USE OF THE PROPENSITY SCORE METHOD FOR RECRUITMENT BIAS REDUCTION IN OBSERVATIONAL STUDIES

Lionel Riou França[1], Stéphanie Payet[1], Katel Le Lay[1], Robert Launois[1]

Correspondence Author
Lionel Riou França
REES France – 28, rue d'Assas – 75006 Paris – France
Tel: +33 1 44 39 16 90
Fax: +33 1 44 3916 92

Source of Financial Support: Health Ministry – DHOS – France

ABSTRACT

PURPOSE To estimate the impact of drotrecogin alfa on intensive care workload in an observational study using propensity score matching to control for recruitment bias.

METHODS PREMISS is a prospective and multicenter pre-post study. Its goal was to compare severe septic patients with multiple organ failure treated either by standard care alone or with drotrecogin alpha (DA). Inclusion of patients took place before and after the drug's market authorization. Recruitment bias was suspected, as treatment assignment was not made at random. Two techniques were considered to compare workload in the two treatment groups. The first fits a multiple regression model on the whole population, adjusting for unbalanced covariates. The second strategy consists on matching patients from both phases with similar characteristics using the propensity score (PS) methodology. The treatment effect was subsequently estimated on the matched sample. The clustering of the patients within the intensive care units was taken into account using random effect modelling.

RESULTS All methods lead to the conclusion that DA increases intensive care workload. The importance of this impact, however, varies from a 28% increase in the crude unadjusted analysis to a 14% increase in the PS-matched adjusted analysis. The PS estimates appear to be more conservative than those due to simple multivariate adjustments. Further exploration indicates that DA can have an impact on workload directly or indirectly, through its effects on adverse events, on the length of stay and on mortality.

CONCLUSION We found similar results with both standard method of adjustment and PS matching. Compared to the whole population analysis, matching increases robustness of results and avoids making too many adjustments, allowing to focus on the treatment effect.

Key words: drotrecogin alfa – propensity score – gamma regression – random effects – intensive care – workload

Topic: Health Technology Assessment

INTRODUCTION

There as been considerable debate on the role of observational studies for the evaluation of an intervention's effect. Many researchers claim that non-randomised studies lead to unreliable results and appeal for the exclusive use of randomised clinical trials (RCTs).[1] The latter are considered as the "gold standard" since they imply the equality of the distribution of the variables measurable at the time of randomisation.[2] On the contrary, non-randomised studies cannot guaranty that the populations being compared share the same distribution of prognostic factors. When the populations differ in some baseline characteristic predictive of the outcome of interest, the estimation of the intervention effect can be biased. We will use the term of recruitment bias to refer to these situations. Despite the risk of producing biased treatment effect estimates, some authors advocate the use of non-randomized studies on the basis that, when correctly conducted, they can lead to results similar to those reported in RCTs.[3] There are indeed some arguments in favour of non-randomised studies. First of all, even if their results are prone to more scepticism than those arising from RCTs, there are some situations where randomisation is infeasible (which is sometimes the case in the field of surgery or living habits), not ethical (for example, if the efficacy of the intervention is already acknowledged) or simply too costly. In a recent review of the question, the NHS concluded that non-randomised studies should only be used in these cases.[4] There are however more possible advantages to non-randomisation. First, some sources of already available observational data can provide valuable information. Moreover, RCTs tend to be conducted under strict protocol-driven conditions, different from what will be the use of the intervention in real practice. Some authors assert that randomised studies do not provide many information relevant to decision makers.[5] Thus, while most clinical researchers see observational studies as exploratory tools whose results need to be confirmed by RCTs, non-randomised studies can also be carried after an RCT in order to assess the external validity of the findings.

The PREMISS study is an example of such an application. After a successful RCT concluded that drotrecogin alfa (DA) reduced the mortality of severe septic patients,[6] the French ministry of health funded in 2002 a study aimed at the estimation of DA's impact in intensive care practice. This study was the result of cooperation between the two professional associations of intensive care practitioners in France: the SFAR (French Society of Anesthesia and Intensive Care) and the SRLF (French Speaking Reanimation Society).

The main objective of the PREMISS study was to estimate the observational costs and efficacy of DA's introduction in French intensive care units (ICUs). Since DA's efficacy had already been explored in a large multicenter RCT, it would have been unethical to randomise patients in a control group. Furthermore, some other features of a RCT, like the blinded allocation of treatments or the highly protocoled patterns of care, would have been incompatible with the study observational objectives. For instance, since DA is known to increase the probability of hemorrhagic events, caregivers aware of the treatment received by their patient may monitor more severely those being administered this product.The PREMISS design was therefore conducted without randomisation.

The propensity score[7] (PS) is an increasingly used method to deal with recruitment bias. It will be illustrated in the estimation of DA's impact on intensive care workload.

