Data Science in Healthcare: Leveraging Electronic Health Records for Predictive Analytics

DR. RUPESH SHUKLA[1], AARYESH SHUKLA[2]
[1] *Principal, Department of Computer Science, ILVA Commerce and Science college, Indore / DAVV Indore, India.*
[2] *Department of CSE, RGPV, Bhopal, India.*

Abstract- ***Analysis of electronic health records, often known as EHR analysis, is a technique that is gaining popularity and is being used increasingly frequently to do research on patient data from the real world. When compared to other approaches to study, the use of data that is routinely obtained offers a variety of advantages, such as fewer administrative costs, the opportunity to update studies when new patterns of behavior emerge, and larger sample sizes. EHR analysis comes with its own distinct set of methodological challenges as a result of the fact that the data in question were not collected with the goal of doing research. In this Viewpoint, we elaborate on the necessity of having an in-depth grasp of clinical procedures and outline six potential pitfalls that should be avoided while working with EHR data. Both of these topics are covered in the context of dealing with electronic health record data. In order to do this, we rely on examples from the research that has already been conducted in addition to our own personal experiences. We provide solutions that may be used to avoid or lessen the impact of each of these six concerns, which are as follows: subjective treatment allocation, sample selection bias, imprecise variable definitions, restrictions to deployment, variable measurement frequency, and model over fitting. In conclusion, we have great expectations that this Viewpoint will serve as a roadmap for researchers to follow in order to further increase the methodological rigour of EHR analysis. This optimism is based on the fact that we have high hopes that this Viewpoint will serve as a roadmap.***

Indexed Terms- ***Healthcare, Electronic, Data Science***

I. INTRODUCTION

Electronic health records, more often referred to as EHRs, are rapidly being utilised to conduct out population health surveys, construct categorization and prediction models for decision support, determine which treatment techniques are the most successful, and even duplicate randomised clinical trials. Other common abbreviations for EHRs are EMRs and EHRs. The use of data that has already been gathered is the primary benefit of EHR analysis when contrasted with the use of other "data sources and research methods," such as cohort studies or randomised controlled trials. This is the most significant advantage of using EHR analysis. Due to the presence of this benefit, it is preferable to the utilisation of several other data sources and research methodologies.[1,2] This helps to reduce the quantity of administrative work that is required, as well as the costs and the possibility of bias in the process of selecting samples. Electronic health records, often known as EHRs, are less likely to be affected by inclusion bias than randomised controlled trials are. This is due to the fact that EHRs are more representative of the overall population that is being targeted. This is as a result of the fact that data are collected from any and all persons who interact with health care in any capacity. As time goes on, electronic health records, often known as EHRs, will be able to combine more specific data from patients and will provide access to datasets with increasing quantities. This is especially true in situations when the EHR study in issue comprises large integrated health-care systems or a network of health-care providers that make use of information systems that are interoperable. Despite the fact that there are still significant epidemiological consequences, such as those that are discussed in this Viewpoint, electronic health records (EHRs) provide increasingly complete data from patients.[3,4] At the end of the day, having a sample size that is sufficient enough might give enhanced

"statistical power to carry out subgroup analyses and reduce the chance of generating type II errors."[5]

- Importance of methodological

The complexity of the data included in EHRs can frequently result in methodological issues, which can restrict both the application and validity of the findings from research conducted using EHRs. Even while the examination of EHR data may offer fruitful research opportunities, the practical implementation and scientific credibility of the resulting conclusions are occasionally impeded by a number of factors.[6] To begin, the potential for variance in the method of data gathering utilised for EHRs is significantly higher compared to that which is utilised for clinical archives and clinical trials.[7] This is due to the fact that the process of evaluating numerous variables makes use of a diverse range of technologies and sample frequencies, which in turn increases the possibility that variation will occur. Second, because the vast majority of EHR data studies are carried out using a retrospective methodology, the cohorts, exposures, and outcomes are all characterised using language that focuses on the past.[8,9] This is due to the fact that the vast majority of EHR data is archived in a retroactive fashion. This retrospective technique makes it possible to modify such criteria in light of the results of the analysis, which might lead to findings that can be interpreted in a way that is unsuitable as a consequence of the testing of a large number of hypotheses without the necessary statistical adjustment. [10] In other words, the findings might be misinterpreted as a result of the testing of the hypotheses.[11] Not to mention the fact that clinical practise patterns not only have a considerable influence on the sort of data that is gathered and the quality of that data, but also on the patients who are selected to participate in the research.[12,13] These patterns add biases into the process of selecting samples, which leads to the formation of misleading linkages between variables. It is vital to keep in mind that the data gathered through electronic health records are not collected with the primary goal of undertaking research.[14] This is something that must be kept in mind at all times. Instead, they serve the aim of maintaining all of the information about patients that is gathered during clinical treatment, or in other cases, they play a role in administration such as invoicing. In addition, they are responsible for the safekeeping of patient records.[15] These two roles are equally essential to the whole. Because of this, the evaluation of the information that is included in EHRs should never be carried out on one's own without the support of investigators who have a degree of knowledge that is comparable to that of an authority on the subject matter that is being studied. This degree of competence need to include everything, from the procedures that go into giving therapy to the processing of data, among other things.[16,17]

