Journal of Privacy and Confidentiality (2012) **4**, Number 1, 127–154

Confidentialising Survival Analysis Output in a Remote Data Access System

Christine O'Keefe*, Ross Stewart Sparks†, Damien McAullay‡, Bronwyn Loong§

1 Introduction

This paper is concerned with the challenge of balancing the competing objectives of allowing statistical analysis of confidential or private data while maintaining standards of privacy and confidentiality. Such standards can include those imposed by relevant privacy legislation and regulation, as well as assurances provided by data custodians to data contributors.

A high level discussion of the problem of enabling the use of sensitive data while protecting privacy and confidentiality typically introduces two broad categories of method, which are often used in combination. The first is *restricted access,* where access is only provided to approved individuals for approved analyses, possibly at a restricted data centre, and possibly with further measures such as restrictions on the types of analyses which can be conducted and restrictions on the types of outputs which can be taken out of the room. The second is *restricted or altered data,* where less than the full dataset is published or the data are altered in some way before publication. Restricting data commonly involves removing identifying attributes (*de-identification*) or other sensitive attributes or observations, aggregating geographic classifications, or aggregating small groups of data. Altering data is commonly carried out with a *statistical disclosure control* method such as rounding, swapping or deleting values, adding random noise to data, or releasing synthetic data designed to be similar to the original data (see Domingo-Ferrer and Torra, 2004, Doyle et al., 2001, Willenborg and de Waal, 2001). A more detailed discussion of the categories of methods is provided in the introduction to O'Keefe and Good (2009), see also Reiter (2004), O'Keefe (2008).

1.1 Remote analysis systems

A *remote analysis system* is designed to deliver useful results of user-specified statistical analyses with acceptably low risk of a breach of privacy and confidentiality. The remote analysis approach differs from de-identification and statistical disclosure control approaches in that datasets are not provided to the user for analysis. Instead, the user accesses an interface to submit statistical queries to be carried out on the original or confidentialised dataset and receives traditional or confidentialised results. The query

*Science Leader for Privacy and Confidentiality, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia, mailto:Christine.O'Keefe@csiro.au

†CSIRO Mathematics, Informatics and Statistics, Australia, mailto:Ross.Sparks@csiro.au

‡CSIRO Mathematics, Informatics and Statistics, Australia, mailto:Damien.McAullay@csiro.au

§Harvard University, Cambridge, MA, USA, mailto:bloong@fas.harvard.edu

could be submitted either as a user-written piece of code or through making selections on a menu-driven interface.

For examples of systems in use in national statistical agencies see Luxembourg Income Study (n.d.), Australian Bureau of Statistics (n.d), O'Keefe (2008), Rowland (2003). Despite the technical challenges in addressing, for example, data quality issues, missing data, outliers, selection bias testing, and assumption checking (Sparks et al., 2005), it seems to be generally agreed upon that remote analysis systems will play an important role in the future of data dissemination (Reiter, 2004).

While remote analysis systems are designed to reduce disclosure risk, they are not completely free from the risk of disclosure, especially in the face of multiple, interacting queries (Gomatam et al., 2005, Reiter, 2003, Reiter and Kohnen, 2005, Reznek, 2006, Sparks et al., 2005, 2008).

1.2 Role of remote analysis as an approach to data confidentialisation

It is unlikely in the foreseeable future that remote analysis systems will completely replace traditional statistical analysis by analysts with full access to the data. This is largely because remote servers significantly reduce flexibility in analysis and conceal details about the data which can be important in designing and carrying out statistical analysis.

However, in some situations an analyst may need to choose between:

1. Navigating a lengthy and sometimes complex application and ethical review process to obtain confidentialised data. Confidentialisation can include removal of sensitive records and data item fields as well as statistical disclosure control procedures.
2. Using a remote analysis server under a lightweight "low risk" application and ethical review process to analyse the raw, unconfidentialised data, but with restricted information present in the system outputs.

It is unclear to date which option enables the analyst to have greater confidence in answering questions of interest, and this paper is a contribution to exploring this open question.

Even if remote analysis servers are not the preferred mode of data access when used alone, it is possible that remote analysis systems may be useful as preparation for traditional statistical analysis in some situations, including:

- Conducting an initial exploration of data under a lightweight "low risk" ethical review process, in order to determine whether a full ethics application for full access to the data would be worthwhile. This is important because full ethics processes can often be quite lengthy.

- Conducting preliminary investigations and obtaining preliminary results, such as assessment of number of cases and statistical power through exploratory data analysis. Funding applications can be more favourably considered if these preliminary results have been obtained.
- Preparation for visiting a secure data laboratory. An analyst could learn as much about the data as possible and formulate some initial analysis approaches without breaching confidentiality. The analyst would then be able to make efficient and effective, informed use of a later session in a secure data laboratory. This is important because of the cost of secure data laboratory access to both the analyst and the administrative organisation.

In all situations, if the remote analysis system user requires more detailed information such as outlier values, event times, and/or and standard errors, then they would need to apply for access to the underlying data.

1.3 Related work

Early proposals for remote access combined a remote server for query restriction with statistical disclosure control on the source data (Duncan and Mukherjee, 1991, Duncan and Pearson, 1991, Keller-McNulty and Unger, 1998, Scouten and Cigrang, 2003). The special case of using a remote analysis system to disseminate marginal sub-tables on a large, high-dimensional contingency table has been investigated in, for example, Dandekar (2004), Karr et al. (2003, 2002).

In early work on remote analysis systems for model fitting, Reiter (2003) noted that users required the ability to check the fit of their models in a manner that did not disclose actual data values to them. In the case of linear regression, Reiter suggested that a remote analysis system should release only synthetic regression diagnostics, i.e., simulated values of residuals and response and explanatory variables. For model fitting involving categorical explanatory variables, in particular logistic and multinomial regressions, the release of grouped diagnostics was proposed by Reiter and Kohnen (2005) as a way to release diagnostics which do not reveal individual data values.

Sparks et al. (2005) proposed a web-based analytical system designed to enable researchers to perform analyses on unconfidentialised datasets behind a firewall and receive confidentialised results. Subsequently, the authors provided details of disclosure risks associated with the results of a single analysis, focusing on exploratory data analysis and model fitting (Sparks et al., 2008). Measures to reduce the described disclosure risks were proposed, which thus reduce the risk of a user reading or inferring any individual record attribute value. Gomatam et al. (2005) describe disclosure risks associated with multiple, interacting queries to model servers, primarily in the context of regression servers, and propose quantifiable measures of risk and data utility. More recent work includes Bleninger et al. (2010), Lucero and Zayatz (2010).

The generality of the treatment in Sparks et al. (2008) does not make it easy to see the range of disclosure risk reduction measures proposed for particular types of analysis.

To address this gap, O'Keefe (n.d.) provided a detailed discussion of the explicit confidentialisation measures in the case of exploratory data analysis, with a comprehensive example comparing confidentialised with unconfidentialised results. O'Keefe and Good (2008, 2009) provided a similar discussion in the case of linear regression, including a side-by-side comparison of the proposed confidentialised residual plots (using parallel boxplots) with plots of synthetic residuals. The current paper addresses the important case of survival analysis.

1.4 In this paper

In this paper we provide explicit confidentialisation measures for survival analysis in a remote analysis system, with examples. The measures are mostly specialisations of the general measures in Sparks et al. (2008).

In Section 2 we discuss survival analysis and confidentiality objectives, and propose measures for reducing disclosure risk in order to achieve these objectives without confidentialising the underlying data. To illustrate the effect of the methods, in Section 2 we give a comprehensive example comparing confidentialised output with traditional output for a range of common survival analyses. An overview of the example was presented at a recent conference, see O'Keefe and Loong (2010), but here we provide additional details and comments. The confidentialised outputs of the survival analyses were produced using the CSIRO *Privacy-Preserving Analytics™ (PPA)* demonstrator software (Sparks et al., 2008) while the traditional output was produced using the R software environment for statistical computing and graphics (R Project for Statistical Computing, n.d.).

