Estimating Time to Progression of Chronic Obstructive Pulmonary Disease with Tolerance

Chunlei Tang* *Senior Member, IEEE*, Joseph M. Plasek*, Xiao Shi*, Meihan Wan, Haohan Zhang, Min-Jeoung Kang, Liqin Wang, Sevan M. Dulgarian, Yun Xiong, Jing Ma, David W. Bates, and Li Zhou

Abstract **We defined tolerance range as the distance of observing similar disease conditions or functional status from the upper to the lower boundaries of a specified time interval. A tolerance range was identified for linear regression and support vector machines to optimize the improvement rate (defined as IR) on accuracy in predicting mortality risk in patients with chronic obstructive pulmonary disease using clinical notes. The corpus includes pulmonary, cardiology, and radiology reports of 15,500 patients who died between 2011 and 2017. Their performance was compared against a long short-term memory recurrent neural network. The results demonstrate an overall improvement by those basic machine learning approaches after considering an optimal tolerance range: the average IR of linear regression was 90.1% and the maximum IR of support vector machines was 66.2%. There was a similitude between the time segments produced by our tolerance algorithms and those produced by the long short-term memory.**

Index Terms **– "pulmonary disease, chronic obstructive," disease progression, machine learning, neural networks, palliative care**

I. INTRODUCTION

Chronic obstructive pulmonary disease (COPD) is the third leading cause of mortality in the United States, affecting an estimated 14.7 million diagnosed patients [1]. The quality of life of patients with COPD deteriorates as the disease progresses. Mild COPD symptoms include shortness of breath, chronic cough, and fatigue; whereas severe COPD can result in death, mostly due to respiratory failure, heart failure, pulmonary infection, and pulmonary embolism [2-3]. Unfortunately, there is currently no cure for COPD.

Clinical notes capture important details about disease progression that cannot be found elsewhere in patients' electronic health records, however they present challenges in accessibility to medical researchers. These challenges have motivated the development of disease progression modeling methods [4-10]. Most of these methods utilize multiple hidden layers, referred to as "deep" neural networks, to seek patient groups with similar disease progression pathways in the investigation of a complex disease [11]. We modified a standard long short-term memory (LSTM) by adding a flatten and dense layer, and it showed gratifying results for learning irregular time lapse segments when incorporating disease conditions mentioned in the clinical notes [12]. Although the results are promising, this deep learning models are without a doubt, complicated. We envision that a data-driven approach may help optimize less complex machine learning models such as linear regression (LR) and support vector machines (SVMs).

Mining and analyzing a large volume of notes requires the development of computational methods that are able to: (*a*) derive summarized clinical information relevant to the disease, and/or (*b*) define a time lapse for each disease stage. These two tasks are complementary and mutually reinforcing [4, 13-14]. Combining the two tasks together, our goal is to develop a general-purpose method for machine learning models, which can optimally learn the time to progression prior to death in patients with COPD using free-text clinical documents.

Determining the optimal duration for each time window in a patient population corresponding to a COPD progression stage is the root issue of this study. Disjoint time windows are often used in temporal segmentation methods to map a specific clinical document to a COPD stage. However, a preset disjoint time window might be biased by human intervention. For example, using a more extended time window (e.g., 360 days) may be problematic as the data within that window may fall within multiple different stages of COPD.

This work was partially funded by the Partners Innovation Fund, the CRICO/Risk Management Foundation of the Harvard Medical Institutes Incorporated, the Shanghai Science and Technology Development Fund No. 19511121204, No.19DZ1200802, and the National Natural Science Foundation of China Projects No. U1636207, No. U1936213.

*These authors contributed equally. Note that Yueyang Hospital of Integrated Traditional Chinese Medicine and Western Medicine is the institutional affiliation of the first author as well.

Corresponding author: Min-Jeoung Kang (mkang6@bwh.harvard.edu).

C. Tang, J. M. Plasek, M. Kang, L. Wang, D. W. Bates, and L. Zhou are with the Division of General Internal Medicine and Primary Care of Brigham and Women's Hospital at Harvard Medical School, Boston, MA, USA, 02115.

X. Shi is with Yueyang Hospital of Integrated Traditional Chinese Medicine and Western Medicine, Shanghai University of Traditional Chinese Medicine, Shanghai, CHN, 200437.