- Goals of this Viewpoint

In the past twenty years, with the ever-increasing use of electronic health record (EHR) data, previous investigations and our research have observed that issues arising during clinical care (for example, people from minority ethnic groups seeking health care less frequently) are also reflected in the EHR data, thus introducing bias in EHR studies. This bias can be traced back to the fact that people from minority ethnic groups seek health care less frequently.[18] This discrimination may be traced back to the fact that persons who belong to ethnic groups that are underrepresented in the majority are more likely to seek medical attention than those who belong to ethnic groups that are underrepresented in the minority.[19] In the corpus of research that already exists, there does not seem to be an integrated and comprehensive evaluation of the ways in which clinical processes and information system design effect EHR data analysis, in our view. This is something that we believe needs to be addressed. In this Viewpoint, we elaborate on these ideas by combining important clinical concerns with machine learning, epidemiological, and statistical factors.[20] We do this by conducting an analysis of the relevant literature and presenting a number of case stories.[21] There are still others that are necessary for any sort of EHR study, despite the fact that the bulk of the risks that we provide are largely important to models that are based on machine learning. In this Viewpoint, we adopt a mixed approach, based both on the assessment of experts and on a study of the relevant literature, to identify six common clinical and methodological blunders.[22] These mistakes include both clinical and methodological errors.[23,24] Our literature review as well as the references of the publications that were discovered were compared to one another in order to check that we had not overlooked any relevant studies. The authors (CMS, SLH, and LAC) were the ones who made the choice of which papers to include in their review. We believe that by

explaining these six pitfalls of critical importance, we will be able to point researchers in the direction of avoiding making the same errors in future studies, and we hope that the solutions that we propose will enhance the scientific robustness of descriptive and predictive models that incorporate EHR data.[25] Even though some of the points that we present here are elaborations or expansions of essential ideas that have already been published in reporting recommendations such as the Strengthening the Reporting of Observational studies in Epidemiology checklist, we also discuss additional dangers and concepts that it is essential to keep in mind. In addition, the provision of solutions that are capable of being implemented is an additional part of this Viewpoint that adds to the total value of the proposition.[26,27]

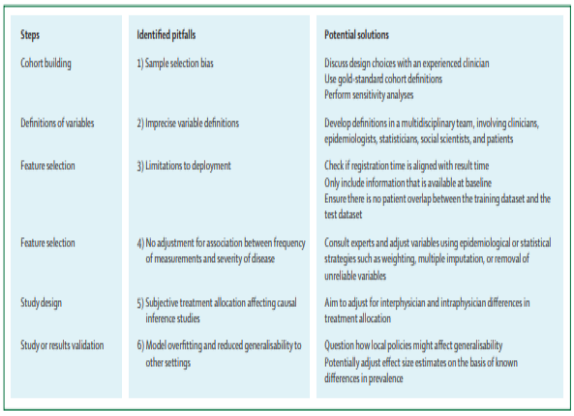
- Pitfalls and solutions

It is possible to make the case that the vast majority of mistakes are not made on purpose or as a result of laziness, but rather because the individual who makes them is unaware of the knowledge gaps in their own understanding. When trying to develop more sophisticated statistical models, it is vital to have an in-depth understanding of how data are gathered and choices are made. [28] This is particularly true when trying to construct more complex statistical models.[29] For instance, it is of the utmost importance to be aware of the patients who are admitted to a hospital or an intensive care unit (ICU), the process by which choices about treatment are made, and the appropriate time for patients to be released from an ICU or from hospital.[30] In spite of the fact that non-medical researchers may be able to gain some insights into these processes by reading about them in the media or by looking them up on the internet, we strongly advise including doctors, nurses, or any other relevant health-care practitioners in the research project at every stage of its development. In addition to this, it is important to note that non-medical researchers may be able to gain some insights into these processes by reading about them in the media or by looking them up on the internet.[31] Because the influence of hospital-specific practises was not taken into account in a number of the earlier study, this oversight ultimately led to issues with the external validity of the studies, and in some instances, with the internal validity of the studies as well. in addition to the need that the influence of domain knowledge on the policy of local health-care unit research be recognised in EHR research.[32,33]Protocols that are adhered to in hospitals may decide which data are collected, when they are collected, and how they are collected. Additionally, these protocols may also influence whether or not certain data are obtained. For instance, in order for the electronic health record (EHR) to provide an odd blood test result, the patient must have previously had a blood test.[34] Therefore, the identification of people with aberrant results can be related with both local processes and different testing frequencies for distinct patient groups. This is because of the fact that the testing frequencies might vary. [35] However, despite the fact that this issue of missing data is extremely common in research, it is often ignored. This results in bias in the selection of samples, which is very seldom adjusted for. The criteria that must be met in order for a patient to be admitted to an intensive care unit (ICU) will have an effect on the findings of any research that involves the analysis of EHR data obtained from patients treated in an ICU. These criteria are different from one intensive care unit (ICU) to the next, and they may even shift based on the circumstances within a single institution.[36] During the COVID-19 pandemic, these changing circumstances were easily evident when the number of patients who required critical care unit beds exceeded the available beds in the hospitals. This caused a shortage of beds. When a patient is ready for departure from the hospital (selective censoring), for instance, or when a patient should be readmitted from the ward to the critical care unit, are both instances of clinical decisions that are sensitive to the subjectivity of the health care practitioners who are making them.[37,38] In the majority of these instances, the decisions about triage and other topics could be impacted in some way by elements of the healthcare practitioners themselves, such as their clinical experience and cultural differences, as well as by current events in general, such as the pandemic. We strongly advise doing regular reviews of the design choices and research assumptions with physicians or other health-care professionals who are familiar with the local practises in the area. Because of this, you will be able to find a solution to the problems that were brought up before. In addition, studies that attempt to explore or predict outcomes that occur from treatment decisions (for example, which patients should be administered renal replacement therapy) should have causal inference frameworks[39]. These frameworks help

