

1 **Widely accessible prognostication using medical history for fetal**
2 **growth restriction and small for gestational age in nationwide insured**
3 **women**

4 Herdiantri SUFRIYANA, MD PhD;^{1,2} Fariska Zata AMANI, MD;³ Aufar Zimamuz
5 Zaman AL HAJIRI, MD;⁴ Yu-Wei WU, PhD;^{1,5} Emily Chia-Yu SU, PhD.^{1,5,6, *}

6 ¹ Graduate Institute of Biomedical Informatics, College of Medical Science and
7 Technology, Taipei Medical University, 250 Wu-Xing Street, Taipei 11031, Taiwan

8 ² Department of Medical Physiology, Faculty of Medicine, Universitas Nahdlatul
9 Ulama Surabaya, 57 Raya Jemursari Road, Surabaya 60237, Indonesia

10 ³ Department of Obstetrics and Gynecology, Faculty of Medicine, Universitas Nahdlatul
11 Ulama Surabaya, 57 Raya Jemursari Road, Surabaya 60237, Indonesia

12 ⁴ Faculty of Medicine, Universitas Nahdlatul Ulama Surabaya, 57 Raya Jemursari Road,
13 Surabaya 60237, Indonesia

14 ⁵ Clinical Big Data Research Center, Taipei Medical University Hospital, 250 Wu-Xing
15 Street, Taipei 11031, Taiwan

16 ⁶ Research Center for Artificial Intelligence in Medicine, Taipei Medical University,
17 250 Wu-Xing Street, Taipei 11031, Taiwan

18 * Corresponding author: Graduate Institute of Biomedical Informatics, College of
19 Medical Science and Technology, Taipei Medical University, 250 Wu-Xing Street,
20 Taipei 11031, Taiwan. Phone: +886-2-66382736 ext. 1515. Email address:
21 emilysu@tmu.edu.tw.

22 Abstract

23 **Objectives:** Prevention of fetal growth restriction/small for gestational age is adequate
24 if screening is accurate. Ultrasound and biomarkers can achieve this goal; however, both
25 are often inaccessible. This study aimed to develop, validate, and deploy a prognostic
26 prediction model for screening fetal growth restriction/small for gestational age using
27 only medical history. **Methods:** From a nationwide health insurance database
28 ($n=1,697,452$), we retrospectively selected visits of 12-to-55-year-old females to 22,024
29 healthcare providers of primary, secondary, and tertiary care. This study used machine
30 learning (including deep learning) to develop prediction models using 54 medical-
31 history predictors. After evaluating model calibration, clinical utility, and explainability,
32 we selected the best by discrimination ability. We also externally validated and
33 compared the models with those from previous studies, which were rigorously selected
34 by a systematic review of Pubmed, Scopus, and Web of Science. **Results:** We selected
35 169,746 subjects with 507,319 visits for predictive modeling. The best prediction model
36 was a deep-insight visible neural network. It had an area under the receiver operating
37 characteristics curve of 0.742 (95% confidence interval 0.734 to 0.750) and a sensitivity
38 of 49.09% (95% confidence interval 47.60% to 50.58% using a threshold with 95%
39 specificity). The model was competitive against the previous models in a systematic
40 review of 30 eligible studies of 381 records, including those using either ultrasound or
41 biomarker measurements. We deployed a web application to apply the model.
42 **Conclusions:** Our model used only medical history to improve accessibility for fetal
43 growth restriction/small for gestational age screening. However, future studies are
44 warranted to evaluate if this model's usage impacts patient outcomes.

45 **Key words:** fetal growth restriction, small for gestational age, machine learning, deep
46 learning, electronic health records, risk prediction.

47 **Introduction**

48 Fetal growth restriction (FGR) and small for gestational age (SGA) are two terms of a
49 single condition with the same diagnostic criterion in principle but different measures in
50 practice.¹ This condition is the second leading cause of preventable perinatal deaths.²
51 The prevention method depends on FGR/SGA predictions with a clinically acceptable
52 predictive performance.³ However, most settings lack accessibility to predictors in
53 existing prediction models.⁴

54 A pregnancy with FGR likely results in delivering low-birth-weight infants,⁵ an
55 indirect cause of neonatal deaths.⁶⁻⁸ Neonatal mortality rates varied from 20 to 30 deaths
56 per 1000 live births worldwide in 2013.⁹ Low-birth-weight infants also need to spend
57 time in a neonatal intensive care unit.¹⁰ But, this requires high costs and is a limited
58 resource in many countries.^{11,12} Prevention of FGR/SGA may reduce neonatal mortality
59 and associated costs.¹³ Several preventive strategies were found to be effective for
60 FGR/SGA,¹⁴ yet, this intervention needs a screening method with a good predictive
61 performance.³

62 Since a low-cost method such as symphysis fundal height was not recommended
63 by a Cochrane review, mainly due to low sensitivity (~17%), there is a trend to employ
64 either ultrasound or biomarker measurements for FGR/SGA screening.¹⁵ Nonetheless,
65 these methods are inaccessible in resource-limited settings.^{15,16} Meanwhile, there was an
66 association detected of FGR/SGA with a woman's medical history.¹⁷ Because a health
67 insurance claim database abundantly records medical histories, this allows proactive
68 screening for FGR/SGA, particularly in countries with universal health coverage.¹⁸
69 Screening by medical history is also independent of the number of pregnancy
70 consultations on which FGR detection depends (hazard ratio 1.15, 95% confidence

71 interval [CI] 1.05 to 1.26).¹⁹ However, studies have yet to develop a screening method
72 for FGR/SGA using only medical history.

73 Prognostic predictions of FGR/SGA using medical histories can be either a
74 prediction model for use in resource-limited settings or a preliminary prediction model
75 before ordering ultrasound and biomarker measurements. Both statistical and
76 computational machine learning can predict pregnancy outcomes in advance,²⁰
77 including deep learning and those using only medical history.^{20,21} We aimed to develop,
78 validate, and deploy a prognostic prediction model for screening FGR/SGA using only
79 medical history in nationwide insured women.

80 **Materials and Methods**

81 Report completeness of this study was according to the transparent reporting of a
82 multivariable prediction model for individual prognosis or diagnosis (TRIPOD)
83 checklist (Appendix A).²² We followed a protocol with the same software and hardware
84 (Tables B1-B3, C1),²³ except those stated otherwise. This study was under a single
85 project that compared a deep-insight visible neural network (DI-VNN) to other machine
86 learning algorithms to predict several outcomes in medicine. The Taipei Medical
87 University Joint Institutional Review Board exempted this project from the ethical
88 review (TMU-JIRB no.: N202106025).