M. Wan and Y. Xiong are with Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University, Shanghai, CHN, 201203.

H. Zhang is with School of Computer Science, Carnegie Mellon University, Pittsburgh, PA USA, 15213.

J. Ma is with the Department of Population Medicine at Harvard Medical School, Boston, MA, USA, 02215.

II. Materials and Methods

We propose a formula based on a specific preset time window for calculating an optimal tolerance range using clinical notes. Our study design is shown in Fig. 1.

A. Setting and corpus

We retrieved 3,001,350 free-text clinical documents corresponding to 15,500 unique patients from the Research Patient Data Registry at Partners Healthcare, an integrated healthcare delivery network located in the Greater Boston area of Massachusetts. Inclusion criteria were: (*a*) date of death between 2011 and 2017 recorded in Massachusetts Death Certificate files, (*b*) the patient received care at any Partners Healthcare facility between 2011 and 2017, with (*c*) at least one COPD diagnosis code was recorded in the EHR. The list of COPD diagnosis codes include International Classification of Diseases (ICD)-9-CM: 490, 491.*, 492.*, 496 and ICD-10-CM: J40, J41.*, J42, J43.*, J44.*. This study was approved by Partners Institutional Review Board (IRB).

Each extracted document contains a header listing information related to patient_num (patient ID), NoteType (i.e., pulmonary notes, cardiology reports, and radiology reports), and CreateLocalDTs (generate time). We constructed a total of four corpora. Clinical notes from each domain or all three domains were merged into a single corpus using a heuristic merger that inserts notes from each domain into the appropriate chronological place in the corpus (see in Appendix Table I, physician interpretation from pulmonary notes, findings from radiology reports, and abnormal ECG from cardiology reports). Such a chronological order results and allows one to predict patient time of patient death. This resulted in each sample is equivalent to a text file containing one-day clinical notes from a domain or all three domains.

B. Optimal enlargement of a preset time window

One potential approach to fix the human bias issue is to consider a permissible tolerance range that could enlarge the preset window appropriately utilizing the data.

Definition 1 (*Tolerance Range*). Tolerance range, we call Δwin, is defined as the distance (or range) from the upper and lower boundaries of an interval.

In the case of COPD, Δwin is defined as the distance of observing similar COPD conditions or functional status from the upper to the lower boundaries of a specified time interval (or a *n*-day window).

Definition 2 (*Global Tolerance Range*). The global tolerance range, *global* Δwin, is based on a "dictionary" learned from the corpus. Each dictionary entry is defined as *[frequency, duration]*, in which the duration is the absolute difference between the generated time of a specific clinical note and the patient's death date, and the frequency is the sum of the number of notes across the same duration.

Definition 3 (*Optimal Tolerance Range*). Given a preset *n*-day window as a time interval (e.g., 0-30 days, or 31-60 days, etc.), the optimized tolerance range (*optimal* Δwin) is:

$$\Delta win = \sigma\left(\frac{e^{Z}-e^{-Z}}{e^{Z}+e^{-Z}}\right)\cdot interval \quad (1)$$

where σ permits a manually specified value, $Z = \text{average}(size)/_{size}$, in which $size$ is the number of notes within a preset *n*-day window and average(*size*) is the total number of notes divided by the total number of preset windows.

Optimal Δwin can be also rewritten as $\sigma \cdot \tanh(Z) \cdot interval$. The reason we use the tanh function is that its output range is from -1 to 1. The $\tanh(Z) \cdot interval$ maps the relative size of text (i.e., the reciprocal of *Z*) to the $[0 * interval, 1 * interval]$ output range because the scale of the horizontal axis is greater than or equals to zero. The $\sigma \cdot (\tanh(z) \cdot interval)$ modifies the time interval based on different values of parameter $\sigma = \{0, 1, 2, 3, \cdots\}$. Note that parameter $\sigma = 0$ means that there is no tolerance range.

Thus, *Z* is proportional to Δwin and inversely proportional to the relative size of texts in a specific time window: when the relative *size* $\rightarrow$ 0, then $Z \rightarrow +\infty$, $\Delta win \rightarrow \sigma \cdot interval$; when the relative *size* $\rightarrow +\infty$, then $Z \rightarrow 0$, $\Delta win \rightarrow 0$. Since interval is fixed by the tolerance range, it must have a blurred border in

