

MACHINE LEARNING FOR MISSING DATA IMPUTATION IN HEALTHCARE RESEARCH: A SYSTEMATIC REVIEW OF METHODS AND APPLICATIONS

AUTHORS

T Porte¹, M Swital¹, N Sedmak¹, C Bouvard¹, A Gougeon^{1,2},
F Roux¹, F Mistretta¹, A Lajoinie¹

AFFILIATIONS

¹ **RCTs**, 38 rue du Plat, 69002 Lyon, France

² **LBBE** (Laboratoire de Biométrie et Biologie Evolutive)
UMR 5558, CNRS, Université Lyon 1, Université de Lyon,
Villeurbanne, France

CONTEXT AND OBJECTIVE

Missing data is a critical issue and a potential source of bias in clinical research, particularly in real-world data (RWD) studies where loss to follow-up and incomplete data are common. Imputing missing data is a significant challenge as it directly affects the validity and reliability of clinical analyses. This literature review aimed to provide an overview of machine learning (ML) imputation methods applied in healthcare and report their performance.

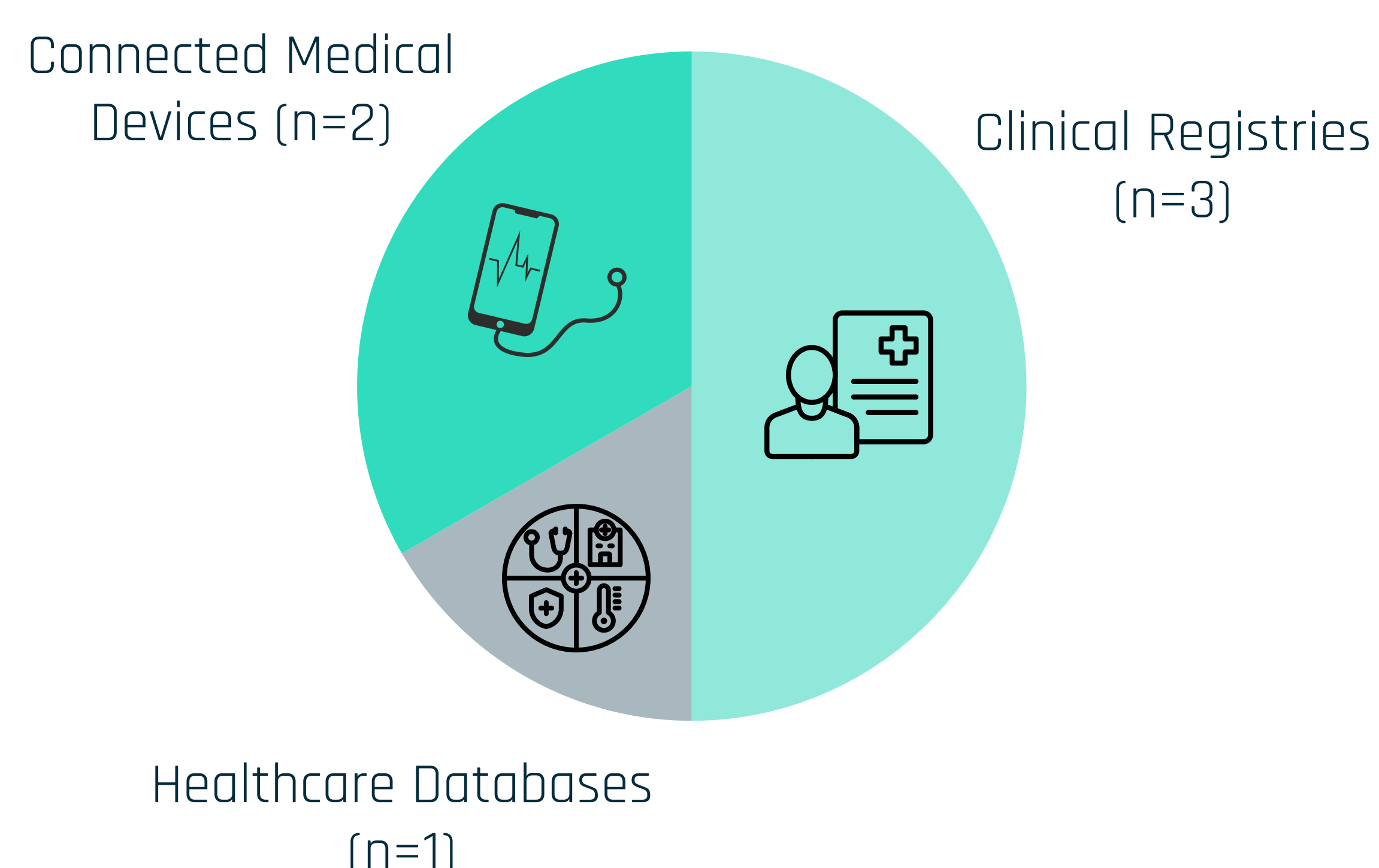
METHODOLOGY

A literature review was conducted on MEDLINE to identify studies published since 2020 on ML-based imputation methods in studies conducted on RWD. Titles and abstracts [Ti/Abs] were screened, followed by full-text review for inclusion.

