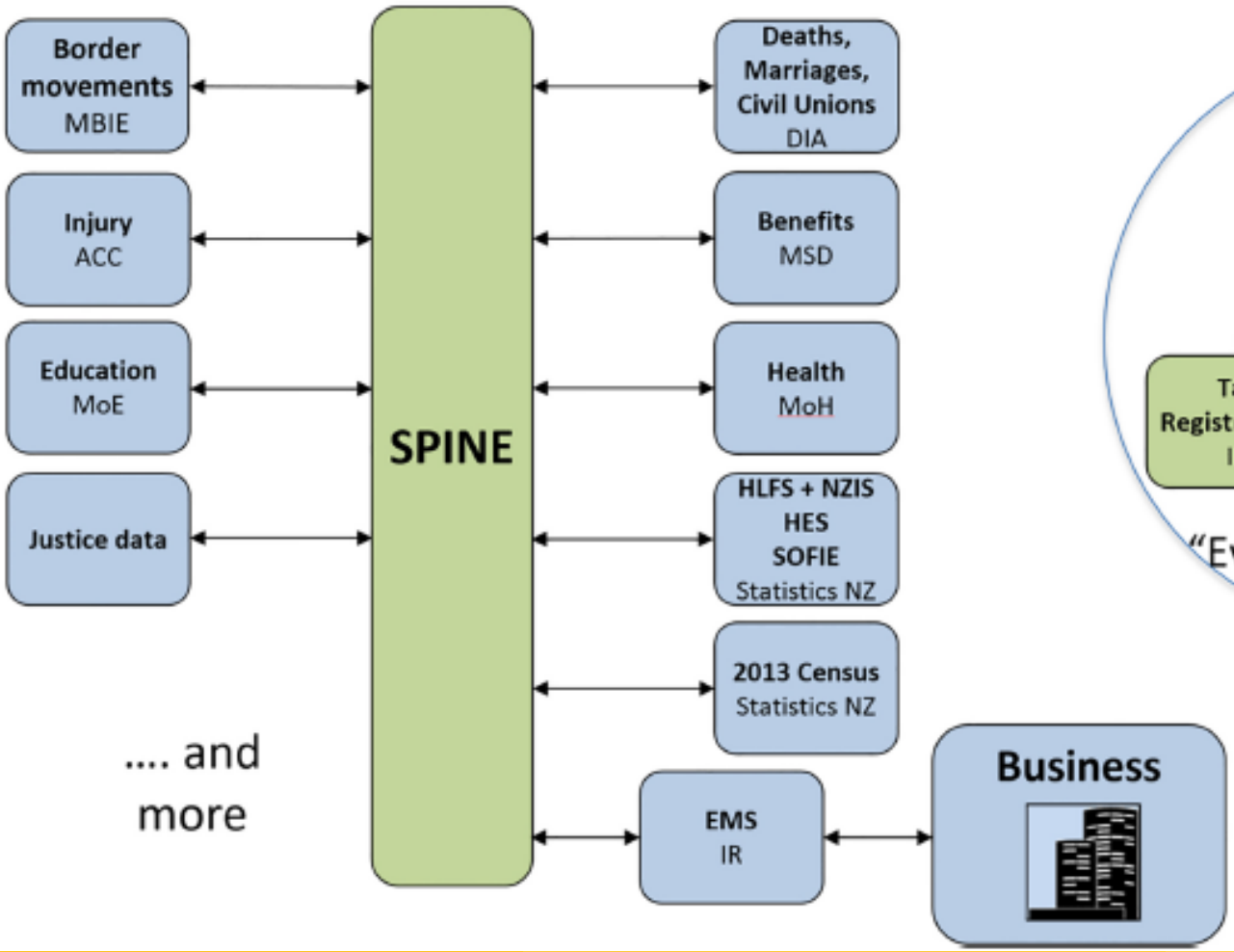
Investigating Linkage Bias in the IDI Using Census and Education data

