Supplementary material

1. Technical details and formulae

1.1 Condition required for a complete case logistic regression to produce an (asymptotically) unbiased estimate of the exposure odds ratio

If R is the response (observation) indicator (such that R=1 for complete cases and R=0 otherwise) and Y, X, and C are the outcome, exposure and confounders, respectively, the complete case exposure odds ratio is asymptotically unbiased provided $P(R=1|Y,X,C) = f(X,C) \times g(Y,C)$ for some functions f(X,C) and g(Y,C).

1.2 General expression for the odds ratio consistently estimated by a complete case logistic regression

Here we let X denote the exposure, Y_{cts} the continuous outcome, and Y_{bin} the binary outcome. As before, let R denote the observation indicator (R=1 if observed, R=0 if not). We assume that R depends only on X and Y_{cts} and, given these, not on Y_{bin}, with

$$P(R = 1|X, Y_{cts}) = \psi(X, Y_{cts})$$

for some function $\psi(X, Y_{cts})$.

The complete case analysis (CCA) consistently estimates the odds ratio among those with R=1. To derive the odds ratio that a CCA consistently estimates, we find an expression for $P(Y_{bin}=1 \mid X=x, R=1)$:

$$P(Y_{bin} = 1|X = x, R = 1) = \int P(Y_{bin} = 1, Y_{cts} |X = x, R = 1)\, dY_{cts}$$

$$= \int P(Y_{bin} = 1|Y_{cts}, X = x, R = 1) f(Y_{cts}|X = x, R = 1) dY_{cts}$$

$$= \int P(Y_{bin} = 1|Y_{cts}) f(Y_{cts}|X = x, R = 1) dY_{cts} \qquad (1)$$

We now expand the second term in the integral as

$$f(Y_{cts}|X = x, R = 1) = \frac{f(Y_{cts}, X = x, R = 1)}{P(X = x, R = 1)}$$

$$= \frac{P(R = 1|Y_{cts}, X = x) f(Y_{cts}|X = x) P(X = x)}{P(R = 1|X = x) P(X = x)}$$

$$= \frac{\psi(x, Y_{cts}) f(Y_{cts}|X = x)}{\int \psi(x, y_{cts}) f(y_{cts}|X = x) dy_{cts}} \qquad (2)$$

Substituting (2) in to (1), we have

$$P(Y_{bin}=1|X=x,R=1)=\int P(Y_{bin}=1|Y_{cts})f(Y_{cts}|X=x,R=1)dY_{cts}$$
$$=\frac{\int P(Y_{bin}=1|Y_{cts})\psi(x,Y_{cts})f(Y_{cts}|X=x)dY_{cts}}{\int \psi(x,Y_{cts})f(Y_{cts}|X=x)dY_{cts}} \quad (3)$$

From this quantity, given specifications for the relationship between the binary outcome and continuous outcome ($P(Y_{bin}=1|Y_{cts})$), for example a logistic model, the distribution of the continuous outcome conditional on the exposure ($f(Y_{cts}|X=x)$), for example normal, and the missingness/observation function $\psi(x,Y_{cts})$, the odds of Y_{bin}=1 in the complete cases among those with X=x can be obtained for a given value of x. From this the odds ratio comparing X=1 to X=0 can be calculated. We note that the expression given in (3) above, and hence also the CCA odds ratio, does not depend on the marginal distribution of the exposure (i.e. the prevalence if X is binary).

To calculate the full population odds ratio the preceding expression can be used, setting

$$P(R=1|X,Y_{cts})=\psi(X,Y_{cts})=1$$

which gives

$$P(Y_{bin}=1|X=x)=\int P(Y_{bin}=1|Y_{cts})f(Y_{cts}|X=x)dY_{cts} \quad (4)$$

Probability of being a complete case depending independently on exposure and continuous outcome

Now consider the special case where the probability of being observed depends independently on X and Y_{cts} such that

$$P(R=1|X,Y_{cts})=\psi(X,Y_{cts})$$
$$=p_1(X)p_2(Y_{cts})$$

for some functions $p_1(X)$ and $p_2(Ycts)$.

