Standardized patient profile review using large language models for case adjudication in observational research

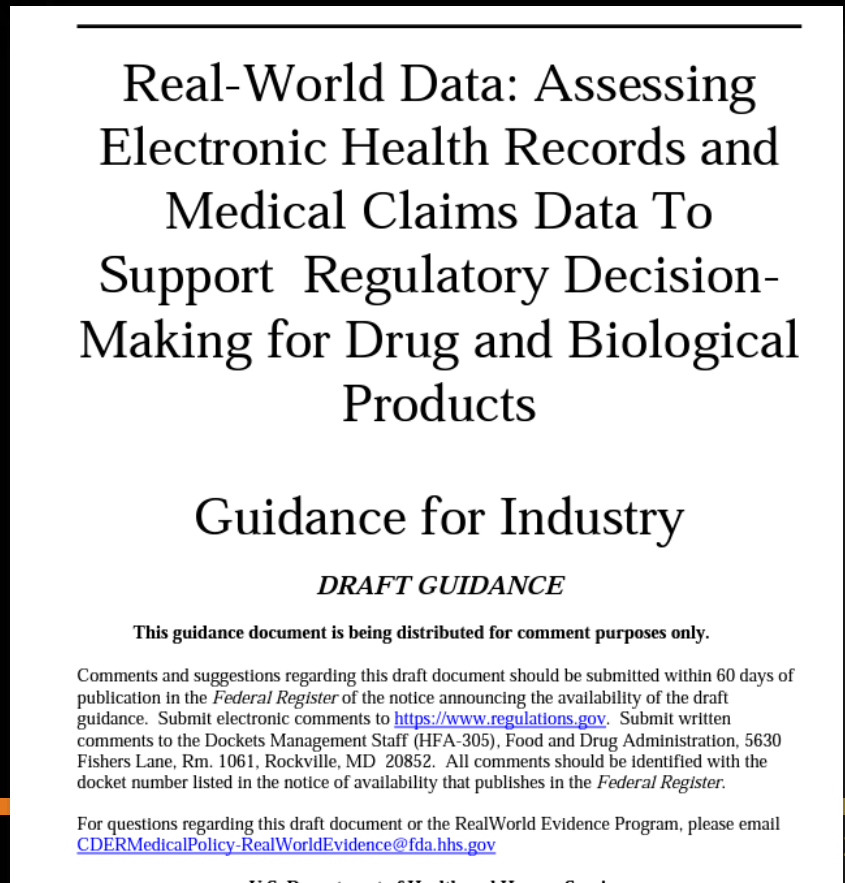
Martijn Schuemie

Johnson & Johnson

UCLA

The need for case validation in observational research

- Electronic Health Records (EHRs) and claims allow answering important clinical questions, but were not collected for research
- To find outcomes, we typically use an operational case definition (aka phenotype algorithm), for example looking for specific codes.
- Outcome misclassification may bias results
- FDA recommends (chart) review of each potential case (full case set review)
- If infeasible, review of sample to measure performance of operational case definition (PPV, sensitivity), and perform quantitative bias analysis
- (In reality: most studies only review identified cases, computing only PPV)



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DRAFT GUIDANCE

This guidance document is being distributed for comment purposes only.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be submitted within 60 days of publication in the *Federal Register* of the notice announcing the availability of the draft guidance.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

Knowledge-Enhanced Electronic Profile Review (KEEPER)

- Review system on structured data from EHR or claims data sources

Anna Ostropolets

Journal of the American Medical Informatics Association, 2024, **31(1)**, 119–129
https://doi.org/10.1093/jamia/ocad202
Advance access publication 17 October 2023
Research and Applications

Research and Applications

Scalable and interpretable alternative to chart review for phenotype evaluation using standardized structured data from electronic health records

Anna Ostropolets, MD, PhD*[,1], George Hripcsak, MD, MS[1,2], Syed A. Husain, MD, MPH[3], Lauren R. Richter, MD, MS[1], Matthew Spotnitz, MD, MPH[1], Ahmed Elhussein, MD, MS[1], Patrick B. Ryan, PhD[1,4]

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

KEEPER information categories:

- 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 in action

(For each disease)
Concept set per
KEEPER category

E.g. ESRD Symptoms:
vomiting, edema, dyspnea

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

Cohort

E.g. Patients with
end-stage renal disease
(ESRD)

