**Journal Requirements:**  
1. We ask that a manuscript source file is provided at Revision. Please upload your manuscript file as a .doc, .docx, .rtf or .tex.

We added the source manuscript to the rebuttal.

**Reviewer #1:** This is an interesting study and the authors have presented a three-step pitfalls retrospective case-control study. However, I believe the paper has some flaws in some text, and I believe this interesting study has not been investigated to its full potential. I've included some comments below.

**REPLY: Thank you for your kind words and suggestions.**

1) Abstraction: The abstract should be rewritten as a technical summary, while the author's summary should be written in a less technical style for a broader audience.  
**REPLY: We rewrote the abstract and technical summary to match their targeted audience.**

***«Abstract   
Causal inference frameworks allow to flexibly estimate treatment effects of medical interventions with machine learning methods from electronic health records (EHRs).***

***However, researchers have to be aware of many pitfalls. We designed a tutorial to showcase common pitfalls and potential solutions by using an example in Medical Information Mart for Intensive Care database (MIMIC-IV) on the effect of albumin on mortality in sepsis with outcome data from randomized controlled trials (RCT) as gold-standard. First, we fixed the study design according to the target trial concept and using the PICOT framework (Population (patients with sepsis), Intervention (combination of crystalloids and albumin for fluid resuscitation), Control (crystalloids only), Outcome (28-day mortality), Time(intervention start within 24h of a dmission)). Second, we selected the confounding variables based on expert knowledge. Third, we fitted multiple models with varying assumptions to asses their influence. Reducing treatment initiation time from 72h to 6h removed immortal time bias. A doubly robust estimator (AIPW) with random forests proved to be the most reliable estimator. By increasingly adding confounders the results converged to the RCT results (0%, 95% CI -1% to 1%, and 0%, 95% -5% to 5%, respectively). Subgroup analyses showed that treatment efficacy of albumin was better for patients >60 years old, males, and patients with septic shock. In conclusion, we show that these steps are all important to build a valid decision-making, causal inference framework and that misspecifications can drastically skew the results.***

***Author summary   
Causal inference frameworks have the potential to inform clinical decision-making using routine care data, especially where randomized controlled trials are not feasible. These frameworks use machine learning predictions and readily available EHR or claims data for tailoring individualized care. However, certain rules need to be followed to account for the heterogeneity of effect as assignment of interventions are confounded by indication, which introduces many pitfalls, particularly in time-varying data. In a tutorial spirit, we make all the code and data openly available. We split our approach into three steps: Step 1) Study design; With the common pitfalls selection bias, i.e. information is not equally collected across treatment and control patients, and immortal time bias, i.e. improper alignment of the inclusion defining event and the intervention time. Step 2) Identification of the causal assumptions and categorization of confounders. Step 3) Estimation of the causal effect of interest by correct aggregation of confounders and selection of an appropriate statistical model. Step 4) Vibration analysis which includes the assessing the analysis’ robustness to assumptions, and finally Step 5) Treatment heterogeneity which includes exploring the treatment effect in subgroups.»***  
  
2) Introduction: The introduction should be improved to better explain the work's motivation and methodology, and the literature review should be more comprehensive, supplementing the advantages and weaknesses of related research. The first sentence of the introduction is hyperbole and not needed.  
**REPLY:** Thank you for your valuable input. We agree that the motivation for our tutorial can be better displayed. Thus, we removed the first sentence of the introduction and reformulated the introduction to better present the motivation. Particularly, we displayed the need to study treatment heterogeneity by showcasing the QRISK example that was very briefly presented in the appendix in our initial submission. We also emphasize our contributions with respect to the previous work.

***«Clinical equipoise extends beyond merely predicting the occurrence of an event; it involves estimating treatment effects in populations, particularly within subgroups. Randomized controlled trials (RCTs) are the gold standard to address these questions, but face issues of their own. In this context, causal inference frameworks emerge as an ideal tool to tackle such questions in observational data. However, numerous considerations must be carefully addressed to ensure the validity of the inferences obtained from these frameworks.***

***Beyond Individualized Predictions: Machine Learning Challenges in Medicine [1–5] Machine learning plays a pivotal role in individualized medicine, showcasing superior performance over traditional rule-based clinical scores in predicting a patient’s readmission risk, mortality, or future comorbidities using Electronic Health Records (EHRs) [1–5]. However, mounting evidence suggests that machine-learning models can inadvertently perpetuate and exacerbate biases present in the data [6], including gender or racial biases [7, 8], and the marginalization of under-served populations [9]. These biases are typically encoded by capturing shortcuts—stereotypical or distorted features in the data [10–12]. For instance, numerous machine learning algorithms rely on post-treatment information [13–16], exemplified by a diagnostic model for skin cancer that depends on surgical marks [11]. A compelling illustration on Intensive Care Unit data is provided in Figure S1 5.***

