WHO Outbreak Toolkit Virtual Assistant

WHO, 16 October 2019

Lucas Deroo, **Stéphane Ghozzi**, Anne Perrocheau, Karl Schenkel

WHO/WHE/HIM/MDC

ghozzis@who.int / ghozzis@rki.de

Outline

1. Overview

2. Input: Data dictionary, references and line lists

3. Score

4. Dimensionality reduction

5. Web application

6. Conclusion and outlook

1. Overview

Objective

Outbreak of **unknown origin**… Help formulating **hypotheses** on disease

Inspiration

Election online helper: Which party is closest to you?

Approach

Use expected answers for diseases of interest = **references**

Compare cases with expectations

Workflow

Input:

1. data dictionary
2. references
3. filled line lists

Output:

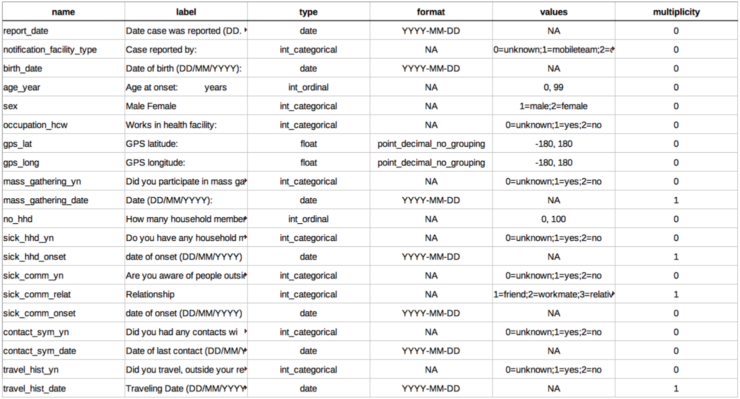
1. **2d visualisation** of cases and references
2. **scores** = similarity between cases and references

2. Input: Data dictionary, references and line lists

Data dictionary

Data dictionary = T0 and T1 questionnaires, i.e. **not** disease specific

Remove free text and other variables that can't be used here: 117 variables left

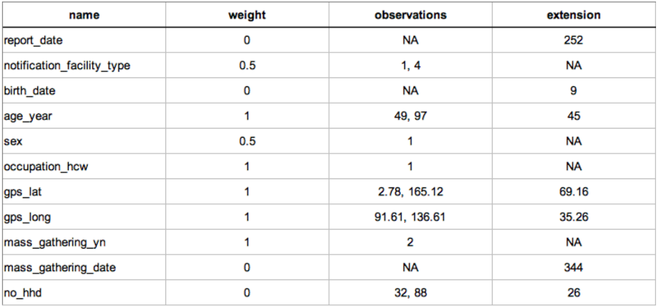


name	label	type	format	values	multiplicity
report_date	Date case was reported (DD.	date	YYYY-MM-DD	NA	0
notification_facility_type	Case reported by:	int_categorical	NA	0=unknown;1=mobileteam;2=	0
birth_date	Date of birth (DD/MM/YYYY):	date	YYYY-MM-DD	NA	0
age_year	Age at onset: years	int_ordinal	NA	0, 99	0
sex	Male Female	int_categorical	NA	1=male;2=female	0
occupation_hcw	Works in health facility:	int_categorical	NA	0=unknown;1=yes;2=no	0
gps_lat	GPS latitude:	float	point_decimal_no_grouping	-180, 180	0
gps_long	GPS longitude:	float	point_decimal_no_grouping	-180, 180	0
mass_gathering_yn	Did you participate in mass ga	int_categorical	NA	0=unknown;1=yes;2=no	0
mass_gathering_date	Date (DD/MM/YYYY):	date	YYYY-MM-DD	NA	1
no_hhd	How many household membe	int_ordinal	NA	0, 100	0
sick_hhd_yn	Do you have any household m	int_categorical	NA	0=unknown;1=yes;2=no	0
sick_hhd_onset	date of onset (DD/MM/YYYY)	date	YYYY-MM-DD	NA	1
sick_comm_yn	Are you aware of people outsi	int_categorical	NA	0=unknown;1=yes;2=no	0
sick_comm_relat	Relationship	int_categorical	NA	1=friend;2=workmate;3=relativ	1
sick_comm_onset	date of onset (DD/MM/YYYY)	date	YYYY-MM-DD	NA	0
contact_sym_yn	Did you had any contacts wi	int_categorical	NA	0=unknown;1=yes;2=no	0
contact_sym_date	Date of last contact (DD/MM/Y	date	YYYY-MM-DD	NA	0
travel_hist_yn	Did you travel, outside your re	int_categorical	NA	0=unknown;1=yes;2=no	0
travel_hist_date	Traveling Date (DD/MM/YYYY	date	YYYY-MM-DD	NA	1