KEEPER package

Time windows
per category

KEEPER output: CSV table with
1 record per person,
1 column per KEEPER category

E.g. Symptoms: -30d to 0d relative to index

Data in Common
Data Model

E.g. US insurance claims
or EHRs

KEEPER experiment overview

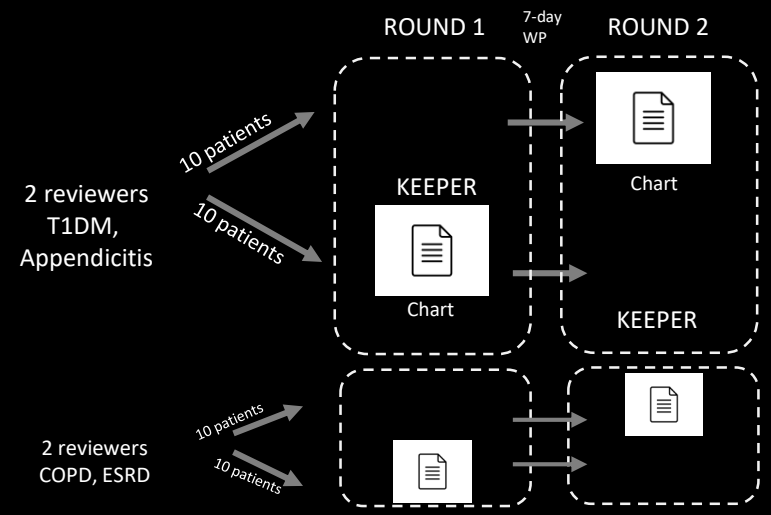
GOLD STANDARD (AO, GH)

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

KEEPER PROFILES

Created KEEPER profiles for 80 patients

EXPERIMENT (AE, LR, MS, SAH)



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

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

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

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

KEEPER 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)

KEEPER 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%*

Combining KEEPER with large language models (LLMs)

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

Can we ask a LLM to review this?

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
- Sheep-Duck-Llama-2-70b-v1.1
  - Further finetuning of Llama-2
  - Sheep-Duck-Llama-2 was at the top of the HF leaderboard at the time
  - Installed on a private machine

All analyses run securely within organizational firewall

Training set

KEEPER PROFILES

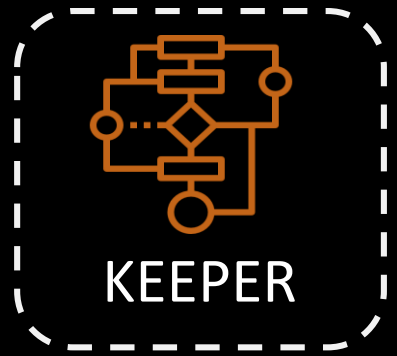
Created KEEPER profiles for 6 conditions (Acute bronchitis, hyperlipidemia, hypoparathyroidism, osteoporosis, rheumatoid arthritis, viral hepatitis type A), focus on hard cases. 358 patients total

EXPERIMENT

1 reviewer

50-100 patients

X 6 diseases



METRICS

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

DATABASE

Optum ClinFormatics® Data Mart (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 [illegible] had the disease or as justification for pe[illegible] whether the patient has the disease. A [illegible] diagnostic procedures may therefore be in[illegible] once. Lack of additional evidence of [DIS[illegible] procedures probably means that the patient [illegible] have [DISEASE]. However, it is unlikely that a patient will be tested many times over, so an abundance of diagnoses will mean the patient has the disease.

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 at the time 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

