Machine Learning Approaches in Inflammatory Bowel Disease

Ileana Scarpino[1][0000−0002−5216−968X]
Rosarina Vallelunga[2][0000−0003−4644−7871]
Francesco Luzza[2][0000−0001−5120−1046]
Mario Cannataro[1][0000−0003−1502−2387]

[1] Data Analytics Research Center, Department of Medical and Surgical Sciences, University of Catanzaro, Italy
[2] Department of Health Science, University of Catanzaro, Italy
{ileana.scarpino,rosarina.vallelunga,luzza,cannataro}@unicz.it

Abstract. The great flow of clinical data can be managed with efficiency and effectiveness, improving the speed of interpretation of information, through Machine Learning (ML) methodologies, aimed at overcoming the barriers present in the diagnosis and treatment processes of patients, such as those affected by Inflammatory Bowel Disease (IBD). In this paper we survey relevant ML applications used for managing the large flow of clinical data and for overcoming the barriers present in the diagnosis and treatment processes of patients, with special focus on IBD. In IBD settings, main data sources include cohort study data, administrative databases, e-Health applications, Electronic Health Records (EHR), medical image data, Omics data, Clinical trial data and social media data. Potential applications for overcoming barriers in the field of IBD are also discussed.

Keywords: Machine Learning · Inflammatory Bowel Disease · Natural Language Processing.

1 Introduction

Medicine needs technological revolution, which allows the identification of new interesting markers that can't be identified with statistical methods. Inflammatory Bowel Disease (IBD), which includes Ulcerative Colitis (UC) e Crohn's Disease (CD), is a complex multifactorial inflammatory disease with common symptoms such as abdominal pain, diarrhea, rectal bleeding, fatigue, and extraintestinal manifestations of the disease [1]. Machine Learning (ML) application in IBD represents a path of research to improve patient health outcomes since it offers patients greater opportunities to access treatment, to understand their state of health, to evaluate prevention, and to receive early diagnoses. IBD gave birth to new challenges that traditional scientific methods have failed to address [2, 3]. The present paper discusses the possibility of ML applications for the characterization of the IBD disease through the extraction of topics from

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_59

public and private sources, such as clinical reports and clinical notes. The rest of the paper is organized as follows. Section II presents an overview of Data Sources and available public databases in IBD. In Section III the background on ML approaches applied on IBD data is introduced. Finally Section IV and V present discussion and conclusion of the paper, opening new challenges of future works.

2 Data Sources in IBD

In the field of IBD, the use of big data has allowed medical researchers to understand the disease and the models related to it and to obtain more information that allow to progress in the clinical practice. The most important data sources in IBD include study cohorts, clinical studies, administrations, medical and electronic health record (EHR) databases, reported results databases, medical imaging. For example, imaging modalities such as colonoscopy, gastroscopy, abdominal ultrasound allow the evaluation of any structural changes in the affected districts. Large data volumes collect both administrative databases and clinical notes, representing structured and unstructured data respectively. Medical data sources include biomarker data, medical images, clinical trials registries, electronic medical records, epidemiological studies, patient-reported health data, omics data, biometric data, data from social media and the Internet [4, 5].
Variety of data sources will continue to grow and the challenges will increase. In the field of IBD research data can be acquired from Administrative databases that are the most straightforward sources. For storing data collected during clinic, hospital, laboratory or pharmacy visits, many countries have developed large databases [6]. Typically, EHRs include both structured and unstructured data [7]. Data analytics in IBD is enhanced by extracting raw data that are processed to be stored, analysed and manipulated, while structured data are in the form of patient demographics, diagnosis codes, laboratory data, vital signs and similar material [8].

2.1 IBD Databases

With the purpose of homogenizing data, many databases collecting a growing number of information in the field of IBD were established. Medical and sociodemographic information from all hospital care and outpatient drug reimbursements can be extracted by Système National d'Information InterRégimes de l'Assurance Maladìe (SNIIRAM) and Programme de Médicalisation des Systèmes d'Information (PMSI) [9–12]. Numerous successful databases have been implemented at European and world level.
General Practice Research Database (GPRD) includes information about incident diagnoses, hospitalizations and surgeries, owing to incomplete records [13].
National Patient Register (NPR) contains data on specialized hospital-based outpatient care as well as data on diagnoses of IBD [14].
Swedish Quality Register (SWIBreg) contains clinical data missing in NPR [15].

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_59

Both SWIBreg and NPR have been validated in clinical studies related diagnoses of IBD [16]. Table 1 shows some publicly available clinical databases.