researchers determine which patients should be treated with renal replacement therapy and how often. It is possible that these frameworks will assist decide which individuals need to have renal replacement therapy delivered to them. We have provided an overview of the many solutions that are conceivable, as well as a list of the probable difficulties that may occur in the future (figure). The dangers are structured in a way that corresponds to the stage of the analysis in which they are most likely to occur. This ensures that the information is easy to find and understand .[40]

- Sample Selection Bias

The step of data analysis known as "building cohorts" is a challenging one since it requires converting a large number of case descriptions into specific data criteria. As a direct consequence of this, this stage is very challenging.[41] The capacity of the researchers to precisely interpret and translate these definitions, as well as the quality of the definition that was used in the literature to correctly identify instances, are both variables that have an impact on the composition of the cohort. Importantly, sample selection bias may occur if these case criteria are either too strict, which may result in the exclusion of a subgroup, or overly wide, which may result in an increase in the number of patients who have been mistakenly identified as being part of the cohort.[42] Both of these scenarios may lead to the same end result: an increase in the number of patients who have been misclassified as being a member of the cohort. Studies that used different methods to identify patients with sepsis are a prominent example of definitions that are too general.[43] These studies include "the quick Sequential Organ Failure Assessment (qSOFA) score, codes proposed by Martin and colleagues32 (also known as the Martin methodology) , codes proposed by Angus and van der Poll33 (also known as the Angus methodology), and the systemic inflammatory response syndrome score. In spite of the fact that it has been shown that these strategies may be useful as proxies for sepsis, the specificity of such approaches is missing. In the meanwhile, it has been shown on several times that the capacity to recognise individuals with sepsis in a range of contexts is varied and somewhat discriminatory.[44] This is the case even though sepsis is a rather common condition. This is shown by areas under the receiver operating characteristic curve (AUROC) values for qSOFA scores of 2 or above that lie between 0.6 and 0.8.38–40 In addition to this, the nature of the sickness itself creates the likelihood of inaccurate classification being applied to the patient.[45] When adopting the gold-standard criterion for sepsis, which is 3.41, this is one illustration of this phenomenon. However, the use of more inclusive definitions of what constitutes a suspected infection rather than restrictive ones (for example, whether a positive nitrite urine test is sufficient or whether other signs of tissue invasion also need to be present) could significantly affect both the size of the cohort and the performance of the model . A positive nitrite urine test is one of the most important components of the Sepsis-3 definition.[46] When carrying out research, it is essential to take into consideration essential criteria such as cohort definitions and external validation. The poor performance of the Epic Sepsis Model serves as an illustration of why this is the case. It is important to note that even when algorithms are used to identify illnesses like sepsis, retrospective examinations of people who are presumed to have sepsis may indicate significant misclassification.[47] This is something that has to be taken into account, so keep that in mind. A further example of the use of criteria that are overly general may be found in the field of mortality prediction. It has been shown that the produced cohorts that are used in the different research projects are comparable to one another in terms of their characteristics.



Steps	Identified pitfalls	Potential solutions
Cohort building	1) Sample selection bias	Discuss design choices with an experienced clinician Use gold-standard cohort definitions Perform sensitivity analyses
Definitions of variables	2) Imprecise variable definitions	Develop definitions in a multidisciplinary team, involving clinicians, epidemiologists, statisticians, social scientists, and patients
Feature selection	3) Limitations to deployment	Check if registration time is aligned with result time Only include information that is available at baseline Ensure there is no patient overlap between the training dataset and the test dataset
Feature selection	4) No adjustment for association between frequency of measurements and severity of disease	Consult experts and adjust variables using epidemiological or statistical strategies such as weighting, multiple imputation, or removal of unreliable variables
Study design	5) Subjective treatment allocation affecting causal inference studies	Aim to adjust for interphysician and intraphysician differences in treatment allocation
Study or results validation	6) Model overfitting and reduced generalisability to other settings	Question how local policies might affect generalisability Potentially adjust effect size estimates on the basis of known differences in prevalence

Figure: 1 In studies that use electronic health records, there are often occurring clinical and methodological difficulties, as well as possible remedies.[48]

variable, despite the fact that all of the studies used the exact "same dataset (Medical Information Mart for Intensive Care III) and investigated the performance of the exact same model (i.e., mortality

prediction). This was the case even though all of the publications were published in the same year.[49]Age limits (for instance, excluding people younger than 18 years of age), the exclusion of individuals who had numerous stays in the intensive" care unit, or the necessity of certain measures, such as those of infection indicators, were some of the factors that contributed to the diversity of the inclusion and exclusion criteria. Other factors that contributed to this diversity included the necessity of certain measures, such as those of infection indicators. We strongly suggest that you always make use of the definitions that have been thoroughly reviewed; nevertheless, you should bear in mind that even these definitions are prone to mistake and are not cast in stone.[50] The examples that were shown before provided the foundation for this advice. In addition, if the performance of the model is supplied, it need to be compared with the algorithms that are now in use, regardless of whether such algorithms are based on the opinions of specialists or on machine learning. In the event that comparisons are made using algorithms that are derived from machine learning, it is essential that differences in terminology be researched. We recommend addressing any possible limits, emphasising any potential implications on the robustness of the findings, and giving a clear explanation of the criteria that were utilised for the selection of cohorts. We also advocate addressing any potential implications on the robustness of the results. In the event that there are no criteria that are universally acknowledged as the benchmark, we suggest doing a sensitivity analysis in order to ascertain which approach is most effective in locating the population that is the subject of the investigation.[51,52]

