Practical Solutions for Working with Electronic Health Records Data – References

EHR Data Structure and Quality

Denny JC. Mining electronic health records in the genomics era. *PLoS computational biology*. 2012;8(12):e1002823.

Forrest, C, Margolis, P, Bailey, C, Marsolo, K, Beccaro, M, Finkelstein, J, Milov, D, Vieland, V, Wolf, B, Yu, F, and Kahn, M., PEDSnet: A National Pediatric Learning Health System, *Journal of the American Medical Informatics Association.* 2014; 21: 602-606.

Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*. 2012;13(6):395-405.

Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, Pacheco JA, Tromp G, Pathak J, Carrell DS, Ellis SB. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *Journal of the American Medical Informatics Association.* 2016;23(6):1046-52.

Meyer D, Hornik K, Feinerer I. Text mining infrastructure in R. *Journal of Statistical Software*. 2008;25(5):1-54.

Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, Xu H. CLAMP–a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association*. 2018;25(3):331-6.

Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association*. 2013;20(1):144-51.

Missing Data and Measurement Error in EHR

Carrigan G, Whipple S, Taylor MD, Torres AZ, Gossai A, Arnieri B, Tucker M, Hofmeister PP, Lambert P, Griffith SD, Capra WB. An evaluation of the impact of missing deaths on overall survival analyses of advanced non–small cell lung cancer patients conducted in an electronic health records database. *Pharmacoepidemiology and Drug Safety*. 2019;doi:10.1002/pds.4758

Chen Y, Wang J, Chubak J, Hubbard RA. Inflation of type I error rates due to differential misclassification in EHR-derived outcomes: Empirical illustration using breast cancer recurrence. *Pharmacoepidemiology and Drug Safety*. 2019;28(2):264-8.

Chubak J, Onega T, Zhu W, Buist DS, Hubbard RA. An Electronic Health Record–based Algorithm to Ascertain the Date of Second Breast Cancer Events. *Medical Care*. 2017;55(12):e81-7.

Duan R, Cao M, Wu Y, Huang J, Denny JC, Xu H, Chen Y. An empirical study for impacts of measurement errors on EHR based association studies. *In AMIA Annual Symposium Proceedings* 2016 (Vol. 2016, p. 1764). American Medical Informatics Association.

Haneuse S, Daniels M. A general framework for considering selection bias in EHR-based studies: what data are observed and why? *eGEMs*. 2016;4(1):16.

Lin KJ, Singer DE, Glynn RJ, Murphy SN, Lii J, Schneeweiss S. Identifying patients with high data completeness to improve validity of comparative effectiveness research in electronic health records data. *Clinical Pharmacology & Therapeutics*. 2018;103(5):899-905.

Methods for Informative Observation Processes

Bůžková P, Lumley T. Semiparametric modeling of repeated measurements under outcome-dependent follow-up. *Statistics in Medicine*. 2009;28(6):987-1003.

Goldstein BA, Bhavsar NA, Phelan M, Pencina MJ. Controlling for informed presence bias due to the number of health encounters in an electronic health record. *American Journal of Epidemiology*. 2016;184(11):847-55.

Lange JM, Hubbard RA, Inoue LY, Minin VN. A joint model for multistate disease processes and random informative observation times, with applications to electronic medical records data. *Biometrics*. 2015;71(1):90-101.

Lin DY, Ying Z. Semiparametric and nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*. 2001;96(453):103-26.

Methods to Account for Phenotyping Error and Measurement Error in EHR

Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. Measurement error in nonlinear models: a modern perspective. *CRC press*; 2006.

Huang J, Duan R, Hubbard RA, Wu Y, Moore JH, Xu H, Chen Y. PIE: A prior knowledge guided integrated likelihood estimation method for bias reduction in association studies using electronic health records data. *Journal of the American Medical Informatics Association*. 2018;25(3): 345-352.

Hubbard RA, Harton J, Zhu W, Wang L, Chubak J. Accounting for Differential Error in Time-to-Event Analyses Using Imperfect Electronic Health Record-Derived Endpoints. In *New Advances in Statistics and Data Science.* 2017 (pp. 239-255).

Hubbard RA, Tong J, Duan R, Chen Y. Reducing bias due to outcome misclassification for epidemiologic studies using EHR-derived probabilistic phenotypes. *Epidemiology. 2020; In press.*

Magder LS, Hughes JP. Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology*. 1997;146(2):195-203.

Meier AS, Richardson BA, Hughes JP. Discrete proportional hazards models for mismeasured outcomes. Biometrics. 2003;59(4), 947-954.

Sinnott JA, Dai W, Liao KP, Shaw SY, Ananthakrishnan AN, Gainer VS, Karlson EW, Churchill S, Szolovits P, Murphy S, Kohane I. Improving the power of genetic association tests with imperfect phenotype derived from electronic medical records*.* *Human Genetics*. 2014; 133:1369-82.

Tong J, Huang J, Chubak J, Wang X, Hubbard RA, Chen Y. An augmented estimation procedure for EHR-based association studies accounting for differential misclassification. *Journal of the American Medical Informatics Association*. 2020; 27(2):244-53.

Pragmatic Clinical Trials Embedded in Healthcare Systems

Coronado GD, Vollmer WM, Petrik A, Taplin SH, Burdick TE, Meenan RT, Green BB. Strategies and opportunities to STOP colon cancer in priority populations: design of a cluster-randomized pragmatic trial. *Contemporary clinical trials.* 2014;38(2):344-9.