Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Follow-up duration, missing data and recording bias in routine primary- care pregnancy records: a data-quality case study of spontaneous abortion in 89,772 SINA records from Mashhad, Iran

Veda Vakili[1]

vakiliv@mums.ac.ir

Mashhad University of Medical Sciences

Navid Eshaghi Moghadam[2]

Mashhad University of Medical Sciences

Ehsan mousa Farkhani[3]

Mashhad University of Medical Sciences

Research Article

Keywords: spontaneous abortion, routinely collected data, electronic health records, data quality, health information systems, missing data, information bias, Iran

Posted Date: November 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11135935/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Routine primary-care information systems are attractive for pregnancy research, but their records are generated during care, and pregnancies that end early leave fewer records. We assessed how follow-up duration, missing data and comparison-group definition affect risk-factor estimates for spontaneous abortion in the SINA system of Mashhad University of Medical Sciences, Iran.

Methods

We analysed 89,772 unique pregnancy records from the project year 1397 (March 2018–March 2019) (5,561 with recorded spontaneous abortion; 84,211 with live fetal status). A naive analysis using all recorded variables was compared with a structure-aware analysis restricted to entry-time variables. We evaluated supplement, laboratory and psychosocial-score records as markers of follow-up, compared three missing-data strategies for BMI, bounded the effect of the heterogeneous comparison group with restrictions to women with advanced or limited follow-up, quantified centre-level recording completeness, and classified the stability of associations across seven analyses.

Results

In the naive analysis iron appeared almost completely protective (OR 0.005) and folic acid strongly harmful (OR 9.49); 85% of women with only folic acid recorded had an abortion versus 0.5% with only iron. Adding supplement records to an entry-data model raised the cross-validated AUC from 0.713 to 0.937. BMI was missing in 2.4% of cases but 18.2% of comparison women, and odds ratios for one living child ranged from 2.85 to 3.52 across missing-data strategies. Only diabetes (OR 1.56, 95% CI 1.23–1.98) and university education (0.81, 0.72–0.90) were stable in direction and size across all analyses. Age < 20 years (2.27, 2.02–2.55), recurrent abortion (1.92, 1.63–2.27), stillbirth (1.38, 1.07–1.79) and living children were stable in direction, but their size depended strongly on how the comparison group was defined, and living children were opposite in direction to previous studies. Age ≥ 40 years, obesity, hard job and smoking changed direction across analyses, and the association of underweight (1.30, 1.15–1.48) depended on assumptions about missing BMI.

Conclusions

Without dates, gestational age and final pregnancy outcome, routine records support only a limited set of entry-time risk factors; variables recorded during follow-up give spurious associations. We propose rules for analysts and minimum data requirements for SINA.

Background

Spontaneous abortion affects about 15–20% of recognised pregnancies [1]; a systematic review estimated the pooled risk at 15.3% (95% CI 12.5–18.7%) [2]. Register-based studies show that maternal age and previous pregnancy history are its major determinants [3, 4], and nationwide cohorts have sought its modifiable risk factors [5], while evidence for other factors, including body mass index (BMI) [6, 7, 8], smoking [9], physically demanding work [10] and vitamin supplementation [11], is less consistent. In Iran, matched case–control studies have identified predictors of miscarriage [12], but large-scale analyses of routine primary-care information systems are scarce.

Since 2016 the health centres of Mashhad University of Medical Sciences have recorded maternal care in the integrated health information system SINA, a resource covering tens of thousands of pregnancies. Such systems are increasingly used for research because they are large and inexpensive, but their fitness for research is task-dependent and can be examined along dimensions such as completeness, correctness, concordance, plausibility and currency [13]. However, they document care, not the natural history of pregnancy: a pregnancy that ends after a first visit leaves a short record, whereas a pregnancy that continues accumulates supplements, laboratory tests and assessments. Analyses that ignore this structure can generate strong but spurious associations, a form of immortal-time bias [14], and missing data may depend on the outcome [15]. More generally, inclusion and completeness of records in electronic health systems are not random but depend on how patients interact with the system [16, 17, 18, 19], and conditioning on pregnancy outcomes can itself induce selection bias [20].

We therefore asked a methodological question, using recorded spontaneous abortion in SINA as a case study: how do follow-up duration, missing data, the definition of the comparison group and centre-level recording completeness shape risk-factor estimates, and which associations survive when these features are addressed?

Methods

Data source and reporting

The analysis used routinely collected data from SINA, completed by midwives and health-care providers at health centres covered by Mashhad University of Medical Sciences (research project No. 971136). The authors had access to the full SINA extract for the study period, obtained through the Vice-Chancellor for Health, with no further sampling; the study used a single database, so no linkage across databases was performed. Reporting follows RECORD, the extension of STROBE for routinely collected data [21, 22].

Records, cases and comparison group

The registry variable “fetal status” classifies each record as aborted or alive; according to the SINA team, “alive” means that the fetus was alive at the last recorded prenatal-care visit. Cases were records with a

spontaneous abortion during the project year 1397 (March 2018–March 2019). The comparison group comprised all other records with a live fetal status from the same year. The extract contains no dates, gestational age or final pregnancy outcome, so the comparison group includes pregnancies that were still in progress at their last recorded visit as well as completed pregnancies. The study size is the entire 1397 extract, not a pre-calculated sample: of 90,504 records, 732 belonged to 699 duplicated member identifiers (41 case and 691 comparison records); the first record per identifier was retained, leaving 89,772 records (Fig. 1). Odds ratios are interpreted as associations with recorded abortion, not as risks.

Variables

Entry-time variables (recorded at or before pregnancy registration) were age, education, nationality, number of living children (recorded during prenatal care and excluding the current pregnancy), recurrent abortion (history of recurrent abortion as recorded in SINA), previous stillbirth, BMI, categorised using standard WHO cut-offs (< 18.5, 18.5–24.9, 25.0–29.9, ≥ 30 kg/m²; pre-pregnancy or at registration; values outside 12–60 kg/m² set to missing), diabetes, hypertension, lifetime smoking, hard job (more than 40 working hours per week), walking, exercise (35% recorded) and blood group (codes 1–8). Gravidity was not used because of inconsistent recording (Results). Follow-up-time variables were tablets of iron, folic acid, iodo-folic acid, multivitamin and vitamin D, haemoglobin, blood pressure, and the psychological assessment (PAR), domestic-violence (HITS) and nutrition scores, which are generated during prenatal visits.