METHODS

The PREMISS study

The intervention to be evaluated is the administration of DA at the recommended dosage of 24 µg/kg/h for 96 hours (but the physicians were free to use different protocols). The inclusion criteria are those defined in the European indications for DA: adult patients (in practice, 13 patients less than 18 years old at their admission were included and kept for analysis. All were more than 15 year old) with severe sepsis with multiple organ failure (MOF) and without contraindications (active internal bleeding, intracranial pathology, neoplasm or evidence of cerebral herniation, concurrent heparin therapy ≥15 International Units/kg/hr, known bleeding diathesis except for acute coagulopathy related to sepsis, chronic severe hepatic disease, platelet count < 30,000 x 10^6/l, even if the platelet count is increased after transfusions, patients at increased risk for bleeding). The PREMISS study follows a quasi-experimental pre/post design. In a “before” phase, from September 2002 to January 2003, prior to DA's marketing autorisation, patients were included in a control group. The “after” phase was carried from January 2003 to November 2004, once DA was available, with patients assigned to the treatment. Analyses were conducted in an intention to treat fashion: all the patients in the “after” phase were considered as treated with DA, even if they did not fulfil the recommended indication. Patients were followed up until their discharge from the hospital.

Data collection

Use of an Internet database

Information was collected in a decentralised fashion using an online case report form (CRF). The information was then centralised in a protected server in a single database. The choice of an Internet questionnaire allowed for the inclusion and non-inclusion criteria to be automatically validated. The data is checked as it is entered, which substantially reduces the number of errors and queries. Quality controls of the information submitted were also automated, making easier the data monitoring process. Confidentiality of the data was ensured by the use of unique password-protected identifiers for each participating ward and by the anonymisation of patients by an alphanumeric code. The CRF, its interface and all associated software was conceived, programmed and maintained by the authors in cooperation with the intensive care practitioners' societies. We consequently were in possession of a tool entirely customisable to the research needs.

The data collected included information about: the ICU itself; the patients'characteristics at the time of their admission on the ward; their severity at the time of inclusion; the administration of antibiotics, corticoids, DA (in the "after" phase) and other drugs; hemorrhagic and transfusion events; the procedures of care given during their stay on the ICU; their survival status at discharge from the ICU, from the hospital and 28 days after sepsis initiation.

Initial characteristics

As treatment allocation was not randomised, we needed to take multiple sources of bias into account. We tried to gather as many data as possible on the patients' initial characteristics. These included demographic variables (age, sex), medical variables (admission category, comorbidities, presence of septic shock…), biological variables (arterial pressure, urine output, platelet count…), microbiological variables (site of infection, type of infection…) and severity indexes (SAPS[8], LODS[9]). On the whole, 46 initial characteristics where measured and considered in the analysis.

Workload measures

The French medical information system was about to be reformed when the study protocol was to be submitted. The new classification of medical procedures was adopted conjointly with the old one in the case report form (CRF). The information system formerly in use in

French ICUs was the Omega score, a tool used to measure intensity of care and shown to have good correlation with workload and costs.[10] This score measures 47 medical procedures performed during the ICU stay.

The new classification comprises more than 7 000 technical acts. Each is arranged hierarchically according to two resources necessary for its realisation: the medical workload (the act complexity) and the financial expenses (the cost of the act). A relative cost index (RCI) is associated with each act as the measure of the intensity of resources (economic or not) it requires. For instance, act ZZQP003 corresponds to the medicalised monitoring of the intra-hospital transfer of a ventilated patient and is associated with a RCI of 60 per transfer. Act EQQP004 corresponds to an hemodynamic supply by cardiopulmonary bypass, its RCI is of 240 per 24 hours.

The use of this classification in intensive care was troublesome: since severe sepsis is only a syndrome, the range of all potential interventions occurring during a patient's stay is quite important. Only the most important acts were included in the PREMISS CRF: essentially, those relative to organ support or monitoring (respiratory, cardiovascular, digestive, renal, haematological, nervous system), those relative to medical imagery (echographia, tomography, magnetic resonance image, endoscopy) and, of course, those related with haemorrhage management. 115 different technical acts were included in the CRF. The ICU workload is estimated as the sum of the number of technical acts weighted by their RCI. One advantage of this approach is that the conversion of this workload to monetary costs is, in theory, possible by multiplying by a conversion factor. However, at the time of the analysis, the regulating authorities have not determined its value.

Statistical analysis

Model specification

In econometrics, it is well known that resource use variables can be particularly skewed. Even if the normality of the explained variable is not a condition of validity of the linear regression model, it can lead to the violation of the hypothesis of normally distributed residuals. There as been large debate over the most appropriate way to deal with such concerns.[11],[12] For simplicity's sake, we will stick to the classical linear model to estimate ICU workload unless its validity conditions are clearly violated. We will also fit a generalised linear gamma model with a log link to confirm the results. If we note y_i the dependent variable for patient i and x_i the vector of his observed covariates, the model is given by:

Log(E[$y_i|x_i$]) = x_i'β, and therefore: E[$y_i|x_i$] = exp(x_i'β).