89 **Study design and data source**

90 We applied a retrospective design to select subjects from a public dataset version 2
91 (August 2019;²⁴ access approval no.: 510/PPID/1223) of a nationwide health insurance
92 database in Indonesia. The dataset was a cross-sectional, random sampling of ~1% of

93 insurance holders within 2 years up to 2016. This sampling included all affiliated
94 healthcare providers ($n=22,024$) at all levels (i.e., primary, secondary, and tertiary care).

95 The inclusion criteria were females aged 12 to 55 years who had visited primary,
96 secondary, or tertiary care facilities. All visits afterward were exprotocolcluded if a
97 woman was pregnant and had a delivery. If a woman became pregnant twice within the
98 dataset period, then different identifiers were assigned to differentiate the pregnancy
99 periods of that woman. To determine a delivery, we used several codes of diagnoses and
100 procedures (Table C2).

101 This study developed a prediction model for detecting in advance a visit by a
102 subject who would be diagnosed with either FGR or SGA. We pursued to achieve an
103 acceptable sensitivity at 95% specificity but using more-accessible predictors.
104 Nevertheless, we compared our prediction models with those from previous studies
105 selected by systematic review methods to evaluate if our predictive modeling was
106 successful. Since there were different policies in choosing a prediction threshold (e.g.,
107 that at 90% vs. 95% specificity), the comparison was conducted using receiver
108 operating characteristics (ROC) curves and the area under the ROC curve (AUROC).

109 The event outcome definition in this study utilized the International
110 Classification of Disease version 10 (ICD-10) codes. These were codes preceded by
111 either O365 (maternal care for known or suspected fetal growth) or P05 (disorders of
112 newborns related to slow fetal growth and fetal malnutrition). Both codes indicating
113 FGR and SGA were assigned with those respectively for mothers and fetuses/newborns.
114 A nonevent outcome was assigned if the end of pregnancy was identified within the
115 dataset period by the codes for determining delivery. Otherwise, we assigned an
116 outcome to a censored one.

117 Candidate predictors were only medical histories of diagnoses and procedures.
118 These were either single or multiple ICD-10 codes. As extensively described in the
119 protocol,²² the preprocessing of candidate predictors consisted of (1) preventing zero
120 variance, perfect separation and leakage of the outcome, and redundant predictors; (2)
121 simulating real-world data; and (3) systematically determining the multiple ICD-10
122 codes for defining latent candidate predictors based on prior knowledge. After this
123 preprocessing (Tables C3-C7), we identified 54 candidate predictors, including four
124 latent candidate predictors of multiple pregnancies, varicella, urinary tract infections,
125 and placenta previa.

126 **Statistical analysis**

127 We developed five models using different algorithms and hyperparameter tuning, as
128 described in the protocol.²² The first applied ridge regression (RR). The second to fourth
129 models used 54 candidate predictors transformed into principal components (PCs). We
130 applied three algorithms using these PCs: (1) elastic net regression (PC-ENR); (2)
131 random forest (PC-RF); and (3) gradient boosting machine (PC-GBM). The fifth model
132 was a deep-insight visible neural network (DI-VNN). However, unlike the protocol,²²
133 we did not limit this model to only 22 of 54 candidate predictors, which had a false
134 discovery rate of $\leq 5\%$ based on differential analyses with Benjamini-Hochberg multiple
135 testing corrections. Instead, we used all 54 candidate predictors considering the
136 feasibility of constructing the data-driven network architecture. In addition, all model
137 recalibration was by either a logistic regression or a general additive model using
138 locally weighted scatterplot smoothing. The recalibration procedure also differed from
139 the protocol.²² This is because the models only sometimes resulted in a wide range of
140 predicted probabilities, as required for recalibration. Unlike the protocol, we chose 100

141 repetitions for bootstrapping, considering the sample size of this study compared to that
142 of the protocol. Details on model development and validation are described in Table B2.

143 For deployment, this model will predict the outcome each time an insured
144 woman visits a healthcare provider. We provided the best model in this study as a web
145 application. A user is only required to upload a comma-separated value (.csv) file
146 consisting a two-column table. It includes column headers of “admission_date” (yyyy-
147 mm-dd) and “code” (ICD-10 code at discharge) from previous to current visits.

148 We computed an uncertainty interval (i.e., 95% confidence interval, CI) for each
149 evaluation metric. This interval inference used subsets of an evaluated set, resampled by
150 bootstrapping and cross-validation. All analytical codes were publicly shared (see “Data
151 sharing statement”).

152 The selection of latent candidate predictors in the first model applied inverse
153 probability weighting for the multivariate analyses, according to the protocol.²² Results
154 were also compared to those by outcome regression. We selected a latent candidate
155 predictor if its association with the outcome had an interval of odds ratio (OR)
156 excluding a value of 1.

157 The evaluation metrics were those for assessing the models' calibration, utility,
158 explainability, and discrimination. To evaluate the model calibration, we assessed (1) a
159 calibration plot with a regression line and histograms of either event or nonevent
160 distribution of the predicted probabilities; (2) the intercept and slope of the linear
161 regression; and (3) the Brier score. We measured the clinical utility using a decision
162 curve analysis by comparing the net benefits of a model with those if we treated all
163 predictions as either positive (i.e., treat all) or negative (i.e., treat none). Clinicians (i.e.,
164 FZA and AZZAH) assessed the explainability. They were given counterfactual

165 quantities for each predictor in a model.²⁵ These consisted of the probability of
166 necessity (PN; Equation 1) and the probability of sufficiency (PS; Equation 2).
167 Eventually, we evaluated the discrimination ability of well-calibrated models by the
168 ROC curve and sensitivity at 95% specificity.

169
$$PN = \frac{\textit{number of predicted nonevents if changing the predictor to negative}}{\textit{number of predicted events with a positive predictor}} \dots\dots\dots \textbf{Equation 1}$$

170
$$PS = \frac{\textit{number of predicted events if changing the predictor to positive}}{\textit{number of predicted nonevents with a negative predictor}} \dots\dots\dots \textbf{Equation 2}$$

171 Furthermore, we compared our models with previous ones identified by a
172 systematic review and meta-analysis (see “Comparison to previous models”). We
173 compared the best model with those from previous studies. These were identified by
174 following 11 of 14 items in section methods of the preferred reporting items for
175 systematic reviews and meta-analyses (PRISMA)-extended checklist statements.²⁶
176 Those items are described in Table B4.

