Improving the reliability and scale of case validation

Patrick Ryan
Johnson & Johnson
Columbia University

Anna Ostropolets
Odysseus Data Services

Martijn Schuemie
Johnson & Johnson
UCLA

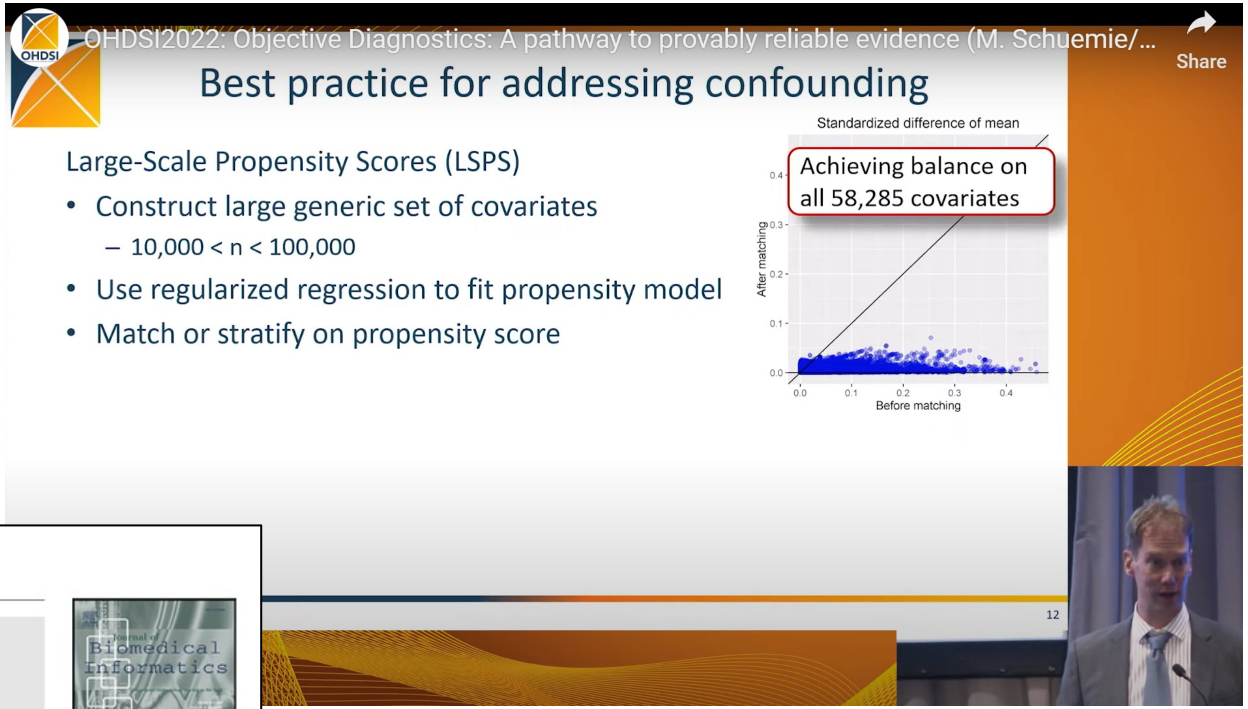
Continuing the journey to reliable evidence

- Bias in observational studies
  - Confounding
  - Selection bias
  - Measurement error

Continuing the journey to reliable evidence

- Bias in observational studies
  - **Confounding**
  - Selection bias
  - Measurement error

OHDSI 2022 Global Symposium



Journal of Biomedical Informatics 134 (2022) 104204

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

ELSEVIER

Original Research

Adjusting for indirectly measured confounding using large-scale propensity score

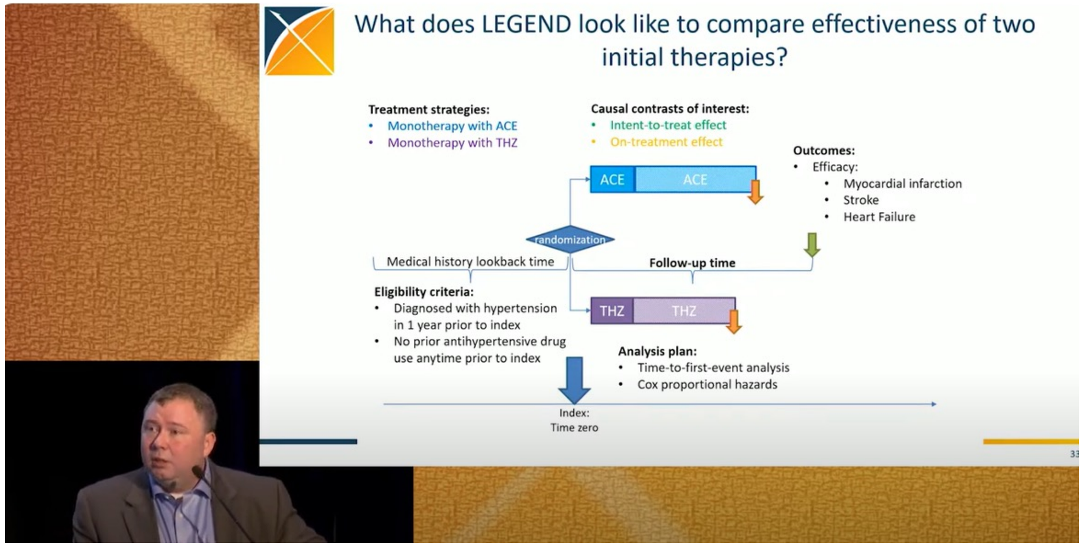
Linying Zhang [a], Yixin Wang [b], Martijn J. Schuemie [c], David M. Blei [d,e], George Hripcsak [a,f,*]

[a] Department of Biomedical Informatics, Columbia University Irving Medical Center, 622 W. 168th Street, PH20, New York, 10032, NY, USA
[b] Department of Statistics, University of Michigan, 1085 S University Ave, Ann Arbor, 48109, MI, USA

Continuing the journey to reliable evidence

- Bias in observational studies
  - Confounding
  - **Selection bias**
  - Measurement error

OHDSI 2019 Global Symposium



Received: 2 August 2018 | Revised: 29 April 2019 | Accepted: 5 May 2019
DOI: 10.1002/sim.8215

RESEARCH ARTICLE

WILEY Statistics in Medicine

A plea to stop using the case-control design in retrospective database studies

Martijn J. Schuemie[1,2,3] | Patrick B. Ryan[1,2,4] | Kenneth K.C. Man[5,6,7,8] | Ian C.K. Wong[5,6] | Marc A. Suchard[1,3,9,10] | George Hripcsak[1,4,11]

Continuing the journey to reliable evidence

- Bias in observational studies
  - Confounding
  - Selection bias
  - **Measurement error**

Learning from FDA guidance on real-world data

Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products

Guidance for Industry

DRAFT GUIDANCE

This guidance document is being distributed for comment purposes only.

Comments and suggestions regarding this draft document should be submitted within 60 days of publication in the *Federal Register* of the notice announcing the availability of the draft guidance. Submit electronic comments to https://www.regulations.gov. Submit written comments to the Dockets Management Staff (HFA-305), Food and Drug Administration, 5630 Fishers Lane, Rm. 1061, Rockville, MD 20852. All comments should be identified with the docket number listed in the notice of availability that publishes in the *Federal Register*.

For questions regarding this draft document or the RealWorld Evidence Program, please email CDERMedicalPolicy-RealWorldEvidence@fda.hhs.gov

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)
Oncology Center of Excellence (OCE)

September 2021
Real World Data/Real World Evidence (RWD/RWE)

1. Definition of Outcomes of Interest

Many outcomes involve diagnoses recorded by physicians as part of routine care. To minimize the effect of variability in practice by different physicians and over time (e.g., using different diagnosis and classification criteria, coding the same event in different ways), FDA recommends defining an outcome of interest based on the clinical, biological, psychological, and functional concepts of the condition, as appropriate. The conceptual definition for the outcome of interest (also referred to as the *case definition*) should reflect the medical and scientific understanding of the condition and might vary by study. For example, for anaphylaxis, the conceptual definition (or case definition) may include the following clinical criteria: sudden onset, rapid progression of signs and symptoms, ≥1 major dermatological criterion, and ≥ 1 major cardiovascular or respiratory criterion. The protocol should include a detailed description of the conceptual definition, including the signs, symptoms, and laboratory and radiology results that would confirm the outcome.

In OHDSI speak: Write a good clinical description upfront….

2. Ascer

To help identify potential cases in the selected data source and study population, operational definitions using diagnosis a
LOINC) and values, or unstr
pathology reports) should be
interest. If the operational d

….and then create a fully specified phenotype algorithm that aims to model the clinical description!

https://www.fda.gov/media/152503/download

Learning from FDA guidance on real-world data

Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products

Guidance for Industry

DRAFT GUIDANCE

This guidance document is being distributed for comment purposes only.

Comments and suggestions regarding this draft document should be submitted within 60 days of publication in the *Federal Register* of the notice announcing the availability of the draft guidance. Submit electronic comments to https://www.regulations.gov. Submit written comments to the Dockets Management Staff (HFA-305), Food and Drug Administration, 5630 Fishers Lane, Rm. 1061, Rockville, MD 20852. All comments should be identified with the docket number listed in the notice of availability that publishes in the *Federal Register*.

For questions regarding this draft document or the RealWorld Evidence Program, please email CDERMedicalPolicy-RealWorldEvidence@fda.hhs.gov

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)
Oncology Center of Excellence (OCE)

September 2021
Real World Data/Real World Evidence (RWD/RWE)

3. Validation of Outcomes

FDA expects validation of the outcome variable to minimize outcome misclassification. Although complete verification of the outcome variable is considered the most rigorous approach, there are scenarios where verifying outcome for every subject might not be feasible and assessing the performance of the operational definition of the outcome might suffice. Outcome validation involves using a clinically appropriate conceptual outcome definition to determine whether a patient's status, classified by an operational definition, truly represents the outcome of interest, typically by r
records in either electronic or pape

Two alternative use cases:
1) full caseset review
2) estimate measurement error

FDA recommends using standardized medical record review processes, including the use of standardized tools, documentation of process, and training of personnel. A standard and reproducible process is critical for minimizing intra- and inter-rater variability, especially for multi-site studies in which medical records usually cannot be shared across systems and a centralized medical record review is n
review, a standardized process helps t
adjudicators or a single adjudicator ov
statistic) is useful to ensure replicabili

Standardized tools to improve evidence reliability? Sounds like a job for OHDSI!

https://www.fda.gov/media/152503/download

Learning from FDA guidance on real-world data

Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products

Guidance for Industry

DRAFT GUIDANCE

This guidance document is being distributed for comment purposes only.

Comments and suggestions regarding this draft document should be submitted within 60 days of publication in the *Federal Register* of the notice announcing the availability of the draft guidance. Submit electronic comments to https://www.regulations.gov. Submit written comments to the Dockets Management Staff (HFA-305), Food and Drug Administration, 5630 Fishers Lane, Rm. 1061, Rockville, MD 20852. All comments should be identified with the docket number listed in the notice of availability that publishes in the *Federal Register*.

For questions regarding this draft document or the RealWorld Evidence Program, please email CDERMedicalPolicy-RealWorldEvidence@fda.hhs.gov

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)
Oncology Center of Excellence (OCE)

September 2021
Real World Data/Real World Evidence (RWD/RWE)

FDA recommends including a quantitative bias analysis in the protocol as a sensitivity analysis to demonstrate whether and how outcome misclassification might affect study results. The protocol should prespecify the indices (e.g., sensitivity, specificity, PPV, NPV) that will be used for quantitative bias analysis and describe how the selected indices will be measured in outcome validation.

…sounds good, but how are we going to estimate sensitivity, specificity, PPV, NPV via source record verification?

https://www.fda.gov/media/152503/download

Measurement error metrics

		Condition (as determined by "Gold standard")		
		Condition Positive	Condition Negative	
Test Outcome	Test Outcome Positive	**True Positive**	**False Positive** (Type I error)	Positive predictive value = $\frac{\Sigma \text{ True Positive}}{\Sigma \text{ Test Outcome Positive}}$
	Test Outcome Negative	**False Negative** (Type II error)	**True Negative**	Negative predictive value = $\frac{\Sigma \text{ True Negative}}{\Sigma \text{ Test Outcome Negative}}$
		Sensitivity = $\frac{\Sigma \text{ True Positive}}{\Sigma \text{ Condition Positive}}$	**Specificity** = $\frac{\Sigma \text{ True Negative}}{\Sigma \text{ Condition Negative}}$	

Case validation in practice

Vaccine 41 (2023) 4666–4678

Contents lists available at ScienceDirect

Vaccine

journal homepage: www.elsevier.com/locate/vaccine

Evaluation of potential adverse events following COVID-19 mRNA vaccination among adults aged 65 years and older: Two self-controlled studies in the U.S.

Azadeh Shoaibi [a,*], Patricia C. Lloyd [a], Hui-Lee Wong [a], Tainya C. Clarke [a], Yoganand Chillarige [b], Rose Do [b], Mao Hu [b], Yixin Jiao [b], Andrew Kwist [b], Arnstein Lindaas [b], Kathryn Matuska [b], Rowan McEvoy [b], Michelle Ondari [b], Shruti Parulekar [b], Xiangyu Shi [b], Jing Wang [b], Yun Lu [a], Joyce Obidi [a], Cindy K. Zhou [d], Jeffrey A. Kelman [c], Richard A. Forshee [a], Steven A. Anderson [a]

[a] *Office of Biostatistics and Pharmacovigilance, Center for Biologics Evaluation and Research, U. S. Food & Drug Administration, 10903 New Hampshire Ave., Building 71, Silver Spring, MD 20993, United States*

[b] *Acumen, LLC, 500 Airport Blvd. Suite 100, Burlingame, CA 94010, United States*

[c] *Centers for Medicare & Medicaid Services, 7500 Security Boulevard, Mail Stop B3-30-03, Baltimore, MD 21244-1850, United States*

[d] *Formerly Affiliated with US Food and Drug Administration, Silver Spring, MD, United States*

ARTICLE INFO

Article history:
Received 19 January 2023
Received in revised form 9 May 2023
Accepted 2 June 2023
Available online 14 June 2023

Keywords:
COVID-19 mRNA vaccines
COVID-19 vaccine safety
COVID-19 Pfizer-BioNTech vaccine
COVID-19 Moderna vaccine
Primary series
Monovalent booster

ABSTRACT

Background: Our near-real-time safety monitoring of 16 adverse events (AEs) following COVID-19 mRNA vaccination identified potential elevation in risk for six AEs following primary series and monovalent booster dose administration. The crude association with AEs does not imply causality. Accordingly, we conducted robust evaluation of potential associations.

Methods: We conducted two self-controlled case series studies of COVID-19 mRNA vaccines (BNT162b2 and mRNA-1273) in U.S. Medicare beneficiaries aged ≥ 65 years. Adjusted incidence rate ratio (IRRs) and 95 % confidence intervals (CIs) were estimated following primary series doses for acute myocardial infarction (AMI), pulmonary embolism (PE), immune thrombocytopenia (ITP), disseminated intravascular coagulation (DIC); and following monovalent booster doses for AMI, PE, ITP, Bell's Palsy (BP) and Myocarditis/Pericarditis (Myo/Peri).

Results: The primary series study included 3,360,981 individuals who received 6,388,542 primary series doses; the booster study included 6,156,100 individuals with one monovalent booster dose. The AMI IRR following BNT162b2 primary series and booster was 1.04 (95 % CI: 0.91 to 1.18) and 1.06 (95 % CI: 1.003 to 1.12), respectively; for mRNA-1273 primary series and booster, 1.01 (95 % CI: 0.82 to 1.26) and 1.05 (95 % CI: 0.998 to 1.11), respectively. The hospital inpatient PE IRR following BNT162b2 primary series and booster was 1.19 (95 % CI: 1.03 to 1.38) and 0.86 (95 % CI: 0.78 to 0.95), respectively; for mRNA-1273 primary series and booster, 1.15 (95 % CI: 0.94 to 1.41) and 0.87 (95 % CI: 0.79 to 0.96), respectively. The studies' results do not support that exposure to COVID-19 mRNA vaccines elevate the risk of ITP, DIC, Myo/Peri, and BP.

Conclusion: We did not find an increased risk for AMI, ITP, DIC, BP, and Myo/Peri and there was not consistent evidence for PE after exposure to COVID-19 mRNA primary series or monovalent booster vaccines.

2.3. Medical record review

To validate the claims-based AE definitions, medical record review (MRR) was conducted for cases identified from the primary series (AMI, PE (all care settings, hospital inpatient setting only), ITP (all care settings), DIC) and booster studies (BP, ITP (hospital inpatient setting only, primary diagnosis) Myo/Peri). For each case definition, medical records were obtained and adjudicated from a random sample of cases identified in both studies. Cases were then classified as true cases, non-cases, and indeterminate using standard clinical definitions when available [12–18]. When not available, case definitions for the AE were developed in consultation with specialist clinicians and consensus literature. For each AE definition, a positive predictive value (PPV) along with a corresponding 95 % confidence interval (CI) was estimated [19]. Table 3 presents classification decisions and PPV estimates by AE. These estimates were used to conduct a quantitative bias analysis (QBA) for each AE to assess the direction and magnitude of event misclassification [20].

Vaccine 41 (2023) 4666–4678

Contents lists available at ScienceDirect

Vaccine

journal homepage: www.elsevier.com/locate/vaccine

Evaluation of potential adverse events following COVID-19 mRNA vaccination among adults aged 65 years and older: Two self-controlled studies in the U.S.

Azadeh Shoaibi [a,*], Patricia C. Lloyd [a], Hui-Lee Wong [a], Tainya C. Clarke [a], Yoganand Chillarige [b], Rose Do [b], Mao Hu [b], Yixin Jiao [b], Andrew Kwist [b], Arnstein Lindaas [b], Kathryn Matuska [b], Rowan McEvoy [b], Michelle Ondari [b], Shruti Parulekar [b], Xiangyu Shi [b], Jing Wang [b], Yun Lu [a], Joyce Obidi [a], Cindy K. Zhou [d], Jeffrey A. Kelman [c], Richard A. Forshee [a], Steven A. Anderson [a]

[a] Office of Biostatistics and Pharmacovigilance, Center for Biologics Evaluation and Research, U. S. Food & Drug Administration, 10903 New Hampshire Ave., Building 71, Silver Spring, MD 20993, United States
[b] Acumen, LLC, 500 Airport Blvd. Suite 100, Burlingame, CA 94010, United States
[c] Centers for Medicare & Medicaid Services, 7500 Security Boulevard, Mail Stop B3-30-03, Baltimore, MD 21244-1850, United States
[d] Formerly Affiliated with US Food and Drug Administration, Silver Spring, MD, United States

A R T I C L E   I N F O

Article history:
Received 19 January 2023
Received in revised form 9 May 2023
Accepted 2 June 2023
Available online 14 June 2023

Keywords:
COVID-19 mRNA vaccines
COVID-19 vaccine safety
COVID-19 Pfizer-BioNTech vaccine
COVID-19 Moderna vaccine
Primary series
Monovalent booster

A B S T R A C T

Background: Our near-real-time safety monitoring of 16 adverse events (AEs) following COVID-19 mRNA vaccination identified potential elevation in risk for six AEs following primary series and monovalent booster dose administration. The crude association with AEs does not imply causality. Accordingly, we conducted robust evaluation of potential associations.
Methods: We conducted two self-controlled case series studies of COVID-19 mRNA vaccines (BNT162b2 and mRNA-1273) in U.S. Medicare beneficiaries aged ≥ 65 years. Adjusted incidence rate ratio (IRRs) and 95 % confidence intervals (CIs) were estimated following primary series doses for acute myocardial infarction (AMI), pulmonary embolism (PE), immune thrombocytopenia (ITP), disseminated intravascular coagulation (DIC); and following monovalent booster doses for AMI, PE, ITP, Bell's Palsy (BP) and Myocarditis/Pericarditis (Myo/Peri).
Results: The primary series study included 3,360,981 individuals who received 6,388,542 primary series doses; the booster study included 6,156,100 individuals with one monovalent booster dose. The AMI IRR following BNT162b2 primary series and booster was 1.04 (95 % CI: 0.91 to 1.18) and 1.06 (95 % CI: 1.003 to 1.12), respectively; for mRNA-1273 primary series and booster, 1.01 (95 % CI: 0.82 to 1.26) and 1.05 (95 % CI: 0.998 to 1.11), respectively. The hospital inpatient PE IRR following BNT162b2 primary series and booster was 1.19 (95 % CI: 1.03 to 1.38) and 0.86 (95 % CI: 0.78 to 0.95), respectively; for mRNA-1273 primary series and booster, 1.15 (95 % CI: 0.94 to 1.41) and 0.87 (95 % CI: 0.79 to 0.96), respectively. The studies' results do not support that exposure to COVID-19 mRNA vaccines elevate the risk of ITP, DIC, Myo/Peri, and BP.
Conclusion: We did not find an increased risk for AMI, ITP, DIC, BP, and Myo/Peri and there was not consistent evidence for PE after exposure to COVID-19 mRNA primary series or monovalent booster vaccines. These results support the favorable safety profile of COVID-19 mRNA vaccines administered in the U.S. elderly population.

Published by Elsevier Ltd.

Table 3
Summary of medical record review case adjudication results and PPVs associated with adverse events.