References

For each disease: Reference = Variable *name*, *weight*, *observations*, *extension*

weight = how indicative of the disease
observations = range of expected answers
extension = expected 50% confidence interval of answers [optional]



name	weight	observations	extension
report_date	0	NA	252
notification_facility_type	0.5	1, 4	NA
birth_date	0	NA	9
age_year	1	49, 97	45
sex	0.5	1	NA
occupation_hcw	1	1	NA
gps_lat	1	2.78, 165.12	69.16
gps_long	1	91.61, 136.61	35.26
mass_gathering_yn	1	2	NA
mass_gathering_date	0	NA	344
no_hhd	0	32, 88	26

5 fake random references
Blue Fever, Danger Fever, Kleptospirosis, Viral Hepatitis Z, West Rhine Virus

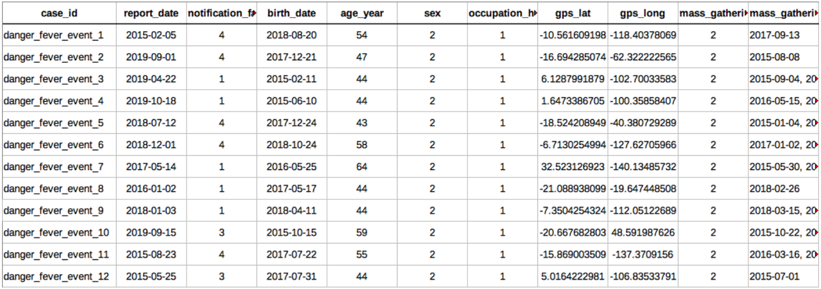
5 real references: rough estimates
Hepatitis A, Hepatitis E, Leptospirosis, Mushroom, Yellow Fever

Line lists (events)

Generate synthetic events:

1. completely **random** cases
2. all cases drawn randomly from **1 disease** reference
3. all cases from **1 disease** but with **noise**
4. cases from **2 diseases** (half/half)
5. cases from **2 diseases** (half/half), with **noise**