Analytic strategy

We contrasted two approaches. The naive analysis, which mimics common practice in analyses of registry data, entered all recorded variables, including follow-up-time variables, with missing values as separate categories. The structure-aware analysis (i) restricted exposures to entry-time variables, (ii) treated missing data explicitly, (iii) used standard errors clustered on health centre (244 centres), and (iv) assessed stability against features of the data structure. Specifically we examined:

1. **(a) Follow-up markers.** The pattern of supplement records and the number of tablets issued in cases and comparison women; associations within strata of haemoglobin (a proxy for gestational stage); the number of assessments contributing to PAR and HITS (identified from fractional values, which are means over visits), with analysis restricted to single assessments; and the cross-validated area under the ROC curve (AUC) of a logistic model using entry-time variables alone, with an indicator of missing BMI, or with follow-up-time records added (10-fold cross-validation).
2. **(b) Missing data.** BMI odds ratios and other estimates from three strategies: a “not recorded” category, complete-case analysis, and multiple imputation (20 imputations by regression with bootstrap-drawn parameters, with outcome, covariates and district in the imputation model, pooled by Rubin’s rules [15, 23]). Because missing-at-random is doubtful, we added delta-adjusted (pattern-

mixture) analyses in which imputed BMI values were shifted by − 3 or + 3 kg/m², following the principle of δ-based controlled multiple imputation [24].

3. **(c) Comparison group.** Because the comparison group mixes pregnancies in progress and completed pregnancies, we bounded its effect with two restrictions that condition on follow-up progress and are therefore bounds, not corrections: comparison women with iron supplementation recorded (advanced follow-up) and comparison women without iron recorded (limited follow-up, more similar to cases), together with analyses stratified by living children. Requiring a minimum amount of recorded data for inclusion can itself bias selection towards women with more follow-up [25].
4. **(d) Centre-level recording.** Random-intercept models for district and centre, the intraclass correlation and median odds ratio [26], case-mix-adjusted observed/expected (O/E) ratios by indirect standardisation, and the correlation between centre O/E and the proportion of women with missing BMI.
5. **(e) Stability classification.** Model 2 (age, education, nationality, living children, recurrent abortion, previous stillbirth, BMI, diabetes, hypertension, smoking, hard job and walking; logistic regression with centre-clustered standard errors) was estimated in six pooled analyses (multiple imputation, complete case, excluding all duplicated identifiers, ages 15–49 years, the comparison group with advanced follow-up, and a random-intercept model) and in a seventh analysis with the limited-follow-up comparison group. An association was classified as stable in direction and size if its direction was identical in all seven analyses, every odds ratio was ≥ 1.10 or ≤ 0.90 and the largest odds ratio was at most 1.6 times the smallest; as stable in direction only if the direction and threshold held but the ratio was larger; as consistent but weak if the direction was identical but the effect smaller; and as not stable otherwise. Effect modification by parity and the non-linear effect of age were examined with restricted splines and likelihood-ratio tests.

Ethics

The study protocol was approved by the Ethics Committee of Mashhad University of Medical Sciences (IR.MUMS.MEDICAL.REC.1398.128). Analyses used routinely collected records that were anonymised before analysis.

Use of AI tools

During the preparation of this manuscript, the authors used ChatGPT (OpenAI) solely to assist with language editing, clarity, grammar, and organization of the manuscript. The authors independently reviewed and verified all AI-assisted content and remain fully responsible for the accuracy, originality, interpretation, and final content of the manuscript. Generative AI was not used to generate, analyze, or interpret the study data, statistical results, or scientific conclusions.

Results

Study population

Of 90,504 records, 89,772 were analysed (5,561 cases; 84,211 comparison women; Fig. 1). Cases were older (median 30 vs 29 years), more often had living children, recurrent abortion or stillbirth, and had BMI recorded far more often (97.6% vs 81.8%) (Table 1). Because the 6.2% abortion prevalence in this extract reflects how records were selected for the extract rather than a population incidence, we report odds ratios throughout and do not translate them into absolute risks.

Table 1
Characteristics of women with recorded spontaneous abortion and comparison women

A naive analysis produces implausible results

When supplement records were entered as exposures, iron supplementation appeared almost completely protective (OR 0.005, 95% CI 0.005–0.006) and folic acid strongly harmful (9.49, 8.28–10.86); iodo-folic tablets (6.01) and multivitamins (0.31) showed similar extreme associations (Table 2). A Cochrane review of 40 randomised trials found that vitamin supplementation did not reduce miscarriage [11], and these estimates are not biologically credible.

Table 2
Naive analysis of supplement records as exposures

Supplement	% recorded, cases / comparison	Naive adjusted OR (95% CI)	Interpretation
Iron tablets	17.6 / 87.4	0.005 (0.005–0.006)	Issued later in follow-up: implausibly “protective”
Folic acid tablets	66.1 / 29.3	9.49 (8.28–10.86)	Issued at first visit: implausibly “harmful”
Iodo-folic tablets	16.0 / 5.0	6.01 (5.14–7.03)	Issued early in follow-up
Multivitamin tablets	1.2 / 17.4	0.31 (0.25–0.39)	Issued later in follow-up
Vitamin D 1000 IU	48.8 / 51.3	1.69 (1.46–1.96)	Weak, mixed timing

Adjusted for the covariates of Model 2 plus exercise (with a “not recorded” category); standard errors clustered on health centre.

Follow-up records mark how far a pregnancy progressed

The pattern of supplement records was almost equivalent to the outcome (Fig. 2A; Additional file 1, Table S3): 85.0% of women with only folic acid recorded had an abortion, versus 14.0% with neither, 3.0% with both and 0.5% with only iron. According to the national maternal care protocol, folic acid is recommended from before conception or from the first prenatal visit, whereas iron is usually started at about week 16 [27]; a pregnancy lost before then has no opportunity for iron supplementation, so the pattern shows how far follow-up progressed. Cases who received iron had received a median of 50 tablets (IQR 30–80), versus 120 (60–180) for comparison women. Stratifying by haemoglobin, a proxy for gestational stage, did not remove the associations (iron OR 0.006–0.013 across strata). Similarly, PAR and HITS are means of one to several assessments, and cases were much more likely to have only a single assessment (HITS 78% vs 62%; PAR 79% vs 46%) or no PAR at all (28.5% vs 3.8%). Restricting to a single assessment attenuated the HITS association (OR per SD from 1.22 to 1.16) but not the PAR association (1.26 to 1.33; Additional file 1, Table S2).

Recording of gravidity showed a similar inconsistency: among women without living children, 77% of comparison women but only 18% of cases had a gravidity of 0, and 63% of cases versus 18% of comparison women had a gravidity of 1, whereas among women with one living child the count was mostly 2 in both groups. The current pregnancy therefore appears to be counted in most records but not consistently entered for nulliparous comparison women, and gravidity was excluded from all models. These markers can produce near-perfect discrimination: a model using entry-time variables alone had a cross-validated AUC of 0.713; adding an indicator of missing BMI raised it to 0.728; adding supplement records raised it to 0.937, and further indicators of haemoglobin, PAR and HITS recording did not change it (0.938; Fig. 2B). Follow-up records therefore leak information about the outcome and must be excluded from predictors and exposures unless their dates are known.