- Imprecise Variable Definitions

After the building of the cohort, the next key pitfall to look out for is the definition of the variables incorrectly. The degree to which the definitions are sensitive and particular has a significant bearing on whether or not the analyses correctly depict how things are done in the actual world. Because of the impact of this factor, definitions need to be subjected to a close and careful analysis. This is also true when forming cohorts.[53] In addition, it is essential to have a clear understanding of the difference between the data that is being collected for the purposes of billing and the data that is being recorded for therapeutic treatment purposes. For instance, it is feasible that the procedure codes may not cover all of the available procedures; rather, they may just include those that can be invoiced. Before settling on a choice about the outcome measure, it is essential to take into account both the epidemiological ramifications and the clinical context of the study. An example that comes up rather often is the conundrum of whether or not to use in-hospital events (like hospital mortality), as opposed to a fixed time point (like 28-day mortality), as a measure of patient outcomes. When investigating the consequences of direct hospital treatments, it is customary to focus on events that take place within the hospital.[54] An effect that should be examined using in hospital events is the influence of using prophylactic heparin on the incidence of venous thromboembolism in patients being treated in hospitals. This is only one example of an effect that should be studied using inhospital events. On the other hand, if you are interested in the incidence of postsurgical thrombosis, you should choose a predetermined time period as your starting point for the study. This is due to the fact that the length of stay a patient has in the hospital is related to the likelihood that they may develop thrombosis during their stay. In general, we recommend consulting the expertise of a group of experts from a variety of fields in order to get advice on definitions. Patients should be included on this team in addition to specialists who are knowledgeable in relevant topic areas. Examples of such experts are medical professionals, epidemiologists, statisticians, and social scientists.

- Limits To Deployment

The fact that the data that is available in EHRs is unable to be simply turned into clinical practise offers a substantial impediment for the implementation and deployment in the actual world of machine learning or other prediction models that were generated from EHR research. This issue arises whenever the data format of EHRs deviates from the practise that takes place in the real world, as well as if there is a lack of time-stamped results. Time-stamped data, for instance, may be judged to be available at the moment of measurement and then reviewed as if it had been gathered at that point in time during a retrospective analysis. This evaluation may be performed as if the data had been obtained at that point in time. On the other hand, it's likely that doctors and other people who make decisions won't really have access to this data at the time that

was first expected. This sort of disparity may appear, for example, when the findings of blood tests or blood cultures are used.[55] This is because the results may be included in the analysis with the timestamp of the registration of the blood draw rather than the time at which they were made available to the doctors. A further illustration of this would be the exclusion of patients who have very long hospital stays since this information is not available early on in the course of the patient's hospital stay (which is when the bulk of algorithms are used). In addition, International Classification of Diseases codes 55 are not often allocated to patients until after they have been released from the hospital or have gone away, and the time stamps that are connected with the occurrence of these codes might vary. Because of this variation in timing, the use of these codes is not something that would be suitable for models that utilise hospital admission as their baseline. The performance of the model may be overstated in many of the aforementioned situations, and it may not be possible to reproduce its results in applications that take place in the real world. The appearance of training data in the test datasets is a separate but related issue, which is connected to the leaking of data, which is a distinct but related one. This is a different issue, yet there is a connection between both. This leakage may often take place if repeated admissions from the same patients are used, or if the timeseries data from a single patient are not restricted to either the training set or the test set, but instead appear in both. Alternatively, this leakage can take place if repeated admissions from the same patients are used. There is a serious flaw in the methodology that is referred to as leakage of data and it has the potential to result in an inaccurately high estimate of the performance of the model. As a direct result of this, the performance of the model as well as its utility in the application of clinical research are both diminished. The challenges that are associated with deployment may be avoided by first acquiring an understanding of when data will become accessible in real time, and then selecting features based on the data that will be made available to physicians. This will allow for the deployment process to go more smoothly. In addition, data from time series as well as data from a large number of admissions to an institution such as a hospital or intensive care unit should be randomly assigned to either the training dataset or the test dataset. Because of this, there will be no leakage of patient data from the training dataset into the test dataset .

