



insight into value



ADVERSE DRUG REACTIONS ON SOCIAL MEDIA: BIAS AND LIMITATION

PhUSE EU Connect 2018 – Paper DH05



KEYRUS
biopharma

AGENDA

Context

- Context
- Previous paper
- Goal

Method

- Data sources
- Languages
- Metadata

Results & Discussion

- Results
- Discussion



AGENDA

Context

- Context
- Previous paper
- Goal

Method

- Data sources
- Languages
- Metadata

Results & Discussion

- Results
- Discussion



CONTEXT – PREVIOUS PAPER

“Patient-generated Health Data (Social Media) is a Potential Source for ADR Reporting”

Goal



*Adverse Drug Reaction

Sources

Twitter[®], a micro-blogging social network forum

Reddit[®], a widely used



reddit

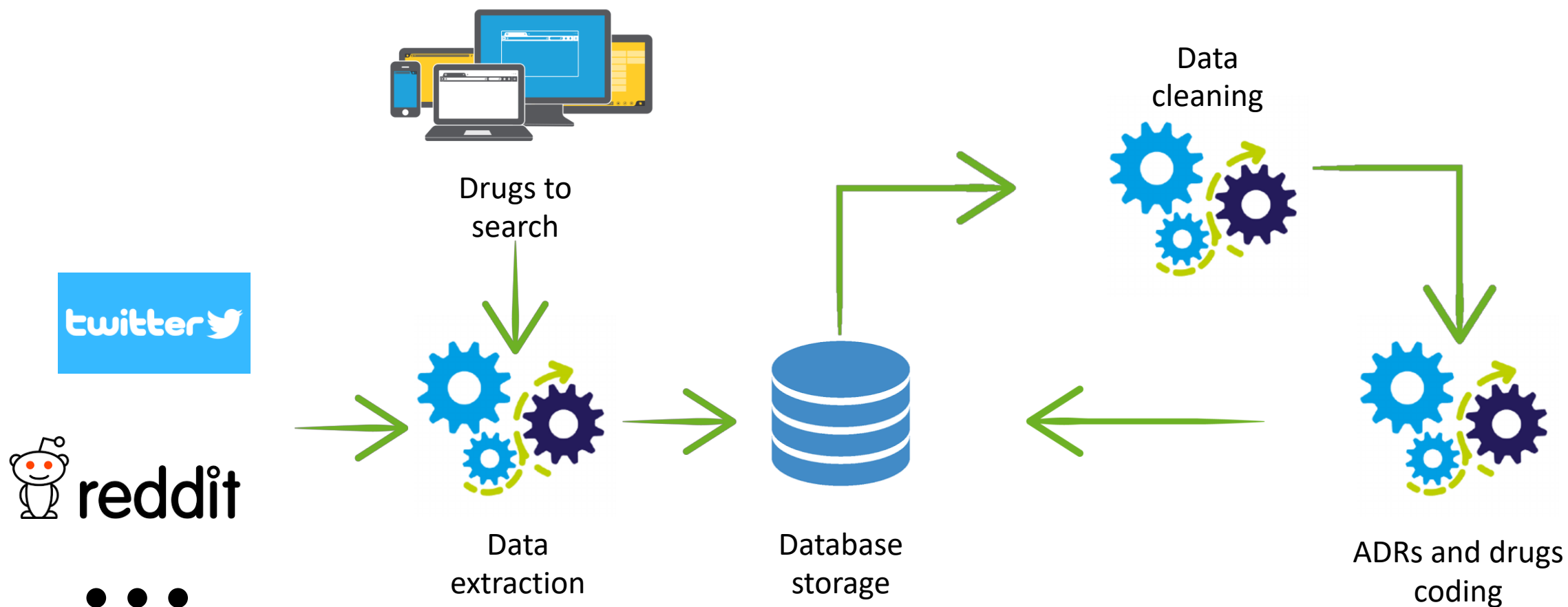
ADR detection possible but subject to limitation:

1) type of sources

2) languages

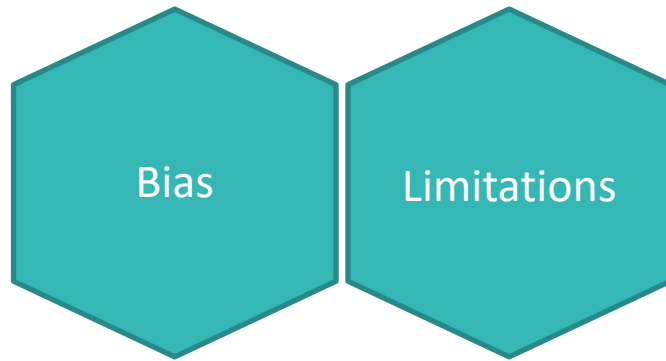
3) method

PROCESS FLOW



CONTEXT - GOAL

Find & Estimate



Of doing ADRs detection on social media

By  *collected data* using:



Source: classic social media and health related websites



Additional language support: using more than just English

Better understand extracted data
Prepare machine learning methods

AGENDA

Context

- Context
- Previous paper
- Goal

Method

- Data sources
- Languages
- Metadata

Results & Discussion

- Results
- Discussion



DATA SOURCES 1/2

- **Twitter®**, a micro-blogging social network
- **Reddit®**, a widely used forum
- **Carenity®**, an online patient community for patient and caregiver



Are all data sources equal



DATA SOURCES 2/2



- 12th *most used* website in the *world*
- Used to send short messages (280 characters)
- Used to *share* news, pictures and *everyday life*



- 9th *most used* website in the *world*
- Used for online discussion
- Contains various *sections* from video games to *drugs usage*



- *Health-related* forum
- Used to share disease experience with other patient
- Contains **1,000+** *pathologies*

LANGUAGES

Languages used for extraction:

