**SUPPLEMENTAL DIGITAL CONTENT**

**TABLE OF CONTENTS**

**SUPPLEMENTAL TEXTS**

Supplemental Text 1: Details of patients and data collection

Supplemental Text 2: Formulation of the CLOGLOG model and the estimation method

Supplemental Text 3: Alternative ways of specifying the level and change variables of TVBIs

Supplemental Text 4: Temporal pattern of the daily hazard of dying

Supplemental Text 5: Dependence of the face value of AUC on the choice of the set of outcomes to be predicted

Supplemental Text 6: How the Day 1 variable of creatinine lost its usefulness in the CLOGLOG model

Supplemental Text 7: Evaluation of the relative importance among TVBIs and the relative importance between the day 1 and change variables of each TVBI

Supplemental Text 8: Insights from the temporal patterns of TVBIs

Supplemental Text 9: Construction and use of mortality risk profiles for individual septic patients

Supplemental Text 10: Validation

Supplemental Text 11: Assessment of the predictive powers of MODS, SOFA and APACHE II

Supplemental Text 12: Assessment of the usefulness of four additional TVBIsand other contextual variables

Supplemental Text 13: Addition of the information on the sites and types of positive cultures had little effect on the predictive power of the CLOGLOG model

Supplemental Text 14: Comparison of the CLOGLOG model with two versions of the Cox model

Supplemental Text 15: Obtaining similar longitudinal insights from a longitudinal logit model or a longitudinal probit model

**SUPPLEMENTAL TABLES**

Supplemental Table 1. Summary of septic patient recruitment from 9 participating centres

Supplemental Table 2. The means and standard deviations of the six TVBIs of the 356 septic patients with mon-missing values as of Day 1 and the last Day: Non-survivors versus survivors

Supplemental Table 3. How the Day 1 Variable of creatinine lost its usefulness in the CLOGLOG Model for 356 septic patients

Supplemental Table 4. Computation of the predictive powers of the predictive variables from differences in the additive terms of log of hazard between (1) the Mean of Non-survivors and (2) the Mean of Survivors, based on the estimated coefficients of the CLOGLOG model

Supplemental Table 5. The predictive powers of the level and change variables of the TVBIs and their Relative Importance

Supplemental Table 6. The primary data of Patient A and the benchmark for constructing his mortality risk profile

Supplemental Table 7. The construction of Patient A's mortality risk profile, based on the CLOGLOG model, with the mean of best 10% of survivors used for setting up the benchmark

Supplemental Table 8. The primary data of Patient **B** and the benchmark for constructing his mortality risk profile

Supplemental Table 9. The construction of Patient B's mortality risk profile, based on the CLOGLOG Model, with the mean of best 10% of survivors used for setting up the benchmark

Supplemental Table 10. Baseline Characteristics of the 28 Septic Patients in the Validation Group

Supplemental Table 11. The Results of the 10 Fold Cross-validation for the Best CLOGLOG Model shown in Table 2 of the Manuscript

Supplemental Table 12. Estimation results of the CLOGLOG model for assessing the predictive powers of MODS, SOFA, and APACHE II

Supplemental Table 13. The estimated coefficients of four additional TVBIs and other contextual variables, obtained by inserting each of them separately into the CLOGLOG model shown in Table 2

Supplemental Table 14. The addition of information on the sites and types of positive cultures had little effect on the predictive power (AUC or Rho-square) of the CLOGLOG model

Supplemental Table 15. Comparison of the CLOGLOG Model with Two Versions of the Cox Model for 356 Septic Patients

Supplemental Table 16. The Computation of Patient A's Probability of Dying in 28 Days and the Construction of His Mortality Risk Profile, Based on the Longitudinal Logit Model, with the Mean of Best Decile of Survivors Used for Setting up the Benchmark

Supplemental Table 17. Simple Correlation Coefficients of the Day 1 and Change Variables of the 6 TVBIs: Based on the 6,724 Person-day Records of 356 Septic Patients

**SUPPLEMETAL FIGURES**

Supplemental Figure 1. The contrast between the exponential form and the power form for the specification of the dependence of the hazard on the current variable of platelet count in the CLOGLOG model

Supplemental Figure 2. Temporal patterns of the observed and predicted daily Hazard of dying

Supplemental Figure 3. The mortality risk profile of Patient B that highlights the relative contribution of each TVBI to the risk of dying

Supplemental Figure 4. The ROC curves for the derivation and validation groups, using the probability of dying in 28 Days as the classifier

Supplemental Figure 5. The mortality risk profile of Patient A that highlights the relative contribution of each TVBI to the risk of dying: generated by *Longitudinal Logit Model*

**SUPPLEMETAL TEXTS**

This Supplement is organized into 15 Supplemental Texts. Supplementary Texts 1-13 provide a more detailed description of the data collection method and the statistical analyses. Supplementary Texts 14 shows that the CLOGLOG model is a versatile version of the Cox model and is better than two versions of the Cox model for achieving longitudinal insights. For researchers who are more comfortable with using logit models, Supplement 15 demonstrates that similar longitudinal insights can be obtained by using a “longitudinal logit model”. Supplemental Tables and Figures are arranged according to the sequence of the Supplemental Texts.

**Supplemental Text 1: Details of data collection methods**

**Patients and inclusion criteria**

Our data are from a prospective, multi-centre observational study of 392 septic ICU patients (the DYNAMICS Study, Clinical Trials.gov Identifier:NCT01355042). The study was approved by the Research Ethics Boards of all participating centers. Written informed consent was obtained from the patient or substitute decision-maker prior to enrolment into the study. When a priori consent was not feasible, a deferred consent approach was used. All septic events were adjudicated by at least 2 experienced ICU physicians. The inclusion criteria were a modification of those defined by Bernard *et al*.(1). Patients must have a confirmed or suspected infection on the basis of clinical data at the time of screening, at least one dysfunctional organ system, ≥3 signs of systemic inflammatory response syndrome (SIRS), and were expected to remain in the ICU for ≥ 72 hours. The presence of organ dysfunction are: (1) SBP ≤90 mm Hg or MAP ≤70 mm Hg or SBP < 40 mm Hg for at least 1 hour despite fluid resuscitation, adequate intravascular volume status, or use of vasopressor in attempt to maintain systolic BP ≥90 or MAP ≥ 70 mm Hg, (2) P/F Ratio ≤ 250 in the presence of other dysfunctional organs or systems, or ≤ 200 if lung is the only dysfunctional organ, (3) acute rise in creatinine > 171 mM or urine output <0.5 ml/kg body weight for 1 hour despite adequate fluid resuscitation, (4) unexplained metabolic acidosis (pH ≤ 7.30 or base deficit ≥ 5 with lactate > 1.5 times the upper limit of normal, and (5) platelet count < 50,000 or a 50% drop over the 3 days prior to ICU admission. The inclusion criteria for septic shock are the same as those for sepsis except that patients must be on vasopressors within the previous 24 hours. Patients were excluded if they were < 18 years old, pregnant or breastfeeding, or were receiving palliative care. Deferred informed consent was obtained for enrolment in 64.8 % of the patients and *a priori* consent was obtained for 34.2 % of the patients.

**Clinical data collection**

Baseline characteristics include demographic information, organ function scores (MODS, SOFA), pre-existing conditions, sites and types of infection, APACHE II score, use of mechanical ventilation, and use of vasopressor/inotropes. Daily data included culture results, organ function, hematologic and other laboratory tests, and type/quantity of resuscitation fluid.

**Blood collection**

Patient blood samples were collected within 24 hours of meeting the inclusion criteria for sepsis. Blood was obtained at baseline, then daily for the first week, followed by once a week for the duration of the patients’ stay in the ICU. The blood was processed within two hours of blood collection. Briefly, blood (4.5 mL) was collected from existing arterial or venous lines (or by venipuncture with a 20-gauge needle) into Becton Dickinson buffered sodium citrate vacutainer tubes (0.105M trisodium citrate). The blood was centrifuged at 1,700 x g for 10 min at 20oC (with the brake off), and the plasma was stored as 200 uL aliquots at -80oC and thawed at the time of assays.

