New approaches to analysis of rates, and survival analysis, in the ONS Longitudinal Study

Andy Sloggett
London School of Hygiene & Tropical Medicine

The ONS Longitudinal Study (ONS LS) is a large record linkage study of 1% of the population of England and Wales. As well as holding individual level census information it contains event data on births, deaths and cancer registrations. This lends itself to analysis of various incidence rates, fertility, birth spacing *etc.* as well as mortality.

However the construction of rates and their analysis, especially in a multivariate way, has not been widely exploited outside of ONS because of technical difficulty, complicated by restriction on release of individual-level data for security reasons.

Within ONS various sophisticated analyses involving construction of rates have been performed over the years, mostly by use of specially written software which can calculate person-years at risk (indeed person-days at risk) from the LS records. "Stage 3" and "Smartie" are two such packages – innovative and useful in their day but not very accessible. So, although these techniques were available to academic researchers on special request take-up was always low, due to complexity of specification, relative scarcity of staff to process, and subsequent potential for delay - real or perceived

Survival analyses and the like have generally therefore been approached in quite simple ways, such as tabulating proportions dying within five years of an event, or by the next census. This has sometimes been taken further by use of logistic regression to compare such proportions in a multivariate way.

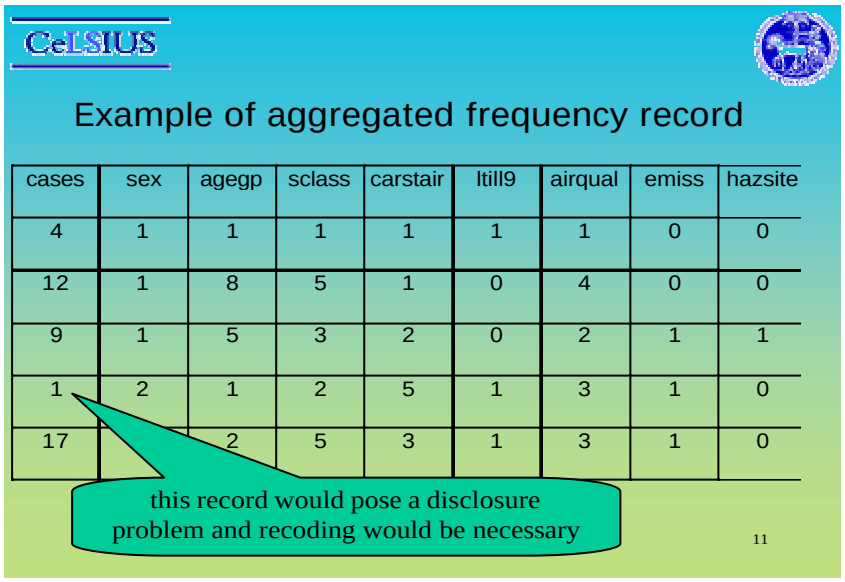
Proportional hazards regression is quite possible with the individual-level data but has had limited appeal to external researchers because the normal way of supplying data to researchers has been by aggregation, thus destroying the individual survival time records required by such analysis.

The speed and sophistication of statistical packages has improved hugely in the last decade and several packages now facilitate record splitting which enables records to be split into segments (of say a year) and the events and person-years of such segments aggregated. One package particularly adept at this process, and which has gained a reputation in this area, is Stata.

Aggregation of data is illustrated in Figure 1. Here one line of data exists for each unique combination of categories across explanatory variables (8 illustrated) and the number of original records that each line represents is given by another variable, in this illustration the variable "cases".

