NCRM National Centre for Research Methods

Data Linkage: An Overview

Natalie Shlomo

University of Manchester

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

UNIVERSITY OF Southampton

MANCHESTER 1824 The University of Manchester

THE UNIVERSITY OF EDINBURGH

Introduction to Data Linkage

Data (record) linkage brings together information from two different records that are believed to belong to the same person based on matching variables

- If two records agree on all matching variables, it is unlikely that they would have agreed by chance, the level of assurance that the link is correct will be high (the pair belongs to the same person)
- If all of the matching variables disagree, the pair will not be linked and it is unlikely that it belongs to the same person
- Intermediate situations where some matching variables agree and some matching variables disagree, need to predict whether the pair is a true match or a non-match

  Often need clerical intervention to determine matching status

Data Linkage is difficult in the presence of errors in collecting data and where no unique high quality identifier is available

Introduction to Data Linkage

Challenges of Data Linkage:

- Errors, variations and missing data on the information used to link records
- Differences in data captured and maintained in different databases, eg. different versions of date of birth compared to age
- Data dynamics and database changes over time, eg. name changes due to marriage and divorce, address changes

Typical problems in strings:

Misspelling, transpositions, fused or split words, missing or extra letters, extraneous information, missing punctuation

Typical problems with numerical variables:

Transposed numbers, insertions, deletions

Introduction to Data Linkage

Data Linkage typically involves three stages:

- Pre-linkage: Editing and data cleaning, parsing, standardizing matching variables
- Linkage: Bringing pairs together for comparison and determining correct matches, i.e. belong to the same person. All pairs are produced within blocks determined by blocking variables
- Post-linkage: Checking residuals, determining error rates, carry out analysis accounting for linkage errors

Properties needed for matching variables:

- Unique; Available; Known; Accurate Stable over time

Introduction to Data Linkage

Context of data linkage to carry out statistical research and inform policy

Focus on two main methods of data linkage and their combination:

- Deterministic (exact) matching
- Probabilistic matching

Deterministic (exact) matching method based on an exact one-to-one character match of matching variables

Probabilistic matching method used if partial identifiers are available, i.e. names and addresses

- A score is computed for each potential pair based on individual probabilities of agreement for each matching variable

Deterministic Linkage

Deterministic (exact matching) method

Records in two datasets must agree exactly on the matching variables in order to conclude that they correspond to the same individual

It can be used when a high quality identifier such as an ID number is available

All matching variables have the same weight associated to them so matching on gender carries the same weight as matching on last name

Incorporating some errors:

In fuzzy matching, exact matching is carried out with a wildcard substituted for characters, eg. *a*a*a can be banana, pajama, etc.

Use transformed data, such as 'Soundex' for names or truncated fields (first 5 letters of a name) which must match exactly

Probabilistic Data Linkage

Does not require that all identifying fields match exactly in order to conclude that the records belong to the same individual

Frequency analysis of data values necessary in order to calculate for each matching variable a weight that indicates for any pair of records how likely it is that they refer to the same entity

Uncommon value agreement stronger evidence for linkage

Large weights assigned to fields that match and small weights are assigned to fields that don't match

Sum the scores over all matching variables and compare the sum to threshold values in order to determine if the pair should be declared a match, a non-match or undetermined for clerical review

Probabilistic Data Linkage

Method relies on calculating scores based on probabilities

Determines agreements between matching variables between a pair of records as well as disagreements

Either from previous experience of record linkage on a similar application or based on a preliminary linkage exercise, how likely is it that the variables which agree between a pair would have done so by chance if the pair were not correctly matched?

Compare this measure to how likely the agreement would be in correctly matched record pairs

Can also use latent class modelling and EM algorithm to estimate the matching probabilities without the need for test data

Probabilistic record linkage more computational demanding and more difficult to program but it reduces the number of overlooked matches by modelling the inconsistencies in the data and taking them into account

Probabilistic Data Linkage

Criterion for good matching variables: agreement between variables which are more typical of correctly matched pairs, rather than those which might have occurred by chance in unrelated records

Example, variables that might agree by chance in unmatched record pairs are those which don’t divide the population unto many subclasses, for example gender

Key technical issues in the development of data linkage procedures

1. Good quality identifiers available to discriminate between the person to whom the record refers and all other persons
2. Deciding whether discrepancies in identifiers are due to mistakes in reporting for a single individual
3. Processing a large volume of data within a reasonable amount of computing processing time

