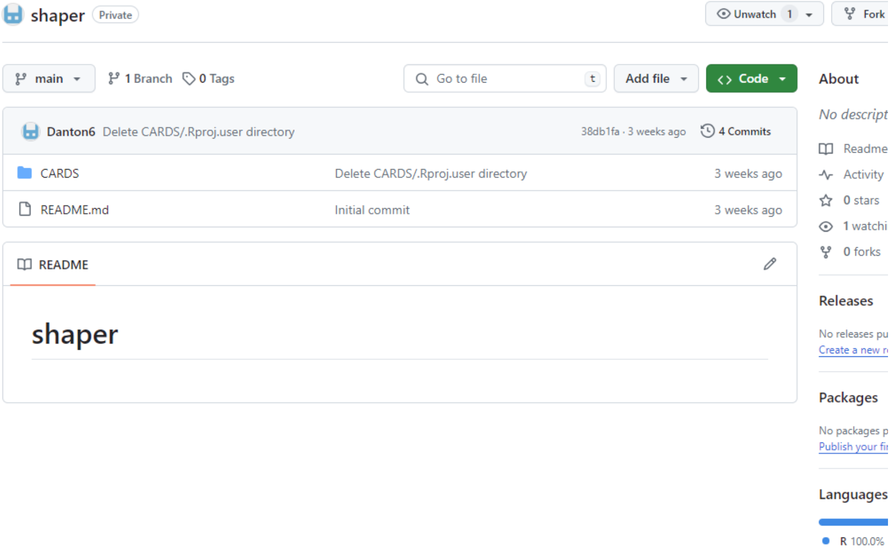
Advantages & Limitations

- **Advantages**
  - Granular data helps better understand our enrollments (who, what, when, where), especially insightful to CTO
  - Complies with HIPAA regulations
  - Relatively simple method to apply
  - Applicable to other data
- **Limitations**
  - Output is only as good as input
  - Aggregated areas are not stagnant geographies

Next Steps

Sharing & Evolving

- Repository of aggregation process
- De-identified patient volumes data
- Continue discussions for how to best use data to inform COE activities



Danton6 Delete CARDS/.Rproj.user directory 38db1fa · 3 weeks ago 4 Commits

CARDS	Delete CARDS/.Rproj.user directory	3 weeks ago
README.md	Initial commit	3 weeks ago

shaper

Thank You

Northwestern Medicine®

Robert H. Lurie Comprehensive Cancer Center of Northwestern University