

Supplementary Material

Article Title: Assessing Suicide Risk Reduction in the Interval before Death by Suicide Using Natural Language Processing

Authors: Maxwell Levis, PhD; Monica Dimambro, MS; Joshua Levy, PhD; Bradley Watts, MD, MPH; Brian Shiner, MD, MPH

DOI Number: 10.4088/JCP.25m16249

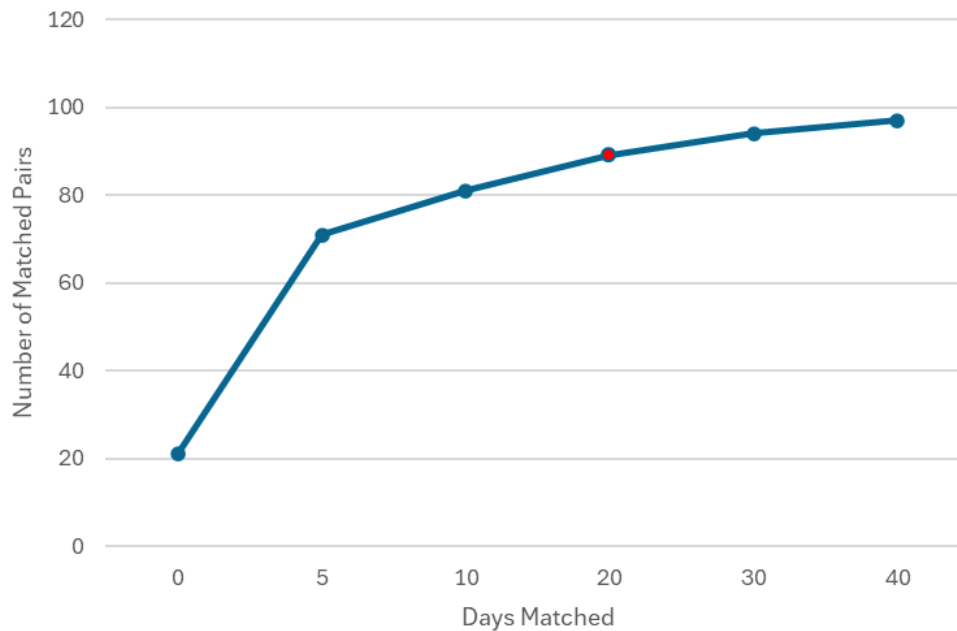
LIST OF SUPPLEMENTARY MATERIAL FOR THE ARTICLE

1. [Figure 1](#) Selection of Caliper Width for Matching of Days Between Flagging and Death Based on Matched Pair Yield
2. [Figure 2](#) Coherence Scores Across Number of Topics
3. [Figure 3](#) The Intertopic Distance Plot and Top 30 Most Relevant Terms for Topic 3 of a 6-Topic Latent Dirichlet Allocation Model
4. [Figure 4](#) The Intertopic Distance Plot and Top 30 Most Relevant Terms for Topic 4 of a 6-Topic Latent Dirichlet Allocation Model
5. [Figure 5](#) The Intertopic Distance Plot and Top 30 Most Relevant Terms Across Topics of a 5-Topic Latent Dirichlet Allocation Model
6. [Figure 6](#) Histograms of Maximum Topic Differences Among Documents, Stratified by Clinical Risk Classification
7. [Table 1](#) Generalized Additive Mixed Model (GAMM) Specifications and Marginal Means Extraction Procedure, as Implemented in R