Data Linkage Parameters

Three key parameters for a successful probabilistic data linkage:

- Quality of the data
- The chance that values of a matching variable will randomly agree
- Ultimate number of true matches that exist in the database

Not all fields for matching give you the same amount of information and uncommon value agreement stronger evidence for linkage

To incorporate the discriminating power of matching fields, the weights are computed as a ratio of 2 frequencies:

- number of agreements of a field in record pairs that represent the same individual
- number of agreements in a field in record pairs that do not represent the same individual

Probabilistic Data Linkage

Need to define the agreement pattern: γ

For example, 3 matching variables with binary comparison tests whether

γ_1 - pair agrees on last name

γ_2 - pair agrees on first name

γ_3 - pair agrees on street name

Simple agreement pattern

$\gamma = (1,0,1)$ and in fact, there would be 8 such patterns

Complex agreement pattern

$\gamma = (0.66,0,0.80)$ and can be based on string comparators

Probabilistic Data Linkage

Data quality is the first parameter of probabilistic linkage – the degree to which the information contained for a matching variable is accurate and stable across time

Data entry errors, missing data, or false dates diminish accuracy and produce low quality

Higher quality data, more likely to make a correct match

Data quality is reflected in one of the probabilities needed for the process – the *m*-probability

Conditional probability that a record pair has an agreement pattern γ given that it is a match (the same person) $m = P(\gamma \mid M)$

This is approximately 1-error rate and is referred as Reliability

Probabilistic Data Linkage

Another parameter depends on the number of random agreements denoted the u-probability

Conditional Probability that a record pair has an agreement pattern γ given that it is not a match $u = P(\gamma | U)$

The third parameter is: $P(M)$ the prior probability of a correct match

Then according to Bayes theorem:

$$P(M | \gamma) = \frac{P(\gamma | M)P(M)}{P(\gamma)}$$

Agreement (or likelihood) Ratio assuming conditional independence:

$$R(\gamma) = \frac{P(\gamma | M)}{P(\gamma | U)} = \frac{P(\gamma_1 | M) \times P(\gamma_2 | M) \times \ldots \times P(\gamma_k | M)}{P(\gamma_1 | U) \times P(\gamma_2 | U) \times \ldots \times P(\gamma_k | U)}$$

Order the comparison vectors by the agreement ratio $R(\gamma)$ and choose upper and lower cut off values for $R(\gamma)$ to determine correct matches and correct non-matches

Probabilistic Data Linkage

Now take the logarithm and we obtain the sum of matching weights for each separate matching variable:

$$\log(R(\gamma)) = \log\left(\frac{P(\gamma_1 | M)}{P(\gamma_1 | U)}\right) + \log\left(\frac{P(\gamma_2 | M)}{P(\gamma_2 | U)}\right) + \ldots + \log\left(\frac{P(\gamma_k | M)}{P(\gamma_k | U)}\right)$$

Example:

P(agree on characteristic x|M)=

0.9 if x=first name, last name, age

0.8 if x=housenumber, streetname

P(agree on characteristic x|U)=

0.1 if x=first name, last name, age

0.2 if x=housenumber, streetname

Probabilistic Data Linkage

P(agree on characteristic Z|M)= 0.9 if Z=first name, last name, age
0.8 if Z=housenumber, streetname, sex

P(agree on characteristic Z|U)=0.1 if Z=first name, last name, age
0.2 if Z=housenumber, streetname,sex

Name	Address	Birth year	Gender
Samantha Smith	435 Main St	1954	M
Sam Smith	435 Main St	1955	F

γ = (disagree first name, agree last name, agree hsnm, agree stnm, disagree birth year, disagree sex) = (0,1,1,1,0,0)

$$\ln(R(\gamma)) = \ln((1-0.9)/(1-0.1)) + \ln(0.9/0.1) + \ln(0.8/0.2)$$
$$+\ln(0.8/0.2) + \ln((1-0.9)/(1-0.1) + \ln((1-0.8)/(1-0.2)) = \; -0.81$$

We can use dictionaries and string comparators which would give partial agreement weight to ‘Sam’ and ‘Samantha’ or a deviation in only 1 year of birth

Probabilistic Data Linkage

Data quality quantified by the m-probability with respect to accuracy and stability of the matching variable

For any given field, the same value for m-probability applies to all records

Distinguishing power is quantified by the u-probability

This can be obtained by the probability that 2 records will randomly agree and is approximately 1/(number of values)