Outcome and Final Case Classifications	Risk and Control Cases Received*	Risk Cases†	Control Cases††
AMI *(Cases Requested: 125)*	**92**	**50**	**42**
Confirmed case	35	20	15
Probable	37	21	16
Possible	15	5	10
Not a case	3	2	1
Unable to be determined	2	2	0
PPV (Confirmed + Probable)¶	80.00 % (95 % CI: 70.59, 86.96)	85.42 % (95 % CI: 72.83, 92.75)	73.81 % (95 % CI: 58.93, 84.70)
PE *(Cases Requested: 179)*	**101**	**59**	**42**
Confirmed case	38	20	18
Probable	5	3	2
Possible	5	3	2
Not a case	46	29	17
Unable to be determined	7	4	3
PPV (Confirmed + Probable)¶	45.74 % (95 % CI: 36.04, 55.78)	41.82 % (95 % CI: 29.74, 54.97)	51.28 % (95 % CI: 36.20, 66.13)
PE (IP) *(Cases Requested: 42)*	**42**	**23**	**19**
Confirmed case	32	19	13
Probable	3	2	1
Possible	4	2	2
Not a case	3	0	3
Unable to be determined	0	0	0
PPV (Confirmed + Probable)¶	83.33 % (95 % CI: 69.40, 91.68)	91.30 % (95 % CI: 73.20, 97.58)	73.68 % (95 % CI: 51.21, 88.19)
ITP *(Cases Requested: 182)*	**91**	**53**	**38**
Confirmed case	2	1	1
Probable	1	1	0
Possible	6	2	4
Not a case	66	39	27
Unable to be determined	16	10	6
PPV (Confirmed + Probable)¶	4.00 % (95 % CI: 1.37, 11.11)	4.65 % (95 % CI: 1.28, 15.46)	3.12 % (95 % CI: 0.55, 15.74)
DIC *(Cases Requested: 128)*	**90**	**48**	**42**
Confirmed case	35	20	15
Probable	0	0	0
Possible	24	12	12
Not a case	23	11	12
Unable to be determined	8	5	3
PPV (Confirmed) ¶	42.68 % (95 % CI: 32.54, 53.48)	46.51 % (95 % CI: 32.51, 61.08)	38.46 % (95 % CI: 24.89, 54.10)
BP *(Cases Requested: 144)*	**79**	**79**	**N/A**
Confirmed case	3	3	N/A
Probable	7	7	N/A
Possible	10	10	N/A
Not a case	40	40	N/A
Unable to be determined	19	19	N/A
PPV (Confirmed + Probable) ¶	12.66 % (95 % CI: 7.02, 21.76)	12.66 % (95 % CI: 7.02, 21.76)	N/A

Abbreviations: AMI, acute myocardial infarction; ITP, immune thrombocytopenia; PE, pulmonary embolism; DIC, disseminated intravascular coagulation; BP, Bell's Palsy; PPV, positive predictive value; CI, Confidence Interval.
N/A Control Cases were not obtained for BP.
* Cases that occurred during either the risk or the control interval.
† Cases that occurred during the risk interval.
†† Cases that occurred during the control interval.
¶ PPV Calculation excludes cases that we are unable to be determined/assigned a case classification based on MRR.

PDS Pharmacoepidemiology & Drug Safety
ispe Official Journal of the International Society for Pharmacoepidemiology

Received: 1 September 2022 | Revised: 27 January 2023 | Accepted: 3 February 2023
DOI: 10.1002/pds.5601

COMMENTARY

WILEY

Validation to correct for outcome misclassification bias

Stephan Lanes | Daniel C. Beachler

Key Points

1. Outcome validation is often requested by regulators to address misclassification bias in database studies of drug safety and comparative effectiveness.
2. Validation studies commonly report only one positive predictive value (PPV) estimate.
3. Since a high value of PPV does not imply misclassification bias is negligible, and a low value of PPV does not imply misclassification bias is important, this approach does not adequately address outcome misclassification bias.
4. Validation should be designed to inform quantitative bias analysis that corrects results for misclassification bias.
5. To correct for misclassification bias, quantitative bias analysis requires parameters for false positive errors and false negative errors in each comparison group.

PDS Pharmacoepidemiology & Drug Safety
ispe Official Journal of the International Society for Pharmacoepidemiology

Received: 1 September 2022 | Revised: 27 January 2023 | Accepted: 3 February 2023
DOI: 10.1002/pds.5601

COMMENTARY

WILEY

Validation to correct for outcome misclassification bias

Stephan Lanes | Daniel C. Beachler

Department of Safety and Epidemiology, HealthCore, Wilmington, Delaware, USA

Correspondence
Stephan Lanes, HealthCore, 123 Justison St., Wilmington, DE 19801, USA.
Email: stephan.lanes@carelon.com

Funding information
HealthCore Inc.

KEYWORDS: misclassification bias, outcome misclassification, outcome validation, quantitative bias analysis

TABLE 1 Conventional approach to validation compared with validation used to support bias analysis for comparative studies of drug safety and effectiveness.

	Conventional validation approach	Validation to support bias analysis
Aim	Assess algorithm performance using PPV	Correct RR estimate for outcome misclassification bias
Method	1. Develop a primary algorithm with a high PPV 2. Apply primary algorithm to study population to identify cases 3. Sample people identified by the primary algorithm 4. Submit sample for outcome classification by gold standard (e.g., clinical expert adjudication of medical records) 5. Calculate PPV of primary algorithm as % cases identified by the algorithm that are confirmed by the gold standard	1. Develop a highly sensitive screening algorithm and a primary (high-PPV) algorithm 2. Apply screening algorithm to study population 3. Stratify people who meet screening algorithm by exposure status (0 = unexposed, 1 = exposed) 4. Sample people identified by the screening algorithm in both exposure groups (ensuring that there is also a sufficient number sampled who meet the primary algorithm) 5. Submit sample for outcome classification by gold standard (blinding adjudicators to exposure status) 6. Calculate PPV_0, PPV_1 of screening and primary algorithms as % confirmed by gold standard 7. Apply primary algorithm to confirmed cases identified by screening algorithm and calculate Se_0, Se_1 for the primary algorithm as % confirmed cases identified by primary algorithm 8. Use bias parameters in each comparison group as inputs for bias analysis to estimate RR for study population corrected for outcome misclassification bias[8,17]
Results	RR estimate uncorrected for outcome misclassification and PPV	RR estimate corrected to the gold standard for outcome misclassification bias
Validation sample	One sample (unspecified exposure status)	Samples for each comparison group (e.g., exposed, unexposed)
Interpretation	Impact of outcome misclassification on reported effect estimate is unknown	Effect estimate is corrected for outcome misclassification

Abbreviations: PPV, positive predictive value; PPV_0, positive predictive value in unexposed group; PPV_1, positive predictive value in exposed group; RR, relative risk; Se, sensitivity; Se_0, sensitivity in unexposed group; Se_1, sensitivity in exposed group.

Case validation to support bias analysis from Lanes PDS 2023

Target

'Highly sensitive screening' algorithm:

Estimate Sensitivity_0 = % of 'true cases' in 'highly sensitive' algorithm contained within 'primary' algorithm

Sample and validate

'Primary (high PPV)' algorithm:

Estimate PPV_0 = % of 'primary' algorithm cases validated to be 'true cases'

Sample and validate

Comparator

'Highly sensitive screening' algorithm:

Estimate Sensitivity_1 = % of 'true cases' in 'highly sensitive' algorithm contained within 'primary' algorithm

Sample and validate

'Primary (high PPV)' algorithm:

Estimate PPV_1 = % of 'primary' algorithm cases validated to be 'true cases'

Sample and validate

OHDSI's progress in estimating measurement error

Journal of Biomedical Informatics 97 (2019)

Contents lists available at ScienceDi

Journal of Biomedical Info

journal homepage: www.elsevier.com/j

ELSEVIER

PheValuator: Development and evaluation of a phenc evaluator

Joel N. Swerdel[a,b,*], George Hripcsak[b,c]

[a] Janssen Research & Development, 920 Route 202, Raritan, NJ 088
[b] OHDSI Collaborators, Observational Health Data Sciences and Info
[c] Columbia University, 622 West 168th Street, PH20, New York, NY

ELSEVIER

Original Research

PheValuator 2.0: Met approach to semi-aut

Joel N. Swerdel[a,c,*], Marti

[a] Janssen Research and Development, Titusville
[b] Columbia University, New York, NY, USA
[c] Observational Health Data Sciences and Infor

PheValuator: noisy labels u then estimate for your 'prim

Examining differential measurement error due to race, age, and sex in mental health disorders using PheValuator.

Presenter: Joel Swerdel

BACKGROUND

- Misclassification of health condition status is a serious threat to validity in research involving observational data from insurance administrative claims data.
- The problem would be exacerbated if there was differential misclassification between population subgroups.
- For example, is the degree of misclassification the same for older vs. younger subjects when examining mental health conditions, such as bipolar disease.
- PheValuator is a methodology within the OHDSI toolstack that uses diagnostic predictive modeling to determine the probability that a subject has a specific health outcome during a specified period of time.(1)
- It was designed to evaluate the performance characteristics, i.e., sensitivity, specificity, and positive and negative predictive value, of phenotype algorithms in observational data.
- The objective of this study was to use the results from PheValuator to estimate subpopulation differences between phenotype algorithm sensitivity and positive predictive value (PPV) across a set of mental health disorders. Populations were subgrouped by race, sex, and age.

METHODS

- We developed phenotype algorithms for eight mental health disorders: anxiety disorder, attention deficit hyperactivity disorder (ADHD), autism, bipolar disorder, depression, post-traumatic stress disorder (PTSD), schizoaffective disorder, and schizophrenia.
- We examined these conditions in three databases which include subjects of all ages: IBM® MarketScan® Multi-State Medicaid Database (MDCD), Optum's Clinformatics® Data Mart (SES), and Optum's Longitudinal EHR repository (EHR).
- We stratified the subjects in the analysis by sex; race, Black and White; and age, 65 years old (YO) and younger and 66 YO and older.
- We used PheValuator (V2.2.6) for the analyses.
- We developed algorithms for each condition using an empirical process previously documented involving the use of the standard OHDSI tools ATLAS, CohortDiagnostics, PHOEBE, and PheValuator.

- **We estimated and compared:**

Sensitivity = true positives/(true positives + false negatives)

PPV = true positives/(true positives + false positives)

for each condition across the three databases.

Researchers may introduce bias into their mental health research if they assume non-differential misclassification by sex, age, or race.

Males: higher sensitivity estimates for:
- ADHD
- Autism
- Schizophrenia
- Schizoaffective disorder

Females: higher sensitivity estimates for:
- Anxiety
- Bipolar disorder
- Depression
- PTSD

Blacks: higher sensitivity estimates for:
- Schizophrenia
- Schizoaffective disorder

Whites: higher sensitivity estimates for:
- Anxiety
- Bipolar disorder
- Depression

Younger Subjects: higher sensitivity estimates for:
- Autism
- Bipolar disease
- PTSD
- Schizophrenia
- Schizoaffective disorder

RESULTS

- **By Sex**: We found higher estimates for sensitivity for female subjects compared to male subjects for anxiety, bipolar, depression, and PTSD as shown by the positive values in each graph. We found lower estimates for sensitivity for female subjects compared to male subjects for ADHD, autism, schizoaffective disorder, and schizophrenia as shown by the negative values in each graph.
- **By Race**: We found large differences in sensitivity estimates for schizoaffective disorder and schizophrenia between Blacks and Whites where the sensitivity for Blacks was higher than that for Whites. We found consistently lower sensitivity estimates for Blacks compared to Whites for anxiety, bipolar disorder, and depression.
- **By Age**: We found that in five of the disorders, autism, bipolar disorder, PTSD, schizoaffective disorder, and schizophrenia, the estimates for sensitivity were much lower in the older age group than the younger age group.
- **PPV**: The differences were much smaller for PPV estimates between the groups compared to the sensitivity estimate differences for race and sex. The differences were larger when comparing age differences.

CONCLUSIONS

- In this study we examined differences in the performance characteristics, sensitivity and PPV, for phenotype algorithms for eight mental health disorders for subgroup populations divided by race, sex, and age.
- **We found large differences in sensitivity estimates for many of the conditions in each of the subgroups.**
- **The results from this study parallel findings in previous research examining sex, race, and age disparities in diagnosis and treatment of different mental health disorders. For example:**
  - Hull et al suggest that females are underdiagnosed for autism compared to males possibly due to the expression of autism in females that do not meet diagnostic criteria.[3] In our estimates the sensitivity of the autism algorithm was significantly lower for females indicating that the number of false negatives, i.e., missing diagnosis codes for autism, was higher in females than males.
  - van Niekerk and colleagues report that autism disorder is underdiagnosed in the older population especially those presenting with comorbid psychiatric disorders.[5] In our current study, we find lower sensitivity for autism in those over age 65.
  - Vanderminden and Esala found that females were more likely diagnosed with anxiety disorder compared to males as were Whites compared to Blacks.[4] This is similar to our findings of higher sensitivity, i.e., fewer missed diagnoses, for females compared to males as well as lower sensitivity in Blacks compared to Whites.
- Future research should be conducted to determine how these differences may affect study results such as those from drug comparative effectiveness analyses.

REFERENCES

Joel N. Swerdel[1,2] and Dmytro Dymshyts[1,2]

Scan to View More Info

Journal of the American Medical Informatics Association, 26(12), 2019, 1664–1674
doi: 10.1093/jamia/ocz094
Advance Access

Same effect size estimate and PPV can yield wildly different true effects based on differential misclassification….

Review

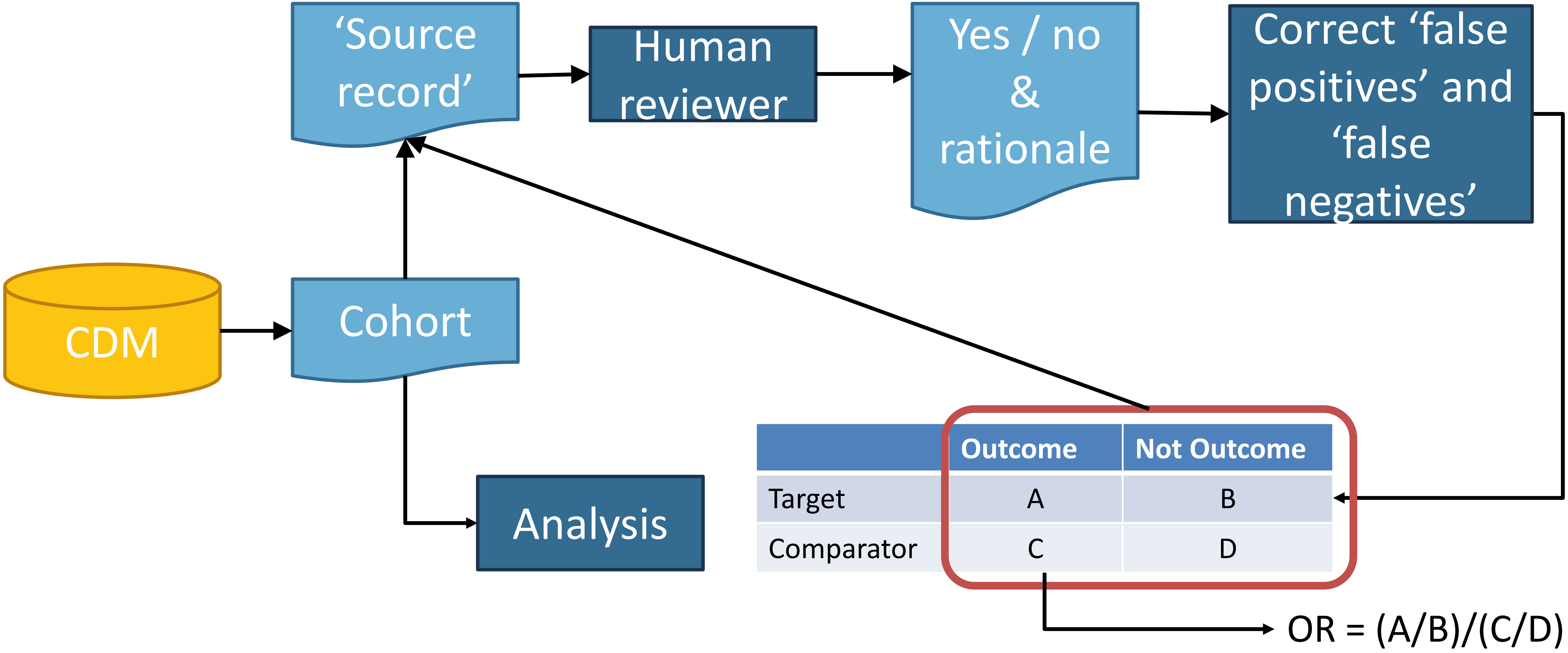
A primer on quantitative bias analysis with positive predictive values in research using electronic health data

Sophia R Newcomer,[1,2] Stan Xu,[2] Martin Kulldorff,[3] Matthew F Daley,[2,4] Bruce Fireman,[5] and Jason M Glanz[2,6]

….so we need to find a more reliable and scalable approach to estimating measurement error (PPV, sensitivity, specificity, NPV) across cohorts

Observed RR	Overall PPV	Stratified PPV	Sensitivity/ specificity	True RR	Impact
0.87 (0.76-0.99)	94%	PPV1 = 93% PPV0 = 94%	SN1=95%; SN0=95% SP1=99.85%; 99.85%	0.86 (0.75-0.98)	No difference
0.87 (0.76-0.99)	93%	PPV1 = 79% PPV0 = 93%	SN1=95%; SN0=90% SP1=99.55%; 99.82%	0.70 (0.60-0.81)	Larger effect
0.87 (0.76-0.99)	93%	PPV1 = 96% PPV0 = 93%	SN1=85%; SN0=95% SP1=99.90%; 99.81%	1.00 (0.88-1.13)	No effect

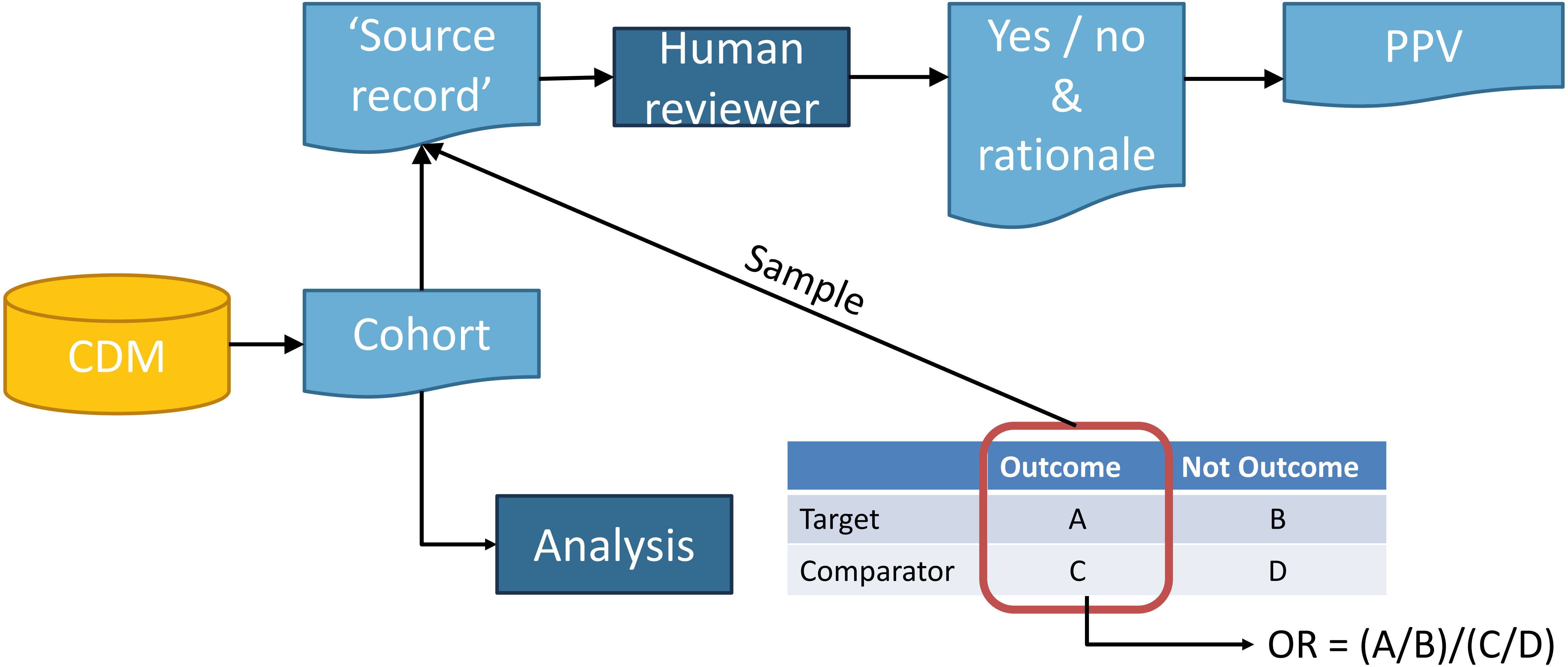
Case validation in the evidence generation workflow: Full caseset review



	Outcome	Not Outcome
Target	A	B
Comparator	C	D

OR = (A/B)/(C/D)

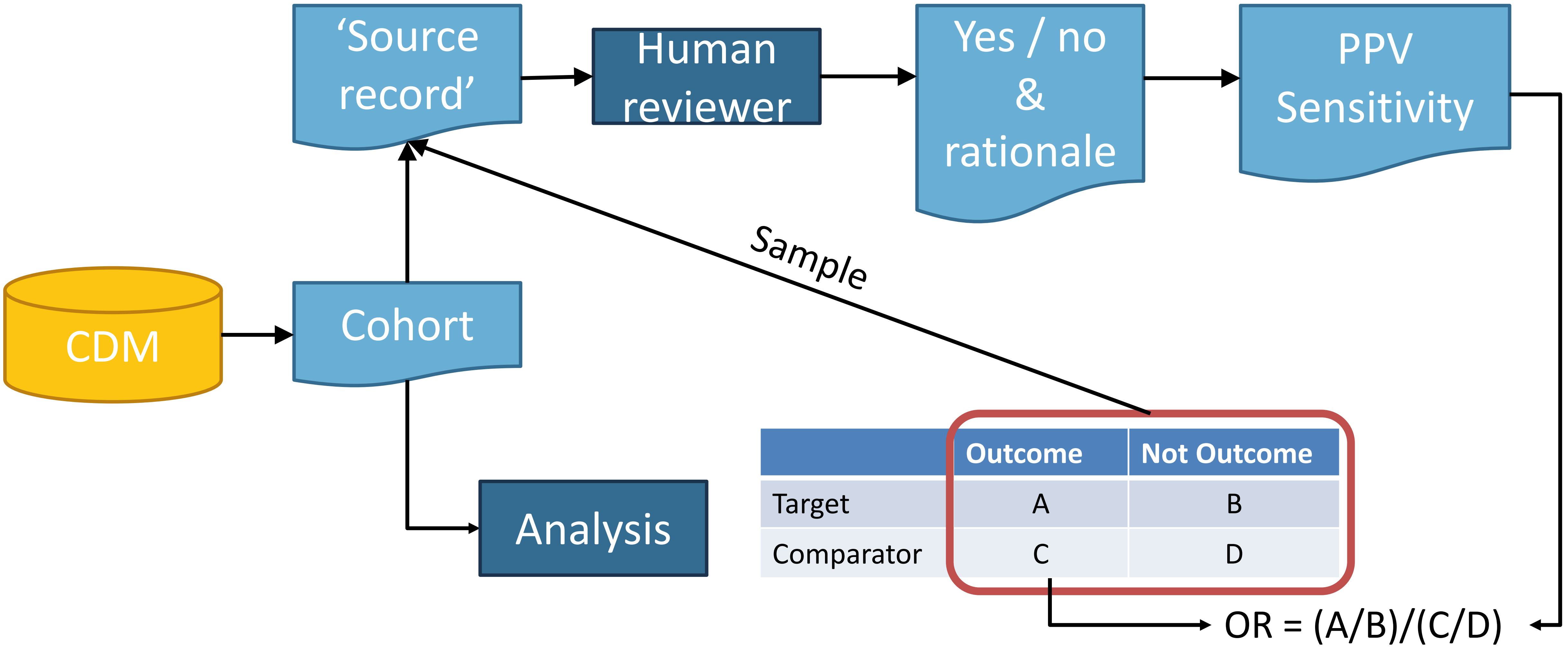
Case validation in the evidence generation workflow: Conventional validation approach



	Outcome	Not Outcome
Target	A	B
Comparator	C	D

OR = (A/B)/(C/D)

Case validation in the evidence generation workflow: Estimate measurement error for quantitative bias analysis



	Outcome	Not Outcome
Target	A	B
Comparator	C	D

Challenges and opportunities

- Case validation is expected for regulatory-grade real-world evidence, but source record verification is time- and resource-intensive and has unknown operating characteristics
- For quantitative bias analysis, estimating positive predictive value is insufficient → need measures of both false positive and false negative errors with target and comparator
- How can we make case validation more reliable?
- How can we make case validation more scalable?