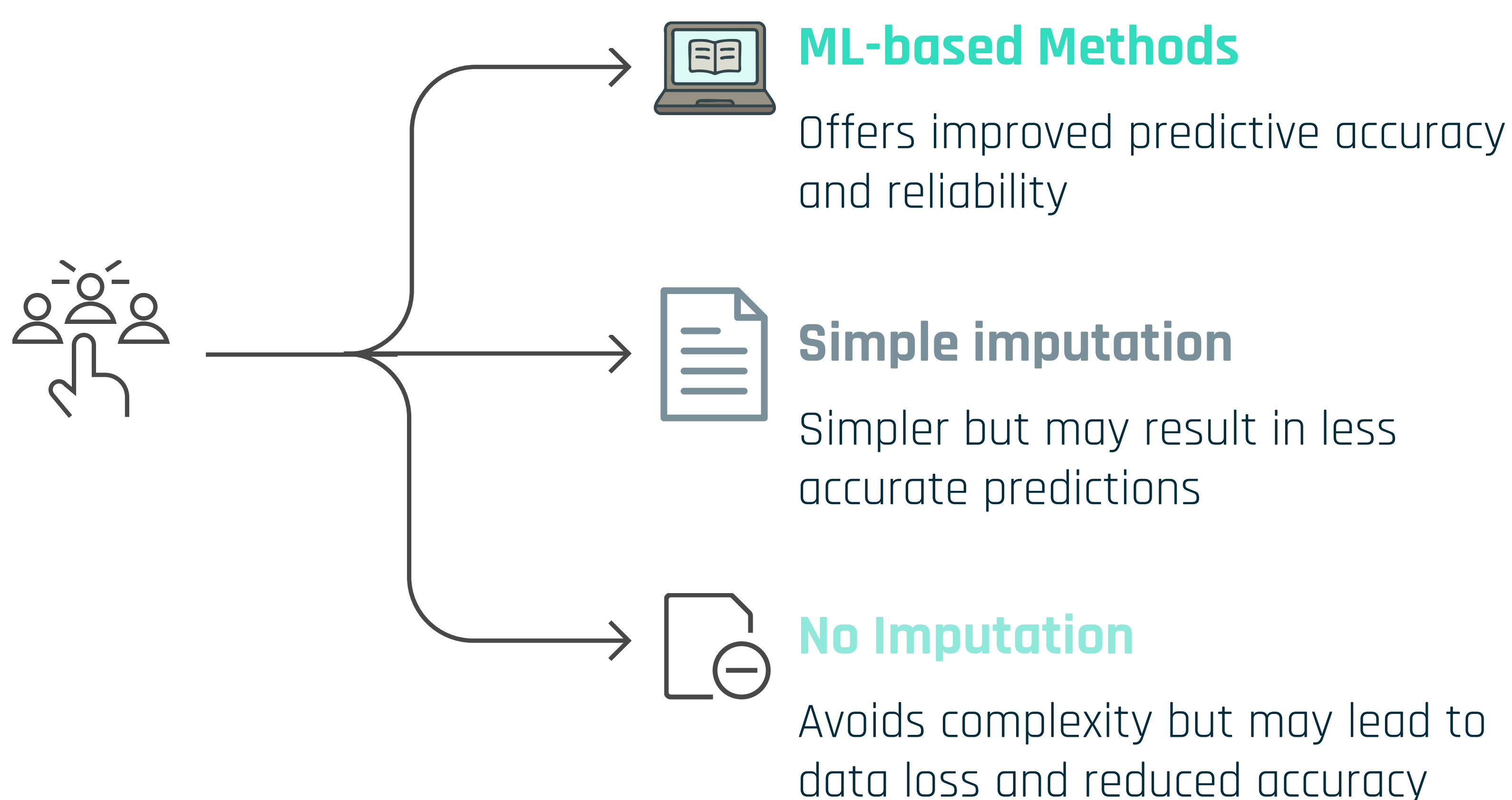
RESULTS/FINDINGS

A total of 166 studies were initially selected through title and abstract screening. After full-text review, 6 studies were included.