Handling of missing data

BMI was missing for 15,463 women: 132 cases (2.4%) versus 15,331 comparison women (18.2%), so 99.1% of women without BMI were comparison women. Women without BMI had odds of abortion of 0.21 (0.17–0.25) relative to women with normal BMI in the “not recorded” category analysis. Estimates for BMI, recurrent abortion and diabetes were similar across these three strategies, but estimates for living children depended on the strategy (Table 3): the odds ratio for one living child was 2.85 with a “not recorded” category or complete cases and 3.52 after multiple imputation, and for two children 3.70, 3.73 and 4.62. Because missingness was so strongly related to the outcome, missing-at-random is doubtful, and we report the range. Delta-adjusted analyses confirmed that this matters for BMI (Additional file 1, Table S6): the odds ratio for underweight ranged from 1.02 (imputed values shifted by − 3 kg/m²) to 1.37 (+ 3 kg/m²) and for obesity from 1.21 to 1.00, whereas odds ratios for living children, recurrent abortion and diabetes did not change.

Table 3
Adjusted odds ratios (95% CI) under three missing-data strategies

Factor	"Not recorded" category (n = 89,772)	Complete case (n = 74,309)	Multiple imputation (n = 89,772)
Age < 20	2.01 (1.79–2.26)	2.02 (1.79–2.27)	2.27 (2.02–2.55)
Age ≥ 40	1.68 (1.48–1.90)	1.64 (1.44–1.86)	1.58 (1.39–1.78)
1 living child	2.85 (2.64–3.07)	2.85 (2.64–3.08)	3.52 (3.26–3.81)
2 living children	3.70 (3.40–4.03)	3.73 (3.42–4.06)	4.62 (4.24–5.04)
Recurrent abortion	1.85 (1.57–2.19)	1.86 (1.58–2.20)	1.92 (1.63–2.27)
BMI < 18.5	1.31 (1.15–1.48)	1.30 (1.15–1.48)	1.30 (1.15–1.48)
BMI ≥ 30	1.14 (1.05–1.23)	1.14 (1.05–1.23)	1.13 (1.04–1.22)
Diabetes	1.46 (1.15–1.85)	1.47 (1.16–1.87)	1.56 (1.23–1.98)
Hard job	1.06 (0.96–1.18)	1.07 (0.96–1.19)	1.13 (1.01–1.25)
Non-Iranian	0.73 (0.63–0.84)	0.72 (0.62–0.83)	0.72 (0.63–0.83)
BMI not recorded (vs 18.5–24.9)	0.21 (0.17–0.25)	—	—

Model 2; standard errors clustered on health centre.

Heterogeneity of the comparison group

Restricting the comparison group to women with iron recorded (advanced follow-up, n = 61,527) left the odds ratios similar (living children 2.48, 3.20 and 2.71). Restricting it instead to women without iron recorded (limited follow-up, n = 7,353) changed them dramatically: living children 17.9, 29.2 and 30.7, recurrent abortion 6.75, stillbirth 2.71 and age < 20 years 3.38, while age ≥ 40 years (0.61) and obesity (0.92) reversed direction; diabetes (1.56), university education (0.81) and underweight (1.91) kept their direction and size (Additional file 1, Table S7). The choice of comparison group therefore matters greatly for parity and pregnancy history, and results for these variables depend on the unknown gestational stage of the comparison women. The women in the limited-follow-up comparison group were almost all nulliparous (94%, versus 39% of cases and 68% of the advanced-follow-up group), 95% had neither iron nor folic acid recorded and 88% had no haemoglobin (Additional file 1, Table S8); this bound is therefore extreme and probably captures very early or minimally documented first pregnancies rather than a representative early-gestation group. Age and recurrent abortion differed by parity (interaction $p < 0.001$, Additional file 1, Figure S1): among women without living children, odds were increased in adolescents (OR 2.24 at 18 years) but not at older ages, and recurrent abortion had a large association (OR 5.60); among women with children, adolescents did not differ (0.96 at 18 years), odds rose steeply after 35 years (2.17 at 40 years), and recurrent abortion was weaker (1.25).

Centre-level recording

After adjustment for case mix, between-centre variation was moderate (SD of random intercepts 0.29 for district and 0.26 for centre; ICC 4.4%; median odds ratios 1.32 and 1.28). Standardised O/E ratios ranged from 0.78 to 2.18 across districts (Additional file 1, Table S4) and, among 118 centres with ≥ 300 records, 36 (31%) fell outside the 95% funnel limits (Fig. 3). Centre O/E was negatively correlated with the proportion of missing BMI (Spearman ρ = −0.43): centres that record less record fewer abortions. Adjusting for indicators of missing data left the variance essentially unchanged (ICC 4.5%), so between-centre differences cannot be interpreted as differences in risk. Similarly, the lower odds of abortion in non-Iranian women (0.72) were mostly confined to Mashhad city centres (96% of non-Iranian women vs 62% of others), and moved to 0.83 (0.70–0.99) after adjusting for centre and recording completeness.

Which associations are stable?

Two associations were stable in direction and size in all seven analyses (Table 4, Fig. 4): diabetes (OR 1.56, 1.23–1.98; range 1.46–1.56) and university education (0.81, 0.72–0.90; range 0.80–0.85). Recorded hypertension (0.81; 88 exposed cases), non-Iranian nationality (0.72; partly explained by recording) and walking (1.29; meaning unclear) were also stable, with the caveats indicated. Age < 20 years (2.27, 2.02–2.55), one, two, or three or more living children (3.52, 4.62 and 3.98), recurrent abortion (1.92, 1.63–2.27) and previous stillbirth (1.38, 1.07–1.79) were stable in direction, but their size depended strongly on the comparison group (Table S7). Underweight (1.30, 1.15–1.48) was stable across comparison definitions but depended on assumptions about missing BMI (Additional file 1, Table S6). Age 35–39 years and ≥ 40 years, overweight, obesity, hard job and lifetime smoking were not stable. Blood group was not associated with abortion (Additional file 1, Table S5).

