Applying Machine Learning in Distributed Networks to Support Activities for Post-Market Surveillance of Medical Products: Opportunities, Challenges, and Considerations

Jenna Wong, PhD
Assistant Professor, Harvard Medical School
Department of Population of Medicine, Harvard Pilgrim Health Care Institute

CBER BEST Seminar Series – June 26, 2024

DEPARTMENT OF POPULATION MEDICINE
Harvard Pilgrim Health Care Institute

COI Disclosure Information
Jenna Wong

- Salary support from: FDA-funded Sentinel Innovation Center activities

- All views discussed in this presentation reflect my own perspectives.

DEPARTMENT OF POPULATION MEDICINE | Harvard Pilgrim Health Care Institute

Collection

Role of Artificial Intelligence and Machine Learning in Pharmacovigilance

Drug Safety (2022) 45:493–510
https://doi.org/10.1007/s40264-022-01158-3

REVIEW ARTICLE

Applying Machine Learning in Distributed Data Networks for Pharmacoepidemiologic and Pharmacovigilance Studies: Opportunities, Challenges, and Considerations

Jenna Wong[1] · Daniel Prieto-Alhambra[2,3] · Peter R. Rijnbeek[3] · Rishi J. Desai[4] · Jenna M. Reps[5] · Sengwee Toh[1]

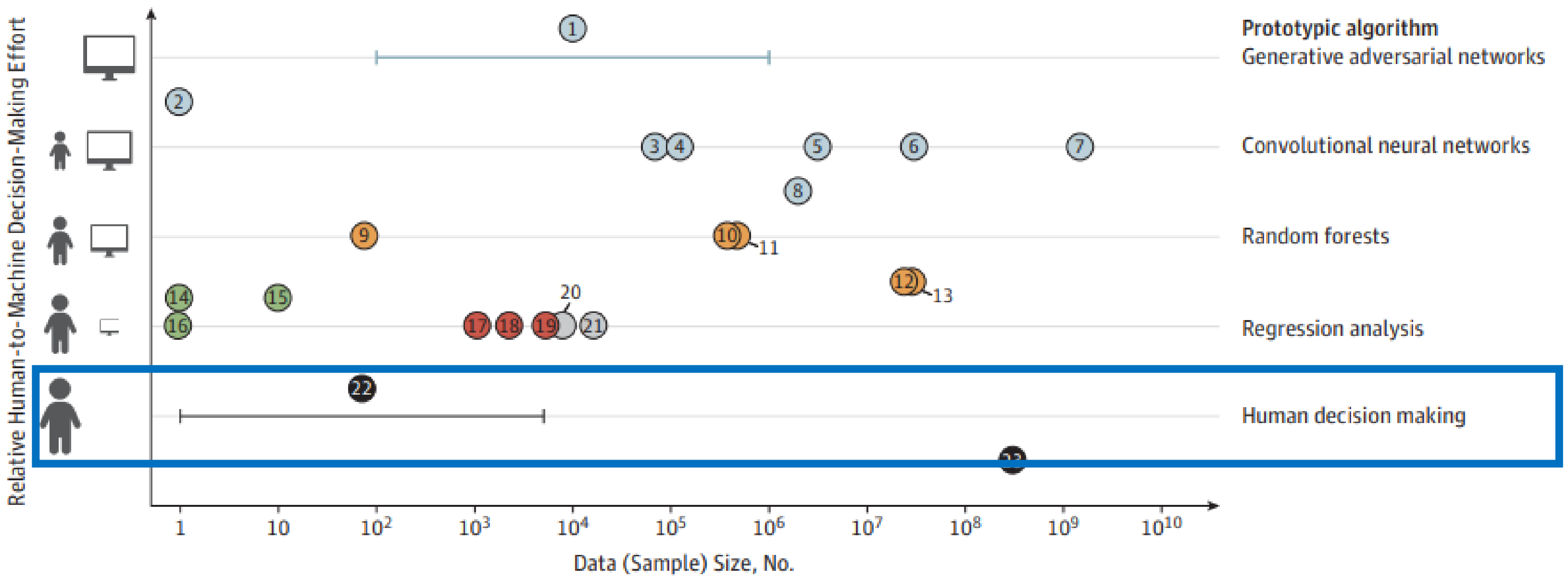
Overview

1. Definitions
2. Key activities of distributed data networks
3. Practical aspects of distributed data networks
4. Four scenarios
5. Additional considerations and conclusions

Figure. The Axes of Machine Learning and Big Data

Beam AL, Kohane IS. JAMA. 2018;319(13):1317-1318.

Figure. The Axes of Machine Learning and Big Data



Deep learning

1. Generative adversarial networks (2014)
2. Google AlphaGo Zero (2017)
3. ATM check readers (1998)
4. Google diabetic retinopathy (2016)
5. ImageNet computer vision models (2012-2017)
6. Google AlphaGo (2015)
7. Facebook Photo Tagger (2015)
8. Prediction of 1-y all-cause mortality (2017)

Classic machine learning

9. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling (2002)
10. EHR-based CV risk prediction (2017)
11. Netflix Prize winner (2006)
12. Google Search (1998)
13. Amazon product recommendation (2003)

Expert AI systems

14. MYCIN (1975)
15. CASNET (1982)
16. DXplain (1986)

Risk calculators

17. CHA_2DS_2-VASc Score for atrial fibrillation stroke risk (2017)
18. MELD end-stage liver disease risk score (2001)
19. Framingham CV risk score (1998)

Randomized Clinical Trials

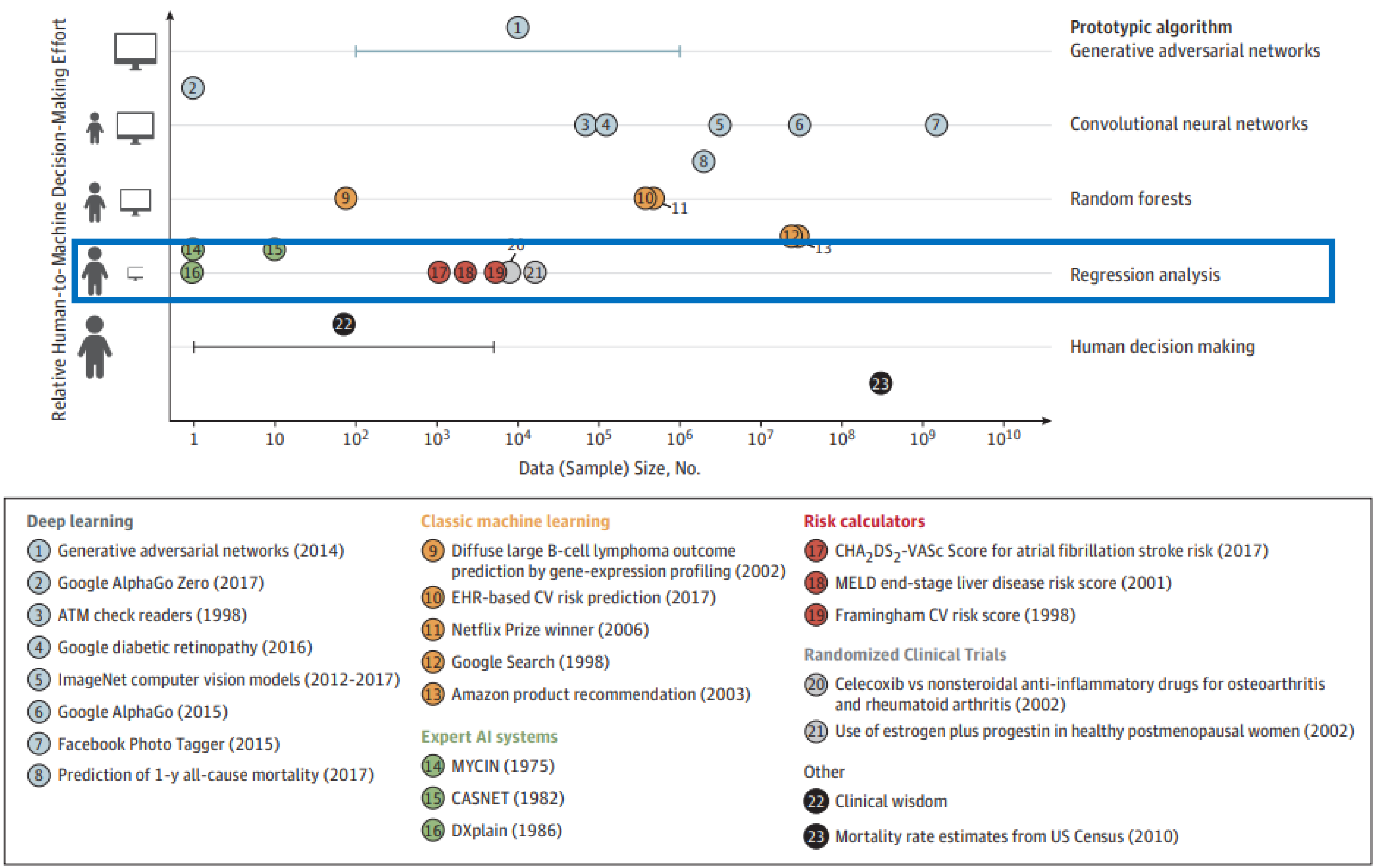
20. Celecoxib vs nonsteroidal anti-inflammatory drugs for osteoarthritis and rheumatoid arthritis (2002)
21. Use of estrogen plus progestin in healthy postmenopausal women (2002)

Other

22. Clinical wisdom
23. Mortality rate estimates from US Census (2010)

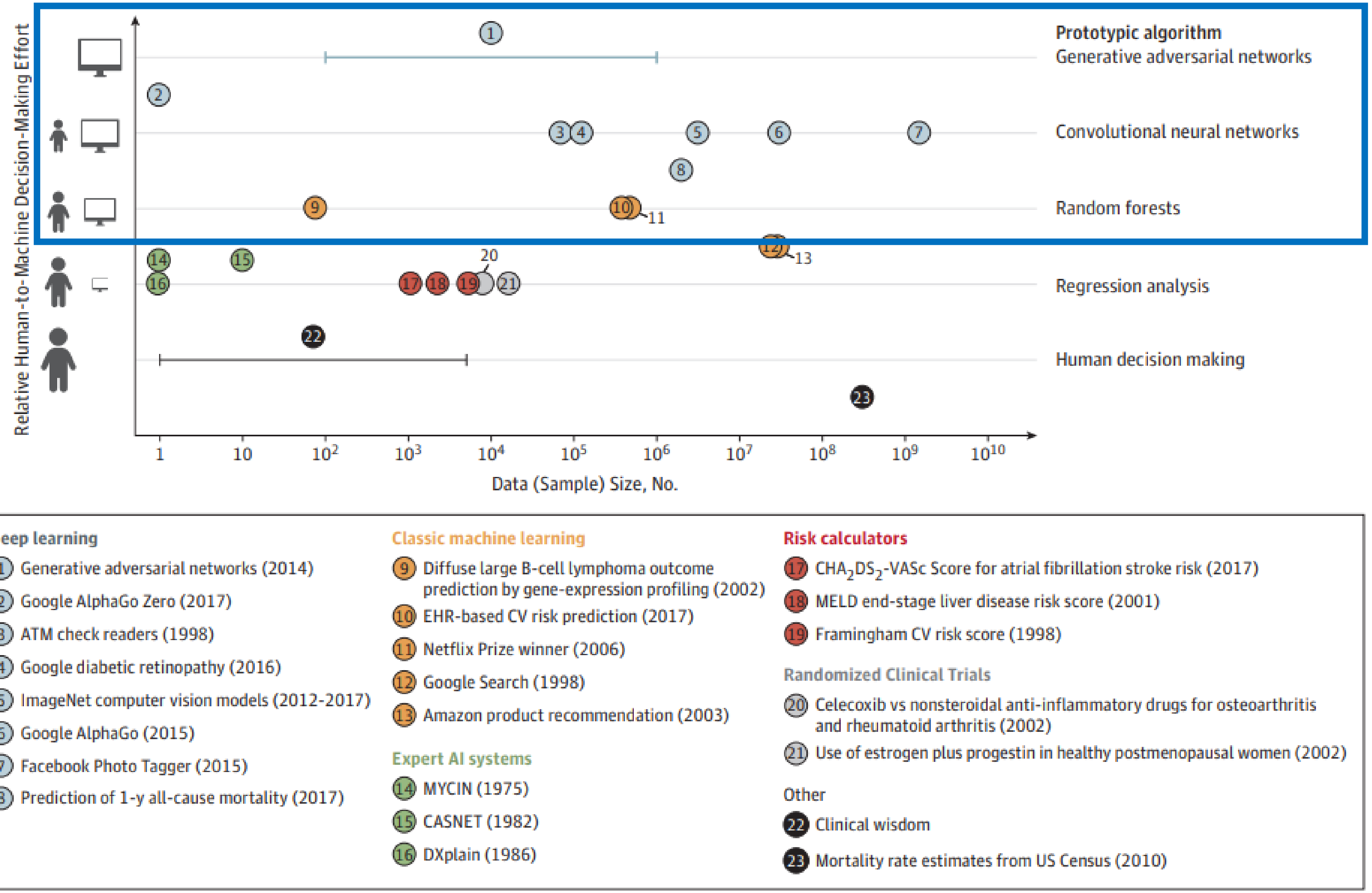
Beam AL, Kohane IS. JAMA. 2018;319(13):1317-1318.

