Jarvis: A Software Tool for Automatically Identifying Substances and Doses From Online Commentary

Michael Chary, MD PhD

Instructor Division of Medical Toxicology, Department of Emergency Medicine
Weill Cornell Medicine, New York Presbyterian Hospital, New York, NY

The Problem

Substance Use is Challenging to Study in Formal Trials

Online Data Provides Real-World Evidence

Prior Findings:

Self-treatment: Kratom, Ibogaine, Ketamine, Psilocybin

Prevalence of cross-titration

Efficacy of gabapentin, not benzos

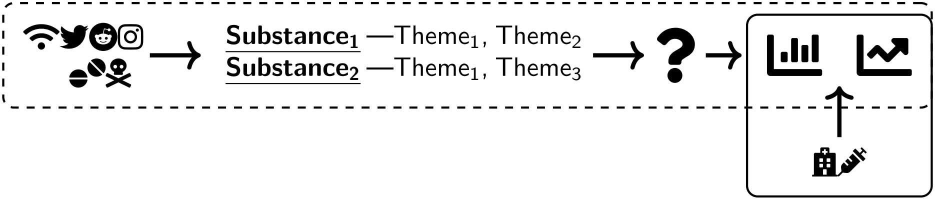
Barriers to MAT, using naloxone kits, accessing care

Limitation to Prior Approaches:

Qualitative, not quantitative

Manual: Time-consuming, not scalable

Key Barrier to Progress:
Qualitative Data from Social Media
Quantitative from Clinical Research



Substance$_1$ —Theme$_1$, Theme$_2$

Substance$_2$ —Theme$_1$, Theme$_3$

If we could estimate the **doses** of substances that people describe using, we could:

Generate hypotheses about the real-world use of medications, including off-label uses, dosing, and side effects.

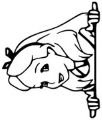
And Test these hypotheses with real-world data.

Weill Cornell Medicine
Emergency Medicine

Jarvis, A Solution in Two Parts: Grammar & Entity Recognition

Grammar allows us to identify drugs and doses without knowing the names of the drugs beforehand.

Drugs are nouns, usually uncountable. If used as countable nous, shorthand for dosage. *Two oxycodone.
Doses are *noun phrases with measure words*



I snort 4 m30s a day. Happy nods.

- ✓ m30s is an *entity* (substance)
- ✓ 4 m30s a day is a *noun phrase* (dosage)
- ✗ snort is a verb
- ✗ happy is an adjective (modifier)

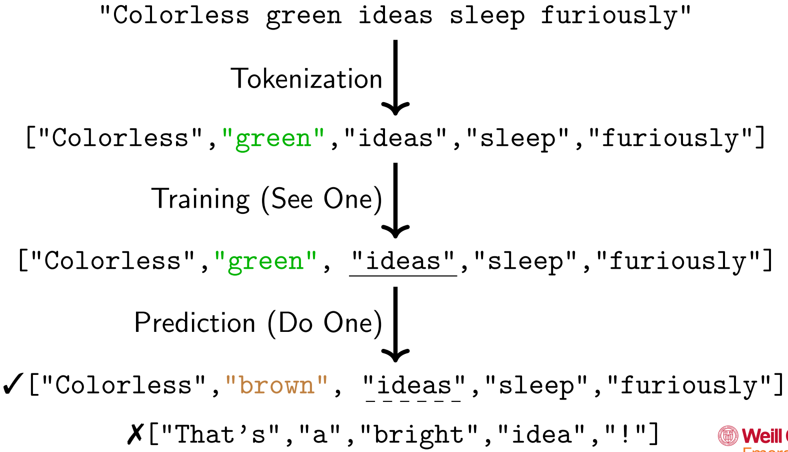
I slithied 4 flors a gorlte.

- ✓ flors is an *entity* (substance)
- ✓ 4 flors a gortle is a *noun phrase* (dosage)
- ✗ slithied is a verb

Weill Cornell Medicine
Emergency Medicine

Jarvis, A Solution in Two Parts: Grammar & Entity Recognition

Entity recognition extracts drugs and doses communicated in in ways our grammar rules don't yet cover.



