

1 **An algorithm to identify patients with rare genetic disorders and its real-world data**
2 **application**

3
4
5 Bryn D. Webb, MD^{1,2*}; Lisa Y. Lau, PhD, MPH^{1*}; Despina Tsevdos, MD³; Ryan A. Shewcraft,
6 PhD¹; David Corrigan, PhD¹; Lisong Shi, PhD¹; Seungwoo Lee, MS¹; Jonathan Tyler, PhD¹;
7 Shilong Li, PhD¹; Zichen Wang, PhD¹; Gustavo Stolovitzky, PhD¹; Lisa Edelmann, PhD¹; Rong
8 Chen, PhD¹; Eric E. Schadt, PhD^{1,4†}; Li Li, MD, MS^{1†}

9
10 ¹GeneDx, Stamford, CT, USA.

11 ²Department of Pediatrics, University of Wisconsin School of Medicine and Public Health,
12 Madison, WI, USA

13 ³Department of Pediatrics, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

14 ⁴Department of Genetics and Genomic Sciences, The Icahn Institute for Genomics and
15 Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

16
17 *Contributed equally as co-first authors

18 † Contributed equally as co-senior authors; corresponding authors

19
20 **Corresponding authors' details**

21 Name: Li Li

22 Address: 333 Ludlow St, Stamford, Connecticut 06902, USA

23 Telephone: 475-533-3720

24 Email: li.li@sema4.com

25
26
27 **Short Title:** Identifying patients with rare genetic disorders

28
29
30 **Conflict of Interest Disclosures:** The authors have no conflicts of interest relevant to this article
31 to disclose.

32
33
34 **Funding/Support:** None. This project was performed in collaboration with GeneDx. GeneDx is
35 a company that integrates genetic testing and data analytics to improve diagnosis, treatment, and
36 prevention of disease. The Icahn School of Medicine at Mount Sinai holds equity in this for-
37 profit company.

38

39 **Abbreviations**

40	CSER	Clinical Sequencing Exploratory Research
41	CS	carrier screening
42	CS-L	carrier screening, large panel
43	CS-M	carrier screening, medium panel
44	CS-S	carrier screening, small panel
45	CT	computed tomography
46	CTICU	cardiothoracic intensive care unit
47	eMERGE	Electronic Medical Records & Genomics EHR
48	EMR	electronic medical record
49	ER	emergency room
50	ICD	International Classification of Diseases
51	MRI	magnetic resonance imaging
52	MSHS	Mount Sinai Health System
53	NICU	neonatal intensive care unit
54	NPV	negative predictive value
55	PPV	positive predictive value

56
57

58 **Article Summary:**

59 Algorithm using EMR data to identify children who have been diagnosed with a genetic disorder
60 or present with illness with increased risk of genetic disorders.

61
62

63 **What's known on this subject:**

64 With over 7000 Mendelian disorders, identifying children with a specific rare genetic disorder
65 diagnosis through structured EMR data is challenging given incompleteness of records,
66 inaccurate medical diagnosis coding, as well as heterogeneity in clinical symptoms and
67 procedures for specific disorders.

68
69

70 **What this study adds:**

71 We developed a digital phenotyping algorithm using electronic medical records (EMR) data to
72 identify children aged 0-3 who have been diagnosed with genetic disorders or present with
73 illness with an increased risk for genetic disorders from a mother-child cohort.

74

75 **Author Contributions Statement Page**

76

77 Dr. Bryn Webb interpreted the data and results, carried out chart review, drafted the manuscript,
78 and critically reviewed and revised the manuscript.

79

80 Dr. Lisa Lau coordinated data collection, analyzed, and interpreted the data; drafted the
81 manuscript, and critically reviewed and revised the manuscript.

82

83 Dr. Despina Tsevdos carried out chart review, drafted, reviewed, and revised the manuscript.

84

85 Dr. Ryan Shewcraft analyzed the data, drafted the manuscript, reviewed, and revised the
86 manuscript.

87

88 Seungwoo Lee, Drs. David Corrigan, Lisong Shi, Jonathan Tyler, Shilong Li, Zichen Wang, Lisa
89 Edelman, and Gustavo Stolovitzky participated in acquisition of data, interpreted the data,
90 participated in drafting the manuscript, and revised the manuscript.

91

92 Dr. Rong Chen designed the study, collected, analyzed, and interpreted the data; critically
93 reviewed and revised the manuscript.

94

95 Drs. Li Li and Eric Schadt conceptualized, designed, and supervised the study; collected the data,
96 interpreted the data and results, drafted the manuscript, critically reviewed and revised the
97 manuscript.