We believe that the example demonstrates that the confidentialised output is still useful for survival analysis, provided the user understands the confidentialisation process and its potential impact.

2 Confidentialising survival analysis outputs

In this section we discuss survival analysis, and propose measures for reducing disclosure risk in a remote analysis system without confidentialising the underlying data. To illustrate the effects of the methods, we provide comprehensive examples comparing confidentialised output with traditional output for the three common survival analysis methods:

1. Non-parametric survival models, of which Kaplan-Meier (Kaplan and Meier, 1958) is the most common
2. Semiparametric regression models, of which Cox's proportional hazards regression model (Cox, 1972) is one of the most important
3. Parametric survival models, of which the Weibull distribution is the most common (see Cox and Oates, 1984, Weibull, 1951)

For introductions to survival analysis, see Anderson and Vaeth (1988), Cox and Oates (1984). In the following we will restrict our attention to the context of clinical trials.

In survival analysis, we have a population and we are interested in comparing survival times for different groups or different treatments. Survival data are almost always *censored*, in that the precise survival time is not observed for those individuals surviving at the end of the study period and for individuals who drop out of the study before the end of the study period. Thus, for some individuals, all that can be said about their survival time is that it exceeds some censoring time (determined by the end of the study period or the time of dropping out). Usually random censorship is assumed, so that survival times and censoring times are independent.

The results of a survival analysis are unlikely to lead to identification of an individual if they:

- Do not reveal identifying information (such as name, address, and health care number), and
- Do not reveal exact values of variables, including hospital procedure dates, diagnoses, and comorbidities.

Dates are particularly disclosive because they are unique and can be used in cross-matching with other datasets. Therefore, these two conditions will be our confidentiality objectives.

The second confidentiality objective is quite strong, but we are interested in exploring whether we can still generate useful survival analysis output with strong confidentiality protection. If the two objectives are achieved, then we believe that the associated disclosure risk would be quite low. However, formal measures of disclosure risk in this situation are not yet available. A data custodian could choose other (possibly less strong) confidentiality settings where appropriate.

The confidentiality objectives will be achieved with a combination of three general types of measures:

1. Use of a predetermined level of sampling from the target dataset, depending on the risk associated with the dataset, the analyst, and the actual analysis conducted
2. Implementation of a web-based user interface which restricts the queries which can be made, and
3. Modifications to confidentialise the output of survival analysis queries

The measures implemented for each type of survival analysis are described in Sections 2.1, 2.2, and 2.3.

For the examples, we will use two publicly available datasets for illustrative purposes only. The first is data regarding survival in patients from the North Central Cancer Treatment Group (NCCTG) with advanced lung cancer (Loprinzi et al., 1994), see also

R Project for Statistical Computing (2009). We will use the variables: survival time in days (time), censoring status (status), and sex (sex).

The second dataset is colon cancer data from the Finnish cancer registry. The dataset contains individual-level data for 15,564 patients, representing all patients diagnosed with localized colon carcinoma in Finland from 1975 to 1994 with follow-up to the end of 1995 (Dickman et al., 1999). We will use the (discrete) variables/factors: sex (SEX), clinical stage at diagnosis (STAGE), and vital status at last date of contact (STATUS); and the (continuous) variables age (AGE) and survival time in completed months (SURV_MM).

In Sections 2.1, 2.2, and 2.3 we show confidentialised output of a survival analysis query as well as the traditional, unaltered output for the same analysis of the same data. Note that the diagnostic plots shown in this paper are a selection of all possible diagnostic plots available. Each subsection closes with a discussion of the differences between the confidentialised and traditional outputs.

2.1 Kaplan-Meier Survival Curves

Let T be the survival time of a randomly selected participant from the population. The survival distribution function $S(t) = \Pr(T > t)$, $t \geq 0$ is used to draw inferences about T. Suppose a study has yielded data of the form $t_1, t_2, \ldots, t_n$, where $0 < t_1 \leq t_2 \leq \ldots \leq t_n$ are survival times for the participants who died during the study as well as censored survival times for those participants who either dropped out or were alive at the end of the study. Let d_j denote the number of participants who died at time t_j and let r_j denote the number of participants alive and in the study just before time t_j, and hence at risk of dying at time t_j.

The Kaplan-Meier estimate of the survival distribution function is:

$$\widehat{S}(t) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{r_j}\right).$$

It is common to use Greenwood standard errors (Greenwood, 1926)

$$\widehat{\sigma}\left(\widehat{S}(t)\right) = \widehat{S}(t) \left(\sum_{t_j \leq t} \frac{d_j}{r_j(r_j - d_j)}\right)^{1/2}$$

and confidence intervals based on them.

Typically the Kaplan-Meier estimates are presented as a survival plot of $\widehat{S}(t)$ versus t. Often, the upper and lower confidence interval limits are also presented as plots on the same diagram. Each of the three plots is a step function with a step occurring at each value of $t_j : j = 1, \ldots, n$, when $d_j > 0$. Censored survival times are indicated with a symbol such as a "+" sign drawn on the survival plot. The plots are typically used to determine overall survival time trends or to compare survival times between groups

such as those receiving different treatments. Knowledge of exact censoring event times and death times is not generally needed for these purposes.

Measures for confidentialising the Kaplan-Meier output are:

1. Suppress the symbols on the survival plot indicating study censoring events. These are event dates which could be used to identify individuals when linked to other databases such as surgery rosters or hospital discharge records.
2. Smooth the survival plot and the confidence interval limit plots, for example, with LOESS (Cleveland, 1979, Cleveland and Devlin, 1988), in order to conceal death times. The times reveal dates which could be used to identify individuals when linked to hospital death records.
3. Add a small amount of noise to the end point of the survival plot to conceal the study end date, as this could be used to identify individuals when linked to other databases. If necessary, additional protection could be provided by terminating the survival fitting earlier than the end of the study.

Example

In this section we are interested in exploring whether there is a difference in survival time between the two sexes in the NCCTG lung cancer patients by calculating and plotting Kaplan-Meier survival estimates. The PPA request screen is shown in Figure 1 and the confidentialised and traditional survival curves are shown in Figure 2. The confidence intervals on the traditional plot are suppressed simply for ease of reading the figure.

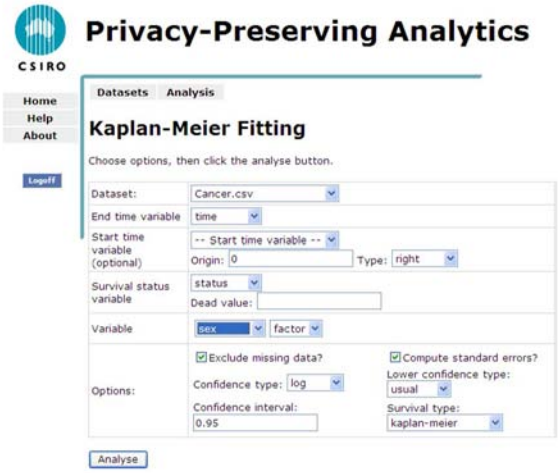
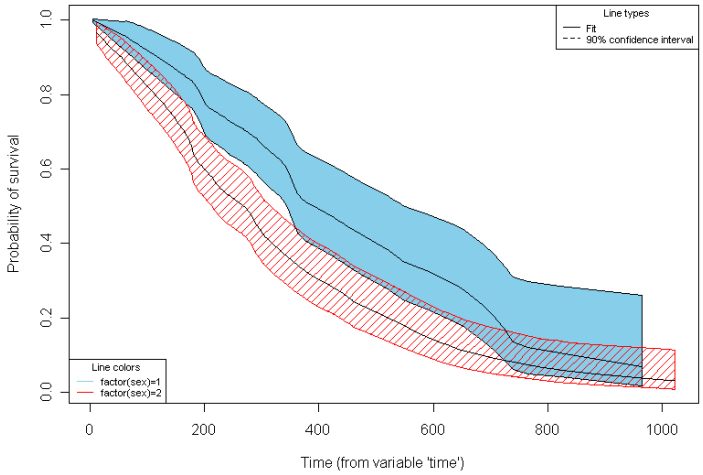
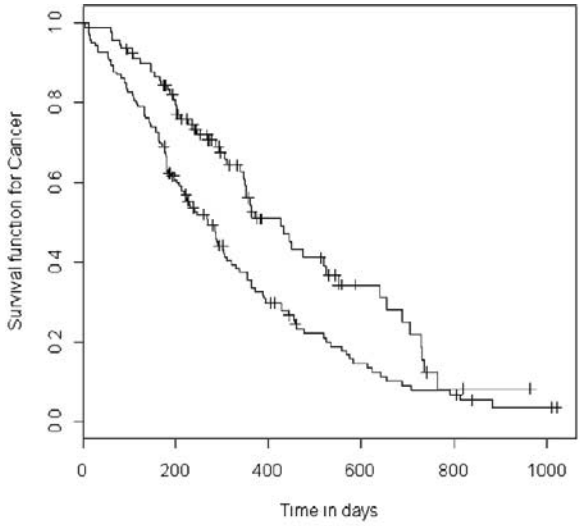