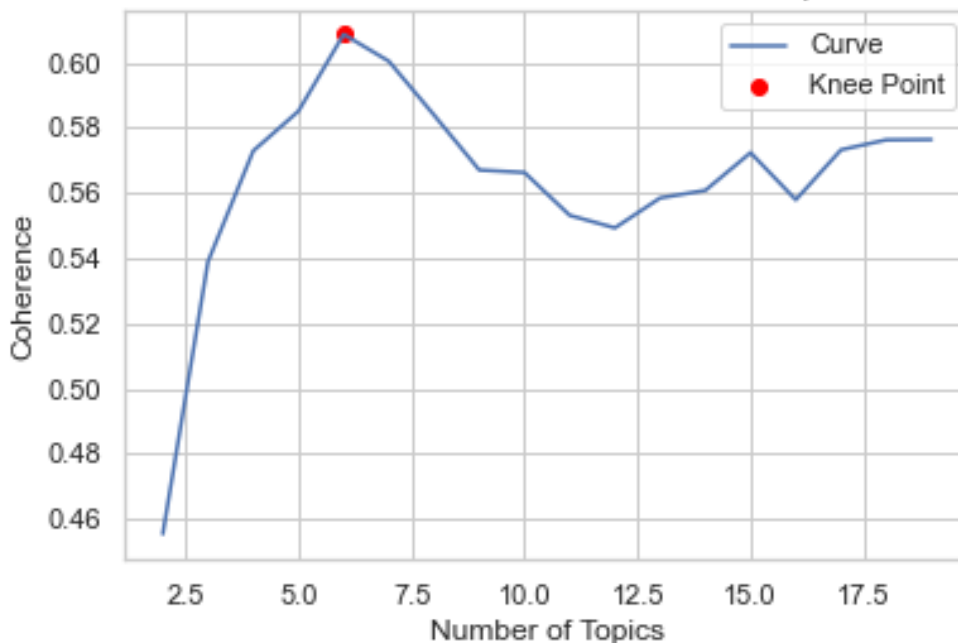
DISCLAIMER

This Supplementary Material has been provided by the authors as an enhancement to the published article. It has been approved by peer review; however, it has undergone neither editing nor formatting by in-house editorial staff. The material is presented in the manner supplied by the author.

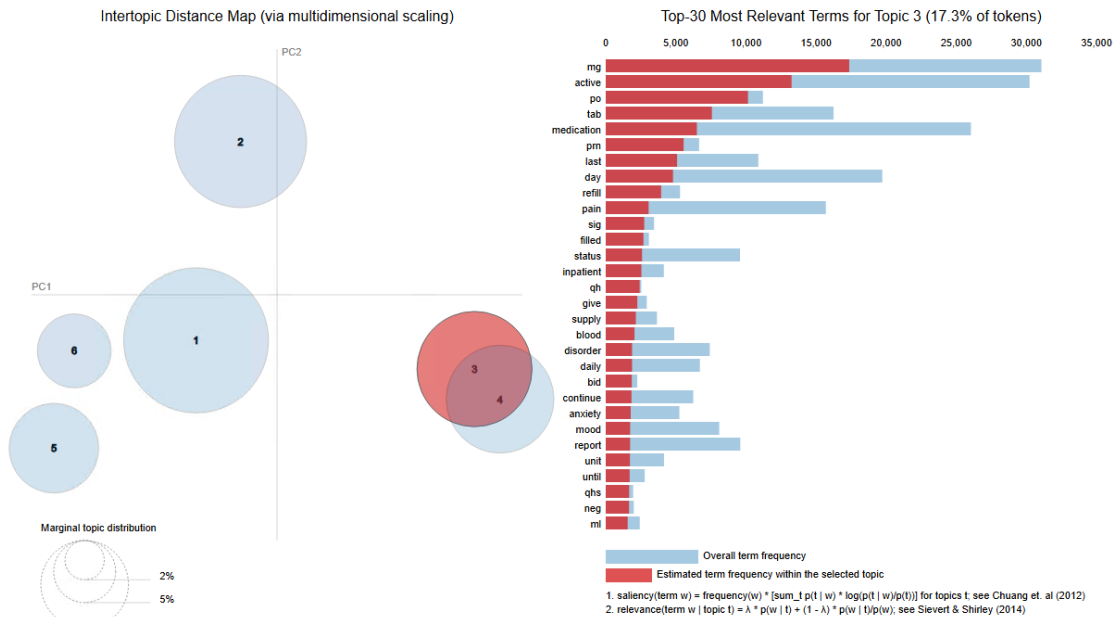
Supplementary Figure 1. Selection of caliper width for matching of days between flagging and death based on matched pair yield. The number of 1:1 matched pairs as a function of matching caliper width (exact through ± 40 days). The plateau region of the curve indicates a range over which the additional caliper relaxation yields diminishing returns in matched pairs; the midpoint of this plateau was selected as the optimal caliper width to balance sample size maximization with matching precision.