The gamma distribution further implies that the variance of the variable is proportional to the square of the mean: V[$y_i|x_i$] = Φ (E[$y_i|x_i$])^2, where *Φ* is the dispersion parameter. The advantages of the gamma model with log link are that it provides a multiplicative model (providing more adequate information in the particular case of workload, which has no absolute meaning) and, above all, that it avoids bias in the interpretation of results: had we chosen to estimate a linear model on the logarithm of workload (a frequent procedure in health economics analyses), we would have been confronted to interpretation issues, since the log of the mean of workload (which is estimated in the gamma model) is not equal to the mean of the log of workload (being modelled in the log transformation).

Adjustments for recruitment bias

Identification of unbalanced initial characteristics

In order to explore the comparability of the "before" and "after" groups, we computed standardised differences. They are frequently used in propensity score methodology as a quantitative measure of bias.[13] A standardised difference is related to the degree of unbalance in a variable means accounting for its degree of variation. The standardised difference of a variable *i* is defined as:

$d_i = 100*(x_{ci} – x_{ti})/\sqrt{}\{s^2_{ci} + s^2_{ti})/2\}$

Where x_{ci} and x_{ti} are the control and treatment sample means of the i^{th} variable and s^2_{ci} and s^2_{ti} are the corresponding sample variances. For binary variables, the sample means are the sample proportions. For polytomous variables, we compute a standardised difference contrasting each category with all the others and we retain the highest one. We will consider a variable as balanced between the groups when its standardised difference absolute value is inferior to 10 %.

Adjusting for recruitment bias using multivariate regression

Multivariate regression models are the most frequently used methods to assess an intervention effect on a quantitative outcome variable. The model will include an indicator of the PREMISS study phase ("before" or "after") as a covariate as well as the variables for which we wish to adjust for. Some problems arise in this model. First of all, any parametric model is more or less robust to violations of its validity assumptions. A misspecified model can lead to erroneous conclusions. Furthermore, the model results will depend on the way the relationship between the variable and the outcome is specified. For example, the association between age and workload can be linear or quadratic. Including age as a quantitative variable or as a qualitative one (e.g. by quartiles) might influence the results. Finally, when controlling for too many covariates, more problems are encountered. The sample size may be too small to accurately estimate all model parameters, and multicolinearity issues can arise.

Adjusting for recruitment bias using propensity score matching methods

An alternative to multiple covariate adjustments is to select a sample of patients comparable in each treatment group. This sample can be obtained with a propensity score approach. The propensity score is defined as the conditional probability of belonging to the "after" group given the initial covariates measured.[14] This probability replaces a large number of covariates by a single scalar. The PS can then be used as a stratification variable, as a matching variable or as an adjustment variable in a multivariate regression (or any combination of the three). Compared to the regression adjustment approach, the PS has the advantage of being less sensitive to assumptions about the functional form of the association of a covariate with the outcome (e.g. linear, quadratic…).[15]

Since there were missing values among the 46 initial covariates, multiple imputation procedures were used.[16],[17] We generated 10 imputed datasets, estimated the PS for each and computed each patient PS as the mean of the 10 imputation-based PS estimates.

The PS was estimated using a logistic regression model. We tested three different fitting strategies. The first model (M1) included all 46 covariates. The second model (M2) included all covariates that remained statistically significant in the logistic model at the 10 % level as well as all the covariates showing signs of unbalance (i.e. with a standardised difference between the treatment groups higher than 10 % in absolute value). The third model (M3) added interaction terms to model (M2).

Once the PS were estimated for each patient in all three models, "after" patients were matched to "before" ones on the basis of their PS using an optimal matching algorithm (the optimality criteria was to minimize the distance between the matched groups. For each pair of matched patients, the absolute value of the difference between their PSs was used as a distance measure). We then selected one of the three matched samples on the basis of the total number of patients matched (the more the better) and the degree of residual unbalance in the sample (in essence, we considered the number of covariates still showing an absolute value of the standardised difference superior to 10 %).

The remaining analysis were performed on the matched sample adjusting only for the covariates still showing signs of unbalance between the two treatment groups.

Accounting for the clustering of patients within the participating ICU's

PREMISS was a multicenter study. One of the other conditions of validity of the linear regression model is that the residuals are independent and identically distributed (iid). Since some patients share the same ICU, we can expect them to be submitted to similar unobserved treatment procedures. Their outcomes are likely to be correlated. In consequence, so will br the residuals of patients from the same ICU and the iid assumption will not be met. We will use random effects models to take into account the clustering of patients among ICUs. In the traditional multiple regression linear model, the outcome y_{ij} observed for patient *i* from ICU *j* can be written as:

$y_{ij} = \Sigma_k x_{ijk} \beta_k + e_{ij}$

where x_{ijk} represents the k^{th} variable observed for patient *i* from ICU *j*, β_k is the estimated effect of the k^{th} variable on the outcome, and e_{ij} is the error assumed to be iid and with null expectation. This model is called the fixed effects model.

