Research and Applications

Impact of possible errors in natural language processing-derived data on downstream epidemiologic analysis

Zhou Lan, PhD¹,² and Alexander Turchin MD, MS³,⁴

¹Center for Clinical Investigation, Brigham & Women’s Hospital, Boston, MA 02115, United States; ²Harvard Medical School, Boston, MA 02115, United States; ³Division of Endocrinology, Brigham & Women’s Hospital, Boston, MA 02115, United States

Abstract

Objective: To assess the impact of potential errors in natural language processing (NLP) on the results of epidemiologic studies.

Materials and Methods: We utilized data from three outcomes research studies where the primary predictor variable was generated using NLP. For each of these studies, Monte Carlo simulations were applied to generate datasets simulating potential errors in NLP-derived variables. We subsequently fit the original regression models to these partially simulated datasets and compared the distribution of coefficient estimates to the original study results.

Results: Among the four models evaluated, the mean change in the point estimate of the relationship between the predictor variable and the outcome ranged from −21.9% to 41.2%. In three of the four models, significance of this relationship was not eliminated in a single of the 500 simulations, and in one model it was eliminated in 12% of simulations. Mean changes in the estimates for confounder variables ranged from 0.27% to 2.27% and significance of the relationship was eliminated between 0% and 9.25% of the time. No variables underwent a shift in the direction of its interpretation.

Discussion: Impact of simulated NLP errors on the results of epidemiologic studies was modest, with only small changes in effect estimates and no changes in the interpretation of the findings (direction and significance of association with the outcome) for either the NLP-generated variables or other variables in the models.

Conclusion: NLP errors are unlikely to affect the results of studies that use NLP as the source of data.

Introduction

Widespread adoption of electronic health records (EHRs) has led to an exponential growth of healthcare data. With this growth come opportunities to leverage these immense amounts of information to streamline processes, decrease costs, discover new treatments, and improve patient outcomes. One distinct component of EHR data is narrative documents. This important data category, which can include clinicians’ notes, radiology reports, and operative summaries, offers richness of detail, nuance of medical uncertainty, and the logic of clinicians’ decision-making.

Modern artificial intelligence (AI) technologies for analysis of text, collectively known as natural language processing (NLP), provide a key to unlock this treasure trove of information. Information supplied by NLP tools has been used in a broad range of medical data analyses, including predictive modeling, analyses of social determinants of health, and health outcomes research, and has often provided unique perspectives not available from other data sources.

An important aspect of NLP-based information extraction is that, unlike for structured data, the process of translating the raw data into discreet analytical fields is not usually perfectly accurate. Recall and precision of state-of-the-art NLP systems vary widely depending on the application, often falling in the range between 60% and 80%. The effect of these errors on the results of clinical investigations that utilize NLP-generated data is not known. We therefore conducted a study of possible effects of NLP errors on the results of multivariable regression models in epidemiologic clinical research.
Methods

Study design
We conducted a Monte Carlo simulation analysis based on data from three previous epidemiologic studies to investigate the impact of possible errors in NLP-derived data on the results of multivariable regression models.

Study settings
All studies included in the analysis were conducted at Mass General Brigham (MGB), a large integrated healthcare delivery network in Massachusetts, which includes several community and specialty hospitals and affiliated outpatient primary care and specialty practices. MGB has maintained a comprehensive EHR system that incorporates a wide range of structured and narrative data sources since 2000. The source of data for all studies used in this analysis was the MGB EHR.