**Plasma samples from healthy controls.**

Plasma samples were obtained from 34 healthy adult volunteers who were not receiving medication at the time of blood sampling. No attempt to match cases and controls was made.

Q**uantification of cfDNA and protein C levels in plasma**

cfDNA was isolated from 200 µL of plasma using the QIAamp DNA Blood Mini Kit (Qiagen, Valencia, CA). The concentration of the DNA was measured by UV absorbance at 260 nm using a spectrophotometer (BioPhotometer Plus, Eppendorf, Mississauga, ON). DNA purity was confirmed by determining the OD260/OD280 ratio. Plasma levels of protein C antigen were quantified by an enzyme immunoassay (Affinity Biologicals Inc., Ancaster, ON).

**Supplemental Text 2: Formulation of the CLOGLOG model and the estimation method**

Since the daily health status of each patient changes with time partly in response to treatments, we used the CLOGLOG model to follow the daily life of each patient. Similar to the Cox proportional hazards model, this model expresses the hazard of dying as a linear-in-coefficient exponential function of observable explanatory variables:

where is the ith patient’s hazard of dying at any time point on day t; is an unknown intercept; is a row vector of unknown coefficients; is a column vector of the explanatory variables that reflect the relevant information on the ith patient up to day t; n is the number of patients in the sample; and is the day when the ith patient died in ICU/hospital, was discharged, or was censored (day 28). A discharge is defined as the transfer of a live patient from the ICU or hospital to home or other institution where the information on mortality status was no longer collected. Since each live patient is censored on day 28, both the day of death and the day of discharge are ≤ 28. We use day as the unit of time and call the *daily hazard of dying*. Implicit in this formulation is the simplifying assumption that remains constant through all time points on day t. We call t the *current day* and the *last day* of the ith patient. We are mainly interested in how changed through time in response to changes in six time-varying biological indicators (TVBIs).

For each of the six TVBIs, we defined the following three analytical variables: (1) the *day 1 variable*, which assumes the same day 1 value of the indicator for all ; (2) the *current variable*, which assumes the observed (directly observed or imputed) value of the indicator on day ; and (3) the *change variable*, which in its simple form is defined as the current variable minus the day 1 variable. Each TVBI is represented by its *day 1* and *change* variables in the CLOGLOG model, with the day 1 variable quantifying the *initial* *level effect*, and the change variable quantifying the *change effect* since day 1. To enhance the model’s predictive power and to obtain biologically meaningful insights in regards to inter-patient differences and temporal changes in TVBI values, we transform some of these variables, as described and explained in Supplemental Text 3.

For any daily value of a current variable that is not directly observed, we imputed it as follows. If the day in question is preceded by at least one day with directly observed value and is followed by at least one day with directly observed value, then it is linearly interpolated from the two closest observed values. Otherwise, it is set to be equal to the nearest observed value. Note that for any TVBI, we do not allow any missing daily value of a patient be imputed from the observed values of any other patient.

To reduce the risk of making misleading inferences from observational data, our model includes the following contextual variables: two preconditions (chronic lung disease and previous brain injury, with each represented by a separate dummy variable), age, duration of stay, and log(duration of stay). In Supplemental Text 4, we provide evidence that the omission of duration and log(duration) from the model resulted in some misleading findings.

To estimate the unknown coefficients of the CLOGLOG model, we use the *maximum likelihood method* in the following way. We first translate the daily hazard of dying into the daily probability of dying , *conditional* on the survival to the beginning of day t. Since is assumed to be constant through all time points on day t, we find that

Let be a dummy variable that assumes the value of 1 if the ith person died on day t. The unknown coefficients are estimated by maximizing the following log-likelihood function

Note that “log” means natural logarithm. To carry out the estimation, we use the Logistic procedure of SAS (Version 9.4) with the option of LINK=CLOGLOG (2).

Mathematically, Eq. (2) is equivalent to

Since the left-hand-side of Eq. (4) is called a complementary log-log function of , this model is called a *complementary log-log model (2)*. The model links the linear-in-coefficient expression of the explanatory variables to the CLOGLOG function of the daily conditional probability of dying. From Eq. (4), the model may be considered as a *discrete-time* model. However, since we choose to formulate it by Eq. (1), it is a *continuous-time* model.

The input data matrix has a simple structure. Each row represents a person-day, in which the information of all explanatory variables is used to enhance the likelihood of the value of the outcome variable (). The original data file for all 392 septic patients contained 7,298 observations (rows). After removing the observations of the 36 patients with missing values for some TVBIs, the number of patients was reduced to 356, and the number of observations of our input data matrix was reduced to 6,724.

The number of patients with missing values was: 2 patients for cfDNA, 3 patients for protein C, 1 patient each for platelets and creatinine, and 32 patients for lactate. The proportion of non-survivors remained at 24% after the removal of the patients with missing values. Little’s test shows that the pattern of the missing values across the 5 TVBIs with missing values is non-random (p=0.035).

Upon finding that our input data matrix has only 85 positive outcomes in as many as 6,724 observed outcomes, a reviewer intuitively feels that by setting the intercept to a negative value of very large magnitude and all other coefficients to 0, we can make all predicted probabilities of dying be practically 0, resulting in 98.7% of all daily outcomes being correctly predicted! But, this nearly perfect prediction has no scientific value, because the sensitivity is 0. Actually, the maximum likelihood method does not attempt to maximize the proportion of all outcomes being correctly predicted. Instead, it maximizes the *likelihood* of the set of observed daily outcomes. In our application, its estimated result can help achieve a balance of sensitivity and specificity at 0.78, with the sum of the two desirable measures being 1.58. The reviewer’s intuition is clearly unhelpful.

Since the Logistic procedure of SAS does not compute the 95% confidence interval for AUC (the area under the curve showing the relationship between sensitivity and 1-specificity in the ROC analysis), we wrote a SAS module to carry out the computation. In our module, the standard deviation of the AUC was computed according to the algorithm developed by Hanley and McNeil (3). To better reflect the effects of the sample size and the number of unknown coefficients on the width of the confidence interval, we took the critical value for constructing the confidence interval from a t-distribution rather than the standard normal distribution.

Since the daily probability of dying considered in this paper is a *conditional probability,* the probability of dying in 28 days should not be computed by adding up 28 daily probabilities of dying. The probability of dying in 28 days implied by the daily hazard of dying *H* is .

The comparisons of the CLOGLOG model with two conventional versions of the Cox model are shown later in Supplemental Text 14, where the justification of the CLOGLOG model as a *versatile* version of the Cox model is also provided

.

**Supplemental Text 3: Alternative ways of specifying the level and change variables of TVBIs**

For some TVBIs like platelet count, a given difference of, say, 50 units between two patients could have a much greater effect on their risk difference if the sicker patient had a level of 60 units rather than 150 units. We thus log-transformed the day 1 variables of protein C, platelet count, and creatinine to enhance the model’s predictive power. Mathematically, the log-transformation changed the dependence of the hazard from an *exponential function* to a *power function*.

Supplemental Figure 1 shows the difference between these two functions for the day 1 variable of platelet count in the CLOGLOG model. In drawing the curves, the values of all other variables in the model were set at the mean of all patients. The shapes of the two functions differ markedly. The power function (red continuous line) has a greater curvature and a flatter tail, suggesting that variations in platelet counts above 250 have little effect on the risk of dying.

