medExtractR: medication extraction from electronic health records

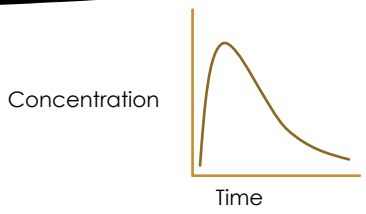
HANNAH WEEKS

VANDERBILT UNIVERSITY

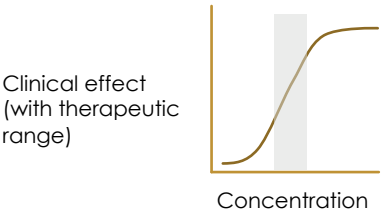
Context

PK/PD studies

- Pharmacokinetics (PK): how the body affects the drug
  - Estimate parameters to explain drug movement within the body

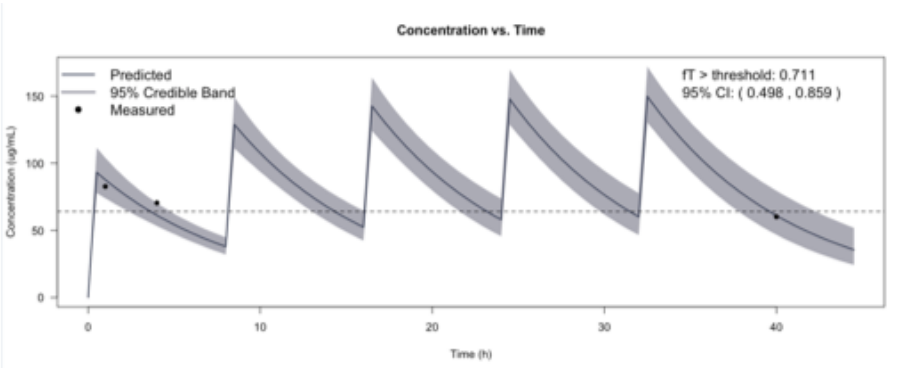


- Pharmacodynamics (PD): how the drug affects the body
  - Dose-response relationships, adverse events



Pharmacokinetic studies

- Feedback through blood concentration measurements



PK/PD studies

Critical information for PK/PD studies:

- Blood concentration measurements
- Dosing information
  - What dosage did the patient take?
  - How often was that dosage taken?
  - At what time was the most recent dose taken?

PK/PD studies

Critical information for PK/PD studies:

- Blood concentration measurements
- Dosing information
  - What dosage did the patient take?
  - How often was that dosage taken?
  - At what time was the most recent dose taken?

Often found as unstructured data in EHR (clinical notes)

"…Patient takes tacrolimus 1mg 2x/day…"

Information extraction

- How do we extract data?
- Natural language processing (NLP)
  - Using computers to understand human language
- Information extraction
  - NLP task that converts unstructured input to structured output

"…Patient takes tacrolimus 1mg 2x/day…"

Drug name	tacrolimus
Dose	1 mg
Frequency	2x/day

medExtractR

- Targeted approach to medication extraction – intended to be used on **a** drug within a dataset
- Customizable through function arguments or modification of source code
- Written in R
  - Widely used for data analysis
  - Available on CRAN

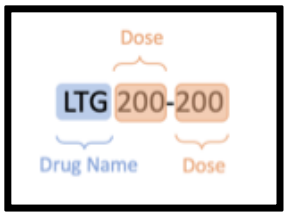
medExtractR

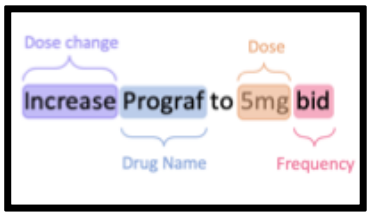
SYSTEM DESCRIPTION

Medication Entities

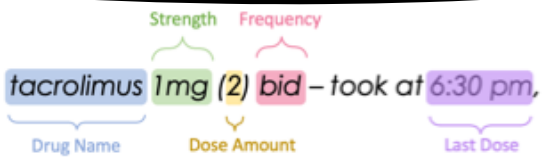
- **Drug name**
- **Strength**: Amount of an individual unit (pill)
- **Dose amount**: number of units taken
- **Dose**: dose given intake (equivalent to strength x dose amount)
- **Frequency**: how often dose is taken
- **Intake time**: relative time of day when dose is taken
- **Dose change**: keyword indicating if dose is an increase, decrease, etc.
- **Last dose**: time at which the last dose was taken