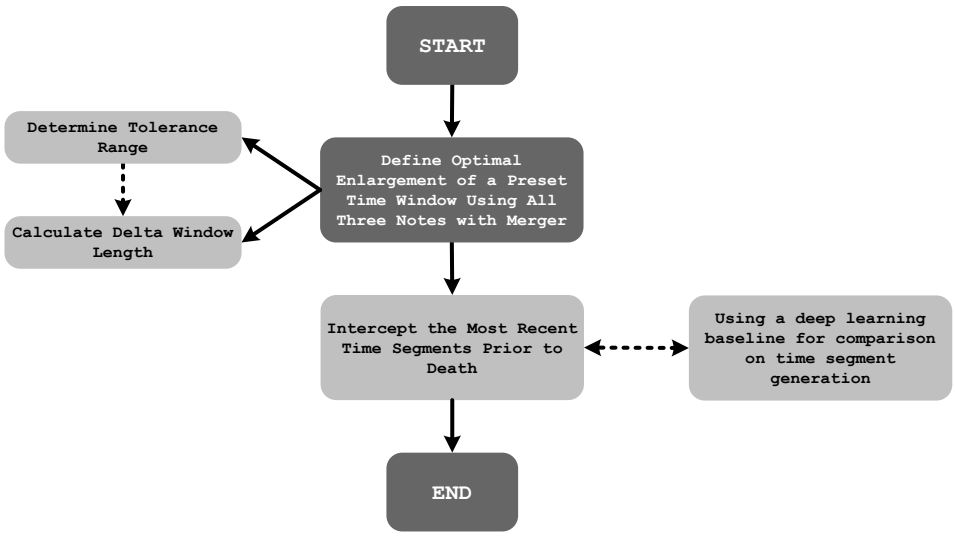
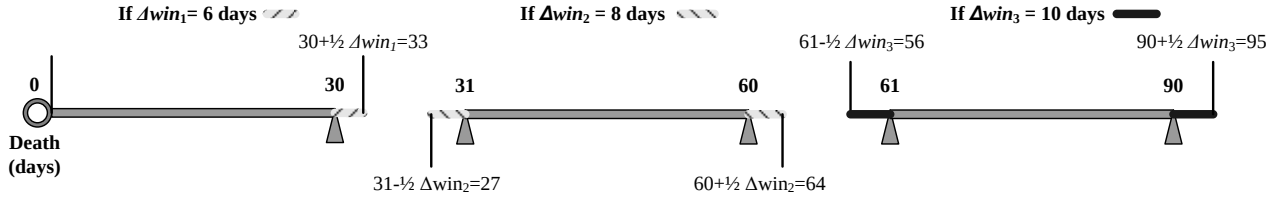
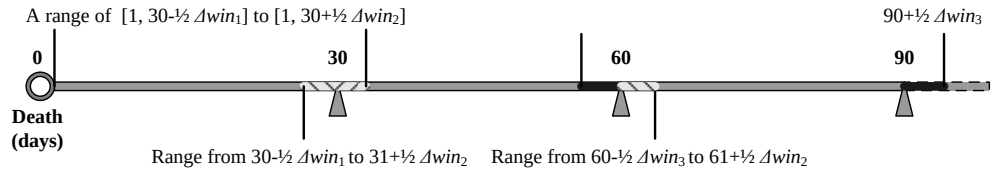


Fig. 1. Overview of the Study Design. Note that COPD is chronic obstructive pulmonary disease and LSTM is long short-term memory.



(a) Optimal *Δwin* for each time window



(b) Optimal enlarged time window by considering all *Δwin* values

Fig. 2. An illustration of how the optimal Δwin works for each time window and for the timeline of disease progression.

the range of $\left[lower\ bound - \frac{1}{2}\Delta win, upper\ bound + \frac{1}{2}\Delta win\right]$.

Each Δ*win* for a given time segment (e.g., 0-30 days, or 31-60 days, etc.) might have a different value, which affects its adjacent segments on the timeline of disease progression (Fig. 2). On each of the four corpora, we specify four preset time windows as 30-day, 90-day, 180-day, and 360-day, respectively.

C. Simple models performance with a tolerance range

To demonstrate that a simple model can perform well by considering a tolerance range, we employ two common machine learning models: linear regression (LR) and support vector machines (SVMs). The reason we use linear regression instead of logistic regression is that the former covers more broadly applications, and the latter is a special type of linear regression. LR attempts to model the relationship between two variables that is dichotomous by finding a linear equation to observed data. For our purpose, LR is used to determine whether the output belongs to a specified time segment or not. An SVM outputs an optimal hyperplane that categorizes testing data. For example, the hyperplane in a two-dimensional space is a line dividing a plane into two parts wherein each class lay on either side. In our experiment, the SVM is used to obtain multiple optimal Δ*win*s produced by all the clinical notes within a specified preset time window.

