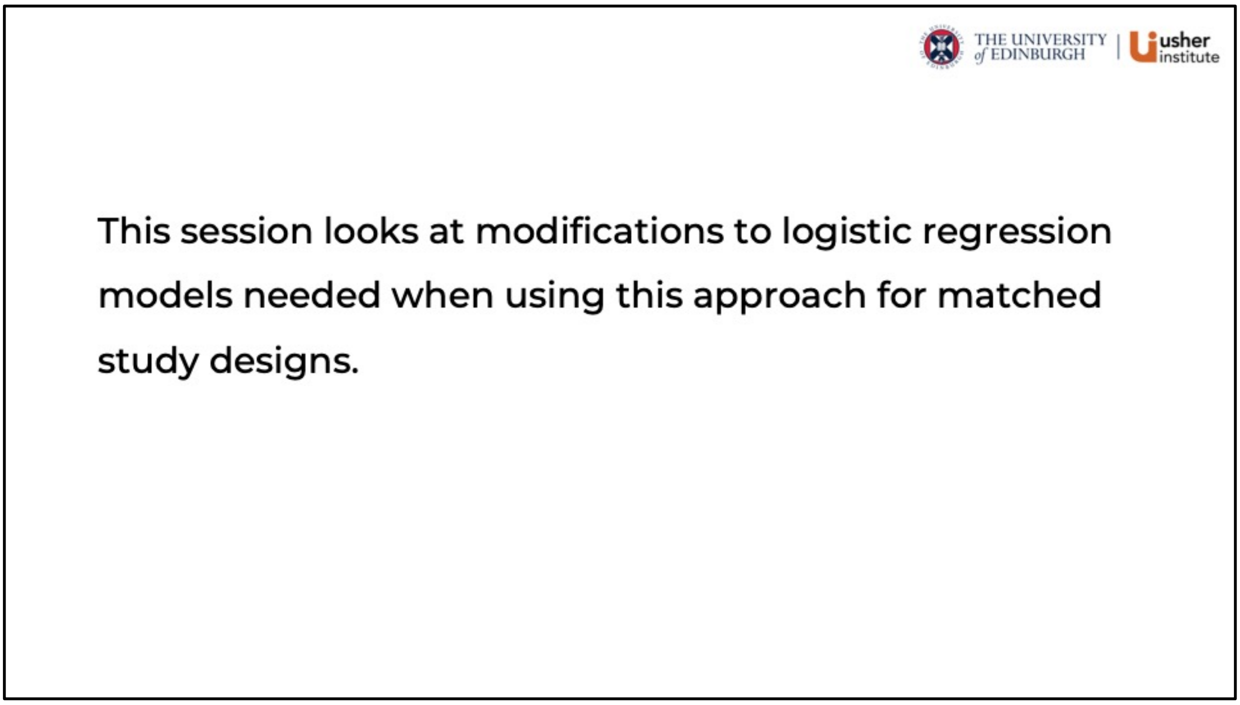
THE UNIVERSITY *of* EDINBURGH | usher institute