Figure. The Axes of Machine Learning and Big Data



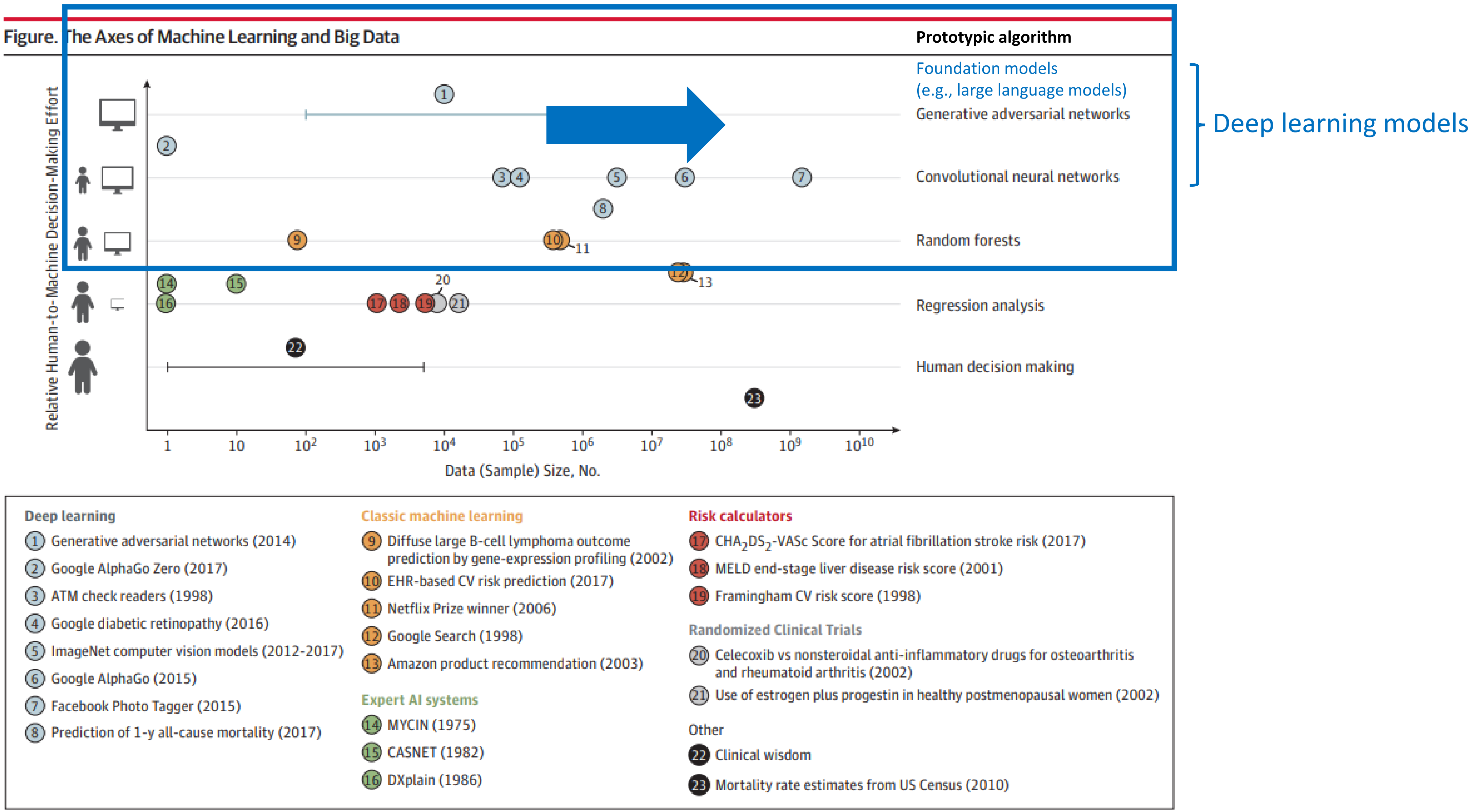
Beam AL, Kohane IS. JAMA. 2018;319(13):1317-1318.

Figure. The Axes of Machine Learning and Big Data



Pre-ChatGPT era

Beam AL, Kohane IS. JAMA. 2018;319(13):1317-1318.

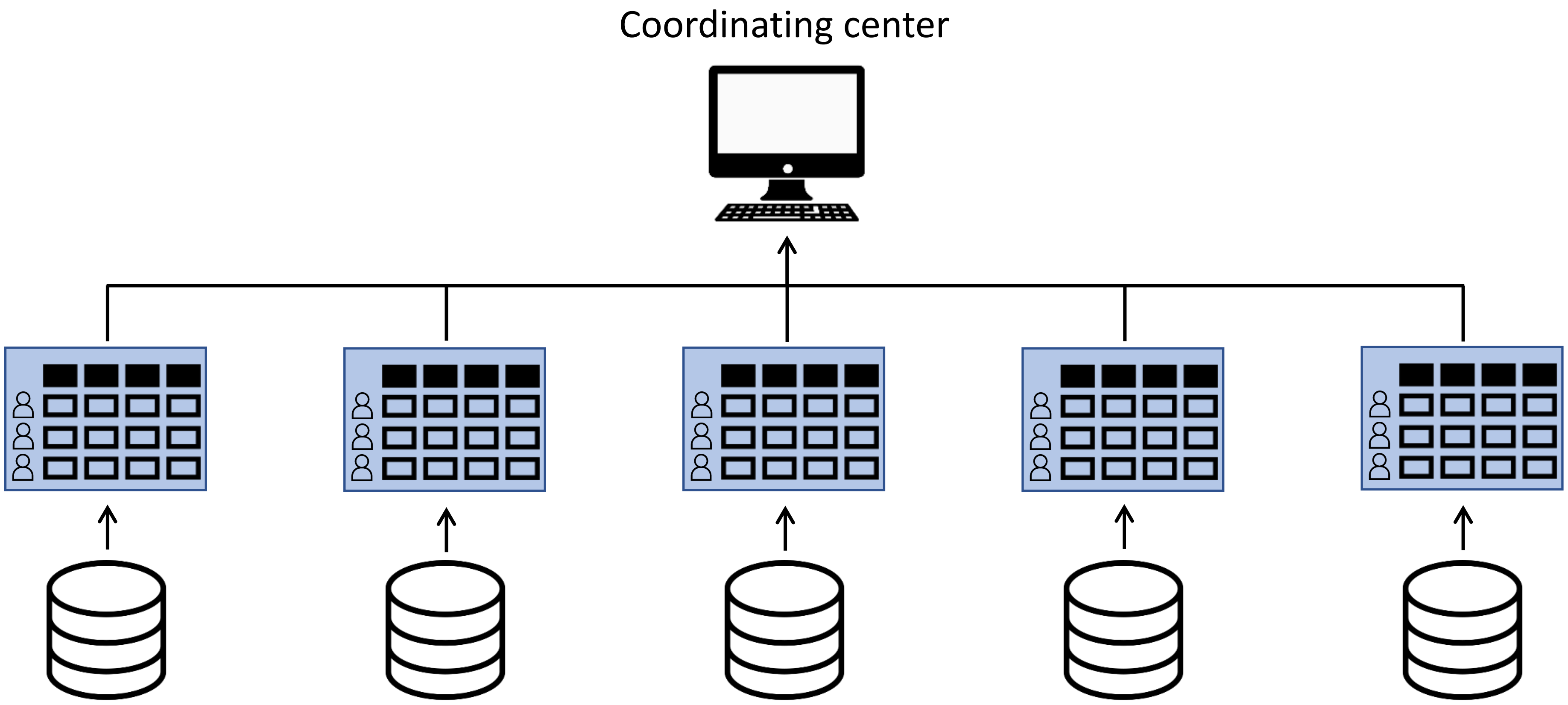


Modified from Beam AL, Kohane IS. JAMA. 2018;319(13):1317-1318.

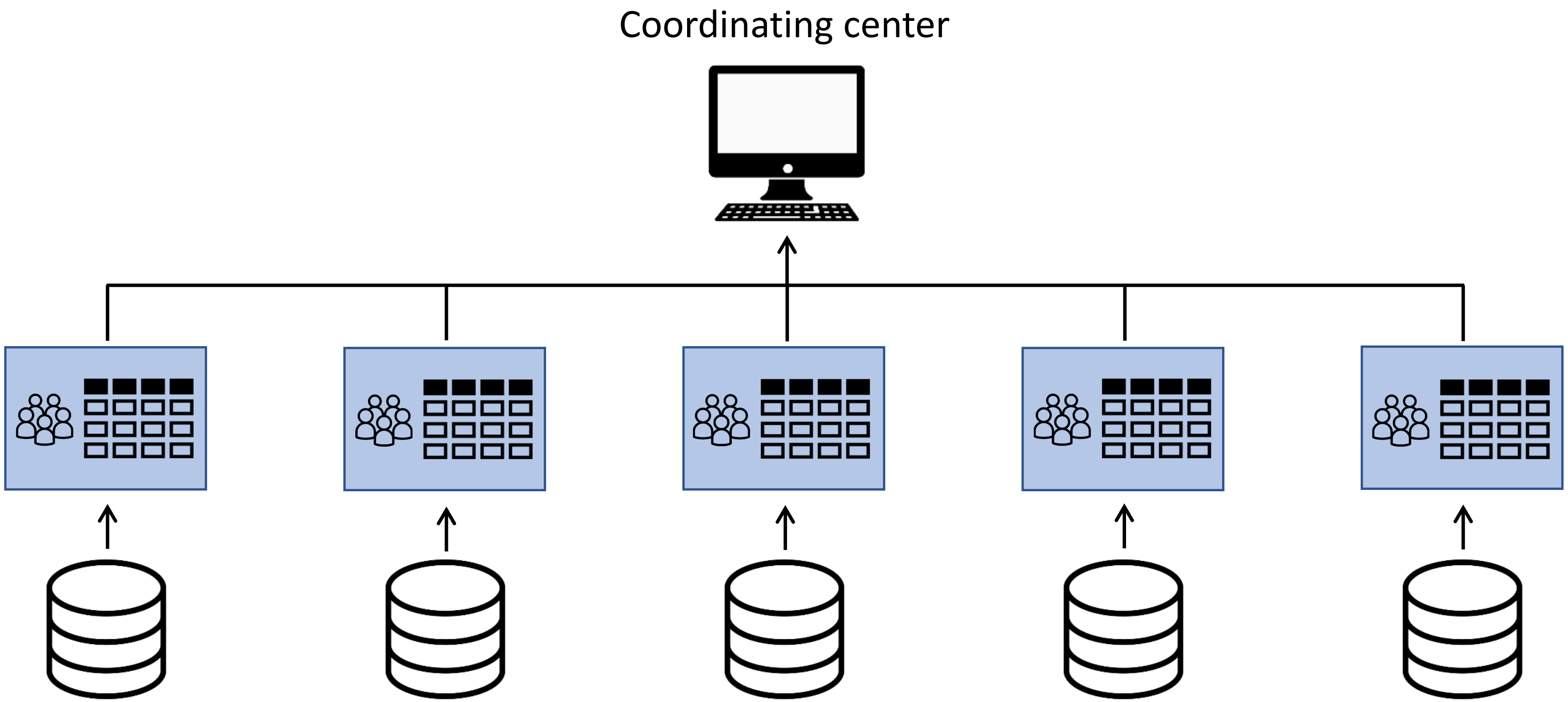
Multi-database studies

- Larger and more diverse populations
  - More precise and generalizable findings
  - Greater capture of rare exposures and outcomes
  - Better suited to investigate heterogenous treatment effects
  - More data for machine learning algorithms