Contributing studies
The present study used data from three previous studies: Turchin et al., Chang et al., and Brown et al. (Table 1). Turchin et al. investigated the association between non-acceptance (decline) of insulin therapy recommendation by patients and time to glycemic control (defined as HbA1c < 7.0%). The primary predictor variable in this study, non-acceptance of insulin therapy recommendation, was identified using NLP analysis of EHR provider notes. The multivariable analysis from Turchin et al. included in the present study used a propensity-weighted Cox proportional hazards model, which adjusted for socio-demographic characteristics, health status, and treatment. Chang et al. explored the relationship between patient-provider discussions of bariatric surgery and subsequent weight changes and receipt of bariatric surgery. The primary predictor variable in this study, documentation of bariatric surgery discussions in EHR provider notes, was identified using NLP analysis. Two multivariable analyses from Chang et al. were included in the present study: a hierarchical mixed linear regression model with random intercepts, which investigated the association between bariatric surgery discussion and changes in BMI (primary analysis in the original study); and a logistic regression model, which analyzed the relationship between bariatric surgery discussion and receipt of bariatric surgery (a secondary analysis in the original study). Both models adjusted for socio-demographic characteristics and relevant comorbidities. Brown et al. explored the relationship between non-acceptance of statin therapy recommendation by patients and time to achieving low-density lipoprotein cholesterol (LDL) levels < 100 mg/dL. The multivariable analysis from Brown et al. included in the present study used a Cox proportional hazards model, which adjusted for socio-demographic characteristics and relevant comorbidities. All NLP tools used to generate data for the original studies were developed using the Canary platform, publicly available at http://canary.bwh.harvard.edu/. Recall of the NLP tools in these studies ranged from 0.88 to 1.0 and precision from 0.76 to 0.95 (Table 1).

Semi-synthetic data-based evaluation
To provide a concise illustration, we used the term “original data” to refer to the data used in one of the three contributing studies. Similarly, the terms “NLP-derived variable” and “outcome” referred to their corresponding terms in one of the three studies. Accuracy of the NLP-derived variable is typically not perfectly accurate. Positive calls may contain false positives (true negatives), and negative calls may contain false negatives (true positives). In real settings, the true underlying values (i.e., not NLP-derived) cannot be accessed. To assess the impact of NLP-derived data on downstream analyses, we simulated the underlying true values using the following approach. Since the positive predictive value (PPV) and negative predictive value (NPV) are defined as PPV = (number of true positives)/(number of positive calls) and NPV = (number of true negatives)/(number of negative calls), respectively, we performed the following two manipulations for each simulation: a) Among the NLP-derived positive calls, we randomly selected N_P × (1 − PPV) calls and changed them to negative calls, where N_P is the number of positive calls; b) Among NLP-derived negative calls, we randomly selected N_N × (1 − NPV) calls and changed them to positive calls, where N_N is the number of negative calls (Figure 1). Following the spirit of Monte Carlo simulation, we performed 500 simulations to mimic all possible scenarios. Since all other variables were kept constant, we referred to the data after manipulation as “semi-synthetic data.” The values of PPV and NPV used for data generation are provided in Table 1. For each of the three studies, we combined the synthetic NLP-derived predictor variable data with the original confounder data (i.e., socio-demographic characteristics and relevant comorbidities) and fit the resulting dataset to the corresponding statistical model. For Cox regression, we used

Table 1. Studies contributing data to the analysis.

<table>
<thead>
<tr>
<th>Reference</th>
<th>NLP-derived medical variable</th>
<th>Outcome</th>
<th>Statistical model</th>
<th>Positive predictive value (PPV)</th>
<th>Negative predictive value (NPV)</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turchin et al.</td>
<td>Non-acceptance of insulin therapy by patients</td>
<td>Time to HbA1c &lt; 7.0%</td>
<td>Propensity score weighted Cox model</td>
<td>0.95</td>
<td>0.99</td>
<td>1.00</td>
</tr>
<tr>
<td>Chang et al.</td>
<td>Patient-provider discussions of bariatric surgery</td>
<td>Change in BMI</td>
<td>Linear mixed model</td>
<td>0.76</td>
<td>0.99</td>
<td>0.89</td>
</tr>
<tr>
<td>Brown et al.</td>
<td>Non-acceptance of statin therapy by patients</td>
<td>Receipt of bariatric surgery</td>
<td>Logistic regression</td>
<td>0.78</td>
<td>0.99</td>
<td>0.88</td>
</tr>
</tbody>
</table>

The table provides the reference of the contributing studies and their key information inclusive of NLP-derived medical variable, outcome, statistical model, PPV, NPV, and sensitivity.
the function `coxph()` in the R package “survival”\(^{18}\) to generate hazard ratios for each variable. For the logistic regression, we used the function `glm()` in the R package “stats” to generate odds ratios for each variable. For the linear mixed model, we used the function `lmer()` in the R package “lme4”\(^{19}\) to generate fixed effect estimates for each variable. Using the R package “performance,”\(^{20}\) we have checked the model assumption of each run of the simulation study. The simulations that did not follow the model assumptions have been excluded from the analysis.