Logistic Regression for Matched Designs

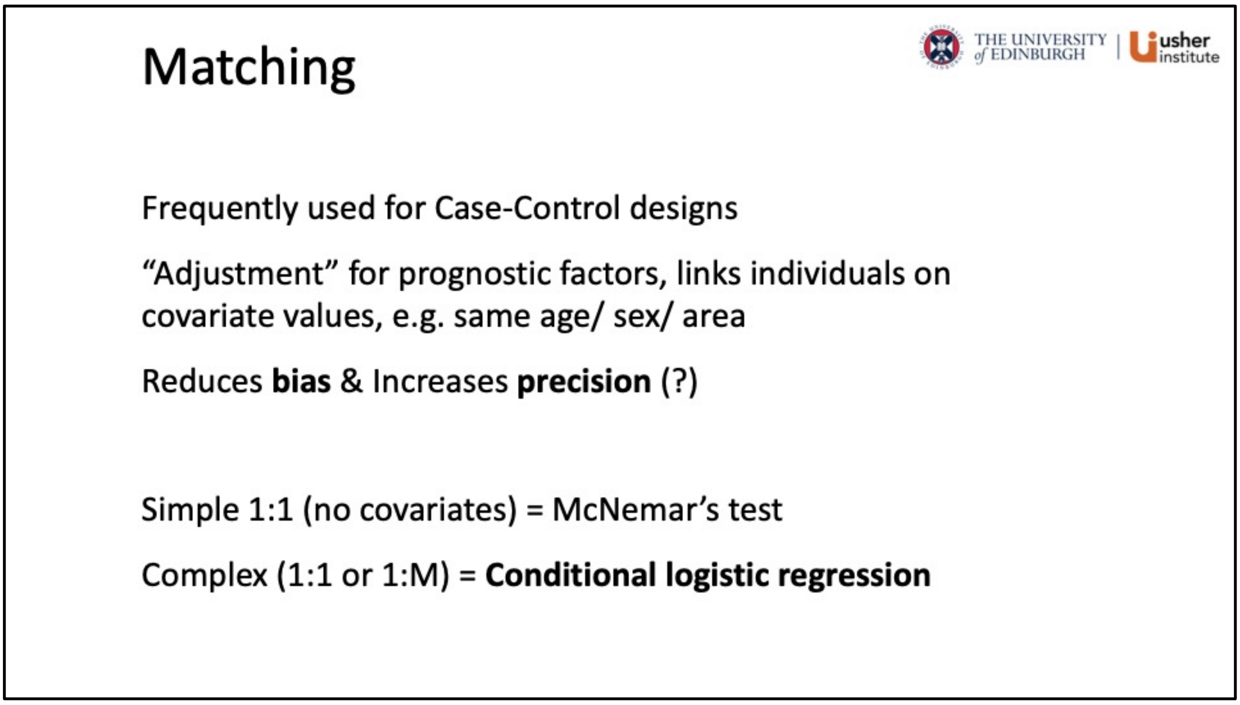
Niall Anderson

Supported by

THE UNIVERSITY *of* EDINBURGH | DDI Data-Driven Innovation



This session looks at modifications to logistic regression models needed when using this approach for matched study designs.



Matching

THE UNIVERSITY of EDINBURGH | usher institute

Frequently used for Case-Control designs

“Adjustment” for prognostic factors, links individuals on covariate values, e.g. same age/ sex/ area

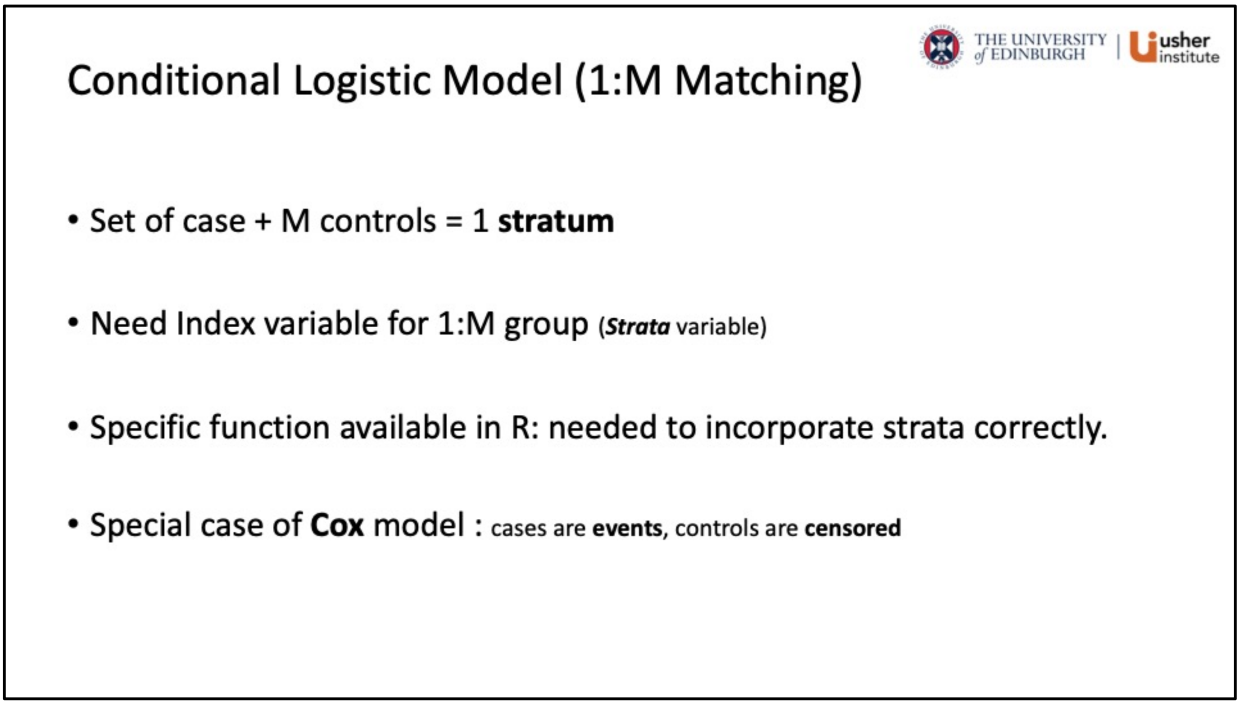
Reduces **bias** & Increases **precision** (?)