- variable measuring frequency

The natural link that exists between the frequency of measurements and the severity of the illness is another issue that is often neglected, despite the fact that it is unavoidably there. When a patient's health is considered to be unstable, practitioners may typically issue orders for more laboratory tests or analyse records of vital sign readings on a more regular basis. It is essential that, regardless of whether one is creating predictive or descriptive models, the link that exists between the frequency of measurements and the severity of the disease be taken into consideration. For example, if patients who had a substantial proportion of missing (that is, not executed) data were deleted from the study, this may have an effect on the validity of the model. Patients who had a significant percentage of missing (that is, not executed) data were omitted from the research. This exclusion often results in a biassed model, which, in the same way as sample selection bias (trap 1) has the ability to overstate the seriousness of the condition, it also has the potential to be misleading. It is vital to keep in mind that imputation, in which the values of the cohort's mean or median are used, is not a technique that may be used to compensate for missing data. This is because the frequency of taking measurements is not a completely arbitrary occurrence. This example highlights how data that are regularly obtained implicitly reflect the judgements of the doctors and how there may be a significant degree of diversity amongst the various providers of medical care. If there is a substantial amount of variation in clinical practises and if the behaviour of doctors were to change at some time in the future, a model that was trained on these kinds of data may be more likely to have poor performance. As a consequence of this, we recommend involving statisticians in order to discuss appropriate epidemiological (for example, weighting) or statistical strategies (for example, multiple imputation, or removal of highly unreliable variables entirely), and we also recommend involving clinicians in order to identify circumstances in which variable measurement frequency could result in a biassed analysis. This is due to the fact that we advocate for incorporating doctors in the process of determining the conditions under which varied measurement frequency may lead to an incorrect interpretation of the data .

There is a large amount of subjectivity involved in the therapy assigning process

In most instances, the goal of research using causal inference is to make an informed prediction as to the unobserved, counterfactual treatment outcome, which then makes it possible to evaluate the impacts of the treatment. In order to quantify these consequences, it is required to first identify and then take into consideration all of the variables that are associated with the treatment allocation. Only then can these impacts be quantified. It has been shown that this approach can function in the same way as randomised controlled clinical trials (RCTs), but despite the fact that it may provide certain advantages, it is not without limitations. Studies that utilise causal inference begin with the supposition that every factor that has a role in treatment allocation is really observed. This is the starting point for these types of investigations. However, previous research has shown that the allocation of treatment can be affected by both differences in interphysician decision making (i.e., different physicians prescribing different treatments in the same clinical setting) and differences in intraphysician decision making (i.e., bias on patient's socioeconomic factors, such as race and ethnicity). In other words, different physicians can prescribe different treatments in the same clinical setting . When conducting studies that are based on EHRs, researchers should employ causal frameworks whenever it is practical to do so in order to minimise introducing bias due to confounding variables. This is because causal frameworks allow for more accurate interpretation of results. Before initiating exploratory studies, it is strongly advised that causal diagrams be developed to illustrate the team's knowledge of the process by which the data were produced (for instance, by making use of directed acyclic networks). This may be done before moving on to the actual exploratory analyses. These diagrams might be helpful in identifying any missing confounding factors or other crucial components that need to be taken into consideration over the course of the investigation. Even though controlling for confounders is a crucial part of replicating randomization, electronic health records (EHRs) often fail to capture important clinical indicators or socioeconomic determinants of health. This is despite the fact that mimicking randomization is a core component of research. In light of the circumstances presented above, we highly recommend giving serious thought to the question of whether or not it is even conceivable to carry out a study involving causal inference. When in doubt, sensitivity studies should be undertaken to investigate the potential effect of unmeasured confounders based on data from earlier research. These analyses should be performed to analyse the possible influence of unmeasured confounders .

- Model Over fitting And Decreased Capacity To Generalise Results

There is a possibility that the results cannot be generalized beyond the data source from which they were produced because of the existence of significant disparities across the various institutions and regions. It is essential to keep in mind that not all research and models need to be generalizable since there is a possibility of model over fitting happening. Therefore, it is necessary to inquire of researchers the answer to the issue of whether or whether the treatments or results of interest are impacted by local practicing patterns. For instance, if researchers working inside of a hospital wanted to determine how the handoff of a patient from one service to another influences clinical outcomes, the best performing model could be constructed without attaching an excessive amount of emphasis to issues resulting from inadequate generalisability. This would allow for the model to have the most predictive power. If, on the other hand, the objective is to solve this problem outside of the limits of a particular institution, then it is necessary to conduct external validation using a separate cohort. In addition, changes in the frequency of important factors or the absence of these variables should be highlighted across clinical settings and correctly accounted for whenever it is possible to do so.[56]
The issue of external validity is not a yes-or-no statement; rather, it is concerned with establishing which specific clinical settings the analysis is appropriate to. This is not a simple yes-or-no argument. When trying to derive the generalizability of models, a causal diagram may be of aid since it makes it evident which correlations in the data are most likely to differ depending on the organisation or the region. Additionally, for the purpose of evaluating the performance of the model, acceptable performance measures that are not affected by class imbalance should be applied. When there is an imbalance in the frequency of a feature or outcome of interest, such as when there is a very low hospital death rate for certain medical conditions, it is

recommended to avoid using metrics that are prone to class imbalance as performance measures for the model . One example of this would be when there is a very low hospital mortality rate for certain medical diseases. Accuracy is a good illustration of this concept. In addition, reporting just aggregate measures of discrimination (such as AUROC) might mask a lower therapeutic usefulness since, for instance, the sensitivity or specificity may not be high enough. It's likely that utilising measurements that are based on a single operating point, like the F1 score, which is aimed to strike a balance between accuracy and recall, would be more informative than using measures that are based on many operating points. In the end, the assessment of the model has to be modified in accordance with the use scenario for which the system was developed (for instance, screening as opposed to therapeutic guidance).

