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

- Manual chart review - gold standard but resource- and time-consuming
- Do nothing (borrow from the literature)
+ other tools in the community (CohortDiagnostics, PheValuator)
Main challenge of chart review for phenotype evaluation

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*

(clinical elements to extract: presentation, plausibility = prior history of disease, demographics, differential diagnosis, diagnostic procedures and labs, treatment and complications)

Principle 2: Dimensionality reduction

*Only extract relevant information*

Principle 3: Standardization

*Both input and output are standardized across data sources and condition*
**KEEPER as an OHDSI package**

**Per disease:**
- Concept sets per KEEPER category

**Cohort**
- Ex: ESRD

**KEEPER data extraction**
- Ex: ESRD Symptoms: vomiting, edema, dyspnea

**Time windows per category**
- Ex: Symptoms: -30d to 0d before index date

**CSV table:**
- record per person, column per element

<table>
<thead>
<tr>
<th>Person_id</th>
<th>Symptoms</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Vomiting and nausea (day -29); Dyspnea (day -11);...</td>
</tr>
</tbody>
</table>
Experiment: Data preparation

**GOLD STANDARD (AO, GH)**
- Random sample of 20 patients per eMERGE algorithm
- Iterative review on full chart + all structured data

<table>
<thead>
<tr>
<th></th>
<th>T1DM</th>
<th>Acute append</th>
<th>COPD</th>
<th>ESRD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case</td>
<td>12</td>
<td>15</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td>Control</td>
<td>8</td>
<td>5</td>
<td>9</td>
<td>7</td>
</tr>
</tbody>
</table>

**KEEPER PROFILES**
- Created KEEPER profiles for 80 patients

Columbia University EHR
Experiment

**EXPERIMENT** (AE, LR, MS, SAH)

2 reviewers
T1DM, Appendicitis

2 reviewers
COPD, ESRD

7-day washout

ROUND 1

KEEPER

ROUND 2

Chart

Dataset:
- 160 patients adjudicated with KEEPER
- 160 patients adjudicated with chart review
Experiment: 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 and agreement

<table>
<thead>
<tr>
<th>KEEPER was:</th>
<th>Chart review</th>
<th>KEEPER</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Twice as fast</td>
<td>67 minutes per 20 patients</td>
<td>30 minutes</td>
</tr>
<tr>
<td>2. More consistent</td>
<td>76.3% (61 pts)</td>
<td>91.2% (73 pts)</td>
</tr>
<tr>
<td>3. As accurate</td>
<td>86.9% (139 pts)</td>
<td>88.1% (141 pts)</td>
</tr>
</tbody>
</table>
KEEPER as an OHDSI package

Input:
1) Cohort defined by the phenotyping algorithm
2) Concept_ids for the clinical elements you want to see

Output:
Record per patient to examine and label
More in the OHDSI Symposium 2023 Plenary

Plenary: Improving the reliability and scale of case validation

Case validation is regarded as a necessary element of regulatory-grade evidence, but conducting case validation through human adjudication of source records is time- and resource-intensive, has unknown performance, and is frequently conducted in such a way that does not enable either full caseset review or proper quantitative bias analysis. In this plenary, OHDSI collaborators presented innovative methodological research and open-source development to improve the reliability and scalability of the case validation process, demonstrating that it may be possible to replace source records through an informatics-enhanced patient profile of structured data from the OMOP CDM (KEEPER), and to supplement human review through the use of large language models to estimate measurement error and identify differential misclassification. KEEPER + LLM was empirically evaluated in 10 diseases across 3 experiments in 2 different data sources, and revealed that there can be substantial heterogeneity in agreement between human reviewers but that LLMs agree with humans as much as humans agree with each other. **Speakers included:** Patrick Ryan, Johnson & Johnson, Columbia University; Anna Ostropolets, Odysseus Data Service; and Martijn Schuemie, Johnson & Johnson, University of California, Los Angeles

https://www.ohdsi.org/ohdsi2023/