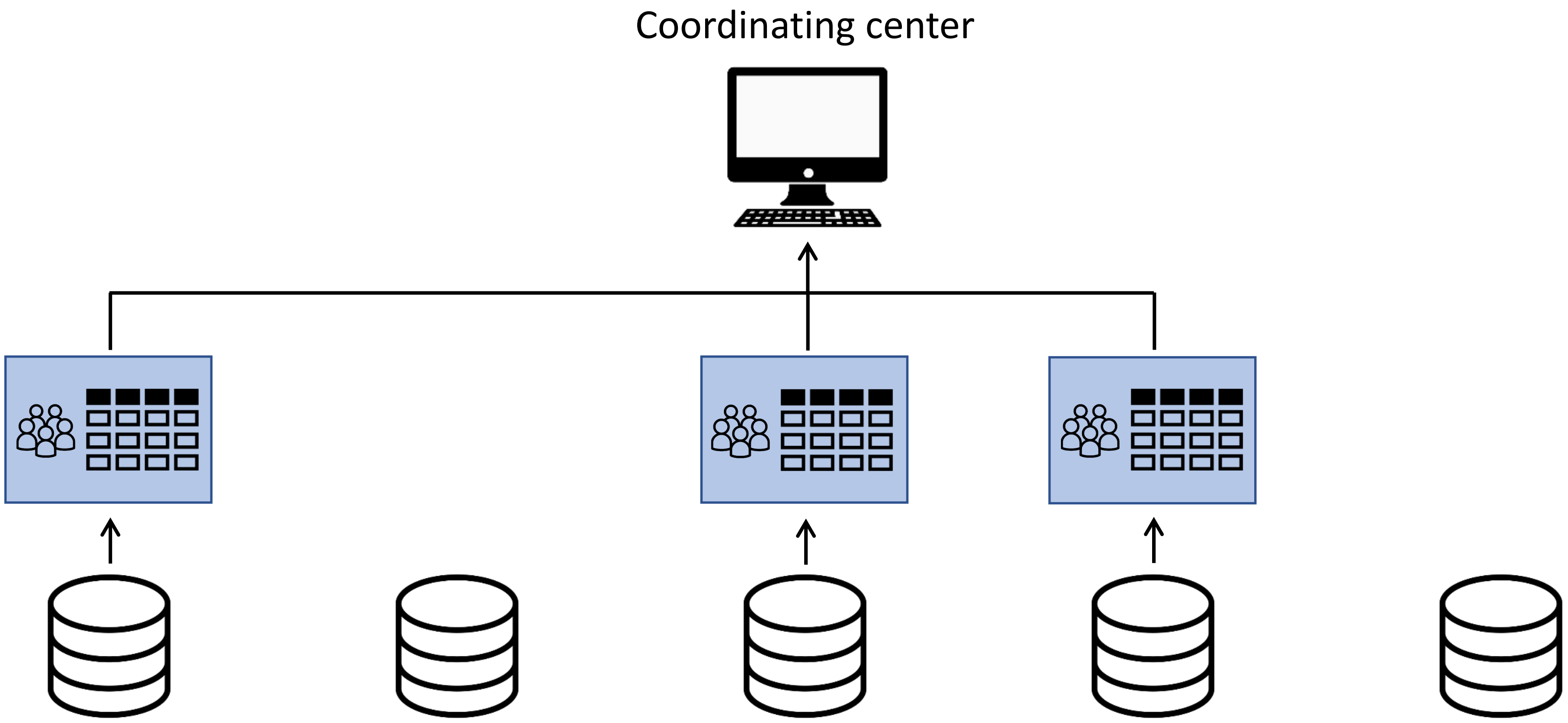
Distributed data networks (DDNs)



Distributed data networks (DDNs)



Distributed data networks (DDNs)

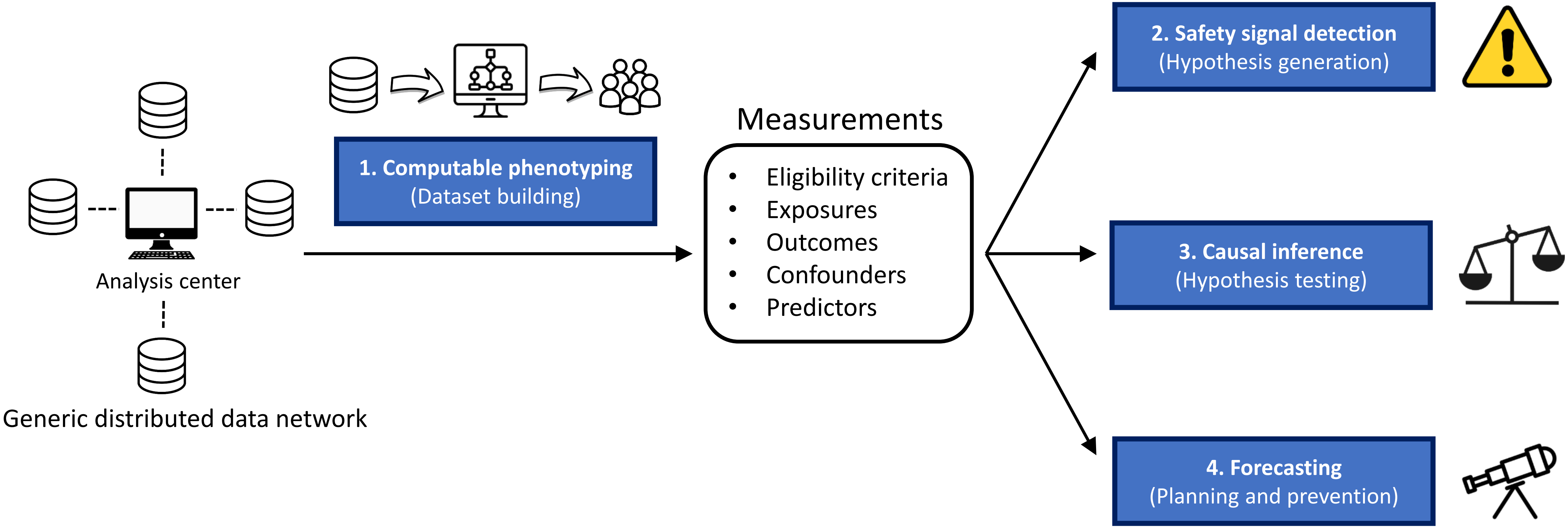


Examples of DDNs that assess the real-world effectiveness and safety of marketed medical products

CNODES = Canadian Network of Observational Drug Effect Studies
PCORnet® = National Patient-Centered Clinical Research Network
OHDSI = Observational Health Data Sciences and Informatics
EHDEN = European Health Data Evidence Network
AsPEN = Asian Pharmacoepidemiology Network

Overview

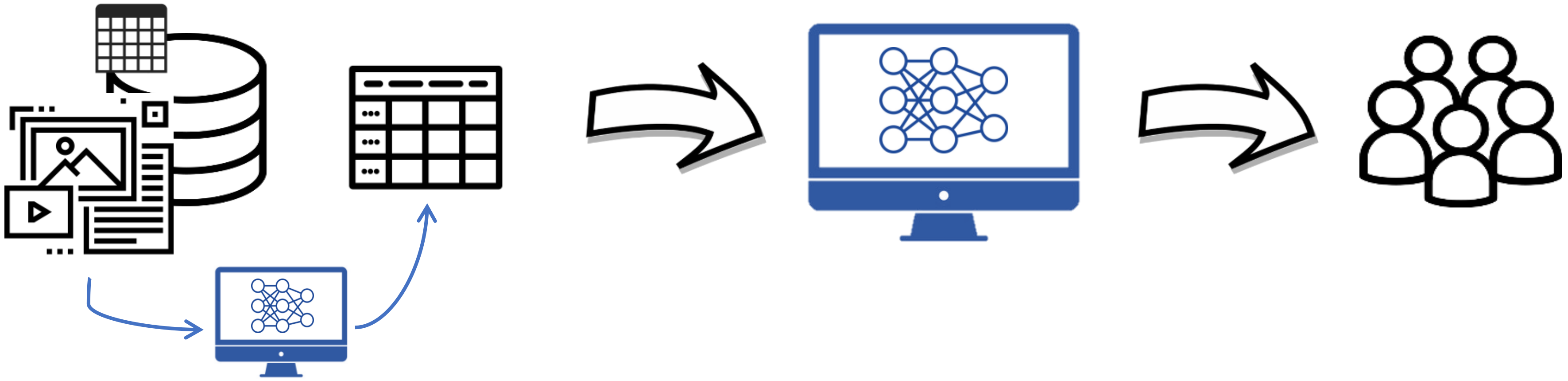
1. Definitions
2. Key activities of distributed data networks
3. Practical aspects of distributed data networks
4. Four scenarios
5. Additional considerations and conclusions



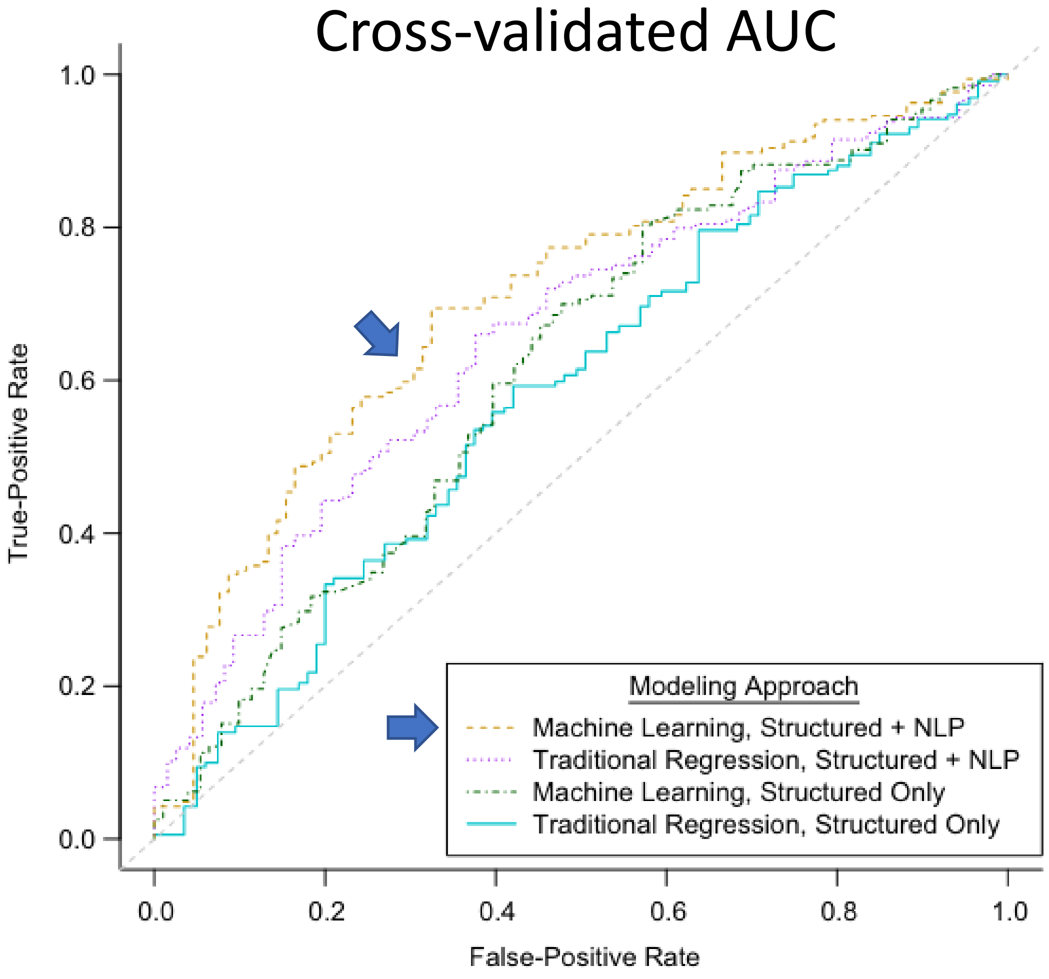
How can machine learning algorithms enhance these activities?