98

99 All authors approved the final manuscript as submitted and agree to be accountable for all
100 aspects of the work.

101 **Abstract**

102 *Objectives*

103 Develop a digital phenotyping algorithm (PheIndex) using electronic medical records (EMR)
104 data to identify children aged 0-3 who have been diagnosed with genetic disorders or present
105 with illness with an increased risk for genetic disorders from a mother-child cohort.

106

107 *Methods*

108 We established 13 criteria for the algorithm where two metrics – a quantified score and a
109 classification – were derived. The criteria and the classification were validated by chart review
110 from a pediatrician and clinical geneticist. To demonstrate the utility of our algorithm in real-
111 world evidence applications, we examined the association between size of carrier screening
112 panel (small/ ≤ 4 genes [CS-S] vs large/ ≥ 100 genes [CS-L]) undertaken by mothers prior to
113 delivery, and children classified as presenting with illness with an increased risk for genetic
114 disorders by our algorithm.

115

116 *Results*

117 The PheIndex algorithm identified 1,088 such children out of 93,154 live births and achieved
118 90% sensitivity, 97% specificity, and 94% accuracy by chart review. We found that children
119 whose mothers received CS-L were less likely to be classified as presenting with illness with an
120 increased risk for genetic disorders and a decreased need to have multiple specialist visits and
121 multiple ER visits, compared to children whose mothers received CS-S.

122

123 *Conclusions*

124 The PheIndex algorithm can help identify when a rare genetic disorder may be present, and has
125 the potential to improve healthcare delivery by alerting providers to consider ordering a
126 diagnostic genetic test and/or referring a patient to a medical geneticist or other specialists.

127

128 **Introduction**

129 The widespread adoption of electronic medical record (EMR) systems has the potential to
130 enable large-scale population-based studies characterizing patients with rare disorders.¹ While
131 identifying genomic information from EMR systems would assist in identifying such patient
132 populations, with groups from Clinical Sequencing Exploratory Research (CSER) and Electronic
133 Medical Records & Genomics © (eMERGE) representing such efforts, they have noted that
134 genetic information is most commonly stored in unstructured formats such as PDF files or in
135 paragraphs of free text, making genetic testing results difficult to locate.^{2,3} Additionally, CSER
136 and eMERGE have not pursued a global approach to identifying patient populations with
137 confirmed genetic disorders, or patients yet to be diagnosed with a genetic condition but rather
138 whose medical records indicate that diagnostic genetic testing is warranted. Indeed, digital
139 phenotyping studies using EMR data have largely focused on identifying populations with
140 specific individual diseases, such as extracting patients with pediatric epilepsy, childhood obesity,
141 or Noonan syndrome.⁴⁻⁷

142 When using EMR data to identify patient populations affected with rare genetic disorders,
143 focusing on a specific rare genetic disorder diagnosis for any given patient is error-prone for
144 many reasons. First, of 6519 rare disorders assessed, only 11% have International Classification
145 of Disease 9 (ICD-9) codes and 21% have ICD-10 codes; some ICD codes are nonspecific, often
146 with multiple phenotypes corresponding to a single ICD code.⁸ Furthermore, physicians and
147 clinicians sometimes log certain ICD codes as they rule in or out a given condition, or when a
148 condition is part of a differential diagnosis, yet still unconfirmed. Diagnosis codes may also be
149 inaccurate or incomplete.⁹

150 Accordingly, algorithms that assess the risk of genetic disorders have the potential to
151 improve healthcare delivery by assisting physicians and clinicians with clinical decision-making,
152 including guiding when to order a diagnostic genetic test and/or refer a patient to a medical
153 geneticist or other specialists may be indicated. Further, such algorithms could also be leveraged
154 to identify rare genetic disorders patient populations to carry out cross-sectional and longitudinal
155 epidemiological studies, assess healthcare utilization, and flag patients who may be considered
156 for participation in specialized undiagnosed disease programs and precision medicine initiatives
157 as underdiagnosis of rare genetic disorders is not uncommon.¹⁰

158 As a collaborative, multidisciplinary team, we developed a digital phenotyping algorithm
159 that used structured EMR data and assessed 13 criteria to identify patients from birth to 3 years
160 of age who have been diagnosed with a rare genetic disorder or who are at high risk for such a
161 diagnosis. We tested our algorithm using a real-world dataset comprised of 93,154 live births
162 with children linked to mothers' medical records in a large academic health system. We
163 validated the algorithm through blinded chart review by a pediatrician and a clinical geneticist.

