83133_Auto_Edited.docx

WORD COUNT	3322	TIME SUBMITTED	16-FEB-2023 03:52PM
		PAPER ID	96676952

Name of Journal: *Artificial Intelligence in Gastroenterology*

Manuscript NO: 83133

Manuscript Type: MINIREVIEWS

Big data and variceal rebleeding prediction in cirrhosis patients

Quan Yuan, Wen-Long Zhao, Bo Qin

Abstract

Big data has convincing merits in developing risk stratification strategies for diseases. The 6 "V"s of big data, namely, volume, velocity, variety, veracity, value, and variability, have shown promise for real-world scenarios. Big data can be applied to analyze health data and advance research in preclinical biology, medicine, and especially, disease initiation, development, and control. A study design comprises data selection, inclusion and exclusion criteria, standard confirmation and cohort establishment, follow-up strategy, and events of interest. The development and efficiency verification of a prognosis model consists of deciding the data source, taking previous models as references while selecting candidate predictors, assessing model performance, choosing appropriate statistical methods, and model optimization. The model should be able to inform disease development and outcomes, such as predicting variceal rebleeding in patients with cirrhosis. Our work has merits beyond those of other colleagues with respect to cirrhosis patient screening and data source regarding variceal bleeding.

INTRODUCTION

Many risk stratification strategies for diseases mainly depend on single-/medium-sized cohort studies or their meta-analysis[1,2], with lead-time bias taken into consideration[3,4]. This type of study method is, by design, well scheduled and well phenotyped but selective for the population sampled, which may not reflect the real-world, pan-subject profile. Real-world patients may have comorbidities, be taking concomitant medications, may be excluded from short-term follow-up, or have poor patient compliance. Direct data acquisition from basic healthcare institutions and cohorts is more representative than limited sampling.

HISTORY OF BIG DATA

Although the use of piles of data in the medical field has a relatively long history[5-7], the term "big data" appeared only in the 1990s, and quickly became popular[8-10]. "Big" is a relative term, especially when it relates to data. Big data usually refers to datasets that exceed the capabilities of commonly used software tools to store, manage, and process that amount of data within a suitable period of time[11]. The term is described by 315 characteristics[12], and fundamentally by the 6 "V"s: volume, velocity, variety, veracity, value, and variability[13-17].

During the recent decade, methods for collecting, storing, and managing big data have evolved[18-20]. We are now entering an era of monitoring health changes using clinical indicators, such as vital signs, serum sugar, lipids, sweating, and bladder fullness, with wearable devices[21]. These changes can reflect physiological change. Constant variation and altered levels may result in different pathological states. Here, we review the applications of big data in predicting disease onset and prognosis.

APPLICATIONS OF BIG DATA

Applications of big data include its use as a tool to monitor the onset of conditions and diseases. Big data have been used for this purpose in relation to hypertension[22], pediatric oncology[23], oral care[24], general practice[25], rheumatic diseases[26], renal diseases[27], mechanical ventilation management in the ICU[28], and cirrhosis and hepatocellular carcinoma morbidity in the nonalcoholic fatty liver disease (NAFLD)/ nonalcoholic steatohepatitis (NASH) population[29]. Situations such as the commencement, development, and control of diseases can be studied and visualized using big data techniques, which is a promising and beneficial approach. With the help of big data, the creation of large, collaborative data can lay a more solid foundation for robust data sharing and scientific discovery in predicting the onset of pediatric oncology. Registry-based research, however, is one of the conventional research methods regarding pediatric cancers. In these studies, a multi-site registry for the study of pediatric patients was utilized, including fields of descriptive epidemiology, survivors, genomics, new registry description, data harmonization, palliative and supportive care, radiology, consensus guidelines, hereditary pediatric cancer, electronic health records, and prospective clinical trials. Limitations of registry-based research include the latest publication time range only, a restricted single publication database and a limited amount of research and registries only if they have yielded publicly-published peer-reviewed papers[23]. With this study strategy, data cannot be automatically mined, cleaned and integrated to perfect the already existing study. When it comes to new subjects, we need to redo the statistical analysis, while modeling and machine study in the big data scenario can perform the whole analysis process.

