Comparative effectiveness research in critically ill patients: risks associated with mischaracterising usual care

Willard N Applefeld,* Jeffrey Wang,* Harvey G Klein, Robert L Danner, Peter Q Eichacker and Charles Natanson

One of us (CN) recently co-authored a critique of design errors in three purportedly usual care trials enrolling critically ill subjects that was published in a United States-based ethics journal. These missteps incorrectly represented unusual care as usual, thereby jeopardising subject safety, confounding trial conclusions, and undermining the consent process. This prior examination was focused on ethical issues arising from these errors and was intended for an audience of medical ethicists. Here, our emphasis is on medical risks for patients emanating from these trial design errors. Previously published data are included to demonstrate how those risks may have affected study results and patient outcomes.

The three case studies presented here challenge the widespread belief that randomised trials of putative usual care, particularly those enrolling critically ill subjects, are invariably safe. Properly designed head-to-head comparisons of contemporary care can improve clinical decision making by better quantifying relative risks and benefits. However, for such research to be informative, at least one arm must be truly representative of current medical practice. Some trials purporting to compare usual care practices may not accurately reflect those practices and instead provide unusual care, either to all enrolled subjects or to important subgroups within a cohort. If subjects are critically ill with a high baseline risk of death, this unusual care may have grave consequences that include but are not limited to the following:

- **Inadequate informed consent**: for research to be ethical and consent to be informed, subjects must understand how their care will change and the potential risks of study enrolment.
- **Suboptimal care**: unusual therapies that are inferior to usual care can put trial subjects at increased risk, particularly when administered to critically ill subjects.
- **Inability to monitor safety**: when a trial compares two unusual therapies, the benchmark for safety monitoring is lost. Without a usual care comparator, harm from either or both unusual care arms may go unrecognised.

**ABSTRACT**

Comparative effectiveness research can help guide the use of common, routine medical practices. However, to be safe and informative, such trials must include at least one treatment arm that accurately portrays current practices. While comparative effectiveness research is widely perceived as safe and to involve no or only minimal risks, these assumptions may not hold true if unrecognised deviations from usual care exist in one or more study arms. For critically ill subjects in particular, such practice deviations may increase the risk of death or injury and undermine safety monitoring. Furthermore, unrecognised unusual care seems likely to corrupt informed consent documents, with underappreciated risks shrouded under the reassuring “comparative effectiveness” research label. At present, oversight measures are inadequate to ensure that research subjects enrolled in comparative effectiveness trials are actually receiving usual and not unusual care. Oversight by governmental and non-governmental entities with appropriate expertise, empowered to ensure that current clinical practice has been properly represented, could help prevent occurrences in clinical trials of unusual care masquerading as usual care.

**Inaccurate and harmful conclusions**: without a usual care comparator arm, if both therapies studied are inferior to usual care, there is a risk that the least harmful intervention will displace usual care in practice guidelines, putting future patients at risk.

In the three examples described below, unusual care misrepresented as usual care resulted in one or more of these unintended harmful consequences. Not accurately capturing current practice is a design error that has been difficult
One study, the AVIOx trial,From the SUPPORT and compared two narrow range targeted in SUPPORT was not usual care was

**Box 1. Reasons for misconceptions of usual care**

- Investigators may make errors in collecting or analysing data or use outdated or inaccurate sources of evidence to document usual care practices.
- Important subgroups within the target population of a study may be misaligned with one or more of the therapies or interventions assigned by randomisation.
- A widely used and accepted clinical practice may be difficult to define.
- Usual care practices may be evolving rapidly and change before approval, implementation or completion of the clinical trial.
- Instead of using accurate and pertinent information to document current practice, investigators rely on expert opinion or guidelines that are poorly supported.
- During a clinical trial, physicians may begin to adopt practices outlined in the treatment arm (ie, the Hawthorne effect) and thereby knowingly or unknowingly alter their perceptions of usual care.

... to recognise; the three protocols presented underwent extensive review processes, conducted by multiple groups at many institutions, before approval and implementation. Failure of the existing system to correct misconceptions of usual care can occur for many reasons (Box 1). However, as we describe at the conclusion of this review, awareness, regulatory guidance, and the development of mitigating procedures can help ensure that future trials intended to study usual care avoid these pitfalls.