The model’s predictive power could also be enhanced by replacing the simple difference between current and day 1 variables of some TVBIs with proportional changes. Let be the day 1 variable, and be the current variable of the TVBI in question. A useful alternative specification of its change variable is , where is the *proportional* change. It is called *change factor* in the tables reporting the estimated coefficients. Mathematically, the switch to this alternative implies the replacement of the exponential function by the power function , where is an unknown coefficient to be estimated. In our model, the change factor is more suitable than the simple change for representing the change variables of protein C, platelets, creatinine, and lactate. In other words, proportional change is better than simple change for quantifying the change effects of these 4 TVBIs.

An important methodological issue is whether the day 1 variable of any TVBI had a non-monotonic effect. For example, in the construction of various versions of APACHE, several variables such as body temperature were assumed to have non-monotonic effects. To deal with this issue, we did the following. Expanding from the best estimated result of the CLOGLOG model reported in Table 2, we let the effect the day 1 variable of each TVBI be quantified by two variables simultaneously: its day 1 variable and the log of the day 1 variable. The combination of these two variables provided a flexibility to allow the data to decide whether the TVBI in question had a non-monotonic effect, which was to be revealed by the two variables having coefficients with opposite signs. This was done for each TVBI in turn. We found that the signs and p-values associated with the two variables were: same sign for cfDNA, with p=0.28 and 0.34; same sign for protein C, with p=0.47 and 0.21; opposite signs for platelets, with p=0.71 and 0.10; same sign for GCS, with p=0.41 and 0.85; opposite signs for lactate, with p=0.21 and 0.90. Thus, we infer that within our observed data range, none of the TVBIs had a significant non-monotonic effect.

**Supplemental Text 4: Temporal pattern of the daily hazard of dying**

In the CLOGLOG model, we use duration and log(duration) as two of the explanatory variables (1) to help capture the temporal pattern of the hazard of dying and (2) to prevent selection biases from resulting in misleading findings. Mathematically, this specification of the time function expresses the dependence of hazard on duration as a product of an exponential function and a power function. It has the advantage of being highly flexible in reflecting the temporal pattern in the data.

Using duration and log(duration) as the only two explanatory variables in the CLOGLOG model, we obtained the estimated function , where t is the duration of stay and is the estimated hazard of dying on day t. This function is represented by the green (continuous) curve in Supplemental Figure 2. The curve peaked in the later part of the first week and then declined.

Although the p-value associated with the variable “Duration” in Table 2 of the paper were somewhat large (0.106), we chose not to set its coefficient to 0 for the following reason. Since “Log(Duration)” is a monotonically increasing function of “Duration”, they must be positively correlated. This correlation contributed to the inflation of the standard error of the estimated coefficient of “Duration” and hence the inflation of the p-value.

From the predicted daily *probabilities* of dying for all 6,724 person-days generated by the SAS/Logistic Procedure for the best CLOGLOG model, we translated them into 6,724 predicted daily *hazards* of dying, using the formula , where is the predicted hazard of patient i on day t and is the corresponding predicted daily probability of dying. The 6,724 predicted hazards were than used to generate 28 predicted *mean* daily hazards of dying by averaging across all patients in the at-risk population on each of the 28 days. These 28 predicted mean hazards were then represented by the blue thick broken curve in Supplemental Figure 2. The curve shows that there is an unusually low hazard of dying on days 1 and 2 followed by an increase in the hazard of dying during days 3 to 5.  This pattern likely reflects the effect of an eligibility criterion for inclusion into the study (ie. patients are to be excluded from the study if they are not expected to remain alive in the ICU for ≥ 72 hours).

Since the estimated result of the best CLOGLOG model also included the lower and upper limits of the 95% confidence intervals for all 6,724 predicted daily *probabilities* of dying, we also translated these limits to those of the predicted daily *hazards* of dying and then averaged them into 28 pairs of limits in the same way. The resulting lower and upper limits of the confidence belt for the hazards are shown in Supplemental Figure 2 by blue thin broken curves with long segments. The large width of this confidence belt reflects that predictions of mortality outcomes on a daily basis are a difficult task. It also reflects that the patients on each day differed markedly in health status.

We are concerned that some researchers may be negatively impressed by the wide confidence belt and then retreat to the conventional versions of the Cox model, resulting in the inability to achieve longitudinal insights. To counter this negative impression, we now consider the confidence intervals for the predicted *mean* daily hazards. Since the standard error of the mean is the standard deviation of the individual values divided by the square root of the sample size, we shrink the widths of the original confidence intervals for predicting individual hazards by dividing them by the square roots of the sizes of the daily at-risk populations. The resulting confidence belt for the predicted *mean* daily hazards is shown to be quite narrow in Supplemental Figure 2 by a pair of blue thin broken curves with short segments.

In general, the temporal pattern of the hazard is hard to interpret, because it absorbs the effects of the balance of various selection biases, such as (1) the exclusion of patients who were not expected to remain alive in the ICU for ≥ 72 hours, (2) the temporal decline in the number of sicker patients due to the death process, and (3) the temporal decline in the number of healthier patients due to the discharge process.  The decline of the hazard after day 7 likely reflects the beneficial effects of the treatments and care received by the patients in ICU.

Keeping selection biases in mind, it is not surprising that the sharp rise in the hazard from day 1 to day 5 is inconsistent with the fact that the daily averages of five of the six TVBIs improved during the same time interval: cfDNA decreased by 3.8%, protein C increased by 25.5%, creatinine decreased by 19.4%, GCS increased by 18.0%, and lactate decreased by 30.7%. This inconsistency was resolved in the CLOGLOG model by the positive estimated coefficient of “log(duration)”. We note that the omission of both duration and log(duration) from the model led to the effects of most TVBIs being underestimated or, in the case of the day 1 variable of lactate,  even non-significant (p=0.221). This finding reveals that the above-mentioned eligibility criterion should not be used any more. It is worth noting that among the patients contributing to the input data, one died on day 1, 4 died on day 2, and 8 died on day 3, suggesting that some clinicians realized that this criterion had the undesirable effect of resulting in not only selection bias but also loss of valuable information and hence chose not to use it. A reviewer commented that the retention of these patients might indicate that “clinical gestalt cannot perfectly predict ICU mortality” and “hence the need for this study”.

Supplemental Figure 2 also shows circles representing the so-called “observed daily hazards of dying”, although hazards are not directly observable. They were computed from the longitudinal data according to the method used for constructing the Kaplan-Meier survival curves, without using any model. The irregular scatter of the circles suggests that predicting the mortality outcome of a patient on any given day is a difficult task.

**Supplemental Text 5: Dependence of the face value of AUC on the choice of the set of outcomes to be predicted**

Since the dependent variable in our CLOGLOG model is the *daily* hazard of dying, the logistic procedure of SAS automatically uses the predicted *daily* probabilities of dying (P1) to conduct the ROC analysis and compute the value of AUC. In other words, the value of AUC was computed by using the daily probability of dying as the *classifier* to predict *the set of 6,724 daily outcomes of all 356 patients*. The face value of the AUC turned out to be 0.865 (95% CI, 0.826 – 0.903). It is important to keep in mind that this face value should not be compared with the AUC value computed in other publications that used a logistic model with the dependent variable being the probability of dying in 28 days (P28), because in the latter case P28 was used as the *classifier* to predict *a smaller set of outcomes, with one outcome per patient*. We conducted simulations that compared the AUC values computed by the two classifiers (the daily probability of dying versus the probability of dying in 28 days) for the same set of the daily hazards of dying. We found that the former was markedly smaller than the latter in most cases. An explanation for this large difference is that predicting whether a patient would die on a given day is more difficult than predicting whether a patient would die in 28 days. In other words, the set of outcome to be predicted by P1 is inherently more difficult to predict than the set of outcomes to be predicted by P28. Our simulations also indicate that the difference in the face values of AUC between the two classifiers tends to become larger when the level of hazards is raised and when the predictive power of the model becomes stronger.