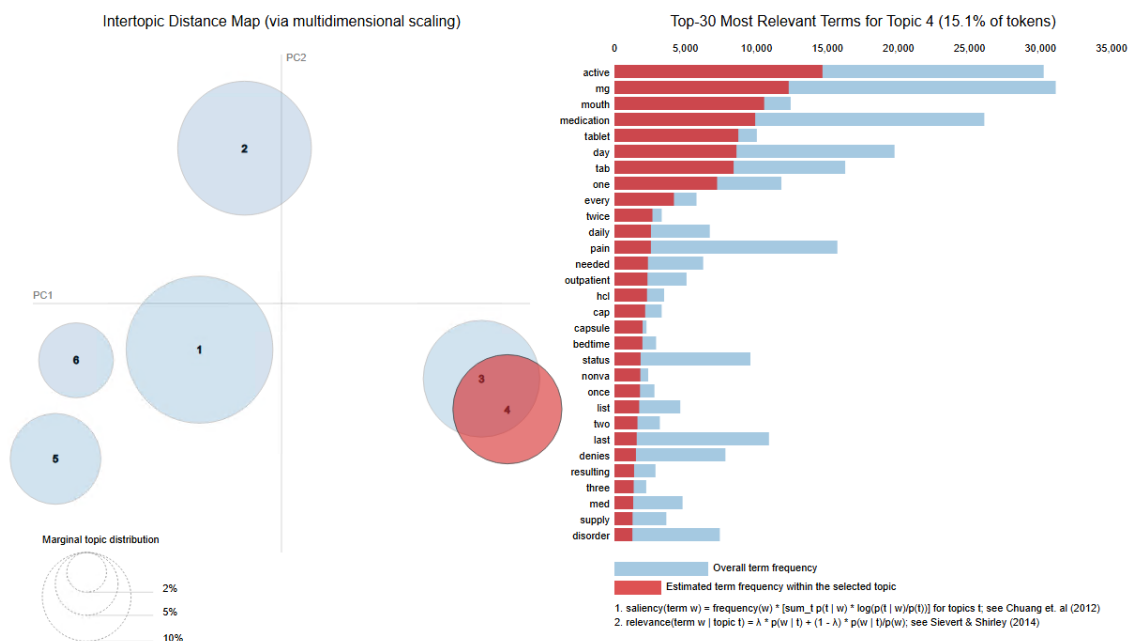
Supplementary Figure 2. Coherence scores across number of topics. The knee point (in red) indicates the optimal threshold for topic number selection. The knee point represents the inflection at which marginal gains in explained coherence diminish.



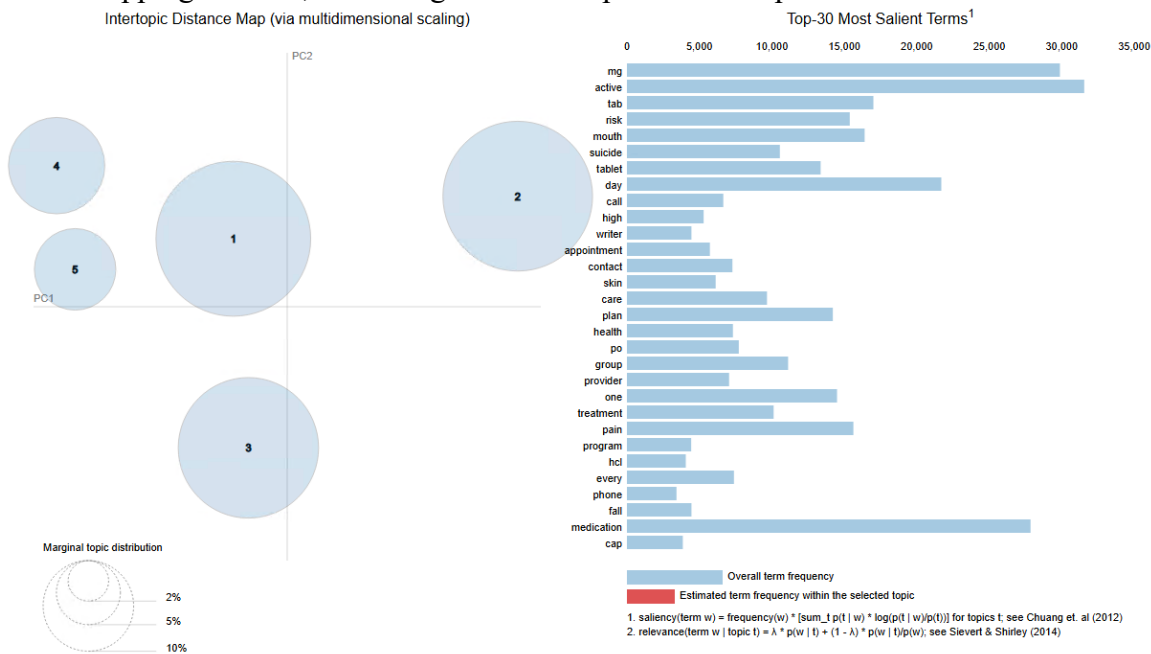
Supplementary Figure 3. The intertopic distance plot (left) and top 30 most relevant terms (right) for Topic 3 of a 6-Topic Latent Dirichlet Allocation Model. Bubble size represents topic prevalence and spatial proximity indicates semantic similarity between topics. Note that the bubbles of Topic 3 and Topic 4 overlap and Topic 3's top words are mostly related to medication.