The simplest form of a random effect model is the random intercept model, which adds a residual cluster effect:

$y_{ij} = \Sigma_k x_{ijk} \beta_k + u_j + e_{ij}$

The cluster effect u_j is also assumed to have a null expectation. More complicated models can be specified when adding random components to the model coefficients.

Random effect models (also called hierarchical or multilevel models) have the advantage of providing a correct modelling of the variation: when fitting fixed effects models to hierarchically structured data, the estimated standard errors associated with the model coefficients will be too small, leading to the overestimation of the significance of the estimates.[18]

RESULTS

Baseline characteristics

Once the database was checked for quality, 1 096 patients were retained in the analysis: 509 in the "before" group and 587 in the "after" one. These patients are nested within 85 different participating ICU's and 64 different hospitals. 49 hospitals are teaching hospitals. The repartition of medical, chirurgical and general ICUs is broadly balanced. The mean age of the patients was of 60.8 years (range: 15.6 to 94.8). 62% of the patients were of male sex. 40% of the patients where admitted in the ICU with an internal transfer, 28% entered the ICU in emergency, 23% with an external transfer. The remaining entered the ICU directly. 72% of the patients are medical ones. 22% of the patients presented a McCabe & Jackson severity of illness score of 2 (ultimately fatal disease), 6% a score of 3 (fatal disease). The mean SAPS II severity score at the time of admission was of 56.6 (range: 7 to 131). Due to the specificity of the inclusion criteria (presence of multiple organ failure), comparisons with other epidemiological studies[19,20] or RCTs,[6] typically focussing on severe sepsis with or without MOF, are difficult. The population recruited in the PREMISS study, compared to other French populations, appears to be comparable in terms of admission categories, age or sex. The severity scores are however slightly higher in our population, which is not surprising since only patients with MOF were recruited.

Assessment of recruitment bias

There is strong evidence of recruitment bias in the PREMISS study. Among the 46 measured initial covariates, 20 have a standardised difference superior to 10 % in absolute value. The first 10 unbalanced covariates are, in decreasing order of unbalance, age (the “after” patients are younger), the PaO_2/FiO_2 (a measure of hypoxemia) ratio (the “after” patients are more frequently ventilated), the McCabe score (the “after” patients are less severe), the Glasgow coma score (the “after” patients are less severe), the delay between hospital admission and ICU admission (shorter for the “after” patients), the presence of a neurological infection site (more frequent among the “after” patients), urine output (values between 0.75 and 1 L/24h are less frequent among the “after” patients), bilirubin (extreme values are less frequent among the “after” patients), the presence of an endo-cardiovascular infection site (less frequent among “after” patients) and the heart rate (extreme low values – < 30 bpm – are less frequent among “after” patients). However, patient severity did not differ between the groups, at ICU admission (SAPS II of 59.93 vs. 56.24 in the “after” group, p=0,54) and at the date of severe sepsis with MOF diagnosis (LODS of 8.62 vs. 8.90, p=0,17).

Workload and DA treatment cost estimation

The workload measured corresponds to the sum of the relative cost indexes associated with the technical acts measured. We will refer to it as the RCI. Two patients have a null RCI: they stayed respectively 3 and 5 days in the ICU and left it alive. The RCI is clearly non-normally distributed (Figure 1). Patients in “After” group seem to have a higher RCI (Figure 2).

Among the “after” treatment group, the mean DA cost was of €6668 (95% confidence interval of [6404 – 6931], for a mean patient weight of 74 kg and a cost of €48.23 per mg of DA). Only 51% of the patients conformed to DA’s indications of use (i.e. posology of 24µg/kg/h, during 96h unless the patient dies before or faces an adverse effect or a treatment contraindication). This figure illustrates the gap between the conditions of care in RCTs and in daily practice.

Figure 1 Distribution of the total RCI among the PREMISS study patients

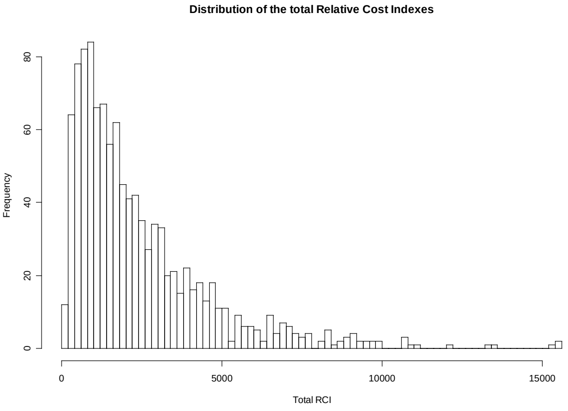
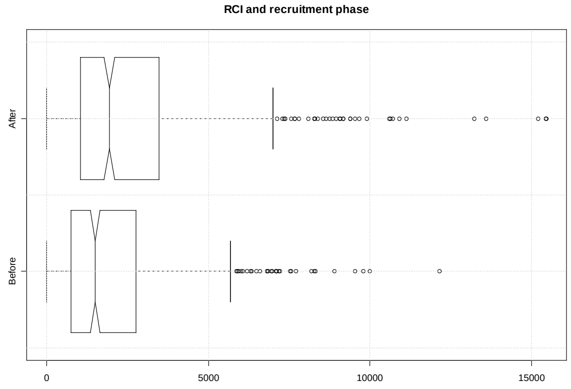