***The Significance of Causal Reasoning in Data-Driven Decision-Making [17] While conventional machine learning relies on retrospective to generate predictions of future effects [18], truly insightful decision-making support necessitates a comparison of potential outcomes with and without the intervention. This involves estimating a causal effect, mirroring the methodology employed in RCTs [17]. However, RCTs encounter challenges such as selection biases [19, 20], difficulties in recruiting diverse populations, and limited sample sizes for exploring treatment heterogeneity across subgroups. Routinely collected data presents a unique opportunity to assess real-life benefit-risk trade-offs associated with a decision [21], with reduced sampling bias and sufficient data to capture heterogeneity [22]. Nevertheless, estimating causal effects from such data is challenging due to the confounding of the intervention by indication. Therefore, dedicated statistical techniques are imperative to emulate a "target trial" [23] from observational data.***

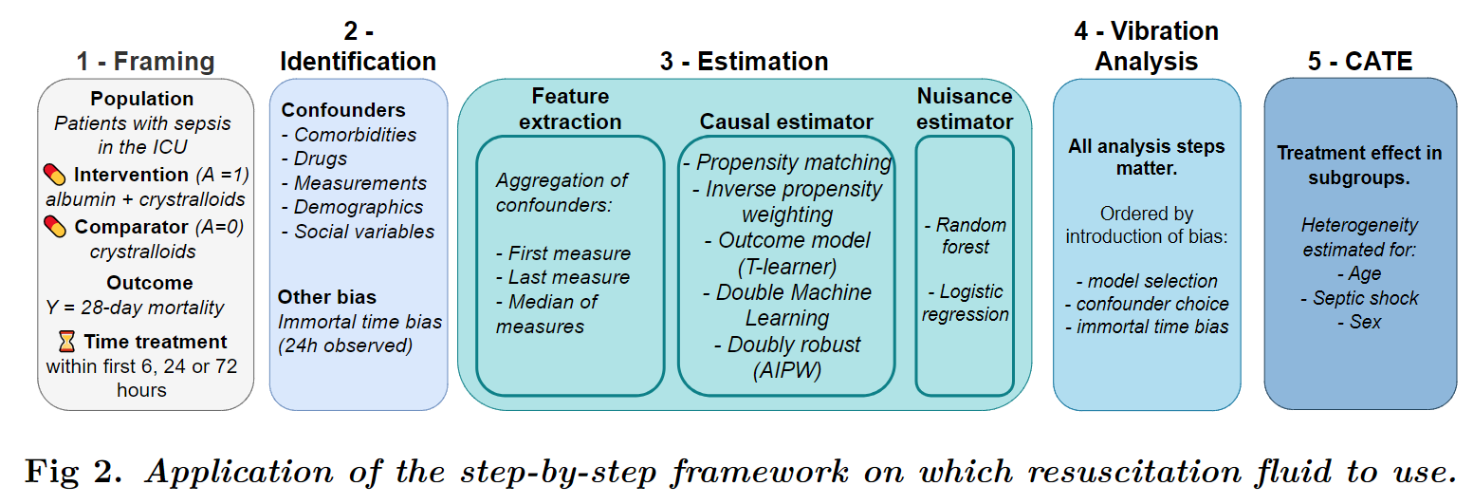
***Population effect vs. individual treatment effect: QRISK cardiovascular risk tool The QRISK is a risk score widely used in the UK to assess 10-year and lifetime cardio-vascular risk [24]. It is well calibrated, recommended by the national regulatory agency [25], and the National Institute for Health and Care Excellence (NICE) argues in favor of systematic screening of cardio-vascular disease (CVD) risks to increase the number of people receiving statins [25], especially if 10-year CVD risk exceeds 10% [26]. These suggestions found on a cost-effectiveness analysis selecting the 10% threshold to optimize quality-adjusted life years at the population level [27]. However, three trials in the NICE analysis weight for more than 70% of the effect (LIPID [28] (32% of the effect), PROSPER [29] (22% of the effect), ALLHAT-LLT [30] (23% of the effect)). All focus on high-risk patients and secondary prevention. All studies operate under the assumption of a consistent statin effect (risk ratios constant) across different CVD risk groups, irrespective of the baseline CVD risk and LDL-cholesterol levels. This aligns with the outcomes of two meta-analyses which observed broadly similar effectiveness across risk levels [31], but is challenged by another study in primary prevention [32]. Meanwhile the US Preventive Services Task Force concluded that the evidence is insufficient to determine the balance of benefits and harms of statin use for the primary prevention of CVD events and mortality in adults 76 years and older [33].***