164 To demonstrate the real-world evidence application of our algorithm, we examined the
165 health outcomes of children whose mothers received carrier screening; specifically, whether
166 there was an association between children who were classified as presenting with illness with an
167 increased risk for genetic disorders by our algorithm, and the size of the carrier screening panel
168 received by the mothers of these children. To the best of our knowledge, we are the first to
169 generate a digital phenotyping algorithm beyond using ICD codes to identify children presenting
170 with illness with an increased risk for genetic disorders and employed this algorithm to assess
171 healthcare outcomes in a large, diverse, pediatric population.

172

173 **Methods**

174 *Construction of mother-child cohort*

175 We obtained de-identified EMR data through June 30, 2020 from the Mount Sinai Health
176 System (MSHS). In total, we identified 93,154 mother-child pairs delivered at MSHS hospitals,
177 covering 68,893 mothers and 93,154 children.¹¹⁻¹³ The newborns in this cohort were born from
178 2007 to 2019, ensuring that all newborns had at minimum one year of follow-up (see also
179 Supplemental Materials). This study was approved by the Mount Sinai institutional review board
180 (IRB): IRB-20-01771.

181

182 *Digital phenotyping algorithm for rare genetic disorders*

183 The *PheIndex* (Phenotype Index) digital phenotyping algorithm was developed based on
184 13 criteria that may be present in children with a rare genetic disorder. These criteria are
185 primarily based on healthcare utilization patterns such as hospital encounters, procedures,
186 specialist visits, and laboratory test orders. Orders that were subsequently cancelled were not
187 considered. Additional criteria that were included were diagnostic codes of developmental delay
188 and metabolic disease, and death. Description of the criteria with the associated scores is listed in
189 Table 1.

190 *PheIndex* combines these criteria in two different ways: (1) “*PheIndex Score*”, a
191 quantified score indicating the severity of illness with a possible range between 0 and 24
192 generated by the sum of the score(s) associated with the criteria met by a child; and (2)
193 “*PheIndex Classification*”, a binary classification of those who present with illness with an
194 increased risk for genetic disorders (*PheIndex Classification* positive) if the following conditions
195 are met: (a) ≥ 2 major criteria, (b) ≥ 1 major criteria and ≥ 1 minor criteria, (c) ≥ 5 minor criteria, or

196 (d) deceased patient; or those who do not present illness with increased risk for genetic disorders
197 (*PheIndex Classification* negative).

198

199 *Chart review verification of the PheIndex digital phenotyping*

200 For the blinded chart review, we selected 200 charts consisting of children who were
201 *PheIndex Classification* positive (N=100) and *PheIndex Classification* negative (N=100). We
202 ensured that the 100 children who were negative covered quantified scores from 0 to 6
203 (inclusive), and from 3 to 21 for 100 children who were positive, based on the distribution of the
204 *PheIndex Score*. Available records for this review were from encounters dated 01/01/2005 to
205 06/30/2020. All criteria determinations were based on available medical records up until three
206 years of age. The review by the pediatrician had two steps: 1) validate the accuracy of the values
207 assigned to each of the 13 criteria for each patient; and 2) summarize diagnostic information
208 from the patient charts. The pediatrician had access to additional delivery notes, progress notes,
209 admission/discharge summaries, and imaging notes. Information on diagnoses available in the
210 notes documented by the pediatrician was then used by a clinical geneticist to decide whether the
211 child presented with illness with an increased risk for genetic disorders. The possible categories
212 of determination were: 1) “Definitively/possibly has genetic disorder diagnosis”, 2) “Does not
213 have a genetic disorder”, 3) “Unknown, insufficient information to make determination on
214 whether a genetic disorder was related with illness.”

215

216 *Statistical analysis*

217 Full details are described in Supplemental Materials.

218

219 **Results**

220 *Distribution of the 13 criteria in PheIndex*

221 Our cohort included 93,154 newborns linked to 68,893 mothers who delivered in the MSHS
222 from 2007 to 2019, with clinical features collected to 2020 (Table 2). We first assessed the
223 frequency of each of the 13 *PheIndex* digital phenotyping criteria in our cohort and summarized
224 the number of children aged 0 to 3 years old that satisfied each of the 13 criteria (Table 3). The
225 most common criteria were multiple ER visits (3,919; 4.22%), followed by developmental delay
226 (3,159; 3.39%), and multiple visits to specialists (3,091; 3.32%). The least common criteria were
227 metabolic disease diagnosis codes (82; 0.09%) and feeding support (132; 0.14%). Figure 1A and
228 1B demonstrate the expected temporal relationship for achieving each criterion.

