Healthcare RegData: Classification of Healthcare Regulations into Topics

# Introduction

Over here at the Quantitative Health Lab (QHL), we are interested in answering questions like “how do the variations in state Medicaid regulations affect access to care, quality of care, and healthcare outcomes?” While simple, answers to such questions are not easy because there is no objective way of measuring Medicaid regulations and how they vary across states. Over the past few years, we have created Healthcare RegData, which builds on the foundations laid by the RegData suite of products. RegData and Healthcare RegData measure regulations (healthcare regulations) by first identifying them and then measuring the restrictions they impose. In addition, we determine the sectors of the economy (as defined by the NAICS codes) that are affected by a regulation. These data still do not answer our primary question. While classifying regulations into sectors is incredibly useful, the sectors are broad. For example, back to our original question, there is no NAICS sector or industry that would perfectly be applicable to Medicaid. We therefore set out to produce the data that would be useful in helping researchers like ourselves answer this question and others like it.

The updated Healthcare RegData now classifies regulations into twenty topics. QHL researchers first created a list of topics that are relevant to scholars. This list is not exhaustive, and we will modify it in future versions of the data, including adding more topics.

# Data Dictionary

The data are in two parts: classification and data measurement. The following tables describe the elements in each dataset.

## Classification

**Figure 1: Classification Data Dictionary**

| **Column** | **Description** | **Notes** |
| --- | --- | --- |
| Document type | The type of document used in creating the data. One of two values *regulations* or *statutes.* |  |
| State | Name of the state. |  |
| Year | The effective year of the document. |  |
| R1\_title | Reference: Title | States differ how they subdivide official documents (regulations and statues in this case) but for simplicity and uniformity, Healthcare RegData has adopted Title->Chapter->Section->Clause  With clause being the smallest unit and the unit of analysis. |
| R2\_chapter | Reference: Chapter |
| R3\_section | Reference: Section |
| R4\_clause | Reference: Clause |
| Label | Topic | One of twenty topics into which the clause (R4\_clause) has been classified. |
| Probability | The probability that the document (R4\_clause) pertains to the topic identified in *Label*. | We only present those topics for which the probability exceeds 50%. Full tables, include those not relevant to healthcare are available via the QuantGov API. |

## Measurement

**Figure 2: RegData Restrictions**

| **Document type** | **Description** | **Notes** |
| --- | --- | --- |
| State | Same as above |  |
| Year |  |
| R1\_title |  |
| R2\_chapter |  |
| R3\_section |  |
| R4\_clause |  |
| Shall | Number of occurrences of the restrictive term “shall.” |  |
| Must | Number of occurrences of the restrictive term “shall.” |  |
| May Not | Number of occurrences of the restrictive term “shall.” |  |
| Prohibited | Number of occurrences of the restrictive term “shall.” |  |
| Required | Number of occurrences of the restrictive term “shall.” |  |
| Restrictions | Number of occurrences of the restrictive term of all restrictive terms (“shall,” “must,” “may not,” “prohibited,” “required”). |  |
| Readability | The readability score of the unit of analysis (R4\_clause). |  |
| Words | Total number of words in the unit of analysis (R4\_clause). |  |

## Combining Data

Use the key fields: R1-R4 to combine the measurement data and the classification data if necessary.

# Methodology

We developed the data using text classification algorithms with state administrative code and laws (statutes) as the primary data sources.

Challenges with classifying regulations into topics – Our goal was to classify the smallest unit possible. We therefore settled on the “clause” – the fourth level in the reference. Title (top/first level) 🡪 Chapter (second level) 🡪 Section (third level) 🡪 Clause (fourth level). The fourth level isn’t clearly marked in all documents and so we developed a series of regular expressions to identify the logical coherent breakdowns of sections.