Figure 1: Screenshot of PPA query input interface for Kaplan-Meier Analysis



(a) Confidentialised Output



(b) Traditional Output

Figure 2: Comparison of confidentialised and traditional Kaplan-Meier Survival Analysis output

There are two main differences between the confidentialised and traditional plots, namely suppression of the censoring event times in the confidentialised plot and smoothing of the confidentialised plot. Although the censoring event times do not appear on the confidentialised plot, they are included in the underlying analysis. Similarly, although the death event times are concealed on the confidentialised plot, they are included in the underlying analysis. Therefore, the survival time distribution shown in the confidentialised plot is just a smoothed version of the survival time distribution shown in the traditional plot. The same conclusions regarding the survival time distributions would be made from the confidentialised plot as from the traditional plot.

In the case of Kaplan-Meier Analysis, the confidentialised output would appear to be suitable for observing overall trends and comparing survival time distributions of different population groups.

2.2 Cox proportional hazards regression model

With the notation introduced in Section 2.1, let the probability density function for survival times T be $f(t)$ and the hazard function be

$$h(t) = \Pr(T = t | T \geq t).$$

For discrete data, $h(t) = f(t)/\Pr(T \geq t)$ and for continuous data $h(t) = f(t)/S(t)$.

In Cox's proportional hazards model (Cox, 1972), the hazard function for participant i with covariates $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is assumed to have the form

$$h_i(t) = h_0(t) \exp\left(\sum_{j=1}^{p} \beta_j x_{ij}\right) = h_0(t) \exp\left(\sum_{j=1}^{p} \beta^T x_i\right),$$

where $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ is a vector of unknown regression coefficients reflecting the influence of the covariates x_i on survival, and $h_0(t)$ is an unspecified function of time representing the baseline hazard (corresponding to the situation in which all covariate values are zero). The fitted model produces the estimate $\widehat{\beta}$ of the vector of coefficients.

The interest in a Cox Proportional Hazards model is mainly in the coefficient estimates $\widehat{\beta}$ rather than the baseline hazard $h_0(t)$. The analyst will also be interested in checking the model fit with diagnostic information and plots.

Our confidentiality objectives require that the values of the covariates x_{ij} for each participant i are not revealed. This also means that the hazard function $h_i(t)$ for each participant i is not revealed.

This is achieved with a combination of measures as follows:

Model selection

1. Conduct each analysis on a random sample of 95% of the observations, where the sampling procedure on the set of observation indices requires a random seed.

The seeds are managed to ensure that each analyst will continue to get the same 95% sample for all similar queries. As soon as they nominate a different response variable, they will be given a different 95% sample for all models fitted with the new response. This strategy reduces the disclosure risk for multiple queries by introducing sampling error, but allows analysts to compare and select models using their favourite criterion, such as AIC.

2. Do not allow new variables (such as a linear combination of other variables or reweighted variables) to be included in the model, with the exception that Box-Cox transformations of continuous variables are allowed. This prevents analysts from manipulating the data in order to discover information about the response variable. For example, an analyst who knows that a certain unit is in the database may be able, through transformations, to turn it into an artificially extreme leverage point. This would reveal the outcome variable for that unit from the predicted value of the fitted regression, since leverage points have a strong effect on the estimated regression and often have a small residual, (see Gomatam et al., 2005).
3. Allow a factor to be included in the model only if each level is observed for at least a minimum threshold value of data items, due to the elevated disclosure risk associated with covariate values for small groups of participants. A threshold of 3 is common.
4. Allow only pairwise interactions of factors, and only those pairwise interactions which are observed for at least a minimum threshold value of data items, due to the elevated disclosure risk associated with covariate values for small groups of participants. A threshold of 3 is common.

Model fitting

5. Use robust estimators (Minder and Bednarski, 1996), which reduce the effect of influential points and outliers on the results. The analyst can have confidence in the results, without needing to know the influential observations and outliers.

Output presentation

6. The coefficient estimates are rounded to introduce uncertainty into reconstructions of observed data values obtained by, for example, attempting to solve for elements of the design matrix. The same rounding is applied to each subset model fitted.
7. Do not disclose standard errors or confidence intervals of coefficients to the analyst. Standard errors can be used to reconstruct response values, (see Sparks et al., 2008, 4.1), and so should not be disclosed. Confidence intervals can be used to reconstruct standard errors, and hence response values, since the $(1-\alpha)$ confidence interval for $\widehat{\exp}(\beta_j)$ is just $\exp(\widehat{\beta}_j \pm z_\alpha \sigma(\widehat{\beta}_j))$.
8. Do not provide accurate p-values, since exact p-values can be used to reconstruct standard errors and hence response values. Provide ranges for the p-values, and consider listing the variables in ranked order of significance. If any confidentialised p-value range indicates an incorrect significance level in comparison with the values found in a traditional analysis (as may happen if the true value is very near a threshold), then replace it with the correct range.

9. Replace each diagnostic dot chart (for example, plot of partial residuals for categorical variables) with a confidentialised boxplot, as follows:
   (a) Winsorise the values, say by discarding observations at distance more than 1.5 times the interquartile range from the median.
   (b) If the difference between the median and upper or lower quartile is zero then add a small amount of noise to the quartile, and if the difference between a winsorised extreme value and the adjacent quartile is zero then add a small amount of noise to the winsorised extreme value.
   (c) Round the final values of the five summary statistics on the interval.
10. Replace each diagnostic pairwise scatterplot with confidentialised parallel boxplots as follows:
   (a) Determine which variable will be on the x-axis and which will be on the y-axis.
   (b) Determine the number of box plots to be constructed, by specifying intervals of the x-axis variable so that each interval has frequency of at least a minimum threshold value.
   (c) On each such interval, if the difference between the median and either the lower or upper quartile of the y-values is zero, amalgamate that interval with an adjacent interval and repeat until all intervals have distinct median, lower, and upper quartiles of y-values.
   (d) For each interval, draw a confidentialised box plot on the y-values, as follows:
      i. Winsorise the y-values, say by discarding observations at distance more than 1.5 times the interquartile range from the median.
      ii. If the difference between the median and upper or lower quartile is zero then suppress or amalgamate the interval with an adjacent interval, and
      iii. Round the final values of the five summary statistics on the interval.

   Individual residuals, when put together with the model, allow the user to reconstruct individual survival times.
11. Replace each Q-Q plot or P-P plot with a confidentialised plot, as follows:
   (a) Remove obvious outliers in the variable of interest from the data.
   (b) Fit a robust non-parametric regression line, which reduces the effect of influential points and outliers on the results. The analyst can have confidence in the results, without needing to know the influential observations and outliers.
   (c) Provide only the fitted regression line as output.