End-stage renal disease: Clinical description

- End-stage renal disease is a terminal illness with a glomerular filtration rate (GFR) of less than 15 mL/min. This is the 5th and final stage of Chronic Kidney Disease (CKD).
- The most common cause of ESRD in the US is diabetic nephropathy, followed by hypertension.
- Other etiologies can include glomerulonephritis, cystic kidney disease, recurrent kidney infection, chronic obstruction, etc.
- The disease can present with nausea, vomiting, metabolic, hematologic, electrolyte derangements, seizures, coma, bleeding diathesis, refractory fluid overload, hypertension unresponsive to pharmacotherapy, uremic pericarditis, etc.
- Vigilant monitoring of GFR and proteinuria in diabetics and non-diabetics is essential for managing disease progression in patients with chronic kidney disease.
- Early referral to specialists is necessary for timely dialysis or renal transplant planning.

https://www.ncbi.nlm.nih.gov/books/NBK499861/

Let’s do some case validation together!

pollev.com/PatrickRyan800

Hypothetical clinical narrative

A 50-year-old male with a history of type 2 diabetes mellitus, hypertension, and chronic kidney disease due to type 2 diabetes mellitus presented for a pharmacy visit followed by an outpatient visit. During the visit, he was diagnosed with chronic kidney disease stage 5, end-stage renal disease, and other related complications. Prior to the visit, he had been diagnosed with chronic kidney disease, anemia, and vitamin D deficiency. After the visit, he was diagnosed with end-stage renal disease, anemia, hyperlipidemia, and hyperparathyroidism due to renal insufficiency. He was prescribed calcitriol for 54 days.

Join by Web **PollEv.com/patrickryan800** Join by Text Send **patrickryan800** to **22333**

Is this a case?

Hypothetical clinical narrative

A 50-year-old male with a history of type 2 diabetes mellitus, hypertension, and chronic kidney disease due to type 2 diabetes mellitus presented for a pharmacy visit followed by an outpatient visit. During the visit, he was diagnosed with chronic kidney disease stage 5, end-stage renal disease, and other related complications. Prior to the visit, he had been diagnosed with chronic kidney disease, anemia, and vitamin D deficiency. After the visit, he was diagnosed with end-stage renal disease, anemia, hyperlipidemia, and hyperparathyroidism due to renal insufficiency. He was prescribed calcitriol for 54 days.

Yes	0%
No	0%

Hypothetical clinical narrative #2

An 80-year-old female patient had an outpatient visit followed by a laboratory visit. The primary diagnosis during the visit was acute renal failure syndrome, while the secondary diagnoses included chronic kidney disease due to hypertension, chronic kidney disease stage 2, essential hypertension, hyperlipidemia, hypothyroidism, proteinuria, renal disorder due to type 2 diabetes mellitus, renal function tests abnormal, and type 2 diabetes mellitus without complication. Prior to the visit, the patient had been diagnosed with hyperlipidemia and hypothyroidism. No treatments were recorded before the visit. Laboratory tests conducted during the visit showed abnormal high levels of creatinine, urea nitrogen, and urea nitrogen/creatinine ratio, while the glomerular filtration rate was normal. The patient's urine creatinine level was normal. After the visit, the patient continued to be diagnosed with hyperlipidemia and hypothyroidism, but there was no evidence of end stage renal disease.

When poll is active respond at **PollEv.com/patrickryan800** Send **patrickryan800** to **22333**

Is this a case?

Hypothetical clinical narrative #2

An 80-year-old female patient had an outpatient visit followed by a laboratory visit. The primary diagnosis during the visit was acute renal failure syndrome, while the secondary diagnoses included chronic kidney disease due to hypertension, chronic kidney disease stage 2, essential hypertension, hyperlipidemia, hypothyroidism, proteinuria, renal disorder due to type 2 diabetes mellitus, renal function tests abnormal, and type 2 diabetes mellitus without complication. Prior to the visit, the patient had been diagnosed with hyperlipidemia and hypothyroidism. No treatments were recorded before the visit. Laboratory tests conducted during the visit showed abnormal high levels of creatinine, urea nitrogen, and urea nitrogen/creatinine ratio, while the glomerular filtration rate was normal. The patient's urine creatinine level was normal. After the visit, the patient continued to be diagnosed with hyperlipidemia and hypothyroidism, but there was no evidence of end stage renal disease.

Yes

0%

No

0%

Hypothetical clinical narrative #3

The patient is a 90-year-old female who had an emergency room visit and an 8-day inpatient stay. She had a history of chronic kidney disease, hypertension, osteoporosis, and other comorbidities. During her visit, she was diagnosed with chronic kidney disease stage 4 and stage 5 due to hypertension. She also had a history of chronic kidney disease stages 2, 3, and 4, as well as malignant hypertensive chronic kidney disease. The patient was treated with furosemide and calcitriol during and after her visit.

When poll is active respond at **PollEv.com/patrickryan800** Send **patrickryan800** to **22333**

Is this a case?

Hypothetical clinical narrative #3

The patient is a 90-year-old female who had an emergency room visit and an 8-day inpatient stay. She had a history of chronic kidney disease, hypertension, osteoporosis, and other comorbidities. During her visit, she was diagnosed with chronic kidney disease stage 4 and stage 5 due to hypertension. She also had a history of chronic kidney disease stages 2, 3, and 4, as well as malignant hypertensive chronic kidney disease. The patient was treated with furosemide and calcitriol during and after her visit.

Yes

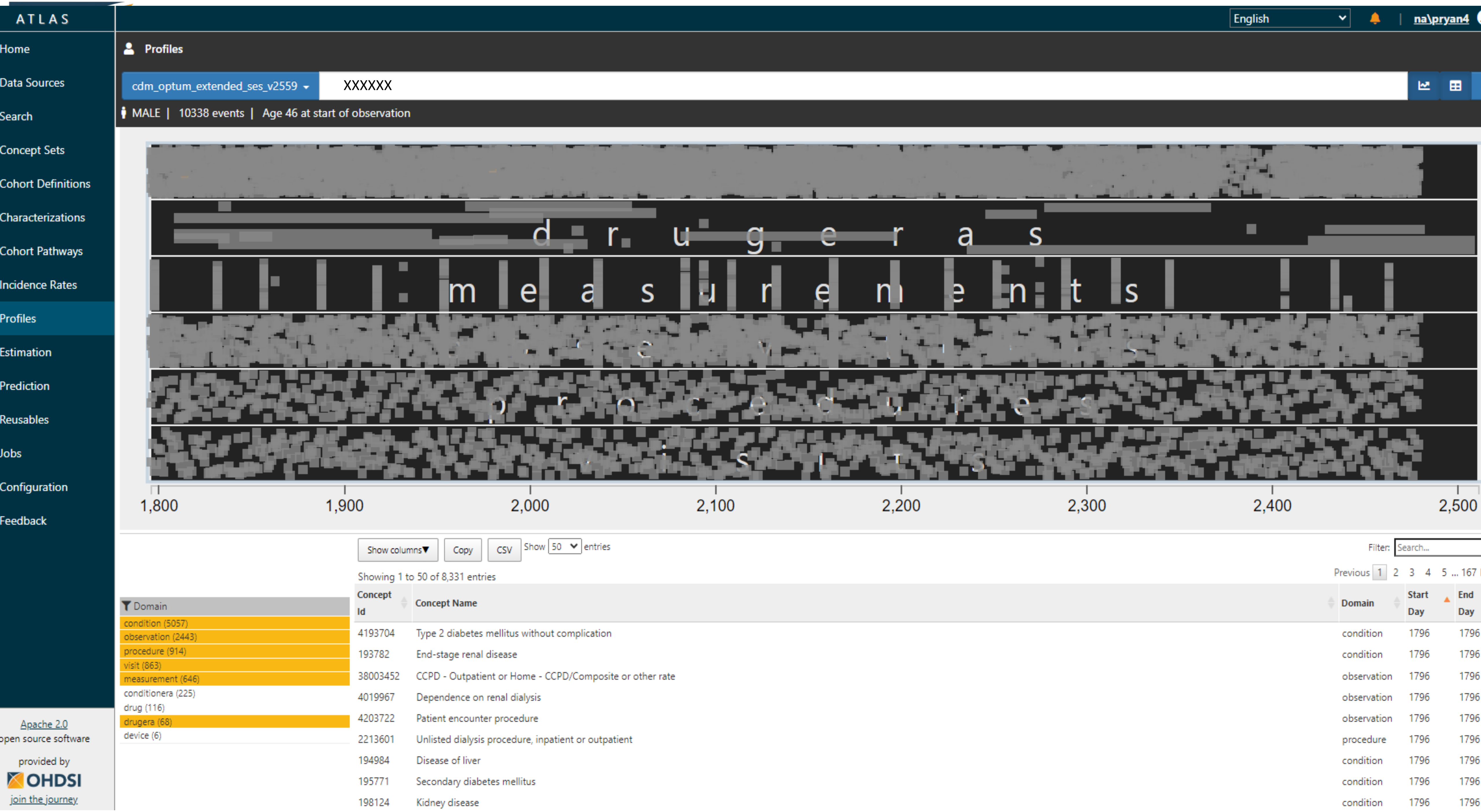
0%

No

0%

Improving reliability in case validation

Anna Ostropolets


ATLAS
English
na\pryan4
Home
Data Sources
Search
Concept Sets
Cohort Definitions
Characterizations
Cohort Pathways
Incidence Rates
Profiles
Estimation
Prediction
Reusables
Jobs
Configuration
Feedback
Profiles
cdm_optum_extended_ses_v2559
XXXXXX
MALE | 10338 events | Age 46 at start of observation
d r u g e r a s
m e a s u r e m e n t s
1,800
1,900
2,000
2,100
2,200
2,300
2,400
2,500
Show columns
Copy
CSV
Show 50 entries
Filter: Search...
Showing 1 to 50 of 8,331 entries
Previous 1 2 3 4 5 ... 167
Domain
condition (5057)
observation (2443)
procedure (914)
visit (863)
measurement (646)
conditionera (225)
drug (116)
drugera (68)
device (6)
Concept Id | Concept Name | Domain | Start Day | End Day
4193704 | Type 2 diabetes mellitus without complication | condition | 1796 | 1796
193782 | End-stage renal disease | condition | 1796 | 1796
38003452 | CCPD - Outpatient or Home - CCPD/Composite or other rate | observation | 1796 | 1796
4019967 | Dependence on renal dialysis | observation | 1796 | 1796
4203722 | Patient encounter procedure | observation | 1796 | 1796
2213601 | Unlisted dialysis procedure, inpatient or outpatient | procedure | 1796 | 1796
194984 | Disease of liver | condition | 1796 | 1796
195771 | Secondary diabetes mellitus | condition | 1796 | 1796
198124 | Kidney disease | condition | 1796 | 1796
Apache 2.0
open source software
provided by
OHDSI
join the journey

Main challenge of patient data review

Challenge: high volume of data, which is hard to navigate and interpret

Solution: KEEPER - Knowledge-Enhanced Electronic Profile Review system on structured data from EHR or claims data sources

KEEPER principles

Principle 1: Adherence to clinical reasoning

KEEPER applies general principles and steps of diagnostic clinical reasoning

Principle 2: Standardization

Both input and output are standardized across data sources and condition

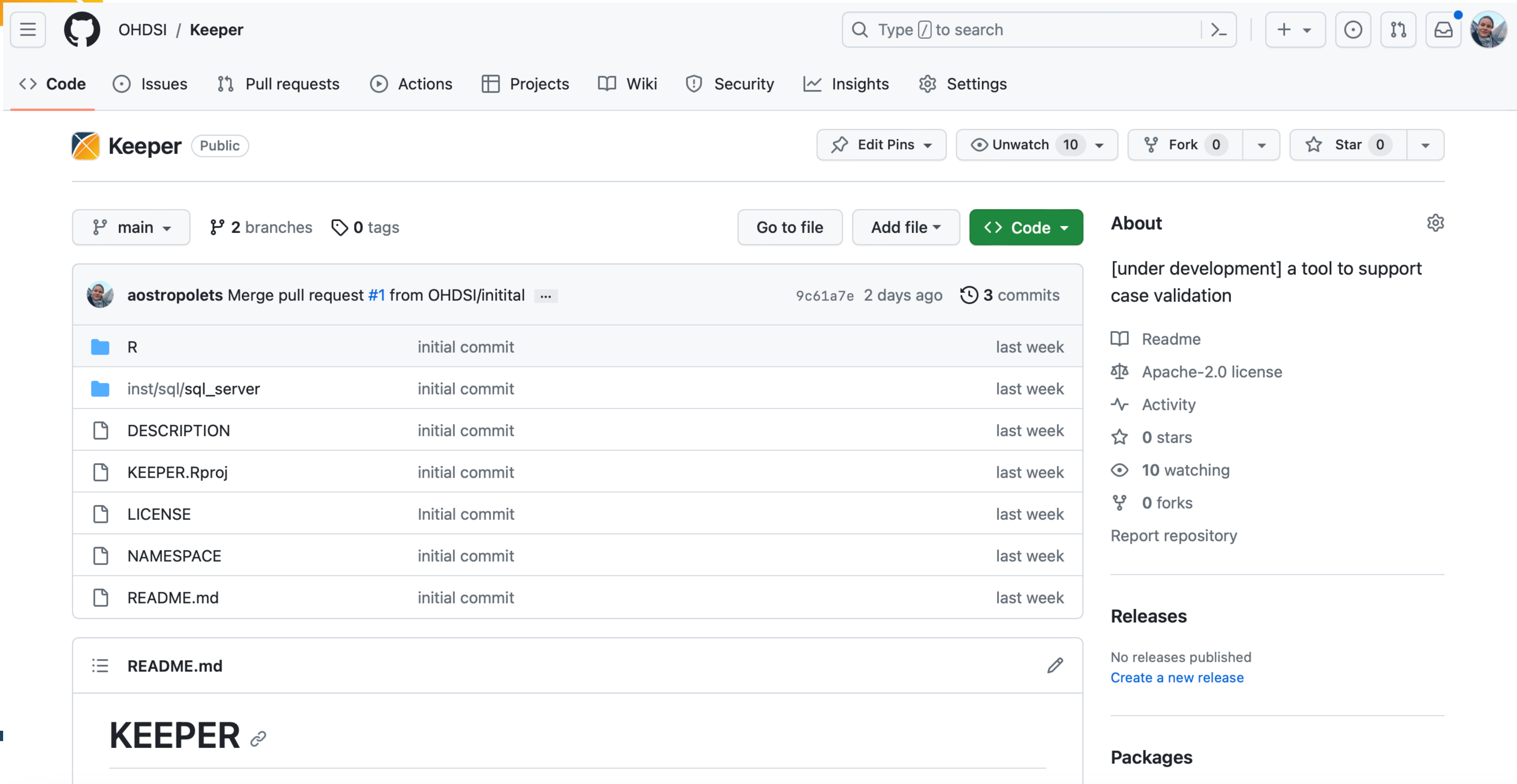
Principle 3: Dimensionality reduction

Only extract relevant information

KEEPER applies general principles and steps of diagnostic clinical reasoning

- Clinical presentation
- Clinical plausibility
  - Demographics
  - Risk factors and co-morbidities
  - Previous history of disease
  - Differential diagnoses
- Diagnostic procedures
- Treatment procedures and medications
- Follow-up care and complications

KEEPER as an OHDSI package



KEEPER as an OHDSI package

Per disease:
Concept sets per
KEEPER category

Ex: ESRD Symptoms:
vomiting, edema, dyspnea

Cohort

Ex: ESRD

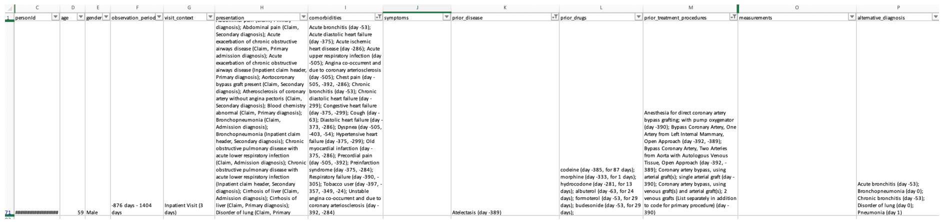
KEEPER
data extraction

Time
windows
per category

Ex: Symptoms: -30d to
0d before index date

CSV table:
record per person,
column per element

Person_id	Symptoms
1	Vomiting and nausea (day -29); Dyspnea (day -11);…



KEEPER output for one case with suspected ESRD

Column in KEEPER	Content of column
Demographics (age, sex)	48 yo, Male
Observation period	-931 days - 315 days
Visit context	Pharmacy visit->Outpatient Visit
Presentation	Chronic kidney disease due to type 2 diabetes mellitus (Primary admission diagnosis); Chronic kidney disease stage 5 (Admission diagnosis); …
Comorbidities	
Symptoms	
Prior disease	Anemia in chronic kidney disease (day –898, -815, -796, -15);...
Prior treatment procedures and drugs	
Diagnostic procedures and labs	
Alternative diagnosis	Acute renal failure syndrome (day -15, 31)
After disease (progression)	CKD stage 5 (day 94, 171, 213, 271); End-stage renal disease (day 1, 896);...
After treatment procedures and drugs	calcitriol (day 287, for 54 days);

KEEPER experiment overview

GOLD STANDARD (AO, GH)

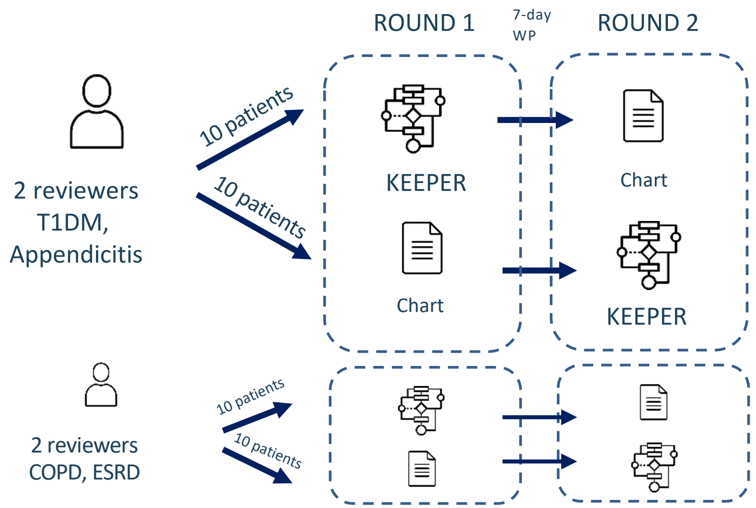
Random sample of 20 patients per eMERGE algorithm
Iterative review on full chart + all structured data

	T1DM	Acute appendicitis	COPD	ESRD
Case	12	15	11	13
Control	8	5	9	7

KEEPER PROFILES

Created KEEPER profiles for 80 patients

EXPERIMENT (AE, LR, MS, SAH)



DM type 1, reviewer 1			
	Time	Positives	Negatives
KEEPER	13 min	15	5
Chart review	28 min	12	8