Healthcare data in some regions are complete and accessible for analysis. Real-world data from primary healthcare facilities in communities in European countries are a good resource, as the primary healthcare service is state-covered and there are few or no co-payments. Therefore, healthcare information and data are collected and stored by state-run big data centers. Most residents are registered at birth and have their complete healthcare information in electronic form, which can be accessed by regional practitioners and analyzed for real-world application scenarios[30]. However, numerous parameters, especially administrative data, mined from patients' inpatient and outpatient Hospital Information System (HIS)/ Electronic Medical Record (EMR) system *via* various algorithms are at risk of information and privacy leaking. Therefore, preliminary selection of data, especially low-dimensional administrative data, is preferable to decrease information leakage and privacy invasion.

Big data boosts the depth and breadth of research in fundamental biology and clinical medicine. There is already impressive progress due to this, including in exome sequencing[31], genomics, and proteomics. Taking the COVID-19 pandemic as an example, primary research, clinical practices regarding treatment, and even trends in media campaigns of whether or not executing lockdown and a positive policy of nucleic acid testing can be swiftly analyzed with big data tools to assist epidemic control[32].

STUDY DESIGN

Study design comprises data source selection, inclusion and exclusion criteria, standard confirmation and cohort establishment, follow-up strategy, and events of interest. A multi-country European real-world study acquired patient data within a set research period mined from central transcription, laboratories, pharmacy offices, medical insurance departments, administrative departments, and other departmental databases *via* an electronic health record data repository along with molecular typing from molecular biology laboratories for preventing outbreaks of hospital infections[33]. Chart presentations can be used to analyze and interpret descriptive data. The Fib-4 score (age, aspartate aminotransferase, alanine aminotransferase, and platelets), which is composed of entirely non-invasive parameters, has been used to detect early liver fibrosis[29].

MODEL DEVELOPMENT AND EFFICACY VERIFICATION

With respect to development and efficiency verification of disease onset and prognosis models, researchers have performed extensive work. Model development is the process of collecting vital parameters (risk factors) of consequence and weighted with varied weight coefficients to form a weighted function. This requires the identification of predominant predictors from a large amount of preselected candidate predictors, assigning proper weights to each predictor to obtain a combined risk score, and assessing the model's predictive performance with statistic methods such as a calibration plot. The latter includes calibration, discrimination, and (re)classification properties, assessing its potential for generalization using internal validation techniques, and if necessary, optimizing the model to avoid over-fitting. Data sources should preferably be prospective cohort(s) with a randomized controlled trial design or real-world medical record data. Preferred outcome choices are those that are related to patients or individuals such as remission time and follow-up period. Methods for outcome verification should be included, and the blind method is preferred.

Regarding the selection of candidate predictors, a surplus should be defined and analyzed before finally including a subset in the final model. Incorporation bias should be avoided by using blinding. Data quality control, missing data processing, continuous predictor modeling, final model development, relative weight assignment for each predictor and internal validation are essential in the process of creating a final prediction model[34].

Choosing appropriate statistical methods during model establishment is vital to guarantee reliability and validity. Regression analysis, including univariate and multivariate regression, is the most commonly used statistical method, especially Cox regression[35] and LASSO[36]. The hazard ratio is used to differentiate cohorts across different conditions and coefficients. Featured with net benefit and threshold probability for more convenient while trusty clinical decision making, decision curve analysis has been used to evaluate whether or not to use a certain prediction model[37]. In this approach, the theoretical relation between the threshold probabilities of a disease - that is to say a disease will take place - and the relative frequency of false positives and false negatives are examined to ensure the validity of a prediction model. The benefits of applying decision curve analysis can be quantified as whether a model can be easily and effectively applied in clinical situations. Its ability to help compare several different models regarding one issue is another advantage[38]. The parameter indicating risk threshold "T-value" has been used to study treatment decisions in risk models. The harm-to-benefit ratio is related to the T value, which is in line with the former. Balancing all benefits and harms in different scenarios is key to determining which T value is reasonable[39]. The net benefit (NB) value, which is a combined "net" effect of the true positives and false positives, was introduced to evaluate the potential clinical application of an estimating tool or a risk-predicting model. Setting the decisive threshold range in modeling is important, which is the boundary to determine whether a patient is judged as positive for a disease or not[40]. However, NB does not directly make up the harms and costs in acquiring the predictors for the chosen model. The focus of NB is to derive the best tradeoff between sufficient indicators and convenience in clinical application[41].