Table 1. Open Access Public Available Clinical Database and Ontologies in IBD

Description	Resource
Database of public and private clinical trials	ClinicalTrials.gov *(1)*
GWAS Catalog	Genome-Wide Association Study *(2)*
Genome-Phenome dataset	Database of Genotypes and Phenotypes *(3)*
PheWAS Catalog	GPhenome-Wide Association Studies *(4)*
Ontology related to inflammatory bowel disease	Mondo Disease Ontology *(5)*
Inflammatory Disease Ontology Browser	Disease Ontology *(6)*

(1) https://clinicaltrials.gov/ct2/results?term=gastroenterology
(2) https://www.ebi.ac.uk/gwas
(3) https://www.ncbi.nlm.nih.gov/gap/?term=gastroenterology
(4) https://phewascatalog.org/phewas
(5) http://purl.obolibrary.org/obo/ $MONDO_0$00525
(6) http://www.informatics.jax.org/disease/612241

The availability of these data could be hampered by several factors, such as intellectual property, fears of different conclusions, confidentiality concerns and lack of resources [17]. When data are analysed, personal information is de-identified but the possibility of recognizing individuals still exists [18].

3 Need for Machine Learning in IBD

Computational techniques can be used to solve problems related to storage, analysis and interpretation caused by enormous amounts of omics data [19] [20].
The arrival of ML into IBD clinical research has allowed researchers to capture complex associations and to increase understanding of disease mechanisms; therefore, ML could play an important role for improving diagnosis. ML algorithms require input data useful for training phase. In IBD input data are those patient biological as gene expression, biomarkers of inflammation in the tissue and blood, gut microbiota composition, endoscopic imaging and histologic imaging [21–25]. ML uses the ability of algorithms to detect predictive patterns, simplifying the interpretation of models at the base of complex medical conditions as IBD.
Table 2 shows some specific types of clinical technologies on IBD in which ML approaches have been applied.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_59

Table 2. Machine Learning Application on IBD Clinical Setting

ML Method	Application	PubMed ID (PMID)
NLP of EHR data	Identification of surveillance colonoscopy in IBD	*23086115 [26]*
NLP of EHR data	Improving case definition of IBD in EHR using NLP approach	*23567779 [27]*
Gaussian Bayesian Network	A probabilistic methods for classification of IBD	*29048458 [28]*
Bayesian ML model using clinical data	Bayesian Machine Learning Techniques for revealing complex interactions in IBD patients	*28269885 [29]*
ML using data from clinical trial	Predictive modeling of endoscopic remission in IBD	*29359519 [30]*
ML model using data set from IBD Genetics Consortium	Advanced machine-learning technique for risk prediction in IBD	*23731541 [21]*
ML model using EHR data	Prediction of outpatient corticosteroid use and hospitalization	*29272474 [30]*
ML model using EHR data	Validation of a Thiopurine Monitoring Algorithm on the SONIC Clinical Trial Dataset	*28838785 [31]*

4 Discussion

Data Mining (DM) and ML algorithms are computational approaches with the aim of extracting knowledge in the medical field. They are used to predict remission in patients with IBD and to analyze if remission predicted by algorithm leads to fewer clinical events [32]. For example Natural Language Processing (NLP) is used to identify arthralgia in electronic health records and to compare the risk of arthralgia between patients with IBD taking vedolizumab and those receiving anti-TNF agents [33]. There are many strengths and limitations of potential data sources from which big data analytics could draw from, in the field of IBD. One of the main challenges is the heterogeneity of data represented by social media posts and unstructured electronic health record notes. Since clinical information is spread across thousands of electronic documents, NLP approaches can reduce the time to organize this information, by overcoming the limitations in understanding documents. Some data sources raise questions of patient privacy and of corrupted, duplicate, missing or inaccurate data that require security solutions. Finally, a discussion about bioinformatics and computational sciences that are essential to adequately manage and integrate data from these components and other sources is reported here [34–36].

5 Conclusions and Future Work

Available data sources in the clinical setting represent the input to apply different analytical methods ranging from traditional statistical methods and advanced methods such as data mining, machine learning, clustering, text analysis and

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_59

image analytics, allowing to improve IBD knowledge and to fill the gaps present in this area. IBD can benefit from ML methodologies useful to understand behavioral drivers and undertake predictive therapeutic approaches. Various ML methods could discover the hidden nature of big data in the gastroenterology field, helping to subtype chronic and complex diseases within the bowel diseases. Performance improvement of NLP will be essential to organize, interpret and recognize patterns from textual data [7], such as unstructured clinical reports. These insights can lead to new discoveries through the extraction of information from medical records and the application of NLP techniques, in particular Text Mining (TM) approaches that can improve the characterization of bowel diseases. For instance, pharmacovigilance can be improved by using text mining, to obtain data on adverse drug events from medical notes [37]. The advantage is that in addition to the PDF format, clinical reports, including endoscopic ones, are often available in plain text and can be processed for NLP analysis. Among the possible future developments and challenges, we are working on the application of NLP and TM techniques to extract useful information from medical records as well as from medical questionnaires, with the aim of a better diagnosis in clinical practice.