Table 4
Stability of associations across seven analyses

Factor	Model 2, multiple imputation OR (95% CI)	Range of ORs across analyses	Classification
Age < 20	2.27 (2.02–2.55)	1.90–3.38	Stable in direction; size depends on comparison group
Age 35–39	1.00 (0.90–1.11)	0.55–1.11	Not stable (direction changes)
Age ≥ 40	1.58 (1.39–1.78)	0.61–1.78	Not stable (direction changes)
1 living child	3.52 (3.26–3.81)	2.48–17.94	Stable in direction; size depends on comparison group
2 living children	4.62 (4.24–5.04)	3.20–29.18	Stable in direction; size depends on comparison group
≥ 3 living children	3.98 (3.52–4.49)	2.71–30.72	Stable in direction; size depends on comparison group
Recurrent abortion	1.92 (1.63–2.27)	1.82–6.75	Stable in direction; size depends on comparison group
Previous stillbirth	1.38 (1.07–1.79)	1.34–2.71	Stable in direction; size depends on comparison group
BMI < 18.5	1.30 (1.15–1.48)	1.29–1.91	Stable in direction and size across comparison definitions; sensitive to assumptions about missing BMI (Table S6)
BMI 25–29.9	1.08 (1.01–1.15)	0.86–1.10	Not stable (direction changes)
BMI ≥ 30	1.13 (1.04–1.22)	0.92–1.17	Not stable (direction changes)
Diabetes	1.56 (1.23–1.98)	1.46–1.56	Stable in direction and size
Hypertension	0.81 (0.65–1.01)	0.63–0.81	Stable in direction and size; few exposed cases (88)
University education	0.81 (0.72–0.90)	0.80–0.85	Stable in direction and size
Non-Iranian	0.72 (0.63–0.83)	0.71–0.79	Stable in direction and size; partly explained by recording (OR 0.83 after adjustment for completeness)
Hard job	1.13 (1.01–1.25)	0.98–1.13	Not stable (direction changes)

Analyses: multiple imputation; complete case; excluding all duplicated identifiers; ages 15–49; comparison restricted to women with iron recorded (advanced follow-up); random-intercept model; comparison restricted to women without iron recorded (limited follow-up). Stable in direction and size: same direction in all seven, every OR ≥ 1.10 or ≤ 0.90, and largest OR ≤ 1.6 times the smallest; stable in direction only: same, but ratio > 1.6.

Factor	Model 2, multiple imputation OR (95% CI)	Range of ORs across analyses	Classification
Walking	1.29 (1.19–1.39)	1.21–1.29	Stable in direction and size; meaning unclear
Lifetime smoking	1.11 (0.87–1.43)	0.85–1.11	Not stable (direction changes)

Analyses: multiple imputation; complete case; excluding all duplicated identifiers; ages 15–49; comparison restricted to women with iron recorded (advanced follow-up); random-intercept model; comparison restricted to women without iron recorded (limited follow-up). Stable in direction and size: same direction in all seven, every OR ≥ 1.10 or ≤ 0.90, and largest OR ≤ 1.6 times the smallest; stable in direction only: same, but ratio > 1.6.

Discussion

Principal findings

Using 89,772 primary-care pregnancy records, we show that variables generated during follow-up, missing data related to the outcome and differences in centre-level recording can each produce large, sometimes implausible, associations. Analyses that included supplement records reported almost complete protection by iron and a nine-fold higher odds with folic acid, and gave a prediction AUC of 0.94 that would be unattainable from entry data (0.71). When the analysis was restricted to entry-time variables and assessed against alternative handling of missing data, comparison groups and clustering, only two associations were stable in direction and size (diabetes and university education); young age, recurrent abortion, stillbirth and living children were stable in direction but their size depended strongly on how the comparison group was defined, and older age and obesity changed direction.

Stable risk factors in context

The associations that were stable in direction (young age, recurrent abortion, stillbirth) agree with register-based evidence on maternal age and previous losses [3, 4] and with nationwide cohort evidence on risk factors for miscarriage [5]; the excess risk at older ages shown by register data [3, 4] was not stable in our analyses (odds ratio for age ≥ 40 years 0.61 to 1.78 across analyses). A Lancet series review lists age under 20 and over 35 years, very low or very high BMI, previous miscarriages, smoking and night-shift work among the risk factors [2]. In the Norwegian register study, risk was lowest at 25–29 years (10%) and rose rapidly after 30 years, reaching 53% at 45 years or older [3], which parallels the nadir at about 30 years in our data. In the Danish register study, high maternal age was a risk factor irrespective of parity and previous miscarriages [4], whereas in our data the excess at older ages was confined to women with living children, which may reflect the composition of our comparison group. Our odds ratio for previous stillbirth (1.38, 1.07–1.79) is close to that reported after a previous stillbirth (1.30, 1.11–1.53), and our odds ratio for recurrent abortion (1.92) lies between the age-adjusted odds ratios of 1.54 and 2.21 after one and two previous miscarriages [3]. In an Iranian matched case–control study in

Hamadan, the odds ratio for previous spontaneous abortion was 3.43 (2.03–5.79) and the odds increased by 1.58 for every five years of age [12]; an Iranian study using records of urban health centres in Zanjan also reported associations with increasing age, low education and previous spontaneous abortion [28]. The associations with obesity [6] and underweight [7] were in the expected direction, and consistent with a nationwide Chinese cohort of 8.8 million parent–child triads in which underweight, overweight and obesity were each associated with a higher risk of spontaneous abortion [8], but they depended on how missing BMI was handled. In cohort studies, the relative risks of miscarriage were 1.08 (1.05–1.11) for underweight, 1.09 (1.04–1.13) for overweight and 1.21 (1.15–1.27) for obesity [7], and a pooled odds ratio of 1.31 (1.18–1.46) has been reported for obesity in spontaneous conception [29]; our estimate for obesity (1.13) is smaller and that for underweight (1.30) larger, and both varied with assumptions about missing BMI. The interaction between age and parity, an adolescent excess in women without children and a steep increase after 35 years in women with children, warrants confirmation in data with complete outcomes, because the comparison group includes pregnancies in progress. The strong positive association with living children deserves particular caution: in the Hamadan study, whose comparison women had completed live births, each live birth was associated with lower odds of miscarriage (OR 0.20, 0.14–0.28) [12], the opposite direction to ours. Because our comparison group includes pregnancies in progress, which are disproportionately first pregnancies, and because the odds ratios for living children depended on how missing data were handled, we regard this association as the one most likely to reflect data structure rather than biology, and it should not be interpreted as a risk factor. Smoking was not significantly associated with abortion; our point estimate (OR 1.11, 0.87–1.43) lies in the direction of the pooled relative risk of 1.23 (1.16–1.30) for active smoking [9], but only about 2% of women reported smoking (88 exposed cases), which suggests under-reporting and limited power. Diabetes was associated with higher odds, which is biologically plausible. In a prospective study of insulin-dependent diabetes, pregnancy loss was not higher in women with good metabolic control but was higher when first-trimester glucose and glycosylated haemoglobin were elevated [30]; glycaemic control was not recorded in SINA and may explain heterogeneity in this association. Higher education was associated with lower odds (OR 0.81), in line with the Zanjan study [28]. The modest association of hard job (more than 40 working hours per week) with higher odds (OR 1.13) is similar in size to the pooled relative risks of 1.12–1.36 reported for long working hours and other physically demanding activities in a systematic review [10], but it was not stable across our analyses. The association of walking with higher odds is unexpected and probably reflects how it was recorded.