Medication Entities



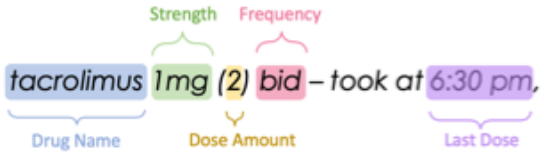


Dictionary-based entities



- Frequency
  - Any expression in the dictionary that is also in the search window will be extracted
  - Can be regular expressions
    - E.g., 'q\\s?day' will match 'qday' or 'q day'
- Default dictionaries: `data(freq_vals)`, `data(intaketime_vals)`, `data(dosechange_vals)`

Rule-based entities



- Strength
  - **'Number unit'**
  - Function argument (unit = 'mg')
- Dose amount
  - **'#** (pill | tablet | capsule)'
  - 'take | takes | taking **#**'
  - '(**#**)'
- Last dose
  - **Time expression**
    - ## am/pm
    - Military time (e.g., 2100)
    - Modifier (e.g., 10 last night)
  - Window includes 'last | took | taken'

medExtractR functionality

Input
clinical note, drug names,
tuning parameters

Internal
- find drug names
- create search window
- identify/extract drug entities

Output
data frame with drug entities

medExtractR example

```
medExtractR(note, drug_names = c("tacrolimus", "prograf", "tac"), unit="mg",
            window_length=60, max_dist=2)
```

"Patient is on tacrolimus 1mg (2) bid – took at 6:30 pm, cellcept 1000mg bid, prednisone 5mg daily."

Entity	Expression	Position
DrugName	tacrolimus	15:25
Strength	1mg	26:29
DoseAmt	2	31:32
Frequency	bid	34:37
LastDose	6:30 pm	48:55

medExtractR

EVALUATION

Data

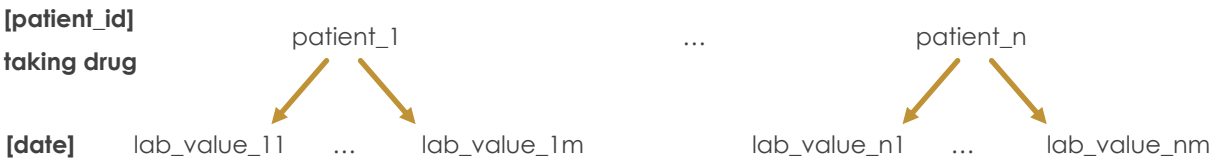
From the Synthetic Derivative – Vanderbilt University de-identified EHR

- Development drugs
  - Tacrolimus and lamotrigine
  - 60 training notes, 50 test notes
- Test drug
  - Allopurinol
  - 110 test notes