These measures are quite strict, consistent with our philosophy of exploring the usefulness of the output under strong confidentiality objectives. A data custodian could choose to implement more or less restrictive measures, based on dataset-specific disclosure risk assessments. Similarly, if the type of risk addressed by Measure 2 above is unlikely, then Measure 2 may not be needed.

Remote analysis systems have the potential to offer extra functionality to a user conducting a Cox Proportional Hazards analysis. For example, in surgery survival

data the name of the surgeon and the hospital are often suppressed as part of the confidentialisation procedure before provision to the analyst. Therefore, the analyst has no information on the potential effect of these covariates. In a remote analysis system, the surgeon and hospital covariates could be included in the analysis, but then all information about the coefficient estimates could be suppressed in the output. In this way, the influence of these covariates would be separated from the influence of the other covariates. An analyst would be able to investigate whether there were significant surgeon or hospital effects, without learning anything about the nature of the difference.

Example

In this section we will investigate the covariates AGE, SEX, and STAGE in a Cox Proportional Hazards Model fitted on the Finnish cancer registry colon cancer data. Figure 3 shows the PPA request screen for the Cox Proportional Hazards analysis, which ensures that queries are restricted as discussed in Section 2.2.

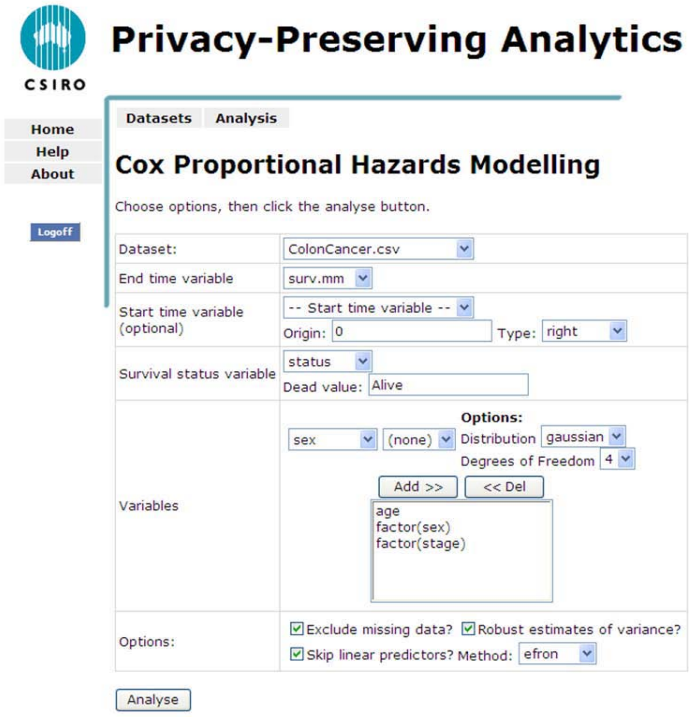


Figure 3: Screenshot of PPA query input interface for Cox Proportional Hazards modelling

Confidentialised and traditional output summary results are shown in Tables 1 and 2, respectively. A selection of corresponding confidentialised and traditional diagnostic plots are shown in Figures 4, 5, and 6.

	coef	**exp(coef)**	**p**
age	0.008	1.008	$p < 0.005$
factor(sex)Male	0.114	1.12	$p < 0.005$
factor(stage)Localised	0.043	1.044	$0.2 < p < 0.5$
factor(stage)Regional	0.255	1.291	$p < 0.005$
factor(stage)Unknown	0.007	1.007	$p > 0.5$

Rsquare= 0.005 (max possible = 0.991)
Likelihood ratio test= 72.2 on 5 df, $p = 3.51e - 14$
Wald test = 72.3 on 5 df, $p = 3.47e - 14$
Score (logrank) test = 72.4 on 5 df, $p = 3.2e - 14$, Robust = 69.2 $p = 1.49e - 13$

Table 1: Confidentialised Cox Proportional Hazards Model Summary Results

	coef	**exp(coef)**	**se(coef)**	**z**	**Pr(>|z)**	
age	0.008158	1.008191	0.001248	6.535	6.35e-11	***
factor(sex)Male	0.115998	1.122994	0.030356	3.821	0.000133	***
factor(stage)Localised	0.045075	1.046106	0.056441	0.799	0.424513	
factor(stage)Regional	0.256935	1.292961	0.066659	3.854	0.000116	***
factor(stage)Unknown	0.013660	1.013754	0.065621	0.208	0.835096	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

	exp(coef)	**exp(-coef)**	**lower 0.95**	**upper 0.95**
age	1.008191	0.991876	1.005728	1.010660
factor(sex)Male	1.122994	0.890477	1.058127	1.191837
factor(stage)Localised	1.046106	0.955926	0.936553	1.168475
factor(stage)Regional	1.292961	0.773419	1.134604	1.473421
factor(stage)Unknown	1.013754	0.986433	0.891405	1.152896

Rsquare= 0.005 (max possible= 0.991)
Likelihood ratio test= 73.25 on 5 df, $p = 2.154e - 14$
Wald test = 73.34 on 5 df, $p = 2.065e - 14$
Score (logrank) test = 73.42 on 5 df, $p = 1.987e - 14$

Table 2: Traditional Cox Proportional Hazards Model Summary Results

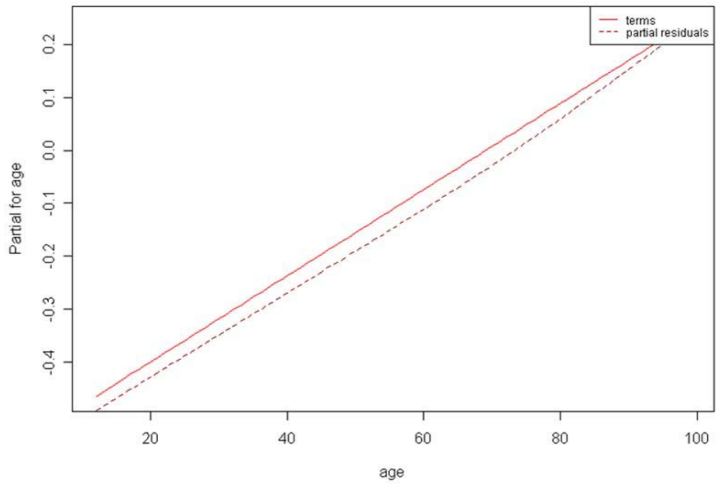
The main differences between the confidentialised and traditional summary results are: suppression of z values, standard errors, and confidence interval widths, as well as reporting of p values in ranges. Numerical differences in the parameter estimates arise from the use of a 95% random sample of the data, the use of robust methods, as well as the rounding of values in the confidentialised case.

In this example, the confidentialised and unconfidentialised outputs indicate the same set of significant coefficients, and approximately the same magnitude of influence in the same direction. The overall model statistics, and their significance, are very similar and should lead the analyst to very similar conclusions about the overall model fit.

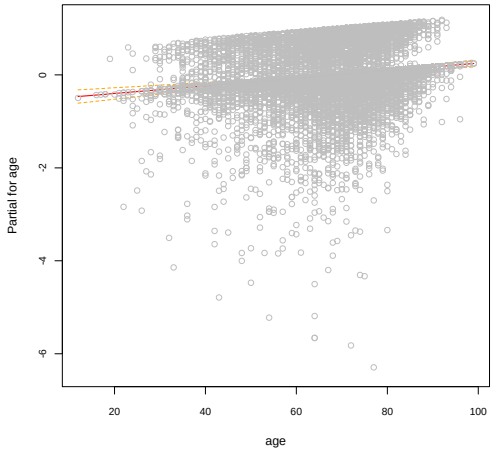
Figures 4, 5, and 6 shows corresponding confidentialised and traditional diagnostic plots, namely the partial residuals for AGE, SEX, and STAGE.