229 We generated a heatmap to show the number and percentage of patients who fell into
230 different major and minor criteria combinations (Figure S1). The distribution for the total
231 number of criteria for each child is given in Figure 1C. A large majority of patients (88.51%) did
232 not meet any of the 13 criteria, and 98.55% met ≤ 2 criteria. We showed the distribution of
233 *PheIndex Classification* – children who presented with illness with an increased risk for genetic
234 disorders or not – stratified by the *PheIndex Score* (Figure 1D), as the *PheIndex Classification*
235 depends on the specific combination of major and minor criteria for each patient. The majority of
236 patients had a *PheIndex Score* ≤ 2 (97.23%), indicating that most children in our study
237 population were not likely to have a rare genetic disorder. With our 13 criteria, the *PheIndex*
238 *Classification* identified 1,088 children who were presenting with illness with an increased risk
239 for genetic disorders out of 93,154 children (1.2%).

240 Hospital utilization patterns are known to vary between pre-term and full-term infants,
241 e.g. pre-term infants often have more prolonged NICU stays. To assess this, we computed the

242 similarity between all pairs of *PheIndex* criteria using the Jaccard index for each group (Figure
243 1E and 1F, Supplemental Materials). In the full-term cohort, *heart surgeries* and *prolonged*
244 *NICU stay* had the highest Jaccard similarity of 0.44, in line with what we would expect to
245 observe clinically. In the preterm cohort, prolonged NICU stay was not chosen to be a criterion
246 because the majority of preterm infants have an extended NICU stay regardless of whether they
247 have a rare genetic disorder or not.

248 249 *Validation of PheIndex: 13 Criteria and Overall Classification*

250 First, we evaluated the accuracy of the values that were extracted from the EMR and
251 assigned to the 13 different criteria for each patient, by comparing *PheIndex*'s identification of
252 each of the 13 criteria against a pediatrician's evaluation directly from the clinical notes for each
253 patient, for a sample of 200 children (Table 4). The 200 children were sampled from those
254 classified as presenting with illness with an increased risk for genetic disorders positive for a rare
255 genetic condition (N=100) and those classified as negative (N=100). From this comparison, our
256 digital phenotyping algorithm achieved an average accuracy of 94% across the 13 criteria.
257 Accuracies were $\geq 90\%$ for all criteria except for "*prolonged NICU stays*", which yielded an
258 accuracy of 81%.

259 Next, we compared the *PheIndex Classification* against the classifications made by a
260 pediatrician/medical geneticist (Table 5). Among the 200 children reviewed, 12 patients did not
261 have sufficient clinical information for the medical geneticist to assess whether a genetic
262 disorder may be present. Ten of these 12 patients were born extremely prematurely (born before
263 28 gestational weeks), which led to uncertainties as to whether the criteria that were met was
264 because of prematurity or because of an underlying genetic disorder (as determined by the
265 medical geneticist). Therefore, these 12 patients were excluded from this performance evaluation.

266 Among the 188 patients remaining (88 classified as positive by *PheIndex* and 100 classified as
267 negative), 85 patients were deemed to be true positives (definitively or possibly has a rare
268 genetic disorder by chart review, 90% sensitivity/recall) and 91 patients were deemed to be true
269 negatives (does not have a genetic disorder, 97% specificity). Three patients who were classified
270 as positive by *PheIndex* were not thought to have a genetic disorder (false positive), and 9
271 patients were thought to definitively or possibly have genetic disorders but were classified as
272 having no genetic disorders by *PheIndex* (false negative), yielding a positive predictive value
273 (PPV) of 97%, negative predictive value (NPV) of 91%, and 94% accuracy. If we considered the
274 prevalence of rare genetic disorders to be 3-3.6% of all livebirths,¹⁴ the adjusted PPV ranges
275 from 48.1% to 48.3%.¹⁵

276 277 *PheIndex Scores by Carrier Screening Gene Panel Size*

278 We examined the association between the *PheIndex Score*, an indicator of disease
279 severity, and the three panel sizes (CS-S, CS-M, and CS-L). We first identified that 3 *PheIndex*
280 criteria (multiple inpatient hospital stays, genetic testing, and developmental delay) were
281 enriched for infants whose mothers had performed only CS-S testing compared with CS-M and
282 CS-L (Table 6). For patients with at least 1 year of follow-up, we observed that the overall
283 *PheIndex Scores* were higher in CS-S (mean=0.70) compared to CS-M (mean=0.38, $p<0.001$)
284 and CS-L (mean=0.57, $p<0.001$) (Figure 3A); and CS-S (mean=0.85) compared to CS-M
285 (mean=0.47, $p<0.001$, Student's T-test) and CS-L (mean=0.70, $p<0.001$) for patients with at least
286 2 years of follow-up (Figure 3D).