Simple 1:1 (no covariates) = McNemar’s test

Complex (1:1 or 1:M) = **Conditional logistic regression**

Matching is quite frequently used in case-control studies as a way to adjust for 1 or more prognostic factors that are not of key interest. This is done by grouping cases and controls who have the same or very similar values of those factors, such that each case is uniquely linked to 1 or more controls who have (for example) the same age/ sex/ area of residence etc. This will therefore remove the influence of those factors from the data (it pre-corrects the analysis for their effects), and may therefore reduce bias and increase precision in some studies – all else being equal, in the same situation, a matched study of equivalent size to a non-matched study should be more powerful.

If 1 case is matched to 1 control and there are no other covariates that need to be incorporated into the model, we can analyse a case-control study via McNemar’s test. More generally, however, we will need to use an adapted form of logistic regression that will take account of the clustering in the {case + control} or {case + M controls} groupings – the particular variant we need is called **Conditional Logistic Regression**.



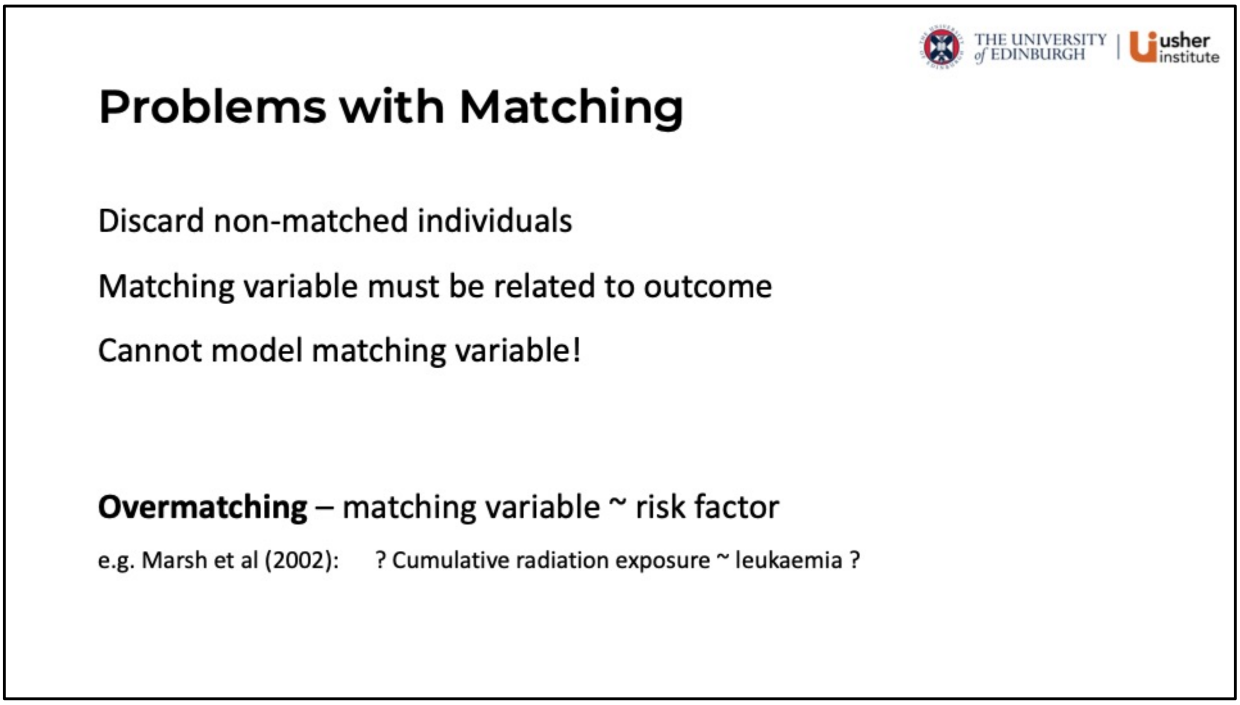
THE UNIVERSITY *of* EDINBURGH | usher institute

Conditional Logistic Model (1:M Matching)

- Set of case + M controls = 1 **stratum**
- Need Index variable for 1:M group (***Strata*** variable)
- Specific function available in R: needed to incorporate strata correctly.
- Special case of **Cox** model : cases are **events**, controls are **censored**

The key difference when fitting a conditional logistic model is that we have to provide our software with information about the linked set of case plus M controls – often called a **stratum**. This is quite simple to achieve in practice – our data set needs to have a simple index variable added, such that each individual in a stratum has the same value, with each value identifying a unique stratum. There is a specific function in R called ***clogit()*** that allows us to fit the correct model, once we have created the index variable for the strata – we must do this for the model to represent the data accurately!