We utilized cross-validation techniques as follows: we split the clinical notes in each corpus into multiple (i.e., 48 *Z*'s) training/testing sets using a fixed ratio (i.e., 2%; specifically, every 2% interval between 2% and 98%, inclusive). For each given clinical note, accuracy was calculated as follows: if the absolute difference between the predicted death from SVM or LR and the actual death date is within the ideal Δ*win* (i.e., global tolerance range), set the accuracy to 1; if not, set the accuracy to 0. For each corpus, we compared the maximum (mean) accuracy and the maximum (σ) accuracy where σ = {1,2,3} for 48 *Z*'s for each corpus. We first calculate the accuracy of the optimal training/testing set among the 48 *Z*'s for each σ (i.e., σ = {1,2,3}, this returns the temporary set {max over the 48 *Z*'s when σ = 1, max over the 48 *Z*'s when σ = 2, max over the 48 *Z*'s when σ = 3}, and then apply the function (i.e., sigma = max, mean) to get our value of interest. For example, if σ = {1,2,3} and the temporary set returns {50%, 75%, 85%} then maximum (σ) would be 85% and maximum(mean) would be (50%+75%+85%) ÷3 = 70%.

Based on maximum accuracy in every corpus, we selected the best performing accuracy from which to calculate the optimal Δ*win* for each experiment. So, for a specified preset time window on each corpus, the best regular time segments can be set up as a preset time window plus the optimal Δ*win*.

D. Evaluation

We invited a panel of subject matter experts (*n*=4) consisting of 3 physicians (including one of the authors of this article) to assist with the evaluation two conducted experiments.

- *Experiment 1: Accuracy by comparing with or without tolerance*

To evaluate the tolerance range, we used LR and SVMs to calculate accuracy using cross-validation. On each of the four corpora, we calculated their improvement rates (*IR*) by the formula as follows for each tolerance range.

$$IR = \frac{Accuracy_{\sigma} - Benchmark\ Accuracy}{Benchmark\ Accuracy} \quad (2)$$

- *Experiment 2: Using a deep learning baseline for comparison on time segment generation*

Previously, we developed a four-layer deep learning model using LSTM to adjust time interval settings and to capture an irregular time lapse segments [12]. The components of the model include (*a*) a pre-processing and word embedding layer to prepare the data, (*b*) an LSTM layer to predict death date, and (*c*) a flatten and dense layer combination to capture the irregular time lapse of segments. We used this LSTM-based deep learning model as an existing baseline for comparison against our tolerance range method.

TABLE II
THE AVERAGE IMPROVEMENT RATE FOR EACH VALUE OF Σ ON FOUR CORPORA

	Support Vector Machine				Linear Regression			
	σ=0	IRσ=1	IRσ=2	IRσ=3	σ=0	IRσ=1	IRσ=2	IRσ=3
Pulmonary notes	11.9%	0	0	5.7%	0.2%	19.3%	148.7%	192.1%
Radiology reports	17.6%	0	7.4%	42.0%	0.1%	0	25.0%	70.1%
Cardiology reports	14.0%	0	0	49.2%	0.1%	47.8%	130.1%	194.7%
Merged notes	18.5%	0	5.8%	66.2%	0.3%	20.5%	92.0%	141.0%

Note that the benchmark accuracy uses SVM σ=0 (see the highlighted row).

TABLE IV
COMPARISON OF OPTIMAL TOLERANCE RANGE

	Time-Window Length	Support Vector Machine			Linear Regression		
		σ=1	σ=2	σ=3	σ=1	σ=2	σ=3
Pulmonary notes	30 days	6.2	12.4	18.6	14.8	26.5	**40.7**
	90 days	22.1	44.3	66.4	40.8	77.4	116.6
	180 days	34.0	67.9	101.9	77.3	154.5	228.9
	360 days	41.2	82.5	123.7	143.9	299.6	442.4
Radiology reports	30 days	15.0	30.0	45.0	14.1	28.1	42.1
	90 days	22.6	45.1	91.9	40.7	81.9	120.4
	180 days	51.6	104.1	165.6	82.5	161.0	239.3
	360 days	116.3	279.5	429.9	136.7	270.1	426.2
Cardiology reports	30 days	13.5	27.0	42.2	13.4	26.5	39.8
	90 days	41.1	85.1	130.0	40.6	82.0	122.1
	180 days	77.8	155.7	253.1	80.2	157.1	239.0
	360 days	154.9	317.5	488.9	154.5	315.6	469.2
Merged notes	30 days	9.5	18.9	31.0	13.8	28.7	41.1
	90 days	25.2	50.3	96.5	43.5	85.6	123.9
	180 days	65.1	130.1	217.7	82.2	159.8	236.0

Note that optimal Δwin length under different values of σ using maximum(σ) accuracy, in which highlighted "40.7" (located in LR's σ=3) is the best optimal Δ*win* for the ***pulmonary notes***.

