

Lexical Concept Distribution Reflects Clinical Practice

Eugene Breydo, PhD^a, Maria Shubina, DSc^b, James W. Shalaby, PharmD^c,
Jonathan S. Einbinder, MD, MPH^{a,b,d}, Alexander Turchin, MD, MS^{a,b,d}

^a*Quality Performance Management, Partners HealthCare, Boston, MA*

^b*Brigham and Women's Hospital, Boston, MA*

^c*PSMI Consulting, LLC, El Cerrito, CA*

^d*Harvard Medical School, Boston, MA*

Abstract

It is not known whether narrative medical text directly reflects clinical reality. We have tested the hypothesis that the pattern of distribution of lexical concept of medication intensification in narrative provider notes correlates with clinical practice as reflected in electronic medication records.

Over 29,000 medication intensifications identified in narrative provider notes and 444,000 electronic medication records for 82 anti-hypertensive, anti-hyperlipidemic and anti-hyperglycemic medications were analyzed. Pearson correlation coefficient between the fraction of dose increases among all medication intensifications and therapeutic range calculated from EMR medication records was 0.39 ($p = 0.0003$). Correlations with therapeutic ranges obtained from two medication dictionaries, used as a negative control, were not significant.

These findings provide evidence that narrative medical documents directly reflect clinical practice and constitute a valid source of medical data.

Introduction

Natural language processing of narrative medical texts is growing in importance as the source of data for research, administrative operations and patient care¹. As the technical challenges of the computational text analysis are being overcome, an important question arises: how valid are the narrative documents as the reflection of clinical reality?

This question represents more than a theoretical concern. While electronic medical records (EMRs) can help providers generate and exchange documents more efficiently, they also provide opportunities for shortcuts that were not available when the documents could only be handwritten or dictated rather than entered electronically. For example, as described by P. Hartzband and J. Groopman: "Many times, physicians have clearly cut and pasted large blocks of text, or even complete notes, from other physicians; we have seen portions of our own notes inserted verbatim into another doctor's note."²

Whether language represents a precise reflection of the world has been the subject of a debate in modern linguistics³. While the actual reasons for discrepancies between language and reality can vary depending on the circumstances, medical narrative does not appear to have completely escaped this dilemma.

In this study, we have investigated the relationship between medical narrative documents and clinical reality on the example of treatment intensification in patients with diabetes. Diabetes is a common, costly, and dangerous disease. Lowering blood glucose, blood pressure and cholesterol in patients with diabetes decreases the risk of complications⁴⁻⁶. However, even though evidence-based guidelines for treatment targets for blood glucose, blood pressure and cholesterol in patients with diabetes have been well publicized, the majority of patients do not meet these targets^{7,8}.

The reasons for this are not well understood but lack of appropriate medication intensification is thought to be a contributing factor. Medication intensification is frequently only documented in narrative provider notes⁹. Therefore, in order to study medication intensification, it is important to understand the relationship between medication intensification recorded in narrative notes and clinical reality.

Medications can be intensified in two ways: a new medication can be started or the dose of an existing medication increased. It would be expected that medications with larger therapeutic range (difference between the maximum and the lowest doses) would have proportionally more dose increases than the medications with smaller therapeutic ranges. In this study we have therefore investigated the relationship between a) the distribution of dose increases vs. new medication initiations among medication intensifications for patients with diabetes recorded in narrative provider notes and b) clinical reality as represented by the therapeutic range of the same medications in

structured EMR medication records. As a negative control, we have also compared this distribution to therapeutic ranges in two medication dictionaries.

Materials and Methods

Design

We carried out a retrospective analysis of EMR data to determine the correlation between a) the fraction of medication dose increases among all intensifications and b) the medications' therapeutic ranges derived from clinical EMR data vs. two medication dictionaries. Individual medication served as the unit of analysis.

Study Medications

We included in our analysis three classes of medications: anti-hypertensive, anti-hyperglycemic and anti-hyperlipidemic. The analysis was restricted to medications for which minimum and maximum doses expressed in the same units were available in all three data sources for medication therapeutic range (see below). We also excluded from our analysis medications that had highly different therapeutic ranges for multiple indications (e.g. nicotinic acid).

Study Measurements

Medication intensification was defined as initiation of a new medication or an increase in the dose of an existing medication¹⁰. Medication intensification events were identified in the text of the notes using a previously validated text analysis algorithm with sensitivity of 83.8% and specificity 95.0%¹¹. The algorithm was subsequently clinically validated by demonstrating correlation of medication intensifications it identified with blood pressure changes¹². The algorithm differentiates between medication initiations and dose increases. The fraction of dose increases was calculated for each medication as the ratio of dose increases to all intensification events.