with noise = 50% of variables of any one case are completely random



case_id	report_date	notification_fa	birth_date	age_year	sex	occupation_h	gps_lat	gps_long	mass_gatheri	mass_gatheri
danger_fever_event_1	2015-02-05	4	2018-08-20	54	2	1	-10.561609198	-118.40378069	2	2017-09-13
danger_fever_event_2	2019-09-01	4	2017-12-21	47	2	1	-16.694285074	-62.322222565	2	2015-08-08
danger_fever_event_3	2019-04-22	1	2015-02-11	44	2	1	6.1287991879	-102.70033583	2	2015-09-04, 20
danger_fever_event_4	2019-10-18	1	2015-06-10	44	2	1	1.6473386705	-100.35858407	2	2016-05-15, 20
danger_fever_event_5	2018-07-12	4	2017-12-24	43	2	1	-18.524208949	-40.380729289	2	2015-01-04, 20
danger_fever_event_6	2018-12-01	4	2018-10-24	58	2	1	-6.7130254994	-127.62705966	2	2017-01-02, 20
danger_fever_event_7	2017-05-14	1	2016-05-25	64	2	1	32.523126923	-140.13485732	2	2015-05-30, 20
danger_fever_event_8	2016-01-02	1	2017-05-17	44	2	1	-21.088938099	-19.647448508	2	2018-02-26
danger_fever_event_9	2018-01-03	1	2018-04-11	44	2	1	-7.3504254324	-112.05122689	2	2018-03-15, 20
danger_fever_event_10	2019-09-15	3	2015-10-15	59	2	1	-20.667682803	48.591987626	2	2015-10-22, 20
danger_fever_event_11	2015-08-23	4	2017-07-22	55	2	1	-15.869003509	-137.3709156	2	2016-03-16, 20
danger_fever_event_12	2015-05-25	3	2017-07-31	44	2	1	5.0164222981	-106.83533791	2	2015-07-01

)

3. Score

1 score for each case and each reference = how similar is this case to the reference?

For each variable:

- answer **within** range or set of reference *observations:* **+1**
- **missing** answer or “unknown”: **+0.5**
- answer **outside** of reference *observations:* **+0**

Score = weighted average over variables using the reference *weights*

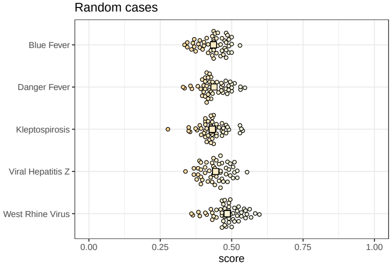
(*extension* not used at the moment for scoring)

Case ~ reference → score = 1
Case outside reference → score = 0
Random case → score ≈ 0.5

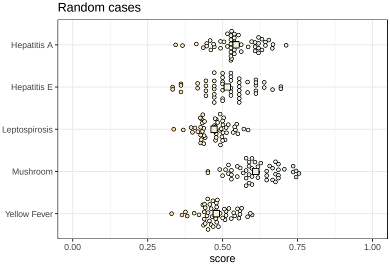
Tested 5 types of events for 2 sets of references = 10 scenarios