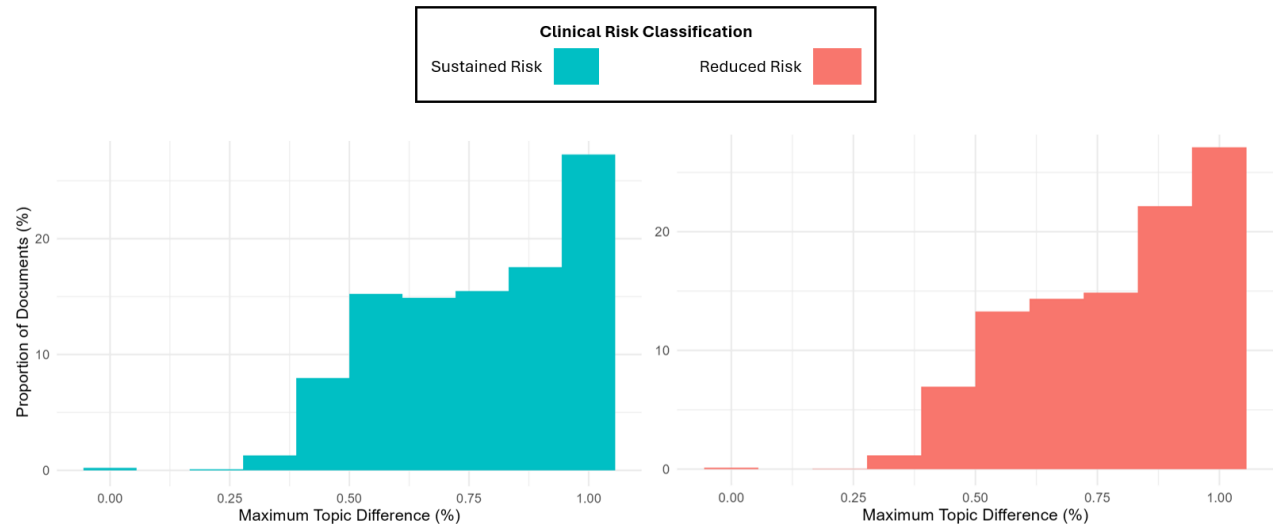
Supplementary Figure 4. The intertopic distance plot (left) and top 30 most relevant terms (right) for Topic 4 of a 6-Topic Latent Dirichlet Allocation Model. Bubble size represents topic prevalence and spatial proximity indicates semantic similarity between topics. Note that the bubbles of Topic 3 and Topic 4 overlap and Topic 4's top words are mostly related to medication.



Supplementary Figure 5. The intertopic distance plot (left) and top 30 most relevant terms (right) across topics of a 5-Topic Latent Dirichlet Allocation Model. Bubble size represents topic prevalence and spatial proximity indicates semantic similarity between topics. Note that there are no overlapping bubbles, indicating semantic separation of topics.



Supplementary Figure 6. Histograms of maximum topic differences among documents, stratified by clinical risk classification. The maximum topic difference is defined as the maximum topic proportion minus the minimum topic proportion, as calculated for each document.



Supplementary Table 1. Generalized additive mixed model (GAMM) specifications and marginal means extraction procedure, as implemented in R.

Task	R Code
Generalized Additive Mixed Model	<code>mod <- mgcv::gamm(topic ~ s(time, by = case_status), random = list(patient_id = ~1), data = df)</code>
Estimated Marginal Means	<code>em <- mod %>% emmeans(~case_status time, at = list(time = seq(min(df\$time), max(df\$time), length.out = 100), type = "response", data = df) %>% as.data.frame() %>% mutate(outcome = outcome)</code>