Fall risk prediction and validation in older adults

Leveraging electronic health records with machine learning

Dormosh, N.

Publication date
2023

Chapter

1

General introduction

This thesis is primarily about how to reliably predict and validate the risk of a future fall of older adults. The prediction is based on routinely-collected data in the Electronic Health Records (EHR) in primary care as well as in the hospital setting using machine learning approaches including the use of Natural Language Processing (NLP) of free-text clinical notes. Below we provide preliminaries on falls and fall-risk factors, fall prediction, EHR and NLP. Next we state the aim and objectives of this thesis and provide an outline of the thesis chapters.

Falls in older adults

Falls in older adults are one of the geriatric giants, prevalent and morbid. One out of every three community-dwelling adults aged 65 and older falls each year, and the risk of falling increases with age (1). Around 10-15% of falls result in fractures, head injuries, and other serious injuries (2–4), which can lead to hospitalization, disability, and even death (5). In addition to the physical consequences, falls can also have emotional consequences, such as fear of falling, loss of confidence, and decreased quality of life (6, 7). Furthermore, falls also occur frequently in hospitals, with approximately 6% of older adults experiencing a fall during their hospital stay (8). In hospital, falls form a significant patient safety concern and can result in prolonged hospital stays, additional medical interventions, and increased healthcare costs (9, 10). In general, falls in older adults impose a staggering burden, not just on the individual level but also the society, as they are associated with increased healthcare utilization and expenses, including hospitalization, rehabilitation, and long-term care (5, 11, 12).

Fall-risk factors and Fall-risk increasing drugs

Falls are complex and multifactorial in nature, and the risk of falling rises with the number of risk factors that are present. Fall-risk factors can be categorized in various ways. The Centers of Disease Control and Prevention (CDC) classifies fall-risk factors into either personal (intrinsic) or environmental (extrinsic) factors (13). Intrinsic risk factors are related to the person themselves such as age, balance and gait, co-morbid health conditions (e.g., arthritis, stroke, incontinence, diabetes, Parkinson's, and dementia) and medication (e.g., polypharmacy, psychotropics and cardiovascular drugs). Extrinsic risk factors, on the other hand, are related to the environment and external factors that may increase the risk of falls. For instance, poor lighting and slippery floors.

Many of the intrinsic and extrinsic fall-risk factors are modifiable and hence can be changed or addressed through medical interventions or lifestyle modifications to reduce fall risk (14). One of the important modifiable risk factors for falls is medication. Medications that have been associated with an increased risk of falls in older adults are called fall-risk increasing drugs (FRIDs). These drugs may cause dizziness, drowsiness, im-

paired balance, and other side effects that can increase the risk of falls. Examples of fall-risk increasing drugs include sedatives, hypnotics, opioids, antipsychotics, antidepressants, and some cardiovascular medications (15–17). Because many FRIDs constitute modifiable risk factors for falls, there is an opportunity for clinicians to review the medications prescribed and potentially adjust dosages or switch to safer alternatives to reduce the fall risk. Medication assessment is a fundamental part of the multifactorial fall risk assessment as outlined in the World Falls Guidelines for falls prevention and management (18). Nevertheless, clinicians frequently face difficulties in medication management and may be reluctant to deprescribe FRIDs for various reasons, such as limited resources or the absence of dedicated tools to support them in addressing medication-related falls (19, 20).

FRIDs are typically grouped based on anatomical, pharmacological, therapeutic, or chemical properties to facilitate interpretation. However, definitions of FRIDs vary between studies and consensus on the exact grouping of these medications into standardized categories is lacking (21), posing challenges in understanding their fall-risk implications. For instance, consider the subclasses of "loop diuretics" and "thiazide diuretics", both used in the treatment of hypertension or heart failure. These subclasses are commonly grouped together under the broader class of "diuretics" based on their pharmacological property. While both are associated with falls, the potential fall risk attributed to "loop diuretics" is significantly higher than that of "thiazide diuretics", a distinction that can be obscured when they are grouped together under the broader class of "diuretics" (15). Therefore, careful consideration of medication grouping is important to ensure accurate assessment and management of fall risk in older adults.