III. RESULTS

A. Influence of parameter value

Using cross-validation on four corpora, we calculated the average improvement rate on different values of σ corresponding to the two models (i.e., LR and SVM) as shown in Table II. Of these, SVM's results were utilized as the benchmark accuracy because LR's accuracy is close to 0 when σ = 0 but increases when σ changes.

We found that, tuning the parameter σ may be used to improve accuracy, with σ = 0 serving as no tolerance range. For example (see detailed information in Appendix Table III), on only pulmonary notes using a tolerance range of σ = 0 (i.e., IR is 210.3%), the maximum accuracy without a tolerance range for 360 days is about 20%, which increased to 63.0% (i.e., 20.3%×(1+210.3%) = 63.0%) using a tolerance range of σ = 3. Another example is, on merged notes using a tolerance range of σ = 3 (i.e., average IR is 103.6%), our approach achieved a max of 83.3% and an average of 74.0% accuracy on the testing set in predicting the individual patient's death within the next 360 days.

B. Optimal Δ*win length calculation*

Table II also indicates that the maximum (σ) performed better than maximum (*mean*) as changes in accuracy were less erratic as the preset time window length increased.

After having chosen 40 days (i.e., the best optimal Δ*win* = 40.7 for linear regression at a 30-day preset time window from the pulmonary notes corpus where parameter σ =3) as the length of optimal Δwin, we obtained an enlarged time window of 70 days (Table IV).

Thus, a sequence of the most recent regular time segments prior to death is {[0, 70], [71,140], [141, 280], [281, 350], …}. Note that choosing LR's σ=2 is also acceptable but not optimal. Table V shows two time segments [140, 210] and [211, 270] corresponding to the tolerance range which are effectively equivalent to the [145, 270] time segment of the LSTM [12, 21].

TABLE V
COMPARISON OF TIME SEGMENTS

Annotated COPD Stage	Linear Regression with Tolerance (σ=3)	Long Short-Term Memory
IV	[0, 70]	[0, 65]
III	[71, 140]	[55, 150]
II-2	[141, 210]	[145, 270]
II-1	[211, 280]	
II-1	[281, 350]	[262, 360]
II-2	[351, 420]	[337, 484]
II-1	[421, 490]	[450, 552]
II-1	[491, 560]	
I	[561, 630]	[449, 630]

Note that two highlighted rows indicate, the results produced by the two methods are inconsistent. COPD = Chronic Obstructive Pulmonary Disease.

IV. Discussion

In this study, we devised a formula of the optimal Δ*win* for finding time segments using simple machine learning algorithms (e.g., linear regression and support vector machines). By considering a tolerance range, the performance of both LR and SVM was improved in terms of accuracy. LR achieved an average IR of 90.1% in predicting the individual patient's death within a specified time. Although lacking stability, SVM had a maximum IR of 66.2% on accuracy prediction. Using a tolerance range also can address the problem of human bias in setting a preset time window. We found that setting parameter σ to 2 or 3 is suitable to get an optimized enlarged time window in our COPD dataset. We identified that longer preset time windows (e.g., 180 or 360 days) might have a clinical significance; for example, obtaining the 1-year relative survival rate for a patient with at least 80% accuracy from a single clinical document.

We achieved very similar results (Table V) to the LSTM method that we previously proposed [9]. Specifically, the most recent nine regular time segments prior to death date are basically in line with the results from the first seven irregular time segments generated by the LSTM. The only inconsistent result was the time segment [145, 270] produced by the LSTM which combined two different COPD substages: II-2 [140, 210] and II-1 [211,280].

Our approach focused on finding the right combination of time and free-text to be able to elicit important information regarding progression stage or timing and to establish the feasibility and usefulness of this approach. Future telemedicine workflows, patient diaries, and monitoring devices may be capable of capturing additional relevant clinical data between clinic visits. These prognostic models and guideline suggestions should be retroactively validated using a large external cohort and then prospectively validated prior to widespread clinical implementation. Our approach may be generalizable to other clinical use cases reliant on free-text longitudinal data (particularly with irregular temporal segments).