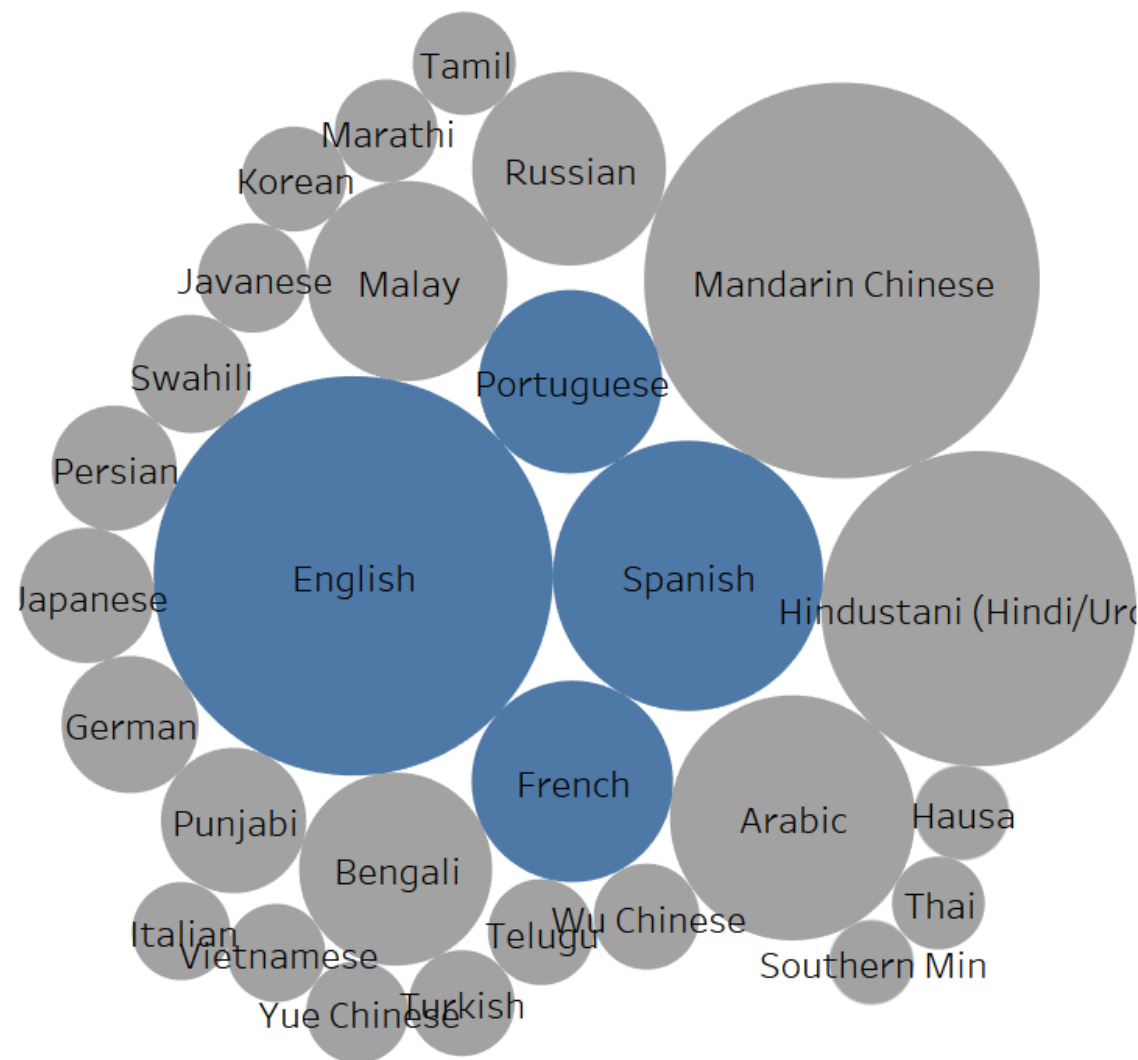
English

French

Spanish

Portuguese

32% of the *world's* population



Language influence on ADRs detection



METADATA

Metadata: data that provides information about other data.

Carency®:



Gender



Age

Reddit®:

- **None available** for anonymization purpose

Twitter®:

- **Not trustworthy** and difficult to exploit

How can we use metadata for ADRs detection



AGENDA

Context

- Context
- Previous paper
- Goal

Method

- Data sources
- Languages
- Metadata

Results & Discussion

- Results
- Discussion



RESULTS

	Twitter	Reddit	Carenity
Period	1 year	2007 - 2018	2011 - 2018
# records	352,285	163,426	12,637

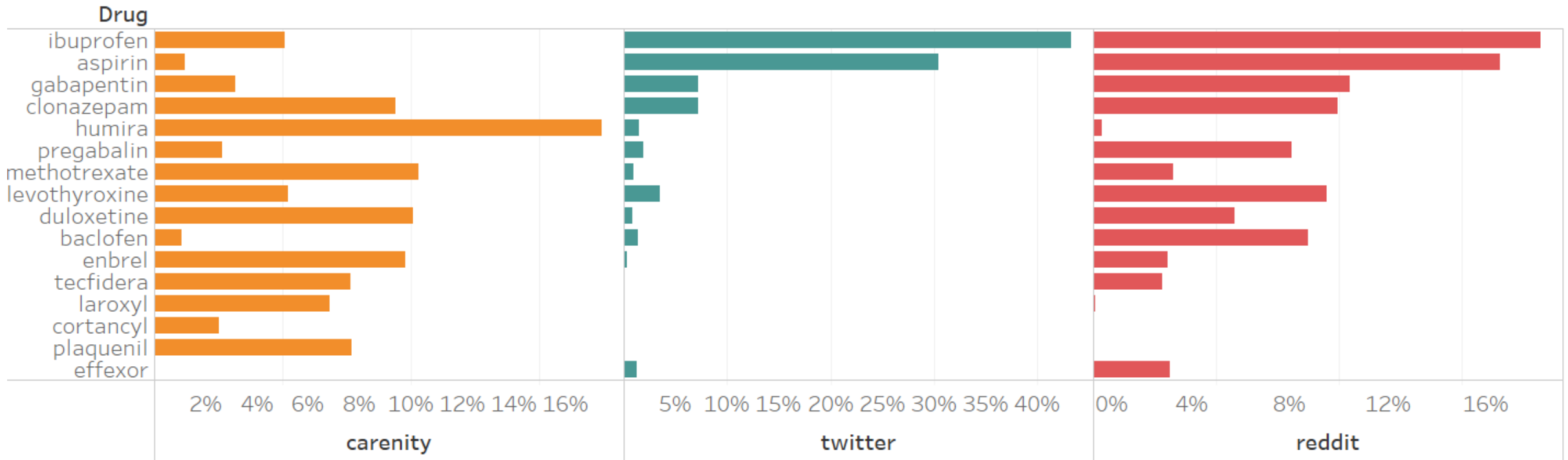
Disclaimer:

The data contained in this paper was obtained for demonstration purposes only using the techniques presented. Anyone analyzing data from social media sites, either public or membership based, should investigate the source terms and conditions relevant to data reuse and obtain any necessary permissions if required.