Figure 2 : Boxplots of the total RCI in the “before” and “after” treatment groups



Assessment of treatment phase’s impact on workload

Full sample multivariate regression

Without any covariate adjustments, the linear model assessing the impact of treatment phase on workload is given by:

$\text{RCI} = 2089.94 + 594.62 * I_{\{\text{Phase=“after”}\}}$

Where $I_{\{Phase=“after”\}}$ is the indicator variable for the “after” treatment phase ($I_{\{Phase=“after”\}} = 0$ if the patient belongs to the “before” group, 1 otherwise). The corresponding gamma regression model is given by:

$(\text{RCI}+1) = \exp(7.64537) * \exp(0.25028 * I_{\{\text{Phase=“after”}\}})$

On the gamma model, we added a constant to the RCI in order to avoid null RCI values (observed for two patients), as the link function in this model is logarithmic. Since the mean RCI is of 2408 points, one RCI point can be considered as negligible. Both models give exactly the same estimates, but the gamma model ones are multiplicative whereas the linear model ones are additive. Without any adjustments for covariate unbalance between the two treatment groups, DA is estimated to increase ICU workload by 28 % (exp(0.25028)=1.28). In both models, treatment phase is significant ($p<10^{-5}$). A quick examination of the linear model residuals shows the normality assumption is not met. According to Akaike's AIC, the gamma model is superior to the linear one (AIC of 19175 and 19969 respectively). We will therefore only present the gamma model results for the remaining of the article.

When adjusting for unbalanced covariates, 6 of them are retained in the model: age, the presence of ventilation (an indicator of respiratory failure), blood urea (high values of blood urea are a sign of renal failure), ICU admission by external transfer, the presence of a neurological infection site and the presence of an urinary tract infection site. A random intercept model was estimated using a penalized quasi-likelihood approach.[21] The models with or without random effects are compared in Table 1.

Table 1 Multivariate fixed-effects and random intercept model for the impact of DA on workload

Variable	**Fixed-effects model** exp(estimate) / (p-value)	**Random intercept model** exp(estimate) / (p-value)
Intercept	1967.1853	1943.3924
"After" phase	1.1813 (0.0031)	1.1910 (0.0020)
Age: [15.6-50.6[	1 (ref)	1 (ref)
Age: [50.6-64.0[	1.3859 (< 0.0001)	1.3673 (< 0.0001)
Age: [64.0-73.3[	1.1332 (0.1128)	1.1293 (0.1134)
Age: [73.3-94.8]	0.9617 (0.6262)	0.9322 (0.3714)
No ventilation[a]	0.4817 (< 0.0001)	0.4740 (< 0.0001)
Blood Urea ≥ 20 mmol/L	1.1956 (0.0044)	1.2121 (0.0016)
External transfer	1.1646 (0.0207)	1.1700 (0.0148)
Neurological infection	0.6205 (0.0003)	0.6331 (0.0003)
Urinary tract infection	0.8063 (0.0244)	0.8160 (0.0283)

a: mechanical ventilation or continuous positive airway pressure (CPAP) ventilation

The model estimates seem coherent: workload is increased for DA treated patients, for patients showing signs of respiratory (i.e. with ventilation) or renal (i.e. with high levels of blood urea) dysfunction. The effect of age on workload is not linear: patients in the second age quartile are the ones requiring more workload. The fact that workload is increased among external transfer admissions is in conformity with the observation that these patients tend to be more severe.[22] The role of the infection sites is uneasier to explain. In the random intercept model, the multiplicative effect of ICUs on baseline workload varies from 0.67 to 1.40, indicative of some variability in hospital practices. However, since the variance component due to individual variation is of 0.6840 (on the log scale) and the one due to cluster correlation is of 0.0398, only about 5.5% of the workload variation is due to variation between the participating ICUs. Once adjusted for unbalanced covariates, DA is estimated to increase ICU workload by 19% (in the random intercept model). This value, lower than the one estimated in the unadjusted model (28%), tends to prove that patients recruited in the "after" phase are less severe than patients in the "before" one.

Propensity score matching

The principal characteristics of the three PS models estimated are given in Table 2. The one with interactions (M3) led to the smallest sample and the greatest number of unbalanced covariates (according to the 10 % threshold in the standardised differences). Model (M1) has a smaller sample size than model (M2) but is more balanced: it was the one elected for the analyses.

