



Ablation of prompts to estimate sensitivity of algorithms in a datasource

Giuseppe Roberto

Senior Pharmacoepidemiology consultant, ARS Toscana

**Validate study variables to reduce
misclassification bias: recent tools and research needs**

HYBRID WORKSHOP

27 March 2025 - 14.30-18.30

ARS Toscana - Villa La Quiete, via Pietro Dazzi 1 - Florence, Italy

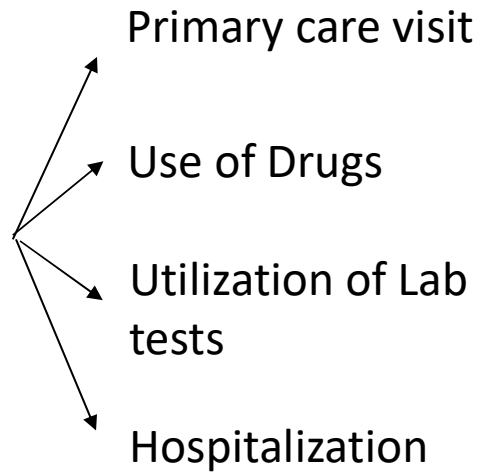
What is a component algorithm?

Let's assume we want to estimate the prevalence of type 2 diabetes in a real word data source....