Detection ~40% of records with potential ADR*

*Potential ADR = ADR and other medical events

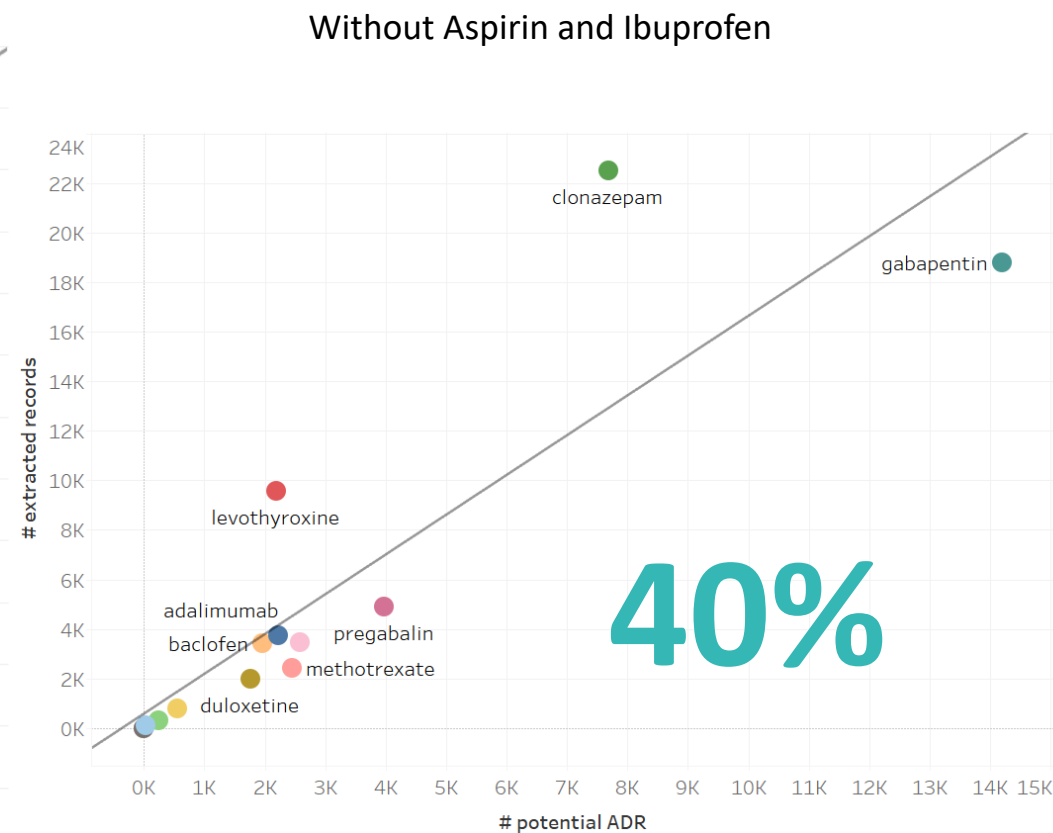
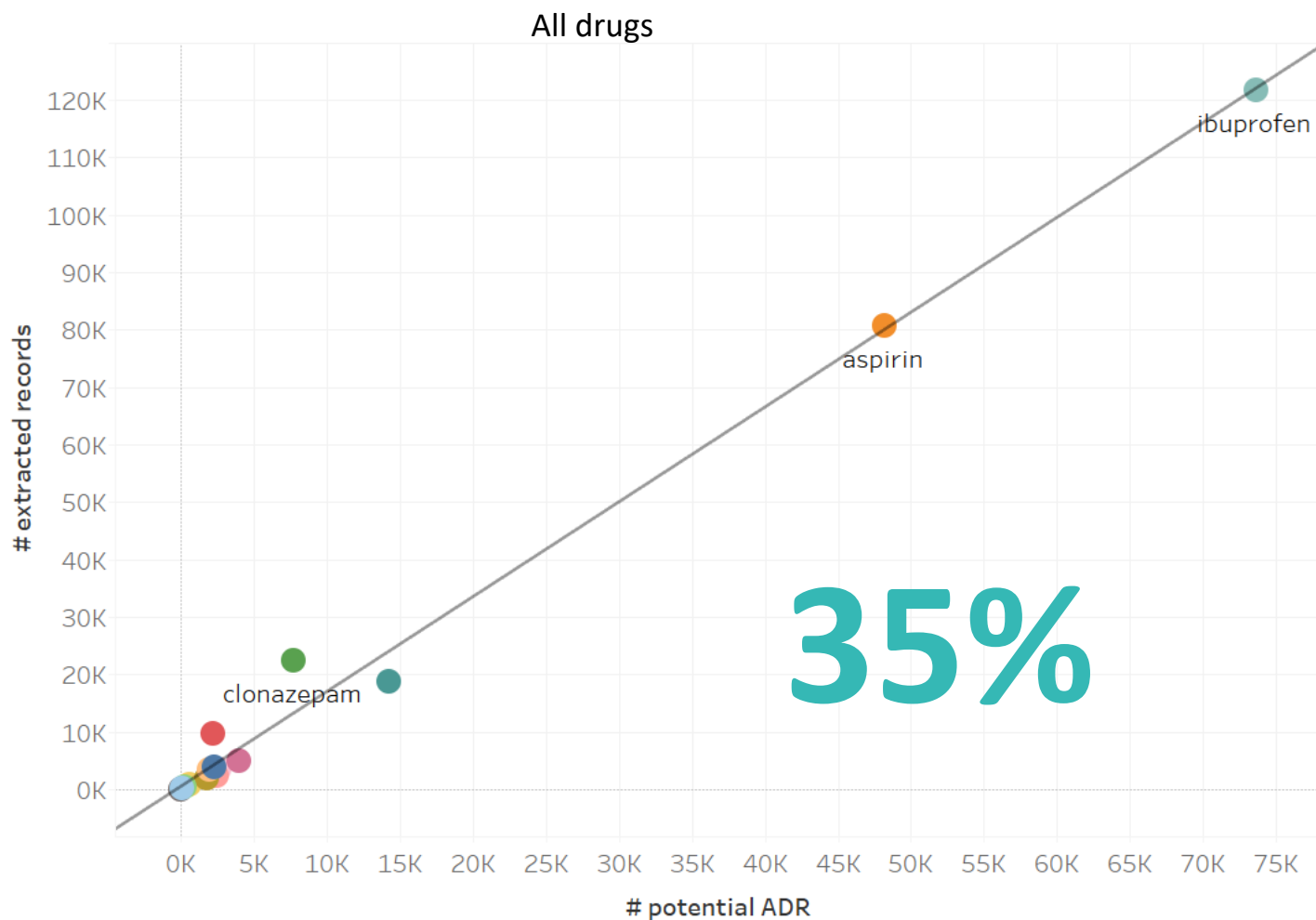
DRUG REPRESENTATION IN TWITTER® VS REDDIT® VS CARENITY®