177 **Results**

178 **Subject characteristics**

179 From the database ($n=1,697,452$), we selected 12-to-55-year-old females ($n=169,746$)
180 that had visited ($n=507,319$) primary, secondary, or tertiary care (Figure 1). After
181 removing subjects with no pregnancy and their visits, we split the selected data for
182 internal and external validation. There were no overlapped visits between the internal
183 and external validation sets. We only used the former to develop the prediction models
184 in this study, including association tests to select candidate predictors.

185 Figure 1

186 To characterize subjects in the internal validation set (Table 1), we also included
187 subjects with uncensored outcomes ($n=26,576$). There were differences between
188 subjects without and those with FGR/SGA based on multiple univariate analyses. These
189 were in terms of subject characteristics, i.e.: (1) maternal age; (2) third vs. first
190 categories of the insurance class; (3) single vs. married categories of the marital status;
191 and (4) private company vs. central-government employee categories of the occupation
192 segment of the householder. We also identified differences in terms of latent candidate
193 predictors. Two of these variables were the risk of adverse pregnancy by maternal age
194 and a low socioeconomic status. The former represented maternal age of either <20
195 or >25 years, while the latter represented either the third insurance class or unemployed
196 householder (Table C7). Differences in the latent candidate predictors implied their
197 associations with the outcome.

198 Table 1

199 **Association tests**

200 To select latent candidate predictors in the prediction models, their associations with the
201 outcome were verified by multivariate analyses using inverse probability weighting (see
202 Table C8 for comparison to those verified by logistic regression). We adjusted
203 associations using confounders (Table 2; Figures B1-B9). Significant associations
204 persisted after adjustment, in which the effect sizes only slightly changed. However,
205 since the effect sizes were small, the selected latent candidate predictors might be weak
206 predictors for the prediction models.

207 Table 2

208 **The best prediction model**

209 Only three of the five models were approximately well-calibrated (Figure 2a): the PC-
210 ENR, PC-GBM, and DI-VNN. Among these models, the PC-GBM was considerably
211 the best-calibrated (intercept -0.00098, 95% CI -0.13098 to 0.12902; slope 0.95, 95%
212 CI 0.46 to 1.44; Brier score 0.0063). Nevertheless, the downstream analyses evaluated
213 all of the well-calibrated models.

214 **Figure 2**

215 The net benefits of these models were higher than those of either the treat-all or
216 treat-none prediction (Figure 2b). It also applied to those using a threshold of 95%
217 specificity. With this threshold, we found the DI-VNN to be the best model in terms of
218 clinical utility with a net benefit of 0.0023 (95% CI 0.0022 to 0.0024).

219 Regarding model explainability, both clinicians chose the DI-VNN among the
220 well-calibrated models. They considered the plausibility of the top-five predictors
221 according to the counterfactual probabilities (Table 3). One of the top predictors in the
222 DI-VNN, i.e., severe preeclampsia, could change most of the predicted events into
223 nonevents (PN 98.57%, 95% CI 98.5% to 98.63%) if the predictors were changed from
224 positive to negative. Most of the nonevents were also changed into events (PS 2.08%,
225 95% CI 2.07% to 2.09%) if the predictors were changed from negative to positive. In
226 addition, we also show the models' parameters (Tables C9-C14) and all counterfactual
227 probabilities (Tables C15-C17).

228 **Table 3**

229 The discrimination ability differed among the well-calibrated models according
230 to the ROC curves (Figure 3) and AUROCs (Figure 4). Based on the internal calibration

231 split, we identified that the best model was also the DI-VNN (AUROC 0.742, 95% CI
232 0.734 to 0.750; sensitivity 49.09%, 95% CI 47.60% to 50.58%). Using external
233 validation, the AUROC of the DI-VNN (0.561, 95% CI 0.558 to 0.564) was
234 considerably robust (i.e., the 95% CI >0.5).

235 Figure 3

236 Furthermore, we compared the best model with those from previous studies.
237 Only three studies fulfilled the eligibility criteria from three literature databases within
238 the last 5 years. All of the studies were systematic reviews. Thus, we also searched
239 eligible articles in the systematic reviews, including those published more than 5 years
240 earlier. This step resulted in 381 records, including the three systematic reviews (Figure
241 B10). We included 27 studies (Tables D1, D2) of these records for the meta-analysis.
242 These studies used only a training set; thus, the evaluation metrics were extracted only
243 from the training set. We categorized these studies based on the publication year such
244 that the trend of the predictor modalities could be differentiated. By estimation, the DI-
245 VNN was outperformed by those using ultrasound only from previous studies published
246 from 1992 to <2002 (Figures 3, 4, Table C18). This finding was according to the
247 sensitivity. However, those models were developed using smaller sample sizes (Figure 3)
248 and were only evaluated using training sets (Figure 4). We also identified the latter issue
249 for the previous model, which used both ultrasound and biomarkers but without other
250 predictors, from a previous study published in a later year. Meanwhile, based on the
251 AUROC using external validation splits, the DI-VNN was also estimated to outperform
252 the previous models, which used either ultrasound or biomarkers without or with other

253 predictors, from previous studies published from 2002 to 2016 (i.e., the two latest
254 groups of publication years).

255 Figure 4

256 Eventually, we chose the DI-VNN to predict FGR/SGA in advance among 12-
257 to-15-year-old females that visited primary, secondary, or tertiary care. Similar to the
258 development pipeline of the prediction model, only a pregnant woman was eligible for
259 the use of the DI-VNN to compute a predicted probability of FGR/SGA. We deployed
260 the DI-VNN as a web application (https://predme.app/fgr_sga/). It can be used for future
261 use or independent validation of the DI-VNN because it is open access.

262 **Discussion**

263 We developed, validated, and deployed a web application to predict FGR/SGA in
264 advance using the medical history of diagnoses and procedures. The prediction model
265 for the web application was the DI-VNN, chosen among five prediction models in this
266 study, using only an internal validation set. However, external validation also
267 demonstrated the robustness of the DI-VNN's predictive performance. It was also
268 comparable to those developed in the previous studies, which used ultrasound and
269 biomarkers without or with other predictors.