If it is high then the field has low distinguishing power, eg. gender

In contrast to the m-probability, a matching variable may have multiple values of u-probabilities each corresponding to a specific value in the matching variable

u-probability typically estimated as the proportion of records with a specific value based on the frequencies seen in the primary data source

Blocking

Number of possible comparisons increases with the product of the file sizes

For large files, it is impractical to link every possible pair in the two files, for example 2 files of size 10,000 will result in 100,000,000 comparisons

We restrict the comparisons to blocks of data where one or more variables need to match exactly and are likely to refer to the same person, thereby reducing the time spent searching the file

Utilizes a deterministic approach to assist the probabilistic method of record linkage

Can block sequentially in an iterative process using different variables

- Start with the most restrictive deterministic matching then proceed to less restrictive models

Example: block on post code and surname, perform clerical review on the set of designated potential matches, and then match on residual files of records not matched using another blocking criteria, such as year of birth

Evaluation and Thresholds

Thresholds for determining match status are based on minimizing errors:

- error of linking unmatched records - Type I error
- error of not linking matched records – Type II error

As in classical decision theory, the thresholds are determined by how much you are willing to be wrong based on the two error types which are pre-determined

High values of overall scores suggest a correct match

Low values of overall scores suggest an incorrect match

Evaluation and Thresholds

What constitutes high and low?

Calculate a frequency distribution to determine critical values of high or low values based on levels of significance (how much you are willing to be wrong)

If we had training data where we knew the correct matching status, we could derive two distributions, for true matches and true non-matches:

Lower distribution for true non-matches typically contain lower values of weights (typically this group is very large as there are many more possible incorrect comparisons than there are correct comparisons)

Upper distribution for true matches contain higher values of weights

Evaluation and Thresholds

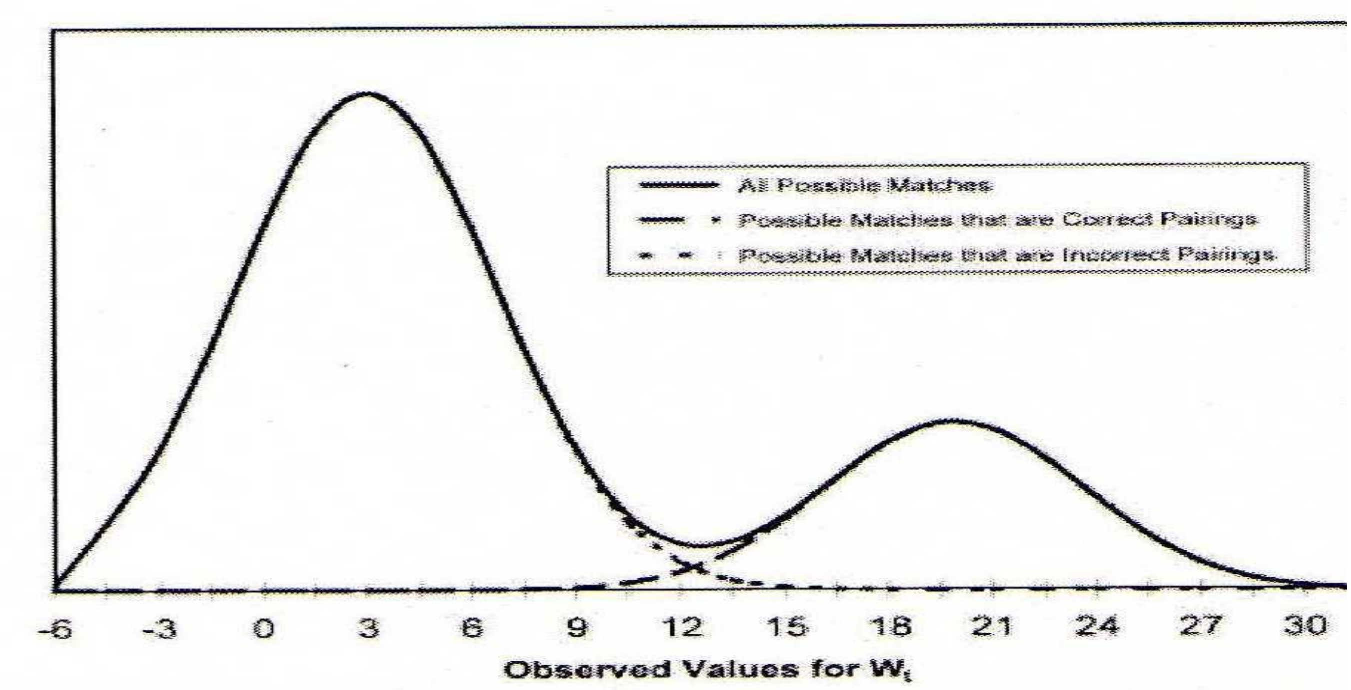


Figure 1. Distribution of w_t.