Toward personalized fall risk estimation

Efficient Identification of older adults at higher fall risk allows for targeted interventions to be implemented to optimally prevent unnecessary falls and reduce the risk of injury. Many guidelines on fall prevention and interventions advocate the use of fall risk stratification tools to identify older adults at high fall risk (20) including the recently introduced World Falls Guidelines for falls prevention and management (18). In many of these guidelines, fall risk stratification is performed by means of algorithms created by clinical experts based on knowledge and evidence from the literature. These algorithms rely primarily on fall history and tests for balance and gait such as the Timed Up and Go Test (TUG) (22), to stratify older adults in coarse risk groups (i.e., low, intermediate or high). However, most of these algorithms are not validated and the TUG has low predictive ability when used in isolation (23).

An alternative approach is to use prediction models to estimate the fall risk. A prediction model is defined as a mathematical formula that uses various patient characteristics (e.g., age, sex, medical conditions and medication) to generate a personalized

risk score for fall risk (24). An important advantage of using prediction models over guideline-based algorithms is that they provide individualized risk scores based on a large number of predictors (some of which may be known risk factors) taking into account the uncertainty and variability associated with each individual predictor. One of the intrinsic proprieties of the probabilities produced by prediction models is that they provide a natural interpretation of their errors. Specifically, when a prediction model indicates a 0.2 probability of fall, it inherently means that for patients who did not fall the error is also 0.2. As such, they are more useful for guiding informative decisions, especially for patients who would fall into an intermediate risk category. Nonetheless, a systematic review examining prediction models for falls in community-dwelling older adults revealed that all current models exhibited a high risk of bias due to shortcomings in statistical analysis, outcome assessment and narrow inclusion criteria, making them unsuitable for clinical application (25). Thus, there is a need for a new and better prediction model for falls in community-dwelling older adults that overcomes important deficiencies present in existing models.

A prediction model is only valuable if it can be effectively implemented in practical settings to reliably inform decision-making or improve outcomes. External validation of prediction models is necessary to justify their implementation in real practice (24, 26). It involves testing the model on a new set of data that was not used to develop the model, where it tends to be less performant due to, for example, differences in case mix and varying outcome rates (26, 27). Therefore, external validation is critical to ensure that the model is generalizable and can be used in different patient populations or clinical settings. Despite the large number of prognostic models for falls in community-dwelling older adults, external validation is lacking (25). That means that most of the existing prediction models for falls cannot be recommended for clinical use (24, 28).

Therefore, in this thesis we attempt the careful development and (external) validation of prediction models for falls in older adults. The focus will be on community-dwelling older adults, although we also address the development and internal validation of a prediction model in the hospital setting.

Electronic health records

The digitization of health records has revolutionized the way patient information is stored, managed, and shared within the healthcare industry. The primary purpose of introducing EHRs in healthcare was to improve the quality, safety, and efficiency of health care delivery (29, 30). However, the wealth of data collected during the clinical care has also opened up opportunities for secondary uses in research (31), facilitated with the advancements in technology and the rapid development in the machine learning community (32).

Developing and validating risk prediction models is an emerging area of secondary

use for EHR data (33). EHR data offer various benefits compared to conventional data sources, such as data gathered in research cohort studies, which make them an attractive source to develop prediction models. A potential advantage of using EHR data is that they are collected under real-world conditions, which can result in greater representation and generalizability (34). Moreover, EHR data are collected as part of routine clinical care offering a broad range of variables and clinical outcome data as well as a relatively large sample size across many time points, while reducing administrative costs (34, 35). Another appealing advantage is that prediction models developed using EHR data can often be implemented and integrated with a clinical decision support system (CDSS) as they rely on readily available variables. By contrast, prediction models based on traditional data sources or guideline-based algorithms require translation to the clinical environment (33). Nevertheless, the analysis of EHR data also presents methodological challenges related to the quality and completeness of the data since they are not collected specifically for research purposes (36). For example, missing data is very common in EHR data as some variables are only collected when they are relevant to the patient (37). Therefore, researchers should take steps to address challenges inherent to EHR data to ensure the validity and reliability of their findings (33, 38).