To make the face values of AUC comparable to those in other papers, we recomputed the AUC in the following way. From the large input file of person-day records, we selected the record of the last day for each patient. For each patient, the selected record was used to compute his/her predicted daily hazard of dying, based the estimated coefficients of the CLOGLOG model. The predicted hazards were then transformed into the predicted probabilities of dying in 28 days by the formula , where was the ith patient’s predicted daily hazard of dying. These predicted probabilities were then used to conduct the ROC analysis on the set of 356 outcomes and then compute the value of the AUC, which turned out to be 0.903 (95% CI, 0.864 – 0.941).

To avoid confusion, in several tables, we use “AUC\_P1” to denote the AUC that is based on using P1 as the classifier, and use “AUC\_P28” to denote the AUC that is based on using P28 as the classifier.

**Supplemental Text 6: How the Day 1 variable of creatinine lost its usefulness in the CLOGLOG model**

Panel 1 of Supplemental Table 3 shows that in the context of the preconditions of chronic lung disease and previous brain injury, age, and duration, the day 1 variable of creatinine, with p=0.1002, had a somewhat uncertain effect on the mortality hazard, although its estimated coefficient had the biologically meaningful positive sign. This finding is misleading because the day 1 and change variables of creatinine had a strong negative correlation (r= -0.57, N=6,724, p<0.001). This negative correlation implied that large improvements in creatinine occurred mostly to patients whose initial values were relatively poor (high), so that the pool of survivors contained increasingly higher proportion of patients whose creatinine was relatively poor on day 1 but had experienced large improvement (reduction) afterwards. To control for the distorting effect of this selective improvement, the proper assessment of the effect of the day 1 variable required the simultaneous inclusion of the corresponding change variable into the model. Since strong negative correlation between day 1 and change variables occurred to all six TVBIs (see Supplemental Table 17), the use of the day 1 variable of each TVBI should be accompanied by the corresponding change variable. Otherwise, the effects of most day 1 variables will be understated.

Panel 2 of Supplemental Table 3 shows that the addition of the change variable of creatinine not only raised the AUC\_P1 markedly from 0.623 to 0.693 but also helped make the corresponding day 1 variable highly significant (p=0.0036) and caused its estimated coefficient to increase markedly from 0.2345 to 0.4320. Thus, controlling for the selective improvement, patients with relatively high day 1 creatinine were found to be at higher risk of dying.

However, Panel 3 of Supplemental Table 3 shows that the further addition of the day 1 and change variables of protein C and platelets caused the day 1 variable of creatinine to become a non-significant explanatory variable (coefficient=0.1593, p=0.3231). This was the consequence of the overlap between the weaker predictive power of the day 1 variable of creatinine and the stronger predictive powers of the day 1 variables of protein C and platelets. Behind this overlap were the significant correlations of the day 1 variable of creatinine with the corresponding variables of protein C and platelets. The two correlations were -0.155 (N=6,724, p<0.0001) for protein C and -0.202 (N=6,724, p<0.0001) for platelets. In short, the loss of usefulness of the day 1 variable of creatinine resulted from a multicollinearity problem. However, it is useful to remember from Panel 2 that a strong correlation between explanatory variables need not result in a multicollinearity problem.

**Supplemental Text 7: Evaluation of the relative importance among TVBIs and the relative importance between the level and change variables of each TVBI**

The hazard ratios computed from the estimated coefficients by exponentiation are not suitable for comparing the relative importance of the explanatory variables because the explanatory variables do not share a common unit. To overcome the comparability problem, we introduced a new way for assessing the relative importance of the explanatory variables. We started by transforming the CLOGLOG model into the following form:

where is the kth element of and is the kth element of . Despite the fact that the explanatory variables have different physical units, the additive terms on the right-hand-side of Eq. (5) have a common unit, log(1/day), so that their magnitudes can be used to evaluate the relative importance among the explanatory variables in determining the log of hazard. We call the *additive contribution to the log of hazard of dying* by the kth explanatory variable.

We then computed the additive contributions to two representative log of hazards: (1) the predicted log of hazard computed for the mean of the non-survivor group, and (2) the predicted log of hazard computed for the mean of the survivor group. For each explanatory variable, the *difference* in its additive contributions to these two representative log of hazards is then considered as its *predictive power* in distinguishing non-survivors from survivors: the greater the difference, the greater the predictive power.

Using the mean values of the 6 TVBIs on day 1 and the last day for non-survivors and survivors in Supplemental Table 2 as well as the contextual variables, Supplemental Table 4 demonstrates the computation of the difference in the additive contributions to the log of hazards between (1) the mean of non-survivors and (2) the mean of survivors for each explanatory variable. For example, the means of cfDNA on day 1 were 6.126 µg/mL for non-survivors and 4.705 µg/mL for survivors. By multiplying these two means by the estimated coefficient of 0.1857, we found that the additive contributions of the day 1 variable of cfDNA to the log of hazard were 1.138 for non-survivors and 0.874 for survivors. Hence, the predictive power of the day 1 variable of cfDNA was 1.138-0.874=0.264, which is shown in the last column of the table. Such computations were done for all explanatory variables in the table. From the last column, we found that with respect to day 1 variables, cfDNA had the greatest predictive power (0.246). With respect to change variables, GCS had the greatest predictive power (0.701).

The combined predictive power of the day 1 and change variables of each TVBI was computed by summing the predictive powers of its day 1 and change variables. For example, the combined predictive power of cfDNA was 0.264+0.028=0.292, whereas the combined predictive power of GCS was 0.021+0.701=0.722. In other words, GCS was much more powerful than cfDNA in distinguishing non-survivors from survivors. The predictive power of GCS came mostly from its change, whereas the predictive power of cfDNA came mostly from its initial level.

The combined predictive power of the day 1 and change variables differed markedly among the six TVBI. GCS is about four times as important as creatinine. This finding raises a question about the appropriateness of giving equal weight to GCS and creatinine in the creation of MODS and SOFA. Since lactate, cfDNA, and protein C were not part of MODS and SOFA but were found to have rather strong predictive powers, it is likely that the combination of our chosen TVBIs would outperform both MODS and SOFA.

In terms of the combined predictive power of the day 1 and change variables, the six TVBIs could be divided into 3 groups: (1) GCS (0.72) and lactate (0.64) on the top; (2) protein C (0.33) and cfDNA (0.29) in the middle; and (3) platelets (0.22) and creatinine (0.18) at the bottom. By summing up the predictive powers of the set of six TVBIs and the set of contextual variables separately, we found that the TVBIs accounted for 93% of the overall predictive power, leaving 7% to the contextual variables.

The information in the last columns of Supplemental Table 4 was transferred to Supplemental Table 5 for assessing the relative importance of the day 1 against the corresponding change variable for each TVBI. The data in this table was used to create Figure 1A in the paper.

The relative importance between the day 1 and change variables also differed markedly among the 6 TVBIs. Further insights about this marked difference will be obtained when we examine the temporal patterns of the 6 TVBIs in Supplemental Text 8. Note that the reason for the change variable to represent 100% of creatinine’s predictive power was that the weaker predictive powers of the day 1 variable of creatinine overlapped with the stronger predictive powers of the day 1 variable of protein C and platelets (Supplemental Text 6).

**Supplemental Text 8: Insights from the Temporal Patterns of TVBIs**

One temporal attribute shared by all six TVBI was that the day 1 variable had a strong negative correlation with the corresponding change variable (r = -0.80 for lactate, -0.70 for GCS, -0.58 for cfDNA, -0.57 for creatinine, -0.36 for platelets, and -0.27 for protein C; N=6,724). This attribute suggested that sicker patients on day 1 tended to benefit more from treatments and to experience greater improvement. Keeping this general attribute in mind, for gaining insights into the temporal pattern of each TVBI, we separated the patients into four quartile groups in terms of the day 1 variable, and then examined the trend of the daily averages of each group. To avoid the selection bias resulting from the death process that could misleadingly exaggerate improvements as the sickest patients in each group were successively removed, the daily records of all non-survivors were removed from the data before the daily averages were calculated. The resulting graphs are shown in Figure 1 in the paper. Except for the less clear evidence for platelets, the worst quartile group experienced the greatest improvement for each TVBI.