"Colorless green ideas sleep furiously"

Tokenization

["Colorless","green","ideas","sleep","furiously"]

Training (See One)

["Colorless","green", "ideas","sleep","furiously"]

Prediction (Do One)

✓["Colorless","brown", "ideas","sleep","furiously"]

✗["That's","a","bright","idea","!"]

The Data for Model Development

Reddit forum (Subreddit)	No. of unique posts
r/opiates	62.138
r/heroin	79.851
r/fentanyl	10.816
r/suboxone	97.551
r/OpiatesRecovery	76.888
r/OurOverusedVeins	79.581
Total	406.825

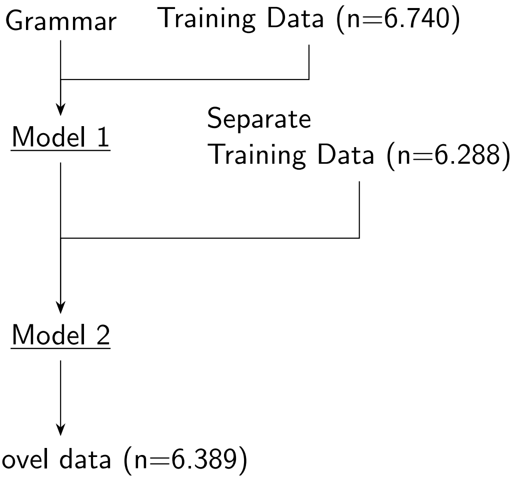
Posts and comments from 2010 to 2023

Excluded posts with fewer than 5 words, no entities, or duplicates

Training & Testing Jarvis

	Substances	**Doses**
Precision	84%	58%
Recall	86%	62%
Sensitivity	92%	49%
Evaluation Data (n=6.436)		

	Substances	**Doses**
Precision	86%	86%
Recall	87%	81%
Sensitivity	96%	89%
Re-Evaluation Data (n=6.389)		