EHR data are rich with information, including information on falls and fall-risk factors, that can be used to develop prediction models for falls. EHR data can be broadly categorized into structured data, which has a consistent format such as demographics, diagnoses, and medications, and unstructured data like clinical notes. Analyzing structured data is relatively straightforward since it is stored in a predefined structured format. However, unstructured data, particularly free-text clinical notes, lack a predefined structure and may contain fragmented sentences, abbreviations, and misspellings, making them more challenging to analyze. Thus, in order to analyze the clinical notes, NLP techniques (described subsequently) are required to convert them into a structured format.

This thesis mainly centers on utilizing EHR data to develop, validate, and improve fall risk prediction models, given their aforementioned advantages, as well as the abundance of falls and fall-related information, including both structured and unstructured data.

Natural Language Processing

A significant portion of patient information contained in EHR data is present in unstructured free-text notes, which can make up to 80% of the data according to some studies (39, 40). As mentioned above, NLP techniques are needed to automatically identify and extract relevant information from these notes. NLP is a branch of artificial intelligence that aims to narrow the gap between human and machine language by using computational techniques to help computers comprehend, interpret, and generate human language. Improvements in computer processing and the accessibility to EHR data opened

up many applications for NLP in healthcare to get insight into hidden information in clinical notes (41). Examples of such applications have been identified in previous systematic reviews, including automatic coding for medical billing (42), patient phenotyping for clinical trials (43) and predicting outcomes (44).

NLP techniques are particularly relevant in fields such as geriatric medicine, where structured data may not adequately capture the nuanced and complex patient care of older adults. A previous study found that several geriatric syndromes, known with their association with falls such as lack of social support, vision impairment, urinary incontinence and walking difficulty, were better identified in unstructured data than structured data (45). Another study found environmental factors (e.g., poor lighting, bed height) to be implicitly documented in clinical notes (46). By using NLP techniques to automatically extract information from clinical notes, we can complement evidence from structured data to improve the predictive performance of fall prediction models, which have traditionally relied exclusively on structured data.

Another application of NLP is to discover patterns and trends in fall risk factors by exploiting the longitudinal aspect of the clinical notes. Most of the existing studies describing risk factors for falls are cross-sectional where data on risk factors are collected at a single point in time, disregarding the influence of time on the risk. However, as mentioned before, falls are complex and involve multiple risk factors that dynamically interact and change over time. For example, mobility function is such a factor possibly improving or deteriorating over time, depending on several factors such as physical activity, environmental factors, medication use and comorbidities. The identification of such trends allows clinicians to anticipate indicators that signal an elevated risk of falls, which can then be addressed through intervention strategies in the early stages before the occurrence of falls.

Topic modelling is a popular NLP technique that can be leveraged to extract abstract topics from clinical notes (47). It assumes that clinical notes contain a mixture of related words that may form a topic, and each clinical note can be seen as a collection of topics varying in prominence of topics. For example, coexistence of the words: “gait”, “walking aid”, “balance” and “limitation” may suggest the topic “mobility limitation”. Topics extracted from the clinical notes can be used as input variables to a machine learning algorithm in order to develop a prediction model. Dynamic topic modelling is an extension of topic modelling that captures changes in topics over time (48). As such, it allows tracking the emergence and diminishment of topics related to fall risk over time.

Accordingly, in this thesis we describe two potential NLP applications for falls in older adults using general practitioners’ (GPs) clinical notes. The first one investigates the incremental predictive value of complementing topics extracted from the clinical notes with traditional clinical variables. The second application employs dynamic topic modeling to uncover topics associated with the onset of falls and track patterns and trends