## Data

The primary data sources for the topic classification are state administrative codes and laws (statutes). For simplicity, we call these documents. The Mercatus Center maintains a database of laws and regulations for all but five states (Vermont, New Jersey, Arkansas). From these data, a team of researchers identified examples of regulations and laws that pertain to each of the topics. The process proceeded as follows. First, two junior researchers used publicly available resources that identified state and federal documents that pertain to a healthcare topic of interest. A document can belong to multiple topics, a concept known as multilabel classification. Concurrently with the first step, a team of developers built an unsupervised model to classify state and federal laws and regulations. Second senior researchers reviewed the classifications from the first step for accuracy and resolved discrepancies between the junior researchers.

## Topic selection

Topic selection was ad-hoc and influenced primarily by research interests. We will continue to expand in subsequent versions of the data. The topics covered are:

| Topic | Description |
| --- | --- |
| Certificate of Need | Require permission to build, expand medical facilities and equipment. |
| Health Facilities | Health facilities including buildings, equipment; may include certificate of need documents. |
| Health Insurance | Health insurance coverage, benefits, mandates, claims, etc. |
| Health Insurance - Mandate | Mandate some form of health insurance coverage or benefits. |
| Medicaid | Documents that pertain to the Medicaid program. It encompasses all Medicaid topics below. |
| Medicaid - Coverage | Rules about what Medicaid covers. |
| Medicaid - Eligibility | Who/how one becomes eligible for Medicaid. |
| Medicaid - Managed Care | Managed care rules. |
| Medicaid - Program Integrity | Rules/laws that ensure the integrity of the Medicaid program, including provider and administrative roles, sanctions, and penalties. |
| Medicare | State Medicare supplementary insurance regulations. |
| Prescription Drugs | Issues of prescription drugs, including formulary, substitutions (mandatory or optional), pharmacies, etc. |
| Prescription Drugs - Substitution | Mandatory or optional substitution of brand name prescription drugs with generic equivalents or biosimilars. |
| Privacy | Patient privacy rules, including personal records. |
| Professional Licensing | Licensing of medical professionals, including physicians, nurses, etc. |
| Professional Regulations | Professional regulations, including licensing, that pertain to medical professionals. |
| Public Health | Public health rules including reporting, vaccinations, quarantines, facilities, etc. |
| Public Health - Disease | Specific diseases that are deemed public health concerns, including tuberculosis, HIV/AIDS. |
| Public Health - Reporting | Public health reporting requirements. |
| Scope of Practice | Restrictions on what medical professionals such as pharmacists and nurses can do. |
| Telehealth | Telehealth rules including benefits, payments, coverage, and services. |

The list of topics is not static, and we will be updating it over time.

## Algorithm

We used a pre-trained deep learning model, fastText, to train the classification system. FastText is an efficient continuous bag of words approach to word embedding.

### Threshold Selection

The classification by the algorithm model is probabilistic. That is, for each document, it returns the probability that it pertains to all the topics listed previously. A document can belong to multiple topics. Probability can be used in one of two ways – binary classification or probability. With binary classification, we selected the conservative threshold of 0.5. That means the default results only include data for topics for which the algorithm assigned a probability of more than 50%. Using the data under this assumption, one can count the number of regulations that pertain to a topic. In addition, together with the measurement data, one can identify the regulatory restrictions (following standard RegData measurement) that pertain to a topic. The complete data, without the thresholds, are available upon request. With the probabilistic results, the metric would be, following RegData nomenclature, topic-relevant restrictions, which is the number of restrictions multiplied by the probability. This is like industry-relevant restrictions.

# Usage

The unique contribution of RegData to policy research has been to provide a unique and replicable means of quantifying regulations. With these data, researchers can and have been examining the role of the regulatory landscape in shaping some economic or political outcomes. Healthcare RegData and its topics derivative can be used in a similar manner. First, we can obtain a quantitative description of how states vary in their healthcare regulations. But even more important, researchers can examine how these differences contribute to healthcare input decisions and healthcare outcomes in various states.