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

Michael Chary, MD PhD

✉ mic9189@med.cornell.edu   🐦 @macacharya

 mac389   🌐 mac389.github.io

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

# Online Data Provides Real-World Data

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:

 Manual: Time-consuming, not scalable
 Qualitative, not quantitative
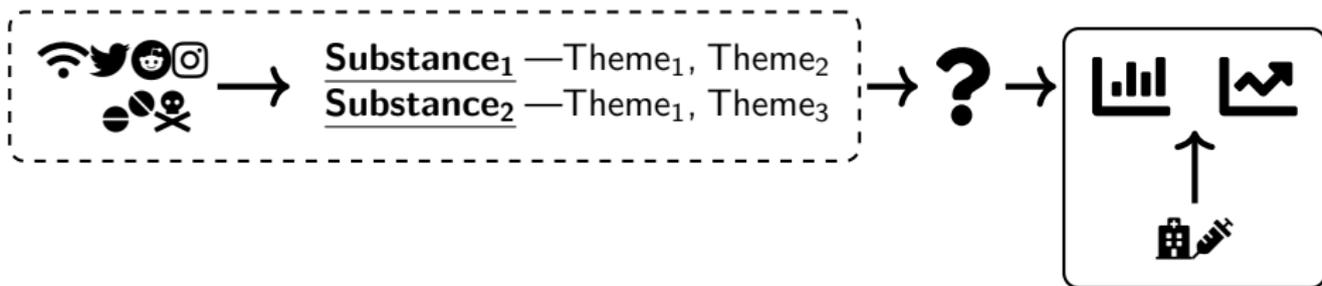
Dose-response curve $\Rightarrow$ substance mediates an effect.

If we could extract dose-effect associations from social media, we could:

directly **compare** real-world data with data from controlled studies

test quantitative hypotheses with real-world data.

ultimately, **infer** substance efficacy from real-world data

Weill Cornell Medicine
Emergency Medicine

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

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

Drugs are uncountable nouns.
If used as countable nouns, shorthand for dosage. *Two oxycodone.
Doses are also *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

```
"I snort 4 m30s every Wednesday."
```

↓

```
["I","snort","4","m30s","every","Wednesday"
```

↓

```
"Every weekend, 4 m30s"
```

Entity recognition is a *supervised machine learning* task.

The Data

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

Creating 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)		

Grammar   Training Data (n=6.740)

Model 1

Separate
Training Data (n=6.288)

Model 2

Apply to novel data (n=6.389)
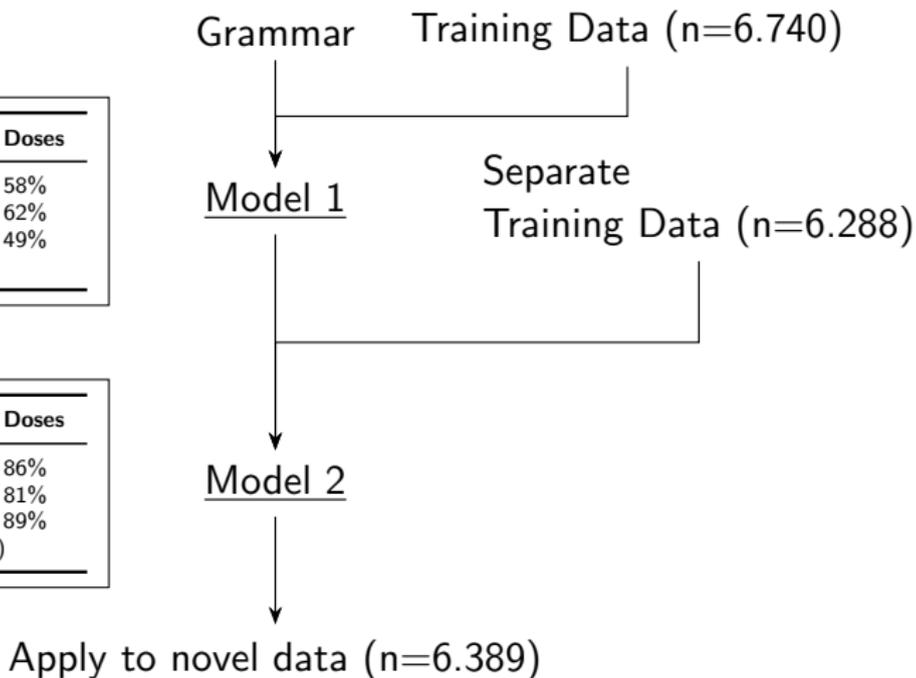
Jarvis Outperforms Other Methods