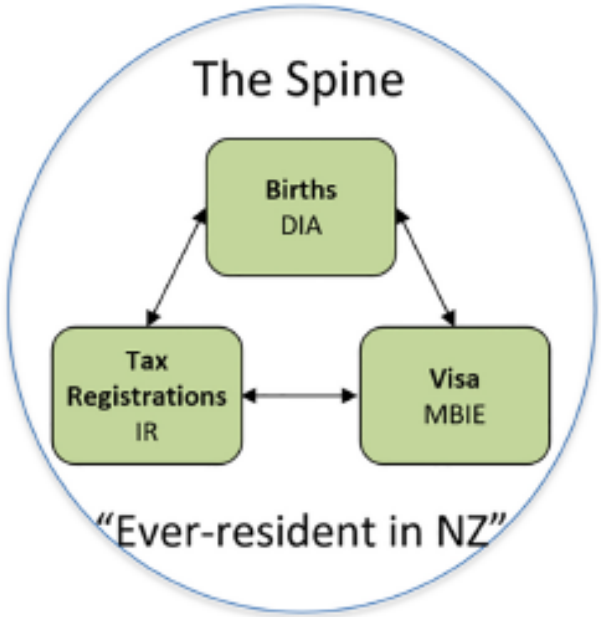
Eileen Li
y.li@auckland.ac.nz

August 2021

Integrated Data Infrastructure (IDI)

IDI: a large research database, holds microdata about people and households. The data is about life events, like education, income, benefits, migration, justice, and health. It comes from government agencies, Stats NZ surveys, and non-government organisations. Each data source contains individual level data, and they can be linked across by a unique identifier for each person.