287 288 *Comparison of time to onset for each criterion for CS-S and CS-L*

289 To investigate the contributions of clinical factors to the *PheIndex* criteria and scores
290 over time, we performed a sub-analysis between the CS-S and CS-L groups. We computed
291 Kaplan-Meier survival curves for each of the 13 criteria for the CS-S and CS-L groups (Figure 4,
292 Supplemental Figure 2) to examine the association of outcomes with time. We found that
293 children in the CS-L cohort were less likely to see multiple specialists ($p < 0.001$, Figure 4A),
294 have multiple visits to the ER ($p < 0.001$, Figure 4B), and less likely to undergo heart surgeries
295 ($p = 0.06$) or die early in childhood ($p = 0.058$), compared to children in the CS-S cohort.

296

297 *Regression Analysis*

298 In the Cox proportional hazards model, we found that for children whose mothers were
299 administered a CS-L panel, a 36% ($p = 0.005$) reduction of being classified as presenting with
300 illness with an increased risk for genetic disorders was estimated, compared to the children
301 whose mothers ordered were administered a CS-S panel test (Supplemental Figure 4).

302

303 **Discussion**

304 Identifying pediatric patients across an entire population with or who possibly has a rare
305 genetic disorder is critical for improving patient outcomes. We and others have attempted to
306 identify patients with specific genetic disorders using EMR data, but have found that such a
307 process is not straightforward, largely due to coding differences, unconfirmed diagnoses,
308 variation in disease names and terminology, and inaccurate information represented in medical
309 records.^{16,17} For most rare genetic disorders, it is difficult to identify patients with specific
310 genetic disorders, given ICD codes are often nonspecific.^{1,18,19} Additionally, seeking to analyze
311 individual diseases, even in EMR databases with millions of patients, would result in

312 underpowered studies given the low frequency of individual rare genetic disorders. However, by
313 using a global metric as opposed to ones derived from specific individual diseases, we were able
314 to identify a large cohort that provided for sufficient statistical power to assess the association of
315 differently sized CS panels with risk of genetic disorders.

316 In this study, we developed a novel, rule-based digital phenotyping algorithm (*PheIndex*)
317 that utilizes 13 criteria to derive a *PheIndex Score* for children from birth to 3 years of age, in
318 order to classify whether a child is presenting with an illness that may be a rare genetic disorder.
319 Importantly, our score is an evaluation of overall health rather than the presence of specific
320 features of individual diseases. To our knowledge, such an approach has not been developed
321 previously. The criteria for the *PheIndex Score* include items that could be extracted from the
322 EMR with a high degree of precision and accuracy. *Our PheIndex Score* may be utilized for
323 various purposes, including its use as a clinical guide to shorten the diagnostic odyssey of hard-
324 to-diagnose patients, timely administration of therapeutics by facilitating more rapid diagnosis,
325 and/or assessing clinical benefit of genetic testing, all of which help enable the practice of
326 precision medicine in a way that may be more accessible to all. Chart review from clinical
327 genetics experts, confirmed that our *PheIndex* algorithm has the following performance
328 characteristics when the numbers of cases and controls are equal: precision of 97%, recall of
329 90%, and accuracy of 94%.

330 To demonstrate the ability of our algorithm to identify an enriched set of patients at risk
331 of harboring a rare genetic condition, we leveraged carrier screening results in mothers who
332 delivered a baby in a large health system. We examined the association between a mother's
333 carrier screening panel size and *PheIndex Score*. We found that CS-L was not only associated
334 with a lower overall *PheIndex Score*, but was also significantly associated with a decreased need

335 for a child having multiple specialist visits and multiple ER visits. In our sub-analysis using a
336 cohort of mother-child pairs whose mother received CS-L or CS-S and with whom the child
337 received at least two years of follow-up, we noted that those in the CS-L group reached the
338 criteria of *multiple specialists* and *genetic diagnostic tests* earlier than those in the CS-S group.
339 This result is notable as it supports that administration of a CS-L panel test may enable earlier
340 diagnosis of genetic disorders in children. Alternatively, testing using a CS-L panel may increase
341 awareness of parental carrier status, thus enabling prenatal diagnostic testing for a larger number
342 of conditions. This increased awareness may also lead to early referral of children manifesting
343 severe illness for rare genetic diagnostic testing and subsequent referrals to the appropriate
344 specialists and potentially earlier treatment. Parental carrier status may also lead to earlier
345 postnatal diagnostic genetic testing and thus confirmation of a particular genetic disorder.