The scale in each confidentialised plot in Figure 6(a) is compressed in comparison with the scale in the traditional plot in Figure 6(b), due to the removal of outliers before plotting. Standardising the scale of the confidentialised plot to match the scale of the traditional plot would indicate to the analyst the presence of outliers, and a poor choice of plot endpoints would reveal approximate values. In most cases knowledge of the presence of outliers is a disclosure risk, which is avoided by not standardising the plot scale. Note that the traditional plots will not be available to the analyst through the remote analysis system, so they will not be able to compare the plots to detect a difference in the scales.

The confidentialised plot of partial residuals for AGE is still suitable to deduce magnitude information as well as observe linear trends in the terms and partial residuals, for the data with outliers removed. The partial residuals for SEX and STAGE are similarly suitable for deducing magnitude and spread information for the data with outliers removed.

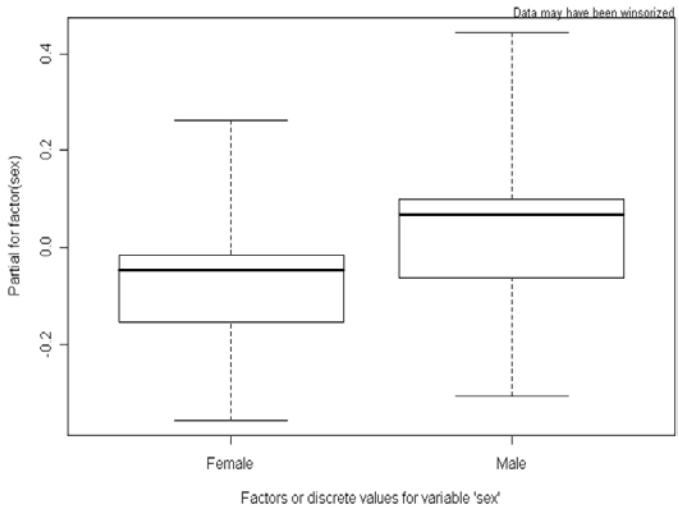


(a) Confidentialised Output

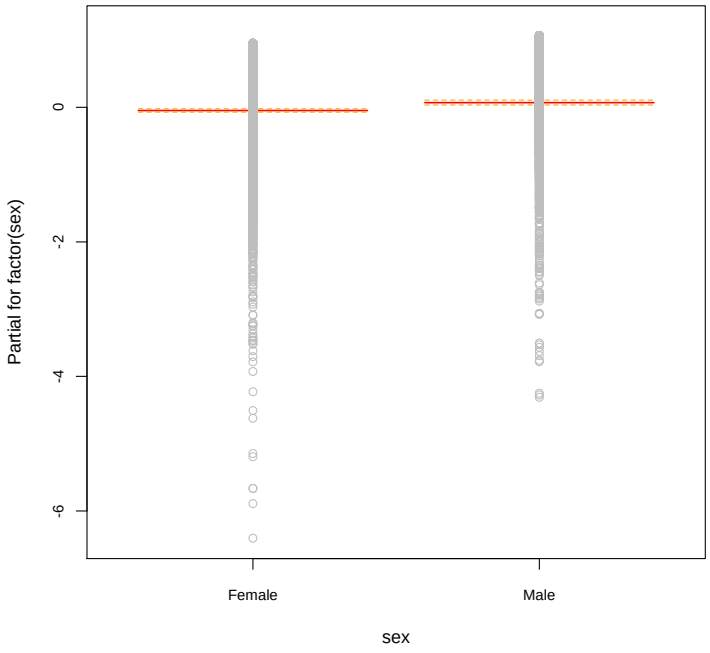


(b) Traditional Output

Figure 4: Comparison of confidentialised and traditional partial residuals for AGE in a Cox Proportional Hazards Model on the Finnish cancer data

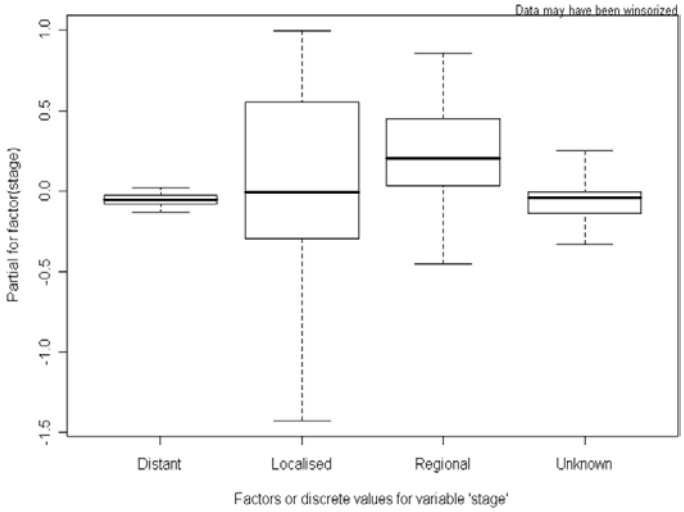


(a) Confidentialised Output

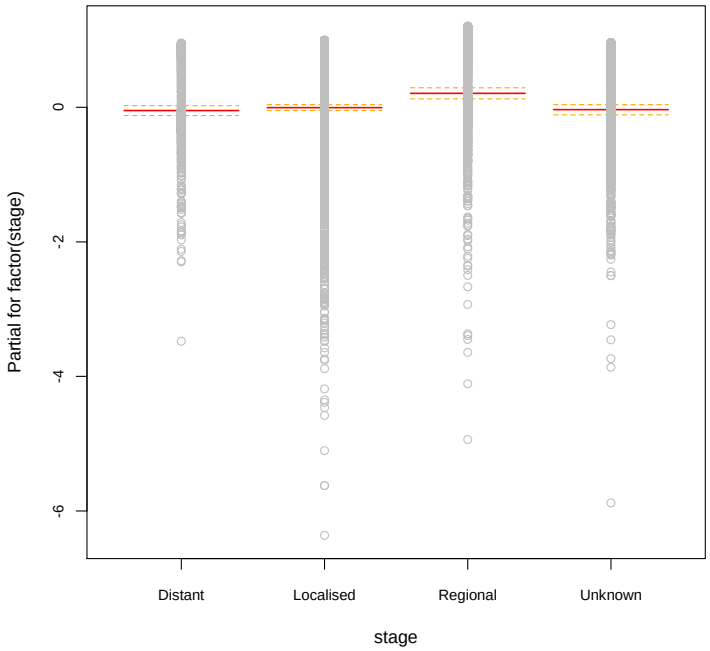


(b) Traditional Output

Figure 5: Comparison of confidentialised and traditional partial residuals for the factor SEX in a Cox Proportional Hazards Model on the Finnish cancer data



(a) Confidentialised Output



(b) Traditional Output

Figure 6: Comparison of confidentialised and traditional partial residuals for the factor STAGE in a Cox Proportional Hazards Model on the Finnish cancer data

2.3 Parametric survival modelling with Weibull distribution

With the same notation as in Sections 2.1 and 2.2, we suppose that the probability density function $f(t)$ of survival time is modelled with a Weibull distribution of the form

$$\begin{aligned} f(t;\alpha,\lambda) &= \left(\frac{\alpha}{\lambda}\right)\left(\frac{t}{\lambda}\right)^{\alpha-1}\left(\exp\left(-\frac{t}{\lambda}\right)^{\alpha}\right) \; for \; t \geq 0 \\ &= 0 \; for \; t < 0, \end{aligned}$$

where $\alpha > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter, (Davison, 2008, see). The hazard function is

$$h(t;\lambda,\alpha) = \frac{\alpha}{\lambda}\left(\frac{t}{\lambda}\right)^{\alpha-1}.$$

For $\lambda = 1$, the hazard function is known as the baseline hazard $h_0(t;\alpha)$. For a participant i with covariates $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ the hazard function is

$$h_i(t;\alpha,\tau(\beta;x)) = h_0\left(t/\tau(\beta;x);\alpha\right)\tau(\beta;x)^{-1},$$

where $\tau(\beta;x) = \exp^{x_i^T\beta}$.