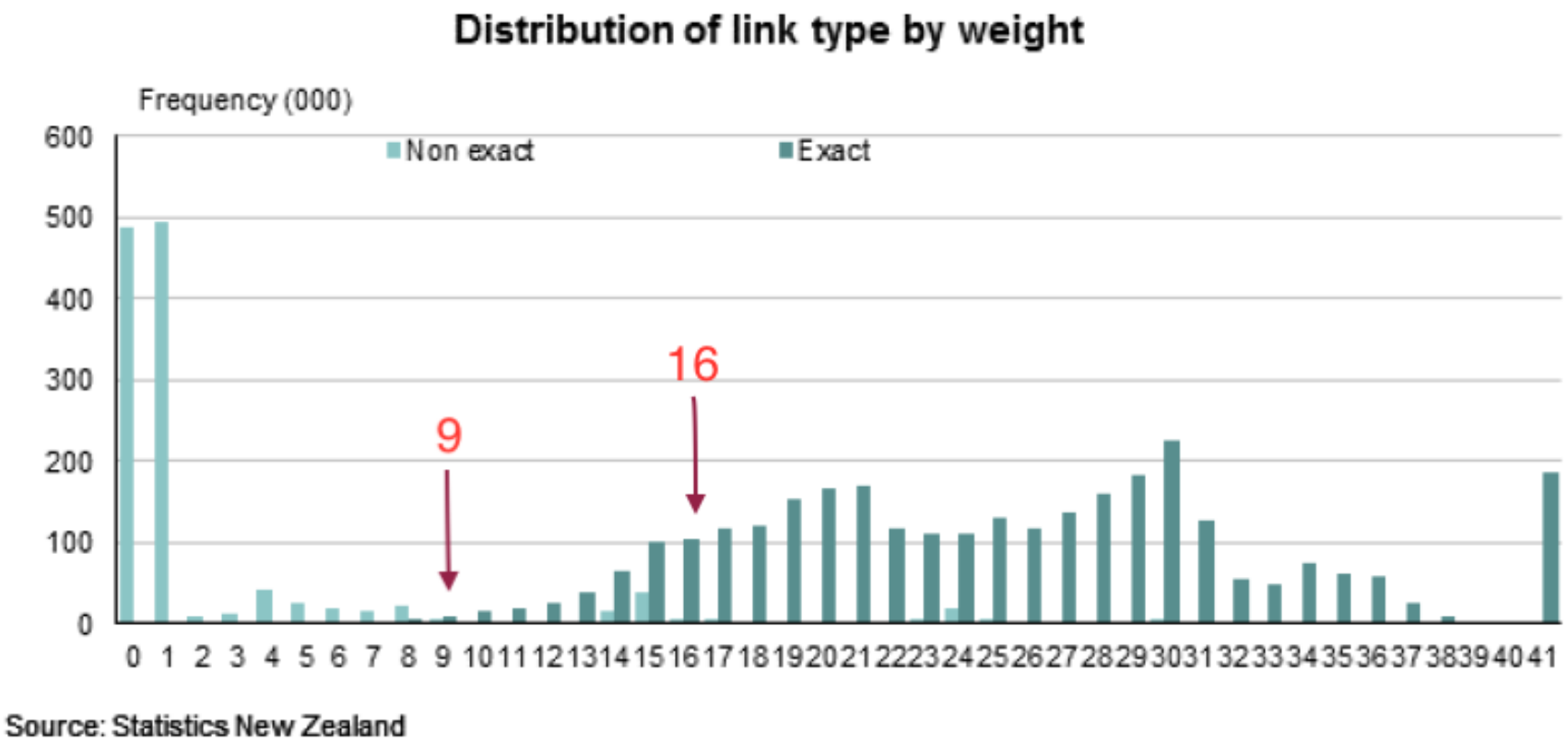


Disclaimer

Disclaimer: The results in this report are not official statistics, they have been created for research purposes from the Integrated Data Infrastructure (IDI), managed by Statistics New Zealand. The opinions, findings, recommendations, and conclusions expressed in this report are those of the authors, not Statistics NZ. Access to the anonymised data used in this study was provided by Statistics NZ in accordance with security and confidentiality provisions of the Statistics Act 1975. Only people authorised by the Statistics Act 1975 are allowed to see data about a particular person, household, business, or organisation, and the results in this report have been confidentialised to protect these groups from identification. Careful consideration has been given to the privacy, security, and confidentiality issues associated with using administrative and survey data in the IDI. Further detail can be found in the Privacy impact assessment for the Integrated Data Infrastructure available from www.stats.govt.nz.

Cut-off selection

Record linkage method developed by Fellegi and Sunter (1969). Suppose there are two populations A and B whose elements ae matched. $A \otimes B = \{(a, b); a \in A, b \in B\}$ is the union of two disjoint sets: $M = \{(a, b); a = b, a \in A, b \in B\}$, $U = \{(a, b); a \neq b, a \in A, b \in B\}$. Let $m_k = P(a_k = b_k|M)$ and $u_k = P(a_k = b_k|U)$. Then weight: $w_k = \log \frac{m_k}{u_k}$ when the variables agree, otherwise $w_k = \log \frac{1-m_k}{1-u_k}$. The overall weight for record pair (a, b) is then $w = \sum w_k$.

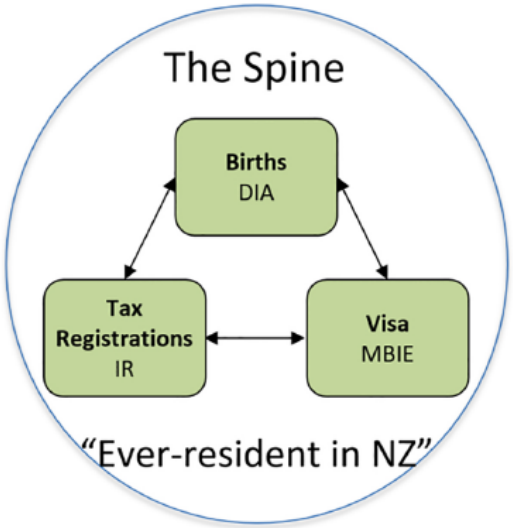


Linkage bias

- If linkage error is low (below 1%) then it is possible to use statistical modelling methods directly without accounting for the linkage error. If the linkage error exceeds 1% and the general statistical models are large or complicated, then linkage error must be accounted for in the modelling (Scheuren and Winkler, 1993).
- Linkage bias
  1. **Missed links** (false negative): links should be made but weren't due to missing linking variables
  2. **False links** (false positive): records actually belonging to two entities but incorrectly linked
- Ways to identify missed and false links:
  missed links $\rightarrow$ use *snz_spine* indicator
  false links $\rightarrow$ challenges...

Missed links

Three contributing data sources are: IRD, DIA and VISA. By comparing with the Census, we can find out who should be linked but are currently not.



From there, we will be able to implement the weighting approaching to increase the weights of people who are more likely to be missed out, and decrease the weights of people who are less likely to be missed out, to create an adjusted population.

False links

We have come up with a few ways to identify false links:

1. spine indicator from personal detail table
2. agency specific uids across refreshes (ACC, DIA, IRD, MoE and MoH)
3. clerical review dataset

We can then build logistic models to predict false links using characteristics including gender, birth year, ethnicity, NZDep, and death indicator.