346

347 *Limitations*

348 While our study population is likely representative of other large, diverse metropolitan
349 areas, it may be less representative of smaller-sized cities and rural areas. Also, we provided an
350 adjusted PPV of 48% based on an estimated prevalence of rare genetic disorders in the general
351 population. However, precise estimates of rare genetic disorder prevalence are unavailable, and
352 may also not reflect the PPV for the target population of our algorithm (i.e. children aged 0-3)
353 due to differences in age of onset.¹⁴ Another potential limitation of our study is that we used only
354 de-identified data available in structured EMR databases, and thus did not include all the
355 information that would be available to physicians, such as clinical notes. However, despite not
356 having access to all available clinical notes, our digital phenotype agreed with physician chart
357 review 94% of the time (under conditions in which the number of cases and controls were

358 sampled to be the same), proving that our algorithm successfully identifies children possibly with
359 rare genetic disorders. In the few occasions where there were discrepancies, this was typically
360 due to incomplete documentation of orders, such as respiratory support and feeding support in
361 *PheIndex* negative children that was uncovered in the notes during chart review. Thus, we
362 believe that our digital phenotype will be more accurate using the *PheIndex* criteria extracted
363 from notes in addition to structured EMR data. With respect to demonstrating the application of
364 *PheIndex* across groups receiving comprehensive carrier screening and those receiving reduced
365 or no carrier screening, the MSHS is unique in that beginning in 2016, CS-L was offered to all
366 women considering pregnancy or already pregnant regardless of the mother's ability to pay or
367 health insurance coverage. Notably, our cohort is comprised of linked mother-child pairs and
368 thus does not directly assess the rate of mothers who chose not to continue pregnancies with an
369 affected fetus; however, the improvement in health of children whose mothers received CS-L
370 may be due to couples choosing to proceed with various reproductive health options such as *in*
371 *vitro* fertilization (IVF), in order to reduce the chances of having a child affected with one of as
372 many as 283 genetic conditions assessed in the CS-L panel described in our study. Additionally,
373 mothers who chose to receive CS-L may be more likely to complete additional genetic testing
374 via chorionic villus sampling or amniocentesis in the setting of advanced maternal age or family
375 history of genetic disorders. Moreover, while we controlled for differential follow-up time and
376 likely confounders, there may still be unmeasured confounding in the Cox regression model.

377

378 **Conclusions**

379 In summary, we utilized a comprehensive EMR to develop a novel digital phenotyping
380 algorithm for identification of a pediatric population with a definitive or possible genetic

381 disorder. Our method utilizes a global approach as opposed to identifying patients in the EMR
382 with each specific genetic disorder, which is fraught for misdiagnoses and error. In addition, our
383 study is the first with adequate sample size and follow up to evaluate the health of children from
384 birth to 3 years of age. Using a mother-child cohort that links children to mothers' genetic carrier
385 screening status, we have identified that *PheIndex Scores* are lower at one or two years of
386 follow-up in children whose mothers received CS-L relative to CS-S. We believe that our
387 *PheIndex* algorithm will address an unmet need to identify children with rare genetic disorders
388 and potentially help overcome well-known obstacles such as underdiagnosis and delayed
389 diagnosis.²⁰

390

391 **ACKNOWLEDGEMENTS**

392 We would like to thank Drs. Mitchell K. Higashi and Paul Kruszka for reviewing the manuscript
393 and Mount Sinai Data Warehouse for EMR data. We also thank the GeneDx IT team for
394 infrastructural and computational support.