T2DM

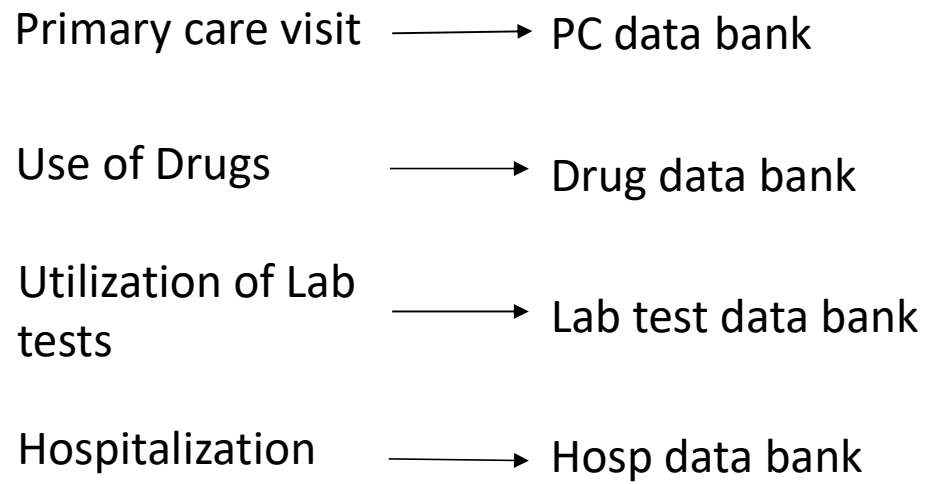
MT2DM

T2DM



MT2DM

T2DM



MT2DM

T2DM

Primary care visit

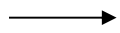
→ PC data bank



PC Diagnosis records

Use of Drugs

→ Drug data bank



Drug utilization records

Utilization of Lab tests

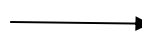
→ Lab test data bank



Lab test records

Hospitalization

→ Hosp data bank

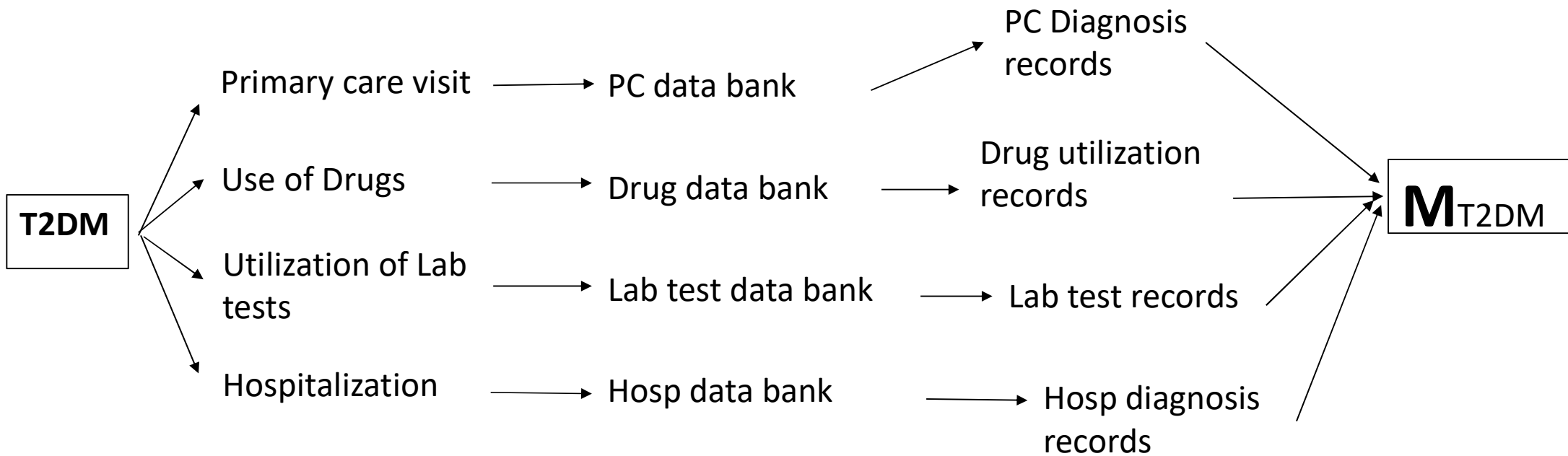


Hosp diagnosis records

M_{T2DM}

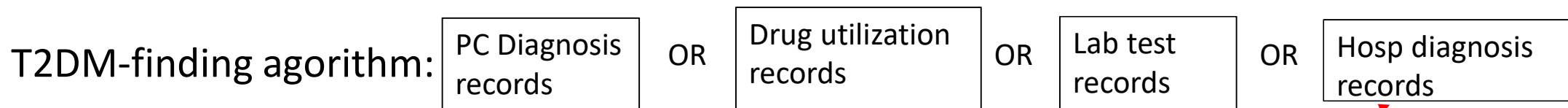


ARS TOSCANA
agenzia regionale di sanità



T2DM-finding algorithm: PC Diagnosis records OR Drug utilization records OR Lab test records OR Hosp diagnosis records

What is a component algorithm?



- Identification algorithms can be broken down in *component algorithms* based on records generated by specific data prompts

How component algorithms can inform on case-finding algorithm sensitivity?

Let's assume we want to estimate the prevalence of type 2 diabetes in 2 diverse* real world data sources....

MT2DM (DATA SOURCE A): observed prevalence 20%

MT2DM (DATA SOURCE B): observed prevalence 15%

* Describing diversity of real world data sources in pharmacoepidemiologic studies: The DIVERSE scoping review. PDS.2024;33:e5787. <https://doi.org/10.1002/pds.5787>

Let's assume we want to estimate the prevalence of type 2 diabetes in 2 diverse* real world data sources....

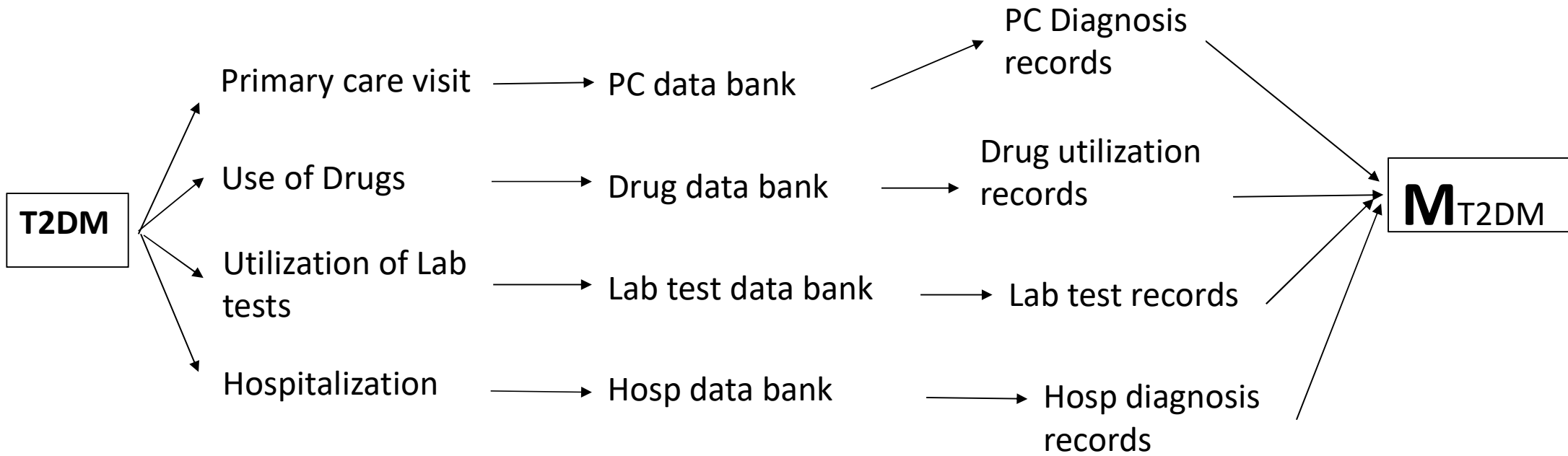
MT2DM (DATA SOURCE A): observed prevalence 20%

