# Can Linked Claims-EHR Data be Used to Generate Synthetic Data? An Evaluation of the MarketScan-Veradigm Linked Claims + EHR Database

Elizabeth R. Packnett, Caroline Henriques, Liisa A. Palmer
Merative, Real World Data Research & Analytics, Ann Arbor, MI, USA

## Background

- Retrospective analyses to evaluate the safety of prenatal exposures may require validation of outcomes or exposures using electronic health records (EHR) [1]. When linkage to EHR records is not available the generation of synthetic EHR data may be used to augment existing data.
- The U.S. Food and Drug Administration (FDA) recently issued guidance on the use of artificial intelligence (AI) to support regulatory decision-making regarding drug safety and effectiveness highlighting that data used by AI should include key data elements and sufficient numbers of representative participants [2].

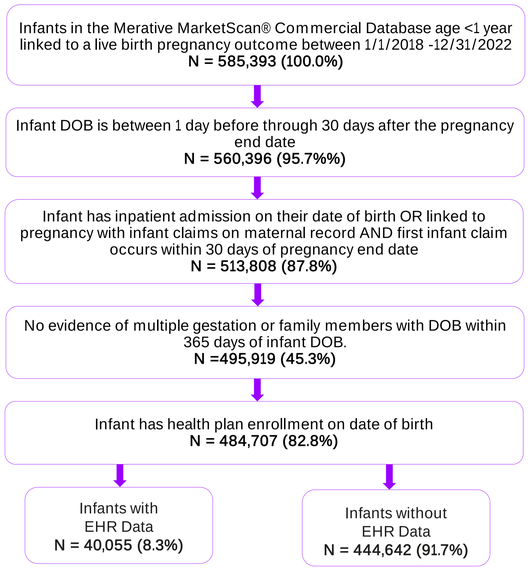
## Objective

- To assess the feasibility of using AI to generate synthetic EHR data by comparing infants in the MarketScan Commercial Database with and without linkage to the Veradigm Network EHR (VNEHR).

## Methods

- This study used data from the Merative™ MarketScan® Commercial Database spanning 1/1/2018-12/31/2022. Infants linked to a live birth pregnancy outcome were required to have health plan enrollment on date of birth (DOB) and DOB and at least one medical claim within -1 to 30 days of the pregnancy end date. Multiples and infants with a family member with a DOB within 365 days were excluded from the study. (Figure 1)
- Infant characteristics were summarized on the date of birth using enrollment data on the DOB; healthcare resource use (HCRU) and costs were tabulated using medical and pharmacy claims in the first year of life using data from the MarketScan Commercial Database.
- Infants with and without data in VNEHR were identified and infant characteristics, HCRU, and costs were compared in infants with and without VNEHR using standardized mean differences (SMD).

Figure 1. Patient Selection




DOB: Date of Birth; EHR: Electronic Health Records

Table 2. HCRU in Infants with and without EHR Data

| | No EHR Data (N=444,642) | | EHR Data (N=40,065) | | |
|---|---|---|---|---|---|
| | N/Mean | %/SD | N/Mean | %/SD | SMD |
| **Inpatient Admission** | | | | | |
| Patients with an Inpatient Admission (N,%) | 338,262 | 76.1% | 29,727 | 74.2% | 0.04 |
| Number of inpatient admissions (Mean, SD) | 5.0 | 22.2 | 4.4 | 18.9 | 0.03 |
| **ER visits** | | | | | |
| Patients with an ER Visit (N,%) | 91,671 | 20.6% | 9,011 | 22.5% | 0.05 |
| Number of ER visits (Mean, SD) | 0.7 | 2.0 | 0.7 | 2.0 | 0.02 |
| **Outpatient Office Visits** | | | | | |
| Patients with and outpatient office visit | 436,908 | 98.3% | 39,768 | 99.3% | 0.09 |
| Number of office visits (Mean, SD) | 12.2 | 7.7 | 13.4 | 7.8 | 0.15 |
| Well baby office visits | | | | | |
| Patients with a well baby visit (N,%) | 434,048 | 97.6% | 39,518 | 98.6% | 0.07 |
| Number of well baby visits (Mean, SD) | 6.9 | 4.4 | 6.8 | 3.9 | 0.03 |
| **Other Outpatient Services (N,%)** | | | | | |
| Patients with other outpatient services (N,%) | 435,951 | 98.1% | 39,436 | 98.4% | 0.03 |
| **Outpatient pharmacy (N,%)** | 245,134 | 55.1% | 25,459 | 63.5% | 0.17 |
| Number of prescriptions (Mean, SD) | 2.4 | 4.0 | 3.3 | 4.8 | 0.19 |

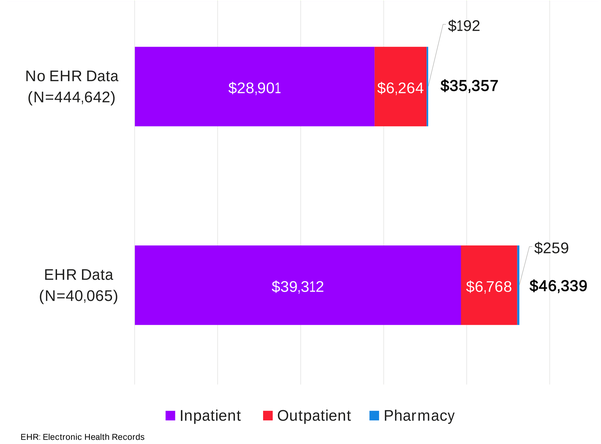
EHR: Electronic Health Records; ER: Emergency Room; SD: Standard Deviation; SMD: Standardized Mean Difference