270 For predicting FGR/SGA, the previous models, as systematically reviewed in
271 this study (Table D2), mainly required either ultrasound or biomarker measurements
272 and a specific range of gestational ages. The models included those which were
273 competitive with the DI-VNN based on the AUROC by internal validation (Figure 4).
274 The models were by Shlossman, et al ²⁷ (nos. 17e, 17f, and 17b), Bednarek, et al ²⁸ (no.
275 21), Valiño, et al ²⁹ (no. 16), and Poon, et al ³⁰ (no. 24). Conversely, external validation

276 estimated that the DI-VNN would outperform the other previous models with the
277 similar requirements. The models were by Bano, et al³¹ (no. 22a), Carbone, et al³² (no.
278 15c), Leung, et al³³ (no. 8c), and Krantz, et al³⁴ (no. 9b). Furthermore, evaluation of
279 the previous models used training sets only, in which the predictive performances might
280 have been overoptimistic.

281 The DI-VNN required neither ultrasound nor biomarkers without or with other
282 predictors. We would expect wider access for FGR/SGA predictions as either (1) a
283 prediction model for use in resource-limited settings or (2) a preliminary prediction
284 model before ordering advanced predictor measurements. However, the DI-VNN needs
285 an impact study to evaluate its effect on patient outcomes in various settings.

286 An effective prevention for FGR was given by ≤ 16 weeks' gestation.³ To widen
287 prevention time window, more clinical trials are needed. These studies are more
288 efficient if they are conducted among pregnant women with higher risk, as predicted by
289 the DI-VNN. Since it did not require a specific range of gestational ages, the DI-VNN
290 opens more opportunities to conduct such trials.

291 One of the strengths of this study were no requirements from our models,
292 including the DI-VNN, for either ultrasound or biomarker measurements to predict
293 FGR/SGA in advance. We could apply our models to a general population of pregnant
294 women. Furthermore, our model did not require a specific gestational age range for
295 computing the predicted probability. Unlike previous studies, we also conducted
296 external validation to estimate the future predictive performance of the DI-VNN.

297 However, we also identified several limitations of this study. The predictive
298 performance of the best model, i.e., the DI-VNN, was considerably moderate according
299 to the AUROC as was the sensitivity at 95% specificity using an internal validation set.

300 However, previous models also achieved similar predictive performances. Another
301 limitation was that medical histories from electronic health records might take time to
302 execute; yet, this is considerably more achievable in many settings. It still needs to be
303 determined if the DI-VNN can improve patient outcomes. Nevertheless, this problem is
304 not exclusive to this study because many previous studies in medicine have yet to
305 evaluate the impacts of their prediction models.³⁵

306 **Abbreviations**

307 AUROC, area under the receiver operating characteristics curve

308 CI, confidence interval

309 DI-VNN, deep-insight visible neural network

310 ENR, elastic net regression

311 FGR, fetal growth restriction

312 GBM, gradient boosting machine

313 ICD-10, International Classification of Disease version 10

314 OR, odds ratio

315 PC, principal component

316 PN, probability of necessity

317 PS, probability of sufficiency

318 RF, random forest

319 ROC, receiver operating characteristics

320 RR, ridge regression

321 SGA, small for gestational age

322 **Appendices**

323 **Appendix A**

324 Transparent reporting of a multivariable prediction model for individual prognosis or
325 diagnosis (TRIPOD) checklist.

326 **Appendix B**

327 Table B1. Guidelines for developing and reporting machine learning predictive models
328 in biomedical research.

329 Table B2. Prediction model risk of bias assessment tools (PROBAST).

330 Table B3. Clinical checklists for assessing the suitability of machine learning
331 applications in healthcare.

332 Table B4. Preferred reporting items for systematic reviews and meta-analyses (PRISMA)
333 2020 expanded checklist.

334 Figure B1. Association diagram of pregnancy-induced hypertension and fetal growth
335 restriction (FGR)/small for gestational age (SGA).

336 Figure B2. Association diagram of multiple pregnancies and fetal growth restriction
337 (FGR)/small for gestational age (SGA).

338 Figure B3. Association diagram of malaria and fetal growth restriction (FGR)/small for
339 gestational age (SGA).

340 Figure B4. Association diagram of varicella and fetal growth restriction (FGR)/small for
341 gestational age (SGA).

342 Figure B5. Association diagram of risk of an adverse pregnancy by maternal age and
343 fetal growth restriction (FGR)/small for gestational age (SGA).

344 Figure B6. Association diagram of urinary tract infections and fetal growth restriction
345 (FGR)/small for gestational age (SGA).

346 Figure B7. Association diagram of placenta previa and fetal growth restriction

347 (FGR)/small for gestational age (SGA).

348 Figure B8. Association diagram of a low socioeconomic status and fetal growth

349 restriction (FGR)/small for gestational age (SGA).

350 Figure B9. Unified association diagram and fetal growth restriction (FGR)/small for

351 gestational age (SGA).

352 Figure B10. Flow diagram to find comparable models from previous studies.

353 **Appendix C**

354 Table C1. R package versions.

355 Table C2. Codes for determining delivery or immediate after-delivery care.

356 Table C3. Candidate predictors with non-zero variances.

357 Table C4. Excluded codes in the training set that may leak outcome information.

358 Table C5. Pair-wise Pearson correlations to identify redundant candidate predictors.

359 Table C6. Results of systematic human learning and data availability for latent

360 candidate predictors.

361 Table C7. International classification of disease (ICD)-10 codes or demographical

362 variables for latent candidate predictors.

363 Table C8. Association tests by logistic regression (LR) and inverse probability

364 weighting (IPW).

365 Table C9. Principal component (PC) weights.

366 Table C10. Principal-component elastic net regression (PC-ENR) weights.

367 Table C11. Principal-component gradient boosting machine (PC-GBM) variable

368 importance.

369 Table C12. Selected predictors based on a differential analysis with a false discovery
370 rate (FDR) of <0.05.

371 Table C13. Deep-insight visible neural network (DI-VNN) intermediate outputs.

372 Table C14. Connections between ontologies under the root of a deep-insight visible
373 neural network (DI-VNN).

374 Table C15. Probabilities of necessity (PN) and sufficiency (PSs) of predictors in a
375 principal-component elastic net regression (PC-ENR).

376 Table C16. Probabilities of necessity (PN) and sufficiency (PS) of predictors in a
377 principal-component gradient boosting machine (PC-GBM).

378 Table C17. Probabilities of necessity (PN) and sufficiency (PS) of predictors in a deep-
379 insight visible neural network (DI-VNN).