Table 2 Comparison of the three competing PS models in the achievement of a balanced subsample

Characteristic	**(M1) – All covariates**	**(M2) – Selected covariates**	**(M3) – Interaction terms**
Sample Size	840	870	748
Standardised differences:			
Age class (≥ 80)	**14.98%**	**14.31%**	**15.56%**
PaO_2/FiO_2	**10.46%**	8.03%	4.42%
Systolic blood pressure	6.92%	**14.44%**	6.95%
McCabe score	8.59%	**11.89%**	6.45%
LODS Renal subscore	9.17%	**10.49%**	**12.64%**
Septic shock	8.21%	5.72%	**12.39%**
Natremia	3.46%	9.28%	**11.39%**
Admission category[a]	6.91%	9.61%	**10.37%**

a: medical, scheduled surgery, emergent surgery, operated traumatism or unoperated traumatism

Although recruitment bias is reduced in the resulting PS matched sample, there are still two covariates that could bias the estimations. The first one is age: although the age distributions cannot be distinguished in both phases (p = 0.33, Mann-Whitney test), the proportion of patients aged more than 80 years remains higher in the “before” group (10% vs 6%, p =0.04, Fisher’s “exact” test). However, our previous analyses showed that the older patients’ workload did not differ from the younger ones (Table 1). The other potential source of bias is related to the ventilation characteristics (as described by the PaO_2/FiO_2 ratio). The higher difference between the treatment groups (as measured by the standardised difference) is related to the “not concerned” category, that is the patients that are not under ventilation. This difference is not statistically significant (7% vs 5%, p = 0.14, Fisher’s exact test). These two covariates were already included in the multivariate gamma regression model.

An unadjusted analysis in the PS matched sample leads to the estimation of an 18 % increase in ICU workload (p = 0.0045, gamma regression model without adjusting for the patient’s clustering among participating ICUs). For comparison purposes, the multivariate model used in the full sample analysis was fitted in the paired sample. The results are summarised in Table 3: DA’s effect on ICU workload is further reduced to 14% (versus 19% in the full sample multivariate analysis).

Table 3 Multivariate PS-matched models for the impact of DA on workload

Variable	**Fixed-effects model** exp(estimate) / (p-value)	**Random intercept model** exp(estimate) / (p-value)
Intercept	1920.8190	1872.8954
“After” phase	1.1374 (0.0316)	1.1438 (0.0235)
Age: [15.6-50.6[	1 (ref)	1 (ref)
Age: [50.6-64.0[	1.4262 (< 0.0001)	1.4295 (< 0.0001)
Age: [64.0-73.3[	1.1530 (0.0996)	1.1558 (0.0812)
Age: [73.3-94.8]	0.9976 (0.9785)	0.9834 (0.8444)
No ventilation[a]	0.3739 (< 0.0001)	0.3784 (< 0.0001)
Blood Urea ≥ 20 mmol/L	1.2398 (0.0016)	1.2480 (0.0008)
External transfer	1.2162 (0.0075)	1.2202 (0.0049)
Neurological infection	0.6641 (0.0168)	0.6543 (0.0096)
Urinary tract infection	0.8029 (0.0336)	0.8069 (0.0305)

a: mechanical ventilation or continuous positive airway pressure (CPAP) ventilation

In the random intercept model, the baseline workload varies by a multiplicative constant ranging from 0.64 to 1.41. 7.0% of the workload variation is due to variation between the participating ICUs. The difference between the PS-matched and the full sample indicates that additional adjustments could have been made in the standard multivariate analysis. The one between the unadjusted and the multivariate PS-matched models indicates that the remaining unbalance between the treatment groups is still influential on the estimations.

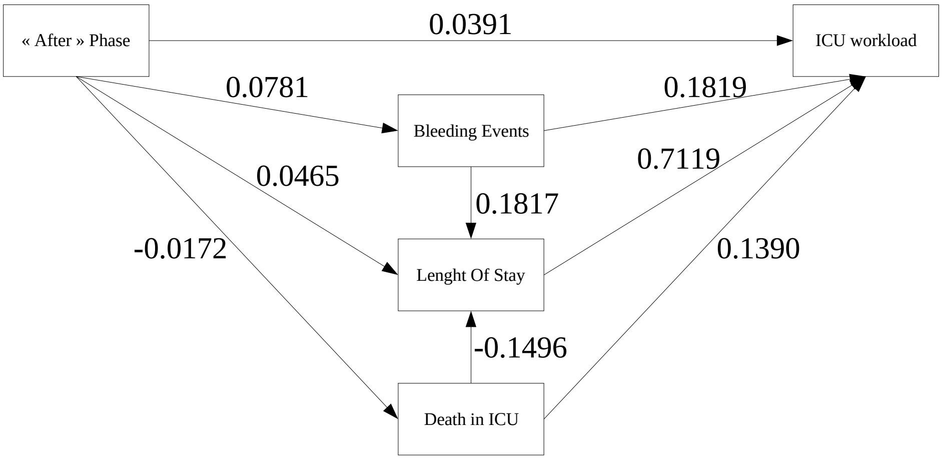
A sketch of DA's effect on workload through structural equation modelling

