Bringing big data into public policy research: Text mining to acquire richer data on program participants, their behavior and services

Abstract

Text mining is a process that has been used in the corporate sector extensively to development meaningful structured data from unstructured text. We discuss how such unstructured data in public sector information systems can be mined to provide data that is not collected in quantitative ways by government agencies. Text mining can enhance administrative data that many researchers have been using for decades and address some of the criticisms of administrative data by adding richness that is often buried in text fields of administrative data or documents of various kinds. This new data includes capturing the behavior of program participants, frontline practitioners, and service providers—not typically thought of being in administrative data. In addition, such data has promise to inform the daily operations and management of the public agencies that have it. The examples discussed are from the human services and juvenile justice fields.

Finding the right data source to study a large public social program is a major challenge for social scientists. Today, the primary options are to conduct primary data collection, typically a survey, or to use multiple administrative databases, or to use a combination of the two. The decision is usually made on the basis of cost, unless, of course, survey data has already been collected by a third party, usually the federal government (e.g., Census Bureau databases). Typically, survey data is seen as richer than administrative data, in that important constructs are not typically collected well or at all in government information systems (Hotz, Goerge, Balzekas, & Margolin, 1998). Efforts that combine the survey and administrative, most notably, but not limited to, the Longitudinal Employer-Household Dynamics project, have been shown to be quite fruitful for research purposes (Lane, 2006).

However, methods associated with big data and the existence of big data in certain domains of public policy have potential benefits for providing richer data without the cost of primary data collection. Text mining, in particular, offers the ability for investigators to extract meaningful information from unstructured text. Computational methods to do this are increasingly used to access social media data and data from government information systems. These methods improve such data sources and lower costs for obtaining richer social, health, and psychological data. Often, the behavior of individuals and their service providers are capture in unstructured text. This paper focuses on data from the public sector sources of text (reports, assessments, case notes) as more of it is being stored in digital format rather than on paper in files or cabinets. In particular, we will present examples in human services and juvenile justice from our own

and others' work to show that text mining of public sector data holds great promise for public policy research, but also for better informing public agency administrators about the characteristics and performance of their agencies. Text mining is already being heavily used in health care and medicine because of the ubiquity of unstructured text data in electronic health records (Evans & Rzhetsky, 2011; Jensen, Jensen, & Brunak, 2012; Suominen, 2014).

The use of big data, narrowly defined in terms of size, in research on public policies and programs, has lagged behind other fields (Coulton, Goerge, Putnam-Hornstein, & DeHaan, 2015). However, if one includes administrative data in the definition of big data, it has been used for the study of social programs for well over two decades (Cancian, Haveman, Meyer, & Wolfe, 2002; Fluke, Edwards, Kutzler, Kuna, & Tooman, 2000; Roos, Nicol, & Cageorge, 1987). Investigators have been using structured data from federal, state, and county information systems (administrative data) to quantitatively study the participation and outcomes of individuals and families in the human service system. In some cases, investigators have used text from these systems to conduct qualitative analysis (Smithgall, Jarpe-Ratner, Gnedko-Berry, & Mason, 2015). Although not big enough to warrant the largest storage devices or most powerful computational platforms, these data are quite complex in that they are used to track a full range of information about cases and, thereby, include data on the characteristics of children and parents, service events, payments, and providers. These data, however, have not been collected primarily for purposes of testing theories or learning more about the behavior of

these individuals and those that provide these services, but for managing these service systems and cases and meeting federal compliance requirements.

Much of the data that originates in the public sector about individuals who have received benefits is sensitive and confidential, therefore making data security an important feature of this kind of analysis. These data contain information about criminal behavior, health conditions, substance abuse, and many other aspects of individuals' lives that are considered to be private. This aspect will be discussed further in the data section below.

As storage has become less expensive, as the need to access a broader range of data continues to grow, and as the public sector has moved away from the use of paper records, more data is now being collected in government and school data systems that may fit into the category of big data by virtue of size. Frontline practitioners—social workers, police officers, and teachers—are being asked to produce text that is saved in a digital format. The format may be anything from word processing files to text fields in relational database management systems or NoSQL databases. Regardless of the format, the fact is that more text is now accessible because these individuals are required to use computers to keep their notes and reports in digital form. As the amount of this data increases, it becomes a significant resource for research and analysis, although perhaps a greater burden for the practitioners

While text mining offers much, there are also important cautions that must be taken in order to not create incorrect meaning. Grimmer and Stewart (2013) provide four

principles that should guide prudent use. The first is that since it is impossible to know exactly the thought behind how text data was generated, one must be cautious about any interpretation in building a quantitative model of text analysis. Second, human interpretation of text is still necessary and computation should augment human judgment. Third, a global method of text analysis has not been found and may never be. Therefore, the method always has to match the purpose of the analysis. Fourth, validation is necessary and there are a number of ways in which to do that. In this article, we use one approach that was found to help accomplish our data development and research goals.