Must chose threshold: value of Score W$^-$ below which automatically classify as incorrect matches

and value of Score W$^+$ above which automatically classify as correct matches

In between W$^-$ and W$^+$ we would need to carry out a clerical review

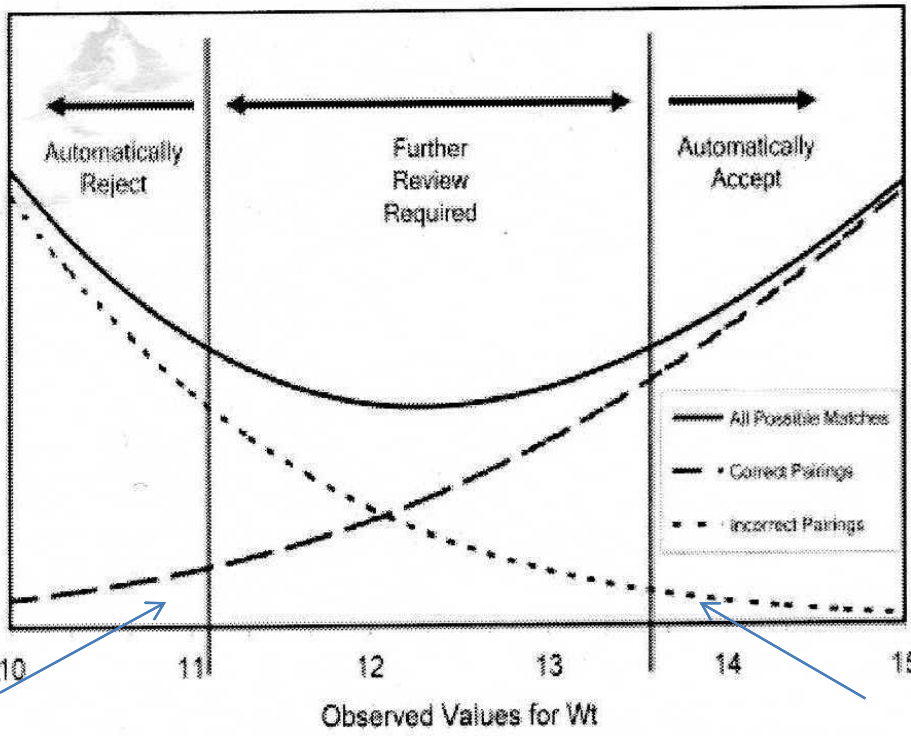
Evaluation and Thresholds

The decision rule is based on 3 types of pairs:

- Believed to be correct matches
- Unknown and might be correct matches
- Unlinked pairs

Empirical distributions can be used to determine thresholds based on training data and

Pre-set Type 1 and Type2 errors



Type 2 error

Type 1 error

Figure 2. w_t cut-offs for accepting and rejecting possible matches.

Evaluation and thresholds

- After record linkage, need to carry out logical checks in the data for evaluation, i.e. might obtain situations such as hospital discharges after a death

- Errors result from poor quality data

Recommended:

On those pairs declared a match - carry out a small random sample and check for accuracy in the matching status, particularly for those near the threshold cut-off values

On those pairs declared a non-match – carry out a small random sample and check for accuracy in the matching status, particularly for those near the threshold cut-off values

Use the errors to compensate for linkage errors when analysing linked data

Evaluation and Thresholds

		True Status	
		Non-Matches (null hypothesis)	Matches (alternative hypothesis)
Decision	Not Linked pairs (fail to reject null)	Not Linked non-matches (True Negative)	Not linked matches Type II error (False Negative)
	Linked pairs (reject null)	Linked non matches Type I error (False Positive)	Linked Matches (True Positive)

Precision=tp/(tp+fp)