Implications for analysts using routine records

The findings suggest six rules for analysis of routine pregnancy records (Box 1). The problems we describe parallel informed presence bias in electronic health records, in which how a person interacts with the health system determines what is recorded [16, 17, 18, 19]. Requiring sufficient recorded data for inclusion, as in our advanced-follow-up bound, can likewise bias selection [25]. We recommend that they be reported alongside any risk-factor or prediction study based on similar systems.

Implications for SINA

To support valid epidemiology, SINA should store the date of each record and visit, gestational age at entry, the date and outcome of every pregnancy, the date each supplement was dispensed, and a woman-level identifier distinct from the pregnancy record. The dates of psychosocial assessments should be recorded together with their scores, and undocumented codes (blood group codes 9–16) and ambiguous variables (recurrent abortion) should be defined. Centre-level completeness should be monitored, because differences in recording, not in risk, otherwise dominate centre comparisons. With these data, time-dependent analyses (landmark or survival methods) would allow unbiased estimation of the association of supplements and other exposures with pregnancy loss.

Strengths and limitations

Strengths include the large sample, the full comparison population, the explicit comparison of naive and structure-aware analyses, and the assessment of stability across seven analyses and three missing-data strategies. According to the SINA team, the extract is a census of pregnancies registered at public health centres in the MUMS catchment during 1397, not a sample; this supports generalisability within that catchment, but women whose entire pregnancy was managed in the private sector without any public health-centre visit are not covered, and the size of this uncovered group is unknown. Limitations: without dates, gestational age or final outcome we could not perform time-dependent analyses, and the comparison group remains heterogeneous; our restriction to progressed pregnancies is a bounding analysis, not a correction. Missing BMI was strongly outcome-related, and no strategy can fully resolve this; delta-adjusted analyses showed that BMI associations are sensitive to it. The variable for recurrent abortion captured recurrent abortion only, gravidity could not be used, the accuracy of the recorded fetal status was not validated, and many potential determinants (medication, infections, genetic causes) were unavailable. Data come from public health centres in one province, and recurrent abortion was defined only through the codebook. Our stability criterion is descriptive, the seven analyses are not independent because they use the same data, and odds ratios should not be read as causal effects.

Conclusions

In a large Iranian primary-care database, variables recorded during follow-up, outcome-related missing data and recording differences between centres produced spurious and implausible associations with spontaneous abortion. Restricting analyses to entry-time variables showed that only diabetes and higher education were stable in direction and size; young age, recurrent abortion and stillbirth were stable in direction but their size depended strongly on the comparison group, living children were opposite in direction to previous studies and probably reflect data structure, and those for BMI, older age and obesity depended on assumptions about missing data or comparison group. Recording dates, gestational age at entry and final pregnancy outcome would make routine data suitable for valid risk estimation.

Abbreviations

AUC

area under the receiver operating characteristic curve
BMI
body mass index
CI
confidence interval
HITS
domestic violence screening score
ICC
intraclass correlation coefficient
MI
multiple imputation
O/E
observed/expected
OR
odds ratio
PAR
psychological assessment rating
RECORD
REporting of studies Conducted using Observational Routinely-collected health Data
SINA
integrated health information system of Mashhad University of Medical Sciences
STROBE
Strengthening the Reporting of Observational Studies in Epidemiology.

Declarations

Ethics approval and consent to participate: The study protocol was approved by the Ethics Committee of Mashhad University of Medical Sciences (IR.MUMS.MEDICAL.REC.1398.128). The analysis used routinely collected records that were anonymised before analysis. The requirement for informed consent was waived by the Ethics Committee. All methods were carried out in accordance with relevant guidelines and regulations, including the Declaration of Helsinki.

Consent for publication: Not applicable.

Availability of data and materials: The data are held by Mashhad University of Medical Sciences and are not publicly available because they are individual-level health records. They can be made available by the corresponding author on reasonable request and with permission of the university. The analysis code (Python) is available from the corresponding author on request [and/or deposited at a public repository, to be confirmed by the authors].

Competing interests: The authors declare that they have no competing interests.

Funding: This work was supported by Mashhad University of Medical Sciences (research project No. 971136). The funder had no role in the design, analysis or writing of the study.

Authors' contributions: VV conceived and designed the study, performed the statistical analysis, interpreted the results and drafted the manuscript. NEM critically reviewed and revised the manuscript. EF provided access to SINA and extracted the data. All authors read and approved the final manuscript.

Acknowledgements: The authors thank the Vice-Chancellors for Research and for Health of Mashhad University of Medical Sciences and the health-centre staff who entered SINA data.

Previous use of data: The cases in this study were analysed in a postgraduate thesis of Mashhad University of Medical Sciences (thesis No. 5620T) with a random sample of 5,603 comparison women. The present analysis uses all comparison records, additional variables and methods, and addresses a different question.

References

1. Regan L, Rai R. Epidemiology and the medical causes of miscarriage. Baillieres Best Pract Res Clin Obstet Gynaecol. 2000;14(5):839–54.
2. Quenby S, Gallos ID, Dhillon-Smith RK, Podesek M, Stephenson MD, Fisher J, et al. Miscarriage matters: the epidemiological, physical, psychological, and economic costs of early pregnancy loss. Lancet. 2021;397(10285):1658–67.
3. Magnus MC, Wilcox AJ, Morken NH, Weinberg CR, Håberg SE. Role of maternal age and pregnancy history in risk of miscarriage: prospective register based study. BMJ. 2019;364:l869.
4. Nybo Andersen AM, Wohlfahrt J, Christens P, Olsen J, Melbye M. Maternal age and fetal loss: population based register linkage study. BMJ. 2000;320(7251):1708–12.
5. Feodor Nilsson S, Andersen PK, Strandberg-Larsen K, Nybo Andersen AM. Risk factors for miscarriage from a prevention perspective: a nationwide follow-up study. BJOG. 2014;121(11):1375–84.
6. Cavalcante MB, Sarno M, Peixoto AB, Araujo Júnior E, Barini R. Obesity and recurrent miscarriage: a systematic review and meta-analysis. J Obstet Gynaecol Res. 2019;45(1):30–8.
7. Balsells M, García-Patterson A, Corcoy R. Systematic review and meta-analysis on the association of prepregnancy underweight and miscarriage. Eur J Obstet Gynecol Reprod Biol. 2016;207:73–9.
8. Liu B, Zhao C, Yang Y, Li J, Wu Y, Liu C, et al. Parental preconception BMI and risk of adverse birth outcomes in 8 million parent-child triads: a nationwide population-based cohort study in China. Lancet Diabetes Endocrinol. 2025;13(9):754–64.
9. Pineles BL, Park E, Samet JM. Systematic review and meta-analysis of miscarriage and maternal exposure to tobacco smoke during pregnancy. Am J Epidemiol. 2014;179(7):807–23.
10. Bonde JP, Jørgensen KT, Bonzini M, Palmer KT. Miscarriage and occupational activity: a systematic review and meta-analysis regarding shift work, working hours, lifting, standing, and physical