***Critics on the efficiency of systematic screening and treatment heterogeneity Conversely, the World Health Organization (WHO) published a systematic review on screening for CVD risk in 2019 [34] asserting the limited effectiveness of such screenings, aligning with a Cochrane review that also reported poor effectiveness [35] resulting in over-treatment. The surge in statin prescriptions has been linked to a divergence in statin prescriptions at specified risk score levels [36]. This leads to over-prescription for low-risk patients and under-prescription for high-risk patients, casting doubt on the overall efficacy as the constancy assumption of the statin treatment effect might be violated. Accounting for sources of heterogeneity in patient profiles would be instrumental in optimizing treatment allocation.   
Moreover, the WHO report critiques suboptimal adoption leading to social biases in the screened populations. In the UK, there are instances of treatment heterogeneity at a given risk level [36] and documented evidence of social inequities in health checks [37]. Evidence suggests that predominantly socially advantaged patients undergo screening, consequently being offered statin treatment. This critique further emphasizes the presence of treatment heterogeneity.»***

3) Related works: The literature review should provide a comprehensive overview of the current state of the arts in the field, outlining the strengths and weaknesses of related research.  
**REPLY:** We agree that the current literature is fragmented between various fields of science. Thus, we rephrased the introduction to give a comprehensive overview.We outlined key and recent references from both scientific community that are involved (epidemiology and causal machine learning) for our problem.  
***«Integrating Multidisciplinary Perspectives for Informed Clinical Decision-Making; Existing literature has only partially delved into the challenges associated with estimating treatment effects using observational data. While epidemiologic studies underscore the importance of the target trial approach [38–42], there, emphasis primarily lies on biases stemming from confounding variables with relatively less attention to issues arising from estimator selection. Recent replications of RCTs using observational data did not explore the impact of modern machine learning methods on the robustness of the results [41, 43].***

***In contrast, machine learning and causal inference literature predominantly concentrates on estimators [44–48]: propensity score matching [49], inverse probability weighting [50], outcome models [51], doubly robust methods [45], or deep learning based models [52]. This literature is often inundated with intricate mathematical details and unverifiable assumptions. Guidelines seldom address time-related biases, or covariate aggregation which frequently emerge in datasets with temporal dependencies [53, 54]. Recently, the machine learning community shifted its focus from EHR data to simulated data, which may not capture the complexities of real-world data [55–58]. In this work, we amalgamate abstract epidemiological concepts with foundational principles from statistical literature, adopting an empirical approach tailored to the practical needs of applied researchers. By establishing a cohesive framework that encompasses the choice of study design, consideration of confounders, and selection of estimators (refer to Section Step-by-step framework for robust decision-making from EHR data), we highlight their equal importance in ensuring the validity of the analysis. To illustrate and compare biases, we investigate the impact of albumin on sepsis mortality using data from a publicly available intensive care database, MIMIC-IV [59] (section Application: evidence from MIMIC-IV on which resuscitation fluid to use). The primary focus of the main section is on accessibility, with technical details expanded in the appendices.»***  
  
4) Results: The experimental section needs clarification and more details, and the analysis and explanation of the test results need to be expanded.  
**REPLY:** Thank you for your helpful input. We agree that we were very brief on our results**.** To amend this deficit, we provided some clarification on the patient’s characteristics and added a supplementary figure that summarizes the different choices that we tested in the analysis.Analogous to our original figure 1, showcasing the use of our principles.

***In MIMIC-IV, these inclusion criteria yield 18,121 patients of which 3,559 were treated with a combination of crystalloids and albumin. While glycopeptide antibiotic therapy was similar between both groups (51.8% crystalloid vs 51.5 % crystalloids + albumin), aminoglycosides, carbapenems, and beta-lactams were more frequent in the crystalloid only group (2.0% vs. 0.7%, 4.3% vs. 2.6%, and 35.5% vs. 13.8%, respectively). The crystalloid only group was more frequently admitted as an emergency (57.3% vs. 30.7%). Vasopressors (80.2% vs 41.7%) and ventilation (96.8% vs 87.0%) were more prevalent in the treated populations, underlying the overall higher severity of patients receiving albumin (mean SOFA at admission 6.9 vs. 5.7). Table 2 details patient characteristics and S5 Fig details the selection flowchart.***



**Reviewer #2:** The Abstract and Author summary provided by the authors are the same. It would be helpful if the authors can create a distinct non-technical summary under the Author summary section. Other sections are concisely presented.  
**REPLY:** Thank you for your review and this suggestion. We rewrote the abstract and technical summary to match their targeted audience. See response to reviewer 1 for the full changes.