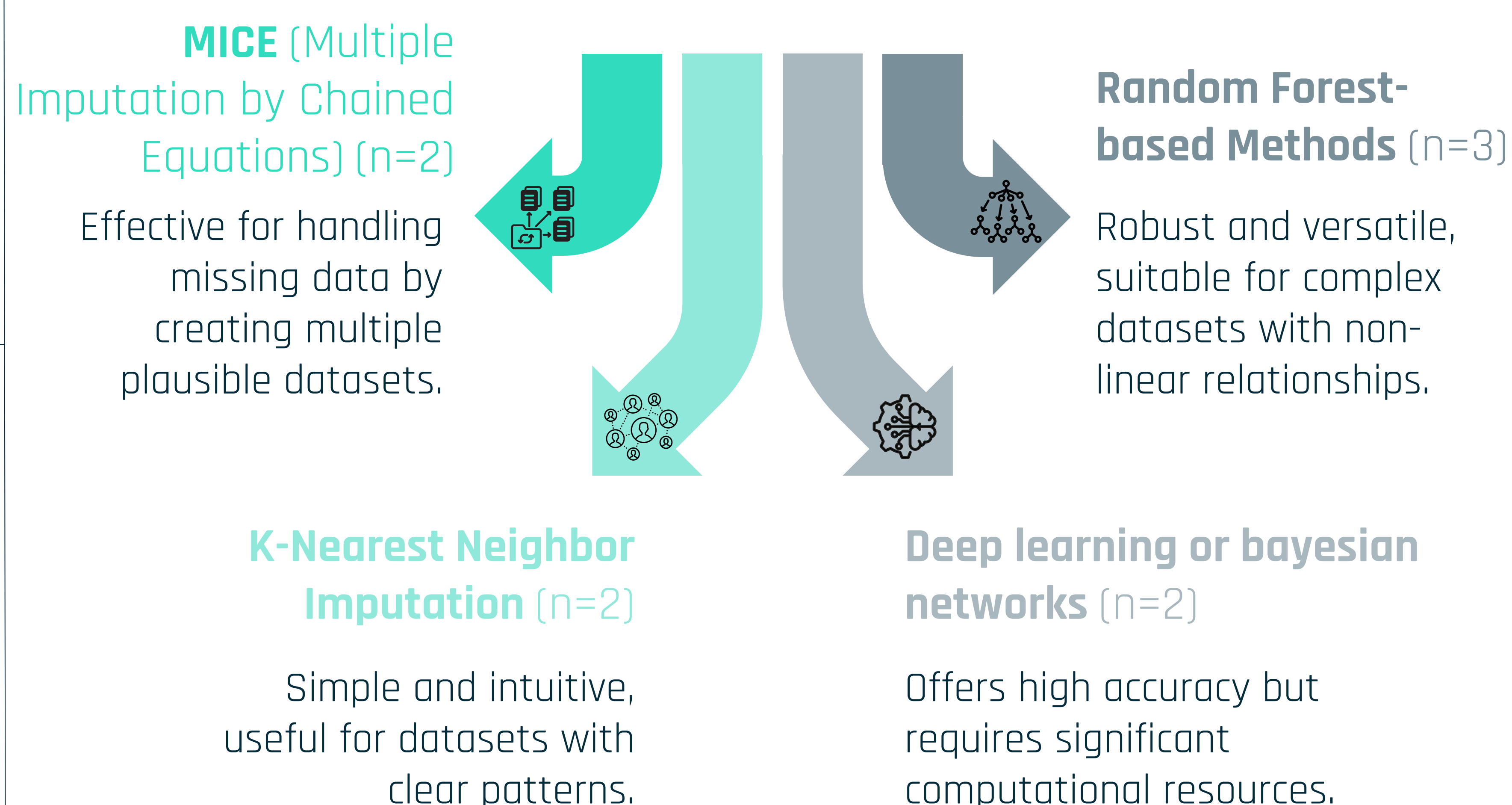
Distribution of Data Sources



Which imputation method can be used for predictive modeling?



Which models to input missing data?



CONCLUSION

Using ML methods for missing data imputation is a promising approach to improving the performance and robustness of predictive models in healthcare. However, the reviewed studies highlight remaining challenges, particularly in cases of high missingness. This warrants cautious interpretation and further methodological refinement.

REFERENCES

Xiong et al., 2024; Kong et al., 2023; Dong et al., 2021; El Badisy et al., 2024; Pavelchek et al., 2023; Gabriel-Vascillica et al., 2024