Computable phenotyping

- **Phenotype definition:** select inputs and learn how to map inputs to phenotype status
- **Information extraction:** extract candidate inputs from unstructured data (e.g., text or images)



Identifying anaphylaxis events from EHR data



Carrell *et al.* Am J Epidemiol. 2023;192(2):283

Safety signal detection (SSD)

- **Disproportionality analysis**
  - Bayesian Confidence Propagation Neural Network (BCPNN) to calculate the Information Component[1]
- **Traditional epidemiological designs**
  - General propensity scores to reduce confounding
- **Other innovative designs**
  - E.g., training a random forest to identify drug-outcome pairs that are adverse drug reactions using features reflecting Bradford Hill causality considerations[2]
- **Information extraction**
  - Extract mentions of adverse drug events from clinical text

Methods for SSD using routinely collected healthcare data

Method	Number of papers using the design[a]
Disproportionality analysis	
PRR	9 (17.3%)
ROR	8 (15.4%)
BCPNN	9 (17.3%)
GPS/MGPS	6 (11.5%)
LGPS/LEOPARD	12 (23.1%)
Other	8 (15.4%)
Subtotal	**52 (100.0%)**
Traditional epidemiological designs	
Self-controlled case series	15 (34.1%)
Self-controlled cohort	5 (11.4%)
New-user cohort	5 (11.4%)
Case–control	13 (29.5%)
Case-crossover	3 (6.8%)
Case-population	3 (6.8%)
Subtotal	**44 (100.0%)**
Temporal association	
Temporal pattern discovery	10 (50.0%)
MUTARA/HUNT	6 (30.0%)
Fuzzy-based logic	4 (20.0%)
Subtotal	**20 (100.0%)**
Sequence symmetry analysis	6 (100.0%)
Sequential testing	
MaxSPRT	4 (66.7%)
CSSP	2 (33.3%)
Subtotal	6 (100.0%)
Tree-based scan statistic	9 (100.0%)
Other designs including machine learning	13 (100.0%)
Lab results	9 (100.0%)
Prescription only methods	5 (100.0%)

Coste *et al.* Pharmacoepidemiol Drug Saf. 2023;32(1):28

[1]Zorych *et al.* Stat Methods Med Res. 2013;22(1):39
[2]Reps *et al.* J Biomed Inform. 2015;56:356

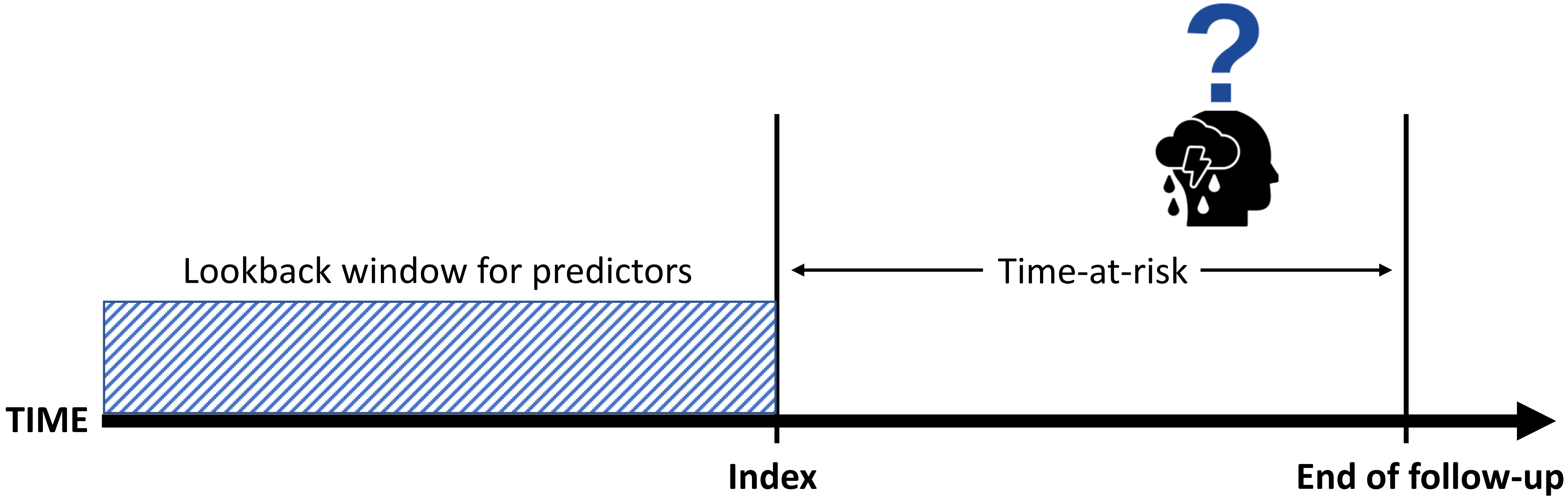
Causal inference

- **High-dimensional confounding adjustment**
  - Estimate “nuisance functions” (e.g., propensity score model and outcome model in targeted maximum likelihood estimation)
  - Prioritize or reduce dimensionality of covariates[1]
- **Information extraction**
  - Extract candidate covariates from unstructured data
- **Counterfactual prediction**
  - Predict potential outcomes for individuals under different treatments[2]

[1]E.g., Weberpals *et al.* Epidemiology. 2021;32(3):378
[2]Feuerriegel *et al.* Nat Med. 2024;30(4):958

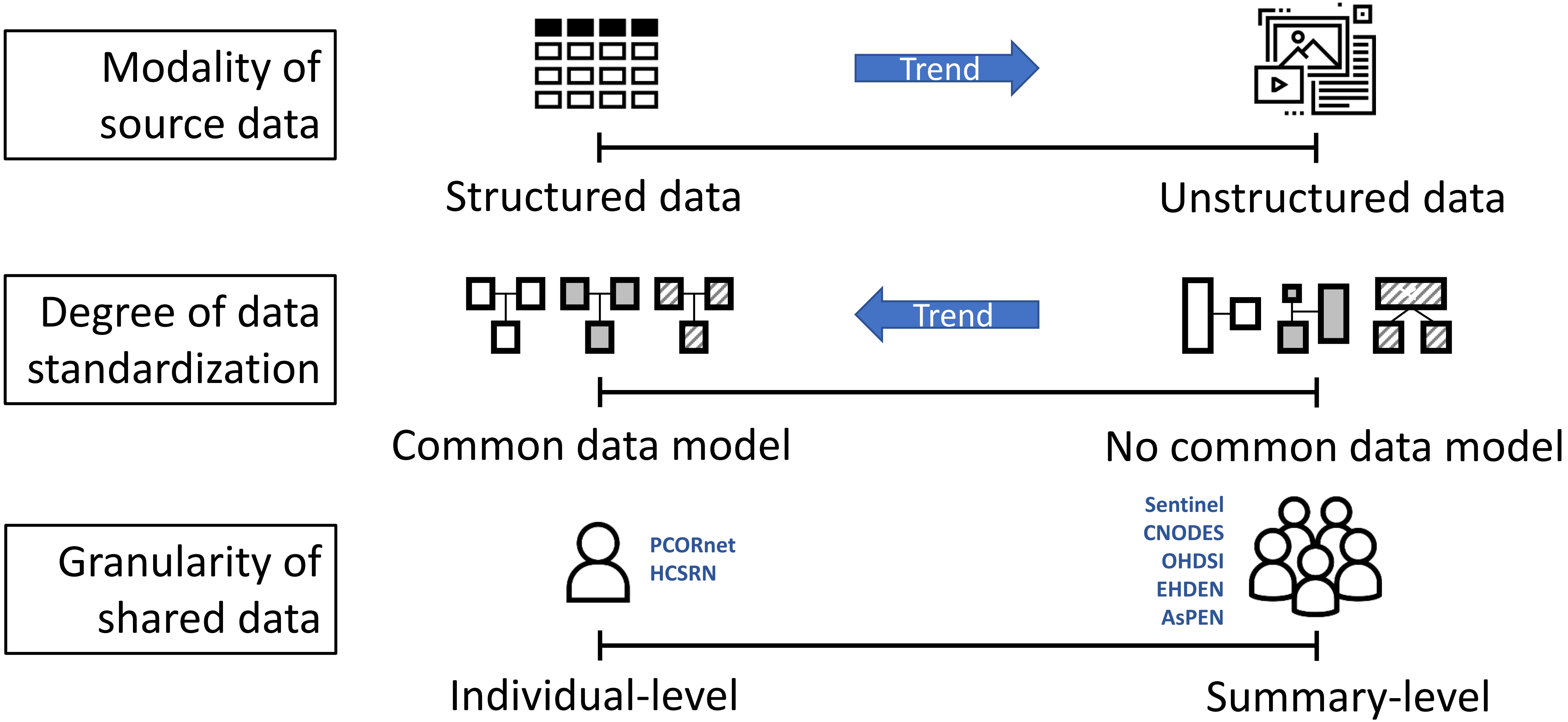
Forecasting

- **Prognostic algorithm:** select predictors and learn how to map predictors to prognosis
- **Information extraction:** extract candidate predictors from unstructured data (e.g., text or images)



Overview

1. Definitions
2. Key activities of distributed data networks
3. Practical aspects of distributed data networks
4. Four scenarios
5. Additional considerations and conclusions


Modality of source data
Trend
Structured data
Unstructured data
Degree of data standardization
Trend
Common data model
No common data model
Granularity of shared data
PCORnet
HCSRN
Sentinel
CNODES
OHDSI
EHDEN
AsPEN
Individual-level
Summary-level

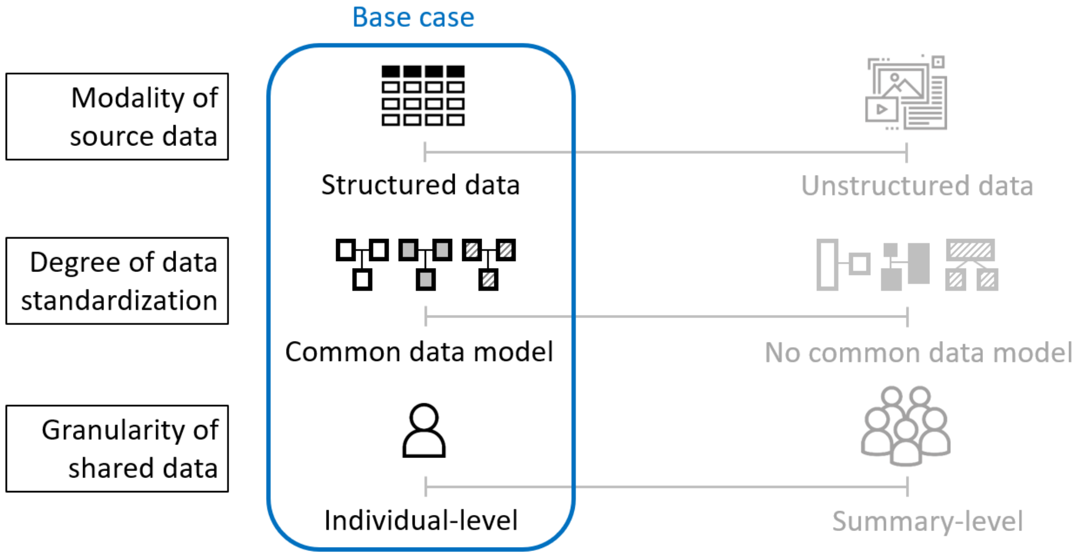
Overview

1. Definitions
2. Key activities of distributed data networks
3. Practical aspects of distributed data networks
4. Four scenarios
5. Additional considerations and conclusions

Scenario	Modality of source data	Degree of data standardization	Granularity of shared data
1 – Base case	Structured data only	Common data model for all inputs	Individual-level data for all sites

Logistically straightforward to apply machine learning in DDNs

Scientifically valid?