As a point of mathematical interest, it turns out that the conditional logistic regression model is a special case of something called the Cox proportional hazards model, a type of model used to analyse survival data. The conditional model assigns cases to be **events** and controls to be **censored**, in the language of survival problems – don't worry if you are not familiar with these types of model, though!



Problems with Matching

Discard non-matched individuals

Matching variable must be related to outcome

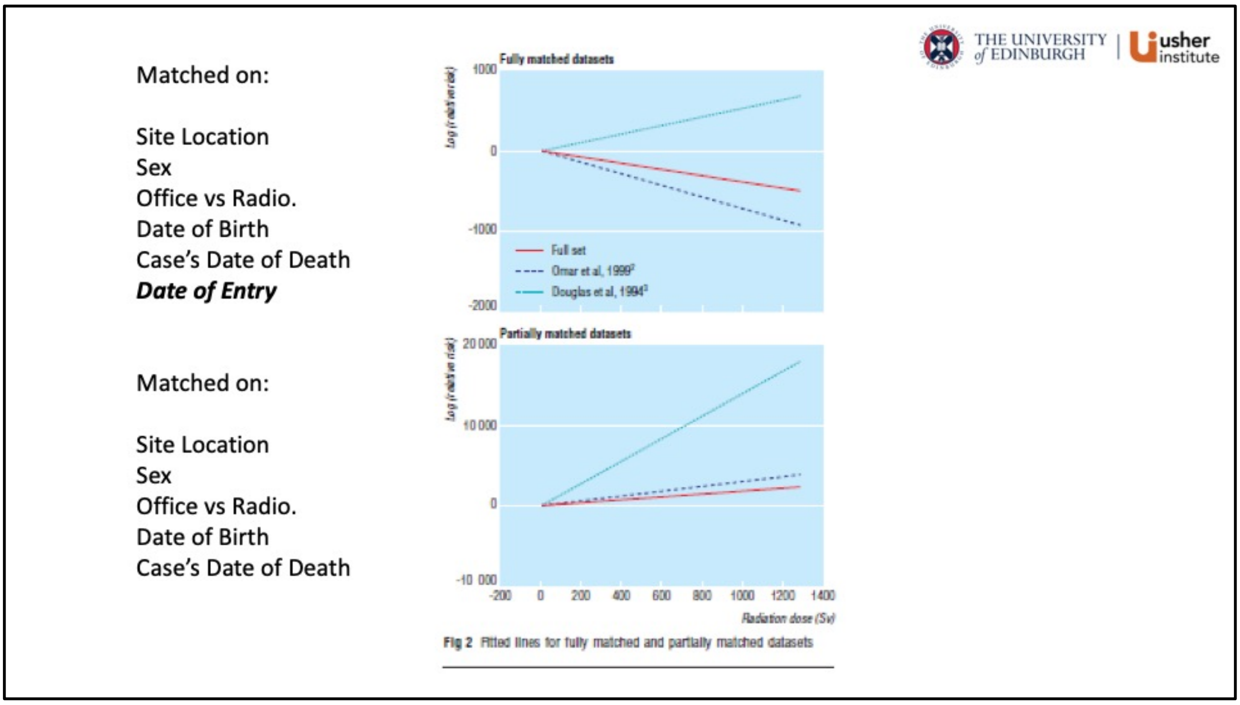
Cannot model matching variable!

Overmatching – matching variable ~ risk factor

e.g. Marsh et al (2002): ? Cumulative radiation exposure ~ leukaemia ?

It should be noted that the matching process can also introduce problems into our analysis, so is not always feasible in practice. The requirement to match cases to controls individually can mean that data on individuals (either cases or controls) who cannot be matched has to be **discarded**, which is wasteful of their information. It is also necessary for the variable(s) on which we match to be **related** to the outcome of interest for the matching process to bring any benefits – if it turns out that there is no relationship, we will have gained no increase in precision and may indeed have lost information by discarding unmatched individuals. An additional caveat is that we cannot include the matching variable in our model, so that we will not be able to determine the scale of its impact or direction of its effect.

An additional issue is the possibility of **overmatching** – an example of confounding in which a variable on which we match turns out to be associated with one of the risk factors we wish to assess via the study. As we have essentially matched on that risk factor, it is corrected out of the model and we would therefore see no effect, even when it is itself strongly related to outcome. This is thought to be a relatively rare phenomenon, but a nice example was outlined in a BMJ Education article in 2002 by Marsh et al – this is linked in the Resources section for this week's activity. The authors re-analysed 2 previous studies that attempted to investigate the relationship between lifetime exposure to radiation and leukaemia in nuclear power installation workers in the UK using a case-control design, but found that their results disagreed with the conclusions previously drawn about a likely link.