DM type 1, reviewer 2			
	Time	Positives	Negatives
KEEPER	33 min	13	7
Chart review	55 min	10	10

PERFORMACE METRICS

- Time to review
- Inter-rater agreement (LR vs MS, AE vs SAH)
- Inter-method agreement (KEEPER vs chart review)
- Agreement with gold standard

DM type 1, reviewer 1 accuracy			
		Gold standard, case	Gold standard, control
KEEPER	Positive	TP = 12	FP = 3
	Negative	FN = 0	TN = 5
Chart review	Positive	TP = 10	FP = 3
	Negative	FN = 2	TN = 5

Data preparation

GOLD STANDARD (AO, GH)

Random sample of 20 patients per eMERGE algorithm
Iterative review on full chart + all structured data

	T1DM	Acute append	COPD	ESRD
Case	12	15	11	13
Control	8	5	9	7

KEEPER PROFILES

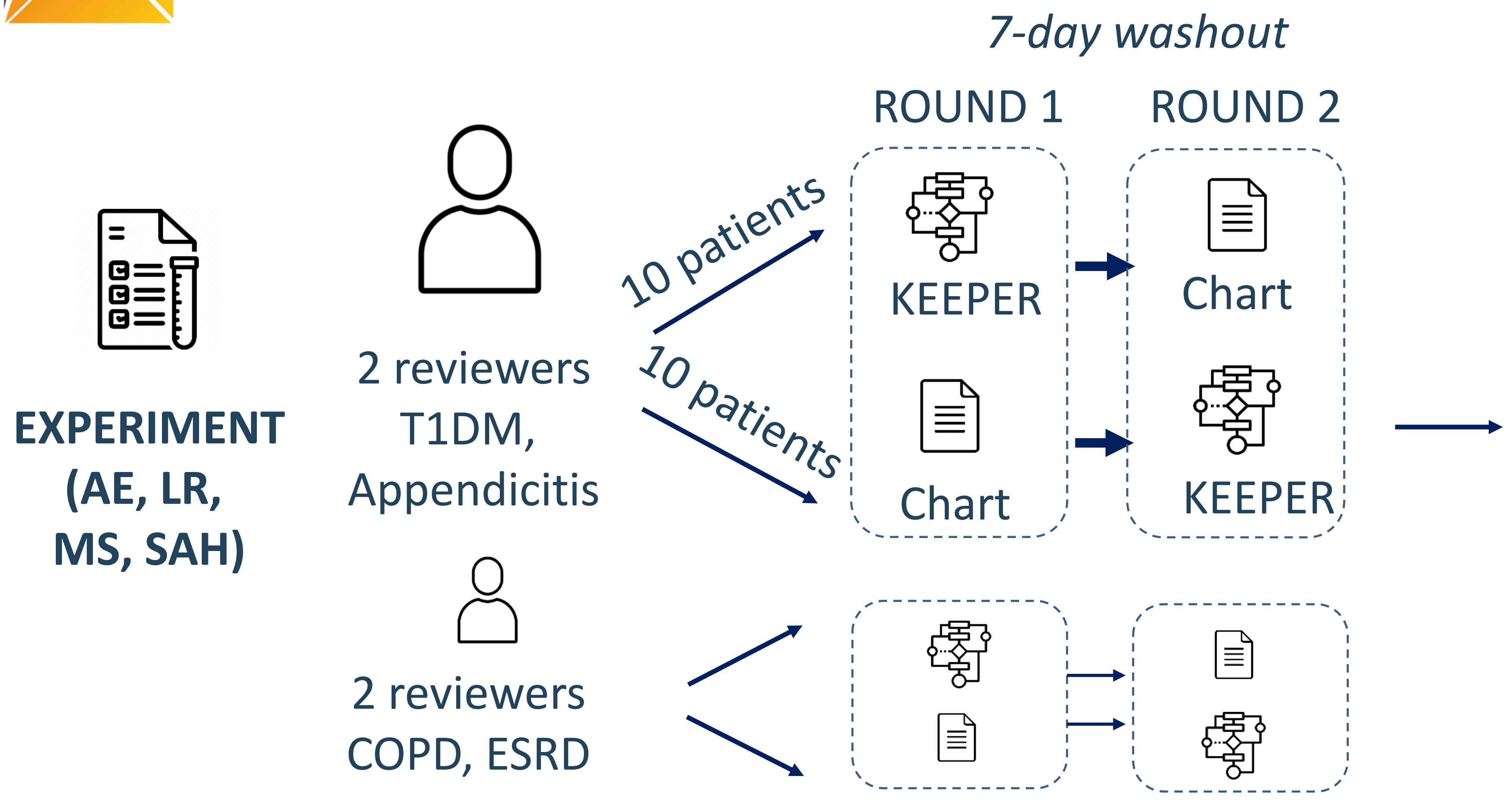
Created KEEPER profiles for 80 patients

Columbia University EHR

Inputs (concepts) for KEEPER

	T1DM	COPD	ESRD	Appendicitis
Symptoms		Cough, chest pain, SOB, wheezing, tachycardia		Abdominal pain
Comorbidities and risk factors	Hypertension, obesity, hyperlipidemia, disorders of pancreas, candidiasis, PCOS	Smoking, disorders of respiratory system, heart failure, IHD	Kidney disorders, multiple myeloma, lupus, HF	
Diagnostic procedures		Spirometry, chest x-ray or CT, bronchoscopy	Ultrasound or CT of kidneys	Ultrasound, CT or X-ray of abdomen and pelvis, laparoscopy
Measurements	Blood glucose, HA1C, insulin and pancreatic antibodies, c-peptide		Creatinine, eGFR, urea nitrogen	Leukocytes
Treatments: - Procedures - Drugs	Insulin, oral glucose lowering drugs	Lung surgery, LABA, SABA, LAMA, steroids	Renal transplant, dialysis, diuretics, tacrolimus, epoetin	Appendectomy, antibiotics
Differential diagnoses	Type II diabetes, pancreatic diabetes, hyperglycemia in other conditions	Asthma, lung cancer, interstitial lung disease, bronchiectasis	Acute renal failure, other stages of CKD	Disorders/Ca of intestine, GERD, hernias, genitourinary disorders
Complications	Diabetic neuropathy, nephropathy, eye disorders	Bronchiectasis, atelectasis, emphysema	Anemia, osteoporosis, hyperkalemia	Disorders of abdomen, abdominal pain

Experiment



Dataset:

- 160 patients adjudicated with KEEPER
- 160 patients adjudicated with chart review

Performance Metrics

PERFORMANCE METRICS

1. Time to review
2. Agreement:
   - Agreement with the gold standard
   - Agreement of manual chart review and KEEPER
   - Agreement among reviewers

Results: time to review

Measured as time to review 20 patients

Manual chart review - 67 minutes (SD = 43)

KEEPER review - 30 minutes (SD = 14, p-value 0.04)

Results: agreement

Hereon, we will focus on pairwise agreement = % of cases for which reviewers have same response for adjudication (both ‘yes’ or both ‘no’)

personId (de-identified)	Reviewer 1	Reviewer 2
1	yes	yes
2	yes	yes
3	yes	yes
4	yes	yes
5	yes	yes
6	yes	yes
7	yes	yes
8	no	yes
9	no	yes
10	no	yes
11	no	yes
12	no	yes
13	no	yes
14	no	yes
15	no	yes
16	no	no
17	no	no
18	no	no
19	no	no
20	no	no

(7 ’both yes’ + 5 ‘both no’) /20 = 60%

**Paper includes kappa statistics*

Results: agreement with the gold standard

Measured as agreement between gold standard (the a priori iterative adjudication by two clinicians) and reviewers' adjudication

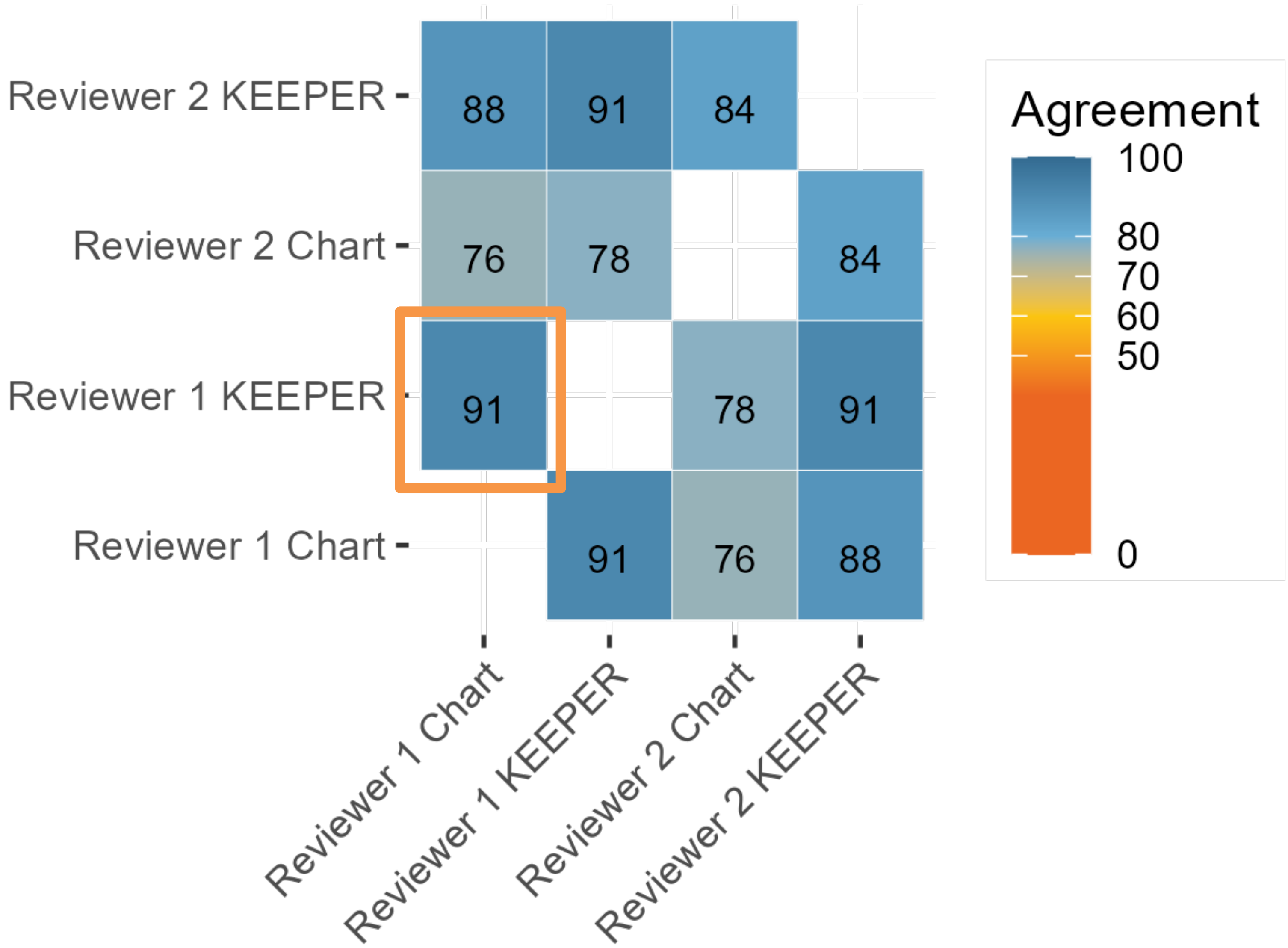
Manual chart review - 86.9% of patients classified similarly to the gold standard

KEEPER review - 88.1% of patients classified similarly to the gold standard

**varied across conditions but KEEPER accuracy always>80%*

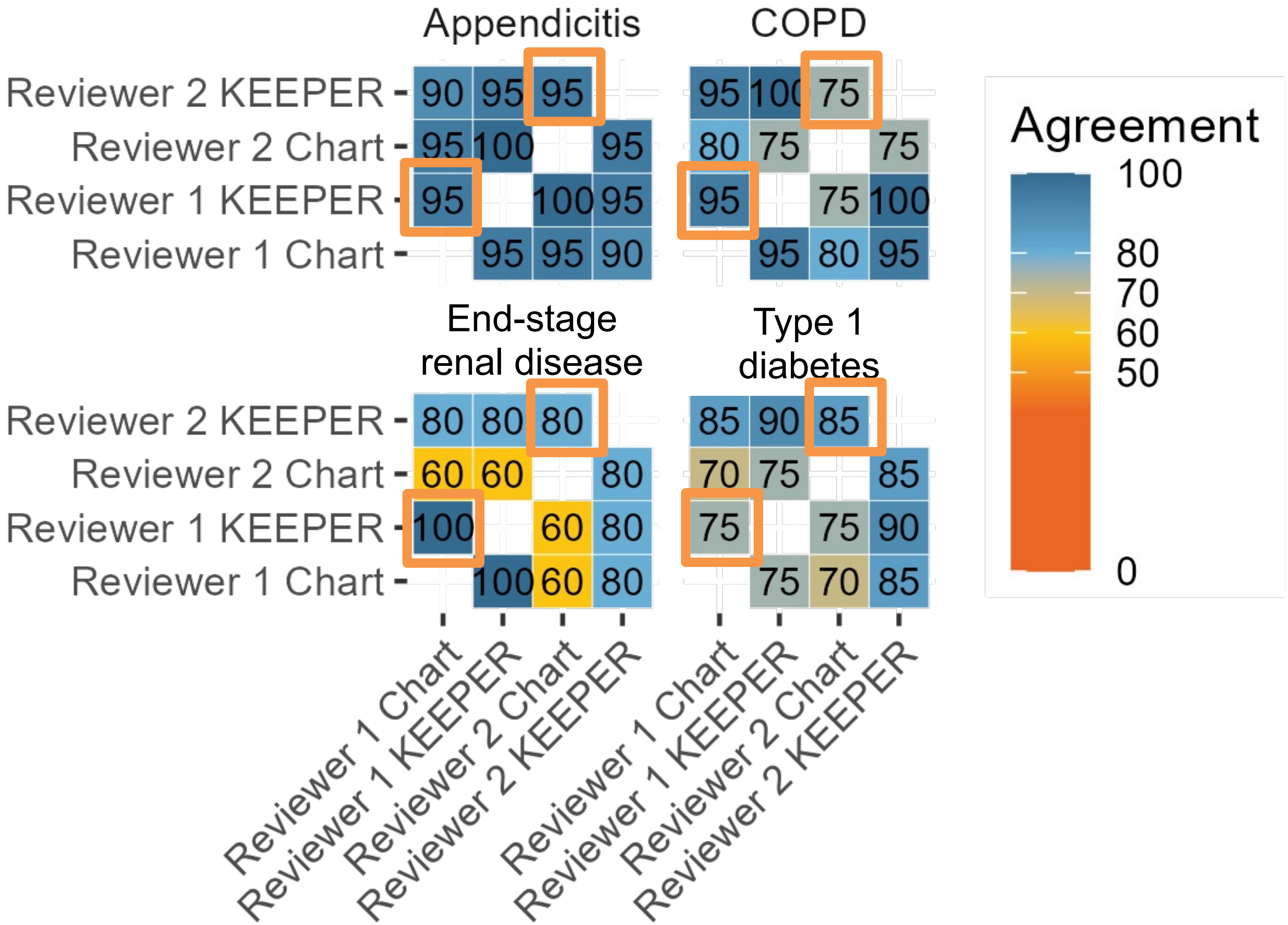
Results: agreement between chart and KEEPER

KEEPER adjudication agreed with manual chart review in 84-91% of the cases



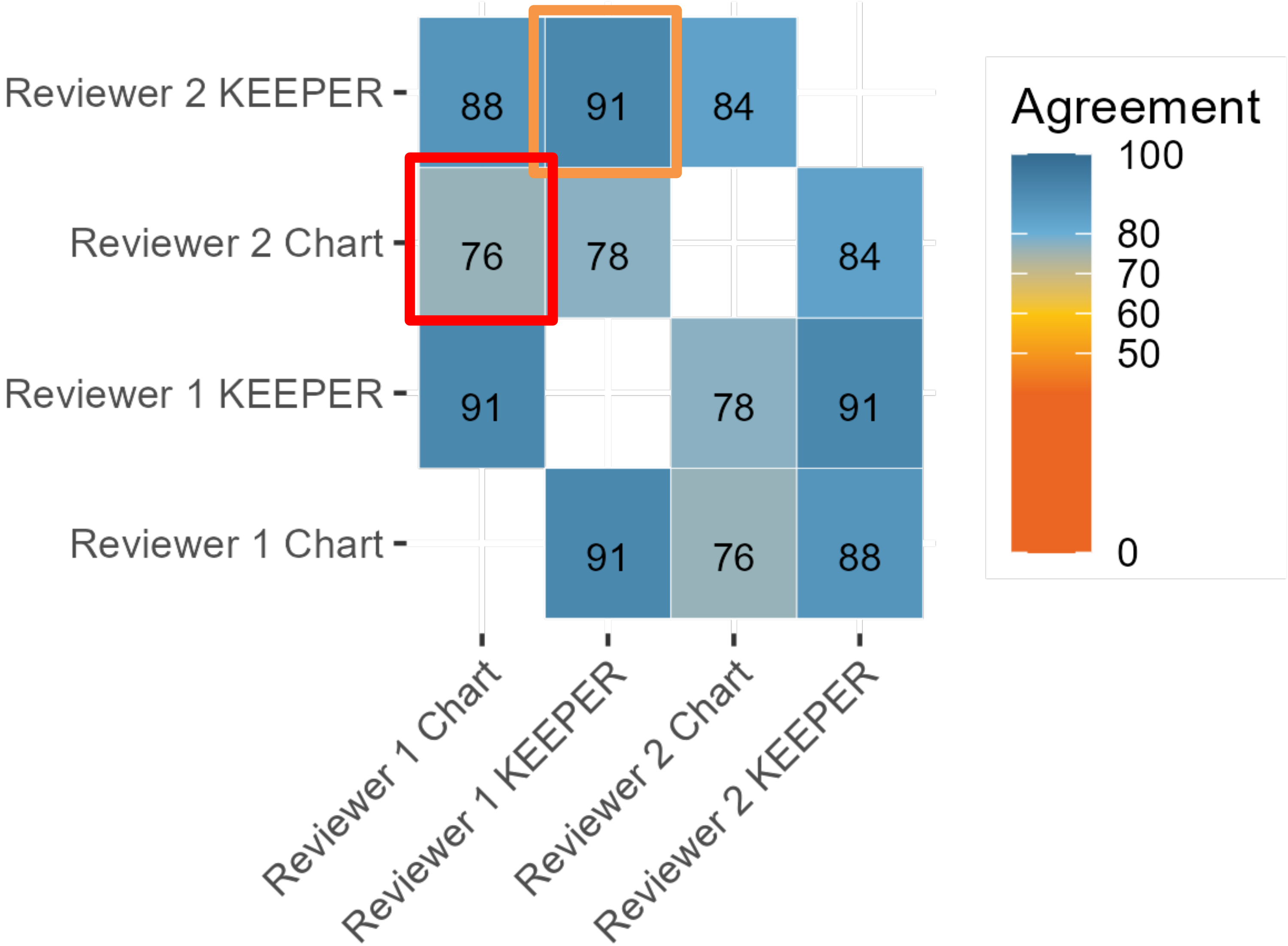
Results: agreement between chart and KEEPER by condition

Agreement between KEEPER and charts was consistently high across diseases (75-100%)



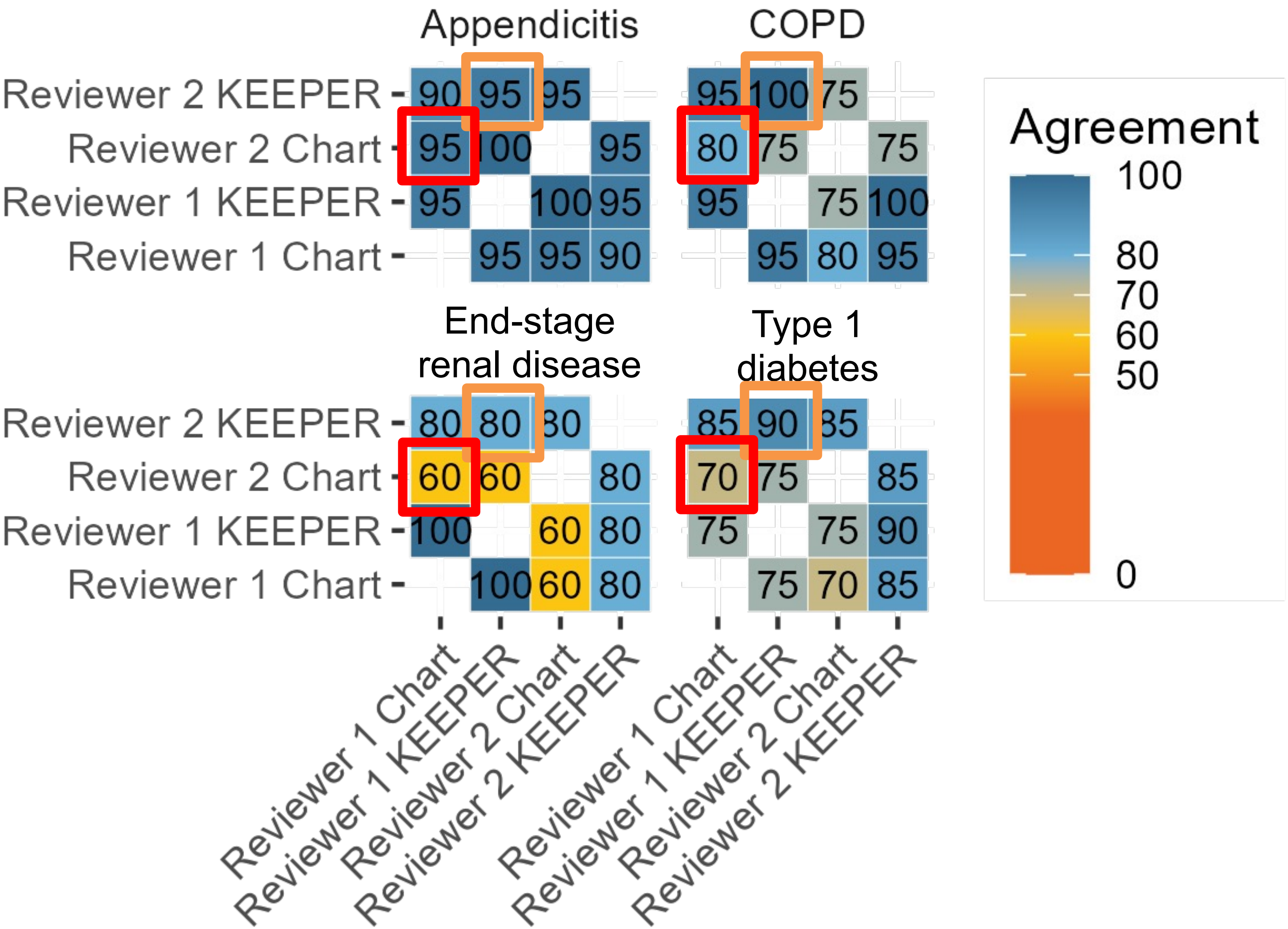
Results: agreement among reviewers

KEEPER adjudication had a significantly higher agreement between reviewers compared to agreement in manual chart review



Results: agreement among reviewers by condition

Heterogeneity of agreement between reviewers across conditions but KEEPER consistently better than chart review



Results: analysis of discrepancies in chart vs KEEPER adjudication

Source of discrepancy	Example
Information interpretation	Chart had a narrative about obstruction caused by cancer (exclusion for COPD), which was not available in KEEPER. Narrative was not supported by objective data.
High chart volume	KEEPER presented colon cancer diagnosis as a relevant alternative diagnosis for acute appendicitis. Finding the diagnosis in chart required extensive exploration.
Missing data in KEEPER	Indicators of specialty and location of visit were missing in KEEPER, which did not allow study reviewers to meaningfully assess discrepancies between specialty diagnoses and GP diagnoses for DM type I.