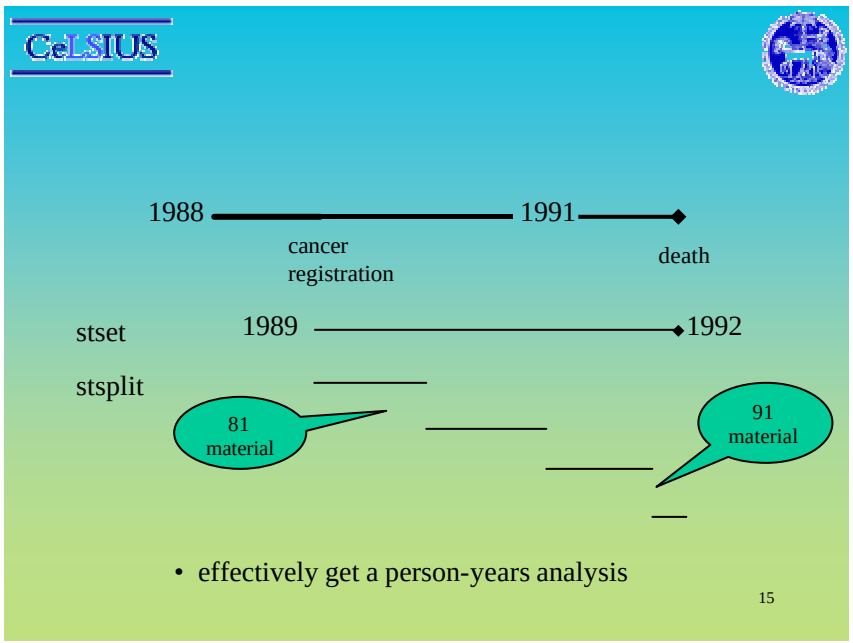
In Figure 1 one record is identified as one that could not be released to a user outside ONS because it is potentially disclosive. Some recoding or further aggregation would be required before this data could be released.

Figure 1



An illustration of how an individual's survival record would be processed to enable a survival analysis is shown in Figure 2. This is an illustrative record of time to death following a diagnosis of cancer (top line).

Figure 2



If we declare a start date and an end date (plus an outcome indicator – dead or censored) to the statistical package it will then allow us, with quite simple commands, to split the record into any number of segments of any chosen length. It is usually convenient to choose segments of one year. Each segment automatically receives its own start date and end date and using these the person-years in the segment is easily calculated. Segments that the person survives right through will normally be one year in length unless they come up against the end of study. Segments in which they die or leave the study will be a fraction of a year, and this fraction can be calculated. Each segment automatically receives its own outcome indicator of dead or censored.

We can then "update" each segment with variables which would normally be time-varying *i.e.* would change with time. In Figure 2 the balloons show that segments before the 1991 census can receive information from the 1981 census (as the best estimate of current status) whereas post-1991 segments can be updated with material from the 1991 census.