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

No. of Explicit Mentions

Emergency Medicine
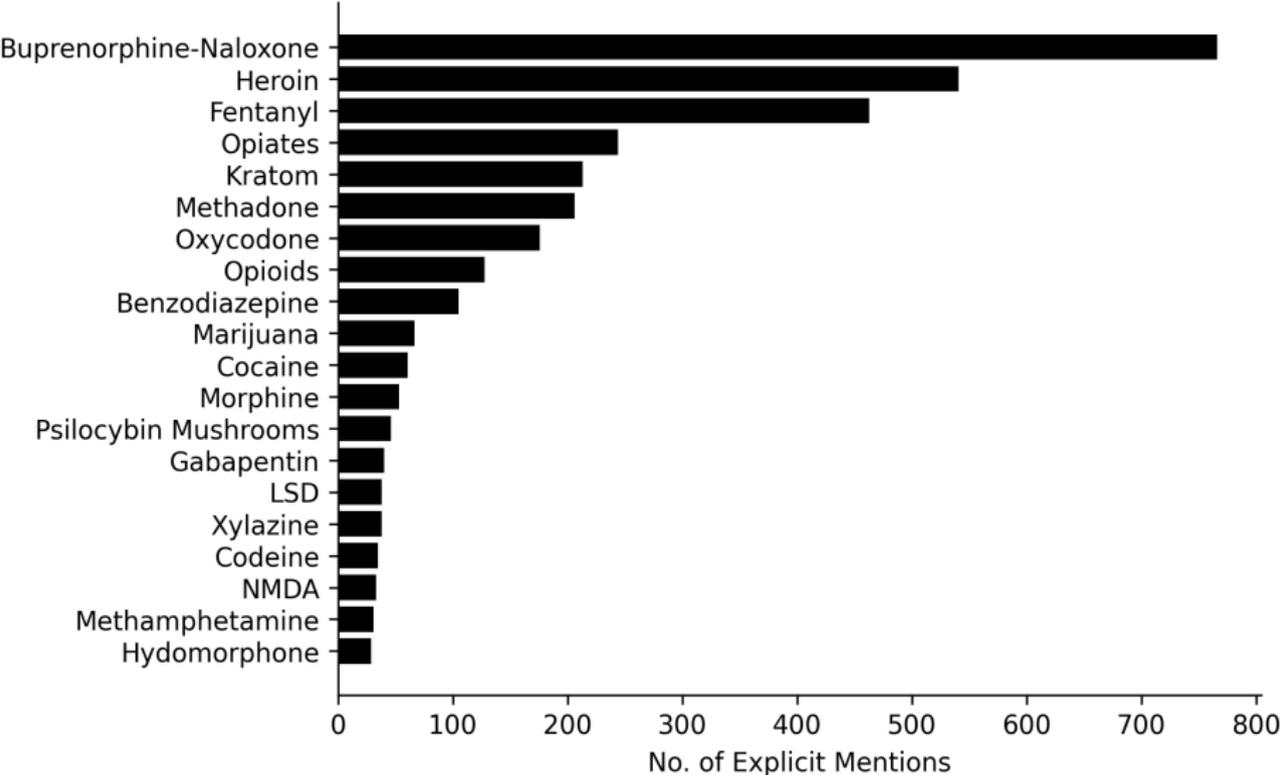
Distribution of Doses for Suboxone

Y-axis: No. of Comments Explicitly Mentioning Dose

X-axis: Dose (mg)

(n=137)

Weill Cornell Medicine
Emergency Medicine
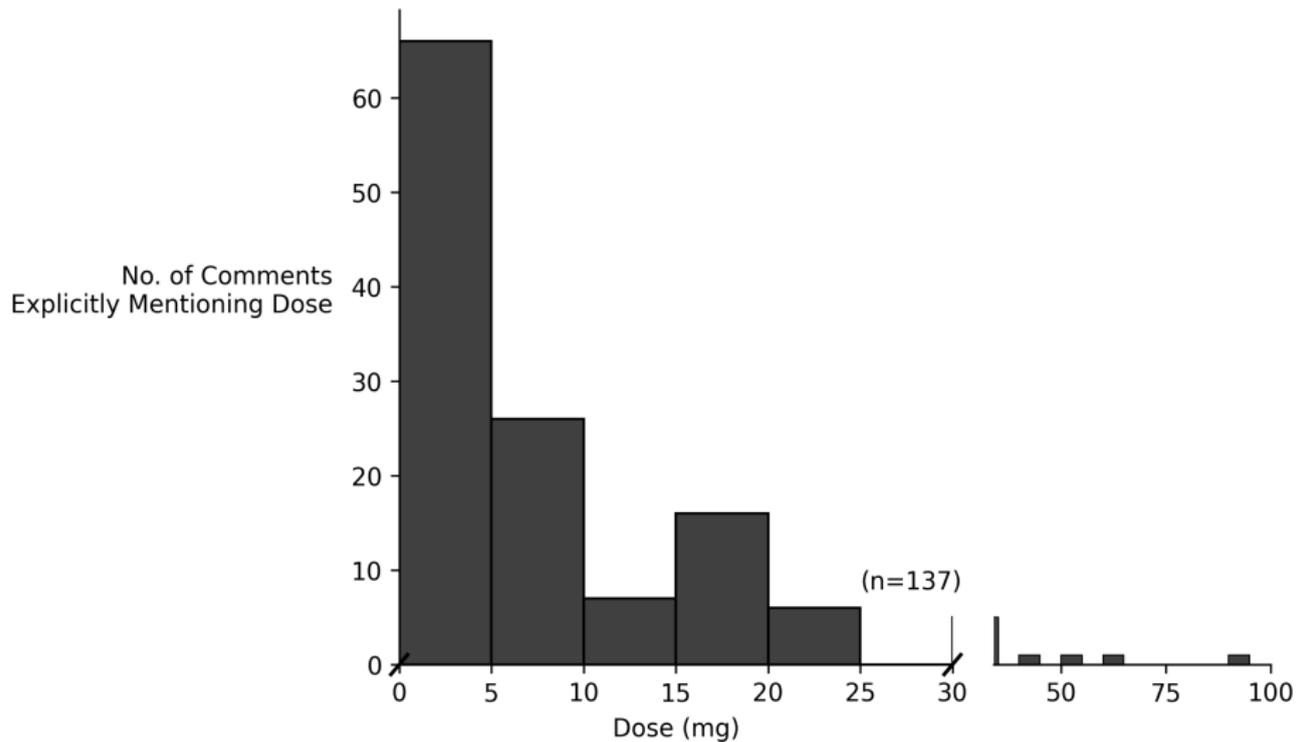
What Doses of Kratom Do People Report Using?