Then substituting into (3) we obtain

$$P(Y_{bin}=1|X=x,R=1)=\frac{\int P(Y_{bin}=1|Y_{cts})p_1(x)\,p_2(Y_{cts})f(Y_{cts}|X=x)dY_{cts}}{\int p_1(x)\,p_2(Y_{cts})f(Y_{cts}|X=x)dY_{cts}}$$
$$=\frac{\int P(Y_{bin}=1|Y_{cts})\,p_2(Y_{cts})f(Y_{cts}|X=x)dY_{cts}}{\int\ p_2(Y_{cts})f(Y_{cts}|X=x)dY_{cts}} \quad (5)$$

Note the $p_1(x)$ term has cancelled, implying that the CCA OR for the exposure effect does not depend on the form of $p_1(X)$.

1.3 Formulae used to calculate the bias in the complete case log odds ratio using the above expressions and in the simulation study

The following were substituted into the equations in 1.2 above to calculate the bias in the CCA log odds ratio; they were also used to generate the simulated datasets.

i. $f(Y_{cts}|X = x)$

The continuous depression score for individual i was generated conditional on smoking such that:

$$\text{Depression score}_i = \beta_0 + \beta_1 \times (\text{smoke\_preg}_i) + \varepsilon_i \quad (6)$$

where *smok_preg* is maternal smoking in pregnancy, coded 0/1, and ε is error, following a normal distribution with mean 0 and variance σ^2, calculated to give the score a variance of 1 marginally.

ii. $P(Y_{bin} = 1|Y_{cts})$

The binary depression measure was assumed to depend on the depression score via logistic regression (Equation 7):

$$\text{logit}(\text{p\_deps}_i) = \alpha_0 + \alpha_1 \times (\text{depression score}_i) \quad (7)$$

where p_deps_i represents the probability that an individual was classified as having depression (in the study data) and with α_0 = -6.9875 and $\alpha_1 = 6.5$, chosen using trial and error to give a prevalence of 15% and such that this logistic function was very steep (Supplementary Figure S1) – i.e. generating a strong relationship between the depression score and the binary depression measure.

The analysis model is given by Equation 8.

$$\text{logit}(\text{p\_deps}_i) = \mu_0 + \mu_1 \times (\text{smoke\_preg}_i) \quad (8)$$

The regression coefficient β_1 for maternal smoking in pregnancy from Equation 6 was set at 0.2317; this was chosen by trial and error to give a log OR for depression of 0.405 (to 3 decimal places) which gave an OR of 1.50 comparing those whose mother smoked to those

whose mother did not smoke during pregnancy). The prevalence of exposure (maternal smoking) was set at 25%; thus, β_0 and σ in Equation 6 were given by:

$\beta_0 = 0 - (0.2317 \times 0.25)$ to give the depression score a mean of 0 and

$\sigma = \sqrt{1 - (0.2317)^2 \times 0.25 \times 0.75)}$ to give it a variance of 1.

(For simulations only): The linked (binary) GP measure of depression was created – using a logistic function – to give different sensitivities in relation to the study's binary measure (Equation 9).

$$\text{p\_GPdep}_i = \frac{1}{1 + \exp(\rho \times (\text{depression score}_i - \theta))} \quad (9)$$

Values of ρ and θ were chosen (using trial and error) to give sensitivities of 25% and 75% and a specificity of 97.5%.

Probability of being a complete case dependent independently on exposure and continuous outcome

iii. $p_1(X)$ and $p_2(Y_{cts})$

As noted in 1.2, the expression for the complete case estimate of the exposure odds ratio does not depend on the form of $p_1(X)$. Thus, we focused only on $p_2(Y_{cts})$: this probability (of the outcome being observed) was assumed to depend on the continuous outcome via logistic regression (Equation 10).

$$\text{logit}(p_{2i}) = \gamma_0 + \gamma_1 \times \text{depression score}_i \quad (10)$$

To show how the bias varied as the strength of association between the continuous outcome and the probability of the outcome being observed and the percentage of missing data varied, we calculated the values of γ_0 that gave a given percentage of missing data for different values of γ_1 [for γ_1= ln(0.90), ln(0.75), ln(0.50) and ln(0.25)]. In the simulations, γ_1 was fixed as ln(0.75).