MT2DM (DATA SOURCE B): observed prevalence 15%

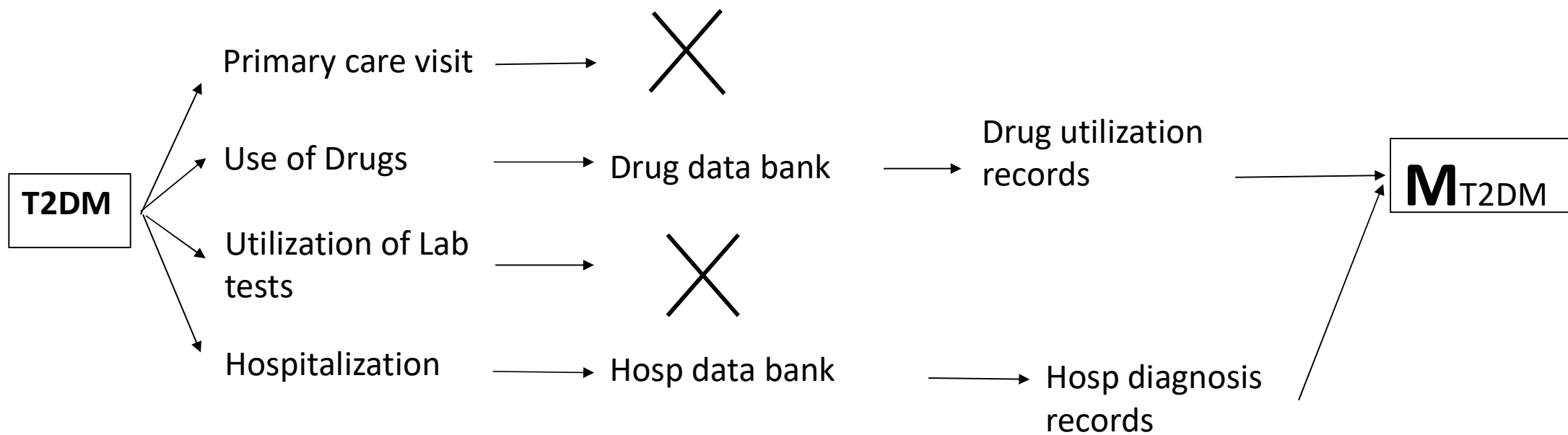
Does true prevalence actually differs in the relevant underlying populations?

* Describing diversity of real world data sources in pharmacoepidemiologic studies: The DIVERSE scoping review. PDS.2024;33:e5787. <https://doi.org/10.1002/pds.5787>

1) DATA SOURCE A



2) DATA SOURCE B



MT2DM(DATA SOURCE A): PC Diagnosis records OR Drug utilization records OR Lab test records⁰ OR Hosp diagnosis records = 20%

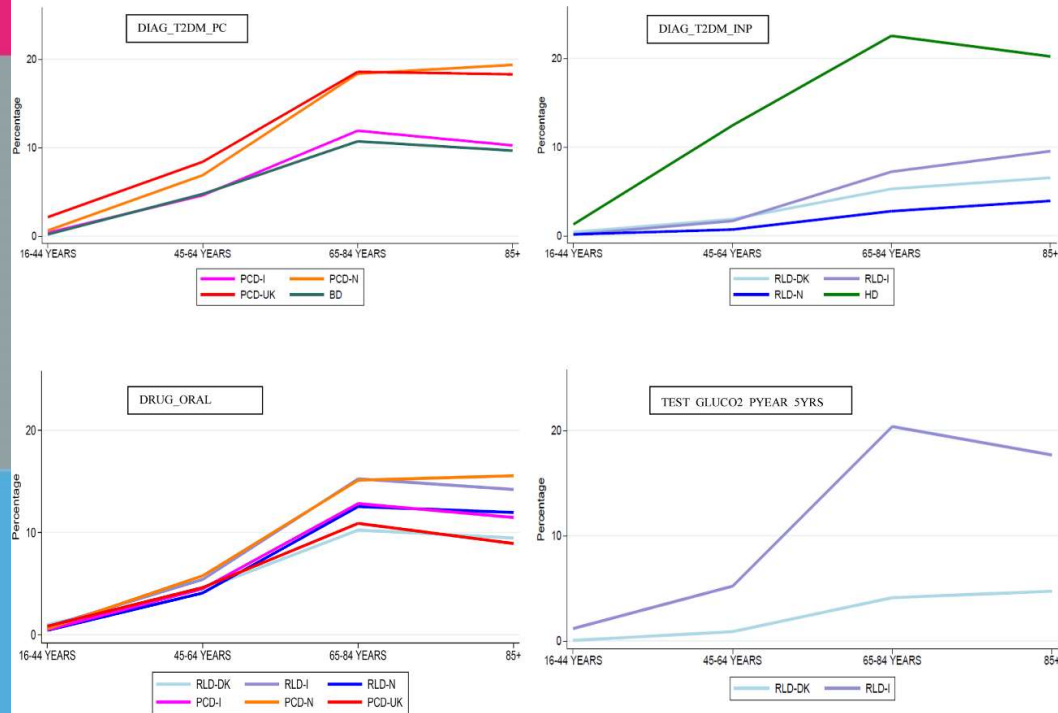
MT2DM(DATA SOURCE B): Drug utilization records OR Hosp diagnosis records =15%