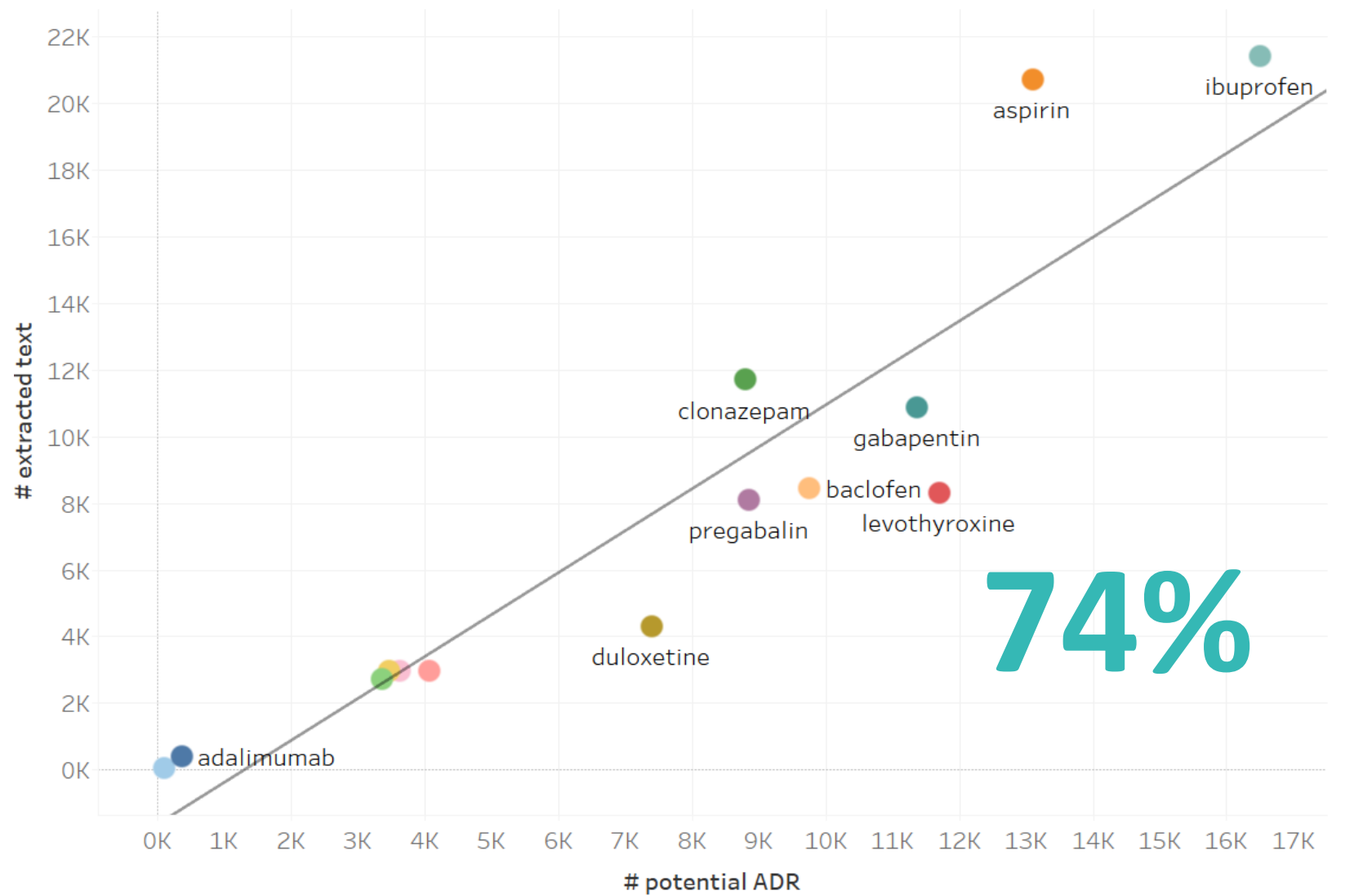
- **Not similar** data *repartition*
- **Aspirin** and **Ibuprofen** over *representation* on Twitter
- **Better** drugs representation on *forums*