**Three case studies**

**Misapplication of a guideline as representative of usual care: the Surfactant, Positive Pressure, and Pulse Oximetry Randomized Trial (SUPPORT)**

Premature, high risk infants have underdeveloped lungs and often require supplemental oxygen therapy at the time of delivery, and sometimes for weeks to months afterwards. However, both the upper and lower ends of the oxygen concentration range used in these infants can have different but potentially serious adverse effects. With too much oxygen, these infants can develop blindness or retinopathy of prematurity; with too little oxygen, they can have neurological injury or death. SUPPORT randomised 1316 premature, low birth weight, high risk infants to a high (91–95%) or low (85–89%) target range of oxygen saturation measured by pulse oximetry (SpO₂) in an attempt to find an SpO₂ target range that minimises these competing risks.² From the SUPPORT study protocol, it does not appear that SpO₂ target ranges in common use at the time of the study were adequately investigated and used to validate the trial design.³ Instead, investigators relied on the American Academy of Pediatrics’ (AAP) guidelines⁴ and compared two narrow SpO₂ target ranges lying at the extreme ends of the much broader range that was actually recommended. To justify

... the lower target range included in SUPPORT, data were cited from a decade-old, retrospective chart review of similar neonates done in northern England by Tin and colleagues.⁵ The methodology employed for the design of study groups in SUPPORT failed for two main reasons. First, the two SpO₂ target ranges at either end of the AAP guideline systematically underrepresented the middle of the recommended range, which better reflected most, if not all, of usual care at the time. Second, the Tin and colleagues’ study⁶ was an inadequate source to justify using the very low SpO₂ range in SUPPORT. It presented only observational outcome data, collected a decade before SUPPORT began enrolling subjects. The study had a mortality rate close to double that seen at the time of SUPPORT, indicating that overall practices had changed substantially by the time SUPPORT was undertaken. Therefore, the Tin et al study was not relevant to care when SUPPORT was designed and conducted (Online Appendix, e-Supplementary Box 1).

To determine usual care at the time of SUPPORT, we performed a comprehensive review of all contemporaneous, medical literature that was available before the trial began describing SpO₂ targets.⁵ One study, the AVIOx trial, specifically examined usual care at the time SUPPORT was designed. Including 84 infants born at less than 28 weeks’ gestation who required oxygen therapy, extensive information was collected on SpO₂ levels over an extended period at 14 neonatal intensive care units (NICUs) in the US, the United Kingdom, Australia and New Zealand.⁷ Importantly, subjects in the AVIOx study met enrolment criteria for SUPPORT. Our review of the AVIOx study found that maintaining infants in the low SpO₂ range targeted in SUPPORT, encompassing the bottom half of the AAP guideline, was not consistent with practices at any of the 14 NICUs included in the study (ie, the low oxygen saturation arm was unusual care in those NICUs) (Figure 1).²,⁶ Only the high SpO₂ target range of SUPPORT was consistent with usual care in AVIOx NICUs. This finding that the low SpO₂ range targeted in SUPPORT was not usual care was confirmed in our systematic review of more than 100 other distinct NICUs reporting such data.⁶

Notably, an observational study of usual care could have been performed before or during the design of SUPPORT or the same information could have been obtained by
Figure 1. Usual care median achieved oxygen saturation measured by pulse oximetry ($SpO_2$) values in 14 care centres for preterm infants receiving oxygen therapy compared with low and high $SpO_2$ arms.

Panel A compares the median achieved $SpO_2$ values and interquartile range of each of these 14 usual care neonatal intensive care units (NICUs), with the intended $SpO_2$ range established in the same NICUs (represented by dark grey vertical bars). Panel B compares median achieved $SpO_2$ values and interquartile range from these 14 NICUs to the target ranges of the low (lower grey-shaded area) and high (upper grey-shaded area) $SpO_2$ arms of the SUPPORT, BOOST II and COT trials. Figure reproduced, with permission, from Cortés-Puch et al. 6

Panel A compares the median achieved $SpO_2$ values and interquartile range of each of these 14 usual care neonatal intensive care units (NICUs), with the intended $SpO_2$ range established in the same NICUs (represented by dark grey vertical bars). Panel B compares median achieved $SpO_2$ values and interquartile range from these 14 NICUs to the target ranges of the low (lower grey-shaded area) and high (upper grey-shaded area) $SpO_2$ arms of the SUPPORT, BOOST II and COT trials. Figure reproduced, with permission, from Cortés-Puch et al. 6
Failure to accurately define usual care in the types of infants studied in SUPPORT exposed trial subjects to abnormally low Sp\(_2\) ranges and increased risks.\(^6\) The result was inaccurate and misleading informed consent documents\(^2,6,8\) and a treatment regimen that inflicted unnecessary harm.