As in Section 2.2, our confidentiality objectives require that the values of the covariates, and hence also the values of the mean survival times, are not revealed. The formula for the mean survival time shows the importance of rounding the values of the coefficient estimates, and checking standard errors for the fitted values, to avoid estimating the survival means too closely. The measures to achieve this are the same as those given in Section 2.2.

Example

In this section, we fit a parametric survival model to the Finnish cancer registry colon cancer data, assuming a Weibull survival distribution. Figure 7 shows the PPA request screen for the Parametric Survival analysis. The menu-driven interface to PPA ensures that queries are restricted as in Section 2.2.

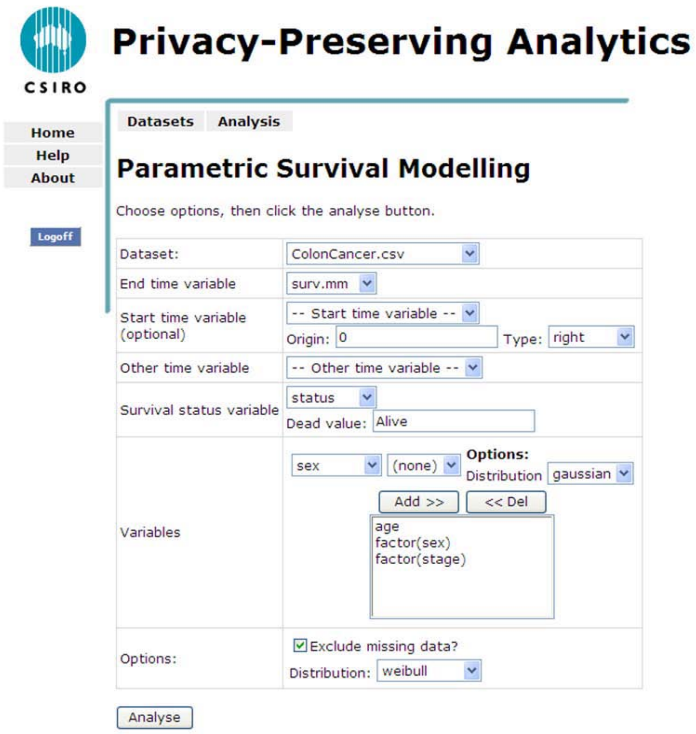


Figure 7: Screenshot of PPA query input interface for a Parametric Survival Analysis

Confidentialised and traditional output summary results are shown in Tables 3 and 4 respectively. A selection of corresponding confidentialised and traditional diagnostic plots are shown in Figures 8, 9, and 10.

	Value	**p**
(Intercept)	5.595	$p < 0.005$
age	-0.033	$p < 0.005$
factor(sex)Male	-0.167	$p < 0.005$
factor(stage)Localised	1.607	$p < 0.005$
factor(stage)Regional	1.414	$p < 0.005$
factor(stage)Unknown	1.071	$p < 0.005$
Log(scale)	-0.962	$p < 0.005$

Scale= 0.382
Weibull distribution
Loglik(model)= -49481.9 Loglik(intercept only)= -26652.8
Chisq= -45658.24 on 5 degrees of freedom, $p = 1$
Number of Newton-Raphson Iterations: 30

Table 3: Confidentialised Parametric Survival Model Summary Results

	Value	**Std.error**	**z**	**p**
(Intercept)	5.6327	0.019150	294.1	0.00e+00
age	-0.0331	0.000255	-129.5	0.00e+00
factor(sex)Male	-0.1675	0.004923	-34.0	1.15e-253
factor(stage)Localised	1.5575	0.007348	212.0	0.00e+00
factor(stage)Regional	1.3664	0.015248	89.6	0.00e+00
factor(stage)Unknown	1.1037	0.008085	136.5	0.00e+00
Log(scale)	-0.9830	0.000000	-Inf	0.00e+00

Scale= 0.374
Weibull distribution
Loglik(model)= -48702.2 Loglik(intercept only)= -26904.1
Chisq= -43596.19 on 5 degrees of freedom, p= 1
Number of Newton-Raphson Iterations: 30

Table 4: Traditional Parametric Survival Model Summary Results

The main differences between the confidentialised and traditional summary results are: suppression of z values and standard errors, as well as reporting of p values in ranges. Numerical differences in the parameter estimates arise from the use of a 95% random sample of the data, the use of robust methods as well as rounding of values in the confidentialised case. In this example, the confidentialised and unconfidentialised outputs indicate the same set of significant coefficients, and approximately the same magnitude of influence in the same direction.

The overall model statistics, and their significance, are very similar and should lead the analyst to very similar conclusions about the overall model fit.

Figures 8, 9, and 10 show confidentialised and traditional residual plots for AGE, SEX and STAGE for the parametric survival model, respectively.

The discussion provided in Section 2.2 regarding the compression of the scale in the confidentialised plot in Figure 10(a) in comparison with the traditional plot in Figure 10(b) is also applicable in this example.

The confidentialised plot of residuals for AGE is still suitable to deduce magnitude information as well as observe the curved trend in the residuals for the data with outliers removed. The residuals for SEX and STAGE are similarly suitable for deducing magnitude and spread information for the data with outliers removed.

In drawing conclusions, the analyst must be aware that outliers have been removed.

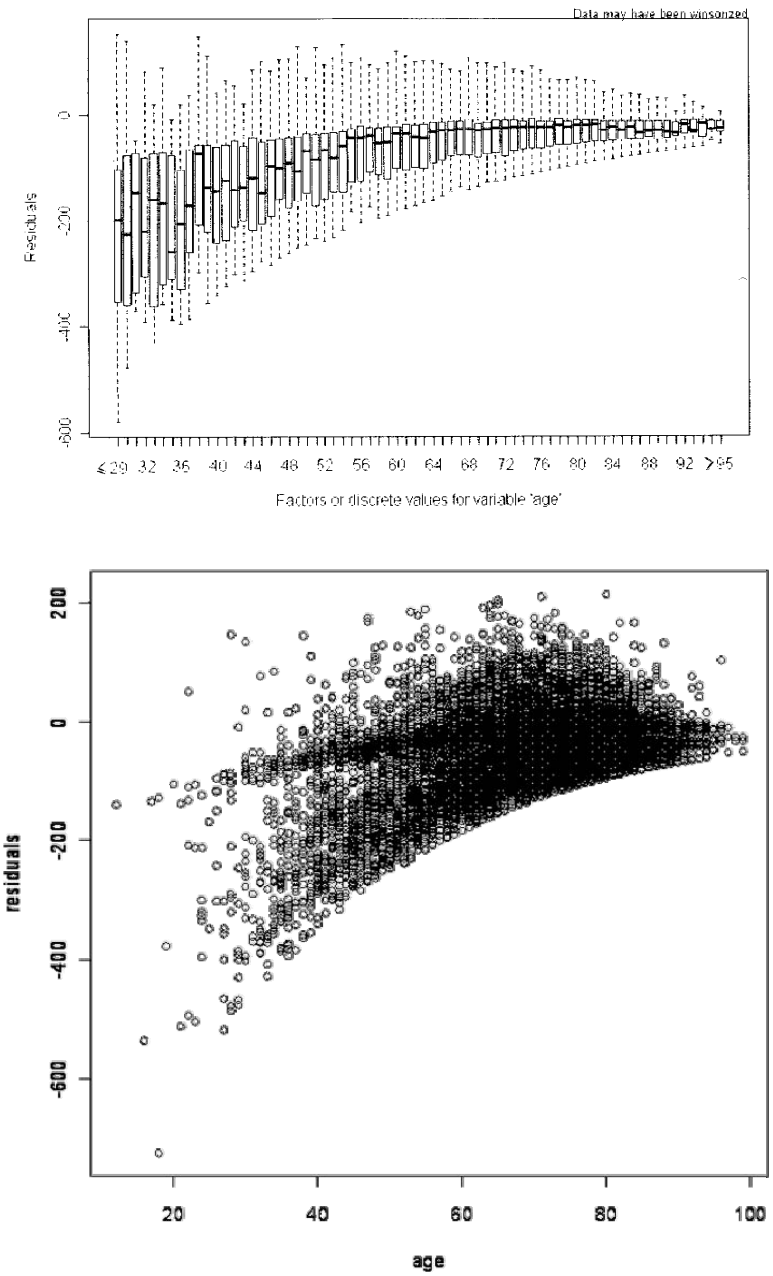


Figure 8: Comparison of confidentialised and traditional residual plots for AGE in the Finnish cancer data, from a Parametric Survival Model