All methods used here lead to the same conclusion: DA use in the treatment of patients with severe sepsis and MOF is associated with an increase in ICU workload (estimated to be between 28 % and 14 %, depending on the method used). There are several ways for DA's to affect ICU workload:

- DA could be linked to the patients' length of stay (LOS). We expect an increase in LOS to be associated with an increase of ICU workload.
- DA could influence ICU workload through its effect on ICU mortality: patients not surviving ICU hospitalisation are known to generate more costs (and therefore more care procedures).[20] On the other hand, surviving patients have a higher ICU LOS.
- DA is associated with adverse events, haemorrhages in particular.[6] The occurrence of an adverse event is expected to lead to additional care and therefore to increase ICU workload. Adverse events could also increase the patients' LOS.
- Finally, DA could directly influence workload, if DA treated patients are more monitored.

Until now, DA's effect on ICU workload has been tested as a whole. We tried to explore the path from treatment phase to ICU workload using a structural equation model.[23] The interpretation of this model should be made with caution, as two of its endogenous variables (treatment phase and ICU mortality) are binary and the others (the number of bleeding events per patient and ICU LOS) have particular distributional shapes (the first is a counting event, the second is particularly skewed). Moreover, although the model is fitted on the PS-matched sample, we have seen that some degree of recruitment bias remains present. Finally the correlation of patients within participating ICUs has not been taken into account. However, the model estimates, presented in Figure 3, provide valuable information.

Figure 3: Structural equation modelling of DA's effect on ICU workload



The higher correlation estimate is between ICU LOS and ICU workload: as expected, most of ICU workload is driven by the duration of care in ICU. The second most influential variable on ICU workload is the occurrence of bleeding events. Bleeding events increase ICU workload directly and also indirectly, through an increase in ICU LOS. The effect of ICU mortality on ICU workload is less obvious: on one hand, deceased patients increase directly the ICU workload; on the other hand, they decrease the ICU LOS. On the whole, mortality remains positively associated with ICU workload (0.1390-0.1496*0.7119=0.0325). This is in agreement with the crude correlation coefficient between ICU mortality and workload, of 0.06. As for DA, it is associated with a decrease in ICU mortality, and it should therefore decrease ICU workload. However, it is also associated with an increase in ICU LOS and in the number of adverse events, both leading to an increase in ICU workload. In addition, there is a positive direct association between DA treatment and ICU workload. The structural model leads to the same conclusion as the regression ones: DA increases ICU workload.

DISCUSSION

The PREMISS study is illustrative of the role of observational studies in medico-economic research. Its objectives where different from those of a randomised clinical trial: the question was not to provide evidence in favour of the efficacy of an innovative treatment, but to gather information on the impact of the introduction of this innovation on a local (the French public hospitals) scale. It is our opinion that observational studies are the best suited for these purposes. Of course, the adoption of a non-randomised design brings its share of methodological issues. However, a careful and rigorous analysis can lead to adequate estimates. We developed a propensity score approach to cope with non-randomisation. There are other methods available,[24] and traditional multivariate regression methods can also perform well in the reduction of recruitment bias.[25] However, the PS methods are easy to implement and force the analysts to explicitly focus on the recruitment biases. Moreover, the use of PS matching can lead to simpler models, leading to more robust conclusions. Our analysis of the relationship between DA and ICU workload is illustrative of the strengths and weaknesses of the PS approach. An (inappropriate) crude analysis on the unmatched sample, without adjustments, leads to the conclusion that DA increases ICU workload by 28%. On the opposite, a multivariate analysis on a PS-matched sample estimates this increase at 14%. This procedure divides by two DA's estimated impact. The figure of 14%, inferior to the one obtained in the crude analysis of the PS matched sample (18%), illustrates one of the characteristics of the PS method: it is expected to reduce recruitment bias, but does not necessarily guarantee perfect balance between the populations being compared. Moreover, contrary to randomisation, the PS cannot assure balance of unobserved or unmeasured variables. When planning a PS analysis, it is important to measure a sufficient number of baseline characteristics. The drawback will be that the more variables are collected, the more the chances to observe differences among one of them.

What have we learned from this analysis? First of all, the presence of a recruitment bias itself deserves attention. The differences between the "before" and "after" groups are a sign that DA treated patients weren't selected on the basis of just the treatment indications (these corresponded to the study's inclusion criteria). As an example, DA treated patients are younger. Since DA is a relatively expensive treatment, the physicians could have judged that DA is not cost effective for older patients. This gap between the theoretical indications and the observed treatment practices could not have been observed in a randomised clinical trial. One other interesting aspect for decision makers is that DA's impact on the ICUs will not be limited to its acquisition cost (at present, DA's cost is reported separately and is reimbursed integrally to the ICU). The treatment will also have an impact on ICU workload and will in particular increase the patients' LOS.

These conclusions are based on a carefully planned analysis, where not only the issues due to non-randomisation were considered, but also other potential sources of bias. We dealt with recruitment bias using propensity score methods, we tried to take the skewness of the workload estimate into account by the use of gamma regression and we accounted for the clustering of patients among the ICUs using random effects modelling.