CONCLUSION

We discovered six prevalent methodological and clinical difficulties that affect the robustness, validity, and reproducibility of research that makes use of electronic health record data (EHR data). The most significant challenges consist of a biassed sample selection, imprecise variable definitions, deployment limits, a lack of adjustment for the relationship between frequency of measurements and severity of sickness, subjective treatment allocation, and limited generalisability of findings. These challenges were encountered because the sample selection process was biassed. Because of all of these different factors, it is challenging to generalise the results. Although this list is not exhaustive and the possible solutions to these worries do not apply in every circumstance, we have penned this Viewpoint in the hopes that it will bring attention to a number of critical risks associated with EHR data and encourage researchers to take them into consideration when working with EHR-based research .

REFERENCES

[1] Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. Sci Data 2019; 6: 96.

[2] Beaulieu-Jones BK, Yuan W, Brat GA, et al. Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? NPJ Digit Med 2021; 4: 62.

[3] Bonomi S. The electronic health record: a comparison of some European countries. In: Ricciardi F, Harfouche A, eds. Information and communication technologies in organizations and society. Lecture notes in information systems and organisation, vol 15. Cham: Springer, 2016: 33–50.

[4] Tambone V, Boudreau D, Ciccozzi M, et al. Ethical criteria for the admission and management of patients in the ICU under conditions of limited medical resources: a shared international proposal in view of the COVID-19 pandemic. Front Public Health 2020; 8: 284.

[5] American Thoracic Society. Fair allocation of intensive care unit resources. Am J Respir Crit Care Med 1997; 156: 1282–301.

[6] Curtis JR, Vincent J-L. Ethics and end-of-life care for adults in the intensive care unit. Lancet 2010; 376: 1347–53.

[7] Piers RD, Azoulay E, Ricou B, et al. Perceptions of appropriateness of care among European and Israeli intensive care unit nurses and physicians. JAMA 2011; 306: 2694–703.

[8] Usman OA, Usman AA, Ward MA. Comparison of SIRS, qSOFA, and NEWS for the early identification of sepsis in the emergency department. Am J Emerg Med 2019; 37: 1490–97.

[9] Singer M, Deutschman CS, Seymour CW, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA 2016; 315: 801–10.

[10] Johnson AEW, Aboab J, Raffa JD, et al. A comparative analysis of sepsis identification methods in an electronic database. Crit Care Med 2018; 46: 494–99.

[11] Wong A, Otles E, Donnelly JP, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med 2021; 181: 1065–70.

[12] Lapsley I, Melia K. Clinical actions and financial constraints: the limits to rationing intensive care. Sociol Health Illn 2001; 23: 729–46.

[13] Trentini F, Marziano V, Guzzetta G, et al. The pressure on healthcare system and intensive

care utilization during the COVID-19 outbreak in the Lombardy region of Italy: a retrospective observational study in 43 538 hospitalized patients. Am J Epidemiol 2022; 191: 137–46.

[14] Jacoba CMP, Celi LA, Silva PS. Biomarkers for progression in diabetic retinopathy: expanding personalized medicine through integration of AI with electronic health records. Semin Ophthalmol 2021; 36: 250–57.

[15] Robles Arévalo A, Maley JH, Baker L, et al. Data-driven curation process for describing the blood glucose management in the intensive care unit. Sci Data 2021; 8: 80. 3 Sauer CM, Gómez J, Botella MR, et al. Understanding critically ill sepsis patients with normal serum lactate levels: results from US and European ICU cohorts. Sci Rep 2021; 11: 20076.

[16] Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med 2018; 24: 1716–20.

[17] 1. Navaneetha Krishnan Rajagopal, Mankeshva Saini, Rosario Huerta-Soto, Rosa Vílchez-Vásquez, J. N. V. R. Swarup Kumar, Shashi Kant Gupta, Sasikumar Perumal, "Human Resource Demand Prediction and Configuration Model Based on Grey Wolf Optimization and Recurrent Neural Network", Computational Intelligence and Neuroscience, vol. 2022, Article ID 5613407, 11 pages, 2022. https://doi.org/10.1155/2022/5613407

[18] 2. Navaneetha Krishnan Rajagopal, Naila Iqbal Qureshi, S. Durga, Edwin Hernan Ramirez Asis, Rosario Mercedes Huerta Soto, Shashi Kant Gupta, S. Deepak, "Future of Business Culture: An Artificial Intelligence-Driven Digital Framework for Organization Decision-Making Process", Complexity, vol. 2022, Article ID 7796507, 14 pages, 2022. https://doi.org/10.1155/2022/7796507

[19] Eshrag Refaee, Shabana Parveen, Khan Mohamed Jarina Begum, Fatima Parveen, M. Chithik Raja, Shashi Kant Gupta, Santhosh Krishnan, "Secure and Scalable Healthcare Data Transmission in IoT Based on Optimized Routing Protocols for Mobile Computing Applications", Wireless Communications and Mobile Computing, vol. 2022, Article ID 5665408, 12 pages, 2022. https://doi.org/10.1155/2022/5665408

[20] Rajesh Kumar Kaushal, Rajat Bhardwaj, Naveen Kumar, Abeer A. Aljohani, Shashi Kant Gupta, Prabhdeep Singh, Nitin Purohit, "Using Mobile Computing to Provide a Smart and Secure Internet of Things (IoT) Framework for Medical Applications", Wireless Communications and Mobile Computing, vol. 2022, Article ID 8741357, 13 pages, 2022. https://doi.org/10.1155/2022/8741357