Unfortunately, two subsequent trials, started after SUPPORT, began enrolling subjects using the same methodology and examining similar extreme ranges of oxygen treatment in high risk, premature infants: one predominantly in Canada (Canadian Oxygen Trial [COT]),\(^9\) and the other in the UK, Australia and New Zealand, (Benefits of Oxygen Saturation Targeting [BOOST II]).\(^10\)

Employing outdated rather than current practices as usual care: Acute Respiratory Distress Syndrome Network Lower Tidal Volume (ARMA) trial

The early stages of acute respiratory distress syndrome (ARDS) are characterised by diffuse alveolar damage for which patients frequently require mechanical ventilatory support. Starting in the late 1980s, concerns grew that excessive tidal volumes and airway pressures associated with mechanical ventilation would aggravate alveolar injury with ARDS.\(^11\) In the latter half of 1990s, the ARMA trial randomised 861 critically ill mechanically ventilated patients with ARDS — independent of any acute need for a change in their ventilator-delivered tidal volume — to receive a fixed large 12 mL/kg predicted body weight (PBW) tidal volume or a fixed small 6 mL/kg PBW tidal volume to compare the impact of these settings on lung damage and mortality.\(^12\) Investigators described the large tidal volume at the time of the study as “traditional” and used it as the usual care control for assessing safety and benefit. “Traditional control” has no accepted definition and may have been confusing to research subjects and their surrogates. On its face, “traditional” suggests care more consistent with treatment in the past. While the ARMA protocol and subsequent publications referred to the large tidal volume arm, somewhat accurately, as “traditional”, consent documents also described this “traditional volume” as being currently or commonly used by clinicians during usual care at the time of the trial. For example, the following statements appear in the ARMA consent forms:

“One [purpose of the study] is to compare two ways of inflating your lungs while on the machine. Doctors currently use both methods to breathe for patients, but it is not known if one method is better [than the other].”\(^13\)

“Presently doctors use different size breaths of oxygen-rich air to inflate the patient’s lungs. It is unknown whether it is better to use large (12 mL/kg) or small (6 mL/kg) [tidal volumes]. Both ways of inflating the lungs are acceptable methods that are commonly used to treat patients with [acute lung injury] and [ARDS].”\(^14\)
“Presently doctors use varying volumes of oxygen-enriched air to inflate the lungs. It is unknown whether it is better to use a large or small volume of oxygen-enriched air to inflate [the] lungs of patients with lung injury. Both ways of inflating a patient’s lungs are considered acceptable methods and are commonly used [in] medical practice.”

An analysis of tidal volumes administered to patients before their enrolment and randomisation in the ARMA trial showed that usual care was not commonly or predominantly 6 mL/kg PBW or 12 mL/kg PBW but actually varied over a wide, normally distributed range related to the level of lung injury and underlying lung compliance (Figure 2 and Figure 3). During usual care pre-enrolment, patients with less compliant lungs received lower tidal volumes and those
with more compliant lungs received higher tidal volumes. This was consistent with a survey published shortly before the ARMA trial, which reported that clinicians were more concerned with excessive airway pressures during mechanical ventilation rather than specific tidal volume levels. To avoid high pressures in less compliant lungs, one common method used by clinicians was to reduce tidal volumes. The “traditional” 12 mL/kg PBW tidal volume chosen for the trial was being used routinely in only about 10–15% of patients with ARDS and the 6 mL/kg in less than 5% of patients with ARDS at the enrolling institutions (Online Appendix, e-Supplementary Box 2).