We used boxplots to visualize the above estimates and compare them to the ones obtained from the original data. The point estimation comparison was used for this measurement. For each variable and model, we also measured the probability (among the 500 simulations) of the estimate changing its direction (sign) and arising or disappearing statistical significance (\(P\)-value < .05) of the relationship between the variable and the model outcome.

**Results**

**Impact on the relationship between the NLP-generated variable and study outcome**

The impact of possible NLP errors on the relationship between the NLP-derived variable and study outcome is summarized in Table 2. In Table 2, the mean % change of the point estimate provided the percentage change between the original estimate and the point estimate for the 500 simulations. Among the four analyses of the relationships the percentage change between the original estimate and the point estimate over all simulations and covariates for the 500 simulations ranged from \(-21.9\%\) to \(4.12\%\) (Table 2). Given that the NLP-derived variables had significant effects on the outcome, we reported the probability that the significance was eliminated among the simulations (Table 2). In one of the four analyses 12% of the 500 simulations showed the relationship between the NLP-derived variable and the study outcome lose its statistical significance; in the other three analyses the relationship maintained its statistical significance (\(P\)-value < .05) of the relationship between the variable and the model outcome.

![Figure 1. Illustration of generation of semi-synthetic data.](image)

**Table 2. Impact of possible NLP errors on the relationship between the NLP-derived variable and study outcome.**

<table>
<thead>
<tr>
<th>Original study</th>
<th>Outcome</th>
<th>Original point estimate</th>
<th>Mean % change of the point estimate (IRQ)</th>
<th>Original (P)-value</th>
<th>Significance eliminated, mean % simulations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turchin et al.(^{14})</td>
<td>Time to HbA1c &lt;7.0%</td>
<td>HR: 0.90</td>
<td>0.61 (0.46, 1.68)</td>
<td>.0122</td>
<td>12</td>
</tr>
<tr>
<td>Chang et al.(^{15})</td>
<td>Change in BMI</td>
<td>Fixed effect: –1.43</td>
<td>(-4.03 (–7.22, –0.714))</td>
<td>&lt;.001</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Receipt of bariatric surgery</td>
<td>OR: 10.23</td>
<td>(-21.90 (–24.3, –19.8))</td>
<td>&lt;.001</td>
<td>0</td>
</tr>
<tr>
<td>Brown et al.(^{16})</td>
<td>Time to LDL &lt;100 mg/dL</td>
<td>HR: 0.57</td>
<td>4.12 (3.36, 4.99)</td>
<td>&lt;.001</td>
<td>0</td>
</tr>
</tbody>
</table>

**Discussion**

In this simulation study of several NLP-based multivariable analyses, the impact of possible NLP errors on the regression model results was modest. Associations of all of the NLP-derived variables with the corresponding study outcomes...
retained statistical significance in all or overwhelming majority of the simulations, and median difference in point estimates did not exceed 25%.

This finding has significant implications for epidemiological studies and outcomes research that use NLP-derived data. Human language is incredibly rich, and it is simply not

---

### Table 3. Impact of possible NLP errors on the relationships between other confounder variables and study outcome.

<table>
<thead>
<tr>
<th>Original study</th>
<th>Outcome</th>
<th>Confounders</th>
<th>Change direction</th>
<th>Mean % change of the point estimate (IQR)</th>
<th>Significance acquired, mean simulations</th>
<th>Significance eliminated, mean simulations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turchin et al.</td>
<td>Time to HbA1c &lt;7.0%</td>
<td>14</td>
<td>0</td>
<td>0.039% (−0.004%, 0.01%)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Chang et al.</td>
<td>Change in BMI</td>
<td>14</td>
<td>0</td>
<td>2.27% (−2.331%, 4.237%)</td>
<td>1.37</td>
<td>0</td>
</tr>
<tr>
<td>Brown et al.</td>
<td>Receipt of bariatric surgery</td>
<td>14</td>
<td>0</td>
<td>0.72% (−0.11%, 2.05%)</td>
<td>1.51</td>
<td>9.25</td>
</tr>
<tr>
<td>Brown et al.</td>
<td>Time to LDL &lt;100 mg/dL</td>
<td>23</td>
<td>0</td>
<td>0.27% (−0.024%, 0.603%)</td>
<td>0.63</td>
<td>0.96</td>
</tr>
</tbody>
</table>