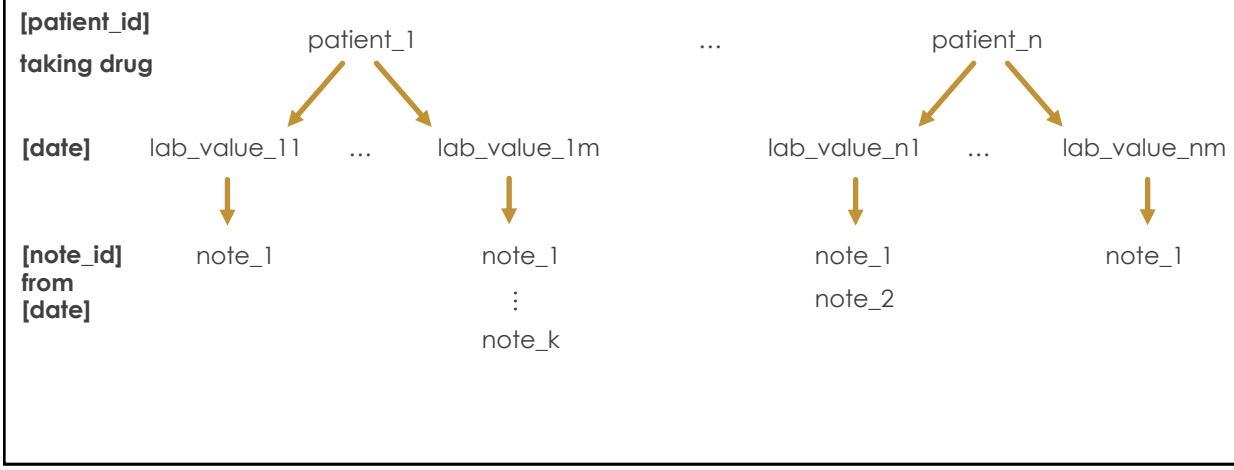
Data: train/test set selection

[patient_id] taking drug	patient_1	…	patient_n

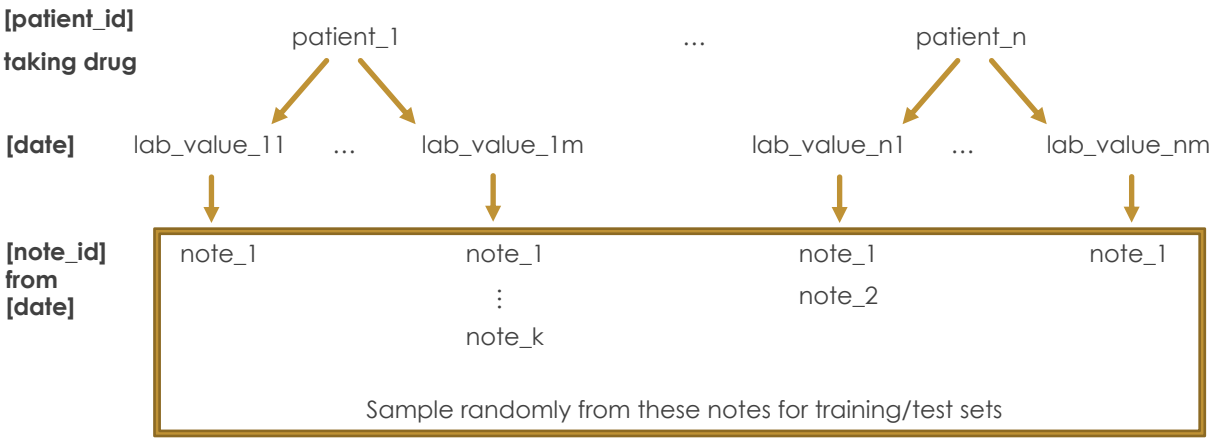
Data


[patient_id]
taking drug
patient_1
...
patient_n
[date]
lab_value_11
...
lab_value_1m
lab_value_n1
...
lab_value_nm

Data


[patient_id]
taking drug
patient_1
...
patient_n
[date]
lab_value_11
...
lab_value_1m
lab_value_n1
...
lab_value_nm
[note_id]
from
[date]
note_1
note_1
⋮
note_k
note_1
note_2
note_1

Data

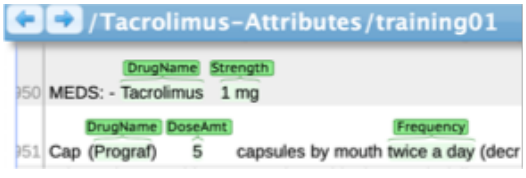


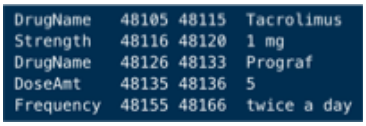
Data: gold standards

BRAT (Brat Rapid Annotation Tool)
used to identify correct drug information

Input: clinical note

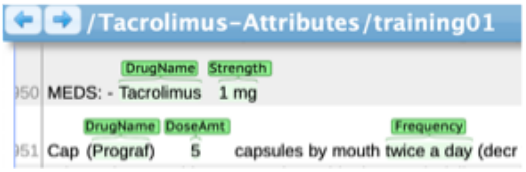
Expected output (for supervised learning):
gold standard annotations





Data: gold standards

- 1. Develop annotation guidelines
  - When to highlight information
  - What defines different drug entities
- 2. Double annotation
  - 2 independent reviewers, evaluate annotation concordance
- 3. Revise guidelines if needed
- 4. Annotate training notes
- 5. Annotate test notes



Performance measures

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

- Positive predictive value
- Fraction of extracted output in gold standard

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

- Sensitivity (true positive rate)
- Fraction of annotations that were correctly extracted

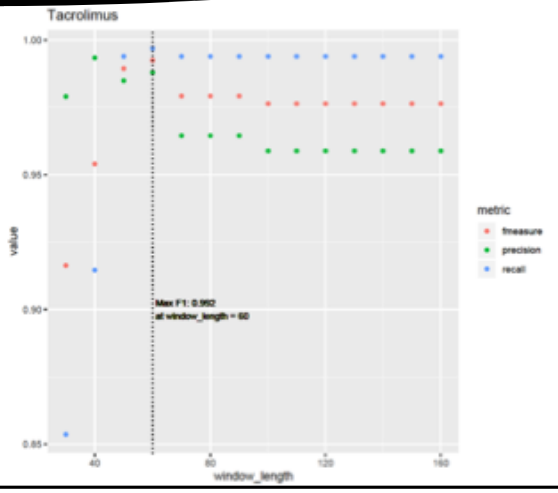
Performance measures

$$\text{F} - \text{measure (F1)} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