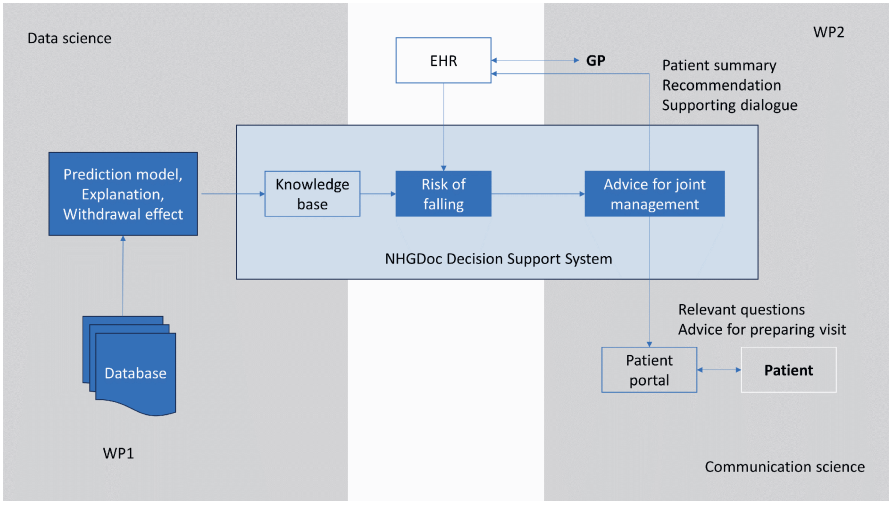

Figure 1: Visualized overview of the main components of the SNOWDROP project, data science (WP1) and communication science (WP2)

in these topics as they change over time leading up to the falls.

SNOWDROP

The research presented in this thesis was part of the SNOWDROP (SeNiors empOWred via big Data to joint-manage their medication-related Risk Of falling in Primary care) project. SNOWDROP is an interdisciplinary initiative focusing on the development and evaluation of a comprehensive data-driven science approach for valid prediction of personalized risk of falling that effectively supports joint medication management between seniors and GPs. The project aimed to accomplish two primary objectives: 1) to develop and validate prediction models for falls in older adults that can be used to estimate individualized fall risk (WP1; data science), and 2) to use these prediction models to provide smart decision support for shared decision-making between GPs and older adults, through a clinical decision support system and a patient portal (WP2; communication science). This thesis is focused on the first objective of the SNOWDROP project. A visual representation of both parts of the project is depicted in Figure 1.

Aim and outline of the thesis

This thesis leverages advanced machine learning techniques and EHR data to create algorithms and tools that can quantify fall risk and provide a sound basis for decision-making regarding interventions that can be effective in reducing fall risk. To accomplish this, the thesis has the following objectives:

1. To examine current prediction models constructed using EHR data for falls in older adults in the community, through a systematic critiquing of current models, in order to gain insights and formulate recommendations for future prediction models in this field.
2. To develop, validate and improve prediction models for falls using routinely-collected EHR data, in primary care as well as in hospital setting.
3. To investigate the potential application of NLP and machine learning in predicting falls, as well as to comprehend the pattern of factors that contribute to the risk of falling.

This thesis is organized as follows. Chapter 2 provides an overview of existing prediction models for falls in older adults living in the community. This review aims to provide a comprehensive understanding of the available models for falls in older adults, and to compare prediction models based on data from prospective research cohorts with those developed using EHR data. The models are evaluated on methodology, bias, applicability and predictive performance. This review also provides recommendations for future prediction models, including those described in later chapters. Chapter 3-5 describe the development and validation of new fall risk prediction models in older adults. In Chapter 3, a prediction model for fall in community-dwelling older adults is presented. The model is developed using pseudonymized primary care EHR data of 50 general practices in the Netherlands who participate in the Academic General Practitioner's Network at Academic Medical Center (AHA AMC). Chapter 4 describes the subsequent validation of the aforementioned model in an external independent cohort of older adults pertaining to 59 general practices registered in the Academic Network of General Practice at VU medical center in Amsterdam (ANH VUmc). By conducting this validation on an independent data, the reliability and generalizability of the model can be established, and its clinical use can be justified as the prediction model is intended for use in clinical practice by general practitioners. As individual characteristics of older adults residing in the community can differ from those who are hospitalized, a specific prediction model for this population is presented in Chapter 5. The model is based on EHR data collected from Amsterdam UMC - location AMC, a large tertiary hospital in the Netherlands. In chapter 6, the impact of describing medications at granularity levels on the predictive performance is investigated. Chapter 7 and 8 describe the potential of NLP and machine learning for falls and fall risk factors, where we apply modern NLP techniques that map words and clinical notes into vectors of numbers in order to capture their meaning. In particular, Chapter 7 investigates the incremental predictive value of incorporating unstructured clinical notes over traditional structured clinical variables, and Chapter 8 aims to discover patterns and trends of fall risk factors over time. Chapter 9 gives an overview over the SNOWDROP project, focusing on addressed problems, challenges, key results and future prospects. Finally, Chapter 10 presents a discussion covering the key contribu-