Multiple linkage: based on logistic modelling, each person has a true positive probability. We replace this person with another person from the entire population based on the person's true positive probability. We repeat this process multiple times, same as we do multiple imputation. In doing so, we get the widest variance of estimated coefficients.

False links - spine indicator

Spine indicator tells us if one person is linked to the IDI spine or not.

Example1: Stable link (sum=6)

Refresh	Refresh 1	Refresh 2	Refresh 3	Refresh 4	Refresh 5	Refresh 6
spine_indicator	1	1	1	1	1	1

Example2: Unstable link (sum=3)

Refresh	Refresh 1	Refresh 2	Refresh 3	Refresh 4	Refresh 5	Refresh 6
spine_indicator	0	1	1	0	1	0

Example3: Never linked (sum=0)

Refresh	Refresh 1	Refresh 2	Refresh 3	Refresh 4	Refresh 5	Refresh 6
spine_indicator	0	0	0	0	0	0

False links - spine indicator

Spine_indicator from personal detail table across refreshes were summed up and values in between the highest and lowest were flagged 1, otherwise 0.

Sum based on 9 refreshes:

spine_ind_sum	0	1	2	3	4	5	6	7	8	9	SUM
Count	317280	5289	1275	40053	9156	2385	61548	2133	4395	3909684	4353198
Percentage	7.29	0.12	0.03	0.92	0.21	0.05	1.41	0.05	0.1	89.81	

Sum based on 6 refreshes:

spine_ind_sum	0	1	2	3	4	5	6	SUM
Count	357381	11880	1431	1755	3162	4266	3973323	4353198
Percentage	8.21	0.27	0.03	0.04	0.07	0.1	91.27	

False links - agency indicator

snz_uid changes refresh by refresh, whereas agency uid does not change across refreshes.

Refresh	**Refresh 1**	**Refresh 2**	**Refresh 3**	**Refresh 4**	**Refresh 5**	**Refresh 6**
snz_uid	11111	11211	11311	11411	11511	11611
ird_uid: A	22222	22222	22222	22222	22222	22222
ird_uid: B	33333	33333	44444	44444	44444	44444

Person B gets ird indicator = 1. Person A gets ird indicator = 0.

False links - agency indicator

Number of uid changes between two adjacent refreshes, by agency:

Refresh 1-2	Refresh 2-3	Refresh 3-4	Refresh 4-5	Refresh 5-6	Refresh 6-7	Refresh 7-8	Refresh 8-9
ird	**ird**	**ird**	**ird**	**ird**	**ird**	**ird**	**ird**
3447	2931	2451	1995	4143	22347	12219	3939
0.09	0.08	0.06	0.05	0.11	0.59	0.32	0.1
acc	**acc**	**acc**	**acc**	**acc**	**acc**	**acc**	**acc**
1362	4170	7437	5397	11358	16350	33858	1452
0.04	0.12	0.21	0.15	0.33	0.48	1.03	0.04
dia	**dia**	**dia**	**dia**	**dia**	**dia**	**dia**	**dia**
5616	3015	3561	2247	14769	14577	5118	3633
0.17	0.09	0.11	0.07	0.45	0.45	0.16	0.11
moe	**moe**	**moe**	**moe**	**moe**	**moe**	**moe**	**moe**
1050	1932	3690	780	24843	8994	2061	23925
0.03	0.06	0.12	0.03	0.82	0.3	0.07	0.8
moh	**moh**	**moh**	**moh**	**moh**	**moh**	**moh**	**moh**
7557	4905	4716	4836	11427	16773	3357	2268
0.2	0.13	0.12	0.13	0.3	0.44	0.09	0.06

Number of 2013 Census individuals with 2 or more agency uids:

ird	moe	msd	acc	moh	dia	nzta
10953	30858	173445	27702	25491	27234	405750

False links - clerical review

Overall, about 92% of all links agree on linking variables with a small tolerance of errors, and 8% of the record pairs did not agree on all linking variables. A random sample of this 8% were reviewed by StatsNZ analysts. This clerical review dataset contains about 35,000 observations, a score is given to each record pair where 1 indicates false positive (false link) and 0 indicates true positive.