[21] Bramah Hazela et al 2022 ECS Trans. 107 2651 https://doi.org/10.1149/10701.2651ecst

[22] Ashish Kumar Pandey et al 2022 ECS Trans. 107 2681 https://doi.org/10.1149/10701.2681ecst

[23] G. S. Jayesh et al 2022 ECS Trans. 107 2715 https://doi.org/10.1149/10701.2715ecst

[24] Shashi Kant Gupta et al 2022 ECS Trans. 107 2927 https://doi.org/10.1149/10701.2927ecst

[25] S. Saxena, D. Yagyasen, C. N. Saranya, R. S. K. Boddu, A. K. Sharma and S. K. Gupta, "Hybrid Cloud Computing for Data Security System," 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), 2021, pp. 1-8, doi: 10.1109/ICAECA52838.2021.9675493.

[26] S. K. Gupta, B. Pattnaik, V. Agrawal, R. S. K. Boddu, A. Srivastava and B. Hazela, "Malware Detection Using Genetic Cascaded Support Vector Machine Classifier in Internet of Things," 2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA), 2022, pp. 1-6, doi: 10.1109/ICCSEA54677.2022.9936404.

[27] Natarajan, R.; Lokesh, G.H.; Flammini, F.; Premkumar, A.; Venkatesan, V.K.; Gupta, S.K. A Novel Framework on Security and Energy Enhancement Based on Internet of Medical Things for Healthcare 5.0. *Infrastructures*2023, *8*, 22. https://doi.org/10.3390/infrastructures8020022

[28] V. S. Kumar, A. Alemran, D. A. Karras, S. Kant Gupta, C. Kumar Dixit and B. Haralayya, "Natural Language Processing using Graph Neural Network for Text Classification," 2022 International Conference on Knowledge Engineering and Communication Systems (ICKES), Chickballapur, India, 2022, pp. 1-5, doi: 10.1109/ICKECS56523.2022.10060655.

[29] M. Sakthivel, S. Kant Gupta, D. A. Karras, A. Khang, C. Kumar Dixit and B. Haralayya, "Solving Vehicle Routing Problem for Intelligent Systems using Delaunay Triangulation," 2022 International Conference on Knowledge Engineering and Communication Systems (ICKES), Chickballapur, India, 2022, pp. 1-5, doi: 10.1109/ICKECS56523.2022.10060807.

[30] S. Tahilyani, S. Saxena, D. A. Karras, S. Kant Gupta, C. Kumar Dixit and B. Haralayya, "Deployment of Autonomous Vehicles in Agricultural and using Voronoi Partitioning," 2022 International Conference on Knowledge Engineering and Communication Systems (ICKES), Chickballapur, India, 2022, pp. 1-5, doi: 10.1109/ICKECS56523.2022.10060773.

[31] V. S. Kumar, A. Alemran, S. K. Gupta, B. Hazela, C. K. Dixit and B. Haralayya, "Extraction of SIFT Features for Identifying Disaster Hit areas using Machine Learning Techniques," 2022 International Conference on Knowledge Engineering and Communication Systems (ICKES), Chickballapur, India, 2022, pp. 1-5, doi: 10.1109/ICKECS56523.2022.10060037.

[32] V. S. Kumar, M. Sakthivel, D. A. Karras, S. Kant Gupta, S. M. Parambil Gangadharan and B. Haralayya, "Drone Surveillance in Flood Affected Areas using Firefly Algorithm," 2022 International Conference on Knowledge Engineering and Communication Systems (ICKES), Chickballapur, India, 2022, pp. 1-5, doi: 10.1109/ICKECS56523.2022.10060857.

[33] Parin Somani, Sunil Kumar Vohra, Subrata Chowdhury, Shashi Kant Gupta. "Implementation of a Blockchain-based Smart Shopping System for Automated Bill Generation Using Smart Carts with Cryptographic Algorithms." CRC Press, 2022. https://doi.org/10.1201/9781003269281-11.

[34] Shivlal Mewada, Dhruva Sreenivasa Chakravarthi, S. J. Sultanuddin, Shashi Kant Gupta. "Design and Implementation of a Smart Healthcare System Using Blockchain Technology with A Dragonfly Optimization-based Blowfish Encryption Algorithm." CRC Press, 2022. https://doi.org/10.1201/9781003269281-10.

[35] Ahmed Muayad Younus, Mohanad S.S. Abumandil, Veer P. Gangwar, Shashi Kant Gupta. " AI-Based Smart Education System for a Smart City Using an Improved Self-Adaptive Leap-Frogging Algorithm." CRC Press, 2022. https://doi.org/10.1201/9781003252542-14.