Model optimization should be conducted in order to reduce the number of predictors and avoid an unmanageable dataset or workload. AMSGrad (“far from the minimum”), a putative optimal method for optimizing models, is commonly used for low-cost cause. By just switching to the direct linear method near the end of the optimization, AMSGrad can do its magic as it has long convergence tails[42]. As for multi-objective racing algorithms with fixed confidence, SPRINT-Race is the first algorithm developed and uses a non-parametric, ternary-decision, dual-sequential probability ratio test to infer a pairwise dominance or non-dominance relationship. In order to minimize the computational effort, by sequentially applying a Holm's step-down family-wise error rate control method, the probability of mistakenly erasing any Pareto-optimal models or returning any clearly dominating models is restricted, which can achieve a pre-estimated confidence level to ensure the quality of the models generated[43]. The quantification of model-to-data correspondence is pivotal to measure a model's performance and future application for the problem at hand. The *Drosophila melanogaster* gap gene system model demonstrates the importance of error quantification, and it is applicable to a wide array of developmental modeling studies[44]. The support vector machine, GLM-Net, generalized linear model, partial least squares, neural network, k-nearest neighbors, random forest and boosted tree are useful tools for establishing the model to predict prognosis in patients with breast cancer[45]. Comparing their differences in performance and necessary model optimization can lead to better and more efficient application in practice.

PREDICTOR SCREENING FOR PROGNOSIS

For predictor screening with regard to disease prognosis, researchers have proposed methods such as the MELD model for cirrhosis-related mortality prediction and the APACHE model for critically ill patients. The clinical data of cirrhosis patients who had early admission, including clinical and socioeconomic factors, were mined from electronic medical records and classified for risk stratification in order to predict readmission within 30 days[46]. Recommended by the European Association of Urology (EAU), the EORTC (European Organization for Research and Treatment of Cancer) risk tables[47], which include six clinical and pathological factors - number of tumors, tumor size, prior recurrence rate, T category, carcinoma *in situ* and grade - were used to separately predict the short-term and long-term risks of progression and recurrence in an individual patient with a non-muscular invasive bladder tumor. It divided patients into four groups with individual recurrent and progression scores (Table 1). However, as EORTC risk tables overestimated recurrence in all risk groups and progression in the high-risk group, the CUETO (Club Urológico Español de Tratamiento Oncológico) scoring model[48] was developed. The well-known new EORTC model[49], or EAU risk groups, was popular in recurrence and progression prediction, in which tumor diameter and extent were key predictors for progression prediction in multistate analyses. The health belief model has been used in risk factors identifying for prostate cancer screening in aged Jordanian adults[50]. Development and validation of a prediction model, including internal and external, temporal and geographical, domain validation and their revision, are all crucial to identify predictors of prognosis[51].

RISK INDICATORS OF VARICEAL REBLEEDING IN CIRRHOSIS

Studies have reported several prediction models that predict variceal rebleeding in patients with cirrhosis. Risk indicators are components of prediction models. Invariably, studies in spotting possible risk indicators of variceal rebleeding among cirrhosis patients require a long time. Child-Pugh score and hepatic-venous pressure gradient are the most significant prognostic factors in stratifying the probability of variceal rebleeding[52]. Antiviral treatment significantly reduced rebleeding in patients with hepatitis B virus (HBV)-related cirrhosis. In-time prophylactic endoscopic treatment of upper gastrointestinal varices after first-time bleeding, including endoscopic varices ligation and gastric fundus varices gluing, is important in postponing variceal rebleeding[53]. Tachycardia, high creatinine level and low albumin level are independent factors which are associated with rebleeding, suggesting a potential predictive role. The transverse of these variables into predictive scores may provide improved prognosis for patients with variceal bleeding[54]. Pre-emptive transjugular intrahepatic portosystemic shunt was independently related to a lower rebleeding rate[55]. Albumin transfusion in patients with low albumin level was positively associated with a decreased rebleeding rate[56]. Five studies showed a lower rebleeding rate after endoscopic variceal ligation (EVL) or drug therapy (non-selective β-blockers ± isosorbide mononitrate [ISMN]), and four trials found decreased variceal rebleeding with combined therapy (EVL+ non-selective β-blockers+ ISMN)[57].