Matched on:

Site Location
Sex
Office vs Radio.
Date of Birth
Case's Date of Death
Date of Entry

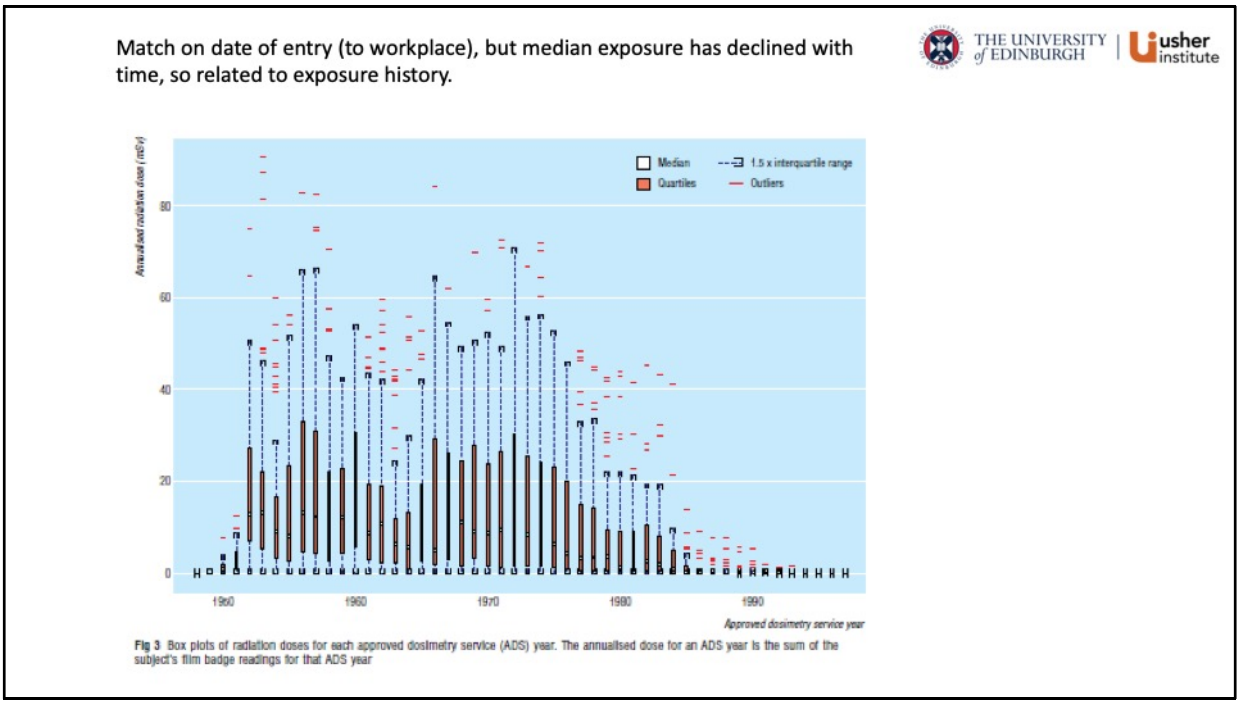
Matched on:

Site Location
Sex
Office vs Radio.
Date of Birth
Case's Date of Death

Fig 2 Fitted lines for fully matched and partially matched datasets

They matched cases and controls on a number of factors, including sex, type of role within the installation, their dates of birth and death and when they began working in the nuclear industry. Their initial results are shown in the top graph, which showed little relationship between risk of leukaemia and increasing radiation dosage (if anything – a protective effect of more radiation, which is counter-intuitive).

They realised that they had over-matched – the date of entry to the industry was associated with radiation dosage received and so removing this from their matching criteria produced the results in the lower graph, which showed a much more sensible dose-response relationship between leukaemia risk and radiation dose.



Match on date of entry (to workplace), but median exposure has declined with time, so related to exposure history.

Fig 3 Box plots of radiation doses for each approved dosimetry service (ADS) year. The annualised dose for an ADS year is the sum of the subject's film badge readings for that ADS year

This graph shows the problem more clearly – it shows the distribution of received radiation dosages over calendar time, from the early stages of the nuclear industry's existence to "present day". Clearly, significant advances in safety and engineering practices dramatically reduced radiation exposure over time, and so when a worker began employment would significantly correlate with the total dose they would be likely to report – hence, this created over-matching when that date of entry variable was included as a criterion.