Heterogeneity can exist in seemingly similar sites

Adults (≥50y) with any diabetes (2011-2020)

Select Patient Characteristics	KPWA (N=74475)	KPNW (N=64231)
Age, mean (SD)	62.8 (9.95)	62.8 (9.91)
Female	36631 (49%)	31461 (49%)
Insulin use	17184 (23%)	12207 (19%)
Elixhauser comorbidity score, mean (SD)	3.59 (2.34)	3.52 (2.28)
Race		
Unknown	20570 (28%)	4168 (6%)
American Indian or Alaska Native	1300 (2%)	925 (1%)
Asian	5776 (8%)	3661 (6%)
Black or African American	3328 (4%)	2495 (4%)
Native Hawaiian or Other Pacific Islander	773 (1%)	855 (1%)
White	42728 (57%)	52127 (81%)
Number of hospitalizations, mean (SD)	0.190 (0.60)	0.198 (0.62)

- KPNW: greater use of "unspecified" codes
- KPWA: greater use of specific codes

ICD-10 codes related to cataract (phecode 366)

Code	Description	Frequency			Adjusted P-value[b]
		KPWA	KPNW	Ratio[a]	
	Any ICD-10 code related to cataract	75535	68658	1.03	
E08.36	Diabetes mellitus due to underlying condition with diabetic cataract	23	0	3.10	6.12x10^-03
E10.36	Type 1 diabetes mellitus with diabetic cataract	92	117	0.75	6.06x10^-02
E11.36	Type 2 diabetes mellitus with diabetic cataract	3065	2996	0.96	5.88x10^-01
H26.40	Unspecified secondary cataract	561	1144	0.46	1.62x10^-39
H26.411	Soemmering's ring, right eye	11	1	1.79	1.26x10^-01
H26.491	Other secondary cataract, right eye	3044	771	3.67	<10^-100
H26.492	Other secondary cataract, left eye	3129	741	3.93	<10^-100
H26.493	Other secondary cataract, bilateral	3952	636	5.76	<10^-100
H26.499	Other secondary cataract, unspecified eye	70	0	7.51	1.55x10^-14
H26.8	Other specified cataract	526	1323	0.38	5.22x10^-27
H26.9	Unspecified cataract	16704	15786	0.99	8.53x10^-01
H59.021	Cataract (lens) fragments in eye following cataract surgery, right eye	47	14	2.23	1.31x10^-01
H59.022	Cataract (lens) fragments in eye following cataract surgery, left eye	78	10	4.13	1.03x10^-02
H59.029	Cataract (lens) fragments in eye following cataract surgery, unspecified eye	1	72	0.13	1.15x10^-06
Z96.1	Presence of intraocular lens	35888	44526	0.76	1.31x10^-79
Z98.41	Cataract extraction status, right eye	3950	199	17.79	<10^-100
Z98.42	Cataract extraction status, left eye	3723	195	17.10	<10^-100
Z98.49	Cataract extraction status, unspecified eye	622	112	4.87	1.15x10^-33

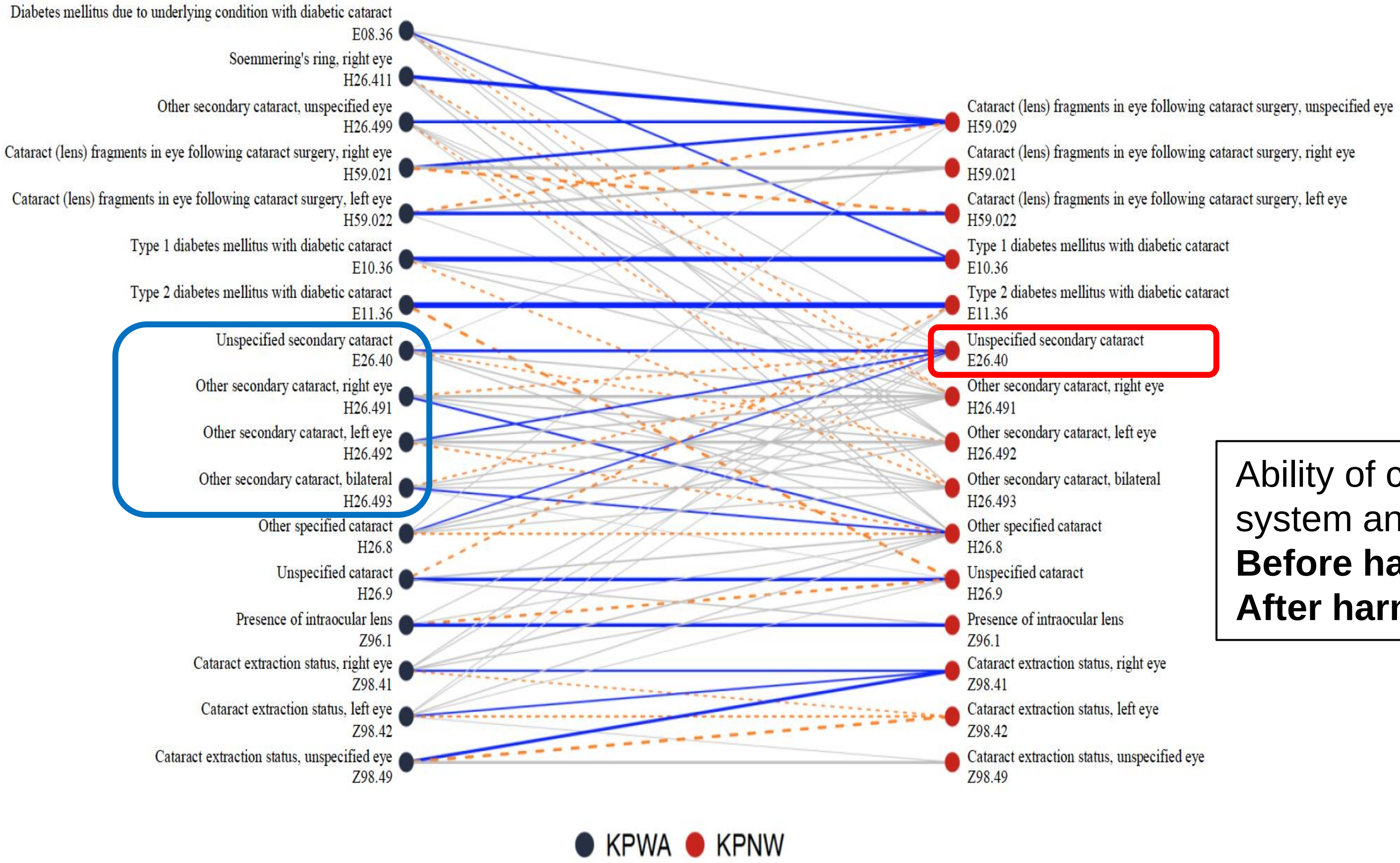
[a]Frequency ratio defined as *(frequency in KPWA + 10)/patient yrs in KPWA* divided by *(frequency in KPNW + 10)/patient yrs in KPNW*; where ratio>1 indicates stronger code endorsement at KPWA and ratio<1 indicates stronger code endorsement at KPNW.
[b]P-value from t-test, adjusted for person-time and baseline patient characteristics (age, sex, insulin, and Elixhauser index)

Source: Sentinel Innovation Center Methods Project (Shi et al., unpublished results)
KPWA = Kaiser Permanente Washington; KPNW = Kaiser Permanente Northwest

Approaches to reduce heterogeneity

- **Approach 1: Fit site-specific models**
- **Approach 2: “Harmonize” the input data**

Unsupervised learning to reduce heterogeneity



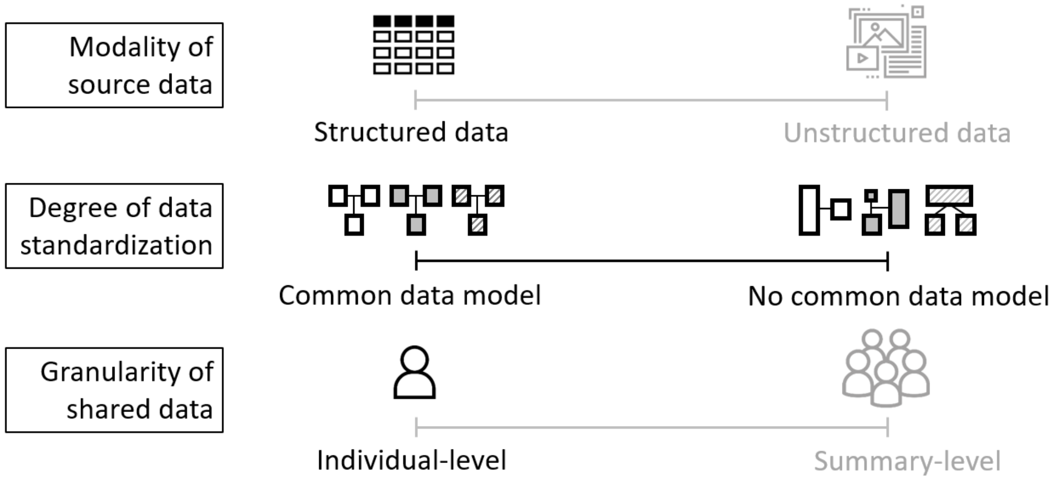
Ability of cataract codes to predict which system an individual was from:
Before harmonization: cv-AUC of 0.72
After harmonization: cv-AUC of 0.59

Direction of harmonization (KPWA to KPNW)

Blue lines = top mapping (code at KPNW with largest similarity)
Orange dashed lines = 2nd top mapping (code at KPNW with 2nd largest similarity)

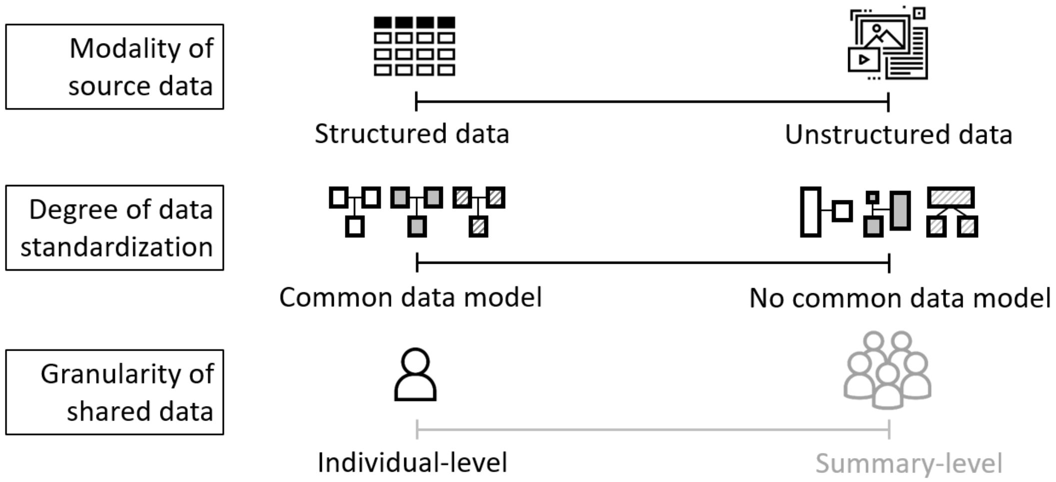
Source: Sentinel Innovation Center Methods Project (Shi et al., unpublished results)
KPWA = Kaiser Permanente Washington; KPNW = Kaiser Permanente Northwest

Scenario	Modality of source data	Degree of data standardization	Granularity of shared data
1 – Base case	Structured data only	Common data model for all inputs	Individual-level data for all sites
2 – Less standardized data available	Structured data only	No common data model for some inputs	Individual-level data for all sites



Scenario	Modality of source data	Degree of data standardization	Granularity of shared data
1 – Base case	Structured data only	Common data model for all inputs	Individual-level data for all sites
2 – Less standardized data available	Structured data only	No common data model for some inputs	Individual-level data for all sites
3 – More complex data modalities used	Structured and unstructured data	No common data model for some inputs	Individual-level data for all sites

Creates challenges for feature engineering



Unavailable structured and unstructured clinical data in the Sentinel Common Data Model were among the top reasons new drug safety concerns could not be evaluated in the FDA's Active Risk Identification and Analysis (ARIA) system.