All data sources are not equal

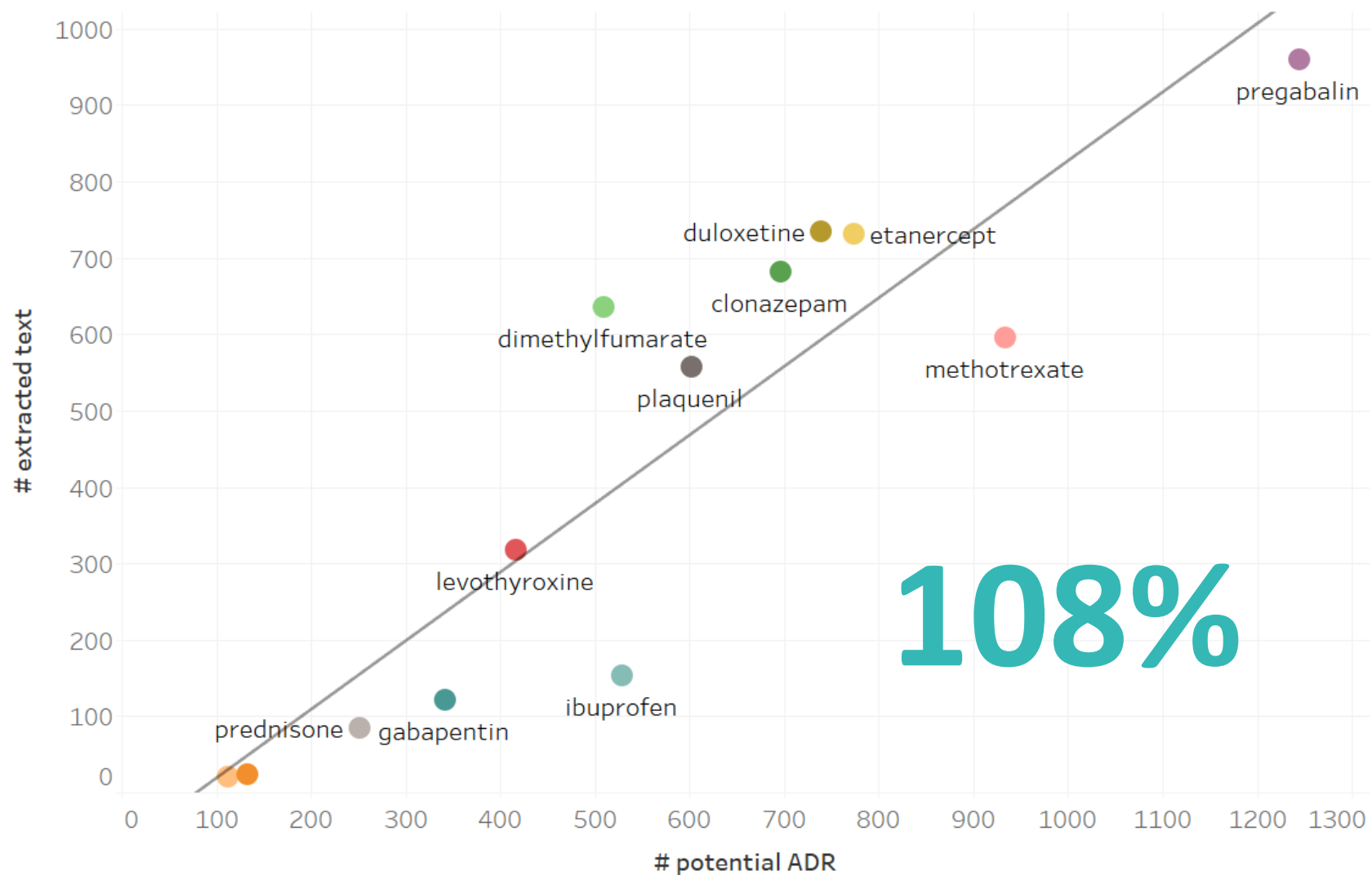
RECORDS AND POTENTIAL ADR RATIO ON TWITTER®