- Quantify uncertainty with 95% bootstrap confidence intervals
  - Bootstrap notes (within drug)
  - Use 2.5 and 97.5 percentiles as interval bounds

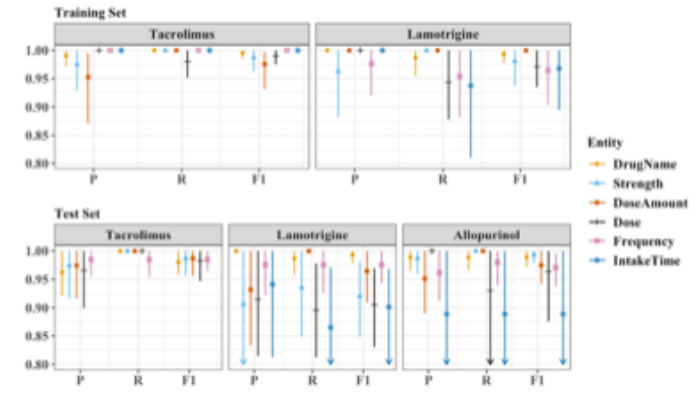
Selecting tuning parameters

- Tune two parameters: window length and maximum edit distance
  - Create a grid of options for each parameter
  - Compute F-measure
  - Select parameters with best performance
- Maximize performance on training set



medExtractR: results

- Tacrolimus prescription patterns are much more simple than lamotrigine
- Allopurinol tested using tacrolimus tuning parameters



medExtractR Extractions	DrugName	Strength	DoseAmount	Dose	Frequency	IntakeTime
Tacrolimus (training)	106	40	42	51	84	6
Lamotrigine (training)	77	27	39	53	44	32
Tacrolimus (test)	79	39	39	29	67	0
Lamotrigine (test)	77	34	44	52	42	39
Allopurinol (test)	187	79	84	35	106	12

medExtractR

IMPLEMENTATION IN MIMIC-III CLINICAL CARE DATABASE

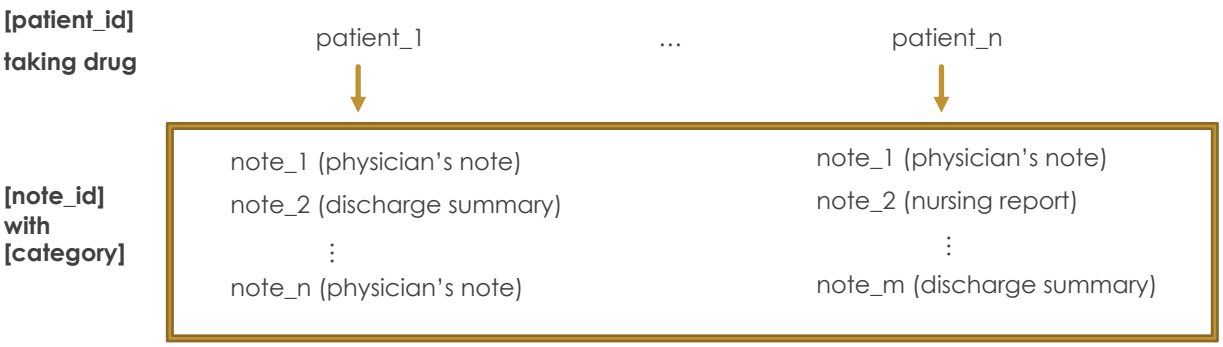
MIMIC-III Clinical Care Database

- De-identified records corresponding to over 60,000 ICU stays
- Over 2 million clinical notes
- Institution: Beth Israel Deaconess Medical Center, Boston, MA

- Dataset available by request through MIT: https://mimic.physionet.org

Johnson AE, Pollard TJ, Shen L, Li-wei HL, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. Scientific data. 2016 May 24;3:160035. doi: 10.1038/sdata.2016.35

Data



Categories have different likelihood of containing dose information

Note sampling procedure

3 drugs: tacrolimus, lamotrigine, oxcarbazepine

- **Tuning set**: 10 notes per drug
  - Randomly select notes one at a time
  - Manually review for presence of dosing information
  - If present, add to tuning set