tions of this thesis, as well as its findings, strengths and limitations, implications, and prospects for future research.

References

[1] David A Ganz and Nancy K Latham. Prevention of falls in community-dwelling older adults. *New England journal of medicine*, 382(8):734–743, 2020.

[2] Sarah D Berry and Ram R Miller. Falls: epidemiology, pathophysiology, and relationship to fracture. *Current osteoporosis reports*, 6(4):149–154, 2008.

[3] Michael C Nevitt, Steven R Cummings, and Estie S Hudes. Risk factors for injurious falls: a prospective study. *Journal of gerontology*, 46(5):M164–M170, 1991.

[4] Richard W Sattin. Falls among older persons: a public health perspective. *Annual review of public health*, 13(1):489–508, 1992.

[5] Elizabeth R Burns, Judy A Stevens, and Robin Lee. The direct costs of fatal and non-fatal falls among older adults—united states. *Journal of safety research*, 58:99–103, 2016.

[6] M. Stenhagen, H. Ekström, E. Nordell, and S. Elmståhl. Accidental falls, health-related quality of life and life satisfaction: a prospective study of the general elderly population. *Arch Gerontol Geriatr*, 58(1):95–100, 2014.

[7] Alice C Scheffer, Marieke J Schuurmans, Nynke Van Dijk, Truus Van Der Hooft, and Sophia E De Rooij. Fear of falling: measurement strategy, prevalence, risk factors and consequences among older persons. *Age and ageing*, 37(1):19–24, 2008.

[8] Prabha Lakhan, Mark Jones, Andrew Wilson, Mary Courtney, John Hirdes, and Leonard C Gray. A prospective cohort study of geriatric syndromes among older medical patients admitted to acute care hospitals. *Journal of the American Geriatrics Society*, 59(11):2001–2008, 2011.

[9] Erin D Bouldin, Elena M Andresen, Nancy E Dunton, Michael Simon, Teresa M Waters, Minzhao Liu, Michael J Daniels, Lorraine C Mion, and Ronald I Shorr. Falls among adult patients hospitalized in the united states: prevalence and trends. *Journal of patient safety*, 9(1):13, 2013.

[10] Catherine A Wong, Angela J Recktenwald, Marilyn L Jones, Brian M Waterman, Mara L Bollini, and Wm Claiborne Dunagan. The cost of serious fall-related injuries at three midwestern hospitals. *The Joint Commission Journal on Quality and Patient Safety*, 37(2):81–87, 2011.

[11] Sven Heinrich, Kilian Rapp, Ulrich Rissmann, Caroline Becker, and H-H König. Cost of falls in old age: a systematic review. *Osteoporosis international*, 21:891–902, 2010.

[12] Curtis S Florence, Gwen Bergen, Adam Atherly, Elizabeth Burns, Judy Stevens, and Cynthia Drake. Medical costs of fatal and nonfatal falls in older adults. *Journal of the American Geriatrics Society*, 66(4):693–698, 2018.

[13] Centers for Disease Control and Prevention. Risk factors for falls. `https://www.cdc.gov/steadi/pdf/Risk_Factors_for_Falls-print.pdf`. Accessed: 2023-04-10.

[14] Lesley D Gillespie, M Clare Robertson, William J Gillespie, Catherine Sherrington, Simon Gates, Lindy Clemson, and Sarah E Lamb. Interventions for preventing falls in older people living in the community. *Cochrane Database of Systematic Reviews*, 2021(6), September 2012.