Histogram with y-axis labeled "No. of Comments Explicitly Mentioning Dose" (0 to 40+) and x-axis labeled "Dose (mg)" (0 to 100). (n=135)

Weill Cornell Medicine
Emergency Medicine
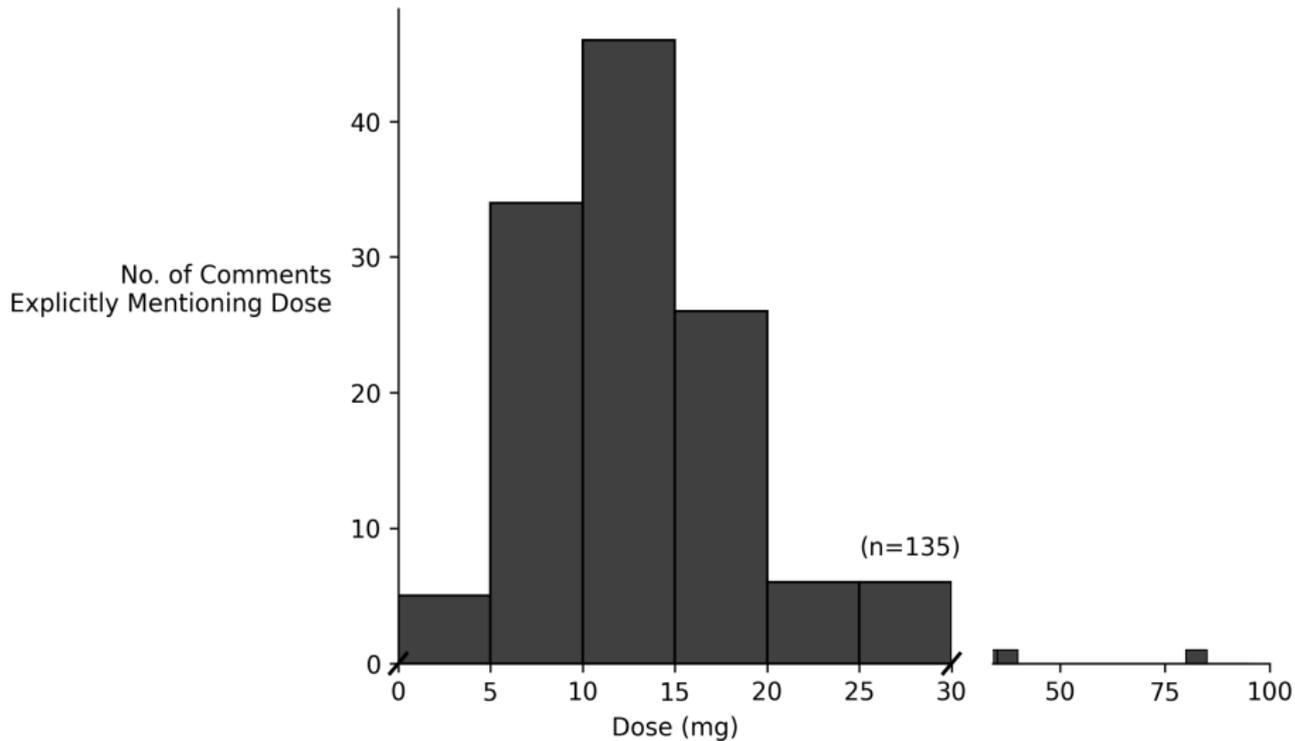
Summary

Extract doses automatically from online commentary

Combination of grammar and entity recognition outperforms either alone

Next Steps:

  Dose-response associations
  Extract effects, descriptions spanning multiple comments
  Comparison to clinical data
  Dosage over time, geographical variation

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