REFERENCES

1. Garcelon N, Burgun A, Salomon R, Neuraz A. Electronic health records for the diagnosis of rare diseases. *Kidney Int.* 2020;97(4):676-686. doi:10.1016/j.kint.2019.11.037
2. Shirts BH, Salama JS, Aronson SJ, et al. CSER and eMERGE: current and potential state of the display of genetic information in the electronic health record. *J Am Med Inform Assoc.* Nov 2015;22(6):1231-42. doi:10.1093/jamia/ocv065
3. Williams MS, Taylor CO, Walton NA, et al. Genomic Information for Clinicians in the Electronic Health Record: Lessons Learned From the Clinical Genome Resource Project and the Electronic Medical Records and Genomics Network. *Front Genet.* 2019;10:1059. doi:10.3389/fgene.2019.01059
4. Yang Z, Shikany A, Ni Y, Zhang G, Weaver KN, Chen J. Using deep learning and electronic health records to detect Noonan syndrome in pediatric patients. *Genet Med.* Nov 2022;24(11):2329-2337. doi:10.1016/j.gim.2022.08.002
5. Cohen KB, Glass B, Greiner HM, et al. Methodological Issues in Predicting Pediatric Epilepsy Surgery Candidates Through Natural Language Processing and Machine Learning. *Biomed Inform Insights.* 2016;8:11-8. doi:10.4137/BII.S38308
6. Lingren T, Thaker V, Brady C, et al. Developing an Algorithm to Detect Early Childhood Obesity in Two Tertiary Pediatric Medical Centers. *Appl Clin Inform.* Jul 20 2016;7(3):693-706. doi:10.4338/ACI-2016-01-RA-0015
7. Wang Y, Wang L, Rastegar-Mojarad M, et al. Clinical information extraction applications: A literature review. *J Biomed Inform.* Jan 2018;77:34-49. doi:10.1016/j.jbi.2017.11.011
8. Fung KW, Richesson R, Bodenreider O. Coverage of rare disease names in standard terminologies and implications for patients, providers, and research. *AMIA Annu Symp Proc.* 2014;2014:564-72.
9. Horsky J, Drucker EA, Ramelson HZ. Accuracy and Completeness of Clinical Coding Using ICD-10 for Ambulatory Visits. *AMIA Annu Symp Proc.* 2017;2017:912-920.
10. Petrikin JE, Willig LK, Smith LD, Kingsmore SF. Rapid whole genome sequencing and precision neonatology. *Semin Perinatol.* Dec 2015;39(8):623-31. doi:10.1053/j.semperi.2015.09.009
11. Zheutlin AB, Vieira L, Shewcraft RA, et al. A comprehensive digital phenotype for postpartum hemorrhage. *J Am Med Inform Assoc.* Jan 12 2022;29(2):321-328. doi:10.1093/jamia/ocab181

12. Zheutlin AB, Vieira L, Shewcraft RA, et al. Improving postpartum hemorrhage risk prediction using longitudinal electronic medical records. *J Am Med Inform Assoc*. Jan 12 2022;29(2):296-305. doi:10.1093/jamia/ocab161
13. Li S, Wang Z, Vieira LA, et al. Improving preeclampsia risk prediction by modeling pregnancy trajectories from routinely collected electronic medical record data. *NPJ Digit Med*. Jun 6 2022;5(1):68. doi:10.1038/s41746-022-00612-x
14. Ferreira CR. The burden of rare diseases. *Am J Med Genet A*. 2019;179(6):885-892. doi:10.1002/ajmg.a.61124
15. Tenny S, Hoffman MR. Prevalence. [Updated 2022 May 24]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan-. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK430867/> [Accessed 2022 Nov 12]
16. Miller KE, Hoyt R, Rust S, Doerschuk R, Huang Y, Lin SM. The Financial Impact of Genetic Diseases in a Pediatric Accountable Care Organization. *Front Public Health*. 2020;8:58. doi:10.3389/fpubh.2020.00058
17. Tisdale A, Cuttillo CM, Nathan R, et al. The IDEaS initiative: pilot study to assess the impact of rare diseases on patients and healthcare systems. *Orphanet J Rare Dis*. 2021;16(1):429. doi:10.1186/s13023-021-02061-3
18. Aymé S, Bellet B, Rath A. Rare diseases in ICD11: making rare diseases visible in health information systems through appropriate coding. *Orphanet J Rare Dis*. 2015;10:35. doi:10.1186/s13023-015-0251-8
19. Navarrete-Opazo AA, Singh M, Tisdale A, Cuttillo CM, Garrison SR. Can you hear us now? The impact of health-care utilization by rare disease patients in the United States. *Genet Med*. 2021;23(11):2194-2201. doi:10.1038/s41436-021-01241-7
20. Zanello G, Chan CH, Pearce DA; IRDiRC Working Group. Recommendations from the IRDiRC Working Group on methodologies to assess the impact of diagnoses and therapies on rare disease patients. *Orphanet J Rare Dis*. 2022;17(1):181. doi:10.1186/s13023-022-02337-2

Figure Captions

Figure 1. Distribution of *PheIndex* criteria of children in the cohort.