References

1. Weersma, R.K., Xavier, R.J., Weersma, R., Barrett, J.C., Vermeire, S., Xavier, R., Anderson, C., Wijmenga, C., Daly, M., Alm, E., et al.: Multiomics analyses to deliver the most effective treatment to every patient with inflammatory bowel disease. Gastroenterology **155**(5), e1–e4 (2018)
2. Bernstein, C.N.: Treatment of ibd: where we are and where we are going. Official journal of the American College of Gastroenterology— ACG **110**(1), 114–126 (2015)
3. Actis, G.C., Pellicano, R., Rosina, F.: Inflammatory bowel diseases: Current problems and future tasks. World journal of gastrointestinal pharmacology and therapeutics **5**(3), 169 (2014)
4. Rumsfeld, J.S., Joynt, K.E., Maddox, T.M.: Big data analytics to improve cardiovascular care: promise and challenges. Nature Reviews Cardiology **13**(6), 350–359 (2016)
5. Lee, C.H., Yoon, H.J.: Medical big data: promise and challenges. Kidney research and clinical practice **36**(1), 3 (2017)
6. Hashimoto, R.E., Brodt, E.D., Skelly, A.C., Dettori, J.R.: Administrative database studies: goldmine or goose chase? Evidence-based spine-care journal **5**(02), 074–076 (2014)
7. Ross, M., Wei, W., Ohno-Machado, L.: "big data" and the electronic health record. Yearbook of medical informatics **23**(01), 97–104 (2014)
8. Raghupathi, W., Raghupathi, V.: Big data analytics in healthcare: promise and potential. Health information science and systems **2**(1), 1–10 (2014)
9. Bezin, J., Duong, M., Lassalle, R., Droz, C., Pariente, A., Blin, P., Moore, N.: The national healthcare system claims databases in france, sniiram and egb: powerful tools for pharmacoepidemiology. Pharmacoepidemiology and drug safety **26**(8), 954–962 (2017)

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_59

10. Moulis, G., Lapeyre-Mestre, M., Palmaro, A., Pugnet, G., Montastruc, J.L., Sailler, L.: French health insurance databases: what interest for medical research? La Revue de médecine interne **36**(6), 411–417 (2015)
11. Tuppin, P., De Roquefeuil, L., Weill, A., Ricordeau, P., Merlière, Y.: French national health insurance information system and the permanent beneficiaries sample. Revue d'epidemiologie et de sante publique **58**(4), 286–290 (2010)
12. Tuppin, P., Rudant, J., Constantinou, P., Gastaldi-Ménager, C., Rachas, A., De Roquefeuil, L., Maura, G., Caillol, H., Tajahmady, A., Coste, J., et al.: Value of a national administrative database to guide public decisions: From the système national d'information interrégimes de l'assurance maladie (sniiram) to the système national des données de santé (snds) in france. Revue d'epidemiologie et de sante publique **65**, S149–S167 (2017)
13. Lewis, J.D., Brensinger, C., Bilker, W.B., Strom, B.L.: Validity and completeness of the general practice research database for studies of inflammatory bowel disease. Pharmacoepidemiology and drug safety **11**(3), 211–218 (2002)
14. Ludvigsson, J.F., Andersson, E., Ekbom, A., Feychting, M., Kim, J.L., Reuterwall, C., Heurgren, M., Olausson, P.O.: External review and validation of the swedish national inpatient register. BMC public health **11**(1), 1–16 (2011)
15. Jin, L., Zuo, X.Y., Su, W.Y., Zhao, X.L., Yuan, M.Q., Han, L.Z., Zhao, X., Chen, Y.D., Rao, S.Q.: Pathway-based analysis tools for complex diseases: a review. Genomics, proteomics & bioinformatics **12**(5), 210–220 (2014)
16. Jakobsson, G.L., Sternegård, E., Olén, O., Myrelid, P., Ljung, R., Strid, H., Halfvarson, J., Ludvigsson, J.F.: Validating inflammatory bowel disease (ibd) in the swedish national patient register and the swedish quality register for ibd (swibreg). Scandinavian journal of gastroenterology **52**(2), 216–221 (2017)
17. Bertagnolli, M.M., Sartor, O., Chabner, B.A., Rothenberg, M.L., Khozin, S., Hugh-Jones, C., Reese, D.M., Murphy, M.J.: Advantages of a truly open-access data-sharing model. The New England journal of medicine **376**(12), 1178–1181 (2017)
18. Genta, R.M., Sonnenberg, A.: Big data in gastroenterology research. Nature Reviews Gastroenterology & Hepatology **11**(6), 386–390 (2014)
19. Schultze, J.L., Rosenstiel, P., et al.: Systems medicine in chronic inflammatory diseases. Immunity **48**(4), 608–613 (2018)
20. Gedela, S.: Integration, warehousing, and analysis strategies of omics data. In: Bioinformatics for Omics Data, pp. 399–414. Springer (2011)
21. Wei, Z., Wang, W., Bradfield, J., Li, J., Cardinale, C., Frackelton, E., Kim, C., Mentch, F., Van Steen, K., Visscher, P.M., et al.: Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. The American Journal of Human Genetics **92**(6), 1008–1012 (2013)
22. Isakov, O., Dotan, I., Ben-Shachar, S.: Machine learning–based gene prioritization identifies novel candidate risk genes for inflammatory bowel disease. Inflammatory bowel diseases **23**(9), 1516–1523 (2017)
23. Iadanza, E., Fabbri, R., Bašić-ČiČak, D., Amedei, A., Telalovic, J.H.: Gut microbiota and artificial intelligence approaches: a scoping review. Health and Technology **10**(6), 1343–1358 (2020)
24. Mossotto, E., Ashton, J., Coelho, T., Beattie, R., MacArthur, B., Ennis, S.: Classification of paediatric inflammatory bowel disease using machine learning. Scientific reports **7**(1), 1–10 (2017)
25. Chen, P., Zhou, G., Lin, J., Li, L., Zeng, Z., Chen, M., Zhang, S.: Serum biomarkers for inflammatory bowel disease. Frontiers in Medicine **7**, 123 (2020)