Table 1. Demographic Characteristics of Infants with and without EHR Data

| | No EHR Data (N=444,642) | | EHR Data (N=40,065) | | |
|---|---|---|---|---|---|
| | N/Mean | %/SD | N/Mean | %/SD | SMD |
| **Sex (N, %)** | | | | | |
| Male | 226,870 | 51.0% | 21,502 | 53.7% | 0.05 |
| Female | 217,772 | 49.0% | 18,563 | 46.3% | 0.05 |
| **Geographic region (N, %)** | | | | | |
| Northeast | 76,935 | 17.3% | 10,164 | 25.4% | 0.20 |
| North Central | 98,117 | 22.1% | 7,781 | 19.4% | 0.07 |
| South | 186,598 | 42.0% | 17,304 | 43.2% | 0.02 |
| West | 80,893 | 18.2% | 4,786 | 12.0% | 0.18 |
| Unknown | 2,099 | 0.5% | 30 | 0.1% | 0.08 |
| **Population density (N, %)** | | | | | |
| Urban | 406,564 | 91.4% | 36,685 | 91.6% | <0.01 |
| Rural | 36,067 | 8.1% | 3,359 | 8.4% | 0.01 |
| Unknown | 2,011 | 0.5% | 21 | 0.1% | 0.08 |
| **Insurance plan type (N, %)** | | | | | |
| Comprehensive/indemnity | 6,193 | 1.4% | 1,092 | 2.7% | 0.09 |
| EPO/PPO | 202,887 | 45.6% | 22,239 | 55.5% | 0.20 |
| POS/POS with capitation | 43,941 | 9.9% | 2,052 | 5.1% | 0.18 |
| HMO | 61,791 | 13.9% | 5,355 | 13.4% | 0.02 |
| CDHP/HDHP | 120,217 | 27.0% | 8,505 | 21.2% | 0.14 |
| Other/Unknown | 9,613 | 2.2% | 822 | 2.1% | 0.01 |
| **Duration of follow-up in claims (Mean, SD)** | 306.1 | 105.8 | 316.9 | 95.9 | 0.11 |
| Median | 365.0 | | 365.0 | | |
| **Duration of follow-up in claims (N, %)** | | | | | |
| <30 days | 9,820 | 2.2% | 447 | 1.1% | 0.09 |
| 31-90 days | 26,769 | 6.0% | 1,961 | 4.9% | 0.05 |
| 91-180 days | 39,372 | 8.9% | 3,098 | 7.7% | 0.04 |
| >180 days | 368,681 | 82.9% | 34,559 | 86.3% | 0.09 |

CDHP: Consumer Driven Health Plan; EHR: Electronic Health Records; EPO: Exclusive Provider Organization; HDHP: High Deductible Health Plan; HMO: Health Management Organization; POS: Point of Service; PPO: Preferred Provider Organization; SD: Standard Deviation; SMD: Standardized Mean Difference

Figure 2. All-Cause Healthcare Costs in Infants with and without EHR Data




EHR: Electronic Health Records

## Results

- A total 484,707 infants linked to a live birth pregnancy outcome were included in the study; 40,065 (8.3%) of infants linked to a pregnancy had data available in the EHR. (Table 1)
- The proportion of males in infants with (53.7%) and without (51.0%) EHR data was similar (SMD=0.05) as was the proportion of infants in urban areas (91.6% vs. 91.4%; SMD <0.01). (Table 1).
- Small differences observed were in the proportion of infants in the Northeast ( 25.4% vs. 17.3%; SMD=0.20) and West (12.0% vs. 18.2%; SMD=0.18) regions and in infants with EPO/PPO (55.5% vs. 45.6%; SMD=0.20) and POS/POS with capitation (5.1% vs. 9.9%; SMD=0.18) health plans. (Table 1)
- Duration of follow-up was also similar in infants with (317 days, SD: 96) and without (306 days, SD: 106) EHR linkage (SMD=0.11). Most infants had at least 6 months of follow-up (83.2%) and proportion of infants with 6 months of follow-up was similar in infants with and without EHR data (86.3% vs. 82.9%, SMD=0.09). (Table 1)
- HCRU patterns were similar in infants with and without EHR data, though the proportion of infants with a pharmacy claim (63.5% vs. 55.1%; SMD:0.17) and number of pharmacy claims (4.8 vs. 2.4; SMD: 0.19) were higher in patients with EHR data. (Table 2)
- Total healthcare costs were similar overall ($46,339 vs. $35,357; SMD=0.01) and by expenditure category (SMD<0.05 for all). (Figure 2)

## Conclusions

- Small differences in geography and health plan type were observed in infants with and without EHR data . These differences are likely due to underlying differences in the MarketScan and Veradigm populations.
- No systematic differences in other patient characteristics, HCRU patterns, or all-cause healthcare costs were observed in infants with and without EHR data, suggesting this database may be appropriate for the generation of synthetic data.

### References

1. FDA. Postapproval Pregnancy Safety Studies Guidance for Industry. May 2019
2. FDA. Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products. January 2025

### Disclosure

All authors are employees of Merative. This study was funded by Merative.