



# HHS Public Access

Author manuscript

*Epilepsia*. Author manuscript; available in PMC 2020 September 01.

Published in final edited form as:

*Epilepsia*. 2019 September ; 60(9): e93–e98. doi:10.1111/epi.16320.

## Investigation of Bias in an Epilepsy Machine Learning Algorithm Trained on Physician Notes

Benjamin D. Wissel, BS<sup>1</sup>, Hansel M. Greiner, MD<sup>2,3</sup>, Tracy A. Glauser, MD<sup>2,3</sup>, Francesco T. Mangano, DO<sup>3,4</sup>, Daniel Santel, PhD<sup>1</sup>, John P. Pestian, PhD, MBA<sup>1,2</sup>, Rhonda D. Szczesniak, PhD<sup>2,5</sup>, Judith W. Dexheimer, PhD<sup>1,2,6</sup>

<sup>1</sup>Department of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

<sup>2</sup>Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, USA

<sup>3</sup>Division of Neurology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

<sup>4</sup>Division of Neurosurgery, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

<sup>5</sup>Division of Biostatistics & Epidemiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

<sup>6</sup>Division of Emergency Medicine, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

### SUMMARY

Racial disparities in the utilization of epilepsy surgery are well documented, but it is unknown if a natural language processing (NLP) algorithm trained on physician notes would produce biased recommendations for epilepsy presurgical evaluations. To assess this, an NLP algorithm was trained to identify potential surgical candidates using 1,097 notes from 175 epilepsy patients with a history of resective epilepsy surgery and 268 patients who achieved seizure freedom without surgery (total N = 443 patients). The model was tested on 8,340 notes from 3,776 patients with epilepsy whose surgical candidacy status was unknown (2,029 male, 1,747 female; median age, 9 years; age range, 0-60 years). Multiple linear regression using demographic variables as covariates was used to test for correlations between patient race and surgical candidacy scores. After accounting for other demographic and socioeconomic variables, patient race, gender, and primary

**All correspondence to:** Judith Dexheimer, PhD, Department of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, MLC 2008, 3333 Burnet Avenue, Cincinnati, OH 45229-3039, USA, Phone: 513-803-2962, Fax: 513-803-2581, Judith.Dexheimer@cchmc.org.

#### AUTHOR CONTRIBUTIONS

B.D.W., H.M.G., J.P.P., and J.W.D. contributed to the conception and design of this study. D.S., J.P.P., and J.W.D. contributed to the acquisition of the data. B.D.W. and R.D.S. performed statistical analyses and interpreted results. B.D.W. and J.W.D. contributed to drafting of the manuscript text. All authors approved the final version of the manuscript.

#### CONFLICTS OF INTEREST

None of the authors report a potential conflict of interest to this study. J.P.P., T.A.G., and H.M.G. report a patent pending on the identification of surgical candidates using natural language processing, licensed to Cincinnati Children's Hospital Medical Center (CCHMC). J.P.P. and T.A.G. report a patent pending on processing clinical text with domain specific spreading activation methods, also licensed to CCHMC.

#### DISCLOSURE

We confirm that we have read the Journal's position on issues involved in ethical publication and affirm that this report is consistent with those guidelines.

language did not influence surgical candidacy scores ( $p > 0.35$  for all). Higher scores were given to patients  $> 18$  years-old who traveled farther to receive care, and those who had a higher family income and public insurance ( $p < 0.001, 0.001, 0.001, \text{ and } 0.01$ , respectively). Demographic effects on surgical candidacy scores appeared to reflect patterns in patient referrals.

### Keywords

Epilepsy surgery; machine learning; natural language processing; clinical decision support

---

## INTRODUCTION

Natural language processing (NLP) algorithms have been reported to incorporate inherent bias when trained on human language.<sup>1, 2</sup> NLP techniques such as word embedding are now used to objectively evaluate gender and ethnic stereotypes in text data.<sup>3</sup> In recent years, there has been unfortunate examples of non-medical NLP and machine learning algorithms that have produced biased recommendations.<sup>4</sup> These setbacks risk jeopardizing physician trust in machine learning-based clinical decision support tools.<sup>5, 6</sup>

The utilization of resective epilepsy surgery varies by patient race, language, insurance status, and distance from specialized epilepsy centers.<sup>7-11</sup> National hospital and billing databases, including the Nationwide Inpatient Sample, were used in many of these studies.<sup>7, 10, 11</sup> It is unknown if free-text notes in electronic health records (EHR) reflect these disparities, or if clinical decision support systems that utilize this information could be biased.