380 Table C18. Extracted data of available evaluation metrics for comparable models.

381 **Appendix D**

382 Table D1. Search and filter results of previous studies.

383 Table D2. Comparable models to evaluate the success criteria.

384 **Author contributions**

385 **HS:** Conceptualization, Methodology, Software, Validation, Formal Analysis,
386 Investigation, Data Curation, Writing—Original Draft, Visualization, Project
387 Administration, Funding Acquisition. **FZA:** Validation, Formal Analysis, Data Curation,
388 Writing—Review & Editing. **AZZAH:** Validation, Formal Analysis, Data Curation,
389 Writing—Review & Editing. **YWW:** Conceptualization, Methodology, Writing—
390 Review & Editing, Supervision. **ECYS:** Conceptualization, Methodology, Resources,
391 Writing—Review & Editing, Supervision, Funding acquisition. All authors have read
392 and approved the manuscript and agreed to be accountable for all aspects of the work in
393 ensuring that questions related to the accuracy or integrity of any part of the work are
394 appropriately investigated and resolved.

395 **Conflict of interest**

396 The authors report no conflict of interest.

397 **Acknowledgments**

398 The BPJS Kesehatan in Indonesia permitted access to the sample dataset in this study.
399 This study was funded by: (1) the Postdoctoral Accompanies Research Project from the
400 National Science and Technology Council (NSTC) of Taiwan (grant no.: NSTC111-
401 2811-E-038-003-MY2), and the Lembaga Penelitian dan Pengabdian kepada
402 Masyarakat (LPPM) Universitas Nahdlatul Ulama Surabaya of Indonesia (grant no.:
403 161.4/UNUSA/Adm-LPPM/III/2021) to Herdiantri Sufriyana; and (2) the Ministry of
404 Science and Technology (MOST) of Taiwan (grant nos.: MOST110-2628-E-038-001
405 and MOST111-2628-E-038-001-MY2), and the Higher Education Sprout Project from
406 the Ministry of Education (MOE) of Taiwan (grant no.: DP2-111-21121-01-A-05) to
407 Emily Chia-Yu Su. These funding bodies had no role in the study design; in the
408 collection, analysis, and interpretation of data; in the writing of the report; or in the
409 decision to submit the article for publication.

410 **Data Statements**

411 The social security administrator provided the data for health or *badan penyelenggara*
412 *jaminan sosial (BPJS) kesehatan* in Indonesia, with restrictions (access approval no.:
413 510/PPID/1223). Data are available from the authors upon reasonable request and with
414 permission of the BPJS Kesehatan. The latter needs a request to the BPJS Kesehatan for
415 their sample dataset published in August 2019 via <https://e-ppid.bpjs-kesehatan.go.id/>.
416 The analytical codes are available at https://github.com/herdiantrisufriyana/fgr_sga.

417 **References**

- 418 [1] Fetal growth restriction: Acog practice bulletin, number 227. *Obstet Gynecol* 2021;**137**:e16-e28.
419 doi: <https://doi.org/10.1097/aog.0000000000004251>.
- 420 [2] Nardoza LM, Caetano AC, Zamarian AC, et al. Fetal growth restriction: Current knowledge.
421 *Arch Gynecol Obstet* 2017;**295**:1061-77. doi: <https://doi.org/10.1007/s00404-017-4341-9>.
- 422 [3] Roberge S, Nicolaides K, Demers S, Hyett J, Chaillet N, Bujold E. The role of aspirin dose on
423 the prevention of preeclampsia and fetal growth restriction: Systematic review and meta-analysis.
424 *Am J Obstet Gynecol* 2017;**216**:110-20.e6. doi: <https://doi.org/10.1016/j.ajog.2016.09.076>.
- 425 [4] Pedrosa MA, Palmer KR, Hodges RJ, Costa FDS, Rolnik DL. Uterine artery doppler in
426 screening for preeclampsia and fetal growth restriction. *Rev Bras Ginecol Obstet* 2018;**40**:287-
427 93. doi: <https://doi.org/10.1055/s-0038-1660777>.
- 428 [5] Mallia T, Grech A, Hili A, Calleja-Agius J, Pace NP. Genetic determinants of low birth weight.
429 *Minerva Ginecol* 2017;**69**:631-43. doi: <https://doi.org/10.23736/s0026-4784.17.04050-3>.
- 430 [6] Lawn JE, Cousens S, Zupan J. 4 million neonatal deaths: When? Where? Why? *Lancet*
431 2005;**365**:891-900. doi: [https://doi.org/10.1016/s0140-6736\(05\)71048-5](https://doi.org/10.1016/s0140-6736(05)71048-5).
- 432 [7] Ausbeck EB, Allman PH, Szychowski JM, Subramaniam A, Katheria A. Neonatal outcomes at
433 extreme prematurity by gestational age versus birth weight in a contemporary cohort. *Am J*
434 *Perinatol* 2021;**38**:880-88. doi: <https://doi.org/10.1055/s-0040-1722606>.
- 435 [8] Tabet M, Flick LH, Xian H, Jen Jen C. Smallness at birth and neonatal death: Reexamining the
436 current indicator using sibling data. *Am J Perinatol* 2021;**38**:76-81. doi:
437 <https://doi.org/10.1055/s-0039-1694761>.