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

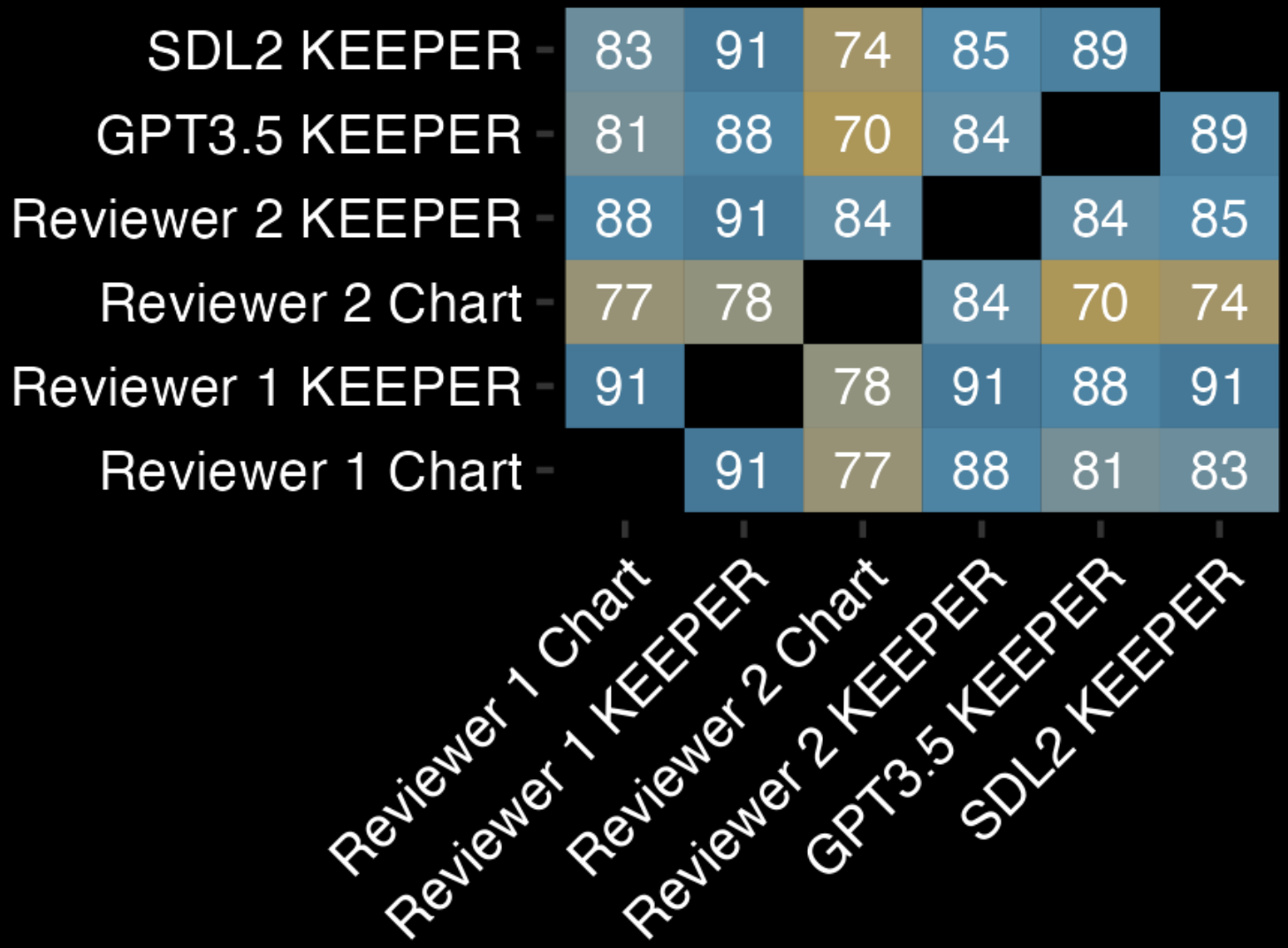
Perturbed patient data

Three evaluation experiments

- Experiment 1: (Original KEEPER evaluation)
  - 4 diseases, 20 cases per disease (acute appendicitis, diabetes mellitus type I, chronic obstructive pulmonary disorder (COPD), end-stage renal disease)
  - Balanced sample (half likely cases, half likely non-cases)
  - Columbia University Irving Medical Center EHR
  - Chart review, KEEPER human review, KEEPER LLM review
- Experiment 2:
  - Same 4 diseases, 25 cases per disease
  - Balanced sample (half likely cases, half likely non-cases)
  - Optum Clinformatics® Data Mart
  - KEEPER human review, KEEPER LLM review
- Experiment 3:
  - 6 diseases, 25 cases per disease (acute bronchitis, hyperlipidemia, hypoparathyroidism, osteoporosis, rheumatoid arthritis, viral hepatitis type A)
  - Random sample of cases
  - Optum Clinformatics® Data Mart
  - KEEPER human review, KEEPER LLM review