workload. Scand J Work Environ Health. 2013;39(4):325–34.
11. Balogun OO, da Silva Lopes K, Ota E, Takemoto Y, Rumbold A, Takegata M et al. Vitamin supplementation for preventing miscarriage. Cochrane Database Syst Rev. 2016;(5):CD004073.
12. Poorolajal J, Cheraghi P, Cheraghi Z, Ghahramani M, Doosti Irani A. Predictors of miscarriage: a matched case–control study. Epidemiol Health. 2014;36:e2014031.
13. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inf Assoc. 2013;20(1):144–51.
14. Suissa S. Immortal time bias in pharmaco-epidemiology. Am J Epidemiol. 2008;167(4):492–9.
15. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393.
16. Phelan M, Bhavsar NA, Goldstein BA. Illustrating informed presence bias in electronic health records data: how patient interactions with a health system can impact inference. EGEMS (Wash DC). 2017;5(1):22.
17. Goldstein BA, Bhavsar NA, Phelan M, Pencina MJ. Controlling for informed presence bias due to the number of health encounters in an electronic health record. Am J Epidemiol. 2016;184(11):847–55.
18. Goldstein BA, Phelan M, Pagidipati NJ, Peskoe SB. How and when informative visit processes can bias inference when using electronic health records data for clinical research. J Am Med Inf Assoc. 2019;26(12):1609–17.
19. Weiskopf NG, Rusanov A, Weng C. Sick patients have more data: the non-random completeness of electronic health records. AMIA Annu Symp Proc. 2013;2013:1472-7.
20. Liew Z, Olsen J, Cui X, Ritz B, Arah OA. Bias from conditioning on live birth in pregnancy cohorts: an illustration based on neurodevelopment in children after prenatal exposure to organic pollutants. Int J Epidemiol. 2015;44(1):345–54.
21. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med. 2007;4(10):e296.
22. Benchimol EI, Smeeth L, Guttmann A, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Med. 2015;12(10):e1001885.
23. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: Wiley; 1987.
24. Cro S, Morris TP, Kenward MG, Carpenter JR. Sensitivity analysis for clinical trials with missing continuous outcome data using controlled multiple imputation: a practical guide. Stat Med. 2020;39(21):2815–42.
25. Rusanov A, Weiskopf NG, Wang S, Weng C. Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research. BMC Med Inf Decis Mak. 2014;14:51.
26. Larsen K, Merlo J. Appropriate assessment of neighborhood effects on individual health: integrating random and fixed effects in multilevel logistic regression. Am J Epidemiol. 2005;161(1):81–8.

27. Ministry of Health and Medical Education. Iran. [Title of the national maternal care protocol — to be completed by the authors]. Tehran: Ministry of Health and Medical Education; 2013. (1392 Solar Hijri).
28. Mosayebi-Molasaraie M, Doosti-Irani A, Pilevari S, Cheraghi Z. Predictors of miscarriage in the west of Iran: a case-control study. J Iran Med Counc. 2023;6(2):362–8.
29. Boots C, Stephenson MD. Does obesity increase the risk of miscarriage in spontaneous conception: a systematic review. Semin Reprod Med. 2011;29(6):507–13.
30. Mills JL, Simpson JL, Driscoll SG, Jovanovic-Peterson L, Van Allen M, Aarons JH, et al. Incidence of spontaneous abortion among normal women and insulin-dependent diabetic women whose pregnancies were identified within 21 days of conception. N Engl J Med. 1988;319(25):1617–23.

Box 1

Box 1. Six rules for risk-factor analysis of routine pregnancy records

#	Rule
1	Use as exposures or predictors only variables recorded at or before pregnancy entry; treat supplements, laboratory results, blood pressure and assessment scores as follow-up markers unless their dates are known.
2	Check whether scores are averages of repeated assessments (fractional values) and restrict to a comparable number of assessments.
3	Compare the missingness of each variable between cases and comparison women before choosing a missing-data method; report results under more than one strategy, including a delta-adjusted analysis, when missingness depends on the outcome.
4	Never add indicators of missing data or follow-up records to prediction models without checking for outcome leakage; report AUC with and without them.
5	Use cluster-aware inference and examine centre-level recording completeness before interpreting differences between centres.
6	Report a stability classification across alternative analyses, and describe how pregnancies in progress were handled in the comparison group.

Table 1

Table 1 is available in the Supplementary Files section.