Score	
0	1
24954	10008

False links - logistic regression

$$\log\left(\frac{p(y_i=1)}{1-p(y_i=1)}\right) = \beta_0 + \beta_1 gender_i + \beta_2 birthyear_i + \beta_3 ethnicity_i + \beta_4 NZDep_i + \beta_5 Death_i + \epsilon_i$$

We build 7 separate logistic models, with spine_ind, acc_ind, dia_ind, ird_ind, moe_ind, moh_ind and score as the outcomes respectively.
For these 7 models, we apply them to 2018 ever-resident NZ population and get predicted probabilities of being one for those indicators. We then check the correlation coefficients among these predicted probabilities.

6 refresh	fitted_cr	fitted_acc	fitted_dia	fitted_ird	fitted_moe	fitted_moh	fitted_spine
fitted_cr	1.000	0.389	-0.090	0.086	0.097	0.329	0.403
fitted_acc	0.389	1.000	-0.059	0.249	0.245	0.469	0.188
fitted_dia	-0.090	-0.059	1.000	0.769	0.074	0.579	0.133
fitted_ird	0.086	0.249	0.769	1.000	0.138	0.816	0.376
fitted_moe	0.097	0.245	0.074	0.138	1.000	0.320	-0.154
fitted_moh	0.329	0.469	0.579	0.816	0.320	1.000	0.325
fitted_spine	0.403	0.188	0.133	0.376	-0.154	0.325	1.000

Conclusion

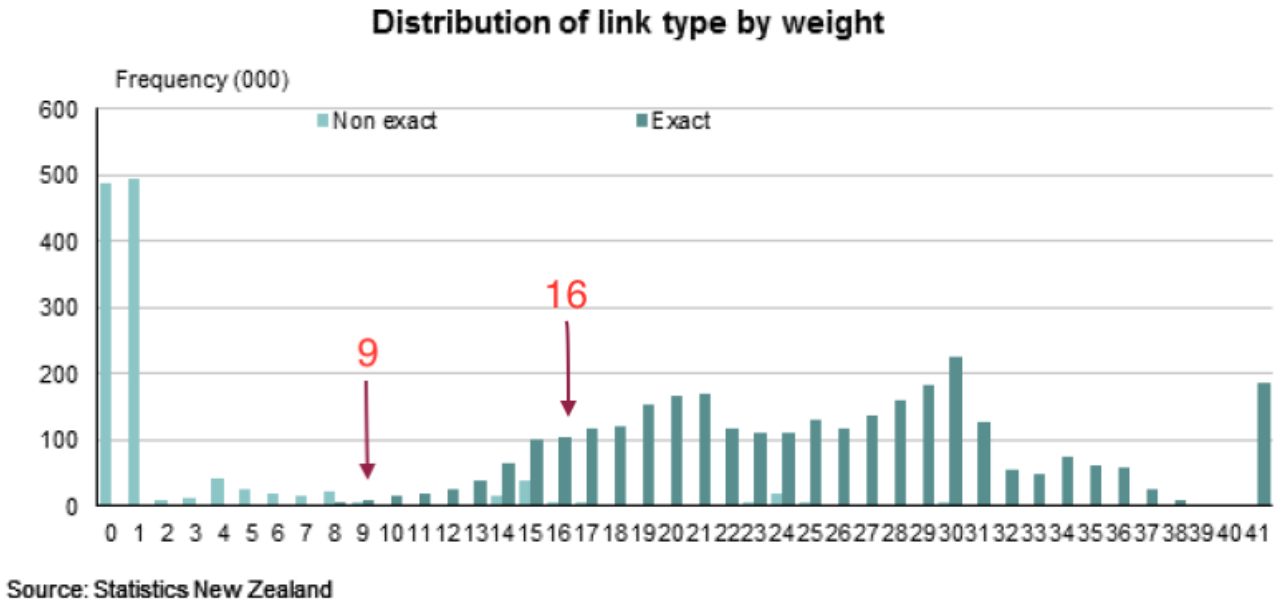
1. The lack of high correlation coefficients among predicted probabilities might suggest that these indicators are not picking up the false links.
2. It might also be that the false links are random.
3. Any ideas or comments?