Experiment 2

KEEPER PROFILES

Created KEEPER profiles for 4 conditions (T1DM, COPD, ESRD, appendicitis), 100 patients total

EXPERIMENT (VK, OZ, SIS, PBR, AO)

5 reviewers

25 patients

X 4 diseases

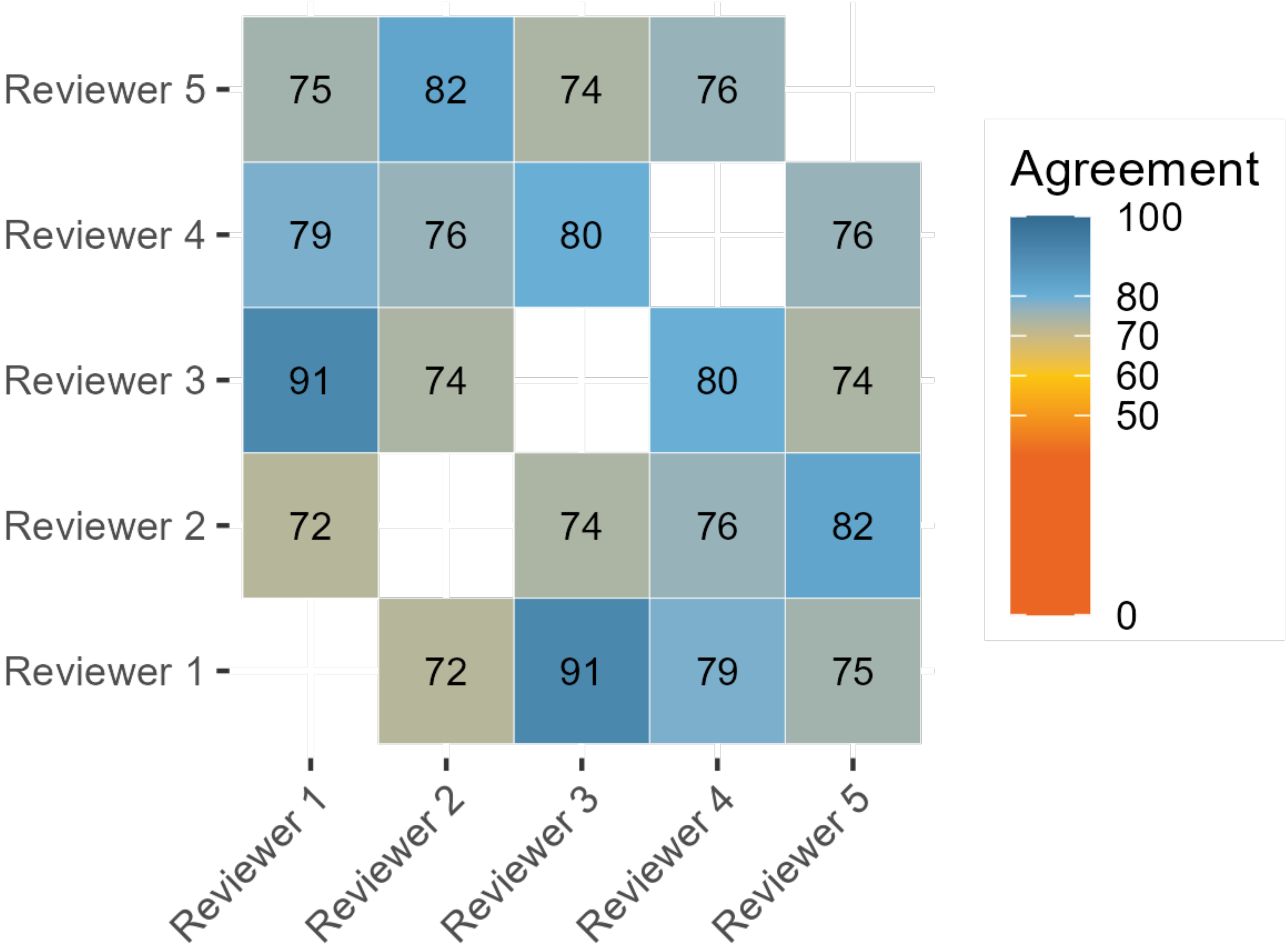
KEEPER

METRICS

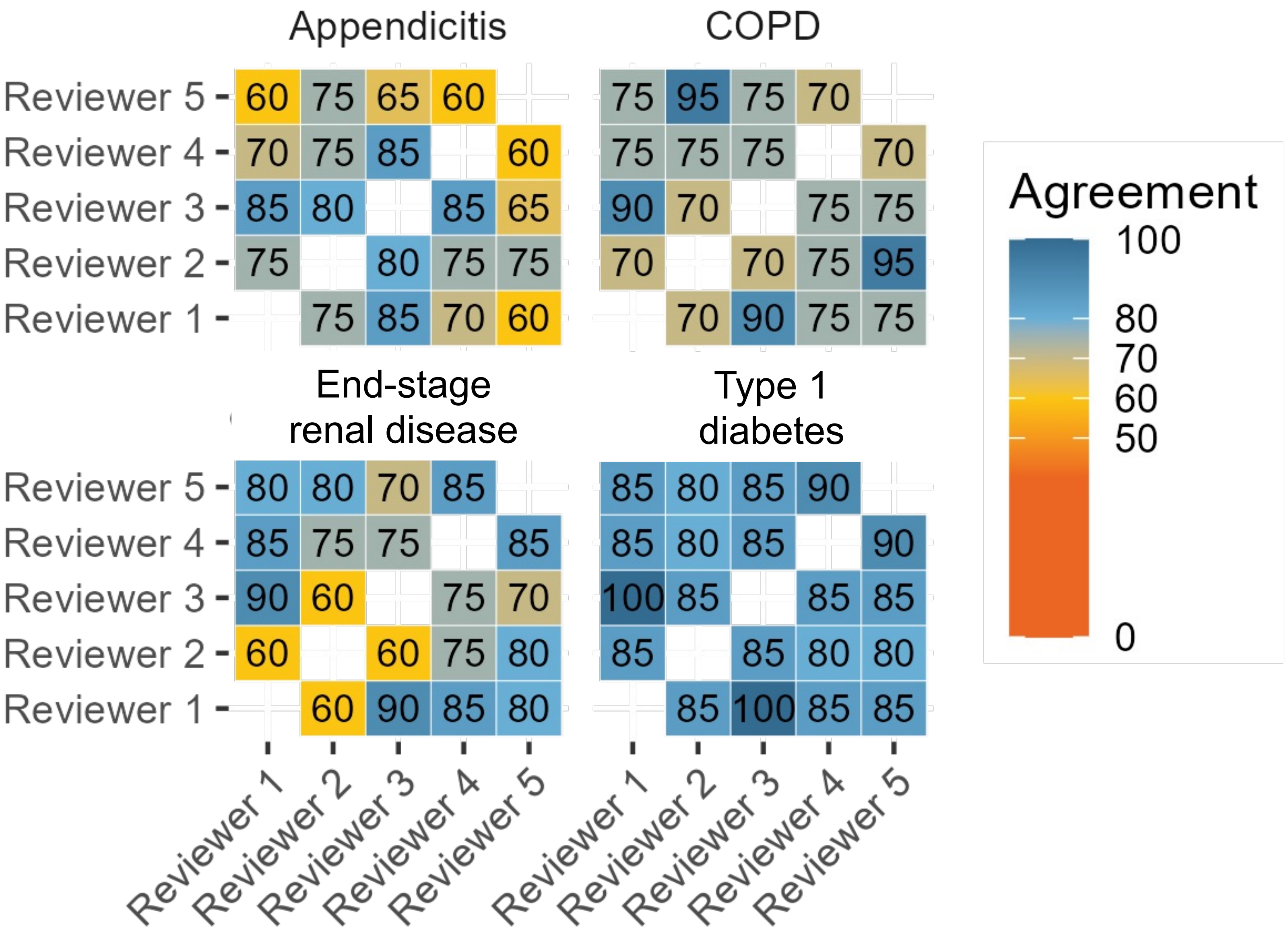
Agreement among reviewers

Optum ClinFormatics (US claims)

Experiment 2 results: agreement among reviewers



Experiment 2 results: agreement among reviewers by condition



Understanding agreement: reviewer adjudication of patient with suspected ESRD in Optum

personId (de-identified)	Reviewer 1	Reviewer 2	Reviewer 3	Reviewer 4	Reviewer 5
1	yes	yes	yes	yes	yes
2	yes	yes	yes	yes	yes
3	yes	yes	yes	yes	yes
4	yes	yes	yes	yes	yes
5	yes	yes	yes	yes	yes
6	yes	yes	yes	yes	yes
7	yes	yes	no	yes	yes
8	no	yes	no	yes	yes
9	no	yes	no	yes	yes
10	no	yes	yes	no	no
11	no	yes	no	yes	no
12	no	yes	no	no	yes
13	no	yes	no	no	yes
14	no	yes	no	no	no
15	no	yes	no	no	no
16	no	no	no	no	no
17	no	no	no	no	no
18	no	no	no	no	no
19	no	no	no	no	no
20	no	no	no	no	no

Unanimous consensus 'positive case'

Disagreement between reviewers

Unanimous consensus 'negative non-case'

Let’s review your PollEverywhere results

Cases

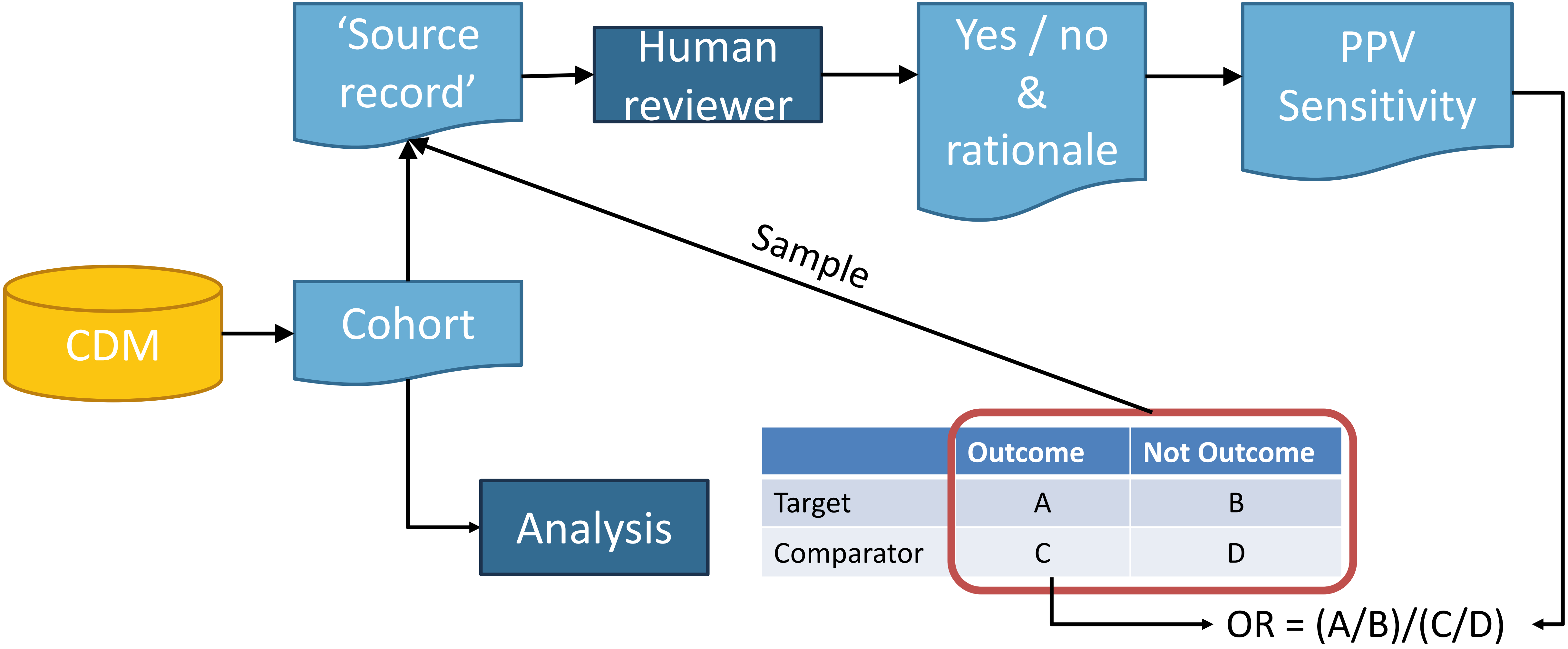
personId (de-identified)	Reviewer 1	Reviewer 2	Reviewer 3	Reviewer 4	Reviewer 5
1	yes	yes	yes	yes	yes
2	yes	yes	yes	yes	yes
3	yes	yes	yes	yes	yes
4	yes	yes	yes	yes	yes
5	yes	yes	yes	yes	yes
6	yes	yes	yes	yes	yes
7	yes	yes	no	yes	yes
8	no	yes	no	yes	yes
9	no	yes	no	yes	yes
10	no	yes	yes	no	no
11	no	yes	no	yes	no
12	no	yes	no	no	yes
13	no	yes	no	no	yes
14	no	yes	no	no	no
15	no	yes	no	no	no
16	no	no	no	no	no
17	no	no	no	no	no
18	no	no	no	no	no
19	no	no	no	no	no
20	no	no	no	no	no

Case #1

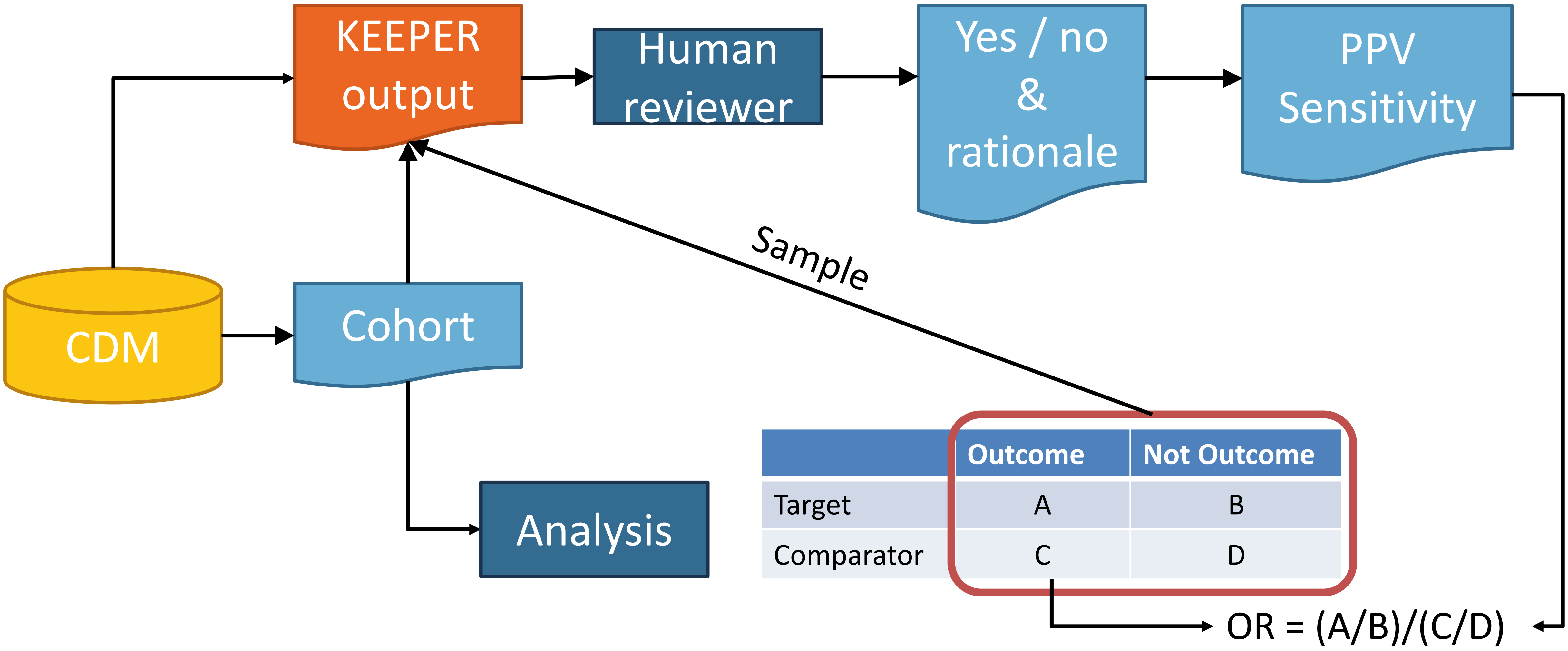
Case #3

Case #2

Case validation workflow



Case validation workflow



Further improvements to the scalability of case validation

Martijn Schuemie

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Generalizability of KEEPER performance

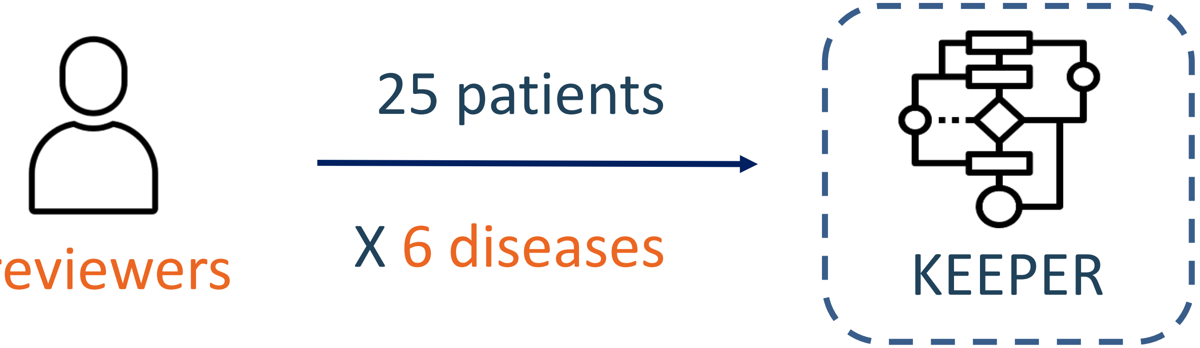
- Previous two experiments used four diseases (ESRD, T1DM, COPD, Appendicitis) with
  - Clear expectations of health utilization
  - Clear markers to use to classify disease status
  - Cases not sampled from a single cohort
- Let's design a new experiment!

