State Health RegData 1.0 User Guide

Kofi Ampaabeng, Stephen Strosko

December 8, 2021

1. Purpose

The State Healthcare RegData (State Health RegData) belongs to the Mercatus Center's RegData suite of products. Using the QuantGov platform, State Health RegData identifies healthcare regulations in 44 US states. The dataset includes the following output, like all RegData products: the probability that the unit of regulation pertains to healthcare, the total number of restrictions, including the types of restrictive terms (shall, must, may not, required, prohibited), wordcounts, the complexity of the text, and the industry relevance.

2. Content

State Health RegData identifies the healthcare regulatory restrictions in all US states with data available in State RegData 2.0. In building State Health RegData, Mercatus researchers used the QuantGov platform to train an algorithm to predict the probability that a unit of regulatory text applies to the provision of healthcare services in a jurisdiction. The training involved defining what constitutes a healthcare regulation. We classify a document as pertaining to healthcare if the subject of the regulation include the provision of healthcare services, the persons who can provide them, creation, and distribution of healthcare products such as drugs, devices, etc. This definition excludes other types of regulations that pertain to health but do not directly regulate healthcare products or services. For example, we exclude occupational, environmental health and safety regulations.

3. Healthcare Probability

The key metric of State Health RegData is the probability that a unit of regulation imposes applies to healthcare, with values ranging from 0 to 1, with one being certainty that the regulation is healthcare. This probability information is included at the document level for all documents in a corpus. This means the data includes even those documents with low probabilities. For convenience, we have included in the API suggested thresholds to classify the regulations. We did not establish the threshold a priori. Rather, each state's threshold was calculated from the probability distribution [Technical notes 1, 2]. These thresholds have also been used to generate state-level summary of healthcare regulatory restrictions. These data can be accessed in the usual manner for RegData.

4. Technical Notes

1. The performance of the classification algorithm varied across states. However, within each state, there is a clear bimodal distribution of predicted healthcare probabilities. This allowed for the development of state-specific thresholds for inclusion.

2. The threshold for a state is calculated as the mode healthcare probability (for a continuous variable, this is the highest value) minus twice the standard deviation of all predicted probabilities above the F1 score.

Table 1: Document Level Variable Description

Variable	Description/Definition
document reference	The reference for the document used as a source for the prediction algorithm
document title	The official title of the document referenced.
healthcare probability	The probability that the unit of regulation imposes some form of occupational licensing restrictions/requirements
restrictions	The total number of restrictions, comprises the sum of “shall”, “must”, “may not”, “required”, “prohibited”.
shall	Occurrences of the word “shall” in the unit.
must	Occurrences of the word “must” in the unit.
may not	Occurrences of the word “may not” in the unit.
required	Occurrences of the word “required” in the unit.
prohibited	Occurrences of the word “prohibited” in the unit.
words	The total number of words in the unit

industry code	Industry classification code (NAICS)
industry probability	The probability that the regulatory unit pertains to industry identified by “industry” variable.