TEXT MINING BASICS

Text mining, most broadly, is the process of algorithmically finding non-trivial patterns from unstructured text data. This leads to structuring and quantifying data that is currently unstructured and unavailable for quantitative analysis. There are many tasks associated with text mining: when Google, for example, ranks how relevant various documents are to a search query, that is based on a type of text mining. When a software program finds and flags terms that appear several times within some text data, that is also a type of text mining. In these and other examples of text mining, a computer program attempts to comprehend textual content by parsing documents and collecting relevant information. The collected data can be used to map words and phrases to concepts and quantitative information, thereby creating structured data from the originally unstructured data. Once patterns have been noted and the data has been organized into a structured format, researchers can analyze the data in various ways.

As textual sources continue to become more plentiful, increasingly sophisticated and powerful text mining methods and tools are being developed and used for a variety of purposes. The tools are often able to efficiently process large amounts of data, in real-time or as batch jobs, and can scale to accommodate increased demands. Developments in combining methods with pattern-oriented methods are able to exploit both the large amounts of text corpora and the accumulated knowledge of subject matter experts in extracting meaning from text

As mentioned, Google, among other web giants, uses text mining to take advantage of the vast amounts of text on the World Wide Web to determine the needs, desires, and characteristics of customers and participants. The text data in SACWIS systems affords an exciting opportunity to do the same. With text mining, one has the opportunity to extract large amounts of data from SACWIS systems and mine for patterns that can be analyzed by researchers to understand what services have been offered or provided to families and what elements of well-being may be changing within a caseload. Not only will researchers be able to use tools already widely accepted, text mining can also enable the use of new tools such as social network analysis.

In addition, text mining allows for large-scale learning, as it is meant for extracting patterns from millions of text entries. Researchers in child welfare must now be satisfied with small file extractions, where analysts do qualitative analysis of at most a few hundred files. With text mining, researchers will be able to analyze data from potentially *hundreds of thousands* of children and their families.

As with any new technology, there will be challenges of security and of adapting text mining to the specialized language of child well-being. However, the authors have extensive experience in addressing these security issues through separating identifying information and segregating identifiers from other data. In fact, text mining will allow us to seek out names and other identifiers and help in the de-identification process.

Machine learning involves developing algorithms that allows "learning" from data and making predictions from that learning (Hindman, 2015). Because text mining methods currently in use combine machine learning methods with pattern-oriented methods, text mining is designed to adapt to specialized language in at least two distinct ways. Through knowledge engineering sessions with subject matter experts, patterns expected to exist within the specialized domain can be identified and operationalized into pattern searching algorithms. Also, on the machine learning side, as text mining software processes documents, it picks up recurring patterns of language, which may or may not be known to subject matter experts; this information can be used to "back out" the patterns in corpora via unsupervised methods, or, in combination with expert coding, the text mining algorithm can learn to identify the patterns in the domain on its own.

CHILD WELFARE SERVICE EXAMPLE

The ability to conduct large-sample rigorous research at the state or local level in child welfare has been mostly limited to the use of administrative data—often linked across programs—in the past 25 years. Since the "discovery" of administrative data by child welfare researchers in the early 1980s (Goerge, Fanshel, & Wulczyn, 1994), many

researchers have taken up the gauntlet of using administrative data to explain entry and re-entry into foster care, duration of care, health care provided to foster children, educational outcomes of children, delinquent behavior of children, and caseworker effects. Administrative data has been linked to many surveys to extend the richness and timeliness of the data. The National Survey of Child and Adolescent Well-Being (NSCAW) is also a significant contribution to the data available to understand developmental and a host of other issues in greater detail (NSCAW, 2007). However, because it is a survey with a relatively small sample, local research is not possible and the data has only been collected for two cohorts. Further, NSCAW did not collect data on biological parents of foster children or the home environment of those parents. Many argue the parents are the true target of interventions in the child welfare system, since their inability to protect their children is the reason that the family came to the attention of the child welfare system in the first place. Also, while very important, NSCAW is unlikely to be replicated often enough to be useful to state and local administrators. With the methods described in this paper, we attempt to address a gap in the methods and available data in child welfare research *and* provide a new source of quantitative (or structured) data that will inform state and local policymakers.