Experiment 3

KEEPER PROFILES

Created KEEPER profiles for 6 conditions representing a range of complexity (Acute bronchitis, hyperlipidemia, hypoparathyroidism, osteoporosis, rheumatoid arthritis, viral hepatitis type A), 150 patients total

EXPERIMENT



METRICS

Agreement among reviewers

Estimate positive predictive value

DATABASE

Optum ClinFormatics (US claims)

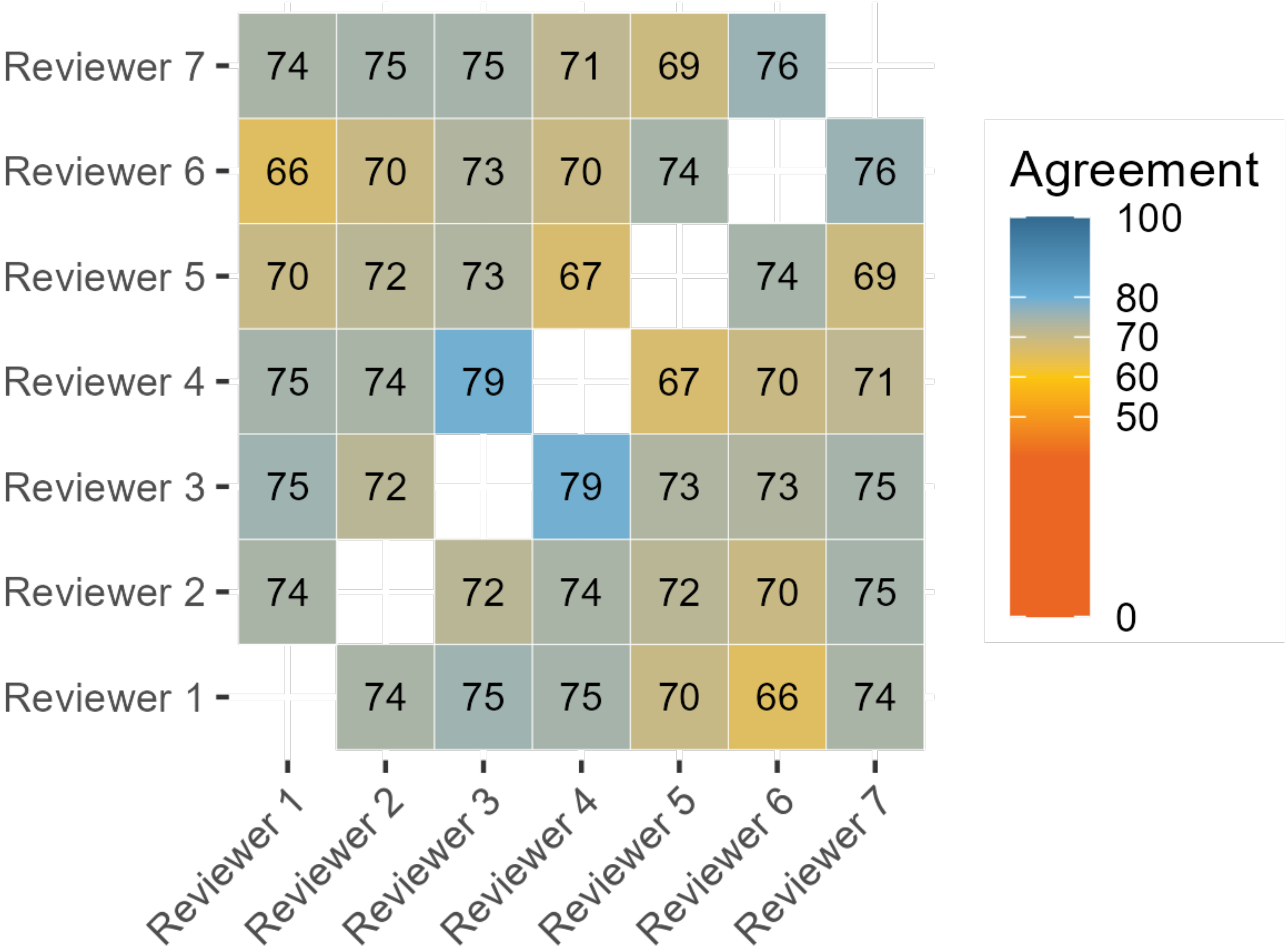
Experiment 3

All diseases have clinical descriptions and phenotype algorithms

Disease	Clinical description	Phenotype algorithm
Acute bronchitis	https://www.ncbi.nlm.nih.gov/books/NBK448067/	https://atlas-phenotype.ohdsi.org/#/cohortdefinition/1218
Hyperlipidemia	https://www.ncbi.nlm.nih.gov/books/NBK559182/	https://atlas-phenotype.ohdsi.org/#/cohortdefinition/1219
Hypoparathyroidism	https://www.ncbi.nlm.nih.gov/books/NBK441899/	https://atlas-phenotype.ohdsi.org/#/cohortdefinition/1220
Osteoporosis	https://www.ncbi.nlm.nih.gov/books/NBK441901/	https://atlas-phenotype.ohdsi.org/#/cohortdefinition/1221
Rheumatoid arthritis	https://www.ncbi.nlm.nih.gov/books/NBK441999/	https://atlas-phenotype.ohdsi.org/#/cohortdefinition/858
Viral hepatitis type A	https://www.ncbi.nlm.nih.gov/books/NBK459290/	https://atlas-phenotype.ohdsi.org/#/cohortdefinition/1222

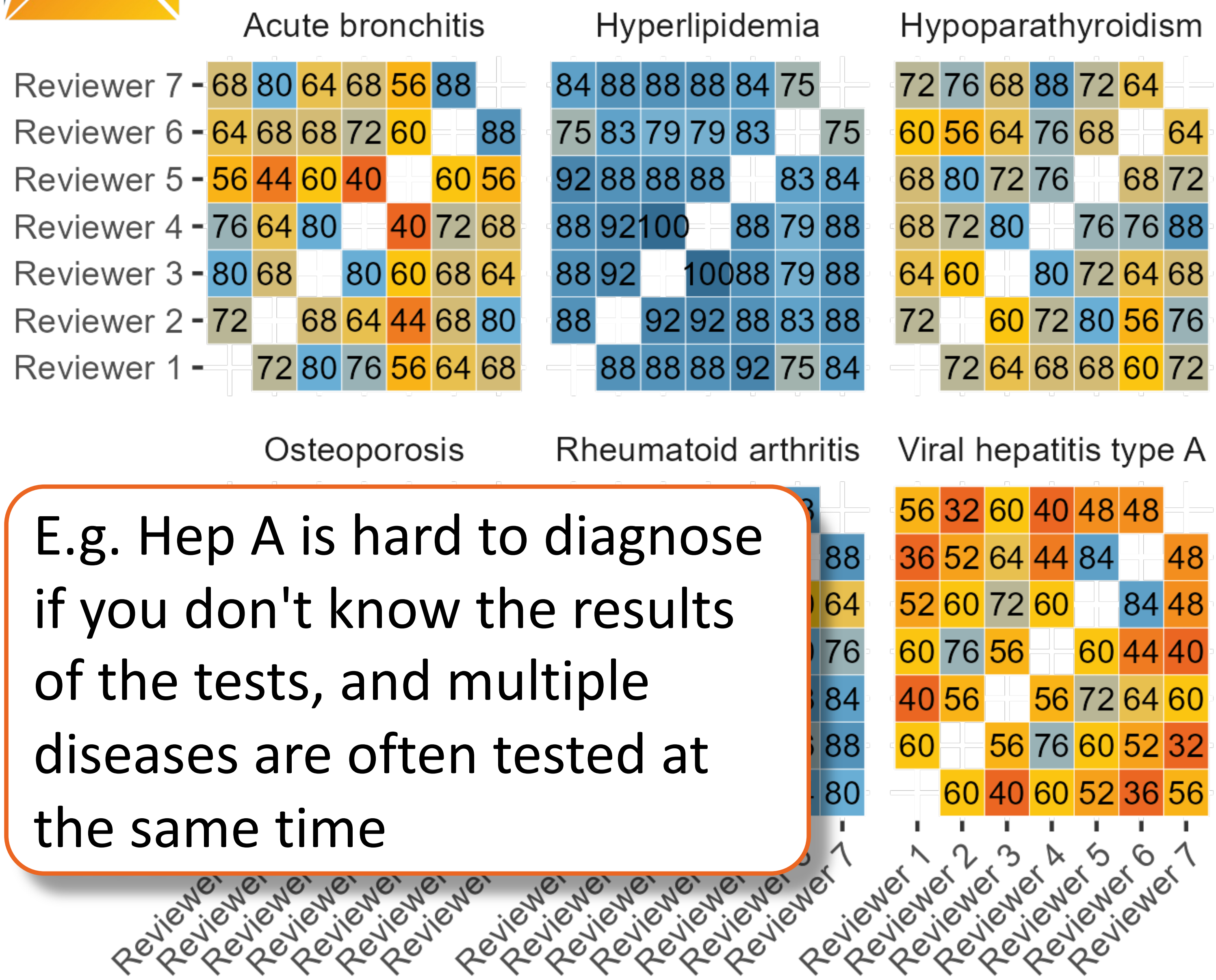
	Acute bronchitis	Hyper-lipidemia	Hypopara-thryroidism	Osteo-porosis	Rheumatoid arthritis	Viral hepatitis type A
Disease history and progression (including complications)	pneumonia, acute respiratory distress syndrome, respiratory failure	coronary artery disease, peripheral artery disease, cerebrovascular accidents, aneurysms, type II diabetes, high blood pressure	Acute Hypocalcemia, Chronic Hypocalcemia	chronic pain and fractures	Anemia of chronic disease, Felty syndrome, Coronary artery disease , lymphoma, Osteopenia, osteoporosis, venous thromboembolic disease	Prolonged cholestasis, Acute renal failure, Autoimmune hepatitis
Symptoms	cough, malaise, difficulty breathing, and wheezing		remote thyroid or other types of head and neck surgery, myalgias, muscle spasms, and in extreme cases, tetany, hypocalcemia, hyperphosphatemia, and increased neuromuscular irritability	loss of height and kyphosis	joint pain and swelling, morning stiffness, Interstitial lung disease, Sjogren syndrome with dry eyes and also dry mouth	nausea, vomiting, right upper quadrant abdominal discomfort, malaise, anorexia, myalgia, fatigue, and fever; pancreatitis, rash, acute kidney injury with interstitial nephritis or glomerular nephritis, pneumonitis, pericarditis, hemolysis, and acute cholecystitis
Diagnostic procedures	Oxygen saturation, pulse rate, temperature, and respiratory rate. Chest x-ray (CXR), A complete blood count and chemistry, Spirometry		Electrocardiogram	dual-energy X-ray absorptiometry scans	Magnetic resonance imaging (MRI) and ultrasonography	serologic testing to detect HAV-specific immunoglobulin (IgM) antibodies , reverse transcriptase-polymerase chain reaction to detect the viral RNA
Measurements		fasting lipid profile	calcium, albumin, serus calcium, parathyroid hormone level, phosphorus, Blood urea nitrogen (BUN) and creatinine, Alkaline phosphatase, 25-hydroxyvitamin D, Urine calcium and creatinine,		RF and ACPA antibodies, Anti-carbamylated protein antibodies, CCP , erythrocyte sedimentation rate (ESR) and levels of C-reactive protein (CRP)	elevated levels of serum alanine aminotransferase, aspartate aminotransferase, bilirubin, alkaline phosphatase, and lambda-glutamyl transpeptidase
Treatments: Drugs, Procedures	dextromethorphan and codeine , Beta-agonists , Analgesic and antipyretic agents	statin , ezetimibe	calcium, vitamine D, calcitriol	risedronate, alendronate, zoledronic acid, or denosumab, Bazedoxifene, teriparatide, raloxifene	DMARDs, NSAIDs, anti-TNF, IL6, CTLA4-Ig, antiCD20, JAK, corticosteroids	liver transplantation, immunoglobulin
Differential diagnoses	Asthma, Acute/chronic sinusitis, Bronchiolitis, COPD, Gastroesophageal reflux disease (GERD), Viral pharyngitis, Heart failure, Pulmonary embolism	familial hypercholesterolemia, familial combined hyperlipidemia, dysbetalipoproteinemia, familial defective apo B-100, and PCSK9 gain of function mutations, obstructive liver disease or biliary obstruction, hypothyroidism, nephrotic syndrome, chronic renal insufficiency, anorexia, obesity, metabolic syndrome, and diabetes	Hypomagnesemia, Postoperative complications of thyroidectomy and other types of head and neck surgery - may be transient or permanent:. Abnormal development of parathyroid tissue, for example, DiGeorge Syndrome, Activating mutations of the calcium-sensing receptor - autosomal dominant hypocalcemia, Activating antibodies of the calcium-sensing receptor, Autoimmune destruction of parathyroid tissue, for example, polyglandular autoimmune syndrome, Type 1, Infiltration of parathyroid tissue, for example, granulomatous disease, hemochromatosis, metastatic disease, Radiation injury, Parathyroid hormone resistance, pseudohypoparathyroidism	Homocystinuria, Hyperparathyroidism, Imaging in osteomalacia and renal osteodystrophy, Mastocytosis, Multiple myeloma, Paget disease, Scurvy , Sickle cell anemia	Osteoarthritis, Psoriatic arthritis, Systemic lupus erythematosus, Sjogren syndrome, Polymyalgia rheumatic, Chronic gouty arthritis	Alcoholic hepatitis, Other Viral hepatitis (B, C, D, E), Autoimmune hepatitis
Comorbidities and risk factors		history of cardiovascular disease, hyperlipidemia, and/or familial hypercholesterolemia; their diet and exercise habits; tobacco, alcohol, or drug use; the presence of coronary artery disease; risk factors or history of CAD;		smoking history and chronic alcohol		

Experiment 3 agreement



- Overall agreement was consistent across all reviewers (66-79%)
- Results are in line but a bit lower than Anna showed in Experiment 2 (72%-91%)

Experiment 3 agreement, by condition



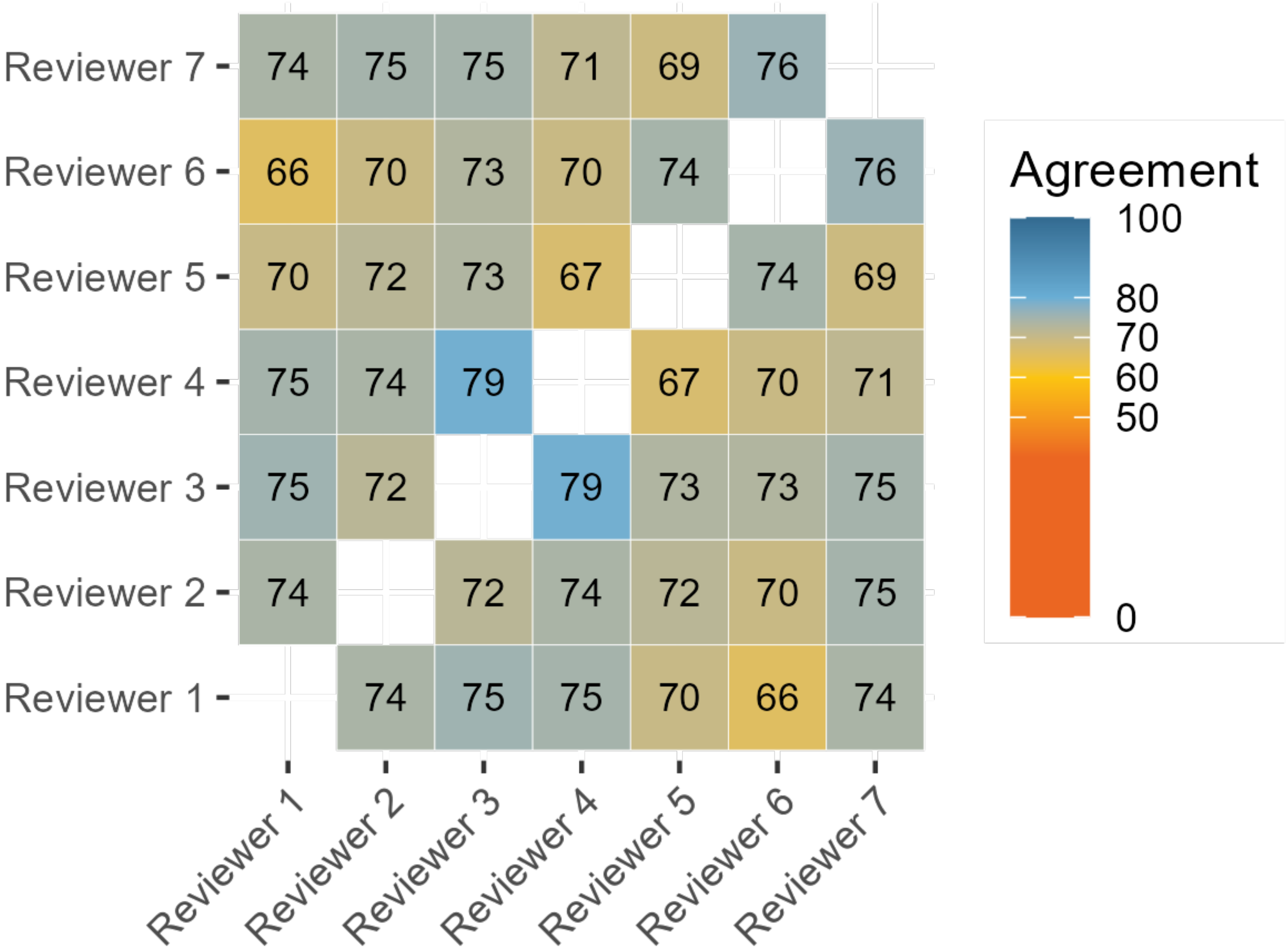
E.g. Hep A is hard to diagnose if you don't know the results of the tests, and multiple diseases are often tested at the same time

- Heterogeneity in agreement across diseases
- Hyperlipidemia and RA had strong agreement across all reviewers
- Hep A and bronchitis had more dis-agreement across all reviewers

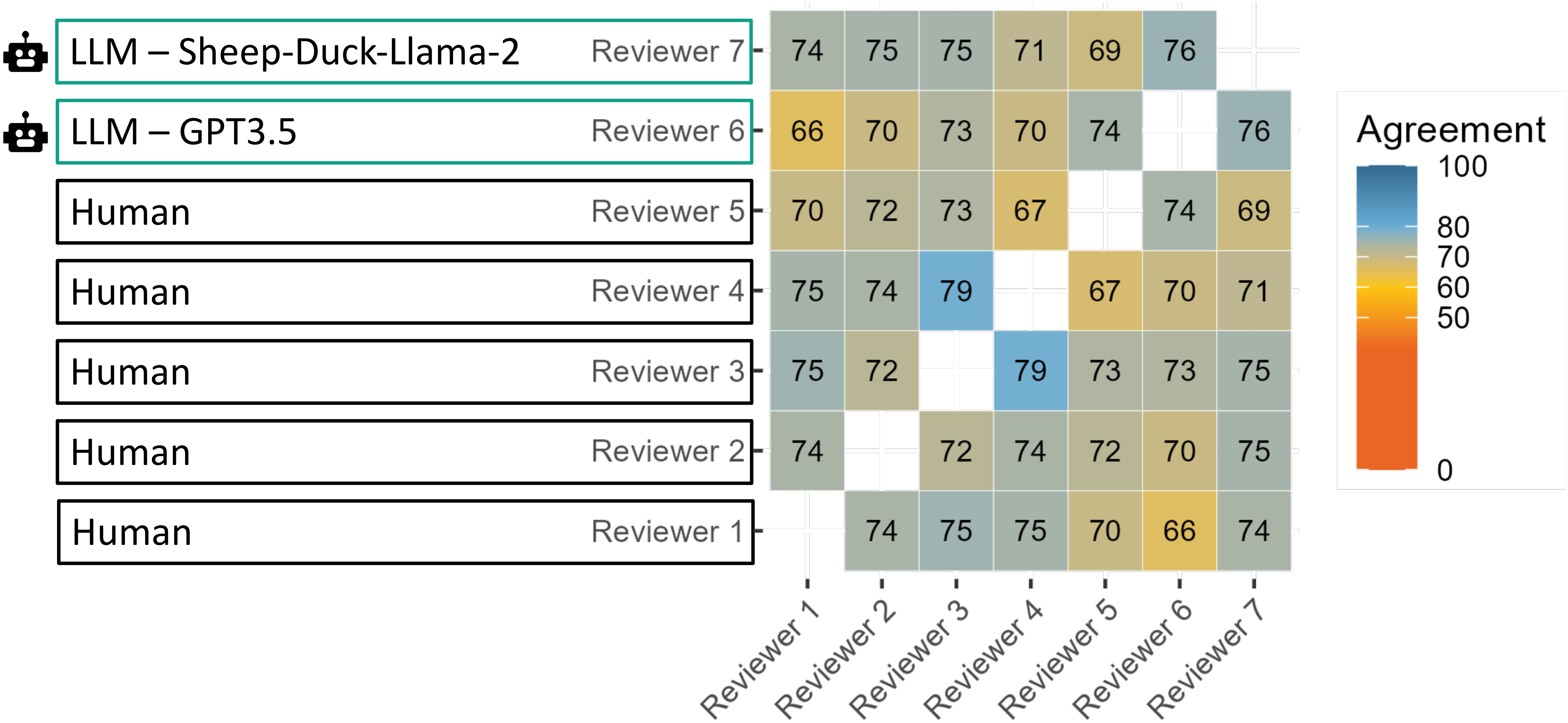
Which reviewer is not a human?

Vote now!

pollev.com/PatrickRyan800

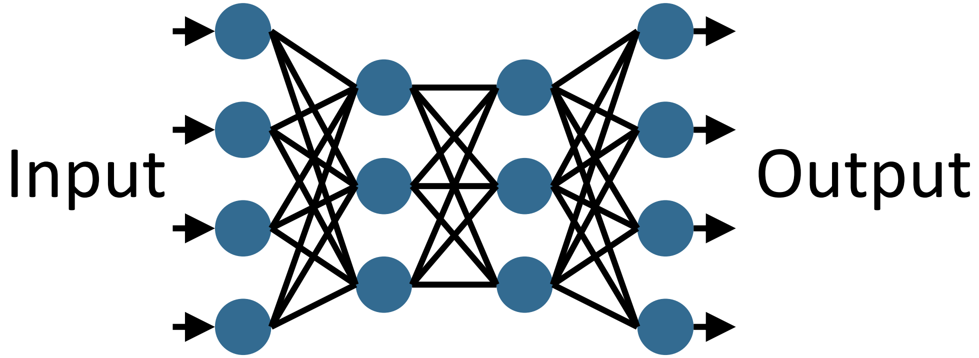


Which reviewer is not a human?