Therapeutic ranges were identified for each medication from three sources: a) clinical practice as represented by the doses found in the EMR; b) First DataBank (FDB) medication dictionary (First DataBank, South San Francisco, CA); and c) Master Drug Dictionary (MDD) – a medication dictionary internally developed and maintained at Partners HealthCare. EMR therapeutic range was calculated as the ratio of the doses in the 95th and 5th percentiles of all doses recorded for the medication during the study period. Therapeutic range for FDB was calculated as the ratio the maximum effective total dose per day (field DR2_MXDOSD) and the lowest effective total dose per day (field DR2_LODOSD) for adults (age $\geq 6,000$ days) in the Dose Range Check Module (DRCM). Therapeutic range for MDD was calculated as the ratio of the maximum and lowest doses specified for the medication in the dictionary.

Data Sources

Text of the notes and EMR medication data were obtained from Longitudinal Medical Record (LMR)¹³ – a CCHIT-certified EMR developed at Partners HealthCare. Partners HealthCare is an integrated healthcare delivery network in eastern Massachusetts that includes founding members Massachusetts General Hospital and Brigham and Women's Hospital, six other academic and community hospitals and a number of affiliated outpatient physician groups. LMR is used (primarily in the outpatient setting) by the majority of physicians affiliated with Partners Healthcare.

Medication intensification data were abstracted from the physician notes of patients with diagnosis of diabetes written between 01/01/2000 and 08/01/2005. Patients with diagnosis of diabetes were identified using a combination of billing data and computational analysis of narrative EMR notes as previously described¹⁴. Intensifications of the specific medication classes were abstracted only from the notes associated with a documented medication target above the treatment goals recommended prior to the onset of the study¹⁵ as follows: a) anti-hypertensive intensification: the lowest blood pressure documented in the note had to be $\geq 130/85$ mm Hg; b) anti-hyperglycemic intensification: the last HgbA1c documented prior to the note $\geq 7.0\%$; c) anti-hyperlipidemic intensification: the last LDL cholesterol documented prior to the note ≥ 100 mg/dL. Notes of physicians in primary care practices affiliated with the Brigham and Women's Hospital and Massachusetts General Hospital were included in the analysis.

Statistical Analysis

Summary statistics were constructed by using frequencies and proportions for categorical data and by using means and standard deviations for continuous variables. Fisher's Exact Test was used for comparison of proportions. To test the hypothesis that the association between the fraction of dose increases and therapeutic dose range is stronger for therapeutic ranges calculated from EMR medication data than from the dictionaries, we applied bootstrap over 10,000 cycles with bootstrap probabilities proportional to the number of intensification records.

IRB

The study protocol was reviewed and approved by the Partners Human Research Committee.

Results

Intensification Patterns among Individual Medications

We analyzed 161,333 notes of 6,142 patients with diabetes who had blood pressure, hemoglobin A1c or LDL cholesterol above the recommended treatment targets. In this dataset, 29,938 medication intensification events were identified. On average, dose increases represented 50.8% of all intensifications. For the majority (78%) of all medications, fraction of dose increases was between 30% and 80% (Figure 1). The highest fractions of dose increase (100%) were found for 2% of medications and 0.01% of number of records. The lowest fractions of dose increases (0%) were found for 8% of medications and 0.06% of number of records. Among medications with more than 100 records in the dataset, the highest fraction of dose increases was 67 % and the lowest 25%.

Intensification Patterns among Medication Classes

Patterns of medication intensification were not distributed equally among different medication classes (Table 1). Anti-hyperglycemic medications had the highest fraction of dose increases at 55.3% while anti-hyperlipidemic medications had the lowest at 32.3% ($p < 0.0001$).

Fraction of Dose Increase and Therapeutic Dose Range in EMR and Medication Dictionaries

To test the hypothesis that medical language reflects clinical reality we analyzed the association of a) the fraction of dose increases among all medication intensifications and b) therapeutic ranges calculated from EMR medication records vs. therapeutic ranges obtained from two medication dictionaries. Therapeutic ranges were calculated from 444,391 EMR medication records (on average 5419 per study medication). At the level of individual medication the correlation between the fraction of dose increases and therapeutic range weighted for the number of intensification records in the study dataset was strongest for the EMR (Figure 2).