Simulations only: probability of being observed dependent multiplicatively on the exposure, continuous outcome and their interaction

The probability of being observed were generated from the logistic model shown in Equation 11, so that the logarithm of the probability of the outcome being observed (being a complete case) depended on exposure, outcome and their interaction. The values of τ_0 were chosen using trial and error to produce given percentages of missing data. In these scenarios with an interaction, τ_1, τ_2, and τ_3 were fixed at ln(0.7), ln(0.9), and ln(1.1), respectively. Note that this interaction on the logit scale implies a multiplicative interaction between the exposure and outcome with respect to the probability of being observed, such that a complete case analysis is not expected to be (asymptotically) unbiased.

$$\begin{aligned}\text{logit}(\text{P}(\text{observed})_i) &= \tau_0 + \tau_1 \times \text{smoke\_preg}_i + \tau_2 \times \text{depression score}_i \\ &+ \tau_3 \times \text{depression score}_i \times \text{smoke\_preg}_i\end{aligned} \quad (11)$$

2. Linkage to GP data

As part of the Secure Anonymised Information Linkage (SAIL) project [1], the NHS Wales Information Service (NWIS) and the Health Informatics Research Unit (HIRU) at the University of Swansea have established a method through which individual level data from multiple sources can be linked and analysed in a secure setting, including data from primary care electronic patient records. ALSPAC, working with the SAIL team, developed two methods to extract GP records which took advantage of the SAIL infrastructure:

Pilot extraction: In 2012 ALSPAC carried out a pilot extraction which included only individuals who had provided explicit consent. The methods for this extraction have been described in a previous paper [2].

Main extraction: The NHS South West Commissioning Support Unit (SWCSU) has developed a governance framework and data extraction mechanism which secured opt-in assent from GP practices for the extraction of records and their use for SWCSU approved purposes. Invitations to participate in this system were made to all practices in the Bristol, North Somerset, Somerset and South Gloucestershire (BNSSSG) clinical commissioning group. The extraction mechanism is provided by EMIS, which supplies software systems to the majority of practices in the BNSSSG area. ALSPAC gained approval from the SWCSU Security and Informatics Group to extract participants' GP records. SWCSU informed all participating practices about this agreement and gave them opportunity to opt-out.

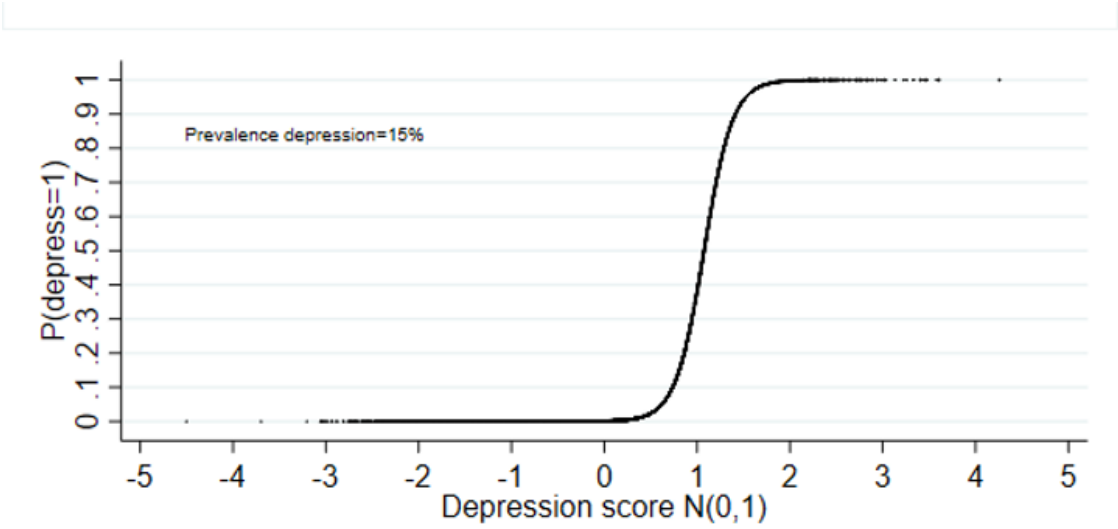
For both the pilot study and the main extraction, the methods after extraction were identical. The extracted records were pseudonymised and securely transferred into the infrastructure at Swansea University using SAIL's "split file" method and adhered to NHS standards of encryption and security, as described previously [2].