Figures

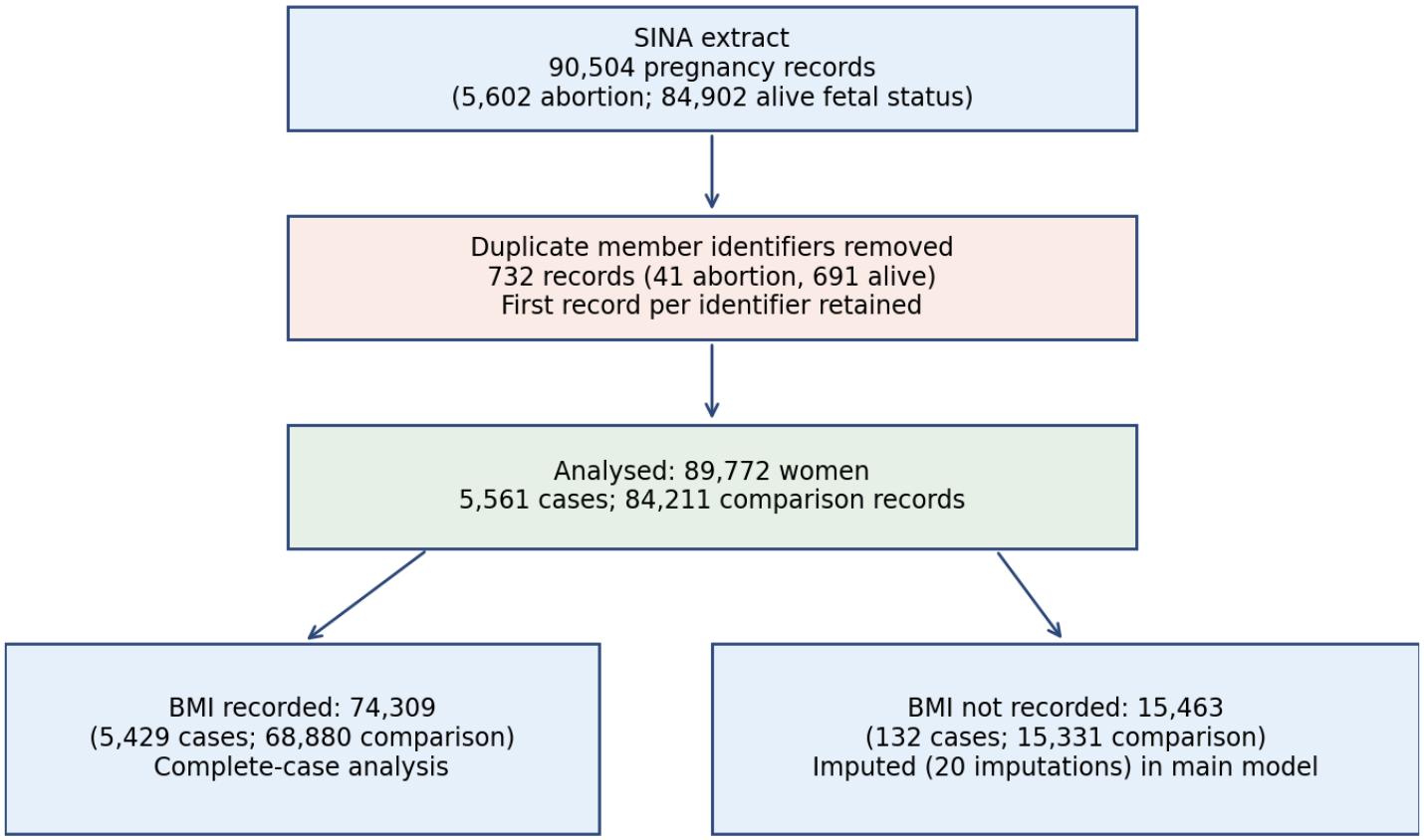


Figure 1

Flow of records into the analysis

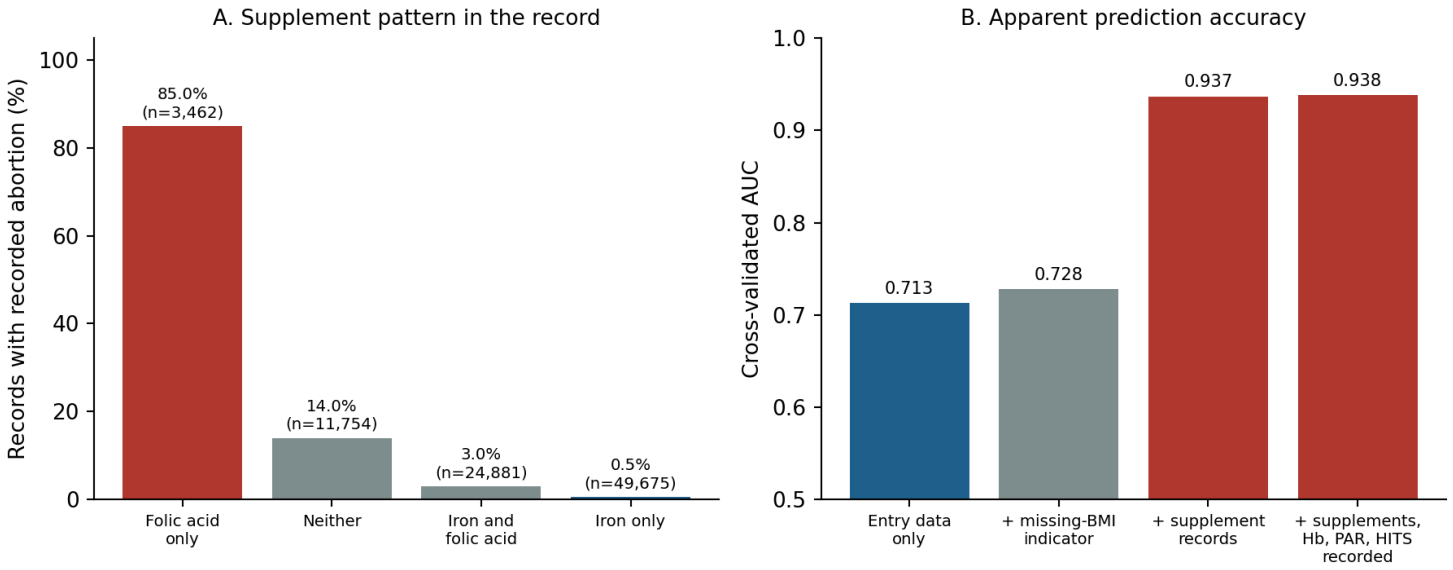


Figure 2

Follow-up records mark how far the pregnancy progressed

A: proportion of records with recorded abortion by supplement pattern. B: 10-fold cross-validated AUC of logistic models with entry-time variables, with an indicator of missing BMI, and with follow-up records added