Thank you

1. Barry Milne and Thomas Lumley
2. Michael Alspach, Anna Lin and Anapapa Mulitalo
3. Sheree Gibb and Craig Wright
4. Public Policy Institute

Cut-off selection

However, this is not how SNZ selects the cut-off value. SNZ used 'near exact' and 'non-exact' links to select cut-off. Near exact links agree on demographic information with some small tolerance for errors. Non-exact links are all other links. Starting with a cut-off of 0, a plot of near-exact and non-exact links by weight is obtained. Then non-exact links around weights following a spike of non-exact links are examined. During examination, if they find lots of actual links are removed, then the cut-off shall be re-set, or the pass will be modified. After cut-off is selected for each pass, they then calculate false positive rate. If the overall FP rate is above 2%, they will have to re-consider the cut-off values so that the overall FP rate is below 2%. This whole process is both scientific and subjective.



FP calculation (manual)

StatsNZ used two approaches for calculating false positive rate:

- manual method
- SoLinks method

An example of manual calculation is given below:

B	C	D	E	F	G	H	I	J	K	L	M	N
Pass	Total number of links	Number of non-exact links	Number of near exact links	Class	Sample size for each class	Size of class	Total sample size per pass	Final weight	Number of false positives per class for non-exact links	Rated up	Total number of false positives per pass for non-exact	False positive rate per pass for all links
	5200	217	4983								14.06	0.270
1	4000	162	3838	1	10	32.4	35	3.24	0	0	12.96	0.324
1	4000	162	3838	2	10	32.4	35	3.24	1	3.24	12.96	0.324
1	4000	162	3838	3	5	32.4	35	6.48	1.5	9.72	12.96	0.324
1	4000	162	3838	4	5	32.4	35	6.48	0	0	12.96	0.324
1	4000	162	3838	5	5	32.4	35	6.48	0	0	12.96	0.324
2	1200	55	1145	1	10	11	25	1.1	1	1.1	1.1	0.092
2	1200	55	1145	2	5	11	25	2.2	0	0	1.1	0.092
2	1200	55	1145	3	5	11	25	2.2	0	0	1.1	0.092
2	1200	55	1145	4	5	11	25	2.2	0	0	1.1	0.092

Where column K, **Number of false positives per class for non-exact links**, is based on clerical review of the sample of non-exact links, such that 1 indicates a false positive, 0 indicates a true positive, and if the analyst is unsure, they may assign a score of 0.5. Column F, **Class**, is used to divide non-exact links into specified number of even sized strata for each pass.

FP calculation (SoLinks)

A modelled approach to false positive estimation: SoLinks.
Initially, proportional model was built based on refresh samples as training data. Over time, logistic model took over and the training data is based on an independent clerical review process.
Key elements of SoLinks model (logistic model):

- Matching status: stratified variable in selecting training dataset; Matching status = 0_1234 where: 0 indicates the total number of agreements in first name, last name, sex and date of birth contained in the record pair. 1 indicates the agreement type for sex. 2 indicates the agreement type for date of birth. 3 indicates the agreement type for first name. 4 indicates the agreement type for last name
- Agreement types: are categorised to account for partial agreements, such that different level of agreements can be presented. For instance, for sex, E is perfect match and 0 is others (including sex missing or not matched)

With these agreement types defined, there are over 200 possible categorisations of matching status of a link.

FP calculation (SoLinks)

Data used for SoLinks come from a two-stage stratified random sample, with the matching status as strata. 30 links per stratum were selected. This clerical review was undertaken outside of refresh cycle and is less time restricted.

The logit transformation of p (probability of true match) has a linear relationship with the agreement types for names, DOB, and sex:
$log(\frac{\hat{p}}{1-\hat{p}}) = \beta_0 + \beta_{1i} Sex_i + \beta_{2j} DOB_j + \beta_{3k} First_k + \beta_{4k} Last_k$

The false positive probability is $1 - \hat{p}$.

Links with same matching status will have the same probability of true positive and probability of false positive. Number of links in each matching status is multiplied by their respective false positive probability. This gives the estimated number of false positives for each matching status. Then these are summed over all matching status, and divided by the total number of links, which yields the overall estimated false positive rate.

SoLinks example

Matching status	Numebr of links	P(TP)	P(FP)	Estimated false positives
4_{EXPE}	100	0.9963	0.0037	0.4
4_{EDPE}	1200	0.8923	0.1077	129.2
3_{E0PC}	10	0.7281	0.2719	2.7
3_{E0PD}	120	0.8908	0.1092	13.1
2_{00DC}	10	0.1617	0.8383	8.4
Total	1440			153.8

Estimate false pisitive rate $= \frac{153.8}{1440} = 10.7\%$