# The Remarkable Benefit of User-Level Aggregation for Lexical-based Population-Level Predictions

**Salvatore Giorgi**[1] **Daniel Preoţiuc-Pietro**[2] **Anneke Buffone**[1]
**Daniel Rieman**[1] **Lyle H. Ungar**[2] and **H. Andrew Schwartz**[3]
[1]Department of Psychology, University of Pennsylvania
[2]Computer and Information Science, University of Pennsylvania
[3]Computer Science, Stony Brook University
sgiorgi@sas.upenn.edu

## 1 Appendix A. Additional Experiments and Results

**Direct aggregation comparison.** In addition to the Pearon $r$ results above we report Mean Squared Error (MSE) in Table 1

| | Income | Educat. | Life Satis. | Heart Disease |
|---|---|---|---|---|
| Tweet to County | 6.67e-3 | 32.9 | 8.81e-4 | 1083 |
| County | 5.49e-3 | 35.1 | 6.49e-4 | 1105 |
| User to County | **3.68e-3** | **20.7** | **5.51e-4** | **913** |

(a) Unigrams + Topics, Mean Squared Error (MSE)

| | Income | Educat. | Life Satis. | Heart Disease |
|---|---|---|---|---|
| Tweet to County | 6.76e-3 | 35.1 | 1.02e-3 | 1289 |
| County | 6.44e-3 | 34.3 | 6.53e-4 | 1232 |
| User to County | **4.18e-3** | **22.9** | **5.75e-4** | **1028** |

(b) Unigrams, Mean Squared Error (MSE)

| | Income | Educat. | Life Satis. | Heart Disease |
|---|---|---|---|---|
| Tweet to County | 6.73e-3 | 38.0 | 6.61e-4 | 1034 |
| County | 6.39e-3 | 32.8 | 6.40e-4 | 1059 |
| User to County | **3.83e-3** | **21.5** | **5.28e-4** | **866** |

(c) Topics, Pearson $r$

Table 1: Prediction results, reported Mean Squared Error (MSE).

**Users per county.** To gain further insight into differences in accuracy, we look at accuracy as a function of our users per county requirement. For this task we consider the "User to County" approach and build a models using unigrams + topics, varying the required number of users. Figure 1 shows the results. We see a general increase in accuracy as the user-threshold is raised. Of course, this happens at the expense of covering fewer counties, varying from 2153 at the 10 user requirement to 626 counties at the 1000 user requirement.

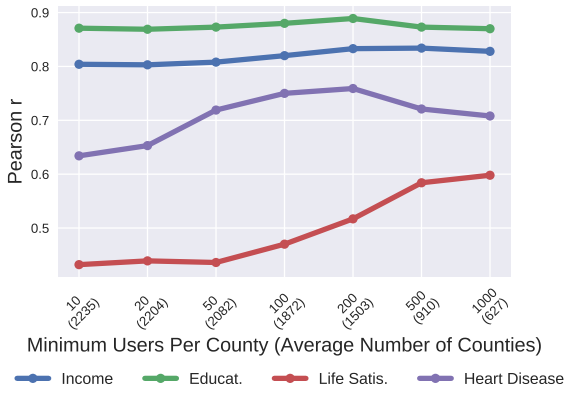


Figure 1: Prediction results (Pearson $r$) using 10% "User to County" (unigrams + topics) when varying the minimum number of Twitter users per county. Parenthetical number indicates the average number of counties across tasks after meeting the minimum user threshold.

**Replication across time.** Here we explore the effect aggregation has on replication over time, considering separate years of age adjusted heart disease mortality rates and aggregate Twitter data over single years (from 2012 and 2013). We chose heart disease as the other variables (income, education and life satisfaction) are not produced every year. Additionally, we chose 2012 and 2013 as the other years in our Twitter sample contain holes or time periods with a 1% sample. Results are shown in Table 2. Here see the same patterns as previous experiments: "County" performs better than "Tweet to County" with "User to County" outperforming both. We also note that all three methods hold across both years with a slight increase across all tasks in 2013, which contains slightly more data (472 million posts in 2013 vs. 455 million in 2012).

| | 2012 | 2013 |
|---|---|---|
| Tweet to County | .50 | .54 |
| County | .52 | .58 |
| User to County | **.58** | **.63** |

Table 2: Replication of Heart disease predictions.