Since the passage of the Adoption and Safe Families Act of 1997, the focus of the child welfare field has been to improve the permanency, safety, and well-being of children in foster care. The measurement of well-being has been a major challenge (Bass, Shields, & Behrman, 2004). On a national level, NSCAW has shown how poor the well-being of foster children is (Planning & Children, 2012). At a local level, through the Child and

Family Service Reviews (CFSR), small samples of cases have been used to understand strengths and weakness of state agencies to improve the three goals of the system. A significant amount of effort goes into collecting these small samples and, again, they have little analytic value below the state level (Courtney, Needell, & Wulczyn, 2004). However, we rely on the experts who study these small samples to assist us in the text mining exercise.

In this paper, we are reporting on our test of the feasibility of using large amounts of text data collected and entered in computer systems by caseworkers to better describe an important decision-making setting—the courts—and the behavior of families of children in the foster care system. We focus on two issues that have been relatively difficult to collect data on in the child welfare system. Akin (2011) found only six studies that addressed whether substance abuse was an important factor in the achievement of permanency. We also look at the impact of court issues on reunification. The effect of court processes has also not been addressed sufficiently in the literature.

DATA

A Statewide Automated Child Welfare Information System (SACWIS) is a system that is "expected to be a comprehensive automated case management tool that meets the needs of all staff (including social workers and their supervisors, whether employed by the State, county, or contracted private providers) involved in foster care and adoptions assistance case management." Under HHS policies, "staff are expected to enter all case management information into SACWIS so it holds a State's 'official case record'—a complete, current, accurate, and unified case management history on all children and

families served by the Title IV-B/IV-E State agency." However, much of the data that describe the well-being of children and other family members, as well as case processes, contained in SACWIS is unstructured text fields and not coded into quantitative information that can be easily analyzed.

Increasingly, state child welfare agencies have such data as new, 21st century systems are built and used. The tradition of social workers maintaining case notes (i.e. narratives) has been included in SACWIS systems through the entry of case notes, assessments, and other unstructured text data. This is a large amount of text, of which a small amount has been analyzed using qualitative methods. We are now able to analyze large amounts of these data to create new structured, quantitative data.

Case notes

Case notes are used to describe events and conditions that occur during contact between caseworkers and any case member. Any time that a caseworker has contact with a case member or engages in any activity on a case, that worker should enter a case note. An example of a case note, below, shows the type of information that is available. (Names of individuals and organizations are masked. Spelling and grammar errors were not fixed.)

> [case note ID] *Today was the continuation of the Shelter Care hearing. This case was just assigned to me. I was informed that all parties were at the court house. I phoned the CDEF office at the court house and asked if all parties could wait for me. I went to the court house and we had a Child Family Team Meeting. When I arrived, the following people were present: CDEF, XXXX, maternal grandmother,*

maternal grandmother, father-XXXXX and XXXXX friend-XXXX. I got phone numbers and addresses for all. We discussed the next steps in the case. I explained the next several court appearances and what they mean. I inquired whether XXXX sees his children or not. He began to make excuses about XXXXX stopping him and not wanting to fight with her.

This note providers information on a court event, individuals that were present, the caseworker's recent assignment to the case, and what took place. The veracity of this data is unknown, although there are characteristics of this data that suggests that it is a reasonable source of information on particular cases. First, there can be dozens of case notes on a particular case. The average is 54 case notes per child. Information that is not available at the time of case opening may become known to the caseworker at a later time. Also, caseworkers on a particular case change often—up to three times in a year (Goerge, 1994). As each case is assigned a new caseworker, it provides a new opportunity to reassess the case and provide additional information. Therefore, there are typically multiple reporters on each case and many more on cases of long duration.

METHODS

Knowledge Engineering

We began the research with two knowledge engineering (KE) sessions with DCFS experts over the course of one day. The purpose of these sessions was to extract as much useful implicit knowledge that was accumulated over the careers of the subject matter

experts in order to derive patterns for our text mining efforts. The KE sessions were divided into two parts. The first KE session involved a set of introductory discussions and an overview of the goals of the project. This session was used to explore what types of text sources and what parts within those sources factored most heavily into the experts' assessments of a case.