Scores: Random cases

fake references



real references

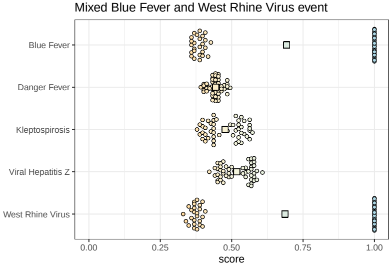


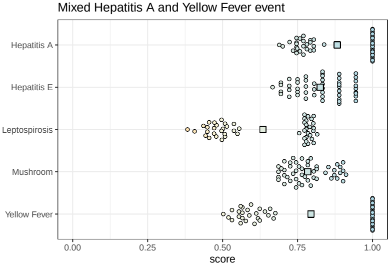
1 dot = 1 case / square = average

Scores: Cases from 2 diseases

fake references

real references

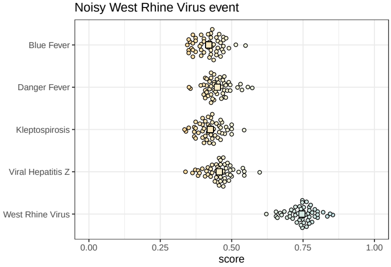




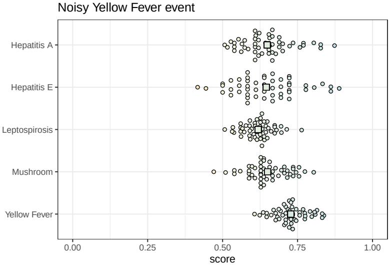
1 dot = 1 case / square = average

Scores: Cases from 1 disease with noise

fake references



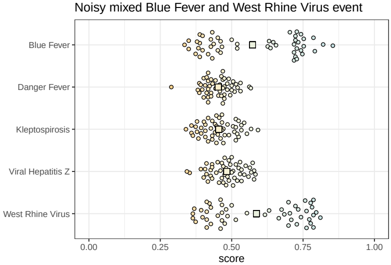
real references



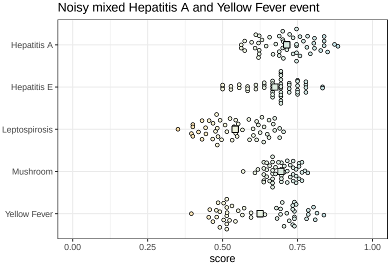
1 dot = 1 case / square = average

Scores: Cases from 2 diseases with noise

fake references



real references



1 dot = 1 case / square = average

4. Dimensionality reduction

Reduce dimensionality form 117 to 2 using the t-SNE algorithm

Needs a **distance** between cases:
similar approach and difficulties as for score

With d_{ij} the distance between case i and case j, and w_v the weight for variable v averaged over all references:

$$d_{ij} = \sqrt{\sum_v w_v \, d_{ijv}^2 / \sum_v w_v}$$

with

$d_{ijv} = \min_{m,n} |v_{im} - v_{jn}| / \max_{i,j}(\min_{m,n} |v_{im} - v_{jn}|)$ if variable is *int_ordinal*, *float* or *date*
$d_{ijv} = 1 - \delta(|\{v_{im}\}_m \cap \{v_{jn}\}_n|)$ if variable is *int_categorical*
$d_{ijv} = 0.5$ if any answer is missing

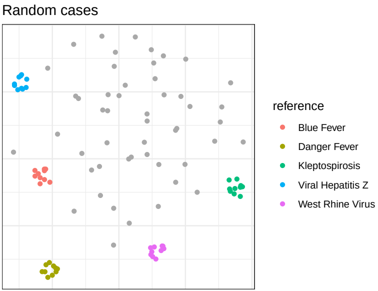
where v_{im} is the m-th observation of variable v for case i, and $\delta(x) = 0$ if $x = 0$, 1 else.

Smallest distance is 0 and largest distance is 1

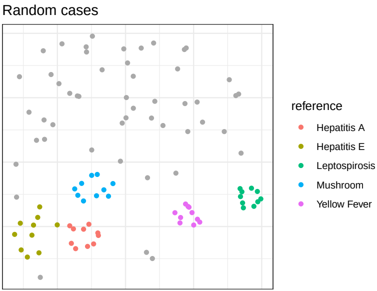
Include **reference cases** for comparison

2d projection: Random cases

fake references



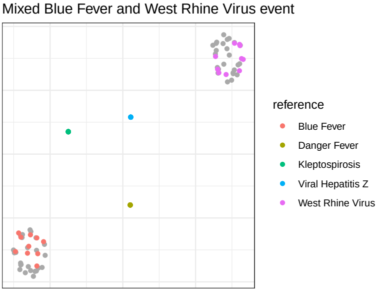
real references



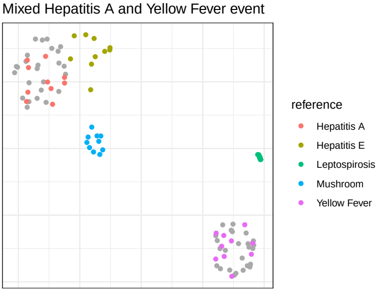
Caution: Cluster size doesn't mean anything