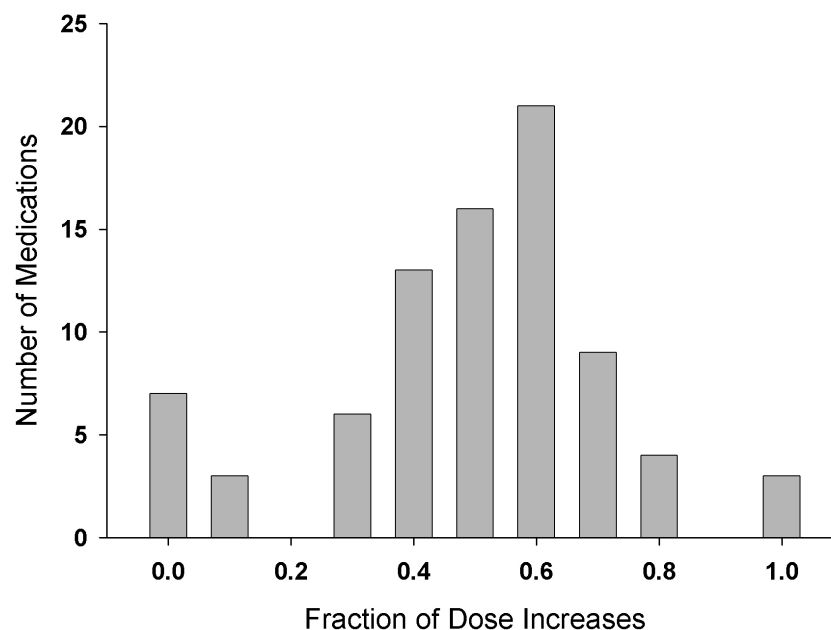


Figure 1. Distribution of Dose Increase Fraction Among Study Medications

Fraction of dose increases among all treatment intensifications was calculated for each of the 82 study medications. Number of medications whose fraction of dose increases fell into the same 10% bin was plotted on the Y axis.

Table 1. Dose Increase Fractions among Medication Classes

Medication Class	Dose Increase	
	Records, N	% (\pm 95% CI)
Anti-hyperlipidemic	1351	32.3 (\pm 1.92)
Anti-hypertensive	5504	51.3 (\pm 0.95)
Anti-hyperglycemic	8341	55.3 (\pm 0.79)

Correlations with the therapeutic ranges from either of the two medication dictionaries did not reach significance (Table 2). The association between fraction of dose increases and therapeutic ranges obtained from EMR was significantly different from both medication dictionaries (Table 2).

Table 2
Correlation between Increase Fraction and Therapeutic Range

Therapeutic Range Source	Correlation with Dose Increase Fraction		Different from EMR (p-value)
	Pearson	p-value	
EMR	0.39	0.0003	
MDD	0.171	0.125	0.013
FDB	0.187	0.0927	0.014