- 438 [9] Lehtonen L, Gimeno A, Parra-Llorca A, Vento M. Early neonatal death: A challenge worldwide.
439 *Semin Fetal Neonatal Med* 2017;**22**:153-60. doi: <https://doi.org/10.1016/j.siny.2017.02.006>.
- 440 [10] Colella M, Frérot A, Novais ARB, Baud O. Neonatal and long-term consequences of fetal
441 growth restriction. *Curr Pediatr Rev* 2018;**14**:212-18. doi:
442 <https://doi.org/10.2174/1573396314666180712114531>.
- 443 [11] Umran RM, Al-Jammali A. Neonatal outcomes in a level ii regional neonatal intensive care unit.
444 *Pediatr Int* 2017;**59**:557-63. doi: <https://doi.org/10.1111/ped.13200>.
- 445 [12] Horbar JD, Edwards EM, Greenberg LT, et al. Racial segregation and inequality in the neonatal
446 intensive care unit for very low-birth-weight and very preterm infants. *JAMA Pediatr*
447 2019;**173**:455-61. doi: <https://doi.org/10.1001/jamapediatrics.2019.0241>.
- 448 [13] Ho T, Zupancic JAF, Pursley DM, Dukhovny D. Improving value in neonatal intensive care.
449 *Clin Perinatol* 2017;**44**:617-25. doi: <https://doi.org/10.1016/j.clp.2017.05.009>.
- 450 [14] Bettiol A, Avagliano L, Lombardi N, et al. Pharmacological interventions for the prevention of
451 fetal growth restriction: A systematic review and network meta-analysis. *Clin Pharmacol Ther*
452 2021;**110**:189-99. doi: <https://doi.org/10.1002/cpt.2164>.
- 453 [15] Audette MC, Kingdom JC. Screening for fetal growth restriction and placental insufficiency.
454 *Semin Fetal Neonatal Med* 2018;**23**:119-25. doi: <https://doi.org/10.1016/j.siny.2017.11.004>.
- 455 [16] Luntsi G, Ugwu AC, Nkubli FB, Emmanuel R, Ochie K, Nwobi CI. Achieving universal access
456 to obstetric ultrasound in resource constrained settings: A narrative review. *Radiography (Lond)*
457 2021;**27**:709-15. doi: <https://doi.org/10.1016/j.radi.2020.10.010>.
- 458 [17] Selvaratnam RJ, Wallace EM, Flenady V, Davey MA. Risk factor assessment for fetal growth
459 restriction, are we providing best care? *Aust N Z J Obstet Gynaecol* 2020;**60**:470-73. doi:
460 <https://doi.org/10.1111/ajo.13147>.
- 461 [18] Wagstaff A, Neelsen S. A comprehensive assessment of universal health coverage in 111
462 countries: A retrospective observational study. *Lancet Glob Health* 2020;**8**:e39-e49. doi:
463 [https://doi.org/10.1016/s2214-109x\(19\)30463-2](https://doi.org/10.1016/s2214-109x(19)30463-2).
- 464 [19] Andreasen LA, Tabor A, Nørgaard LN, et al. Why we succeed and fail in detecting fetal growth
465 restriction: A population-based study. *Acta Obstet Gynecol Scand* 2021;**100**:893-99. doi:
466 <https://doi.org/10.1111/aogs.14048>.
- 467 [20] Sufriyana H, Husnayain A, Chen YL, et al. Comparison of multivariable logistic regression and
468 other machine learning algorithms for prognostic prediction studies in pregnancy care:
469 Systematic review and meta-analysis. *JMIR Med Inform* 2020;**8**:e16503. doi:
470 <https://doi.org/10.2196/16503>.
- 471 [21] Sufriyana H, Wu YW, Su EC. Artificial intelligence-assisted prediction of preeclampsia:
472 Development and external validation of a nationwide health insurance dataset of the bpjs
473 kesehatan in indonesia. *EBioMedicine* 2020;**54**:102710. doi:
474 <https://doi.org/10.1016/j.ebiom.2020.102710>.
- 475 [22] Moons KG, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction
476 model for individual prognosis or diagnosis (tripod): Explanation and elaboration. *Ann Intern*
477 *Med* 2015;**162**:W1-73. doi: <https://doi.org/10.7326/m14-0698>.
- 478 [23] Sufriyana H, Wu YW, Su EC-Y. Human and machine learning pipelines for responsible clinical
479 prediction using high-dimensional data. *Protocol Exchange* 2021. doi:
480 <https://doi.org/10.21203/rs.3.pex-1655/v1>.
- 481 [24] Ariawan I, Sartono B, Jaya C. *Sample dataset of the bpjs kesehatan 2015-2016*. Jakarta BPJS
482 Kesehatan; 2019.
- 483 [25] Moraffah R, Karami M, Guo R, Raglin A, Liu H. Causal interpretability for machine learning-
484 problems, methods and evaluation. *ACM SIGKDD Explorations Newsletter* 2020;**22**:18-33. doi,
485 PMID.
- 486 [26] Page MJ, McKenzie JE, Bossuyt PM, et al. The prisma 2020 statement: An updated guideline for
487 reporting systematic reviews. *Bmj* 2021;**372**:n71. doi: <https://doi.org/10.1136/bmj.n71>.
- 488 [27] Shlossman P, Scisione A, Manley J, Colmorgen G, Weiner S. Doppler assessment of the
489 intrafetal vasculature in the identification of intrauterine growth retardation. Which vessel
490 is 'best' or is a combination better? *American Journal of Obstetrics & Gynecology* 1998;**178**:S88.
491 doi: <https://doi.org/10.1016/j.ajog.2012.01.022>.
- 492 [28] Bednarek M, Dubiel M, Bręborowicz GH. P05.18: Doppler velocimetry in m1 and m2 segments
493 of middle cerebral artery in pregnancies complicated by intrauterine growth restriction.
494 *Ultrasound in Obstetrics & Gynecology* 2004;**24**:300-01. doi:
495 <https://doi.org/https://doi.org/10.1002/uog.1423>.

- 496 [29] Valiño N, Giunta G, Gallo DM, Akolekar R, Nicolaides KH. Biophysical and biochemical
497 markers at 30-34 weeks' gestation in the prediction of adverse perinatal outcome. *Ultrasound*
498 *Obstet Gynecol* 2016;**47**:194-202. doi: <https://doi.org/10.1002/uog.14928>.
- 499 [30] Poon LC, Karagiannis G, Staboulidou I, Shafiei A, Nicolaides KH. Reference range of birth
500 weight with gestation and first-trimester prediction of small-for-gestation neonates. *Prenat*
501 *Diagn* 2011;**31**:58-65. doi: <https://doi.org/10.1002/pd.2520>.
- 502 [31] Bano S, Chaudhary V, Pande S, Mehta V, Sharma A. Color doppler evaluation of cerebral-
503 umbilical pulsatility ratio and its usefulness in the diagnosis of intrauterine growth retardation
504 and prediction of adverse perinatal outcome. *Indian J Radiol Imaging* 2010;**20**:20-5. doi:
505 <https://doi.org/10.4103/0971-3026.59747>.
- 506 [32] Carbone JF, Tuuli MG, Bradshaw R, Liebsch J, Odibo AO. Efficiency of first-trimester growth
507 restriction and low pregnancy-associated plasma protein-a in predicting small for gestational age
508 at delivery. *Prenat Diagn* 2012;**32**:724-9. doi: <https://doi.org/10.1002/pd.3891>.
- 509 [33] Leung TY, Sahota DS, Chan LW, et al. Prediction of birth weight by fetal crown-rump length
510 and maternal serum levels of pregnancy-associated plasma protein-a in the first trimester.
511 *Ultrasound Obstet Gynecol* 2008;**31**:10-4. doi: <https://doi.org/10.1002/uog.5206>.
- 512 [34] Krantz D, Goetzl L, Simpson JL, et al. Association of extreme first-trimester free human
513 chorionic gonadotropin-beta, pregnancy-associated plasma protein a, and nuchal translucency
514 with intrauterine growth restriction and other adverse pregnancy outcomes. *Am J Obstet Gynecol*
515 2004;**191**:1452-8. doi: <https://doi.org/10.1016/j.ajog.2004.05.068>.
- 516 [35] Zhou Q, Chen ZH, Cao YH, Peng S. Clinical impact and quality of randomized controlled trials
517 involving interventions evaluating artificial intelligence prediction tools: A systematic review.
518 *NPJ Digit Med* 2021;**4**:154. doi: <https://doi.org/10.1038/s41746-021-00524-2>.
- 519