Using identified cases only, so allowing computing PPV only

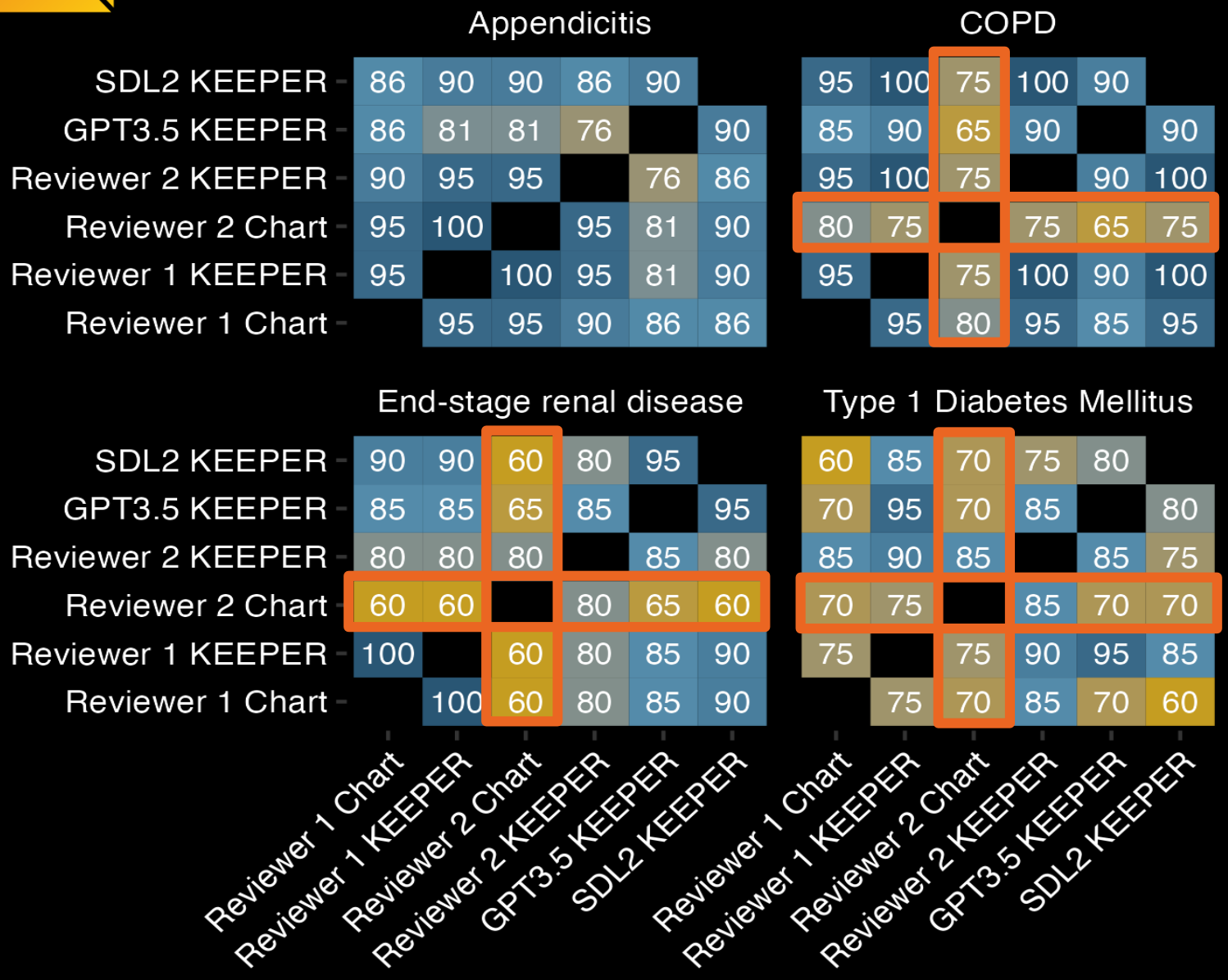
Experiment 1 results: agreement



- Humans agree with humans (median = 86%) as often as humans agree with GPT3.5 (median = 84%) and SDL2 (median = 85%)