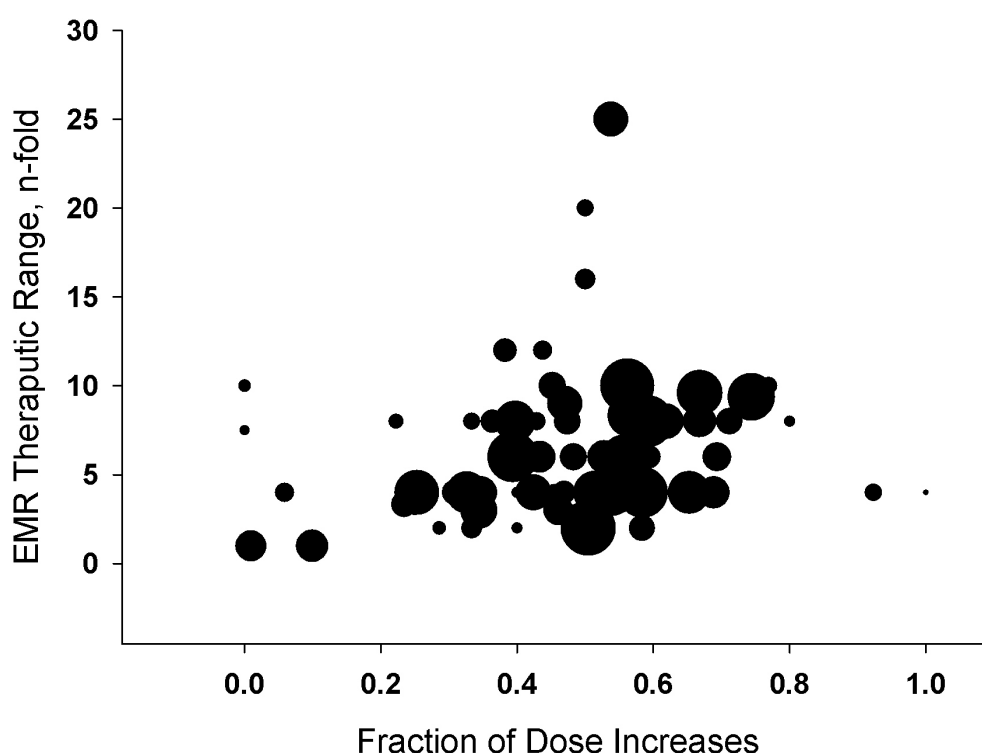


Figure 2. Fraction of Dose Increases and Therapeutic Range obtained from EMR Records

The diameter of each circle is proportional to the natural logarithm of the number of intensification events identified for that medication in the text of physician notes.

Discussion

In this retrospective analysis we showed that the lexical distribution of medication intensification concepts in narrative physician notes correlates directly with clinical reality as represented by electronic medication records. Furthermore, while the fraction of dose increases among medication intensifications in the text correlated well with the therapeutic range calculated from the electronic medication records, the correlation with therapeutic ranges provided by two medication dictionaries (which do not directly reflect clinical practice) did not reach statistical significance.

Language does not always directly reflect reality. Language that is semantically and / or syntactically complex may require transformation to ascertain the reality that it represents. For example, the language of fiction may contain metaphors and the language of diplomats may contain connotations that have been honed by decades or

centuries of protocol¹⁶ but are not known to outsiders. Medical language does not appear to have these extra layers of complexity and can therefore be directly linked to the clinical reality it describes.

At the same time, establishing connection between medical language and reality presents its own set of challenges. Medical language is relatively syntactically poor, leading to ambiguities that require semantic context for their resolution. For example, a phrase “Fenofibrate 48 mg qd” may represent a) a medication the patient was taking in the past if a part of a recorded pre-admission medication list in a hospital admission note; b) a medication the patient is currently taking, if a part of a medication list in a progress note or c) a medication that is being initiated (i.e. a medication intensification event) if it is found in the Plan section of the note.

Connecting medical language to reality also commonly requires background knowledge, or linguistic pragmatics³. For example, by convention vital signs are frequently documented together. Consequently, a combination of two numbers separated by a forward slash in a sentence that refers to “weight” or “pulse” is likely to represent a patient’s blood pressure, even if the word blood pressure or a corresponding acronym are omitted.

As becomes apparent from the above examples, the challenges in linking medical language to reality are primarily of semantic nature. Consequently, semantic theories of language, such as Semantic Frame Theory¹⁷ or Natural Semantic Metalanguage¹⁸, can be helpful when designing NLP systems for medical language. Both approaches were reflected in a number of works in the fields of computational linguistics and general NLP^{19,20}.

In our study, we employed semantic analysis techniques¹¹ to show that the distribution of lexical concepts representing medication intensification corresponds to the therapeutic ranges of the same medications calculated from medication records in the EMR. Our hypothesis was based on the assumption that medications with larger therapeutic range (i.e. greater fold-difference between the maximum and the lowest effective doses) would likely first be started at a lower dose and subsequently increased in a stepwise fashion until the desired effect is achieved or the maximum tolerated dose is reached. On the other hand, medications with narrow therapeutic ranges – some of which may only have one commonly used dose – would have fewer, if any, dose increases recorded, compared with initiations. The hypothesis was confirmed at both individual medication and therapeutic class levels: anti-hyperglycemic (e.g. insulins) and anti-hypertensive medications generally have wider therapeutic ranges than anti-lipidemic agents and had higher fraction of dose increases recorded in our dataset.

We also included in our analysis therapeutic ranges obtained from two medication dictionaries – one internal, another one a commercial dictionary commonly used worldwide – as negative controls. If the fraction of dose increases documented in the notes was related to therapeutic ranges for reasons other than being a reflection of clinical practice, this relationship would have been as strong with the dictionaries as it was with the EMR data. That was not the case. The lower / non-significant correlation with the therapeutic ranges from the dictionaries was primarily due to medications whose maximum and lowest doses as recorded in the dictionary did not adequately represent common clinical practice (e.g. maximum dose of 10 units for insulin lispro).

Our study has a number of strengths. It is one of the first investigations to directly analyze the relationship between medical language and clinical reality. Our results are based on the analysis of over 150,000 patient notes and nearly half a million electronic medication records of 82 medications of three different therapeutic classes. This large-scale investigation therefore provides a fundamental basis for using information obtained from narrative medical documents as a valid data source.