With the information gathered in the first KE session, we designed a case note coding framework that was used in the second KE session. In the second session, the experts coded documents in order to uncover explicit patterns in the case worker notes that they were using to make their assessments of the cases. The coding framework presented individual cases, broken down into individual case worker notes, to the experts. Four categories of coding were used:

- Substance abuse: positive indication
- Substance abuse: negative indication
- Court delays: positive indication
- Court delays: negative indication

The experts were instructed to highlight passages in the notes based on these four categories. One result of the coding was that the experts did not find clear cases of the two negative indication categories and, therefore, in the final coding only the positive indications of substance abuse and court delays were collected. We found that substance abuse, where an individual is abusing a substance or getting treatment, was somewhat easier to code than court delays. In addition, one can use dictionaries of substance abuse terms to help identify drug use and treatment language in the text. Court delays, which

are an issue more specialized to the child protection and juvenile justice fields, were somewhat more difficult to code due to the wide range of phrases that were used by the caseworkers to identify the issue. However, the analysis determined that there were a fairly limited set of terms that made up the majority of the text coding (court delay, continued hearing, etc.). Regardless of the ease or difficulty of coding the text, the immediate benefit was the ability to know exactly when a worker or supervisor made an assessment of substance abuse or court delay, which added richness to the data that had not existed before.

The KE sessions generated patterns and from that pattern syntax was developed. The pattern topics and subcategories are:

- Substance Abuse
    - treatment: substance abuse treatment
    - use: indicating use
    - testing: related to testing for substance abuse
    - drug: drug type
- Court Delay
    - potential delay: general delay-related phrases
    - court: court-specific keywords
    - adoption: indicating adoption topic

The text mining then attempted to identify instances of these categories in the documents using the patterns generated from the KE sessions. In the next section, we describe how

the information we gathered through the KE sessions were used for mining the SACWIS case notes.

Text processing

We conducted text mining on case notes from the cases of 18,964 individual foster children in foster care in 2011. The text mining software was built on top of the free and open source Apache UIMA (http://uima.apache.org) and Apache uimaFIT libraries (Ogren & Bethard, 2009). The software processes each document, in this case an individual case worker note, according to the following processing pipeline:

1. The metadata of the document is extracted. This includes the unique case ID and case note ID, along with the contact date and other details.
2. The regular expression (RE) annotator (see discussion of RE annotator below) is run. The RE annotator annotates phrases that match regular expression patterns with substance abuse or court delay annotations, including the substance abuse or court delay category associated with the regular expression pattern.
3. The document annotation results are inserted into a database, populating the following three database tables:
    a. case_note: includes all the meta data from step 1
    b. substance_abuse: includes the case ID, start and end of the annotation, the category, the underlying matching text, and the pattern that was used for the match
    c. court_delay: same as the substance abuse table but for court delay annotations

Regular Expression (RE) Annotator

The RE annotator performs pattern matching where the pattern is described using a pattern language. The language can specify things like the text to match, optional text to match, word boundaries, and other characteristics of the word in the note. The specific pattern language that was used in the RE annotator is based on Java regular expressions.[1] To this we added the ability to nest the patterns within each other. Each pattern is associated with a category and new patterns can use that category as a placeholder for all patterns of that category. For example, the pattern:

using (drugs|${category.drug})

matches "using drugs" as well as "using" followed by any patterns in the drug category. So, given:

vicodin = category : drug

heroin = category : drug

the "using (drugs|${category.drug})" pattern will match "using vicodin" and "using heroin." See **Error! Reference source not found.** for the specific patterns that were used in the RE annotator.

Analyzing the Annotation Database

Once the database is populated with the information gathered by running the document processing pipeline, it is available for querying and analysis. For each unique case, we can determine substance abuse or court delay characteristics based on the annotation

---

[1] See https://docs.oracle.com/javase/8/docs/api/java/util/regex/Pattern.html for details on Java regular expression patterns.

information on each of the constituent case notes. For the substance abuse component, the existence of substance abuse annotations indicated that substance abuse was being referenced in a case note with high probability. We were able to determine the degree to which a case involved substance abuse by counting the number of substance abuse annotations contained within its case notes. We could also indicate whether the substance abuse was being treated and whether we knew what kind of drug it was. For the case of court delays, we combined cases with potential delays and court patterns to determine if court-related delays were observed.