[15] Vries M, Seppala LJ, Daams JG, Glind EMM, Masud T, Velde N, E.U.G.M.S. Task, and Finish Group Fall-Risk-Increasing Drugs. Fall-risk-increasing drugs: a systematic review and meta-analysis: I. cardiovascular drugs. *Cardiovascular drugs. J Am Med Dir Assoc*, 19(4), 2018.

[16] L.J. Seppala, A.M.A.T. Wermelink, and M. Vries. Fall-risk-increasing drugs: a systematic review and meta-analysis: II. psychotropics. *Psychotropics. J Am Med Dir Assoc*, 19(4), 2018.

[17] L.J. Seppala, E.M.M. Glind, and J.G. Daams. Fall-risk-increasing drugs: a systematic review and meta-analysis: III. others. *Others. J Am Med Dir Assoc*, 19(4), 2018.

[18] Manuel Montero-Odasso, Nathalie van der Velde, Finbarr C Martin, Mirko Petrovic, Maw Pin Tan, Jesper Ryg, Sara Aguilar-Navarro, Neil B Alexander, Clemens Becker, Hubert Blain, et al. World guidelines for falls prevention and management for older adults: a global initiative. *Age and ageing*, 51(9):afac205, 2022.

[19] LJ Seppala, N Van der Velde, T Masud, H Blain, Mirko Petrovic, TJ Van der Cammen, Katarzyna Szczerbińska, Sirpa Hartikainen, RA Kenny, J Ryg, et al. Eugms task and finish group on fall-risk-increasing drugs (frids): position on knowledge dissemination, management, and future research. *European geriatric medicine*, 10:275–283, 2019.

[20] Manuel M Montero-Odasso, Nellie Kamkar, Frederico Pieruccini-Faria, Abdelhady Osman, Yanina Sarquis-Adamson, Jacqueline Close, David B Hogan, Susan Winifred Hunter, Rose Anne Kenny, Lewis A Lipsitz, et al. Evaluation of clinical practice guidelines on fall prevention and management for older adults: a systematic review. *JAMA network open*, 4(12):e2138911–e2138911, 2021.

[21] Lotta J Seppala, Mirko Petrovic, Jesper Ryg, Gulistan Bahat, Eva Topinkova, Katarzyna Szczerbińska, Tischa JM van der Cammen, Sirpa Hartikainen, Birkan Ilhan, Francesco Landi, et al. Stoppfall (screening tool of

older persons prescriptions in older adults with high fall risk): a delphi study by the eugms task and finish group on fall-risk-increasing drugs. *Age and ageing*, 50(4):1189–1199, 2021.

[22] Anne Shumway-Cook, Sandy Brauer, and Marjorie Woollacott. Predicting the probability for falls in community-dwelling older adults using the timed up & go test. *Physical therapy*, 80(9):896–903, 2000.

[23] Gotaro Kojima, Tahir Masud, Denise Kendrick, Richard Morris, Sheena Gawler, Jonathan Treml, and Steve Iliffe. Does the timed up and go test predict future falls among british community-dwelling older people? prospective cohort study nested within a randomised controlled trial. *BMC geriatrics*, 15(1):1–7, 2015.

[24] K.G. Moons, D.G. Altman, and J.B. Reitsma. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): explanation and elaboration. *Ann Intern Med*, 162(1), 2015.

[25] G.V. Gade, M.G. Jørgensen, and J. Ryg. Predicting falls in community-dwelling older adults: a systematic review of prognostic models. *BMJ Open*, 11(5), 2021.

[26] C.L. Ramspek, K.J. Jager, F.W. Dekker, C. Zoccali, and M. Diepen. External validation of prognostic models: what, why, how, when and where? *Clin Kidney J*, 14:49–58, 2021.

[27] T.P.A. Debray, Y. Vergouwe, H. Koffijberg, D. Nieboer, E.W. Steyerberg, and K.G.M. Moons. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol*, 68:279–289, 2015.