[36] Rosak-Szyrocka, J., Żywiołek, J., & Shahbaz, M. (Eds.). (2023). Quality Management, Value Creation and the Digital Economy (1st ed.). Routledge. https://doi.org/10.4324/9781003404682

[37] Dr. Shashi Kant Gupta, Hayath T M., Lack of it Infrastructure for ICT Based Education as an Emerging Issue in Online Education, TTAICTE. 2022 July; 1(3): 19-24. Published online 2022 July, doi.org/10.36647/TTAICTE/01.03.A004

[38] Hayath T M., Dr. Shashi Kant Gupta, Pedagogical Principles in Learning and Its Impact on Enhancing Motivation of Students, TTAICTE. 2022 October; 1(2): 19-24. Published online 2022 July, doi.org/10.36647/TTAICTE/01.04.A004

[39] Shaily Malik, Dr. Shashi Kant Gupta, “The Importance of Text Mining for Services Management”, TTIDMKD. 2022 November; 2(4): 28-33. Published online 2022 November doi.org/10.36647/TTIDMKD/02.04.A006

[40] Dr. Shashi Kant Gupta, Shaily Malik, “Application of Predictive Analytics in Agriculture”, TTIDMKD. 2022 November; 2(4): 1-5. Published online 2022 November doi.org/10.36647/TTIDMKD/02.04.A001

[41] Dr. Shashi Kant Gupta, Budi Artono, “Bioengineering in the Development of Artificial Hips, Knees, and other joints. Ultrasound, MRI, and other Medical Imaging Techniques”, TTIRAS. 2022 June; 2(2): 10–15. Published online 2022 June doi.org/10.36647/TTIRAS/02.02.A002

[42] Dr. Shashi Kant Gupta, Dr. A. S. A. Ferdous Alam, “Concept of E Business Standardization and its Overall Process” TJAEE 2022 August; 1(3): 1–8. Published online 2022 August

[43] A. Kishore Kumar, A. Alemran, D. A. Karras, S. Kant Gupta, C. Kumar Dixit and B. Haralayya, "An Enhanced Genetic Algorithm for Solving Trajectory Planning of Autonomous Robots," 2023 IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS), Raichur, India, 2023, pp. 1-6, doi: 10.1109/ICICACS57338.2023.10099994

[44] S. K. Gupta, V. S. Kumar, A. Khang, B. Hazela, N. T and B. Haralayya, "Detection of Lung Tumor using an efficient Quadratic Discriminant Analysis Model," 2023 International Conference on Recent Trends in Electronics and Communication (ICRTEC), Mysore, India, 2023, pp. 1-6, doi: 10.1109/ICRTEC56977.2023.10111903.

[45] S. K. Gupta, A. Alemran, P. Singh, A. Khang, C. K. Dixit and B. Haralayya, "Image Segmentation on Gabor Filtered images using Projective Transformation," 2023 International Conference on Recent Trends in Electronics and Communication (ICRTEC), Mysore, India, 2023, pp. 1-6, doi: 10.1109/ICRTEC56977.2023.10111885.

[46] S. K. Gupta, S. Saxena, A. Khang, B. Hazela, C. K. Dixit and B. Haralayya, "Detection of Number Plate in Vehicles using Deep Learning based Image Labeler Model," 2023 International Conference on Recent Trends in Electronics and Communication (ICRTEC), Mysore, India, 2023, pp. 1-6, doi: 10.1109/ICRTEC56977.2023.10111862.

[47] S. K. Gupta, W. Ahmad, D. A. Karras, A. Khang, C. K. Dixit and B. Haralayya, "Solving Roulette Wheel Selection Method using Swarm Intelligence for Trajectory Planning of Intelligent Systems," 2023 International Conference on Recent Trends in Electronics and Communication (ICRTEC), Mysore, India, 2023, pp. 1-5, doi: 10.1109/ICRTEC56977.2023.10111861.

[48] Shashi Kant Gupta, Olena Hrybiuk, NL Sowjanya Cherukupalli, Arvind Kumar Shukla (2023). Big Data Analytics Tools, Challenges and Its Applications (1st Ed.), CRC Press. ISBN 9781032451114

[49] Shobhna Jeet, Shashi Kant Gupta, Olena Hrybiuk, Nupur Soni (2023). Detection of Cyber Attacks in IoT-based Smart Cities using Integrated Chain Based Multi-Class Support Vector Machine (1st Ed.), CRC Press. ISBN 9781032451114

[50] Parin Somani, Shashi Kant Gupta, Chandra Kumar Dixit, Anchal Pathak (2023). AI-based Competency Model and Design in the Workforce Development System (1st Ed.), CRC Press. https://doi.org/10.1201/9781003357070

[51] Shashi Kant Gupta, Alex Khang, Parin Somani, Chandra Kumar Dixit, Anchal Pathak (2023). Data Mining Processes and Decision-Making Models in Personnel Management System (1st Ed.), CRC Press. https://doi.org/10.1201/9781003357070

[52] Alex Khang, Shashi Kant Gupta, Chandra Kumar Dixit, Parin Somani (2023). Data-driven Application of Human Capital Management Databases, Big Data, and Data Mining (1st Ed.), CRC Press. https://doi.org/10.1201/9781003357070

[53] Chandra Kumar Dixit, Parin Somani, Shashi Kant Gupta, Anchal Pathak (2023). Data-centric Predictive Modelling of Turnover Rate and New Hire in Workforce Management System (1st Ed.), CRC Press. https://doi.org/10.1201/9781003357070

[54] Anchal Pathak, Chandra Kumar Dixit, Parin Somani, Shashi Kant Gupta (2023). Prediction of Employee's Performance Using Machine Learning (ML) Techniques (1st Ed.), CRC Press. https://doi.org/10.1201/9781003357070

[55] Worakamol Wisetsri, Varinder Kumar, Shashi Kant Gupta, "Managerial Autonomy and Relationship Influence on Service Quality and Human Resource Performance", Turkish Journal of Physiotherapy and Rehabilitation, Vol. 32, pp2, 2021.