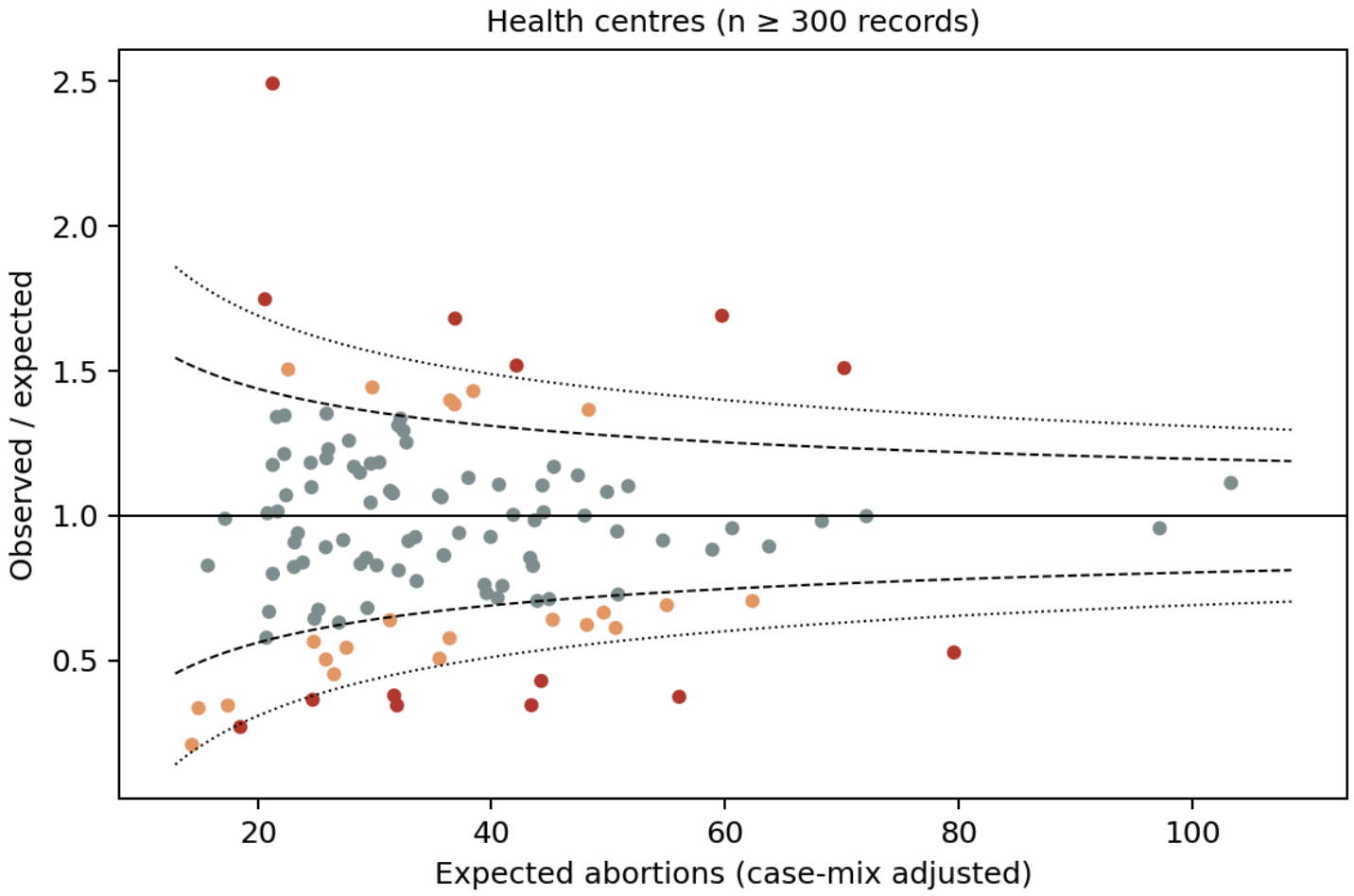


Figure 3

Case-mix-adjusted observed/expected ratio by health centre (≥300 records)

Dashed: 95% limits; dotted: 99.8% limits. Centre O/E correlated negatively with the centre's proportion of missing BMI (ρ = −0.43).

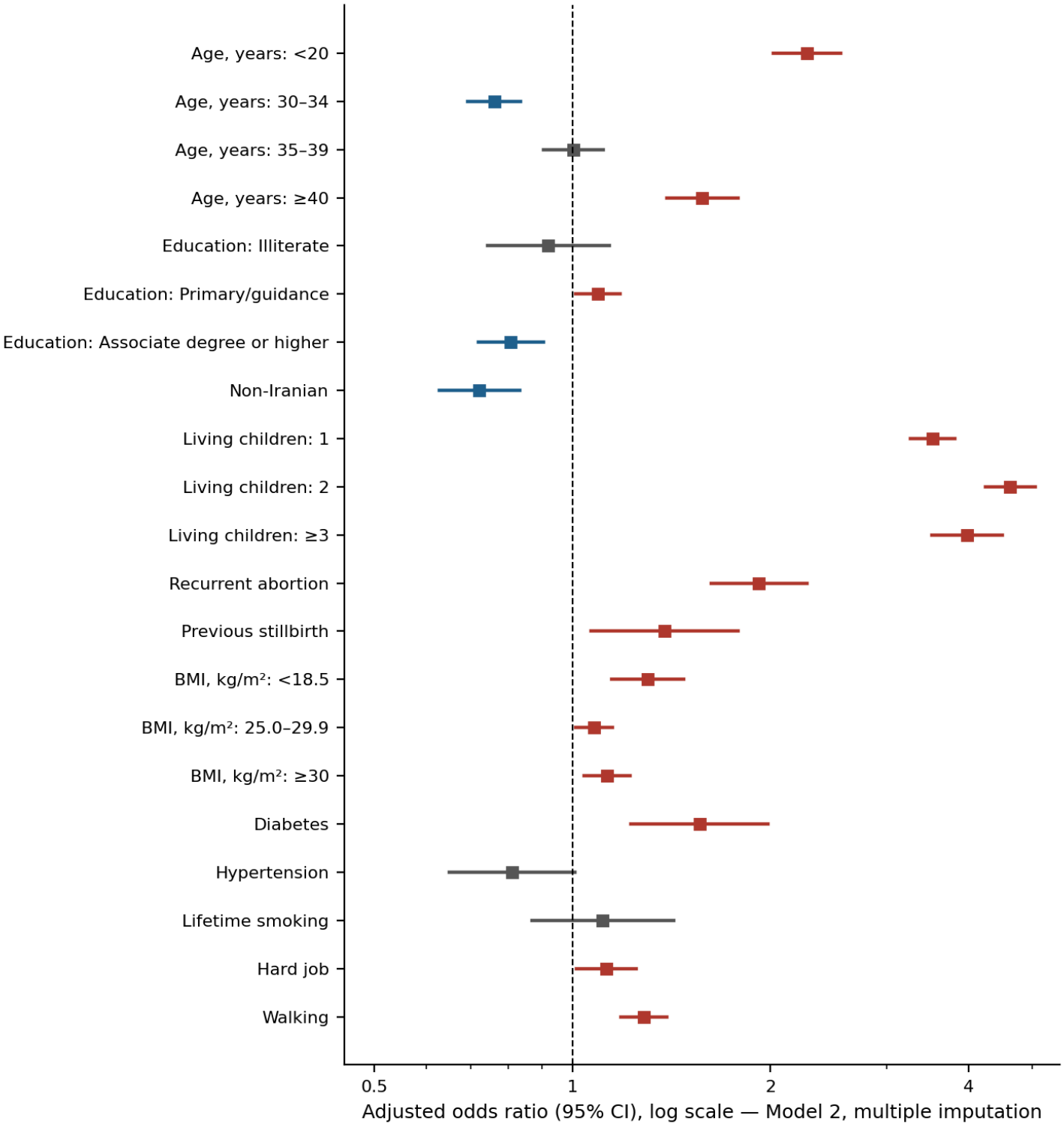


Figure 4

Adjusted odds ratios (Model 2, multiple imputation)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.pdf
- Table1.docx