Clinician decision making for terminal COPD patients in the palliative and hospice care settings could benefit from our approach. The deterioration of patients' quality of life in the end stages of COPD occurs due to a lack of physical, social, and emotional functioning (e.g., extreme dyspnea, shortness of breath, anxiety, depression) [15]. Optimizing patients' quality of life for their desired care can be accomplished through appropriate palliative and hospice care rather than through aggressive treatments that over-estimate prognoses and/or that aren't efficacious [16]. Utilization of our algorithm as a decision aid may help reduce physician over-estimates of prognosis and assist in recommending appropriate palliative or hospice care.

A. Limitation

One limitation was that our study is based on the clinical notes from a single organization that were written using a single language – English, thus, the results may not be generalizable to other institutions, languages, or care settings. Another disadvantage is related to our retrospective cohort design that our data may be incomplete or inadequately captured in the EHR and may not reflect current clinical workflows or meet current documentation standards [17]. Knowledge about COPD can change over time. The clinical progression of the disease discovered in the present study may not reflect the patterns in patients who took advantage of new clinical advances. Further adaptation of the algorithms on new data is necessary in identifying the disease progression patterns after the introduction of new treatments.

V. Conclusion

The main findings of this study were that it is feasible to use the optimal tolerance range as a general baseline to deep learning approaches. The tolerance range effectively addresses the problem of human bias with preset time windows and applies the refined time segments to classify and track disease progression.

References

[1] American Lung Association. Trends in COPD (Chronic Bronchitis and Emphysema): Morbidity and mortality; March 2013. Avariable from: http://www.lung.org/assets/documents/research/copd-trend-report.pdf [Accessed July 2019].

[2] J. Zielinski, W. MacNee, J. Wedzicha, N. Ambrosino, A. Braghiroli, J. Dolensky, P. Howard, K. Gorzelak, A. Lahdensuo, K. Strom, M. Tobiasz E. Weitzenblum. Causes of death in patients with COPD and chronic respiratory failure. Monaldi archives for chest disease= Archivio Monaldi per le malattie del torace, 1997, 52(1): 43-47.

[3] Global Initiative for Chronic Obstructive Lung Disease (GOLD). Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease. 2017. Avariable from: https://goldcopd.org/wp-content/uploads/2017/11/GOLD-2018-v6.0-FINAL-revised-20-Nov_WMS.pdf. [Accessed July 2019].

[4] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, J. Zhou. Patient subtyping via time-aware LSTM networks. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2017). ACM, 2017, p.65-74.

[5] E. Choi, M. T. Bahadori, A. Schuetzy, W. F. Stewarty, J. Sun. Doctor AI: Predicting clinical events via recurrent neural networks. In Machine Learning for Healthcare Conference. 2016, p.301-318.

[6] E. Choi, M. T. Bahadori, A. Schuetzy, W. F. Stewarty, J. Sun. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In Advances in Neural Information Processing Systems. 2016, p. 3504-3512.

[7] C. Esteban, O. Staeck, Y. Yang, V. Tresp. Predicting clinical events by combining static and dynamic information using recurrent neural networks. In 2016 IEEE International Conference on Healthcare Informatics (ICHI). IEEE, 2016, p.93-101.

[8] Z. C. Lipton, D. C. Kale, C. Elkan, R. Wetzell. Learning to diagnose with LSTM recurrent neural networks. arXiv preprint. 2016, arXiv:1511.03677.

[9] J. Zhou, Z. Lu, J. Sun, L. Yuan, F. Wang, J. Ye. Feafiner: Biomarker identification from medical data through feature generalization and selection. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013, p.1034-1042.

[10] J. Zhou, L. Yuan, J. Liu, J. Ye. A multi-task learning formulation for predicting disease progression. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2011, p.814–822.

[11] Z. Che, D. Kale, W. Li, M. T. Bahadori, Y. Liu. Deep computational phenotyping. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015, p.507–516.

[12] C. Tang, J. M. Plasek, H. Zhang, M. Kang, H. Sheng, Y. Xiong, D. W. Bates, L. Zhou L. A Deep Learning Approach to Handling Temporal Variation in Chronic Obstructive Pulmonary Disease Progression. In

Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine, 2018, p. 502-509.