To see the effects of removing non-survivors, we compared the temporal graphs in Figure 1 and the corresponding temporal graphs generated without the removal of non-survivors from the sample. We found that, as expected, the failure to remove non-survivors resulted in the exaggeration of improvements (data not shown).

Being the two TVBIs with the greatest predictive powers, GCS and lactate turned out to show the greatest improvements: the average of the worse quartile group experienced a sharp and rapid improvement, and the averages of all quartile groups converged to a narrow range around a low risk level. This temporal pattern is consistent with the finding that most of the predictive powers of GCS and lactate (97% and 88%) come from their change variables. This finding suggests that the treatments that helped improve GCS and lactate contributed greatly to the reduction of the mortality risk level.

For cfDNA, the average of the worst quartile group remained at the high risk level of 6 ug/mL, after a brief improvement from day 1 to day 4, while the average of the best quartile group increased from less than 3 ug/mL to nearly 4 ug/mL. This temporal pattern corresponded to the finding that a very high proportion of the predictive power (91%) of cfDNA came its day 1 variable, leaving only 9% for its change variable. This finding suggests that novel strategies for reducing cfDNA may improve outcome.

Protein C was the third most powerful predictor of the hazards of dying, with 45% of its predictive power coming from its change variable. There was a general pattern of improvement in protein C for all quartile groups, with the worst quartile group experiencing the greatest improvement. However, by day 28, the gap between the worst and best quartile groups remained quite large, with the average of the worst quartile group being <70% of the normal level. Compared with lactate and GCS, the improvement in protein C was smaller and more prolonged, so that its contribution to the overall reduction of mortality risk was much less. Since protein C is similar to cfDNA in having a strong predictive power for its day 1 variable, a low level of protein C was strongly connected to a high risk of dying. Strategies to increase protein C levels may also improve outcome.

Platelet counts had the second weakest predictive power, with 54% of it coming from the change variable. It has the distinctive feature of a general lack of improvement during the first few days. During the first 3 or 4 days, the average of the worst quartile group remained at the same high risk level of about 100 units, whereas the averages of the 3 better quartile groups all worsened. Beyond day 4, the averages of all quartile groups experienced a prolonged and moderate improvement until around the end of the second week. By day 28, the gap between the worst quartile group (175) and the best quartile group (350) remained large.

Creatinine, the weakest predictor, had all of its predictive power coming from its change variable. The average creatinine of the worst quartile group experienced an improvement from 350 µmol/L on day 1 to 225 µmol/L on day 6. However, with almost no further improvement, the gap between the worst quartile group (220 µmol/L) and the best quartile group (50 µmol/L) was still large on day 28. In summary, the limited improvement in creatinine made a small contribution to the overall reduction in mortality risk.

**Supplemental Text 9: Construction and Use of Mortality Risk Profiles for Individual Septic Patients**

To identify the main TVBIs that contribute to mortality risk and to determine how improvements in TVBIs can reduce the mortality risk of a patient in question, we describe how a *mortality risk profile* is constructed and how it can be used. As a basis for constructing a mortality risk profile, we first selected among the survivors the patients in the best 10% of the predicted probability of dying to serve as the *benchmark* for comparison.

Based on the estimated coefficients of the CLOGLOG model, the construction of the mortality risk profile is demonstrated for a 66-year old male patient who remained alive on day 28 (Patient A). The primary data of Patient A and the benchmark for constructing the mortality risk profile are shown in Supplemental Table 6. In Supplemental Table 7, these data are then used to generate the values of all explanatory variables for both Patient A (in the 2nd numeric column) and the benchmark (in the 4th numeric column). For example, the value of the explanatory variable “Log(ProteinC\_change\_factor)” for Patient A is generated from his day1 and current values of protein C (33 and 47) as log(47/33)=0.35, where log is the natural log function. The remaining computations in the table are identical to those used in Supplemental Table 4.

The elements in the mortality risk profile (last column of Supplemental Table 9) are the predictive powers of the explanatory variables. For example, the predictive power of the day 1 variable of cfDNA is 0.34, which is this variable’s contribution to [(the log of hazard of Patient A) – (the log of hazard of the benchmark)]. The greater is the predictive power, the greater is this variable’s ability to account for the predicted overall mortality gap between Patient A and the benchmark.

The bottom 5 elements of the last column of Supplemental Table 7 are alternative measures of the overall mortality gap between Patient A and the benchmark. The overall difference in log of hazard (1.849) is the sum of the predictive powers of all explanatory variables. The HR representing the mortality gap is 6.4. The predicted probability of dying in 28 days (P28) is 13.3% for Patient A and 2.2% for the benchmark, so that the gap in P28 was 11.1%.

The elements in the last column of Supplemental Table 7 are used to create the top graph in Figure 2 in the paper. Patient A’s mortality risk was mainly associated with his relatively poor values of the day 1 variables of 5 TVBIs. For cfDNA, GCS and lactate, Patient A actually experienced greater improvements than did the benchmark. In other words, if Patient A had not experienced these greater improvements, the mortality gap in P28 would have been greater than 11.1%.

By summing up the predictive powers of the day 1 and change variables of each TVBI, we obtained the combined predictive power of each TVBI shown in the middle graph of Figure 2. We found from this graph that Patient A’s improvements in GCS and lactate were not large enough to fully compensate for the initial disadvantage, although the improvement in his cfDNA was able to do so. For ease of communication, we applied exponential transformation to each bar in the middle graph to make it into a hazard ratio (HR). For ease of visualization, we plotted (HR – 1) for each TVBI in the bottom graph of Figure 2. The HRs used to create this graph are *multiplicative components* of the overall HR of 6.4. Among these components, the HR for protein C (2.1) was the greatest.

Using the Excel version of Supplemental Table 7, we found that an increase of Patient A’s protein C (47) to the level of the benchmark (131) is associated with a decrease of his P28 from 13.3% to 7.2%, and that further increases of his platelets (111) and GCS (11) to the levels of the benchmark (309 and 14, respectively) are associated with a further reduction of his P28 to 3.3%.

To demonstrate how the knowledge of the dynamic nature of the benchmark and the primary data of a patient could facilitate use of the mortality risk profile, we constructed the profile of another male patient (Patient B) who was discharged on day 12. The primary data of Patient B and the benchmark are shown in Supplemental Table 8, the computations are presented in Supplemental Table 9, and his mortality risk profile is shown in Supplemental Figure 4.

Partly as a consequence of having an advanced age of 79 and the precondition of chronic lung disease, his P28 was quite high (50%). The finding that among the 6 TVBIs, lactate had the highest HR of 2.1, which suggested that a lowering his lactate level would most likely be accompanied by a reduction in his mortality risk. However, a closer examination of Supplemental Tables 8 and 9 revealed that Patient B started with a very low level of lactate (1.3 mmol/L) and maintained the same level on the day of discharge, whereas the benchmark started with a rather high lactate level (4.5) and experienced a large decrease (-3.2) to a low level (1.3) on the last day. Thus, the impressive improvement of the benchmark’s lactate could not be replicated by Patient A, because his lactate level was already very low.

In Supplemental Figure 3, cfDNA had the next highest HR (1.8). We found from Supplemental Table 10 that the level of cfDNA on the last day was much higher for Patient B (7.2 ug/mL) than for the benchmark (4.0 µg/mL). Using the Excel version of Supplemental Table 12, we found that a reduction of Patient B’s cfDNA to 4.0 µg/mL and 2.0 µg/mL is associated with a reduction of his P28 from 50% to 32% and 23%, respectively. Together with a reduction of cfDNA to 2.0, an increase of GCS from the current level of 12 to 15 is associated with a further reduction of Patient B’s P28 to 15%.