520 **Figure Captions**

521 **Figure 1. Subject selection by applying a retrospective design and data partitioning**
522 **for internal and external validations.** The set for association tests included censored
523 outcomes. The summation of the internal and external validation numbers differs from
524 the total because: (1) there were subject overlaps; (2) the numbers of subjects and visits
525 in the censored internal validation are not shown; and (3) we excluded subjects with no
526 pregnancy before data analysis. ^a, the first and second pregnancies of a subject within
527 the database period, not parity; ^b, subjects per pregnancy episode; ^c, only subjects in the
528 external random split overlapped with those in the internal validation sets; *n*, sample
529 size; (?), number of censoring; (-), number of nonevents; (+), number of events.

530 **Figure 2. Model calibration (a) and clinical utility (b).** We evaluated both using a
531 calibration split (i.e., ~20% of internal validation set) within the optimal range of
532 predicted probabilities (equivalent to thresholds) across all of the models. This figure
533 shows only the well-calibrated models. Solid lines with gray shading show the
534 regression line and standard errors over point estimates of true probabilities. Dotted
535 lines show a threshold of 95% specificity. CI, confidence interval; DI-VNN, deep-
536 insight visible neural network; ENR, elastic net regression; GBM, gradient boosting
537 machine; PC, principal component.

538 **Figure 3. Model discrimination by receiver operating characteristics (ROC) curves.**
539 The evaluation used a calibration split (i.e., ~20% of the internal validation set) for only
540 the well-calibrated models. The vertical dotted lines show 95% specificity, while the
541 diagonal dotted lines show the area under the ROC curve (AUROC) of 0.5 as a
542 reference. DI-VNN, deep-insight visible neural network; ENR, elastic net regression;
543 GBM, gradient boosting machine; PC, principal components.

544 **Figure 4. Model discrimination by the area under the receiver operating**
545 **characteristics curves (AUROCs).** This figure shows only the well-calibrated models.
546 The vertical dotted lines show AUROCs of 0.5 and the averages using internal
547 calibration split, training set, and external random and non-random splits. See Appendix
548 D for details of eligible models from previous studies. If any predictor list of these
549 models is too long, then it is truncated by "...". β hCG, β -subunit human
550 choriogonadotropin; BMI, body-mass index; Cig., cigarette; CRL, crown-rump length
551 (fetus); CU-R, cerebral-umbilical ratio; DI-VNN, deep-insight visible neural network;
552 EFW, estimated fetal weight; ENR, elastic net regression; GBM, gradient boosting
553 machine; ICA, internal carotid artery; MCA, middle cerebral artery; NT, nuchal
554 translucency thickness (fetus); PAPP-A, pregnancy-associated plasma protein-A; PC,
555 principal component; PIGF, placental growth factor; PI, pulsatility index; RA, renal
556 artery; RI, resistance index; ROC, receiver operating characteristics; SD, systolic-
557 diastolic ratio; sFLT-1, soluble fms-like tyrosinase-1; UA, umbilical artery; UtA,
558 uterine artery.

559 **Table 1. Subject characteristics for association tests and internal validation set.**

Variable		Not FGR/SGA ^a (n=26,459)	FGR/SGA ^a (n=117)	p value
Pregnancy episode within database period ^b	First pregnancy, ^c no. (%)	25,096 (94.85)	109 (93.16)	(reference)
	Second pregnancy, ^c no. (%)	1363 (5.15)	8 (6.84)	0.41
Maternal age	Mean (SD), year	29 (6)	28 (6)	0.006**
Insurance class	First, no. (%)	3604 (13.62)	21 (17.95)	(reference)
	Unspecified, no. (%)	87 (0.33)	1 (0.85)	0.51
	Second, no. (%)	9226 (34.87)	50 (42.74)	0.78
	Third, no. (%)	13,542 (51.18)	45 (38.46)	0.03*
Marital status	Married, no. (%)	16,831 (63.61)	77 (66)	(reference)
	Single, no. (%)	2397 (9.06)	20 (17)	0.02*
	Unspecified, no. (%)	7117 (26.90)	20 (17)	0.05
	Divorced/widowed, no. (%)	114 (0.43)	77 (66)	0.97
Occupation segment of the householder	Central-government employee, no. (%)	7683 (29.04)	20 (17.1)	(reference)
	Private company employee, no. (%)	9611 (36.32)	57 (48.7)	0.002**
	Private company employer or self-employed, no. (%)	7871 (29.75)	35 (29.9)	0.06
	Local-government employee, no. (%)	1278 (4.83)	5 (4.3)	0.42
	Unemployed, no. (%)	16 (0.06)	5 (4.3)	0.98
Pregnancy-induced hypertension	Negative, no. (%)	25,366 (9.6e-01)	98 (8.4e-01)	(reference)
	Positive, no. (%)	1093 (4.1e-02)	19 (1.6e-01)	<0.001***
Multiple pregnancies	Negative, no. (%)	26,271 (9.9e-01)	112 (9.6e-01)	(reference)
	Positive, no. (%)	188 (7.1e-03)	5 (4.3e-02)	<0.001***
Malaria	Negative, no. (%)	26,439 (1.0e+00)	117 (1.0e+00)	(reference)
	Positive, no. (%)	20 (7.6e-04)	0 (0.0e+00)	<0.001***
Varicella	Negative, no. (%)	26,446 (1.0e+00)	117 (1.0e+00)	(reference)
	Positive, no. (%)	13 (4.9e-04)	0 (0.0e+00)	<0.001***
Risk of adverse pregnancy by maternal age	Negative, no. (%)	19,660 (7.4e-01)	93 (7.9e-01)	(reference)
	Positive, no. (%)	6799 (2.6e-01)	24 (2.1e-01)	<0.001***
Urinary tract infection	Negative, no. (%)	26,294 (9.9e-01)	116 (9.9e-01)	(reference)
	Positive, no. (%)	165 (6.2e-03)	1 (8.5e-03)	<0.001***
Placenta previa	Negative, no. (%)	26,187 (9.9e-01)	113 (9.7e-01)	(reference)
	Positive, no. (%)	272 (1.0e-02)	4 (3.4e-02)	<0.001***
Low socioeconomic status	Negative, no. (%)	12,901 (4.9e-01)	72 (6.2e-01)	(reference)
	Positive, no. (%)	13,558 (5.1e-01)	45 (3.8e-01)	0.05*

560 This table shows only latent candidate predictors with significant associations by multivariate analyses (Table 2). *
 561 $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$; ^a, Subject per pregnancy episode (not including censored delivery); ^b, Not
 562 FGR/SGA vs. FGR/SGA (not including those who were not pregnant); ^c, The first and second pregnancies of a
 563 subject within the database period; FGR, fetal growth restriction; SGA, small for gestational age; SD, standard
 564 deviation.