- *Breaking down the identification strategy in standard component algorithms can support the generation of hypotheses on algorithm sensitivity*
- *For example, a lack of sensitivity in DATA SOURCE B can be reasonably hypothesized due to missing Primary Care and Lab Test prompts!!! (e.g. 25% of cases might be missed???)*

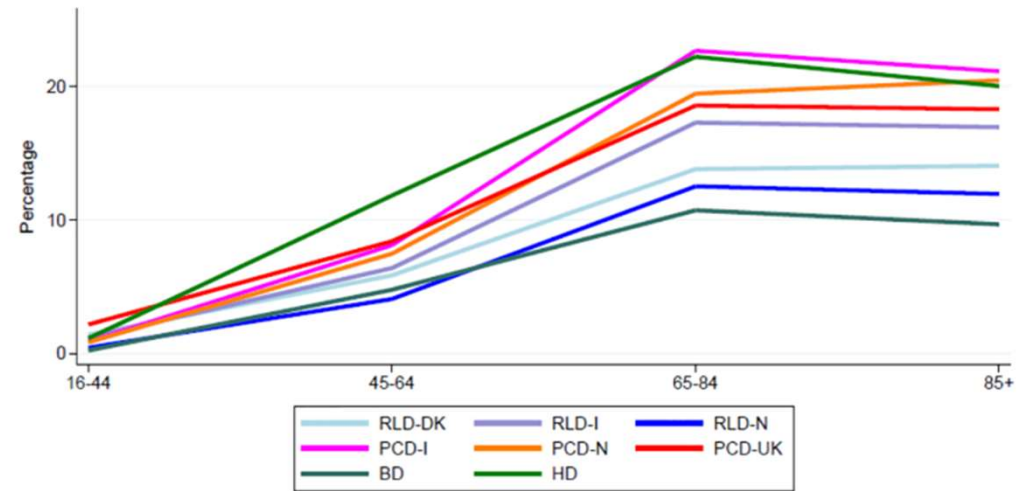
Data DIVERSITY hampers direct comparisons between data sources but the component algorithm strategy remains extremely informative in absence of evidence on algorithm sensitivity!

Prevalence of T2DM

by component algorithm and datasource



Overall prevalence of T2DM by datasource



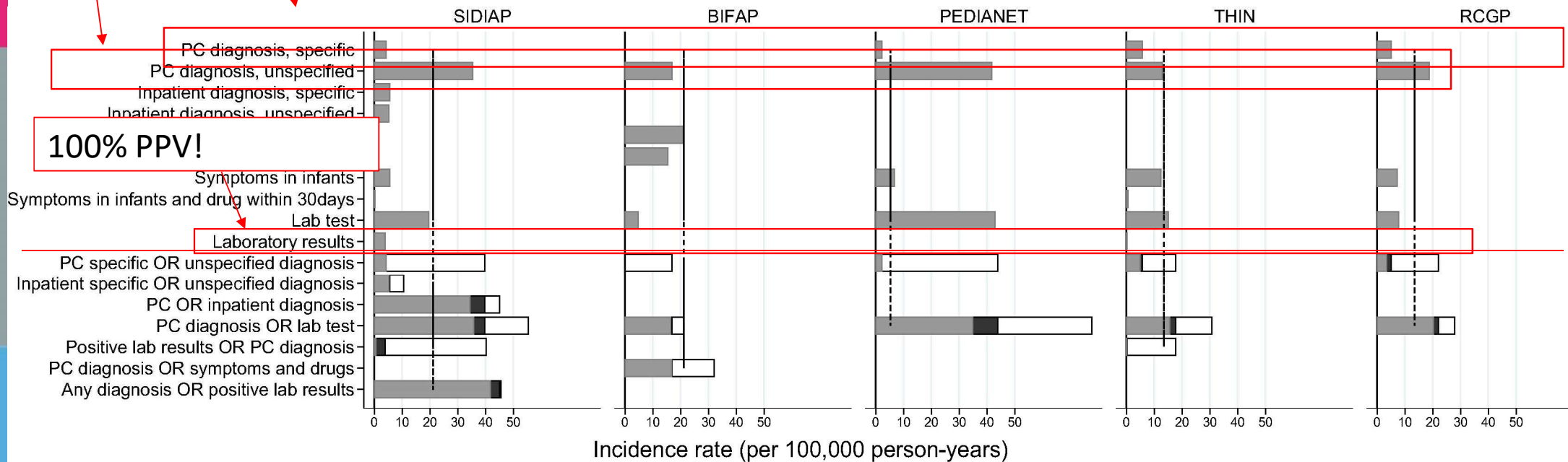
Identifying Cases of Type 2 Diabetes in Heterogeneous Data Sources: Strategy from the EMIF Project. PLoS ONE 11(8): e0160648. <https://doi.org/10.1371/journal.pone.0160648>

Quantifying outcome misclassification in multi-database studies: The case study of pertussis in the ADVANCE project

Very low sensitivity!