RESULTS

We extracted 1,174,989 case notes from SACWIS for calendar year 2011. DCFS experts assisted with coding 104 of these notes for substance abuse and court delay issues over the course of one day. We were prepared to do another round of coding, but found it to be unnecessary.

The table below contains descriptive results about the sample. Seventy-one percent of the children came from families who had substance abuse issues. There was no difference between Cook County (where Chicago is located) and the balance of the state. Also, 76 percent of white children and African American children came from families with substance abuse, while 67 percent of Hispanic children did. There was a slightly higher percentage of children who were placed in homes of relatives from families with substance abuse than regular foster care (72 percent versus 69 percent). As a child's age at first placement increases, the likelihood of their family having substance abuse issues

decreases from 78 percent of infants to 58 percent of youth placed between the ages of 13 and 17.

Relative to court delays, Cook County children had slightly fewer delays than children from the balance of the state (27 percent versus 29 percent). White and African American children were again equal (32 percent). Also, children with a lower age at placement had a greater rate of court delays with 35 percent in families with children placed as infants versus 19 percent with youth placed from 13 to 17 years of age.

Table 1. Existence of substance in families of foster children by race, type of placement, region, age and court activity

	No Substance Abuse (Percent)	Substance Abuse (Percent)	No Substance Abuse (N)	Substance Abuse (N)	Total (N)
All Cases	28.8%	71.2%	5,470	13,494	18964
Race/ethnicity					
White	23.8%	76.2%	1,766	5,646	7,412
Black	24.5%	75.5%	1,596	4,922	6,518
Hispanic	33.1%	66.9%	347	702	1,049
Other	44.2%	55.8%	1,761	2,224	3,985
Type of care					
Foster Care	30.5%	69.5%	1,454	3,312	4,766
HMR	28.1%	71.9%	2,580	6,612	9,192
Other	28.7%	71.3%	1,436	3,570	5,006
Region					
Non-Cook	28.6%	71.4%	4,009	10,000	14,009
Cook County	29.5%	70.5%	1,461	3,494	4,955

Age of child					
0	21.5%	78.5%	1082	3947	5029
1 to 5	26.2%	73.8%	1666	4696	6362
6 to 12	31.6%	68.4%	1381	2985	4366
13 to 17	41.8%	58.2%	1341	1866	3207
Court activity					
Court Delay	39.3%	60.7%	5,313	8,222	13,535
No Court Delay	2.9%	97.1%	157	5,272	5,429

JUVENILE JUSTICE EXAMPLE

Another example involves the analysis of police department "Reports of Investigation." The source of information in this case was 1,005 word processing documents. Although clearly a small number of documents, the methods used were similar to those used in the child welfare system example and could be applied to a much larger number of documents. These "Reports of Investigation" were completed by police officers investigating gang activity. In addition, particular sections were singled out for their importance. For this analysis, we used the "History of Gang Conflict" (HOGC) section that was present in 55 of the documents, filed over a span of approximately two years, from September 2010 to October 2012.

The HOGC section in a document lists a sequence of gang conflict incidents that are relevant to the document and are related to each other. The officer filing the document makes the decision to include the specific set of conflict incidents in a HOGC section based on any information to which they had access—formal or informal. The information was used to generate networks of individuals, based on co-occurrences.

DISCUSSION AND IMPLICATIONS

The ability to take advantage of new data sources to create new information has multiple implications for research on and evaluations of social programs. First, it provides rich information about the individuals involved and the services they receive *that is usually not available in structured data*. Second, it is a dynamic source of information that is collected in real time, is not reliant on long recall periods, and comes from data reporters that are reliable—in part because they are using the data themselves. Third, the number of topics that can be studied with these data are numerous, to say the least. As additional questions arise, these data can either be updated to capture new information. For example, after this initial research was concluded, the investigators re-used the dataset to investigate the topic of youth running away from their foster care placements.

Showing the promise of applying text mining to data collected by frontline practitioners also has implications for the design of future information systems, which may lead to reducing the burden on practitioners to spend precious time on data entry. If most data stored in information systems can be mined from a practitioner's dictation—even race, gender and birthdate—the time that a police officer, social worker, nurse, or teacher spends on data entry can be minimized.