Columbia University Medical Center EHR
4 diseases

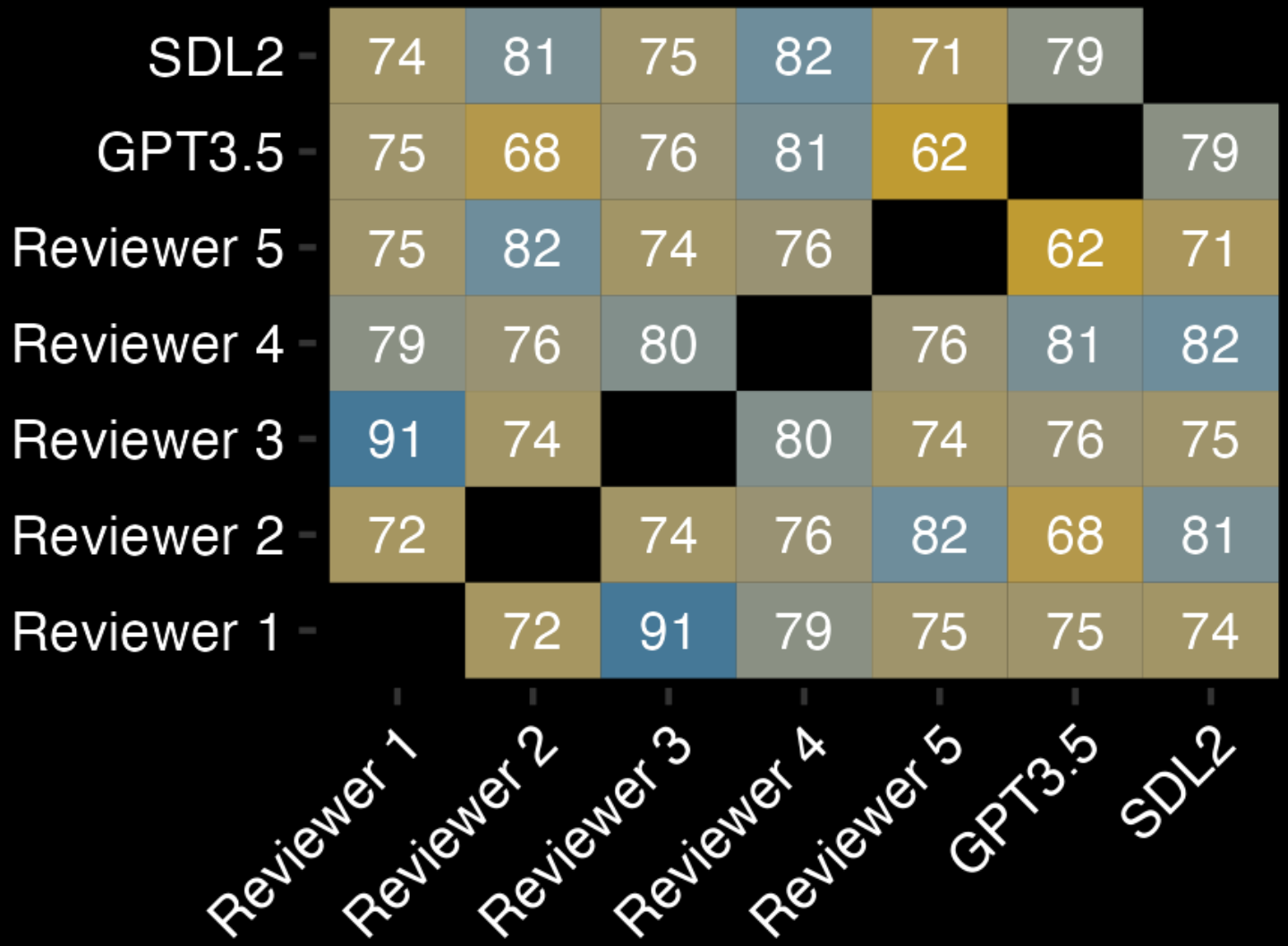
Experiment 1 results : agreement



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

Columbia University Medical Center EHR