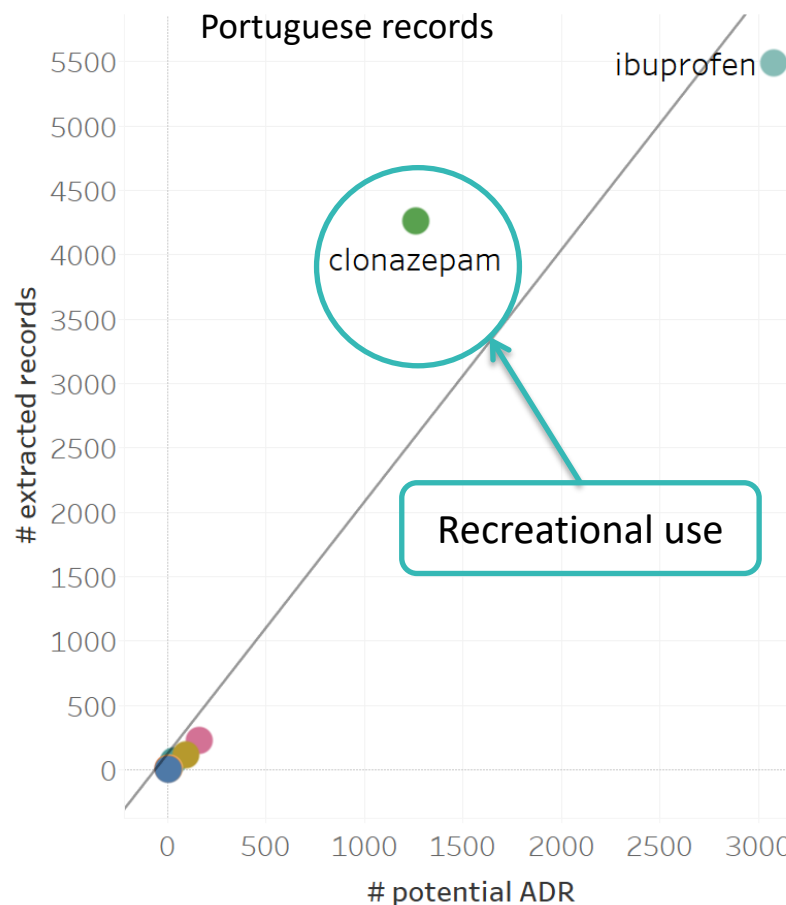
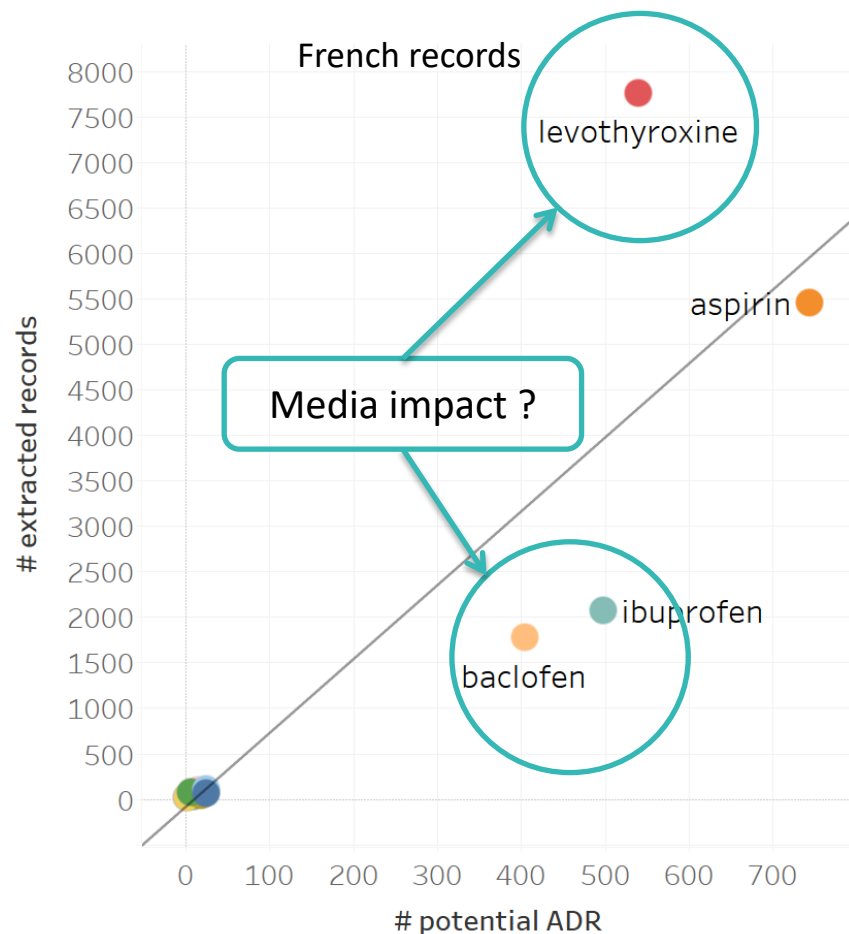
RECORDS AND POTENTIAL ADR RATIO ON REDDIT®



RECORDS AND POTENTIAL ADR RATIO ON CARENITY®



POTENTIAL ADR RATIO ON TWITTER® (FRENCH & PORTUGUESE RECORDS)



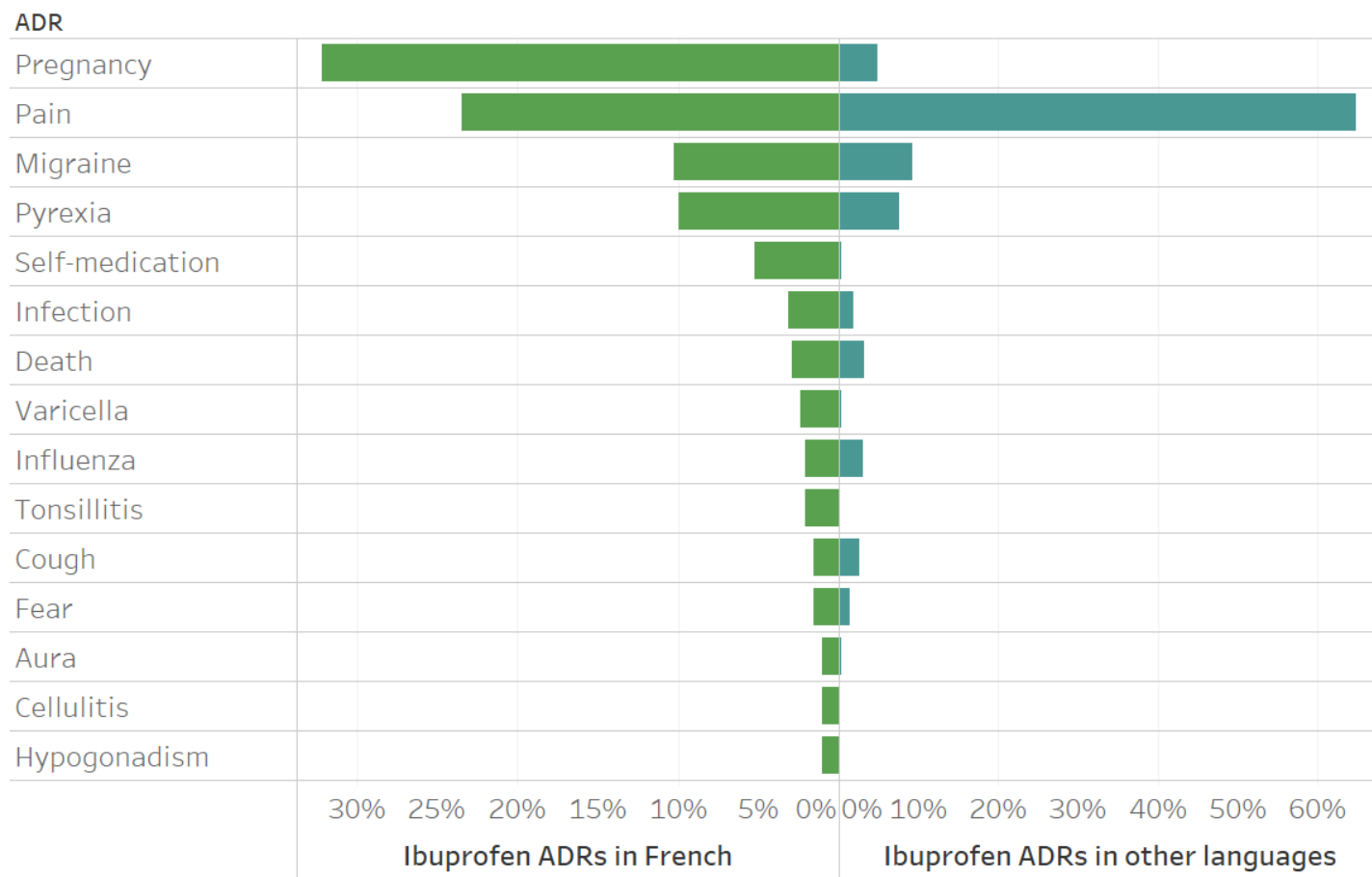
Low ratio (potential ADR/records) *in French*: ~5 to 35%