2d projection: Cases from 2 diseases

fake references



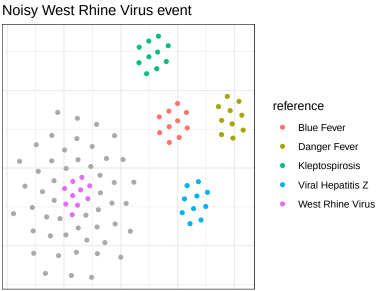
real references



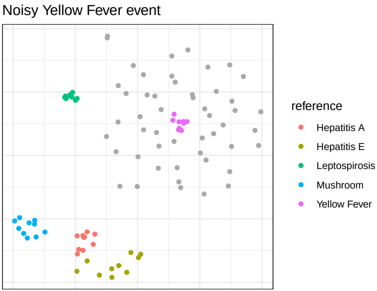
Caution: Cluster size doesn't mean anything

2d projection: Cases from 1 disease with noise

fake references



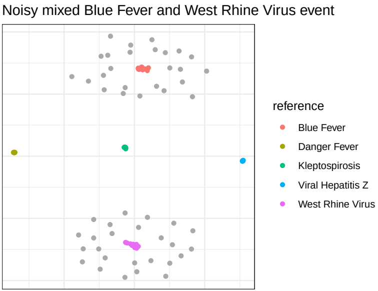
real references



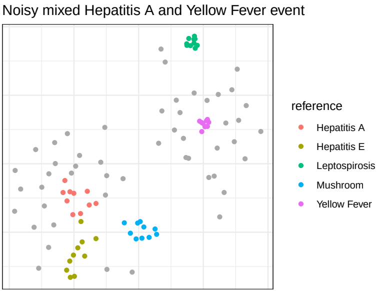
Caution: Cluster size doesn't mean anything

2d projection: Cases from 2 diseases with noise

fake references



real references



Caution: Cluster size doesn't mean anything

5. Web application

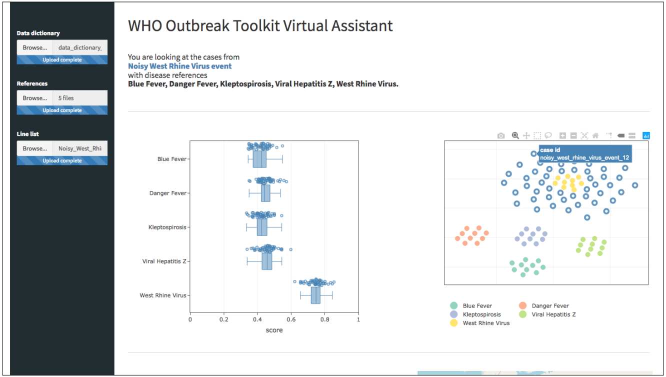
R Shiny prototype

Input: upload data dictionary, set of references and 1 line list
Output: scores, 2d projection, epicurve, map, line list, pivot table

Installation:

local at the moment

WHO cloud (internal) or online (public) in the future



Data dictionary

Browse... data_dictionary.

Upload complete

References

Browse... 5 files

Upload complete

Line list

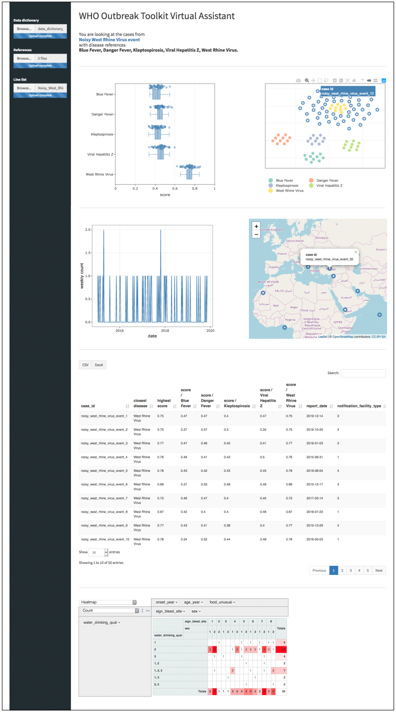
Browse... Noisy_West_Rhi

Upload complete

WHO Outbreak Toolkit Virtual Assistant

You are looking at the cases from
Noisy West Rhine Virus event
with disease references
Blue Fever, Danger Fever, Kleptospirosis, Viral Hepatitis Z, West Rhine Virus.

case id
noisy_west_rhine_virus_event_12



WHO Outbreak Toolkit Virtual Assistant

You are looking at the cases from
Noisy West Rhine Virus event
with disease references
Blue Fever, Danger Fever, Kleptospirosis, Viral Hepatitis Z, West Rhine Virus.