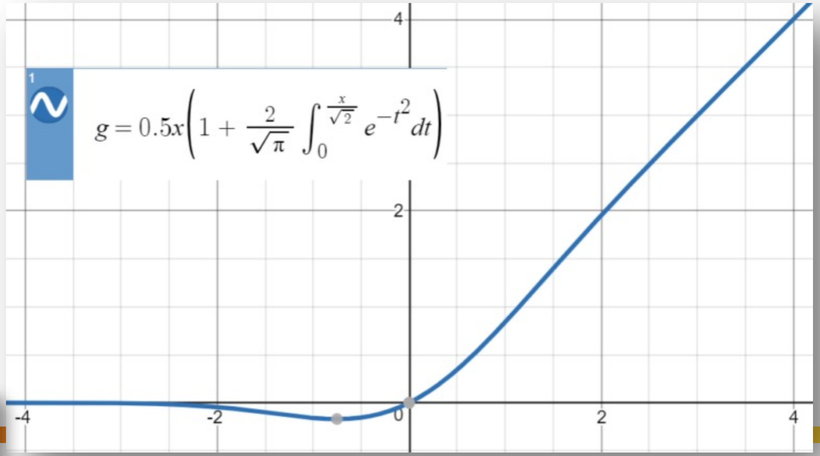


What is a large language model (LLM)?

- A large-language model (LLM) is a neural network / deep learning model
- Consists of nodes and weighted edges



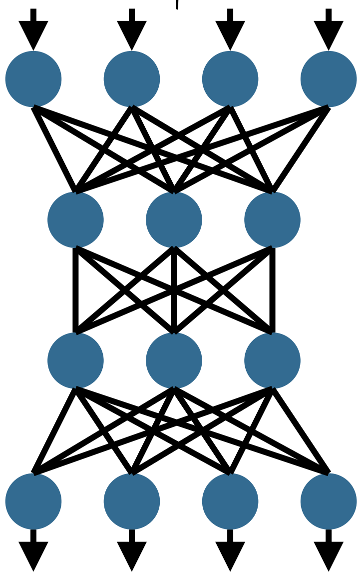
- Each node i in layer j computes its output as: $a_{ij} = g(\sum_k w_{ijk} a_{(j-1)k})$
- Supervised learning:
  - Compute output given input
  - Compare computed output to expected output
  - Adjust weight in small steps to improve output using back-propagation



Note: This is a simplification

Pretraining: predict the next word in a massive corpus

… said the Cat. “I don’t much care where—” said ?



Predicting the next word requires:

- Grammar: next word probably is a (proper) noun
- Semantics: only some things can talk
- Context: this is conversation between the Cheshire Cat and Alice

Most likely next word: Alice

Pre-training scale

- Model parameters:
  - GPT3.5: 175 billion
  - Llama-2: 70 billion
- Corpus size:
  - GPT3.5: 300 billion tokens (token ~ 0.75 word)
  - Llama-2: 2 trillion tokens
- Time to train
  - GPT3.5: (estimated) 355 GPU years
  - Llama-2: 376 GPU years

Fine-tuning world's most expensive auto-completion

- Pre-trained models can be used to predict the next word, and the next, and the next, generating text
- Can be further training to generate answers to questions (chat)
  - Supervised: Human-created training set
  - Reinforcement learning: human corrects output of LLM
- Requires far less training examples if pre-trained (still millions of tokens)
- Initial work shows model learned important concepts in medicine

JMIR MEDICAL EDUCATION Gilson et al

Original Paper

How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment

Aidan Gilson[1,2], BS; Conrad W Safranek[1], BS; Thomas Huang[2], BS; Vimig Socrates[1,3], MS; Ling Chi[1], BSE; Richard Andrew Taylor[1,2*], MD, MHS; David Chartash[1,4*], PhD

[1]Section for Biomedical Informatics and Data Science, Yale University School of Medicine, New Haven, CT, United States

Evaluated large language models

- Azure OpenAI GPT3.5 Turbo
  - Further finetuning of GPT3.5
  - Proprietary
  - Licensed by Johnson & Johnson
- Llama-2-70b-chat-hf
  - Open source
  - Installed on a private machine using HuggingFace library (~50 lines of code)
- Sheep-Duck-Llama-2-70b-v1.1
  - Further finetuning of Llama-2
  - Sheep-Duck-Llama-2 was at the top of the HF leaderboard 2 weeks ago
  - Installed on a private machine

All analyses run securely within organizational firewall

Training set

KEEPER PROFILES

Created KEEPER profiles for 6 conditions (Acute bronchitis, hyperlipidemia, hypoparathyroidism, osteoporosis, rheumatoid arthritis, viral hepatitis type A), focus on hard cases. 358 patients total

EXPERIMENT

METRICS

Sensitivity, specificity, agreement of LLM using human reviewer as gold standard

DATABASE

Optum ClinFormatics (US claims)

Prompt engineering

KEEPER output as text:

Demographics and details about the visit: Female, 70 yo; Visit: Laboratory Visit

Diagnoses recorded on the day of the visit: Rheumatoid arthritis (Primary diagnosis);

Diagnoses recorded prior to the visit: None

Treatments recorded prior to the visit: None

Diagnostic procedures recorded proximal to the visit: Collection of venous blood (day -30, 0, 30)

Laboratory tests recorded proximal to the visit: None

Alternative diagnoses recorded proximal to the visit: None

Diagnoses recorded after the visit: Seropositive rheumatoid arthritis (day 90)

Treatments recorded during or after the visit: None

Perturbed patient data

System prompt: yes / no:

Act as a medical doctor reviewing a patient's healthcare data captured during routine clinical care, such as electronic health records and insurance claims.

Determine whether the patient had [DISEASE].

Use the following format:

Summary: (Only "yes" or "no")

Prompt	Sensitivity	Specificity	Agreement
Yes/no	99.0%	8.9%	64.9%

System prompt: + discuss evidence

Act as a medical doctor reviewing a patient's healthcare data captured during routine clinical care, such as electronic health records and insurance claims.

Determine whether the patient had [DISEASE].

Use the following format:

Evidence in favor of [DISEASE]:

Evidence against [DISEASE]:

Summary: (Only "yes" or "no")

Prompt	Sensitivity	Specificity	Agreement
Yes/no	99.0%	8.9%	64.9%
+ discuss evidence	90.7%	29.0%	67.4%

System prompt: + write narrative

…

Write a medical narrative that fits the recorded health data followed by a determination of whether the patient had [DISEASE]**.**

Use the following format:

Clinical narrative:

…

Observation: LLM always believed diagnosis code was accurate

Prompt	Sensitivity	Specificity	Agreement
Yes/no	99.0%	8.9%	64.9%
+ discuss evidence	90.7%	29.0%	67.4%
+ write narrative	97.1%	21.0%	68.3%

System prompt: + diagnosis insufficient reminder

Remember that recording a diagnosis for a disease could occur either because the patient had the disease or as justification for performing a diagnostic procedure to determine whether the patient has the disease. **y**
diagnostic procedures may therefore be **n**
once. Lack of additional evidence of [I **c**
procedures probably means that the pa **y**
have [DISEASE]**. However, it is unlikely t** **n**
abundance of diagnoses will mean the p

Observation: LLM didn't know how to deal with uncertainty. Would respond 'yes' even though another diagnosis was more likely, or 'no' if there was any (unreasonable) doubt.

Prompt	**Sensitivity**	**Specificity**	**Agreement**
Yes/no	99.0%	8.9%	64.9%
+ discuss evidence	90.7%	29.0%	67.4%
+ write narrative	97.1%	21.0%	68.3%
+ diagnosis insufficient reminder	95.6%	31.5%	71.3%

System prompt: + uncertainty instructions

In your final summary, indicate "yes" if the most probable scenario is that the patient had [DISEASE].

Indicate "no" if it is not the most probable scenario, for example when it is more likely that the patient was tested for the disease but the diagnosis was not confirmed. Also indicate "no" when there is insufficient information to say anything about the relative probability of scenarios.

Prompt	Sensitivity	Specificity	Agreement
Yes/no	99.0%	8.9%	64.9%
+ discuss evidence	90.7%	29.0%	67.4%
+ write narrative	97.1%	21.0%	68.3%
+ diagnosis insufficient reminder	95.6%	31.5%	71.3%
+ uncertainty instructions	82.4%	58.1%	73.2%

System prompt: + provide examples

Added two examples of input and output to the system prompt (few-shot prompt)

Personal preference: picked solution with highest agreement, so not using examples

Prompt	Sensitivity	Specificity	Agreement
Yes/no	99.0%	8.9%	64.9%
+ discuss evidence	90.7%	29.0%	67.4%
+ write narrative	97.1%	21.0%	68.3%
+ diagnosis insufficient reminder	95.6%	31.5%	71.3%
+ uncertainty instructions	82.4%	58.1%	73.2%
+ provide examples	66.7%	73.4%	69.2%

Performance of different LLMs

- Selected optimal prompt using GPT 3.5 for convenience.
- Evaluated optimal prompt on original Llama-2, which did not produce great results.
- Other people have fine-tuned Llama-2. Top of the Huggingface leaderboard two weeks ago was Sheep-Duck-Llama2, by Riiid (under same license).

Large language model	Sensitivity	Specificity	Agreement
GPT 3.5 Turbo	82.4%	58.1%	73.2%
Llama-2-70b-chat-hf	99.0%	12.9%	66.4%
Sheep-Duck-Llama-2-70b-v1.1	90.2%	62.1%	79.6%

Multiple good LLMs are available, but you shouldn't assume they are good until tested

Example prompt

System prompt

Act as a medical doctor reviewing a patient's healthcare data captured during routine clinical care, such as electronic health records and insurance claims.
Write a medical narrative that fits the recorded health data followed by a determination of whether the patient had end stage renal disease.

Remember that recording a diagnosis for a disease could occur either because the patient had the disease or as justification for performing a diagnostic procedure to determine whether the patient has the disease. A diagnosis by itself or accompanied with only diagnostic procedures may therefore be insufficient evidence, even if recorded more than once. Lack of additional evidence of end stage renal disease other than the diagnosis and diagnostic procedures probably means that the patient was only being tested, and does not actually have end stage renal disease. However, it unlikely that a patient will be tested many times over, so an abundance of diagnoses will mean the patient has the disease.

In your final summary, indicate "yes" if the most probable scenario is that the patient had end stage renal disease.
Indicate "no" if it is not the most probable scenario, for example when it is more likely that the patient was tested for the disease but the diagnosis was not confirmed.
Also indicate "no" when there is insufficient information to say anything about the relative probability of scenarios.

Use the following format:

Clinical narrative:

Evidence in favor of end stage renal disease:

Evidence against end stage renal disease:

Summary: (Only "yes" or "no")

Example prompt

Prompt

Demographics and details about the visit: Male, 50 yo; Visit: Pharmacy visit followed by Outpatient Visit

Diagnoses recorded on the day of the visit: Chronic kidney disease due to type 2 diabetes mellitus (Primary admission diagnosis); Chronic kidney disease due to type 2 diabetes mellitus (Primary diagnosis); Chronic kidney disease stage 5 (Admission diagnosis); Complication due to diabetes mellitus (Admission diagnosis); Essential hypertension (Admission diagnosis); Essential hypertension (Secondary diagnosis); Hyperlipidemia (Admission diagnosis); Proteinuria (Admission diagnosis); Renal disorder due to type 2 diabetes mellitus (Admission diagnosis); Renal disorder due to type 2 diabetes mellitus (Secondary diagnosis); Type 2 diabetes mellitus (Admission diagnosis); Type 2 diabetes mellitus (Primary admission diagnosis); Type 2 diabetes mellitus (Primary diagnosis); Vitamin D deficiency (Admission diagnosis); Vitamin D deficiency (Secondary diagnosis);

Diagnoses recorded prior to the visit: Anemia (day -900); Anemia in chronic kidney disease (day -810, -10); Anemia of chronic disease (day -890, -800); Chronic kidney disease (day -860, -820, -10); Chronic kidney disease due to hypertension (day -890, -800, -10); Chronic kidney disease due to type 2 diabetes mellitus (day -890, -10); Chronic kidney disease stage 3 (day -890, -820, -10); Vitamin D deficiency (day -820, -10)

Treatments recorded prior to the visit: None

Diagnostic procedures recorded proximal to the visit: None

Laboratory tests recorded proximal to the visit: None

Alternative diagnoses recorded proximal to the visit: Acute renal failure syndrome (day -10, 30)

Diagnoses recorded after the visit: Anemia (day 210, 290); Anemia in chronic kidney disease (day 170, 310, 1050, 1140, 1230, 1320, 1470); Anemia of chronic disease (day 30); Chronic kidney disease due to type 2 diabetes mellitus (day 30, 90, 250, 280, 920, 1080); Chronic kidney disease stage 4 (day 250); Chronic kidney disease stage 5 (day 90, 170, 210, 270); End-stage renal disease (day 290, 900, 1660); Hyperkalemia (day 30); Hyperlipidemia (day 90, 310); Hyperparathyroidism due to renal insufficiency (day 30, 170, 250, 280); Iron deficiency anemia (day 930); Vitamin D deficiency (day 170, 250, 280)

Treatments recorded during or after the visit: calcitriol (day 290, for 50 days);

Perturbed patient data

Summary: (Only "yes" or "no")

Example prompt

Response by Sheep-Duck-Llama-2

Clinical narrative: A 50-year-old male with a history of type 2 diabetes mellitus, hypertension, and chronic kidney disease due to type 2 diabetes mellitus presented for a pharmacy visit followed by an outpatient visit. During the visit, he was diagnosed with chronic kidney disease stage 5, end-stage renal disease, and other related complications. Prior to the visit, he had been diagnosed with chronic kidney disease, anemia, and vitamin D deficiency. After the visit, he was diagnosed with end-stage renal disease, anemia, hyperlipidemia, and hyperparathyroidism due to renal insufficiency. He was prescribed calcitriol for 54 days.

Evidence in favor of end stage renal disease:

- Diagnosis of chronic kidney disease stage 5 and end-stage renal disease during the visit
- Diagnosis of end-stage renal disease after the visit

Evidence against end stage renal disease:

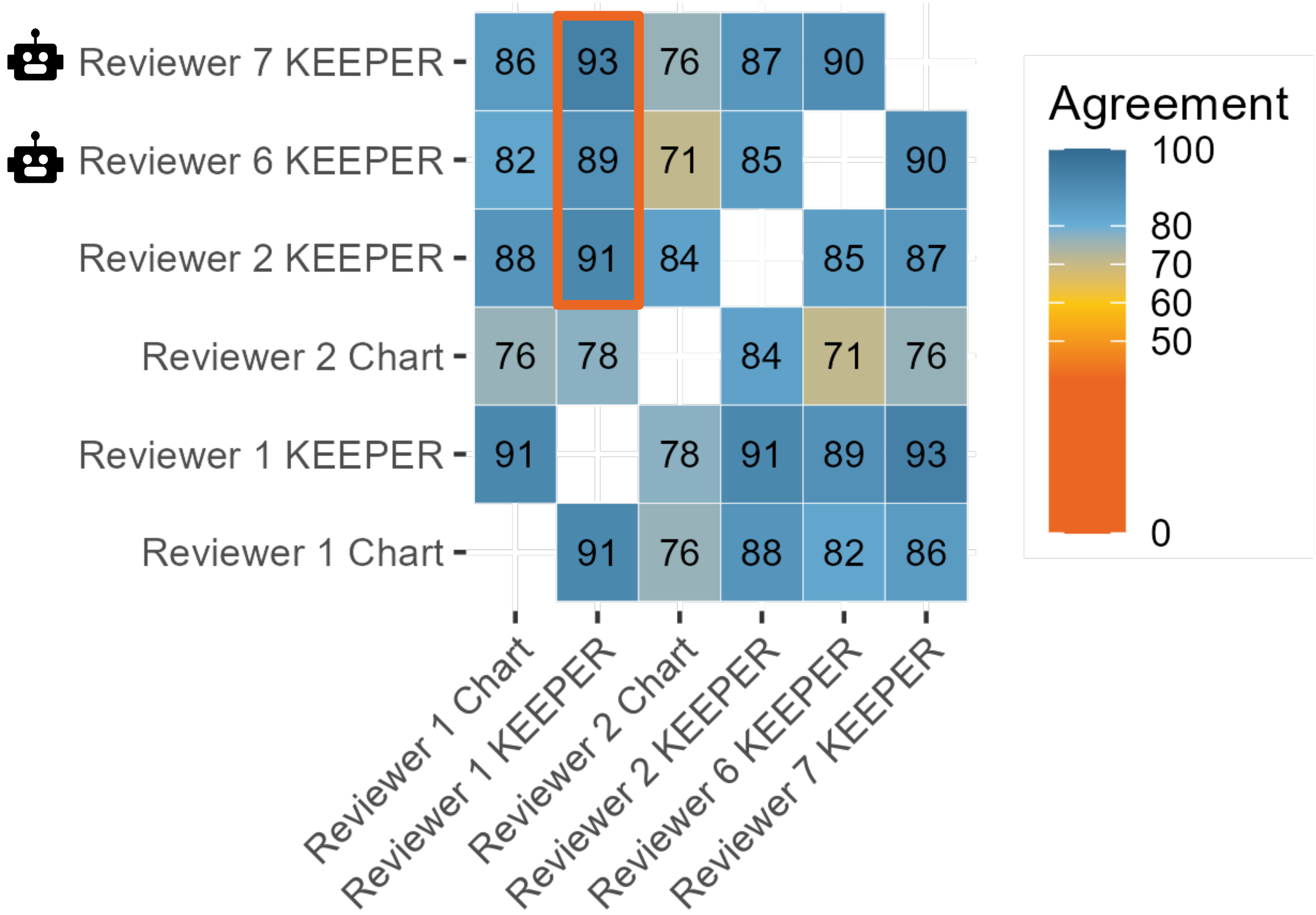
- No evidence against end stage renal disease

Summary: yes

Note: this is the text you reviewed earlier!