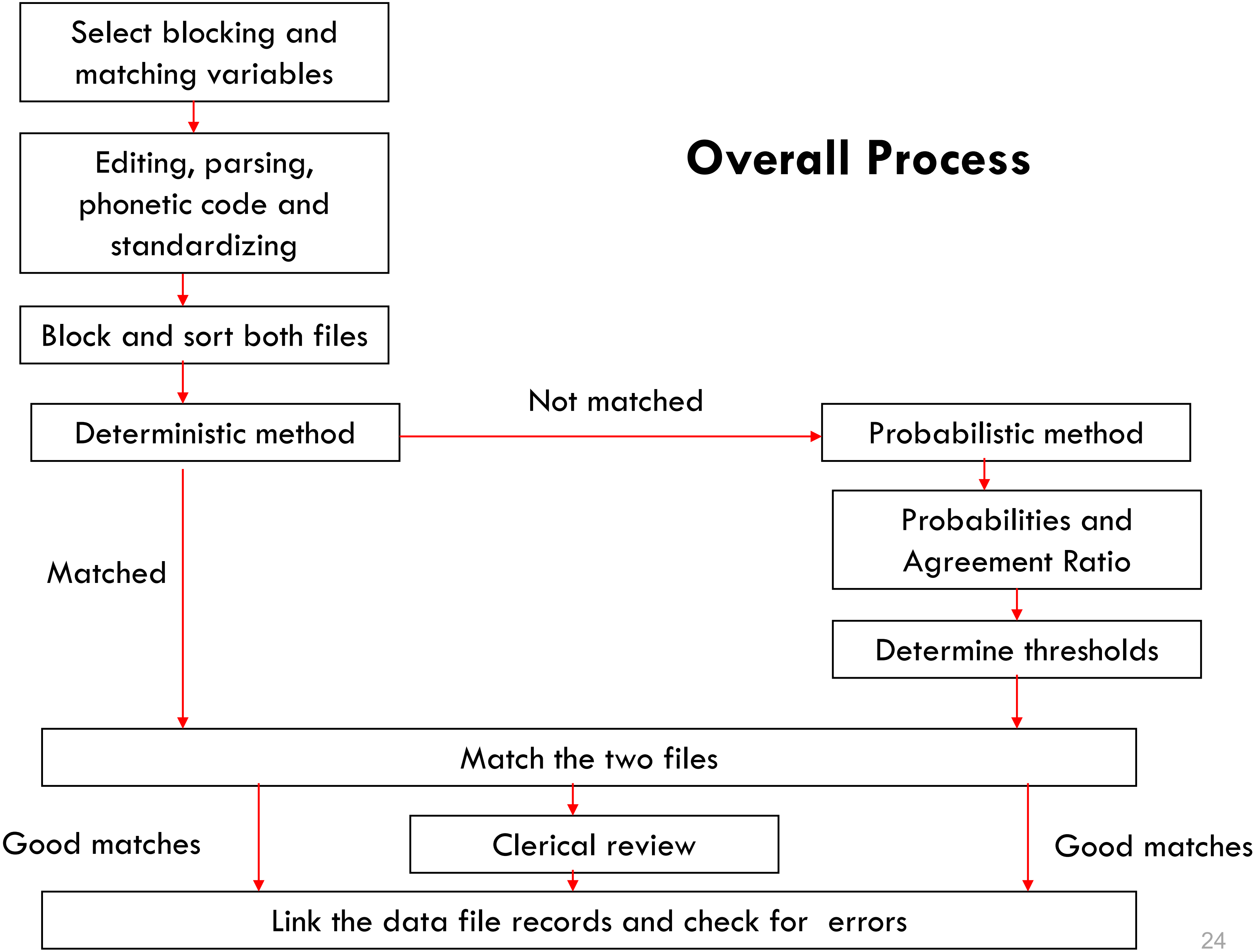
False Positive Rate = fp/(tn+fp)

False Negative Rate= fn/(tp+fn)

Specificity (ability to recognize incorrect matches) = tn/(tn+fp)

Sensitivity or recall (ability to recognize true matches)= tp/(tp+fn)

Overall Process



Select blocking and matching variables

Editing, parsing, phonetic code and standardizing

Block and sort both files

Deterministic method

Not matched

Probabilistic method

Probabilities and Agreement Ratio

Determine thresholds

Matched

Match the two files

Good matches

Clerical review

Good matches

Link the data file records and check for errors

For more information visit

www.ncrm.ac.uk/resources/online