Ideally, the mortality risk profiles of a patient should be created on day 1 and every two days in week 1, and then weekly, because rapid changes tend to occur in week 1. If a patient is admitted to the ICU with sepsis, our model should be directly applicable.

**Supplemental Text 10: Validation**

We used two approaches to validate our CLOGLOG model. First, we used the estimated coefficients of the model in Panel 1 of Table 2 to predict the mortality outcomes of the ICU patients from the 9 Canadian hospitals who were originally non-septic but later became septic in the ICU. Second, we split up the 6,724 records in our original input data randomly into 10 subsets and then applied a 10 fold cross-validation.

In the first approach, we recruited 33 such patients among whom 28 had non-missing values for all the explanatory variables and hence were used to form the validation group. For four of the six TVBIs, one or more patients had missing values after imputation: 1 patient for platelets, 1 patient for creatinine, 1 patient for GCS, and 4 patients for lactate. The null hypothesis that the pattern of missing values across the 4 TVBIs with missing values was random could not be rejected by Little’s test (p=0.191). The removal of patients with missing values caused the proportion of non-survivors to change from 15% to 18%. The baseline data of the 28 patients are shown in Supplement Table 10.

Before conducting the validation, we considered the issue about what should be the definition of day 1: (1) the day of admission or (2) the day of becoming septic. We decided to conduct two validations: (V1) with the admission date as day 1 and (V2) with the day of becoming septic as day 1. For V2, we extended the censored date for any patient who was still alive on day 28 until he/she died, discharged, or remained alive on day 28 since becoming septic. Since there were 5 non-survivors in V1 and 6 non-survivors in V2, the proportion of non-survivors was 18% in V1 and 21% in V2.

For each patient in the validation group, we first found the values of all explanatory variables as of her/his last day. We then computed her/his predicted hazard of dying by inputting these values into our estimated CLOGLOG model. The predicted hazards of all 28 patients were then translated into the probability of dying in 28 days (P28) and the predicted daily probabilities of dying (P1) for conducting an ROC analysis.

For observations in V1, the analysis revealed that AUC\_P28 = 0.939 (95% CI, 0.845-1.000) and AUC\_P1 = 0.901 (95% CI, 0.838-0.963), which are somewhat higher than the correspinding values for the derivation group of 356 patients: AUC\_P28 = 0.903 (95% CI, 0.864 - 0.941) and AUC\_P1 = 0.865 (95% CI, 0.826 - 0.903).

For observations in V2, we found that AUC\_P28 = 0.886 (95% CI, 0.746 – 1.000) and AUC\_P1 = 0.863 (95% CI, 0.748 – 0.979) , which are somewhat lower than the corresponding values of the derivation group. Thus, no matter how day 1 was defined, the validation group was about as well predicted as the derivation group.

Since the validation group had a much smaller sample size than did the derivation group, the 95% confidence intervals of the values of AUC\_P28 and AUC\_P1 for V1 and V2 were much wider than those for the derivation group. Despite the smaller sample size, the validation results were statistically supportive, because the lower limits of all 95%CIs were all much higher than 0.5. The ROC curves for the V1 and V2 validations in terms of P28 are compared with that of the derivation group in Supplemental Figure 4.

The validation group also helped show that our best CLOGLOG model is stronger in predictive power than the CLOGLOG with SOFA representing the effects of TVBIs, which is the strongest alternative model in Panel 2 of Suppemental Table 12. For the patients in V1, this alternative model yielded AUC\_P28 = 0.835 (95%CI: 0.578-1.000) and AUC\_P1 = 0.735 (95%CI: 0.480-0.990). Since both 0.835 and 0.735 are less than the lower limits of the corresponding 95%CIs for our model, we have validated that our model is significantly stronger in predictive power than the CLOGLOG model with SOFA.

In the second approach, the 10 fold cross-validation worked in the following way. First, we used the SurveySelect procedure of SAS to randomly assign the 6,724 records of our original input data into 10 subsets of approximately equal sample size. Second, we set aside the first subset as the first Test set and let the remaining 9 subsets be the first Training set. Third, we used the first Training set to estimate the unknown coefficients of our CLOGLOG model and to generate a value for the goodness-of-fit indicator: AUC\_P1(Training). Fourth, the estimated coefficients were then used to predict the mortality outcomes in the first Test set, yielding a value for the indicator of the predictive power: AUC\_P1(Test). Fifth, steps 2 to 4 were repeated 9 times. In each repetition, we let the Test set go from the second subset to the 10th subset of the data, while using the corresponding remaining 9 subsets as the Training set. In the 9th repetition, the 10th subset became the 10th Test set and the first 9 subsets became the 10th Training set.

The findings of the 10 fold cross-validation are summarized in Supplemental Table 11. Keeping in mind that the AUC\_P1 of our best CLOGLOG model is 0.865 (95% CI, 0.826 - 0.903), we found that the cross-validation revealed that our best model is quite robust. AUC\_P1(Training) turned out to have a mean of 0.865 and a standard deviation of 0.008, while AUC\_P1(Test) turned out to have a mean of 0.854 and a standard deviation of 0.073.

**Supplemental Text 11: Assessment of the Predictive Powers of MODS, SOFA and APACHE II**

For both MODS and SOFA, we created the day 1 and change variables and then assessed the predictive power of each of them in the CLOGLOG model that contained the same set of contextual variables as in our own model. For APACHE II, we only created the day 1 variable because APACHE II was designed to measure disease severity within 24 hours of ICU admission. As shown in Supplemental Table 12, the values of AUC\_P28 are 0.802 (95% CI, 0.746 - 0.858) for MODS, 0.862 (95% CI, 0.817 – 0.907) for SOFA, and 0.774 (CI: 0.723 - 0.826) for APACHE II. All three were lower than the AUC\_P28 achieved by our model: 0.903 (95% CI, 0.864 - 0.941). Although both MODS and in SOFA were constructed from 6 TVBIs, SOFA performed markedly better than did MODS. The better performance of SOFA over MODS probably reflects the fact that one of the TVBIs used in constructing SOFA involved choices of different treatments for hypotension, with worse scores for higher dosages. In other words, the predictive power of SOFA seemed to benefit from physicians’ professional judgements on the relative sickness of patients.

**Supplemental Text 12: Assessment of the Usefulness of Four Additional TVBIs and Other Contextual Variables**

To explore the possibility that adding more TVBIs into our assessment tool can enhance its predictive power, we added 4 additional TVBIs separately into our CLOGLOG model represented by Panel 1 of Table 2 in the paper. Three of the additional TVBIs are the remaining components of the MODS score (bilirubin, PaO2/FiO2 ratio, pressure adjusted heart rate (PAR)). We also examined neutrophil count since neutrophils are a potential source of circulating cfDNA. The day 1 and change variables of each TVBI were entered into the model as a pair because these two variables had a strong negative correlation for each TVBI. Supplemental Table 13 shows that all the variables representing these four TVBIs had p-values greater than 0.20 and hence did not have significant effects on the hazard of dying. Thus, addition of more TVBIs to our tool does not increase its predictive power.

The assessment of the usefulness of additional contextual variables was conducted by inserting each variable separately into the CLOGLOG model shown in Table 2. Also from Supplemental Table 13, we found that among the preconditions, congestive heart failure had a p-value less than 0.05. Despite its small p-value, we did not include it in our model for the following reasons. First, because its predictive power overlapped with that of age, its inclusion inflated the p-value of age from 0.0822 to 0.2604 so that age would be removed from the model. Second, its inclusion had practically no effect on the model’s predictive power, with AUC\_P1 decreasing slightly from 0.865 to 0.862. It is likely that for a larger sample size, congestive heart failure would have a significant effect on the hazard of dying in a multivariate context.