NLP can be used to evaluate epilepsy notes.<sup>12</sup> We developed an NLP algorithm to assign surgical candidacy scores to patients based only on provider EHR notes.<sup>13, 14</sup> This algorithm was incorporated into a pediatric hospital's EHR and is used to alert neurologists when they are scheduled to see a potential candidate for resective epilepsy surgery. Here we conducted a cross-sectional analysis of the surgical candidacy scores to determine whether this algorithm's scores are impacted by patient demographics.

## METHODS

### Algorithm Design and Development

We developed an NLP algorithm to automatically identify patients with epilepsy who may benefit from resection surgery.<sup>13, 14</sup> In prior studies, the NLP algorithm's recommendations were compared to three epileptologists' recommendations (gold standard) to determine how accurately, and how early in the disease course, it could identify candidates for epilepsy surgery.<sup>14</sup> The algorithm was able to achieve equal classification accuracy as the three epileptologists (F1-score = 0.82).<sup>14</sup> The algorithm could accurately classify patients' surgical candidacy using notes from two years before they were referred for a presurgical evaluation.<sup>14</sup>

The algorithm was incorporated into a pediatric hospital's EHR and sends weekly automated alerts to neurology providers when potential candidates for resective epilepsy surgery have

an upcoming visit that week. The system architecture is depicted in Figure 1. The NLP algorithm is trained each week to discriminate between patients with epilepsy (ICD-9 codes 345.\*, 780.3\*, and 779.0) who underwent resective surgery (procedure codes 61510, 61531, 61533–61540, 61542, 61543, 61566, and 61567) and patients who became seizure-free using pharmacotherapy alone. No structured demographic information is included.

The only input to the algorithm is free-text neurology progress notes. Magnetic resonance imaging, electroencephalogram, and genetic reports are not included. Semantic features are extracted using uni-, bi-, and tri-grams. This enables patients with similar language in their notes to be grouped together. Words and phrases are tokenized, normalized to lower case, and stop words are removed.<sup>15</sup> N-grams commonly found in the notes of patients who underwent resective epilepsy surgery are weighted positively, while n-grams more commonly found in the notes of patients who became seizure-free are weighted negatively. Binary normalized n-gram features are rank-ordered according to their discriminatory value using a Kolmogorov–Smirnov test. A support vector machine classifier with a linear or radial basis function kernel assigns a surgical candidacy score to each patient. Scores are evenly distributed and centered around zero, with lower (more negative) scores indicating a higher likelihood of achieving seizure freedom and higher (more positive) scores indicating a higher likelihood that a patient is a candidate for epilepsy surgery. Hyperparameters (the number of rank-ordered features included in the classifier, C, gamma, and kernel type) are selected using 3-fold inner cross-validation.

### Study Design and Participants

The algorithm is used to assign surgical candidacy scores to all patients whose surgical candidacy status is “unknown”. Unknown patients are those who: 1) had an epileptic seizure within the last year (determined using structured and unstructured data in the EHR); 2) have an outpatient neurology visit in the next six months; and 3) have no history of epilepsy surgery or presurgical evaluation. The unknown patients are not included in the algorithm’s training. Eligible patients receive a new surgical candidacy score each week, after the algorithm is re-trained using an updated training set. Patients’ surgical candidacy scores from one randomly chosen week were extracted to test for associations between patient demographics and the NLP-derived surgical candidacy scores. All demographic and socioeconomic variables were extracted separately from the EHR.

### Evaluation of the Algorithm’s Performance During Training and Validation

Receiver operating characteristic curves for the cross-validation were plotted by adjusting the score threshold for surgical candidacy. Sensitivity and specificity were calculated by comparing the algorithm’s scores to known patient outcomes (surgery or seizure freedom). The algorithm’s performance was assessed as the average of sensitivity, specificity, and area under the curve (AUC) obtained from 10-fold cross-validation.