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_59

26. Hou, J.K., Chang, M., Nguyen, T., Kramer, J.R., Richardson, P., Sansgiry, S., D'Avolio, L.W., El-Serag, H.B.: Automated identification of surveillance colonoscopy in inflammatory bowel disease using natural language processing. Digestive diseases and sciences **58**(4), 936–941 (2013)
27. Improving case definition of crohn's disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach
28. Han, L., Maciejewski, M., Brockel, C., Gordon, W., Snapper, S.B., Korzenik, J.R., Afzelius, L., Altman, R.B.: A probabilistic pathway score (props) for classification with applications to inflammatory bowel disease. Bioinformatics **34**(6), 985–993 (2018)
29. Menti, E., Lanera, C., Lorenzoni, G., Giachino, D.F., De Marchi, M., Gregori, D., Berchialla, P., on the Genetics of IBD, P.S.G., et al.: Bayesian machine learning techniques for revealing complex interactions among genetic and clinical factors in association with extra-intestinal manifestations in ibd patients. In: AMIA Annual Symposium Proceedings. vol. 2016, p. 884. American Medical Informatics Association (2016)
30. Waljee, A.K., Liu, B., Sauder, K., Zhu, J., Govani, S.M., Stidham, R.W., Higgins, P.D.: Predicting corticosteroid-free endoscopic remission with vedolizumab in ulcerative colitis. Alimentary pharmacology & therapeutics **47**(6), 763–772 (2018)
31. Waljee, A.K., Sauder, K., Zhang, Y., Zhu, J., Higgins, P.D.: External validation of a thiopurine monitoring algorithm on the sonic clinical trial dataset. Clinical Gastroenterology and Hepatology **16**(3), 449–451 (2018)
32. Waljee, A.K., Sauder, K., Patel, A., Segar, S., Liu, B., Zhang, Y., Zhu, J., Stidham, R.W., Balis, U., Higgins, P.D.: Machine learning algorithms for objective remission and clinical outcomes with thiopurines. Journal of Crohn's and Colitis **11**(7), 801–810 (2017)
33. Cai, T., Lin, T.C., Bond, A., Huang, J., Kane-Wanger, G., Cagan, A., Murphy, S.N., Ananthakrishnan, A.N., Liao, K.P.: The association between arthralgia and vedolizumab using natural language processing. Inflammatory bowel diseases **24**(10), 2242–2246 (2018)
34. De Souza, H.S., Fiocchi, C., Iliopoulos, D.: The ibd interactome: an integrated view of aetiology, pathogenesis and therapy. Nature Reviews Gastroenterology & Hepatology **14**(12), 739–749 (2017)
35. Fiocchi, C.: Integrating omics: the future of ibd? Digestive Diseases **32**(Suppl. 1), 96–102 (2014)
36. Chuong, K.H., Mack, D.R., Stintzi, A., O'Doherty, K.C.: Human microbiome and learning healthcare systems: integrating research and precision medicine for inflammatory bowel disease. Omics: a journal of integrative biology **22**(2), 119–126 (2018)
37. Harpaz, R., Callahan, A., Tamang, S., Low, Y., Odgers, D., Finlayson, S., Jung, K., LePendu, P., Shah, N.H.: Text mining for adverse drug events: the promise, challenges, and state of the art. Drug safety **37**(10), 777–790 (2014)

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_59