However, some indicators have a negative function in preventing rebleeding. A multicenter, double-blind, parallel study of 158 patients indicated that taking simvastatin besides standard prophylaxis (rest, fluid restriction, preventing infection, regular endoscopic examination, anti-HBV therapy, non-selective β-blocker, *etc*.) did not decrease the rebleeding rate[58]. The rate of variceal rebleeding was not reduced after anticoagulation according to a single-center, prospective cohort study[59]. Worsened liver function or insensitive hemodynamic response to non-selective β-blockers indicated an elevated rebleeding rate[52]. A Chinese study of 3,289 hospitalized patients who underwent EVL indicated that male sex, Child-Pugh score > 7.2, and volume of blood vomited before EVL were independent risk indicators of early rebleeding, while albumin concentration > 31.5 g/L was a protective indicator[60]. Bacterial infection in patients with variceal bleeding was strongly positively related to early rebleeding[61]. Acute-on-chronic liver failure is an independent risk factor of variceal rebleeding[55]. The presence of ascites or hepatic encephalopathy, MELD score > 12, or hepatic-venous pressure gradient >20 mmHg indicated an elevated early (less than 6 wk) rebleeding rate[62].

The above indicators were then filtered and optimized by statistic methods, such as Cox regression or LASSO, and systemically integrated into a function with the help of programming or statistical software such as R, Python, SPSS or SAS. This function was actually a preliminary prediction model.

SIGNIFICANCE OF PREDICTION MODELS

Models predicting disease onset and prognosis play an essential and sometimes surprising role as convenient assistants in planning prophylactic, therapeutic and follow-up strategies. Traditionally, medical data such as medical history, results of physical examination, laboratory tests, imaging and endoscopic information, *etc*. were integrated by doctors' clinical comprehension, or into patients' timeline drafted on a paper to identify how disease progressed and predict the possible prognosis according to the trend in medical indicators. Prediction models free doctors from numerous medical data of patients with different diseases, complications, physical, psychological and social-economic situations. All they need to do is to type prescribed parameters into the model, and click! The results of the onset and prognosis of a given disease are then provided.

Prediction models are currently extensively applied in the medical field to inform individuals and healthcare providers on the risks of developing a particular disease, its outcome, and to guide doctors to make better decisions in mitigating these risks. A recent Chinese study indicated that the MELD score and MELD-Na score, including the R score, were useful in predicting variceal rebleeding[63]. Another study indicated that the MELD-Na score model, which indicates liver function, was more efficient than the MELD model and Child-Pugh score model in predicting rebleeding among cirrhosis patients who underwent EVL.

SAFETY AND PRIVACY CONCERNS

Last but not the least, it is worth noting that models using low-dimensional administrative data outperformed in big data analysis with respect to decreasing information safety and privacy invasion, and according to several studies, the models did not improve when high-resolution, privacy-invasive behavioral data were included[64]. De-ID software (De-ID Data) has been used to assign a study identification number to every enrolled patient. Therefore, criteria, included in the informed consent established by the research review board, for exemption from enrollment were met[33]. The *Drosophila melanogaster* gap gene system gives a good example of demonstrating the significance of error quantification, in which model parameters are optimized against *in situ* immunofluorescence intensities. It can be applied to other studies in various fields with regard to model development.

DISCUSSION

Gastrointestinal (GI) rebleeding is a leading cause of mortality in patients with cirrhosis, as massive GI bleeding can induce hemorrhagic shock, disseminated intravascular coagulation, and opportunistic infections, especially pulmonary infection and spontaneous bacterial peritonitis. Thus, reducing or postponing GI rebleeding is significant. A handy tool for clinicians that can be operated on smart phones or other mobile intelligent devices within seconds to evaluate the GI rebleeding rate is interesting and useful for risk grading. Just type in several common laboratory test indicators, and click on "go" and the rebleeding rate and prognosis of a specific patient are provided.