### Statistical Analysis

We performed individual and multiple linear regression using demographic variables as covariates of the NLP’s surgical candidacy scores. All variables shown in the Table were used as covariates in the adjusted regression. Individual and adjusted effect sizes, 95%

confidence intervals (CI), and corresponding p-values were calculated using the 'lm' function in R.<sup>16</sup>  $P < 0.05$  was considered statistically significant.

## RESULTS

Surgical candidacy scores from the week of November 4<sup>th</sup>, 2018 were evaluated. The NLP algorithm was trained on 1,097 notes from 443 patients with epilepsy who were either seizure-free ( $N = 268$ ) or had prior resective surgery ( $N = 175$ ). The algorithm's sensitivity during the 10-fold cross-validation was 0.88 (95% CI: 0.86 to 0.90), specificity was 0.91 (95% CI: 0.87 to 0.95), and AUC was 0.94 (95% CI: 0.92 to 0.96). The trained algorithm was used to evaluate 8,340 notes from 3,776 patients with "unknown" surgical candidacy status. Patient demographics, socioeconomic characteristics, and surgical candidacy scores are shown in the Table. The median (range) surgical candidacy score was  $-0.17$  ( $-1.87$  to  $2.25$ ).

After adjusting for demographic and socioeconomic variables, patient race, gender, and primary language did not influence surgical candidacy scores ( $p = 0.36$  to  $0.72$ ). Patients traveling farther to receive specialty care received higher scores ( $p < 0.001$ ). Patients with public insurance, higher median household income (by zip code), and age  $> 18$  years old received higher surgical candidacy scores ( $p < 0.01$ ,  $0.001$  and  $0.001$ , respectively). Median household income was negatively correlated with distance traveled to receive care (Spearman's  $\rho = -0.35$ ;  $p < 0.001$ ).

## DISCUSSION

The NLP algorithm was trained on free-text physicians notes. High and low surgical candidacy scores were able to identify probable surgical and seizure-free patients, respectively, with a high AUC. We evaluated whether the surgical candidacy scores were influenced by patient demographics. After adjusting for other demographic and socioeconomic variables, patient race, gender, and language spoken did not influence surgical candidacy scores. Higher surgical candidacy scores were given for patients who traveled from outside of the local catchment area, patients who continued care past their 18<sup>th</sup> birthday, and patients with higher family incomes and public insurance. These results appear to be influenced by referral patterns, rather than inherent bias in the model or provider documentation. These unbiased surgical candidacy scores can be used by clinicians as a tool to supplement clinical reasoning. Sharing these scores with clinicians in real time may ultimately facilitate earlier referrals.

African-American race received lower scores than white patients in the unadjusted regression. However, it appeared that race acted as a proxy for other social determinants of health, such as location of residence and family income. After considering these other factors, African-American patients scored similarly to whites, suggesting the algorithm was not racially biased.

Patients traveling farther to the hospital received higher scores, further attenuated after correcting for other demographic and socioeconomic variables. Patients with more severe epilepsy may have been more motivated to travel farther for specialized treatment.

Surprisingly, distance travelled to receive care was negatively correlated with household income. This suggests that their effects on increasing surgical candidacy scores were independent.

Patients over 18 years-old had higher scores. Although this is a majority pediatric (age 0–18 years-old) sample, a substantial minority (20%) of patients analyzed were over 18 years of age. In this center, there are several large multidisciplinary subspecialty practices, including most notably the tuberous sclerosis program, caring for adults with diagnoses that have a high rate of epilepsy. Additionally, many patients transitioning from pediatric to adult care choose to do so after completing undergraduate studies, in their 20s. In addition to differing epilepsy phenotypes, it is possible that higher scores were given to older patients if they delayed surgery for social reasons. The decision to be evaluated for surgical treatment may not have been considered by these patients' parents until their child was old enough to decide for him/herself.

Gender was not associated with differing rates of surgical utilization in other studies,<sup>7, 10</sup> and these findings were reproduced here. A retrospective cohort study reported lower surgical utilization among patients with lower English proficiency.<sup>8</sup> We did not find differences in NLP scores for patients whose first language was not English, suggesting that language's influence on surgical utilization did not influence provider documentation.