4 diseases

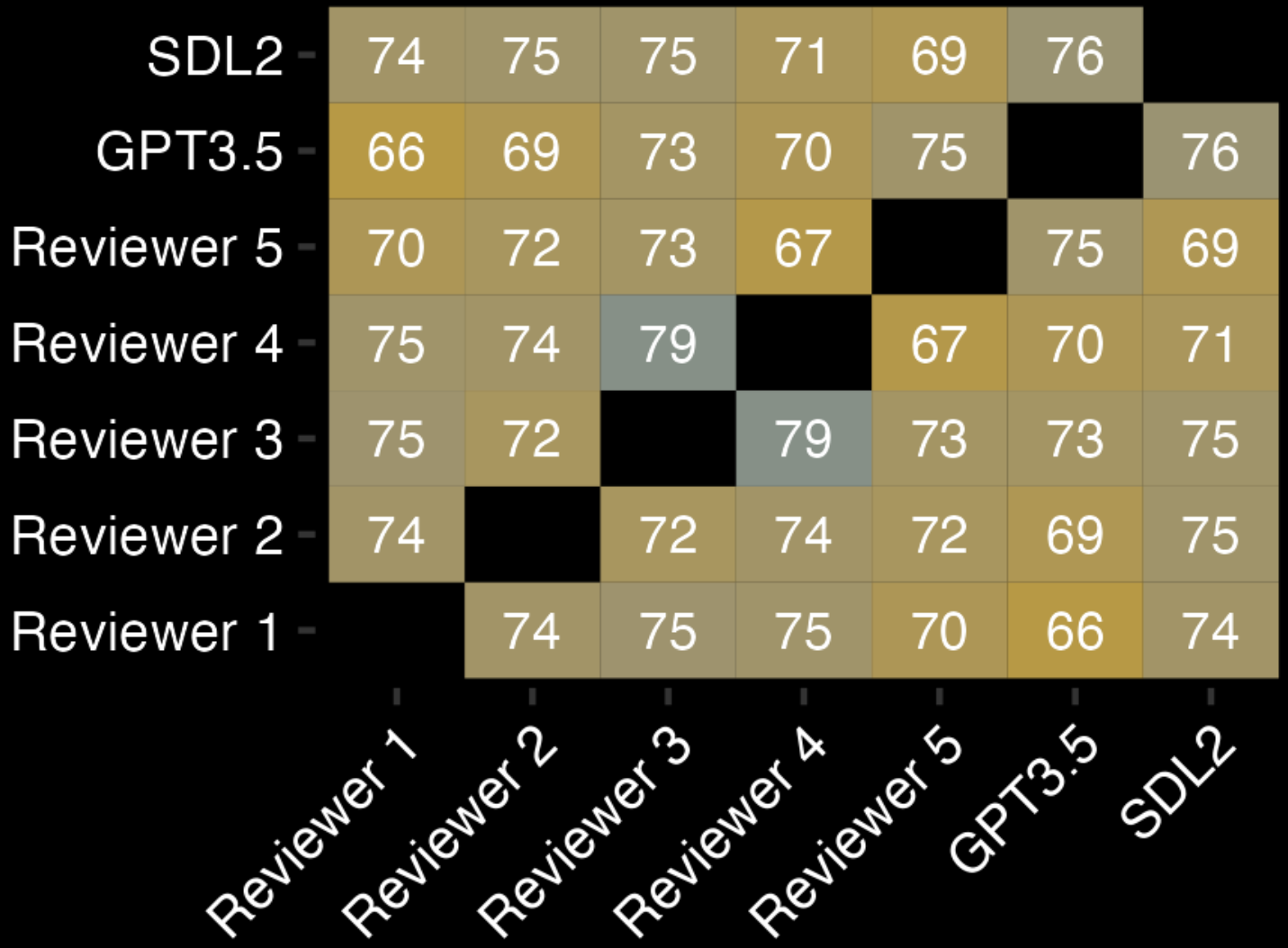
Experiment 2 results : agreement



- Humans agree with humans (median = 76%) as often as humans agree with GPT3.5 (median = 76%) and SDL2 (median = 77%)