References

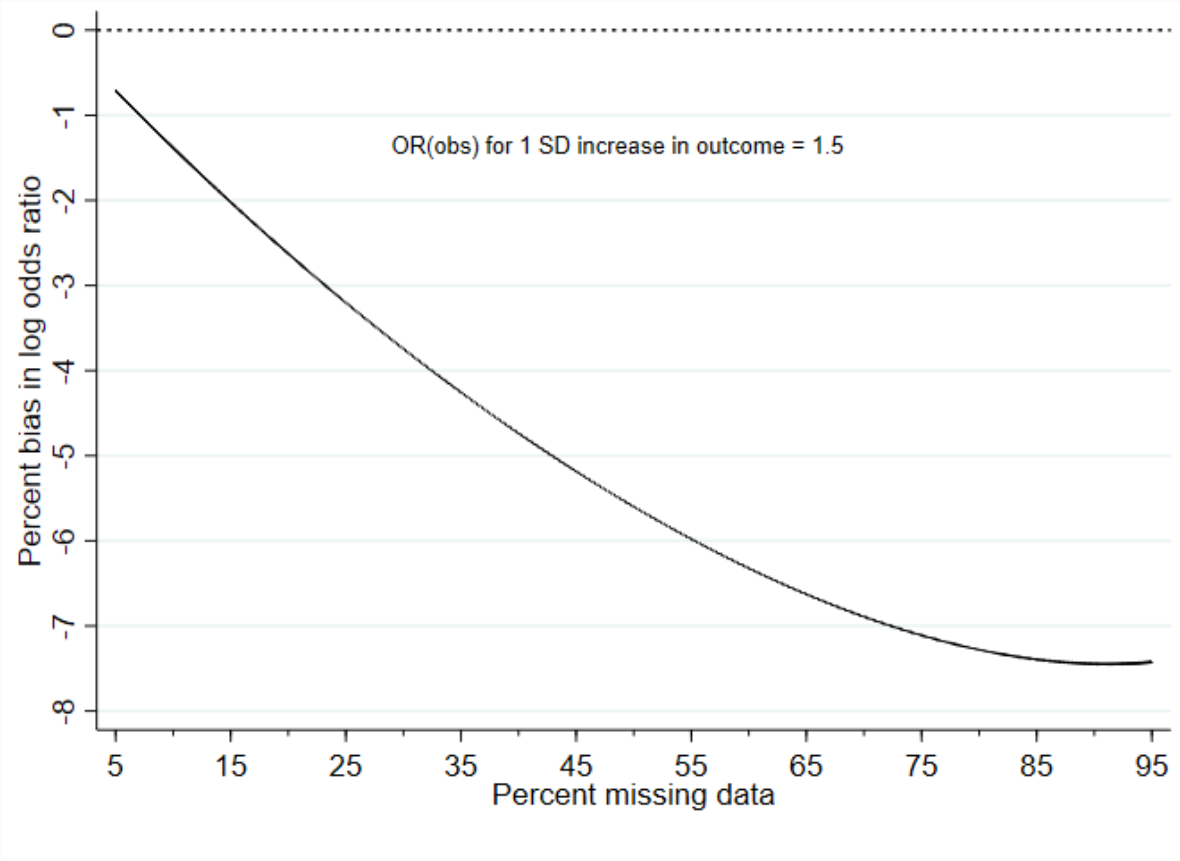
1. Ford, D.V., et al., *The SAIL Databank: building a national architecture for e-health research and evaluation.* BMC Health Serv Res, 2009. **9**: p. 157.
2. Cornish, R.P., et al., *Defining adolescent common mental disorders using electronic primary care data: a comparison with outcomes measured using the CIS-R.* BMJ Open, 2016. **6**(12).