---

**Figure 2.** Comparison of point estimates of hazard ratios for the relationship with time to HbA1c < 7.0% in the simulations of possible NLP errors vs the original study in Turchin et al. Dashed horizontal lines represent the original point estimate. Solid horizontal lines represent in the boxplot the median point estimate for the 500 simulations. Shaded area represents the interquartile range for the 500 simulations. Vertical line and the shaded circles represent the entire range of the estimates for the 500 simulations.
feasible to program—whether by hand or through machine learning methods—all possible permutations that can carry a particular meaning. Furthermore, language is often ambiguous and therefore perfectly accurate interpretation may not even theoretically be feasible. If these inaccuracies translate into dramatically different results in downstream analyses that utilize NLP-derived data, then reliability of the corresponding study results and their implications for patient care would be called into question. The results of this study reassuringly demonstrate that the impact of NLP errors on the results of regression models widely used in epidemiological studies is limited. These results were independent of the regression model used and included some of the most commonly used models—logistic regression, linear regression, mixed models, and Cox proportional hazards models. They will therefore generalize on a large number of observational studies that utilize NLP-derived data.

The role of NLP in epidemiologic and observational studies continues to grow. It has been used in studies across a broad swathe of clinical fields, ranging from radiology to surgery to multiple subspecialties of internal medicine. As the amount of available data continues to explode and new and better technologies are developed, its importance is set to continue to increase. The finding of the present study that errors in NLP-based data acquisition have only a very limited impact on downstream regression analyses removes another important barrier to acceptance of NLP as an essential data source for clinical research.

Figure 3. Comparison of point estimates of (A) regression coefficients for the relationship with BMI change and (B) odds ratios for the relationship with receipt of bariatric surgery in the simulations of possible NLP errors vs the original study in Chang et al. Dashed horizontal lines represent the original point estimate. Solid horizontal lines represent in the boxplot the median point estimate for the 500 simulations. Shaded area represents the interquartile range for the 500 simulations. Vertical line and the shaded circles represent the entire range of the estimates for the 500 simulations.
It is important to note that the present analysis was limited to regression models, where statistical significance (or lack thereof) of the relationship between predictor/confounder variables and the outcome is the most critical finding. Other applications of NLP-derived data where the precise value of the point estimate is essential, such as predictive modeling, may be more strongly affected by NLP errors. For these potentially more error-sensitive applications it would be important to conduct sensitivity analyses similar to this study that would evaluate the potential impact of NLP errors and allow the readers to judge the reliability and calibration of the model.

Monte Carlo simulation methods are widely used in the community of statistics and machine learning to simulate situations that are impractical or impossible to replicate in real-world settings. In medical EHR research, for example, the underlying medical condition may not be readily discernible from the available medical reports. While Monte Carlo simulation has been employed in some medical scenarios, few studies have investigated the impact of possible errors in NLP-derived data on downstream epidemiologic analysis using this approach. Our paper, together with Brown et al., may be among the first papers to systematically examine this issue.

State-of-the-art NLP systems typically achieve recall and precision rates in the range between 60% and 80%. However, our evaluation results suggested that the impact of possible false positives on the model estimates of all variables in the model was minor. For most of the non-NLP-derived variables, their robustness may be attributed to their strong marginal effects. For most of the NLP-derived variables, their robustness may be explained by the existing correlations, which are unaffected by changes in possible false positives.