565 **Table 2. Association between each latent candidate predictor and fetal growth**
 566 **restriction (FGR)/small for gestational age (SGA) by inverse probability weighting.**

Variable of interest	Unadjusted OR (95% CI; <i>p</i> value)	Adjusted OR (95% CI; <i>p</i> value)	Adjustment
Pregnancy-induced hypertension	1.012 (1.011 to 1.013; <i>p</i> <0.001 ***)	1.007 (1.007 to 1.008; <i>p</i> <0.001 ***)	Multiple pregnancies + Risk of adverse pregnancy by maternal age
Multiple pregnancies	1.051 (1.047 to 1.054; <i>p</i> <0.001 ***)	1.048 (1.044 to 1.052; <i>p</i> <0.001 ***)	Risk of adverse pregnancy by maternal age
Malaria	0.993 (0.993 to 0.993; <i>p</i> <0.001 ***)	0.993 (0.993 to 0.993; <i>p</i> <0.001 ***)	Low socioeconomic status
Varicella	0.993 (0.993 to 0.993; <i>p</i> <0.001 ***)	0.993 (0.993 to 0.993; <i>p</i> <0.001 ***)	(no adjustment)
Risk of adverse pregnancy by maternal age	0.996 (0.996 to 0.996; <i>p</i> <0.001 ***)	0.996 (0.996 to 0.996; <i>p</i> <0.001 ***)	(no adjustment)
Urinary tract infection	1.068 (1.064 to 1.073; <i>p</i> <0.001 ***)	1.137 (1.128 to 1.146; <i>p</i> <0.001 ***)	Risk of adverse pregnancy by maternal age
Placenta previa	1.028 (1.026 to 1.031; <i>p</i> <0.001 ***)	1.022 (1.02 to 1.024; <i>p</i> <0.001 ***)	Risk of adverse pregnancy by maternal age
Low socioeconomic status	0.999 (0.999 to 1; <i>p</i> =0.05*)	0.999 (0.999 to 1; <i>p</i> =0.05*)	(no adjustment)

567 * *p*≤.05; ** *p*≤.01; *** *P* ≤.001; CI, confidence interval; OR, odds ratio.

568 **Table 3. Model explainability by clinical assessments based on counterfactual**
 569 **probabilities.**

Model	Top-five predictor	PN (95% CI)	PS (95% CI)	Clinician 1	Clinician 2
DI-VNN ^{a, b}	M791, Myalgia	97.63% (97.51% to 97.76%)	1.7% (1.7% to 1.71%)	Plausible	Implausible, only a general symptom
	O141, Severe preeclampsia	98.57% (98.5% to 98.63%)	2.08% (2.07% to 2.09%)	Implausible	Plausible, especially early-onset preeclampsia
	O410, Oligohydramnios	98.22% (98.11% to 98.33%)	1.34% (1.33% to 1.34%)	Plausible	Plausible
	O470, False labor before 37 completed weeks of gestation	98.41% (98.15% to 98.67%)	0.59% (0.59% to 0.59%)	Plausible	Plausible
	O48, Prolonged pregnancy	98.33% (98.18% to 98.48%)	0.82% (0.82% to 0.82%)	Plausible	Implausible, FGR/SGA mostly preterm and term
PC-ENR ^b	Placenta previa ^c	98.2% (98.15% to 98.25%)	8.39% (8.39% to 8.39%)	Implausible	Plausible
	E86, Volume depletion	98.08% (97.93% to 98.22%)	8.44% (8.44% to 8.44%)	Implausible	Plausible
	K021, Caries of dentine	99.9% (99.88% to 99.91%)	9.31% (9.3% to 9.32%)	Implausible	Plausible
	O410, Oligohydramnios	98.4% (98.32% to 98.49%)	8.47% (8.47% to 8.47%)	Implausible	Plausible
	O624, Hypertonic, uncoordinated, and prolonged uterine contractions	99.43% (99.37% to 99.49%)	8.44% (8.44% to 8.44%)	Implausible	Implausible, after FGR/SGA onset and only during labor
PC-GBM	Urinary tract infection ^c	98.96% (98.85% to 99.06%)	5.7% (5.68% to 5.71%)	Implausible	Plausible
	E282, Polycystic ovarian syndrome	99.82% (99.79% to 99.85%)	2.34% (2.34% to 2.34%)	Implausible	Plausible, PCOS mostly with infertility which is likely undergoing ovarian stimulation, subsequently resulting in twin pregnancy and FGR/SGA
	E86, Volume depletion	99.08% (98.97% to 99.19%)	16.46% (16.43% to 16.49%)	Implausible	Plausible
	N832, Other and unspecified ovarian cysts	98.68% (98.51% to 98.86%)	8.22% (8.2% to 8.24%)	Implausible	Implausible, only large-size cysts compete with fetal growth, yet, unspecified cysts are likely small, corpus-luteum cysts
	Z349, Supervision of normal pregnancy, unspecified	98.65% (98.61% to 98.7%)	1.76% (1.76% to 1.77%)	Implausible	Implausible, no risk of FGR/SGA in normal pregnancy

570 The clinicians assessed only the well-calibrated models without information on the predictive performances; the top-
 571 five predictors had either a top probability of necessity (PN) or probability of sufficiency (PS); ^a, chosen by clinician
 572 1; ^b, chosen by clinician 2; ^c, latent predictor (see Table 2). CI, confidence interval; DI-VNN, deep-insight visible
 573 neural network; ENR, elastic net regression; FGR, fetal growth restriction; GBM, gradient boosting machine; PC,
 574 principal component; PCOS, polycystic ovarian syndrome; PN, probability of necessity (probability of predicted
 575 outcomes would have been nonevents among samples with a positive predictor and an event if changing the predictor
 576 to negative); PS, probability of sufficiency (probability of predicted outcomes would have been events among
 577 samples with a negative predictor and a nonevent if changing the predictor to positive); SGA, small for gestational
 578 age.