Maro *et al.* Clin Pharmacol Ther. 2023;114(4):815

Table 4 Reasons for determinations of ARIA insufficiency

Reasons for insufficiency	Number of determinations	Example	Direction of future development
Insufficient supplemental structured clinical data	89	Lack of laboratory, imaging, or vital signs data	Addressable with the addition of EHR data elements into ARIA[35,36]
Inability of ARIA tools to perform required analysis	82	Insufficient signal identification tool	ARIA has integrated signal identification abilities (**Figure 1**)[16–18]
Study requires data elements captured in unstructured clinical data, such as clinical notes	73	Lack of radiology or pathology findings in notes	Addressable with development of feature engineering capabilities to extract and structure these data[37]
Absence of validated code algorithm	72	No gold-standard chart review was performed for outcome of interest	Sentinel has performed several gold standard chart validations[38–42] but these require substantial resources. Efforts underway to investigate rapid silver standard reviews.
Identification of clinical concepts with available code algorithms/terminologies is not possible or inadequate	60	Codes do not exist for concept or validated performance characteristics are inadequate	Potentially addressable with added EHR elements but if outcome is not well-defined or new (e.g., long COVID), there may be substantial hurdles to identification
Inadequate sample size	57	Low uptake of drug	Non-actionable as ARIA is the largest system of its kind
Requires linkage to additional data source that is unavailable	52	Inability to ascertain cause of death	Additional linkages are possible with significant financial resources
Insufficient observation time available	44	Inability to follow patients across healthcare plans or systems	Actionable with substantial further research and development and resolution of data governance issues[43]
Insufficient mother-infant linkage	24	Lack of ability to connect mothers and infants	Resolved with 2018 integration of Mother-Infant Linkage table[15]
Insufficient inpatient data	18	Inability to access granular inpatient pharmacy information	Resolved with partnerships with inpatient healthcare systems[10]
Inability to identify over-the-counter medication use	8	Over-the-counter medication use not captured	Inherent limitation of both claims and EHR data
Insufficient race capture of information on race	3	Race is not well-captured	FDA is working with Data Partners to understand approaches for better capture of this data
Insufficient representation of the population of interest	1	Limited generalizability based on commercial claims data	Sentinel added Medicare data in 2018 and Medicaid in 2022

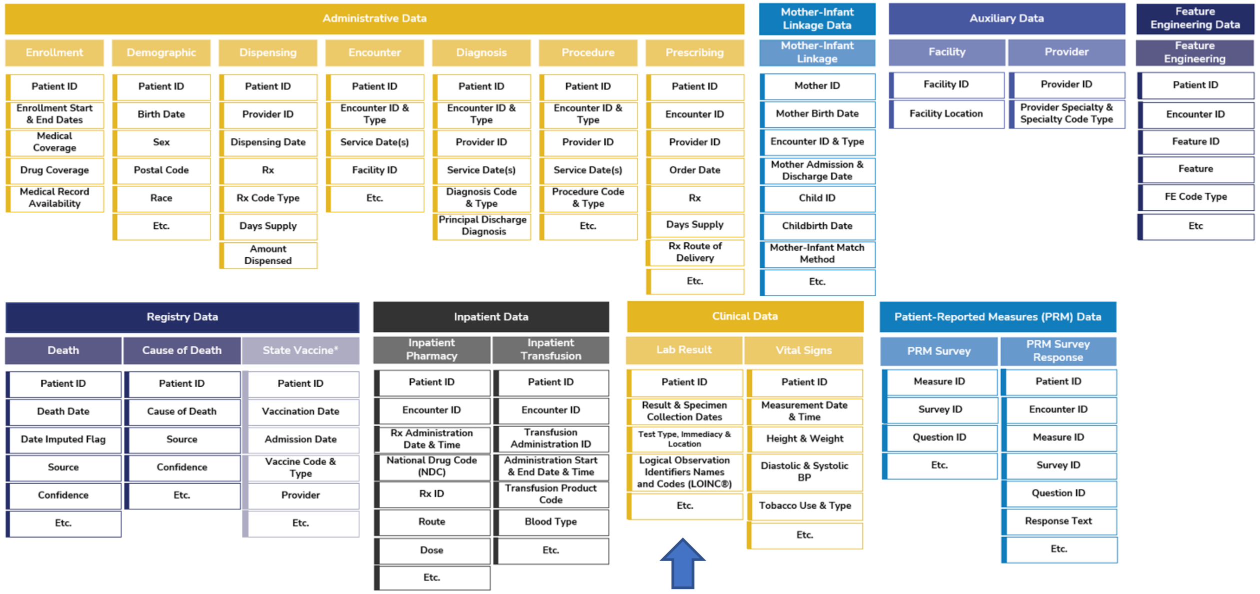
ARIA, Active Risk Identification and Analysis; COVID, coronavirus disease; EHR, electronic health record; FDA, US Food and Drug Administration.

When desired information is outside the CDM

- **Approach 1: Standardize the unstandardized information**
  - Invest time and resources upfront
  - Some considerations:
    - How easily can the information be added?
    - How frequently will the information be used?
    - How urgently is the information required?

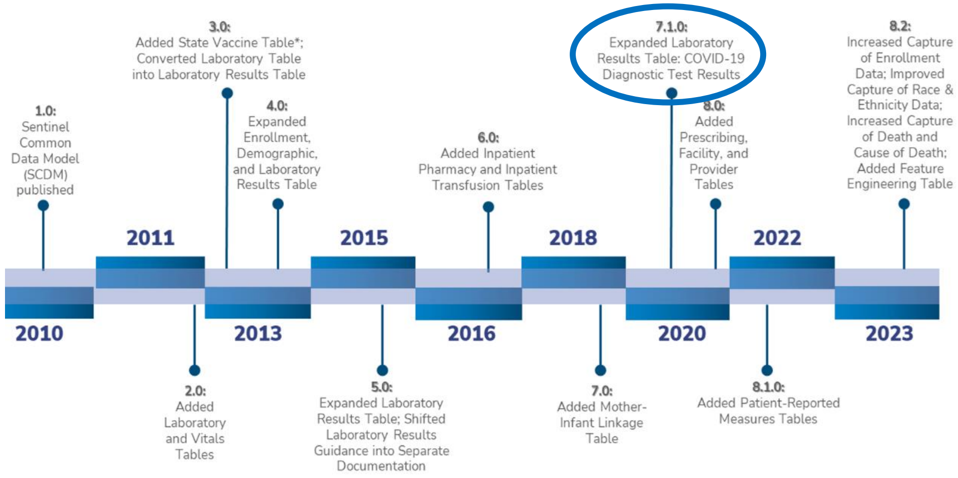
The Sentinel Common Data Model over time

Latest version (SCDM v8.2.0)



https://www.sentinelinitiative.org/methods-data-tools/sentinel-common-data-model

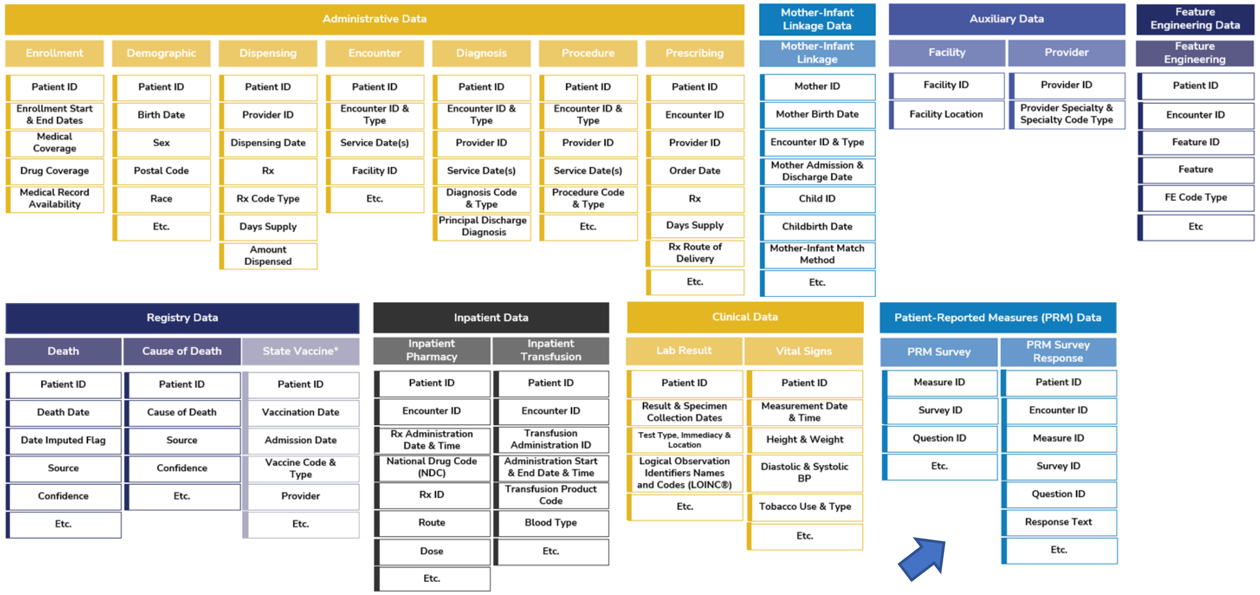
*The State Vaccine table has not been in use since SCDM v6.0.



https://www.sentinelinitiative.org/methods-data-tools/sentinel-common-data-model#enhancements-to-sentinel-common-data-model

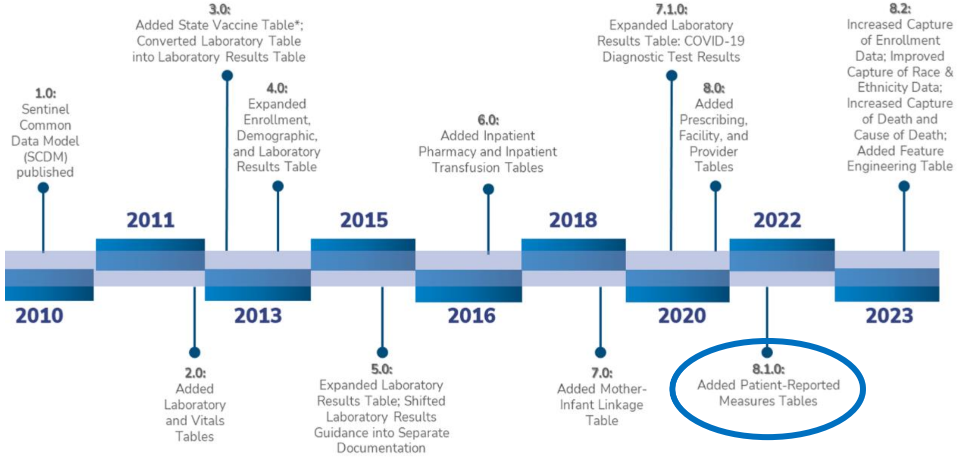
The Sentinel Common Data Model over time

Latest version (SCDM v8.2.0)



https://www.sentinelinitiative.org/methods-data-tools/sentinel-common-data-model

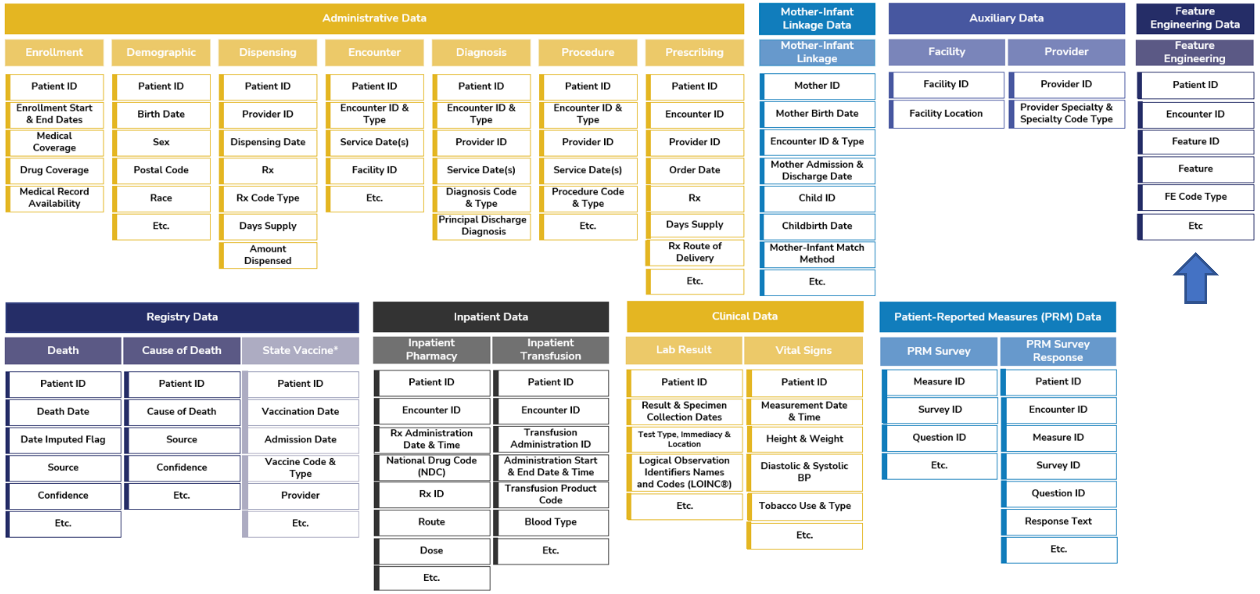
*The State Vaccine table has not been in use since SCDM v6.0.