![Figure 3](https://academic.oup.com/jamiaopen/article/6/4/ooad111/7502595 by guest on 15 April 2024)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at study entry</td>
<td>0.980-0.977</td>
</tr>
<tr>
<td>BMI-Baseline</td>
<td>1.105-1.095</td>
</tr>
<tr>
<td>Charlson Co-morbidity Index</td>
<td>1.055-1.045</td>
</tr>
<tr>
<td>Coronary artery disease</td>
<td>1.35-1.30</td>
</tr>
<tr>
<td>Diabetes</td>
<td>1.300-1.275</td>
</tr>
<tr>
<td>English as Primary Language</td>
<td>1.055-1.025</td>
</tr>
<tr>
<td>Female</td>
<td>1.71-1.68</td>
</tr>
<tr>
<td>Hypertension</td>
<td>1.12-1.08</td>
</tr>
<tr>
<td>Married</td>
<td>1.23-1.21</td>
</tr>
<tr>
<td>Median household income by zip code ($1,000)</td>
<td>0.99990-0.99875</td>
</tr>
<tr>
<td>Nongovernment insurance</td>
<td>1.00-0.98</td>
</tr>
<tr>
<td>Race-White</td>
<td>1.17-1.15</td>
</tr>
<tr>
<td>Receiving bariatric surgery discussion within 1 year of study entry</td>
<td>1.19-1.19</td>
</tr>
<tr>
<td>Smoking</td>
<td>1.15-1.13</td>
</tr>
<tr>
<td>Stroke</td>
<td>1.3-1.3</td>
</tr>
</tbody>
</table>
The present study had a number of strengths. It included a broad range of regression models, from logistic to linear regression to Cox proportional hazards. The study analyzed the effects of two types of errors—false positives and false negatives—in combination, as they would be expected to occur in real data. The large number of simulations utilized in the study and large datasets evaluated—tens of thousands of patients—increase reliability of our findings. Finally, the study evaluated the impact of possible NLP errors on the relationships of a broad range of variables, including demographics, comorbidities, vital signs and laboratory measurements with several different types of outcomes (laboratory tests, body measurements and receipt of treatment), increasing its generalizability.

The findings of the present study should be interpreted in the light of its limitations. Data for the analysis originated in a single integrated healthcare system in eastern Massachusetts, and therefore the findings may not be applicable to other settings. The range and types of errors simulated might not fully capture the complexity and variability of errors that occur in real-world NLP applications; errors could also vary between patient subgroups. This could limit the generalizability of the findings. Only four types of regression models were included in the analysis, and therefore the impact of possible NLP errors on other types of studies that use NLP data, such as predictive models or sentiment analysis, should be analyzed in future research. Machine learning-based NLP methods were not included in this evaluation. However, since the nature of NLP errors—false positives and false negatives—is the same irrespective of the specific NLP technique—our findings would be expected to generalize to all NLP methodologies. Finally, the

Figure 4. Comparison of point estimates of hazard ratios for the relationship with time to LDL <100 mg/dL in the simulations of possible NLP errors vs the original study in Brown et al. Dashed horizontal lines represent the original point estimate. Solid horizontal lines represent in the boxplot the median point estimate for the 500 simulations. Shaded area represents the interquartile range for the 500 simulations. Vertical line and the shaded circles represent the entire range of the estimates for the 500 simulations.
studies included in this analysis focused on short to intermediate-range outcomes. Further investigations should be conducted to determine whether the impact of NLP errors could be different in analyses of epidemiology of long-term clinical outcomes.

Conclusion
In this study that evaluated effects of NLP accuracy on a range of regression models in several real-world evidence studies, we have found that the impact of possible NLP errors on the point estimates and their statistical significance in regression models was modest. This finding supports broad use of NLP in real-world evidence investigations that utilize regression models. Further research should be conducted to assess possible impact of NLP errors on other analytical methods.

Author contributions
ZL and AT contributed to the conception and design of the work and acquisition of data. ZL analyzed the data and drafted the manuscript. AT revised the manuscript, obtained funding, and supervised the work. All authors approved the final manuscript and agree to be accountable for all aspects of the work.

Funding
This study was funded in part by Patient-Centered Outcomes Research Institute (ME-2019C1-15328).

Conflicts of interest
A.T. reports research support from Eli Lilly and Novo Nordisk and consulting for Novo Nordisk and Proteomics International. Z.L. reports no conflicts of interest.

Data availability
This repository at https://github.com/lanzhouBWH/NLP-Downstream-Impact provides R codes (.R) and job scheduler codes (.lsf) for the Monte Carlo experiments to analyze Impact of Possible Errors in NLP-Derived Data on Downstream Epidemiologic Analysis.

References