Patients with public insurance received higher surgical candidacy scores. This is in contrast to other studies that reported patients with public insurance receive resection surgery less often than those with private insurance.<sup>7, 10</sup> Higher NLP scores observed here suggests that these patients had a higher seizure burden. This apparent discrepancy may be explained by evidence that suggests patients with public insurance face higher barriers to specialty care.<sup>17</sup> Once at a specialty center, these patients appeared to be better surgical candidates.

These findings should be interpreted within the context of the study's limitations. There were fewer non-white than white patients in this large cohort. However, this cohort's demographic makeup was roughly consistent with the average across the United States.<sup>18</sup> Second, the algorithm's scores were not exact. The probability that a known surgical candidate scored higher than a non-surgical candidate was only 94% (AUC = 0.94). Third, this algorithm was not been tested outside of a tertiary center. The generalizability of the algorithm to non-specialty or adult centers outside of Cincinnati remains to be determined.

## Conclusions

Collectively, these results demonstrated that epilepsy surgery candidacy scores from our NLP algorithm were not biased by patient demographics. NLP surgical candidacy scores in this large, diverse cohort of pediatric epilepsy patients reflect relationships between tertiary center referral patterns and patient location, insurance, age, and income.

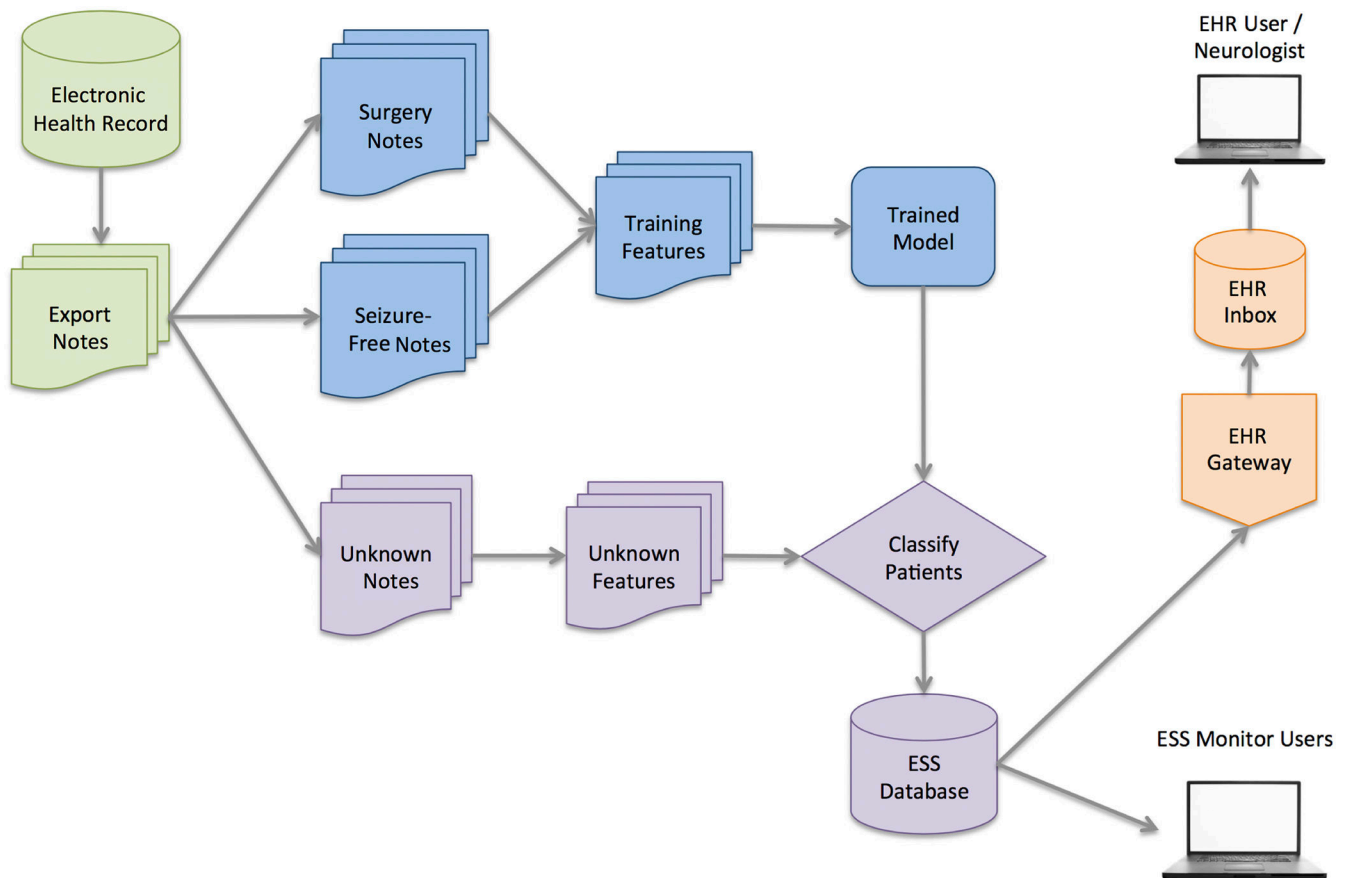
## ACKNOWLEDGEMENTS

We thank the Cincinnati Children's Research Foundation, the University of Cincinnati Biomedical Informatics Graduate Program, and the providers who gave excellent care to the patients included in this study.

This study was funded by a grant from the Agency for Healthcare Research and Quality (AHRQ 1 R21 HS024977–01). Effort by RS was supported in part by the National Institutes of Health (K25 HL125954).

## REFERENCES

1. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. *Science* 2017;356:183–186. [PubMed: 28408601]
2. Zou J, Schiebinger L. AI can be sexist and racist—it’s time to make it fair. Nature Publishing Group; 2018.
3. Garg N, Schiebinger L, Jurafsky D, et al. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc Natl Acad Sci* 2018;115:E3635. [PubMed: 29615513]