Optum Clinformatics® Data Mart
4 diseases

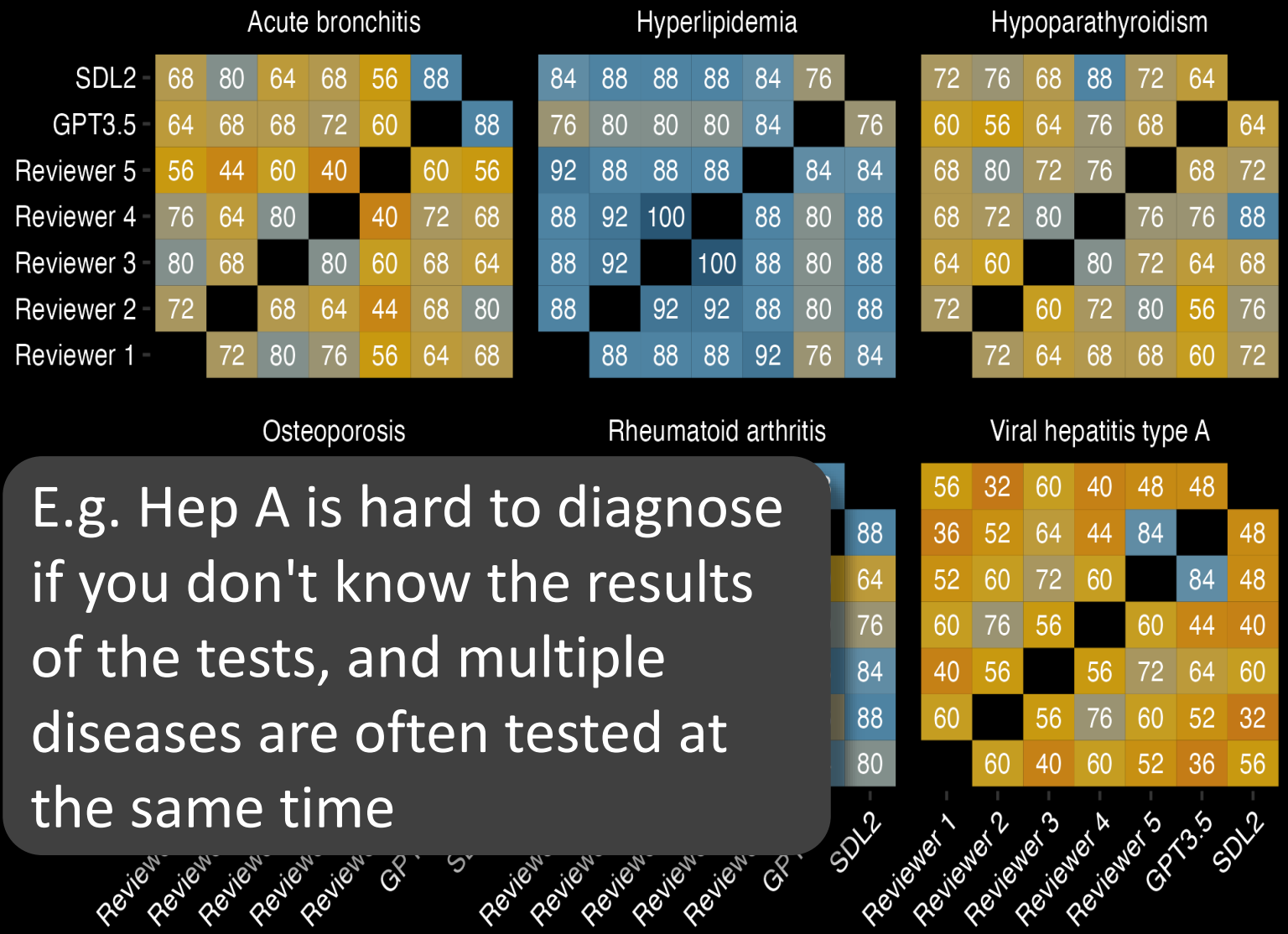
Experiment 3 results : agreement



- Humans agree with humans (median = 74%) as often as humans agree with GPT3.5 (median = 72%) and SDL2 (median = 74%)

Optum Clinformatics® Data Mart
6 diseases

Experiment 3 results : agreement



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

Optum Clinformatics® Data Mart
6 diseases

Computing sensitivity: using a highly sensitive cohort

- Created highly sensitive cohort for RA: any diagnosis or symptom or treatment or complication or lab test
  - Database: Optum Clinformatics® Data Mart
- Sampled 25,000 persons
- Validate using KEEPER with GPT 3.5
  - Took 40 hour
  - Cost $15
  - Classified 527 persons as cases (2.1%)
- Used annotated sample to compute performance of RA phenotype algorithm (#196 in the OHDSI Phenotype Library)
  - PPV = 70.3% (66% - 74%)
  - Sensitivity = 79.1% (75% - 83%)

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

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

Thanks to

- Anna Ostropolets
- Oleg Zhuk
- Vlad Korsik
- Marc Suchard
- George Hripcsak
- Patrick Ryan