REFERENCES

[1] Dunn D, Babiker A, Hooker M, Darbyshire J. The dangers of inferring treatment effects from observational data: a case study in HIV infection. Control Clin Trials. 2002 Apr;23(2):106-10.

[2] Abel U, Koch A. The role of randomization in clinical studies: myths and beliefs. J Clin Epidemiol. 1999 Jun;52(6):487-97.

[3] Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med. 2000 Jun 22;342(25):1887-92.

[4] Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, Petticrew M, Altman DG; International Stroke Trial Collaborative Group; European Carotid Surgery Trial Collaborative Group. Evaluating non-randomised intervention studies. Health Technol Assess. 2003;7(27):iii-x, 1-173.

[5] Heckman JJ, Smith JA. Assessing the case for social experiments. Journal of Economic Perspectives. 1995;9(2):85-110.

[6] Bernard GR, Vincent JL, Laterre PF, LaRosa SP, Dhainaut JF, Lopez-Rodriguez A, Steingrub JS, Garber GE, Helterbrand JD, Ely EW, Fisher CJ Jr; Recombinant human protein C Worldwide Evaluation in Severe Sepsis (PROWESS) study group. Efficacy and safety of recombinant human activated protein C for severe sepsis. N Engl J Med. 2001 Mar 8;344(10):699-709.

[7] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41-55.

[8] Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a european/north american multicenter study. JAMA 1993 ; 270 : 2957-63

[9] Le Gall JR, Klar J, Lemeshow S, Saulnier F, Alberti C, Artigas A, Teres D. The Logistic Organ Dysfunction system. A new way to assess organ dysfunction in the intensive care unit. ICU Scoring Group. JAMA. 1996 Sep 11;276(10):802-10.

[10] Sznajder M, Leleu G, Buonamico G, Auvert B, Aegerter P, Merliere Y, Dutheil M, Guidet B, Le Gall JR. Estimation of direct cost and resource allocation in intensive care: correlation with Omega system. Intensive Care Med. 1998 Jun;24(6):582-9.

[11] Diehr P, Yanez D, Ash A, Hornbrook M, Lin DY. Methods for analyzing health care utilization and costs. Annu Rev Public Health. 1999;20:125-44.

[12] Manning WG, Mullahy J. Estimating log models: to transform or not to transform? J Health Econ. 2001 Jul;20(4):461-94.

[13] Normand ST, Landrum MB, Guadagnoli E, Ayanian JZ, Ryan TJ, Cleary PD, McNeil BJ. Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: a matched analysis using propensity scores. J Clin Epidemiol. 2001 Apr;54(4):387-98.

[14] D'Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med. 1998 Oct 15;17(19):2265-81.

[15] Drake C. Effects of misspecification of the propensity score on estimators of treatment effect. Biometrics 1993;49:1231-6.

[16] Schafer JL. Analysis of incomplete multivariate data. London: Chapman and Hall, 1997.

[17] Raghunathan TE. What do we do with missing data? Some options for analysis of incomplete data. Annu Rev Public Health. 2004;25:99-117.

[18] Localio AR, Berlin JA, Ten Have TR, Kimmel SE. Adjustments for center in multicenter studies: an overview. Ann Intern Med. 2001 Jul 17;135(2):112-23.

[19] Brun-Buisson C, Meshaka P, Pinton P, Vallet B; EPISEPSIS Study Group. EPISEPSIS: a reappraisal of the epidemiology and outcome of severe sepsis in French intensive care units. Intensive Care Med. 2004 Apr;30(4):580-8. Epub 2004 Mar 02.

[20] Adrie C, Alberti C, Chaix-Couturier C, Azoulay E, De Lassence A, Cohen Y, Meshaka P, Cheval C, Thuong M, Troche G, Garrouste-Orgeas M, Timsit JF. Epidemiology and economic evaluation of severe sepsis in France: Age, severity, infection site, and place of acquisition (community, hospital, or intensive care unit) as determinants of workload and cost. J Crit Care. 2005 Mar;20(1):46-58.

[21] Venables WN, Ripley BD. Modern Applied Statistics with S. Fourth edition. Springer, 2002.

[22] Guidet B, Aegerter P, Gauzit R, Meshaka P, Dreyfuss D; CUB-Rea Study Group. Incidence and impact of organ dysfunctions associated with sepsis. Chest. 2005 Mar;127(3):942-51.

[23] Kline RB. Principles and practice of structural equation modeling. Second edition. New York: The Guilford Press, 2005.

[24] Klungel OH, Martens EP, Psaty BM, Grobbee DE, Sullivan SD, Stricker BH, Leufkens HG, de Boer A. Methods to assess intended effects of drug treatment in observational studies are reviewed. J Clin Epidemiol. 2004 Dec;57(12):1223-31.

[25] Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. J Clin Epidemiol. 2005 Jun;58(6):550-9. Epub 2005 Apr 19.