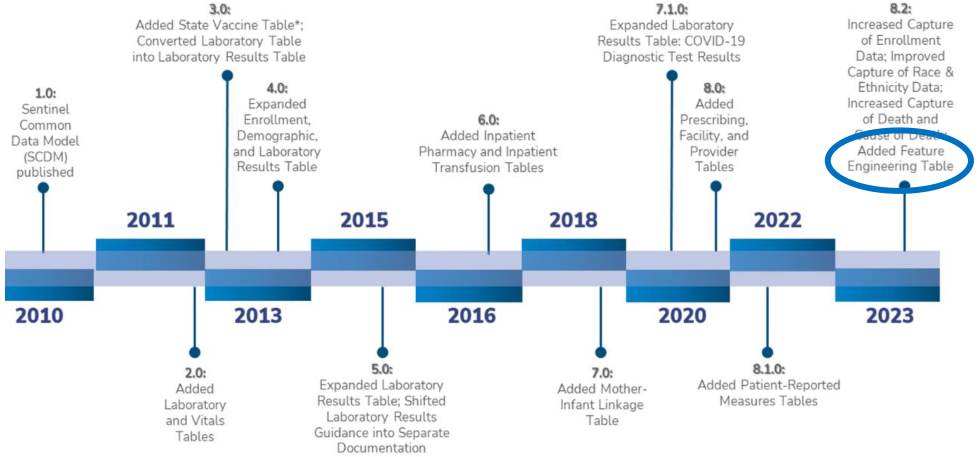


https://www.sentinelinitiative.org/methods-data-tools/sentinel-common-data-model#enhancements-to-sentinel-common-data-model

The Sentinel Common Data Model over time

Latest version (SCDM v8.2.0)





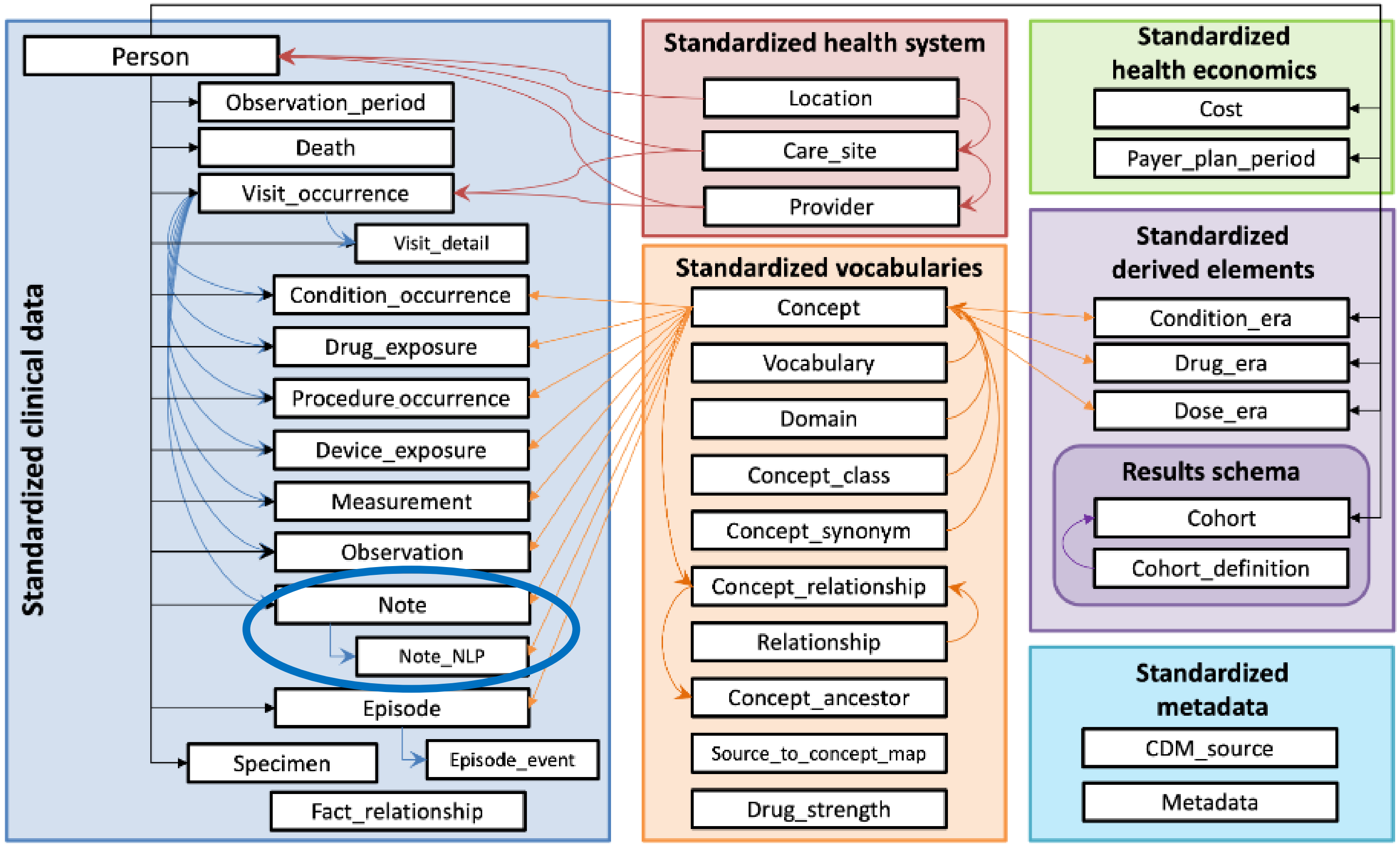
https://www.sentinelinitiative.org/methods-data-tools/sentinel-common-data-model

*The State Vaccine table has not been in use since SCDM v6.0.

https://www.sentinelinitiative.org/methods-data-tools/sentinel-common-data-model#enhancements-to-sentinel-common-data-model

Observational Medical Outcomes Partnership (OMOP) Common Data Model

Latest version (OMOP CDM v5.4)

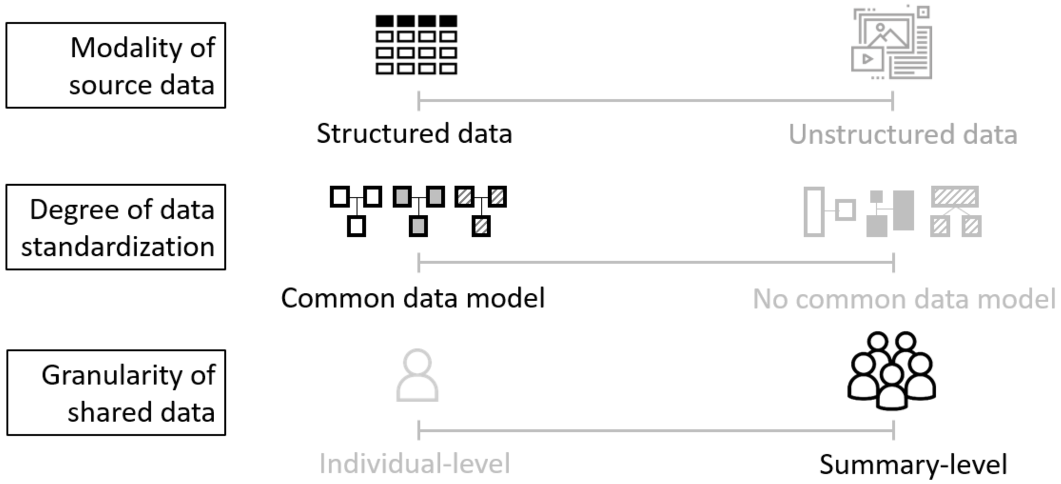


https://ohdsi.github.io/CommonDataModel/

When desired information is outside the CDM

- **Approach 2: Do a site-specific analysis** (using a common protocol)
  - May be especially preferred when:
    - Desired information captured only at some sites
    - Added value of desired information for the model is uncertain

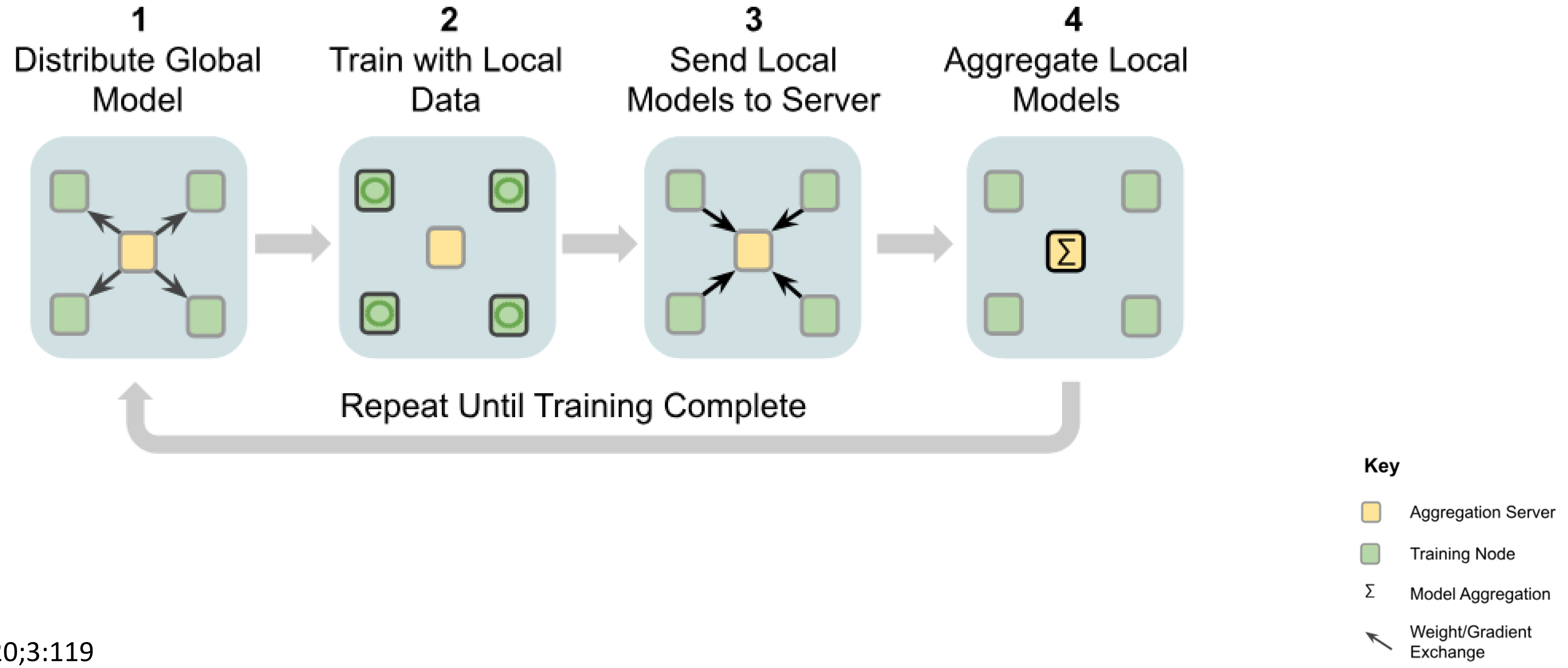
Scenario	Modality of source data	Degree of data standardization	Granularity of shared data
1 – Base case	Structured data only	Common data model for all inputs	Individual-level data for all sites
2 – Less standardized data available	Structured data only	No common data model for some inputs	Individual-level data for all sites
3 – More complex data modalities used	Structured and unstructured data	No common data model for some inputs	Individual-level data for all sites
4 – Less granular data shared	Structured data only	Common data model for all inputs	Summary-level data for all sites



Impacts how machine learning models can be trained

Training models with only summary-level data

- **Approach 1: Collaboratively train a global model** (federated learning)