4. Dastin J Amazon scraps secret AI recruiting tool that showed bias against women, 10 9, 2018 Available at: <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>. Accessed November 8, 2018.
5. Verghese A, Shah NH, Harrington RA. What this computer needs is a physician: Humanism and artificial intelligence. *JAMA* 2018;319:19–20. [PubMed: 29261830]
6. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. *N Engl J Med* 2019;380:1347–1358. [PubMed: 30943338]
7. McClelland S 3rd, Guo H, Okuyemi KS. Racial disparities in the surgical management of intractable temporal lobe epilepsy in the United States: a population-based analysis. *Arch Neurol* 2010;67:577–583. [PubMed: 20457957]
8. Betjemann JP, Thompson AC, Santos-Sanchez C, et al. Distinguishing language and race disparities in epilepsy surgery. *Epilepsy Behav* 2013;28:444–449. [PubMed: 23891765]
9. Szaflarski M, Szaflarski JP, Privitera MD, et al. Racial/ethnic disparities in the treatment of epilepsy: what do we know? *Epilepsy Behav* 2006;9:243–264. [PubMed: 16839821]
10. Englot D, Ouyang P, Garcia P, et al. Epilepsy surgery trends in the United States, 1990–2008. *Neurology* 2012;78:1200–1206. [PubMed: 22442428]
11. Pestana Knight EM, Schiltz NK, Bakaki PM, et al. Increasing utilization of pediatric epilepsy surgery in the United States between 1997 and 2009. *Epilepsia* 2015;56:375–381. [PubMed: 25630252]
12. Barbour K, Hesdorffer DC, Tian N, et al. Automated detection of sudden unexpected death in epilepsy risk factors in electronic medical records using natural language processing. *Epilepsia* 2019;60:1209–1220. [PubMed: 31111463]
13. Matykiewicz P, Cohen K, Holland KD, et al. Earlier identification of epilepsy surgery candidates using natural language processing. *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing* 2013;1–9.
14. Cohen KB, Glass B, Greiner HM, et al. Methodological Issues in Predicting Pediatric Epilepsy Surgery Candidates Through Natural Language Processing and Machine Learning. *Biomed Inform Insights* 2016;8:11–18. [PubMed: 27257386]
15. Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch. *J Mach Learn Res* 2011;12:2493–2537.
16. Team RC. R: A language and environment for statistical computing 2013.
17. Bisgaier J, Rhodes KV. Auditing access to specialty care for children with public insurance. *N Engl J Med* 2011;364:2324–2333. [PubMed: 21675891]
18. Colby SL, Ortman JM. Projections of the size and composition of the US population: 2014 to 2060: Population estimates and projections 2017.