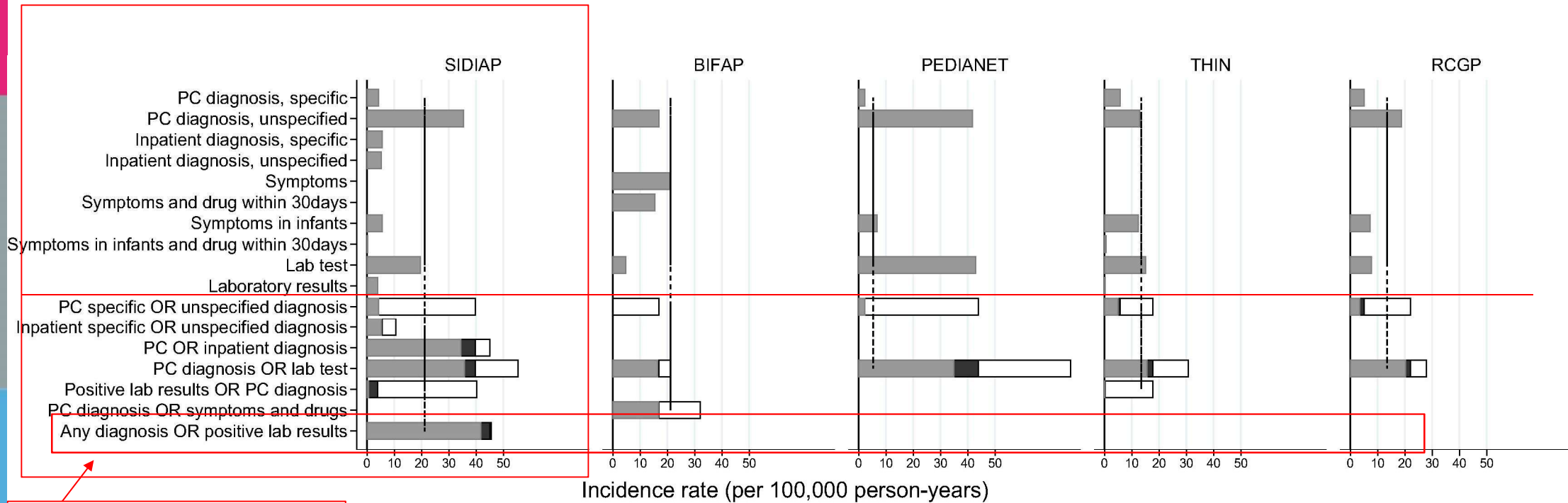
High false positives!

100% PPV!



Legend:
■ Left-hand side component only
■ Both components
□ Right-hand side component only
----- TESSy

Quantifying outcome misclassification in multi-database studies: The case study of pertussis in the ADVANCE project



100% sensitivity?

-hand side component only ■ Both components □ Right-hand side component only
 ----- TESSy

Quantifying outcome misclassification in multi-database studies: The case study of pertussis in the ADVANCE project

- In SIDIAP, the *component algorithm strategy* was used to substantiate assumptions on unknown validity indices of component and composite algorithms and apply algebraic formulae to obtain approximate estimates of algorithm validity, including *sensitivity*, under two distinct scenarios.
- Benchmark data from an external reference (π), the Tessy surveillance system were used

Known parameters	Formula to derive another parameter
One algorithm	
PPV and SE	$\pi = \frac{P \times PPV}{SE}$
PPV and Π	$SE = \frac{P \times PPV}{\pi}$
SE and Π	$PPV = \frac{SE \times \pi}{P}$
Two algorithms A and B	
SE of A, of B, and of A AND B	$SE_{AORB} = SE_A + SE_B - SE_{AANDB}$
Π and PPV of A, of B, and of A AND B	$SE_{AORB} = \frac{P_A \times PPV_A}{\pi} + \frac{P_B \times PPV_B}{\pi} - \frac{P_{AANDB} \times PPV_{AANDB}}{\pi}$
SE of A OR B, and PPV of A, of B, and of A AND B	$\pi = \frac{P_A \times PPV_A + P_B \times PPV_B - P_{AANDB} \times PPV_{AANDB}}{SE_{AORB}}$
PPV of A, of B, and of A AND B	$PPV_{AANDB} = \frac{P_A \times PPV_A + P_B \times PPV_B - P_{AANDB} \times PPV_{AANDB}}{PPV_{AORB}}$

Quantifying outcome misclassification in multi-database studies: The case study of pertussis in the ADVANCE project

In order to obtain an approximate estimate of algorithm validity, we explored two scenarios in SIDIAP, corresponding to different assumptions for PPV of 'PC diagnosis, unspecified' and of 'inpatient diagnosis, unspecified': in the first scenario, PPV was 70%, in the second scenario, PPV was 50%. As a consequence, in the first scenario 'PC specific OR unspecified diagnosis' had a PPV of 72% (or, in the second scenario: 54%) and a sensitivity of 85% (or, in the second scenario: 83%). Based on this estimate, the adjusted IR of BorPer in the SIDIAP study population was 35.5 per 100,000 PY (or, in the second scenario: 25.9) vs the TESSy surveillance system IR 21.2.