[28] D.G. Altman, Y. Vergouwe, P. Royston, and K.G.M. Moons. Prognosis and prognostic research: validating a prognostic model. *BMJ*, 338:1432–1435, 2009.

[29] Hilal Atasoy, Brad N Greenwood, and Jeffrey Scott McCullough. The digitization of patient care: a review of the effects of electronic health records on health care quality and utilization. *Annual review of public health*, 40:487–500, 2019.

[30] Deepa Wani and Manoj Malhotra. Does the meaningful use of electronic health records improve patient outcomes? *Journal of Operations Management*, 60:1–18, 2018.

[31] Tabinda Sarwar, Sattar Seifollahi, Jeffrey Chan, Xiuzhen Zhang, Vural Aksakalli, Irene Hudson, Karin Verspoor, and Lawrence Cavedon. The secondary use of electronic health records for data mining: Data characteristics and challenges. *ACM Computing Surveys (CSUR)*, 55(2):1–40, 2022.

[32] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604, 2017.

[33] Benjamin A Goldstein, Ann Marie Navar, Michael J Pencina, and John P A Ioannidis. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 24(1):198–208, May 2016.

[34] Lars G Hemkens, Despina G Contopoulos-Ioannidis, and John PA Ioannidis. Routinely collected data and comparative effectiveness evidence: promises and limitations. *Cmaj*, 188(8):E158–E164, 2016.

[35] Sophie H Bots, Rolf H H Groenwold, and Olaf M Dekkers. Using electronic health record data for clinical research: a quick guide. *European Journal of Endocrinology*, 186(4):E1–E6, 03 2022.

[36] Mark G Weiner and Peter J Embi. Toward reuse of clinical data for research and quality improvement: the end of the beginning? *Annals of internal medicine*, 151(5):359–360, 2009.

[37] Irene Petersen, Catherine A Welch, Irwin Nazareth, Kate Walters, Louise Marston, Richard W Morris, James R Carpenter, Tim P Morris, and Tra My Pham. Health indicator recording in uk primary care electronic health records: key implications for handling missing data. *Clinical epidemiology*, pages 157–167, 2019.

[38] Christopher M Sauer, Li-Ching Chen, Stephanie L Hyland, Armand Girbes, Paul Elbers, and Leo A Celi. Leveraging electronic health records for data science: Common pitfalls and how to avoid them. *The Lancet Digital Health*, 2022.

[39] Hyoun-Joong Kong. Managing unstructured big data in healthcare system. *Healthcare informatics research*, 25(1):1–2, 2019.

[40] Travis B Murdoch and Allan S Detsky. The inevitable application of big data to health care. *Jama*, 309(13):1351–1352, 2013.

[41] K. Kreimeyer, M. Foster, and Pandey A. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *J Biomed Inform*, 73:14–29, 2017.

[42] Rajvir Kaur, Jeewani Anupama Ginige, and Oliver Obst. A systematic literature review of automated icd coding and classification systems using discharge summaries. *arXiv preprint arXiv:2107.10652*, 2021.

[43] Betina Idnay, Caitlin Dreisbach, Chunhua Weng, and Rebecca Schnall. A systematic review on natural language processing systems for eligibility prescreening in clinical research. *Journal of the American Medical Informatics Association*, 29(1):197–206, 2022.

[44] T.M. Seinen, E.A. Fridgeirsson, and Ioannou S. Use of unstructured text in prognostic clinical prediction models: a systematic review. *J Am Med Inform Assoc*, 29:1292–302, 2022.

[45] H. Kharrazi, L.J. Anzaldi, and Hernandez L. The value of unstructured electronic health record data in geriatric syndrome case identification. *J Am Geriatr Soc*, 66:1499–507, 2018.

[46] R.I. Bjarnadottir and R.J. Lucero. What can we learn about fall risk factors from ehr nursing notes? a text mining study. *EGEMs (Generating Evidence and Methods to Improve Patient Outcomes*, 6(1):21, 2018.

[47] Rob Churchill and Lisa Singh. The evolution of topic modeling. *ACM Computing Surveys*, 54(10s):1–35, 2022.

[48] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120, 2006.