[13] X. Wang, D. Sontag, F. Wang. Unsupervised learning of disease progression models. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD 2014). ACM, 2014, p.85-94.

[14] H. Xiao, J. Gao, L. Vu L, T. Deepak. Learning temporal state of diabetes patients via combining behavioral and demographic Data. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2017). ACM, 2017, p.2081-2089.

[15] J. M. Gore, C. J. Brophy, M. A. Greenstone. How well do we care for patients with end stage chronic obstructive pulmonary disease (COPD)? A comparison of palliative care and quality of life in COPD and lung cancer. Thorax, 2000, 55(12): 1000-1006.

[16] Where do Americans die? Available: https://palliative.stanford.edu/home-hospice-home-care-of-the-dying-patient/where-do-americans-die. [Accessed July 2019].

L. G. Portney, M. P. Watkins. Foundations of clinical research: applications to practice. 3rd ed. Upper Saddle River, N.J.: Pearson/Prentice Hall; 2009.

APPENDIX

See Table I and IV.

TABLE I
REAL-WORLD CHRONIC OBSTRUCTIVE PULMONARY DISEASE CORPUS DESCRIPTION

	Summary			Example	
	Total number	Unique Patient	Average time span of patient (days)	Section Name	Section Text
Pulmonary notes	78,489	2,431	724.4	Physician Interpretation	FEV1, FVC, and FEV1/FVC are reduced. TLC is normal. FRC is increased. RV is increased. RV/TLC ratio is increased. Single breath diffusion capacity is reduced. DL/VA is reduced. These data demonstrate a very severe (FEV1 <35 of predicted) obstructive ventilatory deficit, with gas trapping and mild hyperinflation, that is worse (17 decline in FEV1) since [DATE].
Radiology reports (chest X-ray)	1,893,498	13,414	843.8	Findings	Comparison was made to prior study of [DATE]. The cardiomediastinal silhouette is stable with tortuous aorta. No confluent opacities to suggest pneumonia. There is no pleural effusion or pneumothorax. Wedge compression deformity of upper thoracic vertebra is present.
				Impression	No acute cardiopulmonary process.
Cardiology Reports	1,029,363	13,918	2,459	Abnormal ECG	When compared with ECG of [DATE], (unconfirmed) Junctional rhythm has replaced Sinus rhythm.

Note that we set up a total of 4 datasets (i.e., pulmonary notes, radiology reports, cardiology reports, and their merger).

TABLE IV
ACCURACY ON FOUR DATASETS

		Support Vector Machine				Linear Regression			
		σ=0	σ=1	σ=2	σ=3	σ=0	σ=1	σ=2	σ=3
Pulmonary notes	30 days	5.5%	5.5%	5.5%	5.5%	0.0%	4.1%	5.5%	6.8%
	90 days	10.8%	10.8%	10.8%	13.5%	0.7%	11.6%	19.9%	28.1%
	180 days	10.8%	10.8%	10.8%	10.8%	0.0%	18.9%	40.5%	40.5%
	360 days	20.3%	20.3%	20.3%	20.3%	0.0%	21.9%	52.1%	63.0%
Radiology reports	30 days	12.4%	12.4%	12.4%	14.4%	0.0%	3.4%	4.8%	6.6%
	90 days	9.3%	9.3%	9.3%	17.6%	0.0%	11.9%	19.3%	28.4%
	180 days	13.8%	13.8%	13.8%	18.6%	0.5%	13.9%	24.9%	35.5%
	360 days	34.7%	34.7%	44.9%	49.0%	0.0%	26.5%	38.8%	49.0%
Cardiology reports	30 days	8.7%	8.7%	8.7%	13.0%	0.3%	3.9%	7.7%	11.3%
	90 days	7.3%	7.3%	7.3%	9.7%	0.0%	11.0%	20.3%	29.9%
	180 days	16.7%	16.7%	16.7%	27.1%	0.0%	25.5%	38.3%	48.9%
	360 days	23.4%	23.4%	23.4%	34.0%	0.0%	42.6%	63.0%	75.3%
Merged notes	30 days	8.3%	8.3%	8.3%	14.6%	0.3%	4.5%	7.3%	9.1%
	90 days	15.4%	15.4%	16.1%	23.1%	0.0%	13.5%	27.1%	38.5%
	180 days	16.8%	16.8%	16.8%	20.4%	1.0%	22.9%	36.5%	46.9%
	360 days	33.3%	33.3%	39.6%	64.6%	0.0%	47.9%	70.8%	83.3%

Note that each dataset with different preset time windows is changed by the value of parameter σ.