Ablation of data prompts to estimate lack of sensitivity

Ablation of data prompts to estimate lack of sensitivity

- Identification of type 2 diabetes:


MT2DM(DATA SOURCE A): PC Diagnosis records OR Drug utilization records OR Lab test records OR Hosp diagnosis records = 20%

Ablation of data prompts to estimate lack of sensitivity

- Identification of type 2 diabetes:

MT2DM(DATA SOURCE A): PC Diagnosis records OR Drug utilization records OR Lab test records OR Hosp diagnosis records = 20%

Prompt ablation

MT2DM(DATA SOURCE A):  OR Drug utilization records OR Lab test records OR Hosp diagnosis records = 15%

Ablation of data prompts to estimate lack of sensitivity

- Identification of type 2 diabetes:

MT2DM(DATA SOURCE A): PC Diagnosis records OR Drug utilization records OR Lab test records OR Hosp diagnosis records = 20%

Prompt ablation

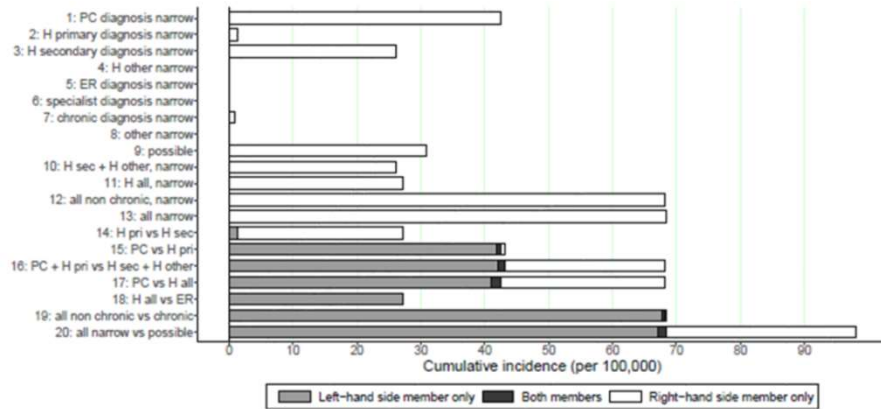
MT2DM(DATA SOURCE A): [Redacted] OR Drug utilization records OR Lab test records OR Hosp diagnosis records = 15%

A data source that lacks data prompted by PC visits could possibly miss a significant share of cases!!!

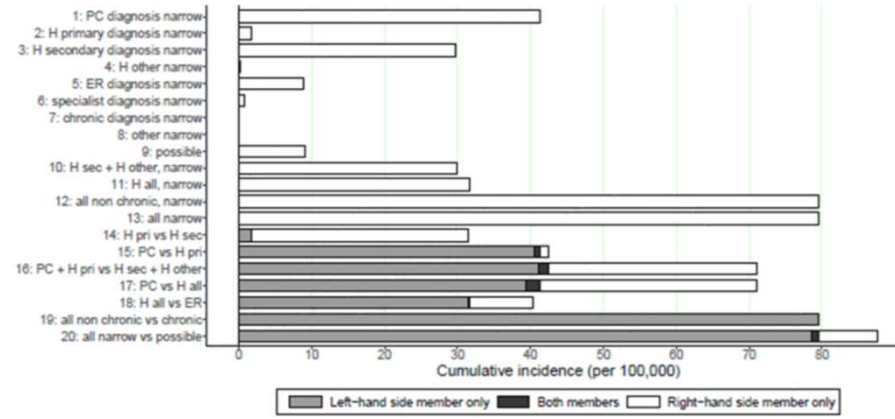
An example from the COVID-19 Vaccine Monitoring project

- The component algorithm strategy was applied to assess the impact of diversity of data sources on background incidence rates of 39 adverse events of special interest (AESI) in a set of DIVERSE data sources
- each data source was queried 20 times, first using separately each component algorithm, and then reassembling them back in composite algorithms (see next slide)
- Cumulative incidence of AESI occurrence was calculated for each component and composite algorithm

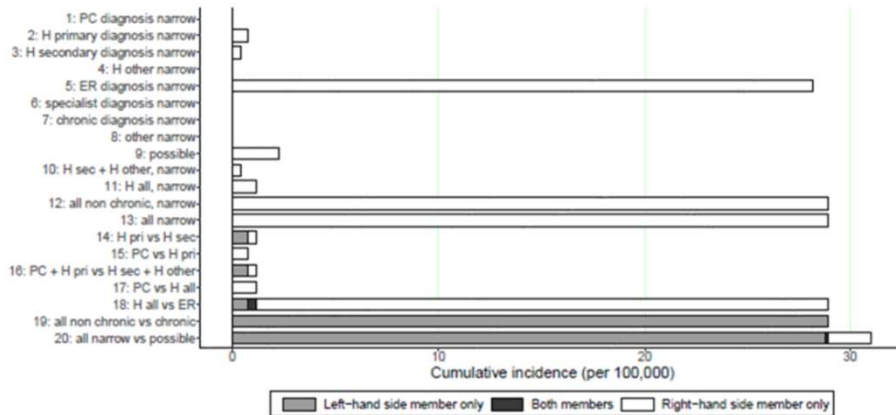
ES-BIFAP-PC-HOSP



ES-SIDIAP



IT-ARS



Incidence was highest in PC, 40.0 in SIDIAP and 42.5 in BIFAP

H on top of PC: +60.6% in BIFAP, and +71.9% in SIDIAP

ER on top of H: >999% in ARS, and +27.5% in SIDIAP

PC on top of H+ER: +96.8% in SIDIAP

Background incidence estimate might be significantly underestimated if any prompt among PC, H or ER is missing!!!