For about 80% of ARMA subjects assigned to the “traditional” 12 mL/kg PBW arm, ventilator tidal volumes had to be increased after randomisation to reach this level. Accordingly, airway pressures also increased in a substantial number of these participants beyond levels believed safe. Contemporaneous patients with ARDS who met the eligibility criteria but were not enrolled for various reasons, and received usual care in the same intensive care units, had a significantly lower mortality rate than participants in the “traditional” arm (Figure 4). As pre-randomisation ARMA data showed, usual care for these unenrolled patients with ARDS entailed adjusting tidal volumes to keep airway pressures on average at levels considered safe. The absence of a truly representative usual care arm receiving titrated care deprived the data and safety monitoring board of an accurate benchmark against which to compare mortality and serious adverse events during the trial. Without such usual care data, the data and safety monitoring board lacked a clear basis to stop the trial early, when it would have otherwise become evident during interim analyses that the large “traditional” tidal volumes were harmful (Figure 4 and Online Appendix, e-Supplementary Box 2).

Ignoring usual care, resulting in harm to subgroups receiving care opposite to usual care treatment: Transfusion Requirements in Critical Care (TRICC) trial

The TRICC trial randomised 838 critically ill but stable surgical patients who were not actively bleeding to either a “restrictive” or a “liberal” red blood cell (RBC) transfusion strategy. The restrictive strategy withheld RBCs transfusions until haemoglobin concentrations decreased to low levels (transfusion trigger, 7.0 g/dL; maintenance, 7–9 g/dL). The liberal strategy triggered transfusion when patients still had relatively high haemoglobin levels (transfusion trigger, 10.0 g/dL; maintenance, 10–12 g/dL). The TRICC trial reported that overall hospital mortality was significantly higher for patients assigned to the high RBC transfusion trigger arm compared with the low one. Based on this finding, the study concluded that critically ill surgical patients should receive RBC transfusions only if haemoglobin levels fall to low concentrations similar to the trial’s restrictive arm.

Randomising subjects to two fixed haemoglobin thresholds meant that transfusion strategies were no longer based on patient individual needs, comorbidities or underlying physiology. Instead, patients were randomly assigned to transfusion regimens chosen primarily to conserve blood supplies. Before beginning the trial, investigators surveyed 193 physicians to ascertain existing practices (Online Appendix, e-Supplementary Box 3). This survey found that physicians used a wide range of haemoglobin levels to trigger transfusion but these triggers were not random. Physicians transfused RBCs to achieve higher haemoglobin levels in older patients and in those with particular comorbidities. Only 3% of physicians reported that they would use a haemoglobin trigger as low as 7 g/dL for patients with known cardiovascular disease and only 12% reported they would use one as high 10.0 g/dL for healthy young patients. This titration of care was consistent
Box 2. Recommendations to safeguard participants and ensure the scientific integrity of usual care trials

- Collect sufficient data or conduct research to accurately describe and fully understand usual care in the population of patients to be studied before designing a trial and writing the protocol.
- Ideally, institutions participating in the study should be included in the characterisation of current medical practices for the condition under investigation. To be considered “usual”, this care may need to account for differences in practice within and across enrolling institutions as well as regionally, nationally and possibly internationally. Where usual care varies nationally or internationally, trial designs might include, if appropriate, components that accurately account for this diversity in usual care practices.
- All information used to characterise usual care should be documented in the research protocol or supplemental material submitted for the trial approval process, detailing all sources of data contributing to the assessment (surveys, prospective observational studies, chart reviews, or literature searches with systematic reviews).
- Funding agencies should require documentation of current clinical practices in grant applications as a condition of support. Government granting agencies, such as the National Institutes of Health in the United States, should ensure that their internal review processes adhere to this requirement.
- Government guidance and regulations as well as protocol approval processes should be strengthened and further developed to ensure that claims about usual care practices under study are well characterised and accurate.
- Peer reviewers and journal editors should require the submission of supplemental material that supports the way usual care is characterised in clinical trials and assess whether control arms were designed in a manner that protects patients and minimises the potential for harm.

with consensus guidelines on transfusion practices at the time of the trial, which recommended individualising the administration of RBC transfusions.23

The TRICC trial did not include a usual care group with transfusion triggers based on an individualised assessment of patient needs. The design of the trial heeded neither the results of the investigator survey nor the existing consensus guidelines on transfusion practices, instead making both arms unusual care for important patient subgroups by setting arbitrarily fixed triggers.22,23 This approach simplified the design but did not minimise risks.