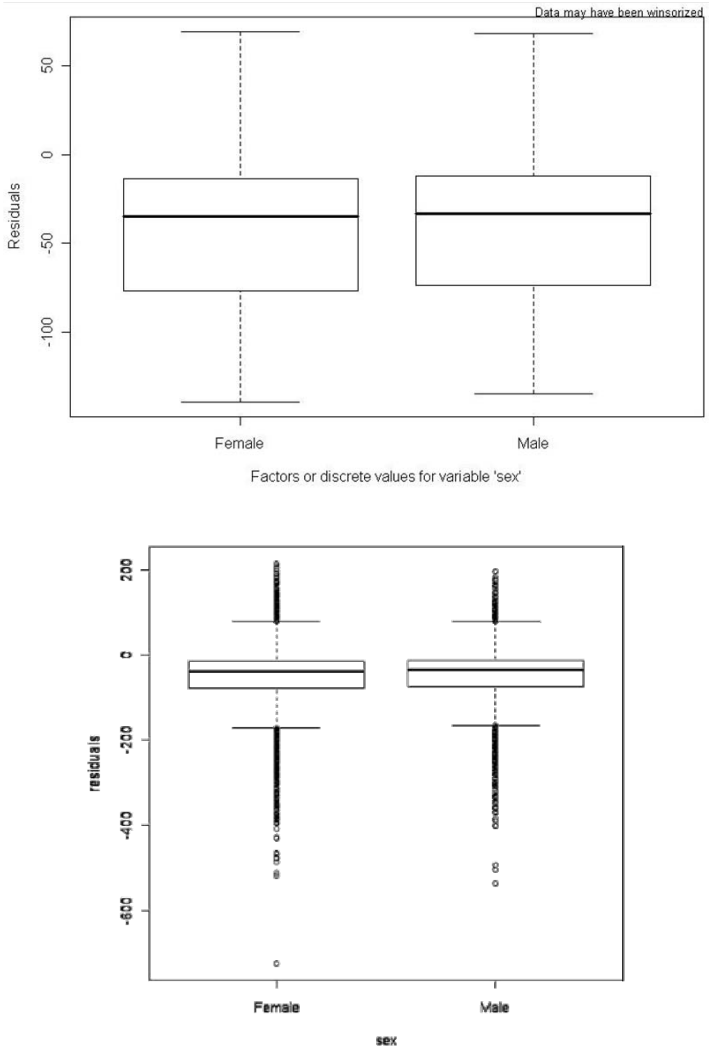
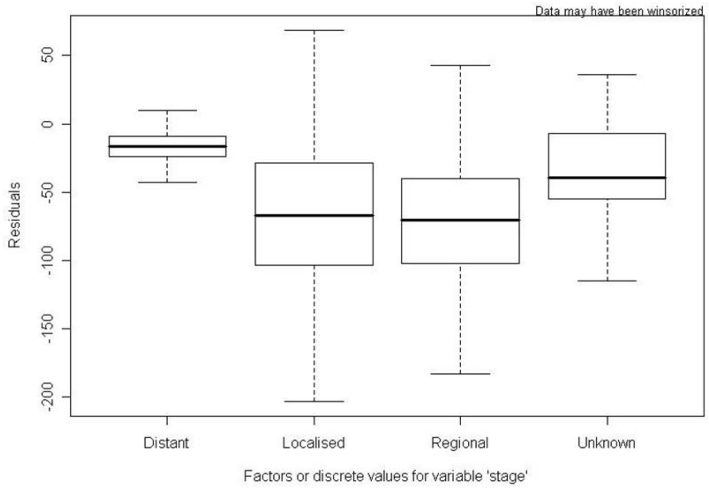
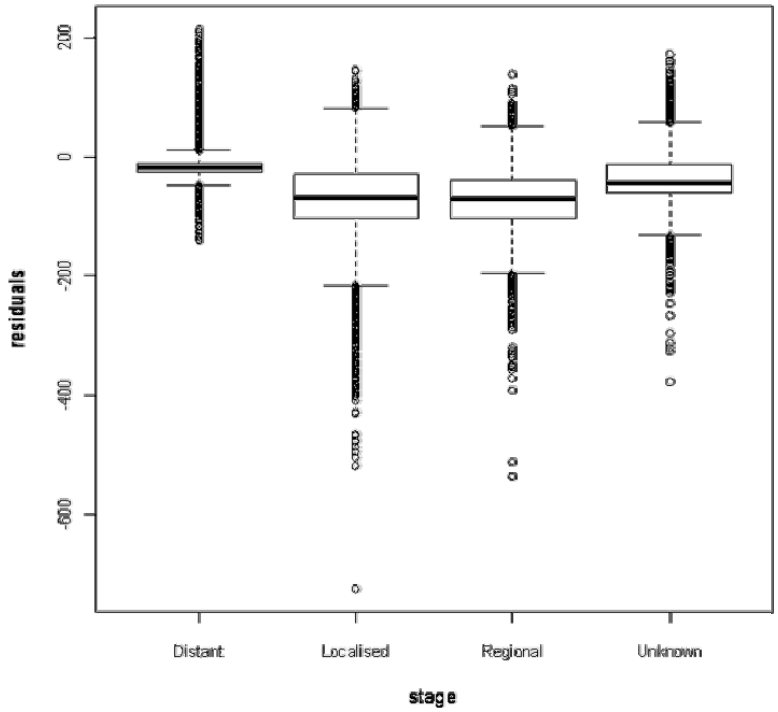


Figure 9: Comparison of confidentialised and traditional residual plots for SEX in the Finnish cancer data, from a Parametric Survival Model



(a) Confidentialised Output



(b) Traditional Output

Figure 10: Comparison of confidentialised and traditional residual plots for STAGE in the Finnish cancer data, from a Parametric Survival Model

3 Discussion and conclusions

In this paper we have described the implementation of a remote analysis system allowing survival analysis on confidential data, including defining confidentiality objectives for the system output, and measures for achieving them. To illustrate the effects of the methods, we provide a comprehensive example comparing confidentialised output with traditional output for a range of common survival analyses.

In the case of Kaplan-Meier Analysis, the confidentialised output would appear to be suitable for observing overall trends and comparing survival curves of different population groups. The analyst should be aware that censoring event times and death times are not shown on the plot.

For both the Cox Proportional Hazards Analysis and the Parametric Survival Analysis, the following are observed:

1. The use of a 95% random sample of the data, the use of robust estimators and rounding of results in the confidentialised output may lead to parameter estimates and overall model statistics which are different from the traditional estimates. However, the examples demonstrated confidentialised and unconfidentialised outputs indicate the same set of significant coefficients, and approximately the same magnitude of influence in the same direction. Furthermore, the overall model statistics, and their significance, are very similar and should lead the analyst to very similar conclusions about the overall model fit.
2. The confidentialised output provides significance only up to a given interval, so explanatory variables with p-values in the same interval cannot be ranked in order of significance.
3. The confidentialised model diagnostic plots are constructed as smoothed curves or parallel boxplots on the residuals—but with outliers removed. This has the effect of compressing the range of the plots. However, the confidentialised output would appear to still be suitable to observe magnitude information, trends and curvature of the data without outliers in order to check model fit.