CSV Excel

Search:

case_id	closest disease	highest score	score / Blue Fever	score / Danger Fever	score / Kleptospirosis	score / Viral Hepatitis Z	score / West Rhine Virus	report_date	notification_facility_type
noisy_west_rhine_virus_event_1	West Rhine Virus	0.75	0.47	0.47	0.4	0.47	0.75	2019-12-14	3
noisy_west_rhine_virus_event_2	West Rhine Virus	0.75	0.37	0.57	0.5	0.34	0.75	2019-10-20	3
noisy_west_rhine_virus_event_3	West Rhine Virus	0.77	0.47	0.46	0.42	0.41	0.77	2016-01-03	3
noisy_west_rhine_virus_event_4	West Rhine Virus	0.76	0.49	0.41	0.43	0.5	0.76	2015-08-31	1
noisy_west_rhine_virus_event_5	West Rhine Virus	0.78	0.43	0.42	0.43	0.45	0.78	2019-08-04	4
noisy_west_rhine_virus_event_6	West Rhine Virus	0.66	0.37	0.53	0.48	0.49	0.66	2015-12-17	3
noisy_west_rhine_virus_event_7	West Rhine Virus	0.72	0.46	0.47	0.4	0.45	0.72	2017-03-14	3
noisy_west_rhine_virus_event_8	West Rhine Virus	0.67	0.42	0.4	0.4	0.46	0.67	2018-07-23	1
noisy_west_rhine_virus_event_9	West Rhine Virus	0.77	0.43	0.41	0.38	0.4	0.77	2015-12-29	4
noisy_west_rhine_virus_event_10	West Rhine Virus	0.78	0.34	0.52	0.44	0.48	0.78	2016-05-03	1

Show 10 entries

Showing 1 to 10 of 50 entries

Previous 1 2 3 4 5 Next

	sign_bleed_site	1		2	3		4		5		6		7		8		Totals
	sex	1	2	2	1	2	1	2	1	2	1	2	1	2	1	2	
water_drinking_qual																	
1				1					1		1		1		1	1	6
2		3	5				1	3	1	2	2	2	1	4	2	1	27
3			1		1				1					1			4
1, 2										1				1			2
1, 2, 3			1				2				1			1		2	7
1, 3						1						1					2
2, 3										1						1	2
Totals		3	7	1	1	1	3	3	3	4	4	3	2	7	3	5	50

6. Conclusion and outlook

Conclusion

Already simple approaches are promising and deliver useful insights

2d visualisation: discriminates better
scores: single number per case / for the whole event

Usefulness for potential users can be probed: Are the *functionalities* interesting?

What's next? 2 possible directions

1. Improve and evaluate method

- computation time
- distinguish between suspected, confirmed and excluded cases
- find real line lists and convert to data-dictionary format
- evaluate the tool
- establish the reference for each epidemic prone disease
- machine learning (classification)

Resources needed: at least 1 data scientist & 1 medical expert

2. Prospective piloting

- build as software *or* integrate in other tools (Go.Data?)
- feasibility study needed

Resources needed: at least 1 software engineer

Ideas for collaborations:

- make an R package and share it with community, e.g. RECON
- Cambridge working group on same topic
- WHO Medical expert group, Antwerp (medical experts), READY project (USAIDS), RKI

Thank you!

Advertisement:

Anomaly detection for EIOS
Machine learning and natural language processing to detect signals in stream of articles

Signale team at the Robert Koch Institute
data science for infectious-disease epidemiology
rki.de/signale-project