While not all state or county systems will be able to produce extracts as easily as in Illinois, this is becoming more common. The juvenile justice example shows that individual documents are an acceptable data source for text mining. Clearly, the future of using public sector unstructured administrative data for research depends on having

subject matter experts, access, computational resources, the ability to extract the data and financial resources. But these hurdles are minor given that the richness and volume of text that is being collected in a daily basis.

References

Bass, S., Shields, M. K., & Behrman, R. E. (2004). Children, Families, and Foster Care: Analysis and Recommendations. *The Future of Children*, *14*(1), 4. Retrieved from https://www.questia.com/library/journal/1G1-115696228/children-families-and-foster-care-analysis-and

Cancian, M., Haveman, R. H., Meyer, D. R., & Wolfe, B. (2002). Before and After TANF: The Economic Well-Being of Women Leaving Welfare. *Social Service Review*, *76*(4), 603–641. doi:10.1086/342997

Coulton, C., Goerge, R. M., Putnam-Hornstein, E., & DeHaan, B. (2015). *Stepping up to harness big data for social good: A grand challenge for social work*. Baltimore, MD.

Courtney, M. E., Needell, B., & Wulczyn, F. (2004). Unintended consequences of the push for accountability: the case of national child welfare performance standards. *Children and Youth Services Review*, *26*(12), 1141–1154. doi:10.1016/j.childyouth.2004.05.005

Evans, J. A., & Rzhetsky, A. (2011). Advancing science through mining libraries, ontologies, and communities. *The Journal of Biological Chemistry*, *286*(27), 23659–66. doi:10.1074/jbc.R110.176370

Fluke, J., Edwards, M., Kutzler, P., Kuna, J., & Tooman, G. (2000). Safety, permanency, and in-home services: Applying administrative data. *CHILD WELFARE*, *79*(5), 573–595. Retrieved from http://apps.webofknowledge.com.proxy.uchicago.edu/full_record.do?product=WOS&search_mode=GeneralSearch&qid=32&SID=2BQfSWRMnbbCsoB2XvW&page=1&doc=3

Goerge, R. M. (1994). The effect of public child welfare worker characteristics and turnover on discharge from foster care. *Child Welfare Research Review*, *1*, 205–217.

Goerge, R, Fanshel, D. Wulczyn, F. (1994). Foster Care Research Agenda for the ’90s. *Journal of the Child Welfare League of America*, *73*(5), 525 – 549.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*, 267–297. doi:10.1093/pan/mps028

Hindman, M. (2015). Building Better Models: Prediction, Replication, and Machine Learning in the Social Sciences. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 48–62. doi:10.1177/0002716215570279

Hotz, V. J., Goerge, R., Balzekas, J., & Margolin, F. (1998). Administrative data for policy-relevant research: Assessment of current utility and recommendations for development. *Report of the Advisory Panel on Research Uses of Administrative Data of the Northwestern University/University of Chicago Joint Center for Poverty Research*.

Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, *13*(June), 395–405. doi:10.1038/nrg3208

Lane, J. (2006). Integrated Employer-Employee Data: New Resources for Regional Data Analysis. *International Regional Science Review*, *29*(July), 264–277. doi:10.1177/0160017606289897

NSCAW. (2007). NSCAW Report # 16: A summary of NSCAW findings, (16), 6–9.

Ogren, P. V., & Bethard, S. J. (2009). Building test suites for UIMA components. In *Proceedings of the Workshop on Software Engineering, Testing, and Quality Assurance for Natural Language Processing (SETQA-NLP 2009)* (Vol. Proceeding, pp. 1–4). Boulder, CO: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1621947.1621948

Planning, O., & Children, A. (2012). Child Well-Being Spotlight, 2012.

Roos, L. L., Nicol, J. P., & Cageorge, S. M. (1987). Using administrative data for longitudinal research: comparisons with primary data collection. *Journal of Chronic*

Diseases, *40*(1), 41–9. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3805233

Smithgall, C., Jarpe-Ratner, E., Gnedko-Berry, N., & Mason, S. (2015). Developing and testing a framework for evaluating the quality of comprehensive family assessment in child welfare. *Child Abuse & Neglect*. doi:10.1016/j.chiabu.2014.12.001

Suominen, H. (2014). Text mining and information analysis of health documents. *Artificial Intelligence in Medicine*, *61*(3), 127–30. doi:10.1016/j.artmed.2014.06.001