Triggers of alerts of lack of sensitivity

Trigger= missing prompt that contributed with $\geq 25\%$ of cases in other data sources

Primary care (PC) in 27 AESIs

Hospitalisation secondary diagnosis (H sec) in 21 AESIs

Hospitalisation primary diagnosis (H pri) in 16 AESIs

Emergency Room (ER) in 17 AESIs

Relevance: data source *fit-for-purpose* assessment and identification of false safety signals in *Obs/exp analysis!!!*

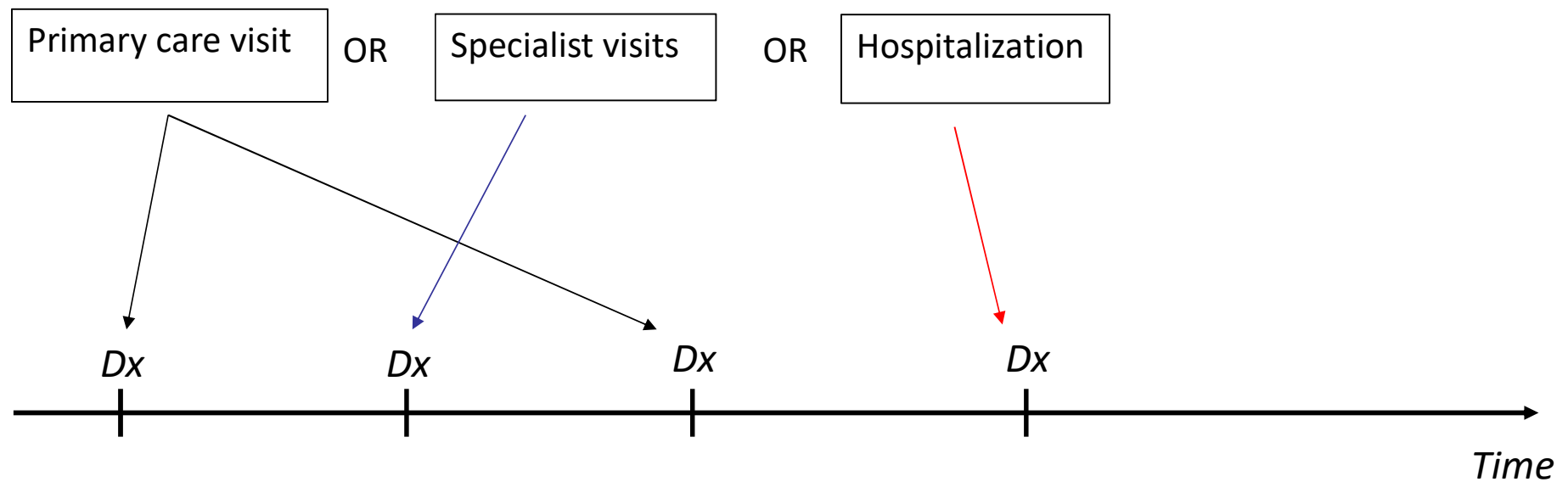
Limitation: case validation was not performed, therefore some cases may be false positives

Recommendation: component-level validation can avoid false safety alerts or missing true ones