Our results also have several limitations. They were focused on a narrow topic of medication intensification and the findings may not be applicable to other clinical domains. Narrative provider notes in the EMR served as the sole data source. Consequently, the results may not be generalizable to other types of narrative medical documents including discharge summaries, imaging and pathology reports, etc. Electronic medication records themselves may not be a precise reflection of clinical practice. This may have in part accounted for the fact that the correlation coefficient, while highly significant, was only 0.39. Further research is necessary to validate our findings on other sources of data on clinical practice and in other areas of medical language processing.

Conclusion

In this large-scale retrospective analysis of medication intensification data obtained from narrative provider notes we were able to show that the patterns of lexical concept distribution for medication intensification differ between individual medications and medication classes. There is a significant correlation between the fraction of dose increases among all medication intensifications identified in narrative text and clinical practice as represented by the therapeutic range calculated from electronic medication records. These findings provide evidence for a direct link between narrative medical text and clinical reality and support the use of narrative text as a valid source of medical data.

Acknowledgements

This research was supported in part by funding from the Agency for Healthcare Research and Quality (R18 HS017030).

References

1. Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. Unlocking clinical data from narrative reports: a study of natural language processing. *Ann Intern Med* 1995;122:681-8.
2. Hartzband P, Groopman J. Off the record--avoiding the pitfalls of going electronic. *N Engl J Med* 2008;358:1656-8.
3. Wierzbicka A. Cross-cultural pragmatics: The semantics of human interaction: Walter de Gruyter; 1992.
4. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. The Diabetes Control and Complications Trial Research Group. *N Engl J Med* 1993;329:977-86.
5. Tight blood pressure control and risk of macrovascular and microvascular complications in type 2 diabetes: UKPDS 38. UK Prospective Diabetes Study Group. *Bmj* 1998;317:703-13.
6. Colhoun HM, Betteridge DJ, Durrington PN, et al. Primary prevention of cardiovascular disease with atorvastatin in type 2 diabetes in the Collaborative Atorvastatin Diabetes Study (CARDS): multicentre randomised placebo-controlled trial. *Lancet* 2004;364:685-96.
7. Resnick HE, Foster GL, Bardsley J, Ratner RE. Achievement of American Diabetes Association clinical practice recommendations among U.S. adults with diabetes, 1999-2002: the National Health and Nutrition Examination Survey. *Diabetes Care* 2006;29:531-7.
8. Grant RW, Buse JB, Meigs JB. Quality of diabetes care in U.S. academic medical centers: low rates of medical regimen change. *Diabetes Care* 2005;28:337-442.
9. Turchin A, Shubina M, Breydo E, Pendergrass ML, Einbinder JS. Comparison of Information Content of Structured and Narrative Text Data Sources on the Example of Medication Intensification. *J Am Med Inform Assoc* 2009.
10. Berlowitz DR, Ash AS, Hickey EC, et al. Inadequate management of blood pressure in a hypertensive population. *N Engl J Med* 1998;339:1957-63.
11. Turchin A, Kolatkar NS, Grant RW, Makhni EC, Pendergrass ML, Einbinder JS. Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes. *J Am Med Inform Assoc* 2006;13:691-5.
12. Turchin A, Shubina M, Breydo E, Pendergrass ML, Einbinder JS. Comparison of information content of structured and narrative text data sources on the example of medication intensification. *J Am Med Inform Assoc* 2009;16:362-70.
13. Shah NR, Seger AC, Seger DL, et al. Improving acceptance of computerized prescribing alerts in ambulatory care. *J Am Med Inform Assoc* 2006;13:5-11.
14. Turchin A, Kohane IS, Pendergrass ML. Identification of patients with diabetes from the text of physician notes in the electronic medical record. *Diabetes Care* 2005;28:1794-5.
15. Standards of medical care for patients with diabetes mellitus. *Diabetes Care* 2000;23 Suppl 1:S32-42.
16. Fenton-Smith B. Diplomatic condolences: ideological positioning in the death of Yasser Arafat. *Discourse & Society* 2007;18:697.
17. Fillmore CJ. Frame semantics. *Cognitive Linguistics: Basic Readings* 2006:185-238.
18. Wierzbicka A. *Semantics: Primes and universals*: Oxford University Press; 1996.
19. Talmy L. *Toward a cognitive semantics*: MIT; 2000.
20. Nirenburg S, Raskin V. *Ontological semantics*: MIT Press Cambridge, Mass; 2004.