Figure 3

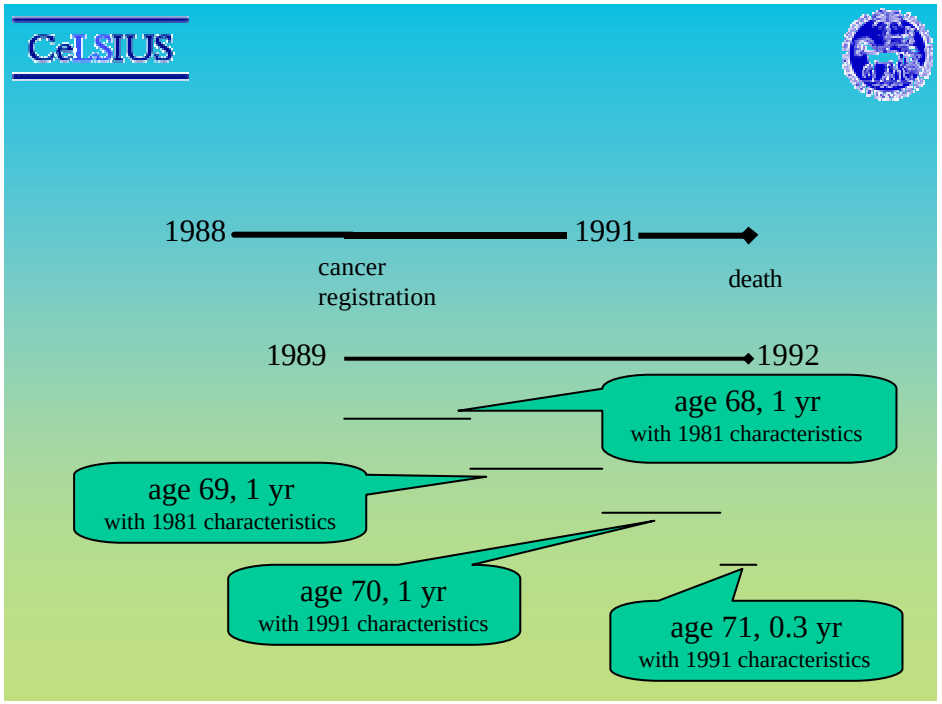
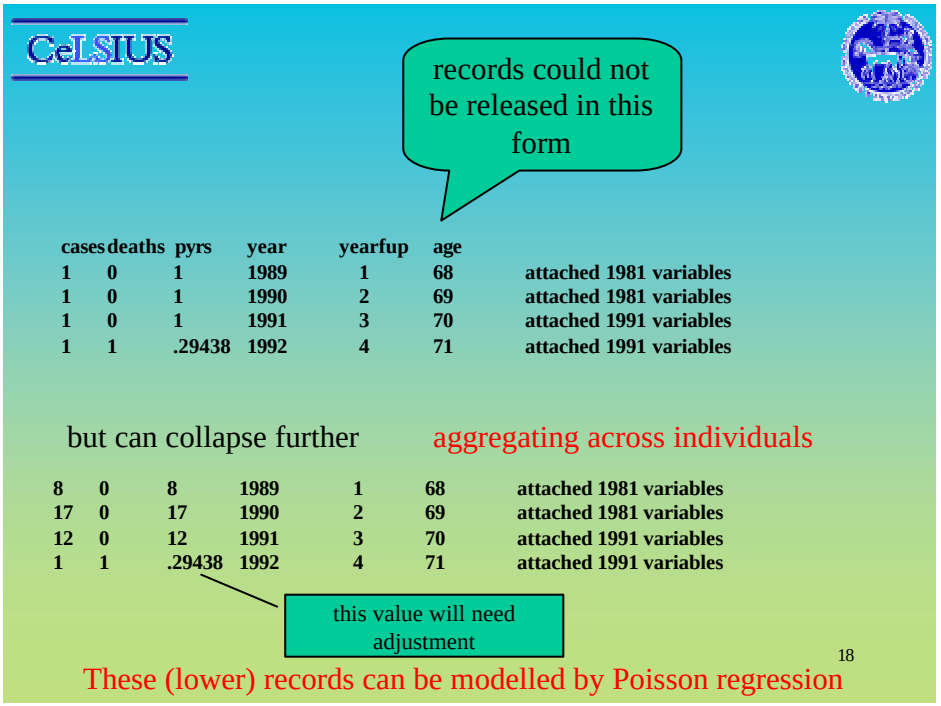


Figure 3 gives a similar and even more common example of updating age. As each year passes we all get one year older – sad but true!

Following this process the records can be aggregated, separately summing the events and the person years. Figure 4 gives examples of unaggregated and aggregated records. The former could be analysed but could not be released outside ONS because the sparse records are potentially disclosive. After aggregation they are far closer to being "releasable" but some records could still be considered disclosive – the final record is an

example – because the accuracy of the person-years figure pinpoints the exact date of death, which is disclosive. This could be disguised by adding or subtracting a small random element from the person years in this record, which would disguise the date of death but really not affect the analysis. Despite this procedure the record would still require clearance from ONS before being released to a user.

Figure 4



The aggregated records can be modelled in several packages, including Stata, in the form of a Poisson regression model, with deaths as outcome and person-years as exposure. In the aggregated form these models run very quickly, despite them representing huge volumes of data. Interactions and goodness of fit tests are all accommodated easily. Year of follow-up should always be included in such a model.