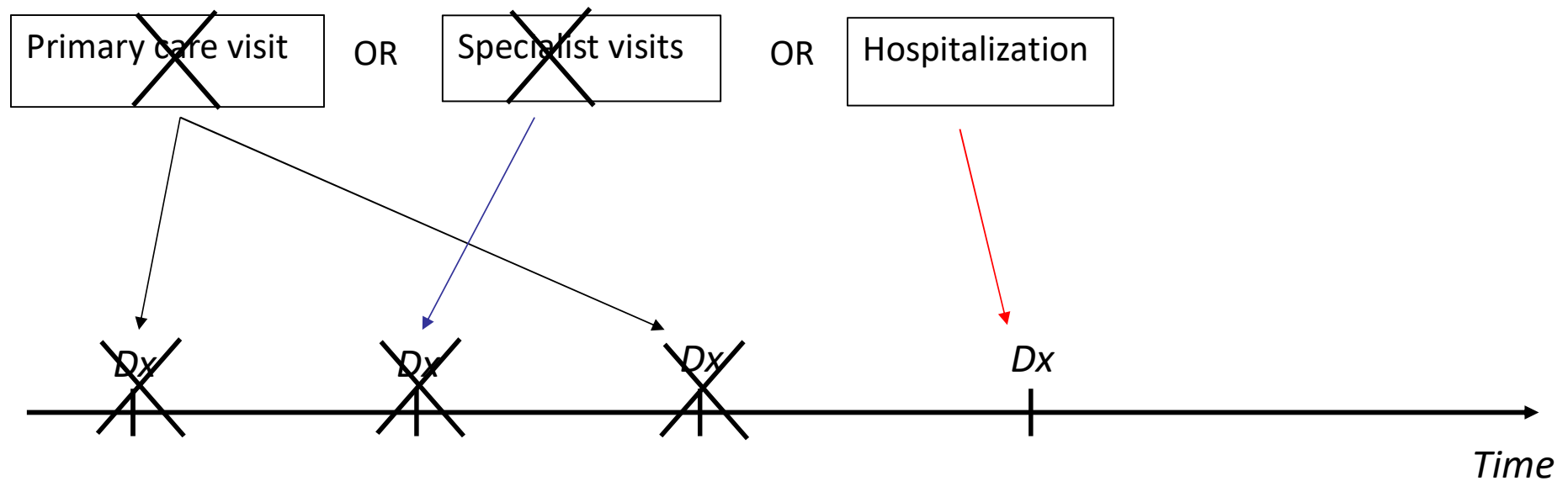
AESI	Missing H pri	Missing H sec	Missing PC	Missing ER	Missing possible
B_COAGDIS_AESI					
B_DIC_AESI					
B_HAEMOPHAGOLYMPHOHISTIO_AESI					
B_TP_AESI					
C_ARRH_AESI					
C_CAD_AESI					
C_MYOCARD_AESI					
C_PERICARD_AESI					
D_LIVERACUTE_AESI					
D_PANCRACUTE_AESI					
E_THYROIDAUTOIMM_AESI					
E_THYROIDSUBACUTE_AESI					
G_KIACUTE_AESI					
Im_ANAPHYLAXIS_AESI					
Im_KAWASAKI_AESI					
M_RHABDOMYOLISIS_AESI					
N_ADEM_AESI					
N_BELLP_AESI					
N_CONVULSION_AESI					
N_CVST_AESI					
N_GBS_AESI					
N_HEARINGLOSS_AESI					
N_MENINGOENC_AESI					
N_MYEELITISTRANSV_AESI					
N_NARCOLEPSY_AESI					
N_STROKEHEMO_AESI					
O_DEATHSUDDEN_AESI					
R_ARDS_AESI					
SO_ANOSMIAAGEUSIA_AESI					
Sk_ERYTHMULTI_AESI					
Sk_SCAR_AESI					
V_CHILBLAIN_AESI					
V_MICROANGIO_AESI					
V_THROMBOSISARTERIALALGOR_AESI					
V_VASCULITISSINGLEORG_AESI					
V_VTEALGORITHM_AESI					

Ablation of data prompts to estimate delay of case identification

- Prompts of diagnosis records for identification of Ulcerative colitis:



- Diagnosis record prompts for identification of Ulcerative colitis:



An explorative analysis with ulcerative colitis:

- Data instances from 4 diverse data sources from the Safety VAC project (<https://vac4eu.org/safety-vac-project/>) were used
- Impact of ablation of prompts on cohort entry was assessed as an explorative analyses not requested from the technical specification of the project

- Median number of days from start of study period to cohort entry=

Primary care<Specialist<Inpatient

- Ablation of prompts can inform time-to-event analysis as well as studies where disease stage can strongly affect the outcome of interest!!!

Days from start of study period to cohort entry, ablation analysis keeping "PC", median (IQR)

1313 (648-1943)	1107 (879-1465)	1288 (632.25-1939.5)	968 (452-1670)
--------------------	--------------------	-------------------------	-------------------

Days from start of study period to cohort entry, ablation analysis keeping "Inpatient or Specialist", median (IQR)

-	1308 (990-1599)	-	1058 (514-1610)
---	--------------------	---	--------------------

Days from start of study period to cohort entry, ablation analysis keeping "Inpatient", median (IQR)

1362.5 (688-1981)	1311 (989.75-1598)	1422 (731-1996)	-
----------------------	-----------------------	--------------------	---

Conclusions and Recommendations

Recommendations for the PhD student (Xabi):

- **Use the DIVERSE framework** to represent and describe data sources' diversity
- **Apply the component strategy to key study variables** (e.g. eligibility event / outcome) to support the generation of approximate estimates of validity
- **Apply ablation of prompts** based on component strategy to:
 - i) identify triggers of lack of sensitivity, or
 - ii) assess data source fitness-for-purpose, or
 - iii) estimate delay in case identification/cohort entry
- Validate key study variables, whenever feasible, per prompt/component algorithm

Requests for the Professors (Ersilia and Robert):

- Develop methods for designing validation studies that measure delay of event identification
- Develop statistical methods to use estimates of both validity and/or delay to adjust study results