The precondition of liver disease had a p-value somewhat lower than 0.10, suggesting that it might have some effect on the hazard of dying. However, its coefficient was negative. Diabetes and chronic renal insufficiency also had *negative* but non-significant coefficient. It is not clear whether these preconditions had spurious life-saving effects, resulting from the medications for their treatments. Ischemic heart disease, chronic dialysis, and cancer had positive coefficients but their p-values were too large to be considered for the inclusion in the model.

With respect to gender, being female appears to be associated with an elevated hazard of dying.  However, its p-value is 0.0782.  So far, we have not found physiological reasons for female septic patients to have a higher mortality risk than their male counterparts.  We also observed that there were no gender-specific differences in the administration of treatment (e.g. use of mechanical ventilation, vasopressors/inotropes, fluids) or in the length of stay in the ICU.

**Supplemental Text 13: The Addition of the Information on the Sites and Types of Positive Cultures Had Little Effect on the Predictive Power of the CLOGLOG Model**

From the raw data of the 356 septic patients who contributed information to the input data matrix of the CLOGLOG model, we found that some sites and types of positive cultures were associated with relatively high crude death rates. For example, the 53 patients whose *site* of positive cultures was *urinary tract* had a crude death rate of 34%, and the 72 patients whose *type* of positive cultures was *mixed* had a crude death rate of 31%, compared with the overall crude death rate of 24%. To determine if information on the sites and types of positive cultures would enhance the predictive power of the CLOGLOG model, we represented each site or type of positive cultures by a dummy variable and added it into the CLOGLOG model reported in Panel 1 of Table 2 in the paper. In terms of the values of the AUC that was based on the daily probability of dying as the classifier, we found that the addition of each of the dummy variables representing the 8 sites and 6 types of positive cultures had little effect on enhancing the model’s predictive power (Panel 1 of Supplemental Table 14). Paradoxically, the values of the AUC *decreased* as a consequence of adding each of 5 sites of positive cultures (pleural cavity, blood, peritoneal, skin, and “other”) into the CLOGLOG model. This finding reveals that the intuitively appealing AUC is not a completely consistent measure of predictive power. Here, the complete consistency of a measure is defined as the property that adding an explanatory variable into the model can never result in a worse value for the measure.

A completely consistent measure of predictive power for a CLOGLOG or logit model is the *Rho-square*, which is defined as 1 – A/B, where A is the maximum log-likelihood of the model in question, and B is the maximum log-likelihood of the null model, which has the intercept as the only unknown coefficient 5. In panel 2 of Supplemental Table 14, we see that the addition of any of the dummy variables did not result in a decrease in the value of Rho-square. This complete consistency is the same as the complete consistency of the R-square for regression models. An important difference between them is that as demonstrated in Panel 2, a value of about 0.2 for Rho-square can represent a very high predictive power, whereas such a value for R-square usually indicates a low predictive power.

To impose a penalty on adding explanatory variables that either contain mostly random noises or are highly redundant, the Rho-square is modified into the *Adjusted Rho-square*, which is 1 – (A-k-1)/(B-1), where k is the number of explanatory variables (4). This is analogous to the Adjusted R-square for keeping regression models parsimonious. We see from Panel 3 of Supplemental Table 14 that most of the sites and types of positive cultures contributed negatively to the Adjusted Rho-square and hence should not be added to the CLOGLOG model.

**Supplemental Text 14: Comparison of the CLOGLOG Model with Two Versions of the Cox Model**

In our notation, the *Cox proportional hazards (PH) model* for the ith patient accounts for the hazard of dying in the following form:

where is a function of duration (t) but does not vary with patients, and is a column vector containing *time-invariant* explanatory variables relevant to patient i. The unknown coefficients are estimated by maximizing the *partial likelihood*, which is a product of *hazard ratios*, each of which is the hazard of a non-survivor divided by the sum of the hazards of a set of all patients who were exposed to the mortality risk on the day the non-survivor died. With the likelihood being a product of hazard rations, the estimation method wipes out (cancels out) the part of the hazard function from consideration. If several patients died on the same day, the specification of the partial likelihood becomes a challenge, because it is difficult to guess who among these patients should be included in the denominator. Based on different assumptions, several methods have been proposed to deal with these ties. Irrespective of how the technical difficulty is dealt with, the estimation method estimates only the coefficients in , so that there is insufficient information to compute the predicted hazard and hence any probability of dying. To apply the Cox PH model to our data, we let the explanatory variables in be the day 1 variables of the 6 TVBIs as well as age and the two preconditions. The input data matrix contains exactly one record per patient, which is the record on the patient’s LAST day. The estimation was carried out by the PHREG procedure of SAS (version 9.4), with the option of TIES=EXACT (2). The estimated coefficients are shown in Panel 2 of Supplemental Table 15 for comparison with those of the CLOGLOG model (Panel 1). The estimated coefficients of the day 1 variables of GCS and lactate as well as the associated p-values turned out to be highly misleading. Without an estimated intercept, it is impossible to compute the predicted hazard and hence P28 for each patient to conduct ROC analysis. By using the linear-in-coefficient expression as the classifier to perform to ROC analysis, we found that the Cox PH model achieved a modest AUC\_P28 of 0.766 (0.708 - 0.825) in predicting a single mortality outcome for each of the 356 patients in 28 days. Clearly, it is a poor choice for our empirical problem.

We then modified the Cox PH model into the following *Cox time-varying (TV) model*:

where is allowed to have the flexibility of including both time-invariant and time-varying variables. The maximum partial likelihood estimation method continued to wipe out from consideration. In applying this model, we let the variables in include the day 1 and change variables of the TVBIs as well as age and the two preconditions. Letting each patient be represented by one and only one record in the input data file, and using the PHREG procedure, we obtained the estimated coefficients of this model and show them in Panel 3 of Supplemental Table 15. Comparing with the CLOGLOG model, the magnitudes of most coefficients were underestimated, although the AUC\_P28 of 0.916 (0.878 - 0.954) was slightly higher.

For gaining longitudinal insights, we found the CLOGLOG model to be preferable to these two versions of the Cox model for the following reason. The former can be used easily to generate of the predicted hazard for any patient on any day, whereas the latter cannot. The predicted hazards can be easily transformed into easily interpretable probabilities on multiple days so that clinicians can obtain concrete longitudinal images of the mortality risks of individual patients that can change markedly within a few days. Also, the practical usefulness of hazard ratios as measures of the strengths of effects depends on the magnitude of the reference hazard.

Finally, we show that the Cox TV model can be made into the CLOGLOG model by making the following three changes. First, we make an explicit function of duration in the following flexible form:

where and are unknown coefficients to be estimated. Second, we replace the maximum partial likelihood method by the maximum likelihood method for estimation. Third, we expand the input data matrix to include all person-day records.

By letting t and log(t) be part of and letting and be part of , Eq. (7) becomes identical to Eq. (1). In other words, the Cox TV model becomes mathematically identical to the CLOGLOG model. The replacement of the estimation method and the expansion of the input data matrix allow the estimation of the full set of unknown coefficients and the generation of the predicted hazards for all patients on all days. Remember that keeping t and log(t) in the CLOGLOG model helped resolve an inconsistency problem resulting from a selection bias in the data (Supplemental Text 4). In sum, *the CLOGLOG model is* *a versatile version of the Cox model for achieving longitudinal insights*.

**Supplemental Text 15: Obtaining Similar Longitudinal Insights from a Longitudinal Logit Model or a Longitudinal Probit Model**

For researchers who are more comfortable with using logit models, we demonstrate here that similar longitudinal insights can be obtained by using a logit model of the following form:

where is the ith patient’s probability of dying on day t, conditional on surviving to the beginning of day t; is an unknown intercept; is a row vector of unknown coefficients; is a column vector of the explanatory variables that reflect the relevant information on the ith patient up to day t; n is the number of patients in the sample; and is the day when the ith patient died in ICU/hospital, was discharged, or was censored (day 28). Due to its usefulness for gaining longitudinal insights, we offer the name *Longitudinal Logit Model* (L-Logit Model, for short). The unknown coefficients are estimated by maximizing the log-likelihood function shown in Eq. (3). The input data matrix is identical to that of the corresponding CLOGLOG model.