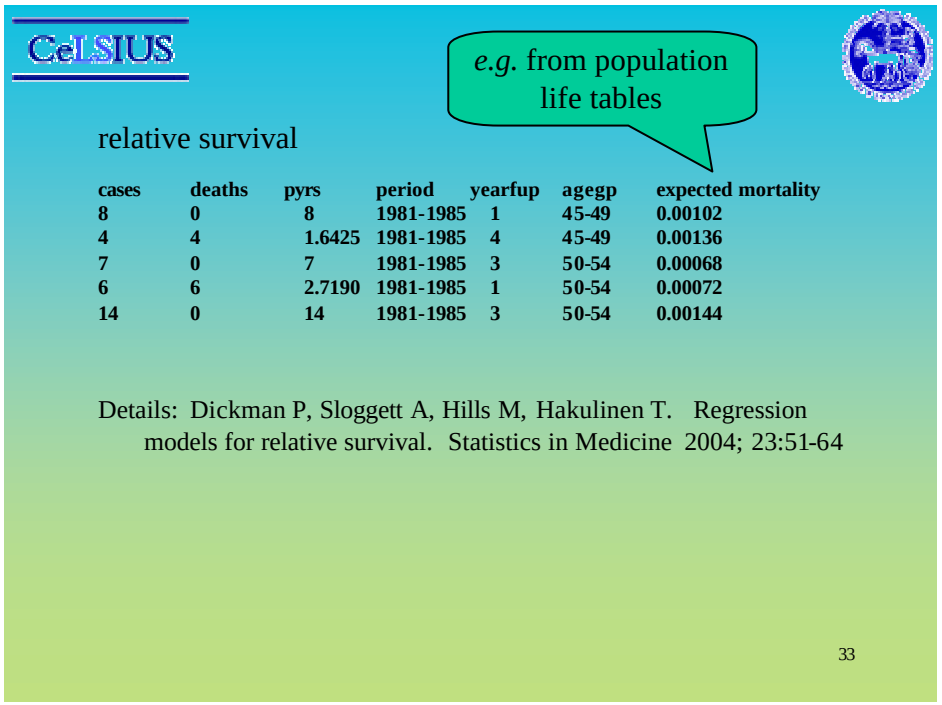
Models for incidence can be similarly organised. A slightly more complicated procedure of data handling is necessary to prepare the records since exposure continues after an event and repeated events are possible. However, with thought, the correct procedure can be derived.

An interesting extension of survival analysis, most relevant to mortality, although other analyses could be accommodated, is the analysis of ***relative survival*** or ***excess mortality***. This is mortality ***over and above*** that expected in the general population. To effect this we need expected probability of an event happening. In the case of mortality this is relatively easily provided by merging on to the unaggregated data a risk of death (${}_nq_x$ value), by age, sex, calendar period and anything else available, from a population life

table. When aggregated, with these risks summed, they provide expected mortality or expected deaths. Figure 5 shows records with expected mortality risk.

To use these records in the Poisson model we have to employ a user-defined link function, which effectively re-scales the model to work with excess deaths only. Once this is done the model behaves like any other. This form of mortality statistic is now widely used in the cancer epidemiology field and is beginning to catch on in other fields as well.

Figure 5



Results using cancer survival data from the ONS LS, processed in this way, have been very comparable with results using data from cancer registries. The latter is complete data from the whole population whereas the LS data is, of course, only a 1% sample. This can bring problems with small numbers unless some cancers are appropriately grouped – so it is unlikely to be of use for rarer cancers. Nevertheless the rich social data of the LS has no equivalent in registry data and some interesting analyses are possible with high incidence cancers.

Survival analysis, both simple and relative, by means of Poisson modelling is growing in popularity. Modelling of incidence is similarly possible and may hold more potential. Analyses do not need to be restricted to mortality. The models are very flexible and powerful and all normal regression diagnostics can be applied. This paper demonstrates how they can be effected using LS data, subject to the normal safeguards on disclosure.