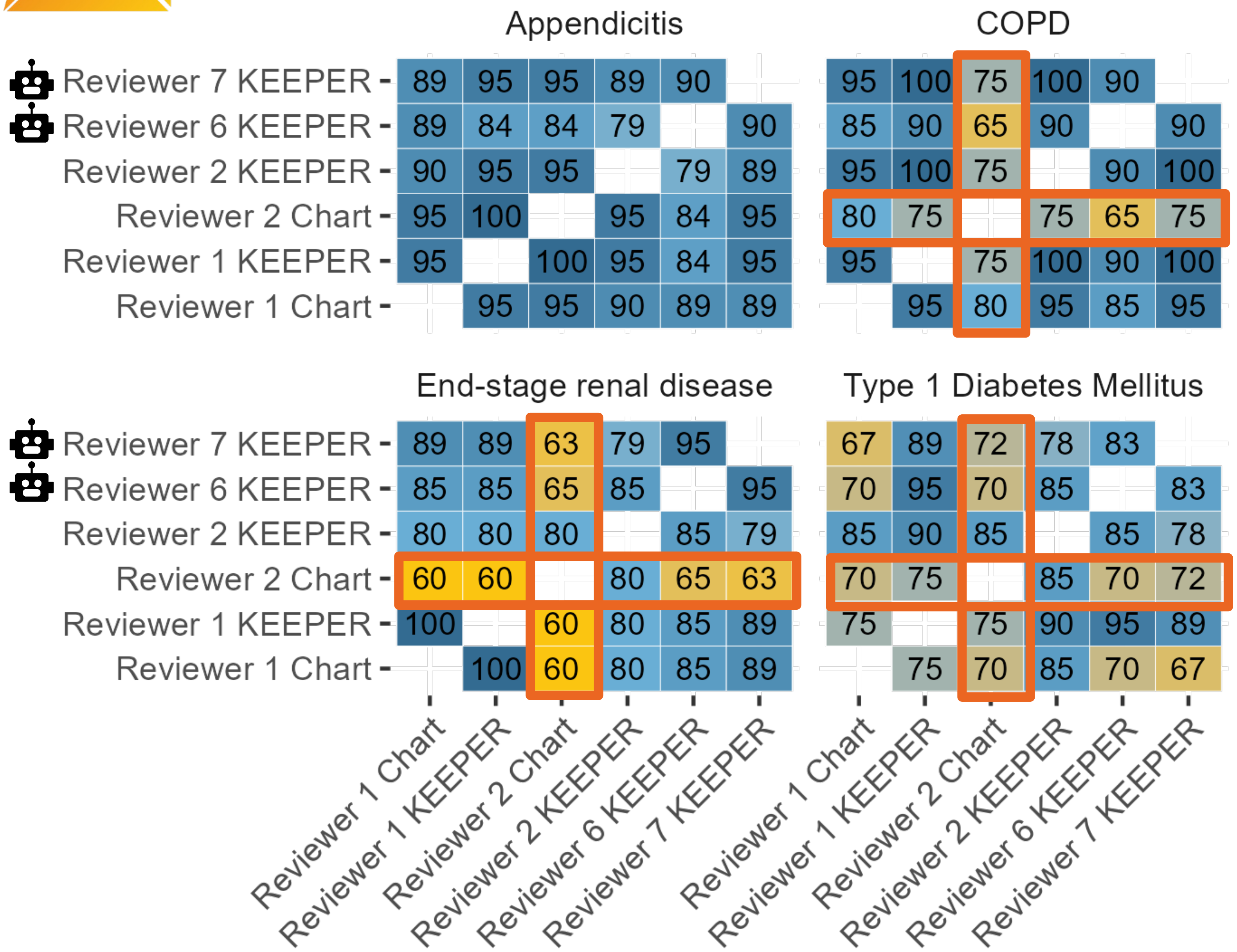
Perturbed patient data

Revisiting Experiment 1: Agreement using CUIMC EHR



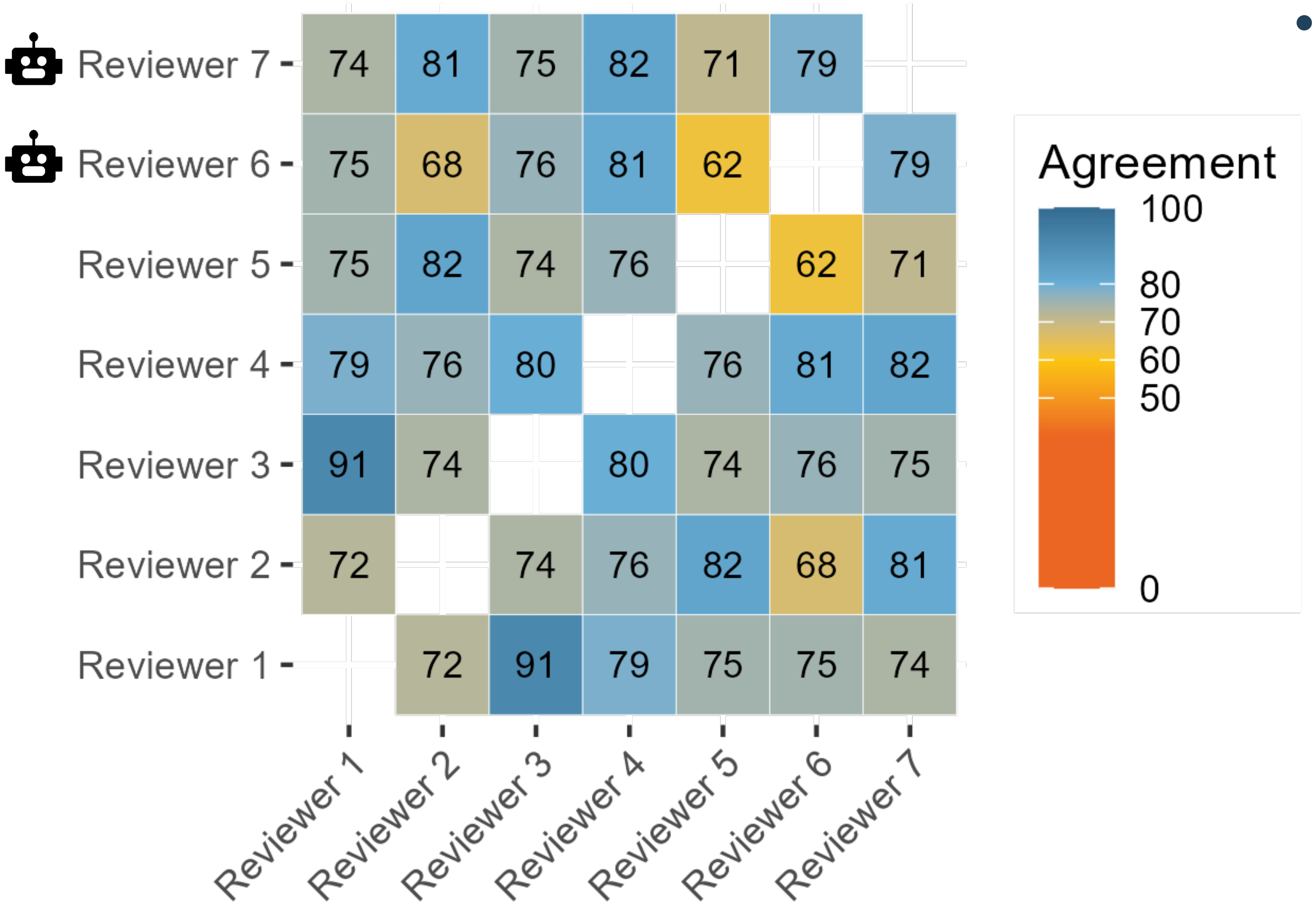
- Overall agreement was consistent across all human and LLM
- KEEPER agreement between Reviewer 1 and reviewer 2 (91%) aligns with Reviewer 1 agreement with both LLMs (89-93%)

Revisiting Experiment 1: Agreement using CUIMC EHR



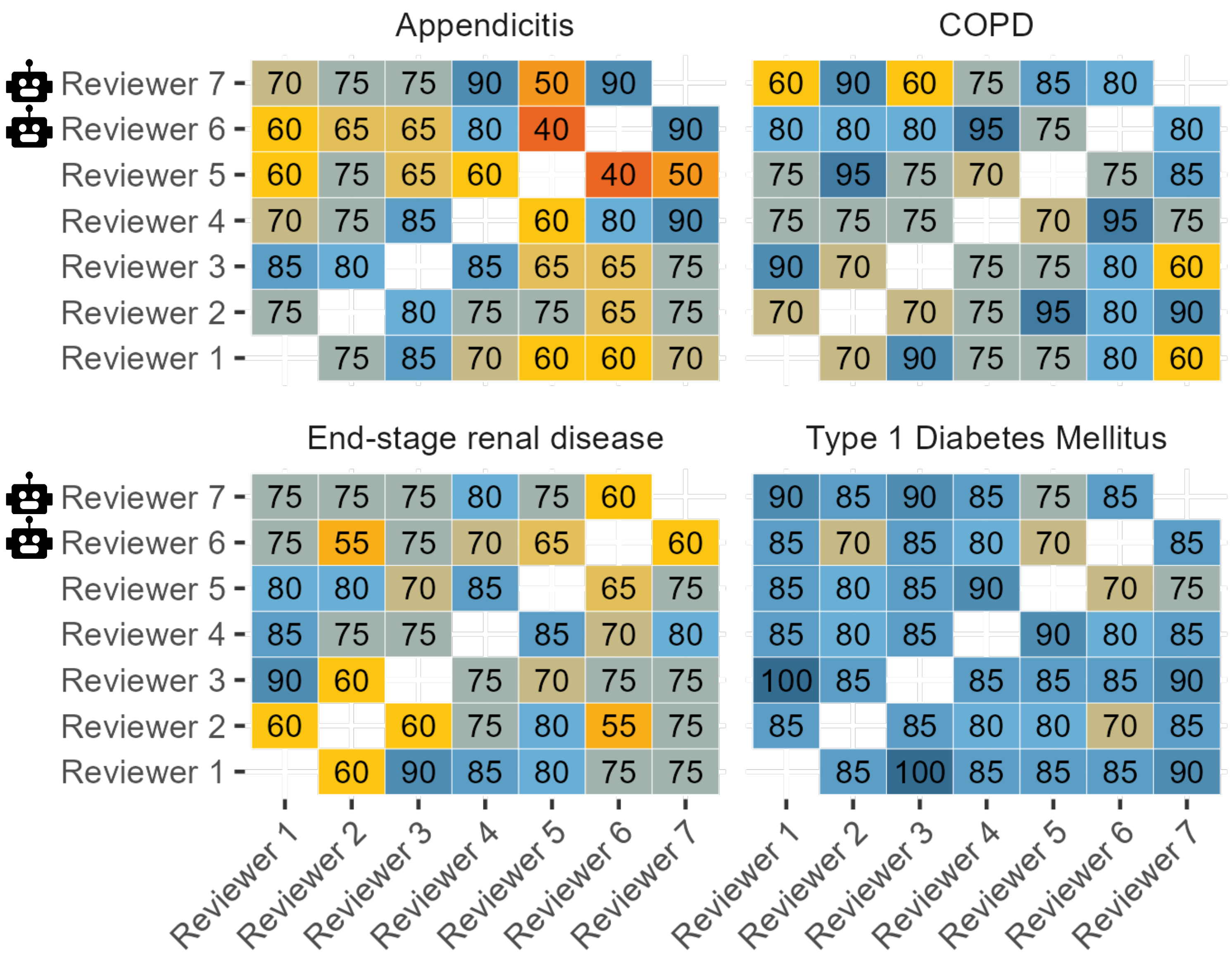
- Overall agreement was consistent across all human and LLM in each disease
- Reviewer 2 using chart was equally inconsistency with humans and LLMs

Revisiting Experiment 2: Agreement using Optum claims



- LLM agree with humans (62%-82%) about as often as humans agree with other humans (72%-91%)

Revisiting Experiment 2: Agreement using Optum claims



- Heterogeneity in agreement across diseases
- LLM performance varied by disease
  - GPT3.5 (Reviewer 6) better for COPD
  - Sheep-Duck-Llama-2 (Reviewer 7) better for others

Estimating positive predictive value: reviewer responses to Rheumatoid Arthritis in Optum

Reviewer

Case	1	2	3	4	5	6	7
	yes	yes	yes	yes	yes	yes	yes
	yes	yes	yes	yes	yes	yes	yes
	yes	yes	yes	yes	yes	yes	yes
	yes	yes	yes	yes	yes	yes	yes
	yes	yes	yes	yes	yes	yes	yes
	yes	yes	yes	yes	yes	yes	yes
	yes	yes	yes	yes	yes	yes	yes
	yes	yes	yes	yes	yes	yes	yes
	yes	yes	yes	yes	no	yes	yes
	yes	yes	yes	yes	no	yes	yes
	yes	yes	yes	yes	no	yes	yes
	yes	yes	yes	yes	no	yes	yes
	yes	yes	yes	no	yes	yes	yes
	no	yes	yes	yes	no	yes	yes
	yes	yes	yes	no	no	yes	yes
	no	yes	no	no	yes	no	yes
	no	yes	no	no	no	no	yes
	no	no	no	no	no	yes	yes
	no	no	no	no	no	yes	yes
	no	no	no	no	no	yes	no
	no	yes	no	no	no	no	no
	no	no	no	no	no	no	no
	no	no	no	no	no	no	no
	no	no	no	no	no	no	no
	no	no	no	no	no	no	no
PPV	56%	72%	60%	52%	40%	72%	76%
LB 95% CI	37%	54%	41%	32%	21%	54%	59%
UB 95% CI	75%	90%	79%	72%	59%	90%	93%

- PPV varies by reviewer: 40%-76%

Estimating positive predictive value: reviewer responses to Rheumatoid Arthritis in Optum

Reviewer

Case	1	2	3	4	5	6	7	All yes	Any yes	Majority
	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	yes	yes	yes	yes	no	yes	yes	no	yes	yes
	yes	yes	yes	yes	no	yes	yes	no	yes	yes
	yes	yes	yes	yes	no	yes	yes	no	yes	yes
	yes	yes	yes	yes	no	yes	yes	no	yes	yes
	yes	yes	yes	no	yes	yes	yes	no	yes	yes
	no	yes	yes	yes	no	yes	yes	no	yes	yes
	yes	yes	yes	no	no	yes	yes	no	yes	yes
	no	yes	no	no	yes	no	yes	no	yes	no
	no	yes	no	no	no	no	yes	no	yes	no
	no	no	no	no	no	yes	yes	no	yes	no
	no	no	no	no	no	yes	yes	no	yes	no
	no	no	no	no	no	yes	no	no	yes	no
	no	yes	no	no	no	no	no	no	yes	no
	no	no	no	no	no	no	no	no	no	no
	no	no	no	no	no	no	no	no	no	no
	no	no	no	no	no	no	no	no	no	no
	no	no	no	no	no	no	no	no	no	no
PPV	56%	72%	60%	52%	40%	72%	76%	32%	84%	60%
LB 95% CI	37%	54%	41%	32%	21%	54%	59%	14%	70%	41%
UB 95% CI	75%	90%	79%	72%	59%	90%	93%	50%	98%	79%

- PPV varies by reviewer: 40%-76%
- Alternative strategies to combine reviewers will impact PPV estimates

Estimating positive predictive value: reviewer responses to Rheumatoid Arthritis in Optum

Reviewer

Case	1	2	3	4	5	6	7	All yes	Any yes	Majority
	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	yes	yes	yes	yes	no	yes	yes	no	yes	yes
	yes	yes	yes	yes	no	yes	yes	no	yes	yes
	yes	yes	yes	yes	no	yes	yes	no	yes	yes
	yes	yes	yes	yes	no	yes	yes	no	yes	yes
	yes	yes	yes	no	yes	yes	yes	no	yes	yes
	no	yes	yes	yes	no	yes	yes	no	yes	yes
	yes	yes	yes	no	no	yes	yes	no	yes	yes
	no	yes	no	no	yes	no	yes	no	yes	no
	no	yes	no	no	no	no	yes	no	yes	no
	no	no	no	no	no	yes	yes	no	yes	no
	no	no	no	no	no	yes	yes	no	yes	no
	no	no	no	no	no	yes	no	no	yes	no
	no	yes	no	no	no	no	no	no	yes	no
	no	no	no	no	no	no	no	no	no	no
	no	no	no	no	no	no	no	no	no	no
	no	no	no	no	no	no	no	no	no	no
	no	no	no	no	no	no	no	no	no	no
PPV	56%	72%	60%	52%	40%	72%	76%	32%	84%	60%
LB 95% CI	37%	54%	41%	32%	21%	54%	59%	14%	70%	41%
UB 95% CI	75%	90%	79%	72%	59%	90%	93%	50%	98%	79%

- PPV varies by reviewer: 40%-76%
- Alternative strategies to combine reviewers will impact PPV estimates
- 25 cases provides PPV with wide confidence intervals, need more power!

LLM use cases

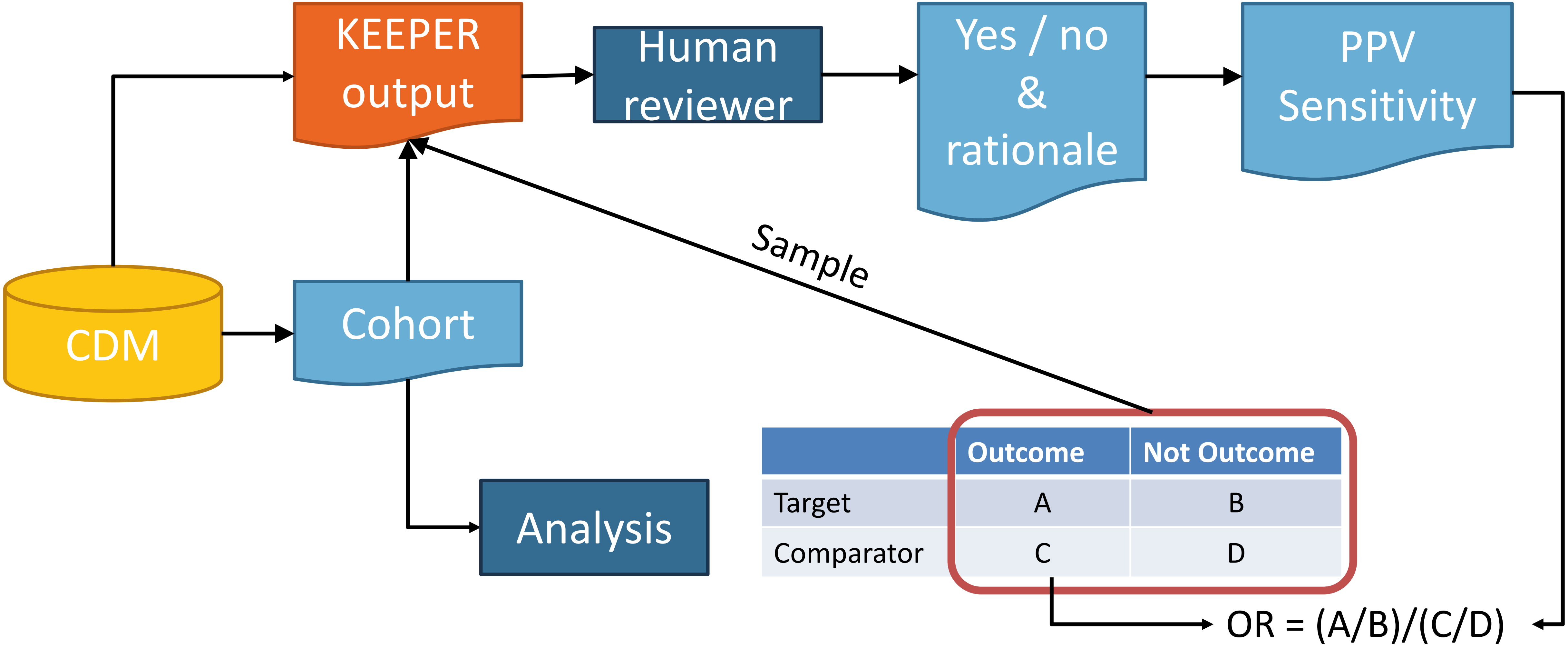
Depending on your preference, you can use the LLM

- As a **co-pilot**, to generate an assessment that a human can use as starting point to save time
- To **validate the full cohort**, and perform the observational analysis using only the confirmed cases
- To **estimate operating characteristics** of the phenotype algorithm in the database
  - PPV
  - Sensitivity!

LLM validation of highly sensitive cohort

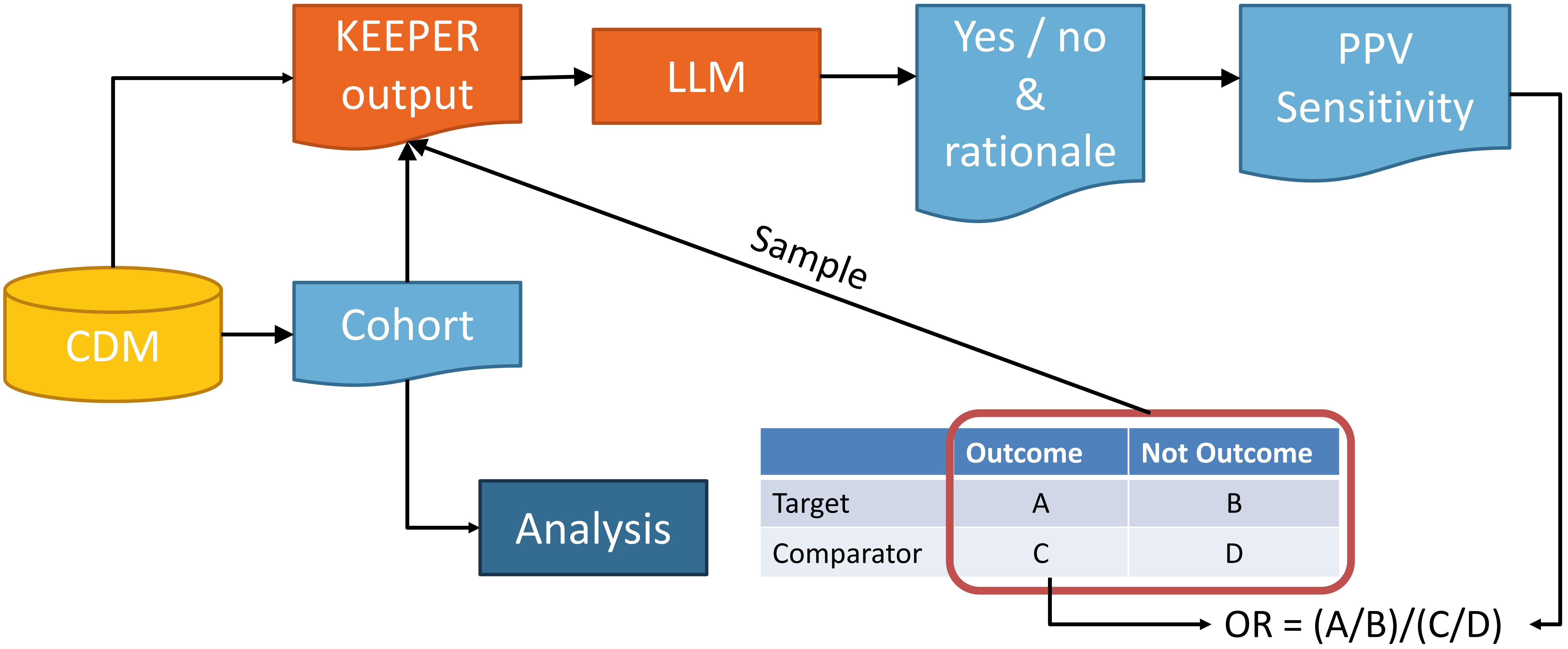
- Created highly sensitive cohort for RA: any diagnosis or symptom or treatment or complication or lab test
  - Database: Optum Clinformatics
- Sampled 25,000 persons
- Validate using KEEPER with GPT 3.5
  - Took 40 hours
  - Cost $15
- Used annotated sample to compute performance of RA phenotype algorithm (#196 in the OHDSI Phenotype Library)
  - PPV = 70.3% (0.66 - 0.74)
  - Sensitivity = 79.1% (0.75 - 0.83)

Case validation workflow



	Outcome	Not Outcome
Target	A	B
Comparator	C	D

Case validation workflow



Conclusions on LLMs

- Across all three experiments, LLMs agree with humans as much as humans agree with humans
  - LLMs have the potential to increase scale of case validation without sacrificing reliability
  - Scaling up means more precise PPV estimate, and allows estimating sensitivity, to fully enable quantitative bias analysis
- LLM performance depended strongly on choice of prompt and LLM
  - Zero-shot prompt showed good results
  - Fine-tuning would require a much larger training set
- While use of LLMs for clinical care remains controversial, our use case of increasing reliability of evidence from observational data seems promising and low risk

Overall conclusions

- Case validation is expected to be part of the evidence generation process to ensure reliability
- OHDSI has developed and evaluated standardized tools for case validation
  - Successfully applied across multiple data sources
  - Inter-rater agreement varies by disease (even when using full chart review)
- Results show
  - Standardized KEEPER output from the OMOP CDM provides a reliable and more efficient alternative to source records
  - KEEPER + LLMs provide a more scalable alternative with similar agreement to human review

Thank you to all the humans, sheep, ducks, llamas for all the case validations for this research!

Experiment	Diseases	Human reviewers per disease	Cases to review	Total cases
CUIMC KEEPER	4	2	20	160
CUIMC Chart	4	2	20	160
Optum KEEPER	4	5	20	400
Optum KEEPER	6	5	25	750
LLM Training set	6	1		358
Total validated by humans				**1,828**
Total validated by LLMs: 2 x 1,828 + 5 x 358 + 25,000 =				**30,446**

Thank you for reviewers:
Matt, Lauren, Ahmed, Ali, Oleg, Vlad, Seung In, Anna, Patrick

Thank you