Our work has merits beyond those of other colleagues. According to our literature retrieval on PubMed, except for only one article published last year indicating the degree of liver stiffness is consistent with GI rebleeding rate in cirrhosis patients[65], there are no other studies on the prediction and prognosis analysis of GI rebleeding. However, the above mentioned exclusive study has limitations. Firstly, it was a prospective cohort study with only 289 patients enrolled in the final analysis, even if PASS 15 was applied to calculate the statistically minimum sample size. In our study, we obtained real-world data from a big data platform collecting many more indicators from 6 hospitals, which were automatically collected. Secondly, our study included patients with esophageal and gastric fundus varices rebleeding, which were the most common varices presented in cirrhosis patients, and the other study only included esophageal varices rebleeding. Finally, the previous study only included patients with HBV-related decompensated cirrhosis, while our data were collected from cirrhosis patients with alcohol-related cirrhosis, autoimmune-related cirrhosis, primary biliary cirrhosis and lipogenic cirrhosis in addition to HBV-related cirrhosis. Following parameter filtering and modeling, our study used a visual nomogram to demonstrate correlations among risk indicators, occurrence and prognosis of GI rebleeding, which provides clinicians with a more explicit demonstration of all indicators and their effects on one page to easily and rapidly evaluate a patient to establish a strategy for further management and follow-up.

CONCLUSION

Modeling is popular using regression analysis and has vast applications in predicting disease occurrence and prognosis. However, modeling and its validation are not the ultimate objective in terms of healthcare providers' clinical participation and patients' health outcomes. They need to be applied and provide convenience for clinical practice. Studies on the application and optimization of these models should be designed and conducted, focusing on the utilization of existing and updated models and their impact on behavior and (self-) management of physicians, healthcare providers, and general individuals[66-67], especially in patients with decompensated cirrhosis at high risk of mortality. For diagnostic and prognostic modeling with higher consistency and efficiency in predicting, treating, and following up decompensated cirrhosis, more comprehensive data and a clearer display mode are needed. We still have a long way to go.

ACKNOWLEDGEMENTS

We would like to express our sincere gratitude to Sylvester RJ, van der Meijden AP, Oosterlinck W, Witjes JA, Bouffioux C, Denis L, Newling DW, Kurth K, Liu L, Liu Q, Xiao N, Zhang Y, Nie Y and Zhu X for allowing us to cite the tables and figures in their studies.

83133_Auto_Edited.docx

ORIGINALITY REPORT

5%
SIMILARITY INDEX

PRIMARY SOURCES

1	www.researchgate.net Internet	44 words — 1%
2	www.ncbi.nlm.nih.gov Internet	38 words — 1%
3	Nuur Azreen Paiman, Azian Hariri, Ibrahim Masood. "Risk prediction model: Statistical and artificial neural network approach", AIP Publishing, 2017 Crossref	21 words — 1%
4	bibbase.org Internet	16 words — < 1%
5	Luca Carmignani, Stefano Picozzi, Robert Stubinski, Stefano Casellato, Giorgio Bozzini, Luca Lunelli, Domenico Arena. "Endoscopic resection of bladder cancer in patients receiving double platelet antiaggregant therapy", Surgical Endoscopy, 2011 Crossref	15 words — < 1%
6	A. J. Vickers. "Decision Curve Analysis: A Novel Method for Evaluating Prediction Models", Medical Decision Making, 11/01/2006 Crossref	14 words — < 1%
7	Tiantian Zhang, Michael Georgiopoulos, Georgios C. Anagnostopoulos. "Pareto-Optimal Model	12 words — < 1%

Selection via SPRINT-Race", IEEE Transactions on Cybernetics, 2018
Crossref

8	bmcurol.biomedcentral.com Internet	12 words — < 1%
9	pubmed.ncbi.nlm.nih.gov Internet	12 words — < 1%

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	< 12 WORDS