We performed an analysis using data from the original TRICC trial publication combined with subsequently published trial data and demonstrated that subjects with pre-existing severe cardiovascular disease had a significantly different response to the two fixed transfusion thresholds compared with patients without that condition.17,24,25 A more recent meta-analysis that included 15 subsequent studies randomly allocating patients to receive RBC transfusions triggered by fixed low or high haemoglobin levels and enrolling thousands of additional patients found the same result (Figure 5).25 For a restrictive compared with liberal strategy in both analyses, subgroups with severe cardiovascular disease had increased mortality—an outcome opposite to those patients without this comorbidity. Subgroup analysis of the original TRICC trial by its investigators also indicated that younger, healthier patients with low severity of illness scores primarily drove the significant mortality increase in the liberal arm.21

Randomisation to high and low haemoglobin thresholds in the TRICC trial ensured that one subgroup (younger, stable patients) in the high trigger arm would receive transfusions when not necessarily clinically indicated.17,21,24 In contrast, another subgroup (older patients with cardiovascular disease) in the low trigger arm would receive fewer transfusions. For these different specific subgroups nested within the overall study population, one or the other of the two transfusion approaches represented unusual care. Excess mortality within the two subgroups—young, stable patients and older patients with cardiovascular disease—occurred in opposite arms of the study, rendering the overall comparison misleading. The young healthier patients in the liberal arm were exposed to unnecessary transfusions and volume challenges, while older patients with cardiovascular disease in the restrictive arm had an increased number of cardiac ischaemic events. A usual, individualised care arm might have had lower mortality than either arm. However, lack of an arm representing current medical practice, as identified in the original survey of clinicians, deprived the TRICC trial of a supported scientific basis for recommending a fixed transfusion threshold for most patients.

Remedies to improve usual care research

The three case studies above illustrate the pitfalls of designing usual care trials without first carefully documenting and understanding contemporaneous clinical practices at participating institutions and elsewhere. As shown for the three examples reviewed here, misunderstanding or oversimplifying usual care can render results uninformative, lead to inaccurate conclusions, obscure research risks, and fail to adequately protect research participants. Unfortunately, the methods employed by SUPPORT, ARMA and TRICC have been used in a large number of subsequent trials that similarly lacked treatment arms closely reflecting
usual care practices. With the growth of comparative effectiveness research and its role in seeking to improve outcomes and cost-effectiveness, it has become essential to carefully and rigorously document contemporaneous practices and incorporate that knowledge into trial designs and the processes for approving human subjects research. As previously described, adherence to the six recommendations listed in Box 2 may help safeguard participants by improving both the scientific and ethical integrity of comparative effectiveness research.

Conclusion
Randomised clinical trials of purported usual care currently have inadequate regulatory and institutional safeguards to guarantee that protocol-determined care has not become in fact unusual. Recognising and understanding risks associated with an inaccurate or incomplete representation of usual care is a critical first step toward preventing such errors in future trials. This awareness is necessary for developing regulations, guidance, and protocol approval processes that increase the rigour with which investigators, sponsors and funding agencies define usual care. Ultimately, comparative effectiveness research can only improve both resource utilisation and patient outcomes if trials have at least one study arm that is truly representative of medical practice at the time of the trial. Addressing pitfalls stemming from the incomplete, oversimplified or imperfect implementation of usual care promises to make such trials safer, more informative and better able to improve patient outcomes. Adherence to scientific rigour and accuracy in defining and delivering actual usual care is an essential cornerstone for such research to minimise risks to human subjects and realise its full potential.

Acknowledgements: We thank Michael Carome, for reviewing versions of this article; Ruth Macklin, for her work on the unrevised version of this manuscript; and Juli Maltagliati and Kelly Byrne, for help editing and submitting the manuscript. This research was supported by the Intramural Research Program of the National Institutes of Health Clinical Center. The opinions expressed in this article are the authors’ own and do not represent any position or policy of the National Institutes of Health, the Department of Health and Human Services, or the United States Government. The corresponding author confirms that they had access to all the data in the study and had final responsibility for the decision to submit for publication.

Competing interests
None declared.

Author details
Willard N Applefeld.1
Jeffrey Wang.1
Harvey G Klein2
Robert L Danner1
Peter Q Eichacker1
Charles Natanson1

1 Critical Care Medicine Department; Clinical Center, National Institutes of Health, Bethesda, MD, USA.
2 Department of Transfusion Medicine, National Institutes of Health, Bethesda, MD, USA.

* Equal first authors.

Correspondence: cnatanson@nih.gov

References