- Language complexity
- French market oriented

Low ratio (potential ADR/records) *in Portuguese*: ~25 to 40%

Languages difference

ADRS DETECTED FOR IBUPROFEN: FRENCH VS OTHER LANGUAGES IN TWITTER

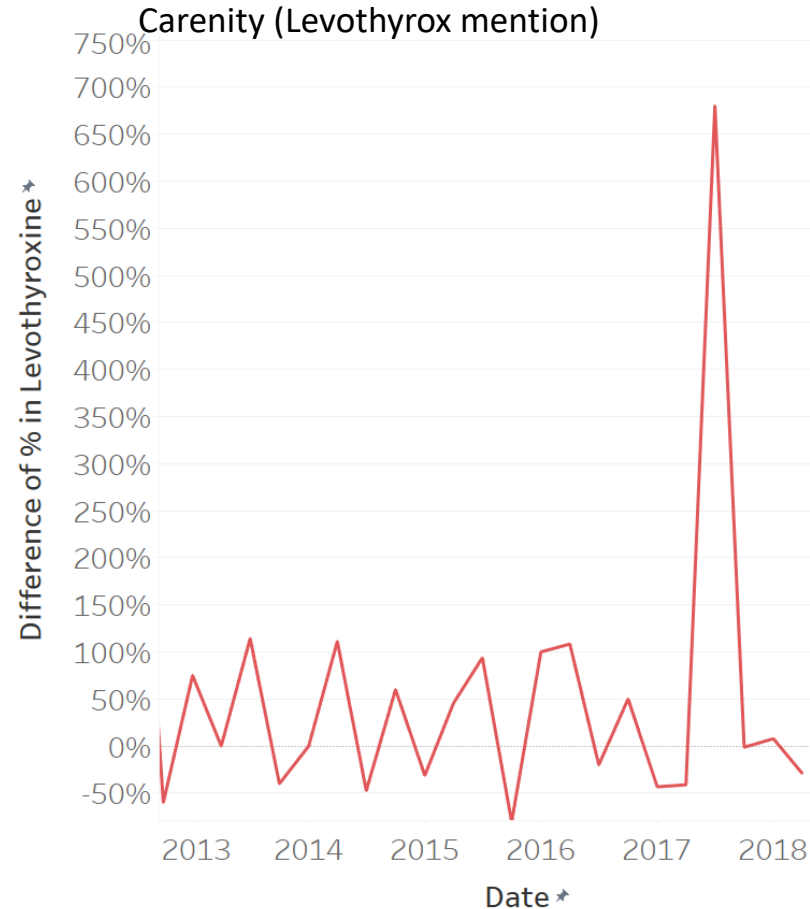
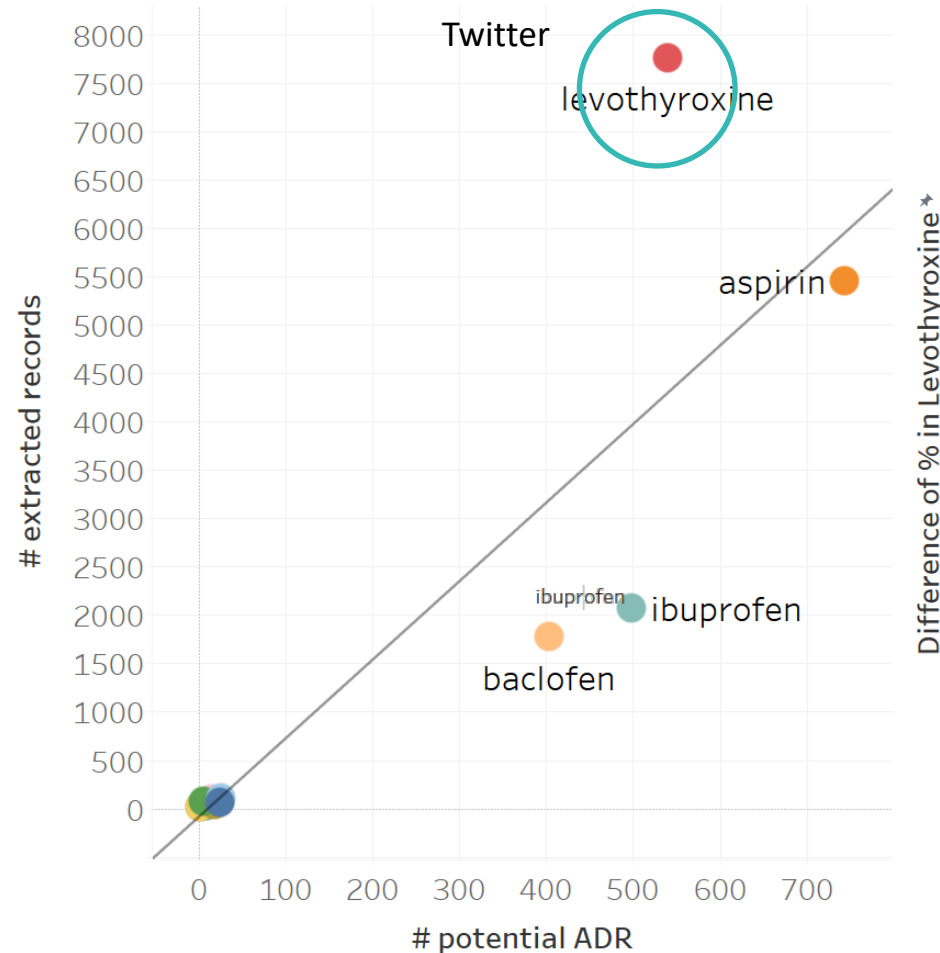