Using the SAS logistic procedure with “LINK=LOGIT”, we obtained the estimated coefficients shown in Supplemental Table 16. With AUC\_P28= 0.904 (95% CI, 0.865 - 0.942), this model’s predictive power was about the same as that of the corresponding CLOGLOG model.

To demonstrate that the L-Logit model could be used to reveal similar insights as those obtained via the CLOGLOG model, we used it to create the mortality risk profile of Patient A who was used for demonstration in Supplemental Text 9. Analogous to the predicted log of hazard in the CLOGLOG model, the predicted

logit in the L-Logit model can be decomposed into additive terms in the following way:

where is the ith patient’s *predicted logit of dying* on day t, is the ith patient’s *predicted probability of dying* on day t, and is the additive contribution of the kth explanatory variable to the predicted logit.

With the estimated coefficients (in the first numeric column of Supplemental Table 16), we computed the additive contributions to the predicted logits of (1) Patient A and (2) the benchmark (in the 3rd and 5th columns of Supplemental Table 16). The values of the difference in the contributions to the predicted logits between (1) and (2) were then computed (in the last column of Supplemental Table 16). These values are interpreted as the *predictive powers* of the explanatory variables in accounting for the patient’s mortality risk relative to the benchmark that represented the mean of the best decile of the survivors. The predictive powers attributable to the day 1 and change variables of each of the 6 TVBIs were plotted in Panel 1 of Supplemental Figure 5. The combined predictive power of the day 1 and change variables of each TVBI was then plotted in Panel 2 of Supplemental Figure 5. We then translated predictive powers (the differences in logit) to the corresponding *odds ratios* by exponentiation and then plotted the values of (odds ratio – 1) in Panel 3 of Supplemental Figure 5.

It should not be surprising that the patterns of the bars in the 3 panels of Supplemental Figure 5 are very similar to those in Figure 3 in the paper, because *logit* and *odds ratio* in the L-Logit model are analogous to *log of hazard* and *hazard ratio* in the CLOGLOG model. We expect that by realizing that the L-Logit model is actually analogous to the CLOGLOG model, many users who are experienced in using logit models would adopt the L-Logit model.

This demonstration helps reveal better that the novelty of our methodology lies in the *combination* of the following features. First, the creation of the input data matrix with detailed longitudinal depth. Second, the formulation of a longitudinal model in which each TVBI is represented by a *day 1 variable* and a *change variable*. The use of these two variables helped reveal a serious blind spot of previous predictive attempts that ignored the possibility of marked and rapid improvements among sicker patients. Third, the use of a flexible time function that helped prevent the balance of various selection biases from distorting the coefficients of the day 1 and change variables. Fourth, the decomposition of the overall predictive power into the additive terms of the linear-in-coefficients expression in the model so that the importance of different explanatory variables can be compared. Fifth, the use of the 10% of the survivors with the lowest predicted risks of dying as the *benchmark* for constructing a *mortality risk profile* for each patient in question.

Finally, by realizing that *probit* and *risk ratio* (i.e. ratio of two predicted probabilities) in a probit model are analogous to *logit* and *odds ratio* in a logit model, researchers who are experienced in using probit models could use our demonstration to formulate a *longitudinal probit model* for gaining similar longitudinal insights.

**SUPPLEMENTAL TABLES**

**Supplemental Table 1.** Summary of septic patient recruitment from 9 participating centres

|  |  |  |  |
| --- | --- | --- | --- |
| **Academic ICU** | **Survivors** | **Non-survivors** | **Mortality rate** |
| Hamilton General Hospital, Hamilton, ON | 81 | 24 | 22.8% |
| St. Joseph’s Healthcare, Hamilton, ON | 52 | 18 | 25.7% |
| St. Michael’s Hospital, Toronto, ON | 49 | 14 | 22.2% |
| Ottawa Civic Hospital, Ottawa, ON | 15 | 8 | 34.8% |
| Ottawa General Hospital, Ottawa, ON | 42 | 18 | 30.0% |
| Laval Hospital, Quebec City, QC | 8 | 4 | 33.3% |
| St. Paul’s Hospital, Vancouver, BC | 35 | 4 | 10.3% |
| Calgary Foothills Hospital, Calgary, AB | 10 | 5 | 33.3% |
| London Health Sciences Centre, London, ON | 4 | 1 | 20.0% |



























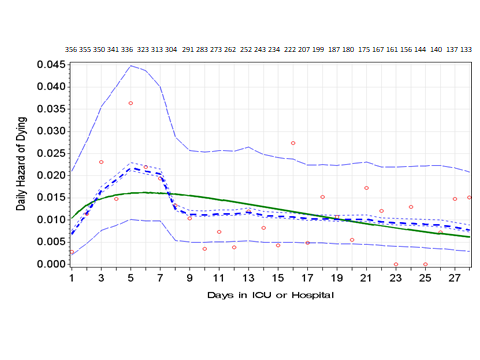






**SUPPLEMETAL FIGURES**

**Supplemental Figure 1.** The exponential form (blue line) and the power form (red line) for the specification of the dependence of the hazard on the current variable of platelet count in the CLOGLOG model. The power form has a greater curvature and a flatter tail.



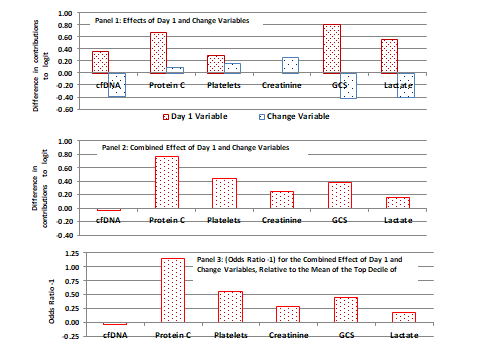
**Supplemental Figure 2.** Temporal Patterns of the Observed and Predicted Daily Hazards of Dying (DHD). Circles represent the observed daily hazards of dying. Green continuous curve traces the DHD predicted by the null model, which contains duration and log(duration) as the only two explanatory variables. Thick blue broken line traces the DHD predicted by the best estimation result of the best CLOGLOG model. For this model, two 95% confidence belts are shown by thin blue broken lines: (1) for the prediction of the daily hazards of individual patients (lines with long segments), and (2) for the prediction of the daily *mean* hazards (lines with short segments). The sizes of the daily at-risk populations are shown across the top of the graph.



**Supplemental Figure 3.** The mortality risk profile of Patient B that highlights the relative contribution of each Time-varying Biological Indicator (TVBI) to the risk of dying.



**Supplemental Figure 4.** The ROC curves for the derivation and validation groups using the probability of dying in 28 days as the classifier. Black curve: derivation group, N=356, AUC= 0.903 (95% CI, 0.864 - 0.941). Blue curve: validation group, N=28, without redefining day 1, AUC= 0.939 (95% CI, 0.849 – 1.000). Red curve: validation group, N=28, with day 1 redefined, AUC= 0.886 (95% CI, 0.746 – 1.000).



Supplemental Figure 5. The mortality risk profile of Patient A that highlights the relative contribution of each TVBI to the risk of dying: generated by *Longitudinal Logit Model.*

Reference List

1. Bernard GR, Vincent JL, Laterre PF et al: Efficacy and safety of recombinant human activated protein C for severe sepsis. *N Engl J Med* 2001; 344:699-709

2. Allison PD: *Survival Analysis Using SAS: A Practical Guide,* *Second Edition*. *SAS Institute Inc* 2010158-172

3. Hanley JA and McNeil BJ: The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; 143:29-36

4. Ben-Akiva M and Lerman SR: Discrete choice analysis: Theory and application to travel demand. *Cambridge, MA: The MIT Press* 198590-91