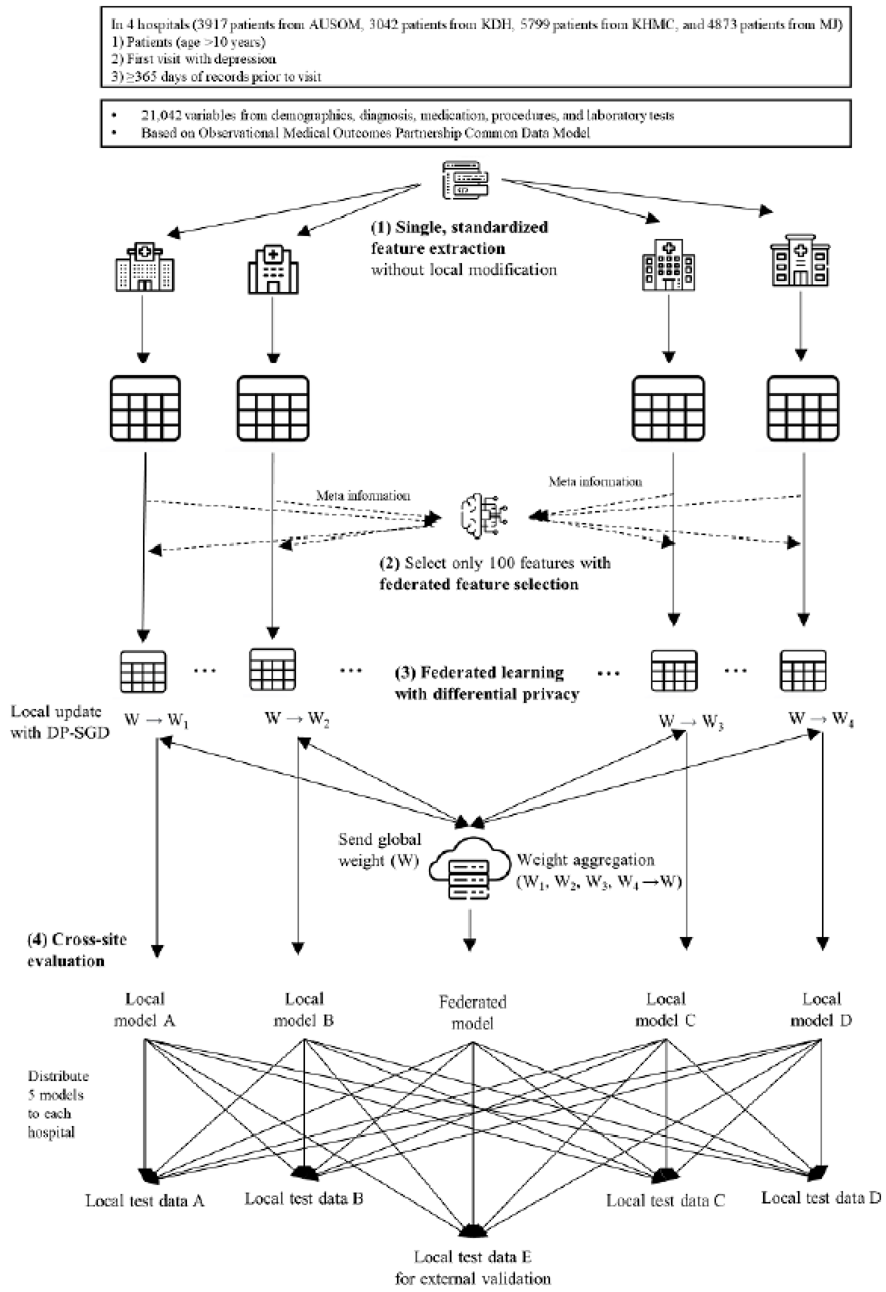


Rieke *et al.* NPJ Digit Med. 2020;3:119

Training models with only summary-level data

- **Approach 1: Collaboratively train a global model** (federated learning)

Advantages	Disadvantages
• Train more robust and generalizable models by using data from multiple sites	• Privacy leakage concerns • Coordination and implementation challenges (e.g., hardware and infrastructure requirements, communication costs) • Global model may not converge or perform well if data across sites are too heterogeneous

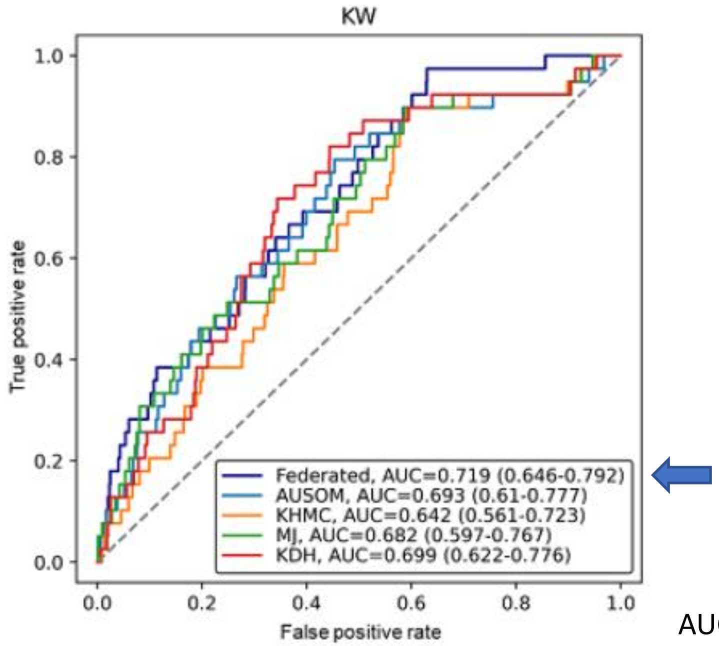


Lee *et al.* J Med Internet Res. 2023 Jul 20;25:e46165.

AUC of federated vs local models in test sets (Cross-site evaluation)

	Test set				
	AUSOM	KHMC	MJ	KDH	Mean
Federated	0.819 (0.74-0.897)	0.731 (0.584-0.931)	0.707 (0.618-0.794)	0.649 (0.497-0.801)	0.726
Local models: AUSOM	0.816 (0.742-0.89)	0.65 (0.447-0.853)	0.579 (0.472-0.686)	0.524 (0.355-0.692)	0.642
KHMC	0.663 (0.473-0.854)	0.736 (0.55-0.923)	0.641 (0.535-0.747)	0.606 (0.452-0.761)	0.662
MJ	0.766 (0.656-0.875)	0.732 (0.571-0.893)	0.715 (0.634-0.795)	0.614 (0.421-0.807)	0.707
KDH	0.811 (0.685-0.937)	0.654 (0.453-0.855)	0.598 (0.487-0.709)	0.705 (0.497-0.912)	0.692

AUC of federated and local models in an external database



AUC = area under the curve

Training models with only summary-level data

- **Approach 2:** Train a local model independently at each site

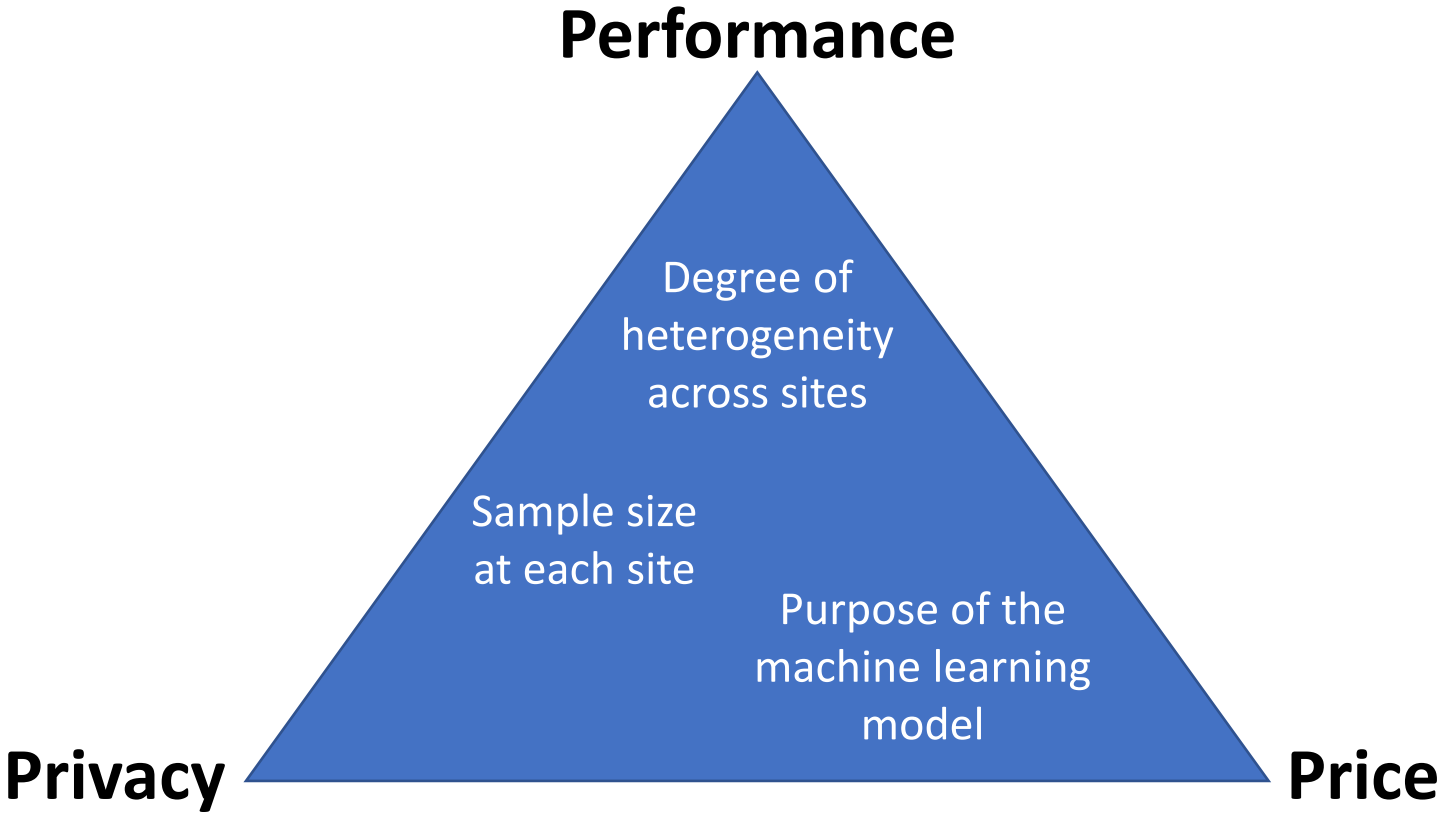
Advantages	Disadvantages
• Can be easily externally validated in other sites[1] • Do not have to use the same inputs as other sites • Transportability of local models can be improved using simpler federated learning approaches[2]	• Does not harness the full potential of the network to train more robust and generalizable models

[1]Reps *et al.* BMC Med Res Methodol. 2020;20(1):102.
[2]Reps *et al.* BMC Med Inform Decis Mak. 2022;22(1):142

Overview

1. Definitions
2. Key activities of distributed data networks
3. Practical aspects of distributed data networks
4. Four scenarios
5. Additional considerations and conclusions

Choice of approach is a balancing act



Other benefits of machine learning in DDNs

Issue	Single database	Distributed data network
Generalizability	External validation of models is rare and slow	External validation of models can be done more quickly and easily
Transparency	Less impetus to document finer-grain details	High transparency required to enable data partners to replicate process
Interpretability	Less impetus to interpret and explain model outputs	Unusual or discrepant results across data partners require ability to interpret and explain model outputs

Conclusions

- **Many opportunities** exist for machine learning to enhance the activities of DDNs for post-market medical product surveillance.
- The diverse and siloed storage of data in DDNs create **unique challenges** for applying machine learning.
- **Various approaches** can be considered to address these challenges.
- Rapid rise of **LLMs and generative AI** may accelerate the ability of DDNs to address some challenges (e.g., incorporate information from unstructured data into the CDM), but may also raise new challenges and considerations.
- Machine learning will continue to play an important role in **advancing the capabilities of DDNs** for post-market surveillance in the years to come.

- **Many opportunities** exist for machine learning to enhance the activities of DDNs for post-market medical product surveillance.
- The diverse and siloed storage of data in DDNs create **unique challenges** for applying machine learning.
- **Various approaches** can be considered to address these challenges.
- Rapid rise of **LLMs and generative AI** may accelerate the ability of DDNs to address some challenges (e.g., incorporate information from unstructured data into the CDM), but may also raise new challenges and considerations.
- Machine learning will continue to play an important role in **advancing the capabilities of DDNs** for post-market surveillance in the years to come.

Questions?

jenna_wong@hphci.harvard.edu

DEPARTMENT OF POPULATION MEDICINE

Harvard Pilgrim Health Care Institute