AUTHOR INFORMATION

Chunlei Tang is a research associate at Harvard Medical School. Dr. Tang, author of The Data Industry: The Business and Economics of Information and Big Data, is a recognized advocate for the Data Economy. Tang's work begins with an innovative data-mining method built upon her doctoral work and applicable to various fields (e.g., finance, economics, insurance, bioinformatics, and sociology). She then endeavored to better match data products with their marketing capabilities by tackling business innovation. Inspired by business, she implemented a theoretical innovation for Data Science and proposed the (now-mainstream) definition of the data industry in 2013, stating that the data industry aims to bridge the gap between data science and economics. Her current focus is on applying innovation in healthcare, and she believes precision medicine is an application of the data industry.

Joseph M. Plasek is a health data scientist and clinical informatician. Joseph joined the MTERMS research lab at Brigham and Women's Hospital in 2010. His research focuses on clinical applications of natural language processing, machine learning, and dynamic systems theory.

Xiao Shi, an attending physician at Shanghai Yueyang Integrated Medicine Hospital, severs as the Chief of the hospital's division of geriatrics. She also is a professor in osteoporosis, geriatrics, and rheumatoid at Shanghai University of Traditional Chinese Medicine. Shi has extensive experience in multiple chronic conditions among elderly people on how the bridging of Traditional Chinese Medicine and Western Medicine.

Meihan Wan is a postgraduate majoring in computer science at Fudan University. Her research area is data mining, representation learning and clinical applications using machine learning.

Haohan Zhang is a graduate student majoring in computer science at Carnegie Mellon University. Her research interest lies in distributed systems and scalability in machine learning.

Min Jeoung Kang is a research fellow at Harvard Medical School and Brigham and Women's Hospital. She was a former emergency room nurse and taught in nursing college. Her research focuses on improving patient safety in clinical settings using electronic health records.

Liqin Wang is a research fellow at Harvard Medical School and Brigham and Women's Hospital. She received PhD in Biomedical Informatics. Her research focuses on developing and implementing new methodologies to drive high-value care, by leveraging large-scale healthcare data, biomedical literature, medical knowledge bases (e.g., UMLS), and novel artificial intelligence technologies including natural language processing and machine learning.

Sevan M. Dulgarian is a research assistant at Brigham and Women's Hospital. She received a degree in Public Health from the University of Massachusetts Amherst Commonwealth Honors College. Her research interests are in the areas of prevention and healthcare quality and safety.

Yun Xiong received the PhD degree in computer and software theory from Fudan University. She is a professor of computer science at Fudan University, Shanghai, China. Her research interests include data science and data mining.

Jing Ma is an associate professor and the director of the China Center in the Department of Population Medicine and the Harvard Pilgrim Health Care Institute at Harvard Medical School. Ma has extensive experience in prospective and longitudinal population research in the United States. Ma has been conducting an extensive high-quality study on nutritional, hormonal and metabolic biomarker and genetic markers as well as the impact of environmental and lifestyle exposures on cancer risk and progression, drawing on the data of Physicians' Health Study and Nurses' Health Study, among others. She has also been closely involved in fostering international research collaborations with multiple cohorts among many countries.

David W. Bates is an American-born physician, biomedical informatician, and professor, who is internationally renowned for his work regarding the use of health information technology to improve the safety and quality of healthcare, in particular by using Clinical Decision Support. Dr. Bates has done especially important work in the area of medication safety. He began by describing the epidemiology of harm caused by medications, first in hospitalized patients and then in other settings such as the home and nursing homes. Subsequently, he demonstrated that by implementing computerized physician order entry (CPOE), medication safety could be dramatically improved in hospitals. This work led the Leapfrog Group to call CPOE one of the four changes that would most improve the safety of U.S. healthcare. It also helped hospitals to justify investing in electronic health records and in particular CPOE.

Li Zhou is an Associate Professor at Harvard Medical School and a Lead Investigator at Brigham and Women's Hospital (BWH). Her research interests are in the areas of data science, natural language processing (NLP), medication safety, and clinical decision support. She has served as PI and co-investigator on numerous collaboratives, multiple-site research grants. She directs the MTERMS lab at BWH. Her team has designed and implemented a generic and highly configurable NLP system known as MTERMS, which has been effectively applied to various clinical domains and documents.