- Determine changes to annotation guidelines
- Annotate gold standards

- **Validation set**: 100 notes per drug
  - Randomly sample 50 discharge summaries
  - Randomly sample 50 from all other note categories

- Annotate gold standards after tuning

MIMIC-III: Starting point

- Tuning set errors motivate next steps
- e.g. *Tacrolimus 1 mg: One (1) capsule q daily*

Position	medExtractR	Gold Standard
1:11	Tacrolimus	Tacrolimus
12:16	1 mg	1 mg
18:21	<NA>	One
23:24	1	1
34:41	<NA>	q daily

False negatives

Evaluation method

Present performance for different quantities for each drug:

- 1. No modification – "out of box" performance based on SD development
- 2. Tuning only
  - Smaller changes (dictionary updates, parameter selection)
- 3. Tuning plus customization
  - Adding or changing rules in the source code
  - Requires more advanced coding ability

Evaluation method

Present performance for different quantities for each drug:

- 1. No modification – "out of box" performance based on SD development
- 2. Tuning only
  - **Add 'q daily' to frequency dictionary**
  - **Select parameter values for function arguments**
- 3. Tuning plus customization
  - **Text number followed by (digit) is a dose amount**

MIMIC-III evaluation

- Performance with no tuning is not ideal
- F-measures still above 0.80 benchmark

	Tacrolimus (n = 423 annotations)		
	No tuning	Tuning only	Tuning plus customization
Precision	.96 [.92, .99]	.93 [.89, .96]	.95 [.91, .98]
Recall	.77 [.71, .83]	.81 [.76, .85]	.89 [.84, .94]
F-measure	.85 [.81, .89]	.86 [.83, .90]	.92 [.88, .95]
	Lamotrigine (n = 381 annotations)		
	No tuning	Tuning only	Tuning plus customization
Precision	.87 [.82, .92]	.93 [.89, .97]	.94 [.90, .98]
Recall	.81 [.77, .85]	.83 [.78, .87]	.92 [.87, .96]
F-measure	.84 [.81, .87]	.88 [.84, .91]	.93 [.89, .96]
	Oxcarbazepine (n = 375 annotations)		
	No tuning	Tuning only	Tuning plus customization
Precision	.79 [.72, .86]	.97 [.94, .99]	.97 [.95, .99]
Recall	.83 [.79, .87]	.85 [.80, .89]	.92 [.88, .96]
F-measure	.81 [.76, .85]	.90 [.87, .93]	.95 [.92, .97]

MIMIC-III evaluation

- Some improvement with tuning alone
  - Higher for lamotrigine and oxcarbazepine
- Largest improvement with tuning plus customization

	Tacrolimus (n = 423 annotations)		
	No tuning	Tuning only	Tuning plus customization
Precision	.96 [.92, .99]	.93 [.89, .96]	.95 [.91, .98]
Recall	.77 [.71, .83]	.81 [.76, .85]	.89 [.84, .94]
F-measure	.85 [.81, .89]	.86 [.83, .90]	.92 [.88, .95]
	Lamotrigine (n = 381 annotations)		
	No tuning	Tuning only	Tuning plus customization
Precision	.87 [.82, .92]	.93 [.89, .97]	.94 [.90, .98]
Recall	.81 [.77, .85]	.83 [.78, .87]	.92 [.87, .96]
F-measure	.84 [.81, .87]	.88 [.84, .91]	.93 [.89, .96]
	Oxcarbazepine (n = 375 annotations)		
	No tuning	Tuning only	Tuning plus customization
Precision	.79 [.72, .86]	.97 [.94, .99]	.97 [.95, .99]
Recall	.83 [.79, .87]	.85 [.80, .89]	.92 [.88, .96]
F-measure	.81 [.76, .85]	.90 [.87, .93]	.95 [.92, .97]

Takeaways

- Without tuning, medExtractR performance is likely to be less than ideal, especially if building datasets for medication studies.
- Recommend at least performing tuning steps when using medExtractR for a new study. Customization is ideal, when possible.
- **medExtractR approach provides a compromise between relying on "out-of-box" performance of existing medication extraction systems and having to manually create a validated dataset.**

Contact

- Email: hannah.L.weeks@Vanderbilt.edu
- Weeks HL, Beck C, McNeer E, Williams ML, Bejan CA, Denny JC, Choi L. medExtractR: A targeted, customizable approach to medication extraction from electronic health records. *J Am Med Inform Assoc.* 2020; 27(3):407-18.