Detected ADR are not similar:

- French first potential ADR: Pregnancy
 - First potential ADR for other languages: Pain
- Cultural and not grammatical characteristic

Country specific

MEDIA IMPACT IN FRENCH



- Twitter: *lots of Levothyroxine* records
- Carenity: *high peak* of levothyroxine mention in *2017*

→ *Media impact*

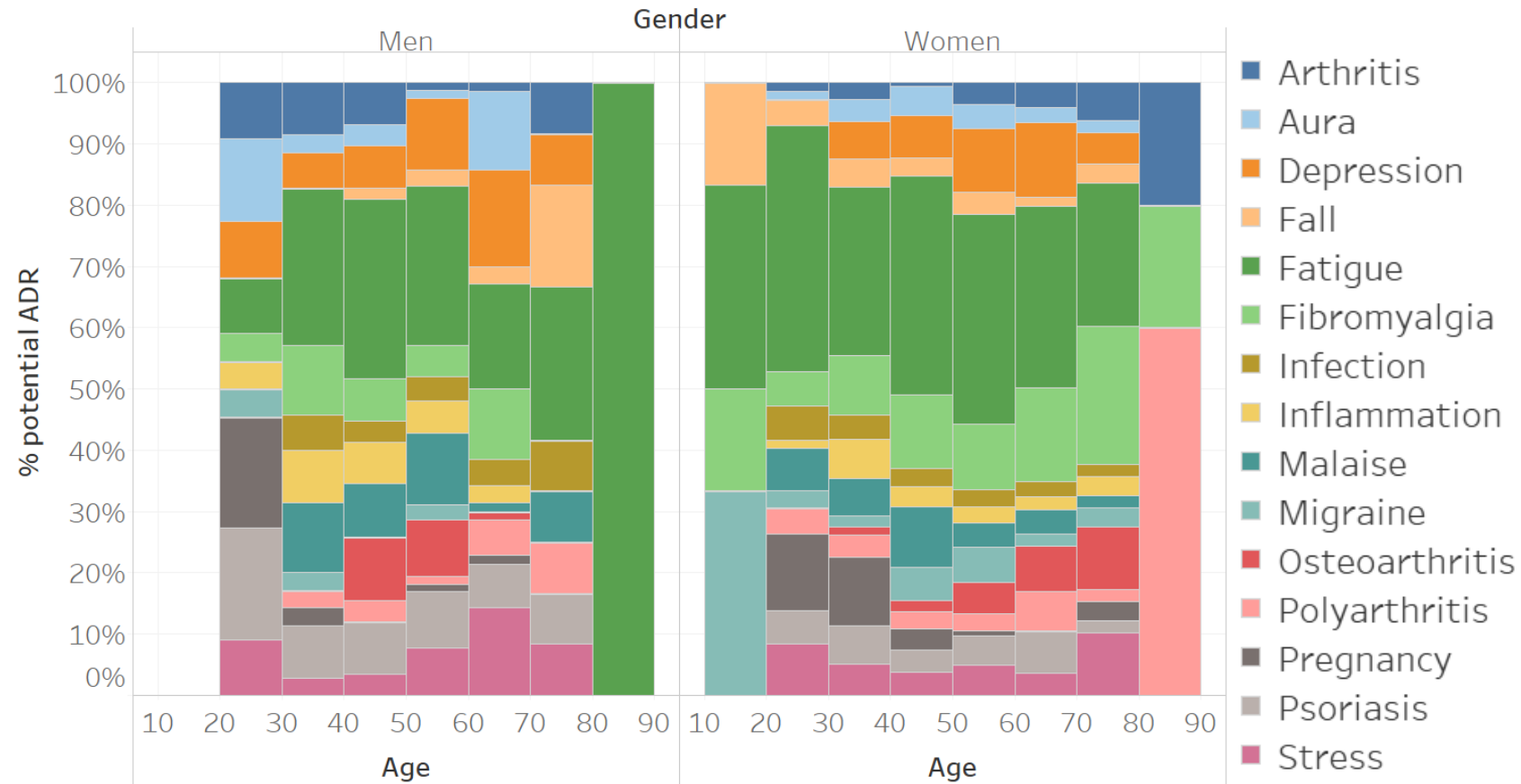
Levothyrox : l'Agence du médicament reconnaît un défaut d'information aux patients

5 000 signalements d'effets inhabituels ont été enregistrés fin août. Parmi les 3 millions de patients traités, certains se reportent sur des solutions prévues pour les enfants.

LE MONDE | 01.09.2017 à 11h35 • Mis à jour le 02.09.2017 à 06h39 |
Par Pascale Santi

All sources subject to media impact

POTENTIAL ADRS DETECTION BY AGE AND GENDER ON CARENITY



Close but *not similar* distribution between *men* and *women*

ADRs and pathologies not similar between gender

AGE & GENDER REPARTITION ON CARENITY



- *Consistent* age distribution
- *Women: 80%* of the records
- Women *~35yo* mention *more ADRs*

Women more willingly to share their experiences



REAL ADR VS CONDITION (SUBSET FOR METHOTREXATE)

	Reddit® En	Twitter® En	Carenity® En	Carenity® Fr
# detected ADR	318	318	318	318
# real ADR % of the total	31 10%	17 5%	54 17%	23 7%
# ADR missed % of missed	6 16%	4 19%	4 7%	32 58%

- *Carenity* and *Reddit* in *English*: most *accurate* sources
- *Twitter*: *not* a lot of *useful* information
- *Carenity* in *french*: *lot* of ADR *missed*
→ Language barrier ?

Forums seems more valuable, especially health-related ones

CONCLUSION

Objectives

Detect bias and limitation by:

- sources number
- number of languages



Results

Sources quality
Languages specification
Gender influence



method
false positives
false negatives

Next
steps

QUESTIONS



Erwan Le Covec



Lise Radoszycki



Stéphane Chollet



Keyrus Biopharma