3. Additional figures and tables

Supplementary Figure S1: Simulated relationship between the continuous depression score and the probability of depression (p_deps) being equal to 1 for a prevalence of 15%.



Supplementary Figure S2: Percent bias in log odds ratio from a complete case analysis for a full population exposure log odds ratio of 0.405 when the probability of being observed only depends on the continuous outcome and this probability increases as the continuous outcome increases



Supplementary Table S1: Simulation results - complete case and MI estimates of the log odds ratio (full population log odds ratio = 0.405) when the probability of being observed depended only on the continuous outcome

Factor 1: % missing	Complete case		Factor 2: sensitivity	MI			
	Mean estimate (empirical SE)	% bias (mcse)		Mean estimate (empirical SE)	% bias (mcse)	Gain in precision	FMI
20%	0.409 (0.070)	0.9% (0.5%)	25	0.410 (0.069)	1.2% (0.5%)	2%	19%
			75	0.409 (0.064)	1.0% (0.5%)	8%	11%
40%	0.414 (0.084)	2.2% (0.7%)	25	0.415 (0.080)	2.4% (0.6%)	5%	38%
			75	0.414 (0.070)	2.1% (0.6%)	21%	23%
60%	0.416 (0.110)	2.8% (0.8%)	25	0.420 (0.102)	3.8% (0.8%)	8%	58%
			75	0.414 (0.081)	2.1% (0.8%)	37%	41%
80%	0.422 (0.162)	4.2% (1.3%)	25	0.422 (0.148)	4.2% (1.2%)	8%	78%
			75	0.412 (0.109)	1.7% (0.9%)	50%	65%

Supplementary Table S2: Simulation results - complete case and MI estimates of the log odds ratio (full population log odds ratio = 0.405) when the probability of being observed depended multiplicatively on exposure, continuous outcome and their interaction

Factor 1: % missing	Complete case		Factor 2: sensitivity	Factor 4: 25% missing in linked variable	MI			
	Mean estimate (empirical SE)	% bias (mcse)			Mean estimate (empirical SE)	% bias (mcse)	Gain in precision	FMI
20%	0.432 (0.072)	7% (0.5%)	25	No	0.426 (0.070)	5% (0.5%)	3%	20%
			75	No	0.409 (0.068)	1% (0.5%)	6%	11%
			75	Yes	0.417 (0.070)	3% (0.5%)	4%	14%
40%	0.465 (0.082)	15% (0.6%)	25	No	0.453 (0.078)	12% (0.6%)	5%	41%
			75	No	0.418 (0.074)	3% (0.6%)	11%	25%
			75	Yes	0.432 (0.077)	7% (0.6%)	7%	30%
60%	0.505 (0.107)	25% (0.8%)	25	No	0.485 (0.102)	20% (0.8%)	5%	62%
			75	No	0.427 (0.083)	5% (0.7%)	29%	43%
			75	Yes	0.443 (0.088)	9% (0.7%)	21%	48%
80%	0.544 (0.155)	34% (1.2%)	25	No	0.517 (0.146)	28% (1.1%)	6%	81%
			75	No	0.437 (0.109)	8% (0.9%)	42%	67%
			75	Yes	0.453 (0.116)	12% (0.8%)	34%	71%

Supplementary Table S3: Characteristics of the ALSPAC-enrolled sample and complete cases