The issue of suppression of standard error values is particularly problematic for analysts. The confidentialised output only gives general information such as: if the p-value is less than 0.001 then the standard error is less than the value that would correspond to p=0.001. Providing rounded standard errors is disclosive, and determining a protective level of rounding which still provides reliable information is perhaps not possible. This issue provides concrete evidence that remote analysis is unsuitable for particular applications, as discussed generally in Section 1.2.

In summary, we believe that the confidentialised output is still useful for survival analysis, provided the user understands the confidentialisation process and its potential impact. If the analyst is concerned about the impact of the confidentialisation process, or requires more detailed information, they could seek approval for access to the underlying data.

References

Anderson, P. K. and Vaeth, M. (1988). Survival analysis. In S. Kotz and N. L. Johnson, editors, *Encyclopedia of Statistical Sciences*, volume IX, 119–129. New York: John Wiley & Sons.

Australian Bureau of Statistics (n.d). Remote access data laboratory (RADL). `http://www.abs.gov.au`.

Bleninger, P., Drechsler, J., and Ronning, G. (2010). Remote data access and the risk of disclosure from linear regression: An empirical study. In J. Domingo-Ferrer and E. Magkos, editors, *Privacy in Statistical Databases*, vol. 6344 of *LNCS*, 220–233. Springer.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368): 829–836.

Cleveland, W. S. and Devlin, S. J. (1988). Locally-weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403): 596–610.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society B*, 34(2): 187–220.

Cox, D. R. and Oates, D. (1984). *Analysis of Survival Data*. Monographs on Statistics and Applied Probability. Chapman & Hall.

Dandekar, R. A. (2004). Maximum utility-minimum information loss table server design for statistical disclosure control of tabular data. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Databases*, vol. 3050 of *LNCS*, 121–135. Springer.

Davison, A. C. (2008). *Statistical Models*. New York: Cambridge University Press.

Dickman, P. W., Hakulinen, T., Luostarinen, T., Pukkala, E., Sankila, R., Söderman, B., and Teppo, L. (1999). Survival of cancer patients in Finland 1955–1994. *Acta Oncol*, 38(Suppl. 12): 1–103.

Domingo-Ferrer, J. and Torra, V. (eds.) (2004). *Privacy in Statistical Databases*, vol. 3050 of *LNCS*. Springer.

Doyle, P., Lane, J. I., Theeuwes, J. J. M., and Zayatz, L. (eds.) (2001). *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*. Amsterdam: North Holland.

Duncan, G. T. and Mukherjee, S. (1991). Microdata disclosure limitation in statistical databases: Query size and random sample query control. In *Proceedings of the 1991 IEEE Symposium on Security and Privacy*, 278–287.

Duncan, G. T. and Pearson, R. W. (1991). Enhancing access to microdata while protecting confidentiality: Prospects for the future. *Statistical Science*, 6: 219–239.

Gomatam, S., Karr, A. F., Reiter, J. P., and Sanil, A. (2005). Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access systems. *Statistical Science*, 20: 163–177.

Greenwood, M. (1926). The natural duration of cancer. Reports of Public Health and Medical Subjects 33. Technical report, Her Majesty's Stationery Office, London. Pp. 26.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53: 457–481.

Karr, A. F., Dobra, A., and Sanil, A. P. (2003). Table servers protect confidentiality in tabular data releases. *Communications of the ACM*, 46.

Karr, A. F., Lee, J., Sanil, A. P., Hernandez, J., Karimi, S., and Litwin, K. (2002). Web-based systems that disseminate information but project confidentiality. In W. M. McIver and A. K. Elmagarmid, editors, *Advances in Digital Government: Technology, Human Factors and Public Policy*, 181–196. Amsterdam: Kluwer.

Keller-McNulty, S. and Unger, E. A. (1998). A database system prototype for remote access to information based on confidential data. *Journal of Official Statistics*, 14: 347–360.

Loprinzi, C. L., Laurie, J. A., Wieand, H. S., Krook, J. E., Novotny, P. J., Kugler, J. W., Bartel, J., Law, M., Bateman, M., and Klatt, N. E., et al. (1994). Prospective evaluation of prognostic variables from patient-completed questionnaires. *Journal of Clincial Oncology*, 12(3): 601–607. North Carolina Cancer Treatment Group.

Lucero, J. and Zayatz, L. (2010). The microdata analysis system at the U.S. Census Bureau. In J. Domingo-Ferrer and E. Magkos, editors, *Privacy in Statistical Databases*, vol. 6344 of *LNCS*, 234–248. Springer.

Luxembourg Income Study (n.d.). `http://www.lisproject.org`.

Minder, C. E. and Bednarski, T. (1996). A robust method for proportional hazards regression. *Statistics in Medicine*, 15(10): 1033–1047.

O'Keefe, C. M. (2008). Privacy and the use of health data—reducing disclosure risk. *electronic Journal of Health Informatics*, 3(1): e5.

— (n.d.). Confidentialising exploratory data analysis output in remote analysis. Preprint.

O'Keefe, C. M. and Good, N. M. (2008). A remote analysis system—what does regression output look like? In J. Domingo-Ferrer and Y. Saygin, editors, *Privacy in Statistical Databases*, vol. 5262 of *LNCS*, 270–283. Springer.

— (2009). Regression output from a remote analysis system. *Data & Knowledge Engineering*, 68: 1175–1186.

O'Keefe, C. M. and Loong, B. (2010). Remote access in action: Comparison of confidentialised and traditional survial analysis outputs. In *Privacy in Statistical Databases* Conference proceedings, PSD2010, 22–24. Corfu, Greece. `http://www.csiro.au/resources/pf2cc.html`.

R Project for Statistical Computing (2009). NCCTG Lung Cancer Data. `http://cran.r-project.org/web/packages/survival/survival.pdf`.

— (n.d.). `http://www.r-project.org`.

Reiter, J. P. (2003). Model diagnostics for remote-access regression systems. *Statistics and Computing*, 13: 371–380.

— (2004). New approaches to data dissemination: A glimpse into the future (?). *Chance*, 17: 12–16.

Reiter, J. P. and Kohnen, C. N. (2005). Categorical data regression diagnostics for remote systems. *Journal of Statistical Computation and Simulation*, 75: 889–903.

Reznek, A. P. (2006). Recent confidentiality research related to access to enterprise microdata. Prepared for the Comparative Analysis of Enterprise Microdata (CAED) Conference, Chicago, IL, USA.

Rowland, S. (2003). An examination of monitored, remote access microdata access systems. In *National Academy of Sciences Workshop on Data Access*.

Scouten, B. and Cigrang, M. (2003). Remote access systems for statistical analysis of microdata. *Statistics and Computing*, 13: 371–380.

Sparks, R., Carter, C., Donnelly, J., Duncan, J., O'Keefe, C. M., and Ryan, L. (2005). A framework for performing statistical analyses of unit record health data without violating either privacy or confidentiality of individuals. In *Proceedings of the 55th Session of the International Statistical Institute*. Sydney, Australia.

Sparks, R., Carter, J., C. Donnelly, O'Keefe, C. M., Duncan, J., Keighley, T., and McAullay, D. (2008). Remote access methods for exploratory data analysis and statistical modelling: Privacy-preserving analytics™. *Computer Methods and Programs in Biomedicine*, 91(208–222).

Weibull, W. (1951). A statistical distribution function of wide applicability. *Journal of Applied Mathematics*, 18(3): 293–297.

Willenborg, L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control*, vol. 155 of *LNCS*. Springer.