**Figure 1.**

Architecture of the natural language processing (NLP) algorithm. Electronic health record (EHR) data from all epilepsy patients were exported from the operational database (upper left). Patients with epilepsy were split into three groups: 1) patients who had a history of epilepsy surgery; 2) patients who were seizure free; and 3) “unknown” patients. The NLP algorithm was trained using patients from the first two groups (“Surgery” and “Seizure-Free”), and the trained model was used to assign patients in the “Unknown” group a surgical candidacy score. The scores were then stored in the epilepsy surgery software (ESS) database, where they could be accessed by the study team or used to send alerts to neurologists (shown on right).

**Table 1.**

The effect of demographic and socioeconomic patient characteristics on the NLP algorithm's surgical candidacy scores.

	Individual Effect on NLP Score (95% CI)	P-Value	Adjusted Effect on NLP Score (95% CI)	P-Value *
<b>Race</b>				
White (N = 3,064)	Reference	Reference	Reference	Reference
African American (N = 558)	-0.10 (-0.16, -0.04)	0.001	0.03 (-0.03, 0.09)	0.36
Other (N = 154)	-0.03 (-0.13, 0.08)	0.62	-0.01 (-0.11, 0.10)	0.92
<b>Distance from CCHMC</b>				
0-25 miles (N = 1,920)	Reference	Reference	Reference	Reference
26-50 miles (N = 608)	0.16 (0.10, 0.22)	<0.001	0.18 (0.12, 0.25)	<0.001
51-100 miles (N = 641)	0.20 (0.14, 0.26)	<0.001	0.25 (0.19, 0.31)	<0.001
Over 100 miles (N = 607)	0.30 (0.25, 0.36)	<0.001	0.37 (0.31, 0.44)	<0.001
<b>Age</b>				
0-4 years old (N = 614)	Reference	Reference	Reference	Reference
5-9 years old (N = 924)	0.02 (-0.05, 0.09)	0.56	0.02 (-0.04, 0.09)	0.52
10-14 years old (N = 928)	0.00 (-0.06, 0.07)	0.90	0.01 (-0.05, 0.08)	0.71
15-17 years old (N = 558)	0.03 (-0.04, 0.11)	0.43	0.06 (-0.02, 0.13)	0.12
>18 years old (N = 752)	0.22 (0.15, 0.29)	<0.001	0.24 (0.17, 0.31)	<0.001
<b>Gender</b>				
Male (N = 2,029)	Reference	Reference	Reference	Reference
Female (N = 1,747)	0.02 (-0.02, 0.06)	0.42	0.01 (-0.03, 0.05)	0.53
<b>Language</b>				
English (N = 3,703)	Reference	Reference	Reference	Reference
Non-English (N = 73)	-0.01 (-0.16, 0.14)	0.89	0.03 (-0.13, 0.19)	0.72
<b>Insurance</b>				
Private (N = 1,875)	Reference	Reference	Reference	Reference
Public (N = 1,697)	0.03 (-0.01, 0.07)	0.20	0.06 (0.02, 0.11)	0.01
Self-Pay (N= 150)	-0.05 (-0.16, 0.06)	0.39	-0.04 (-0.15, 0.07)	0.43
Other (N = 51)	0.01 (-0.17, 0.19)	0.91	0.01 (-0.16, 0.19)	0.87
<b>Median Household Income (N = 3,776)</b>	0.004 per \$10,000 (-0.007, 0.02)	0.43	0.03 per \$10,000 (0.02, 0.06)	<0.001

\* Adjusted p-values from the multiple linear regression.

NLP: natural language processing; CI: confidence interval; and CCHMC: Cincinnati Hospital Children's Medical Center.