Characteristic		Enrolled singletons and twins, alive at one year, not subsequently withdrawn (n=14,566)[1]	Complete cases (n=2,718)	Those with GP data needed to measure depression (n=10,560)
Sex	Male	7,645 (51%)	802 (43%)	5,297 (50%)
	Female	7,902 (49%)	1,067 (57%)	5,263 (50%)
Maternal age	<20	647 (5%)	29 (1%)	495 (5%)
	20-24	2,679 (19%)	312 (11%)	1,904 (19%)
	25-29	5,358 (39%)	1,038 (38%)	3,896 (39%)
	30-34	3,809 (27%)	979 (36%)	2,758 (28%)
	35+	1,371 (10%)	360 (13%)	953 (10%)
Parity	0	5,728 (45%)	1,346 (50%)	4,104 (44%)
	1	4,491 (35%)	796 (29%)	3,272 (35%)
	2+	2,601 (20%)	576 (21%)	1,895 (20%)
Smoking in pregnancy	No	7,645 (68%)	2,308 (85%)	5,568 (68%)
	Yes	3,582 (32%)	410 (15%)	2,582 (32%)
Maternal education	O level/lower	7,967 (65%)	1,346 (50%)	5,903 (66%)
	A level	2,766 (22%)	796 (29%)	1,958 (22%)
	Degree/higher	1,579 (13%)	576 (21%)	1,037 (12%)
Family occupational social class	Non-manual	9,184 (81%)	2,433 (90%)	6,581 (80%)
	Manual	2,222 (19%)	285 (10%)	1,679 (20%)
Housing tenure	Mortgaged/owned	9,473 (73%)	2,413 (89%)	6,952 (74%)
	Private rented	921 (7%)	92 (3%)	555 (6%)
	Other	2,523 (20%)	213 (8%)	1,828 (20%)
Number of rooms in home	Median (IQR)	5 (4-6) [n=12,786]	5 (4-6)	5 (4-6) [n=9,244]
Maternal depression score (EPDS)	Median (IQR)	6 (3-10) [n=11,875]	6 (3-9)	6 (3-10) [n=8,631]
Paternal depression score (EPDS)	Median (IQR)	3 (1-6) [n=9,614]	3 (1-6)	3 (1-6) [n=6,957]
Maternal anxiety score	Median (IQR)	4 (2-7) [n=11,945]	4 (2-6)	4 (2-7) [n=8,679]
Paternal anxiety score	Median (IQR)	2 (1-4) [n=9,564]	2 (1-4)	2 (1-5) [n=6,909]

1. Denominators vary because the variables come from different questionnaires and have different completion rates.

Supplementary Table S4: Predictors of the odds of observing ALSPAC-measured depression: covariates (n= 7,027 with complete covariate data)

Factor	Level	OR (95% CI)[1]	p-value
Smoking in pregnancy	Yes vs no	0.64 (0.56, 0.73)	p<0.001
Sex	Female vs male	1.59 (1.45, 1.75)	p<0.001
Mother's education	O level/lower	1.00	
	A level	1.43 (1.27, 1.61)	
	Degree	1.59 (1.35, 1.85)	p<0.001
Mother's age at birth	<20	1.00	
	20-24	1.82 (1.19, 2.86)	
	25-29	1.96 (1.28, 3.03)	
	30-34	2.56 (1.64, 3.85)	
	35+	2.78 (1.79, 4.35)	p<0.001
Parity	0	1.00	
	1	0.78 (0.69, 0.88)	
	2+	0.61 (0.52, 0.71)	p<0.001
Maternal depression	Per 1 point increase	1.00 (0.99, 1.02)	p=0.7
Maternal anxiety	Per 1 point increase	0.99 (0.97, 1.01)	p=0.3
Paternal depression	Per 1 point increase	0.99 (0.97, 1.01)	p=0.4
Paternal anxiety	Per 1 point increase	1.02 (1.00, 1.04)	p=0.1
Housing tenure	Mortgaged /owned	1.00	
	Private rented	0.51 (0.39, 0.65)	
	Council/HA/other	0.75 (0.62, 0.90)	p<0.001
Number of rooms	Per 1 room increase	1.09 (1.04, 1.14)	p<0.001
Family occupational social class	Manual vs non-manual	0.80 (0.68, 0.93)	p=0.005