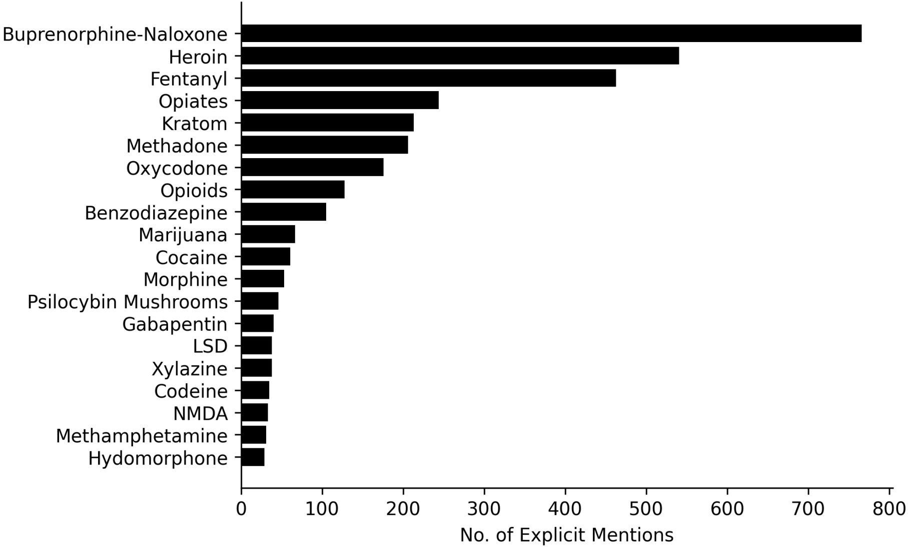


Comparison with Other Methods

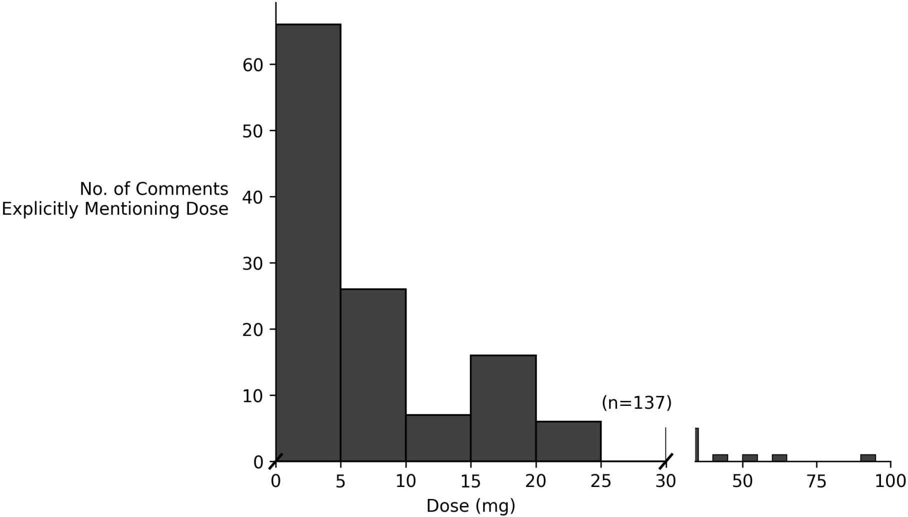
Method	Substances	Doses
Jarvis (NER + Grammar)	85%	86%
Spacy NER	78%	68%
Stanford NER	72%	62%
ClinicalNLP	54%	52%

F1 score, geometric mean of precision and recall.

Frequency of Substances

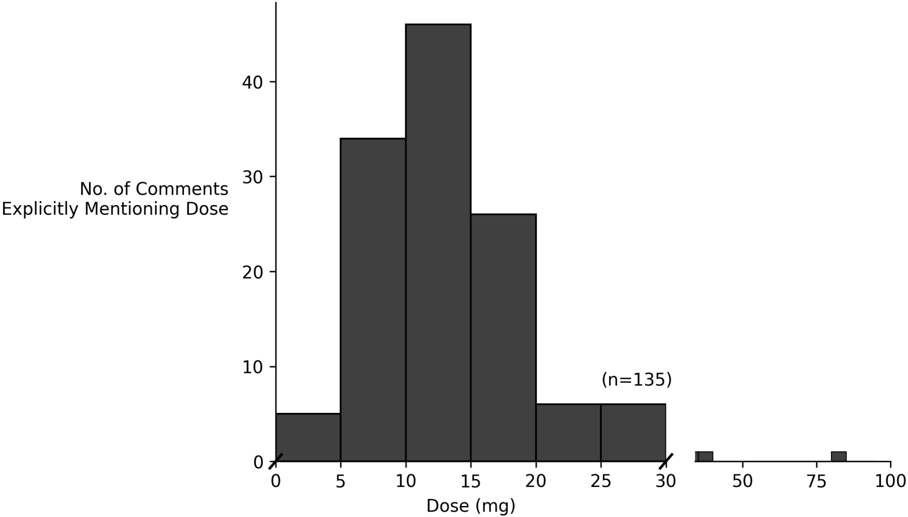


Distribution of Doses for Suboxone



Weill Cornell Medicine
Emergency Medicine

What Doses of Kratom Do People Report Using?



Weill Cornell Medicine
Emergency Medicine

Summary

First tool to extract doses automatically from online commentary

Largest samples of real-world distribution of doses and medications

Combination of grammar and entity recognition outperforms either alone

Next Steps:

Dosage over time, geographical variation

Dose-response associations

Extract effects, descriptions spanning multiple comments

Comparison to clinical data

Acknowledgements

Jai Kapoor

Ali Abdelati MD (PGY-1), Roland Zemla, MD PhD (PGY-1), Caitlin House, Svetlana Ross, Aiden Peleg (MS1)

WCM: Rahul Sharma, MD, MBA; Junaid Razzak, MBBS DrPH, Judy Poremba

ACMT: Paul Wax, MD; Kim Aldy, MD

Ed Boyer, MD, PhD; Alex Manini, MD, MS

Weill Cornell Medicine
Emergency Medicine