(A,B) Cumulative distribution of time when patients first meet each of the 13 *PheIndex* criteria. Only patients that met each criterion within the three-year limit were included in each cumulative distribution. (A) is sorted by the percentage of patients meeting the criteria at 200 days (least number of patients at the top).

(C) Bar graph showing the showing the number and percentage of patients with passing different numbers of *PheIndex* criteria.

(D) Distribution of *PheIndex* scores for children within the mother-child cohort.

(E,F) Clustered heatmap showing the Jaccard index between possible pairs of *PheIndex* criteria in the pre-term (E) and full-term (F) cohorts. The number and percentage of patients for each criterion are labeled.

Figure 2. Summary statistics of the genetic carrier screening status for newborns in the mother-child cohort.

(A) Top: Numbers of newborns whose mothers were tested with different genetic carrier screens arranged by years of birth (YOB).

Bottom: Percentages of newborns whose mothers were tested with different genetic carrier screens arranged by YOB.

(B) Histogram showing the distribution of genetic carrier tests dates (by month) relative to delivery (inset, weekly).

Figure 3. *PheIndex* Scores across the carrier screening (CS) testing cohort stratified by length of follow-up.

(A) Average *PheIndex* score from the digital phenotype for all children with ≥ 1 year of follow-up in each CS testing cohort. Error bars show 95% confidence interval. *** denotes $p < 0.001$.

(B) Violin plot of *PheIndex* scores using the same categories as (A), but only including children labelled as “positive” from digital phenotype.

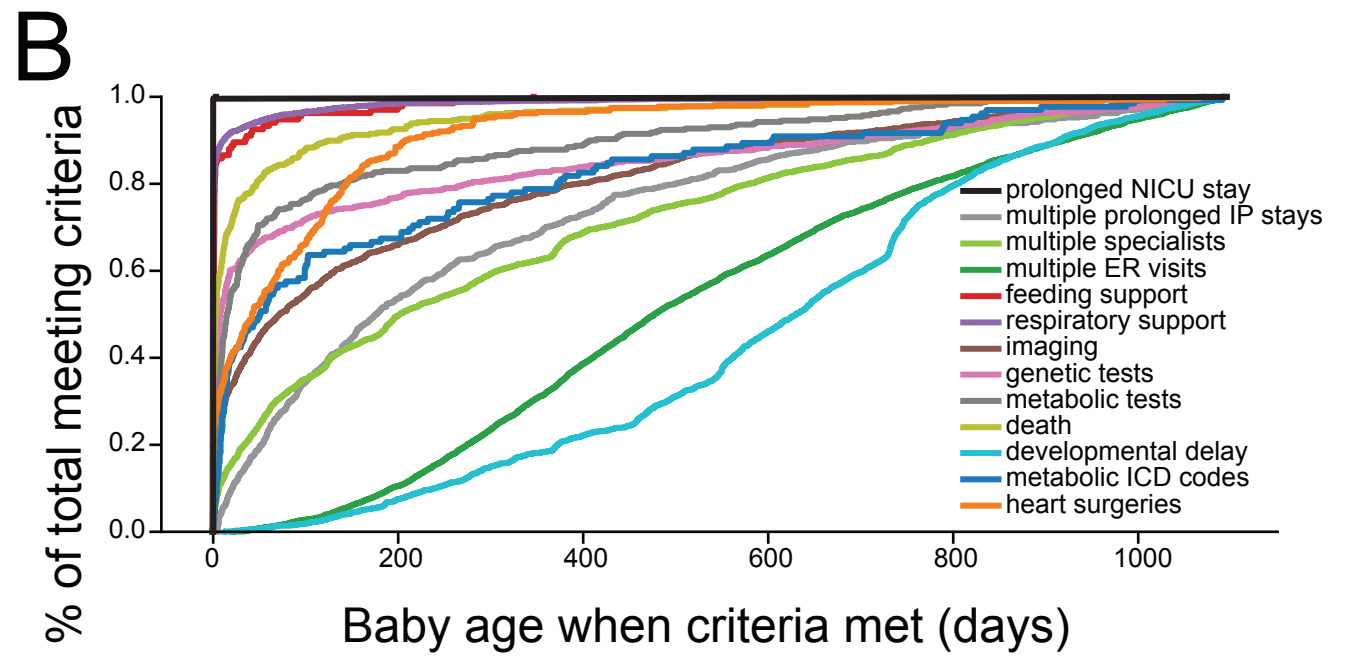
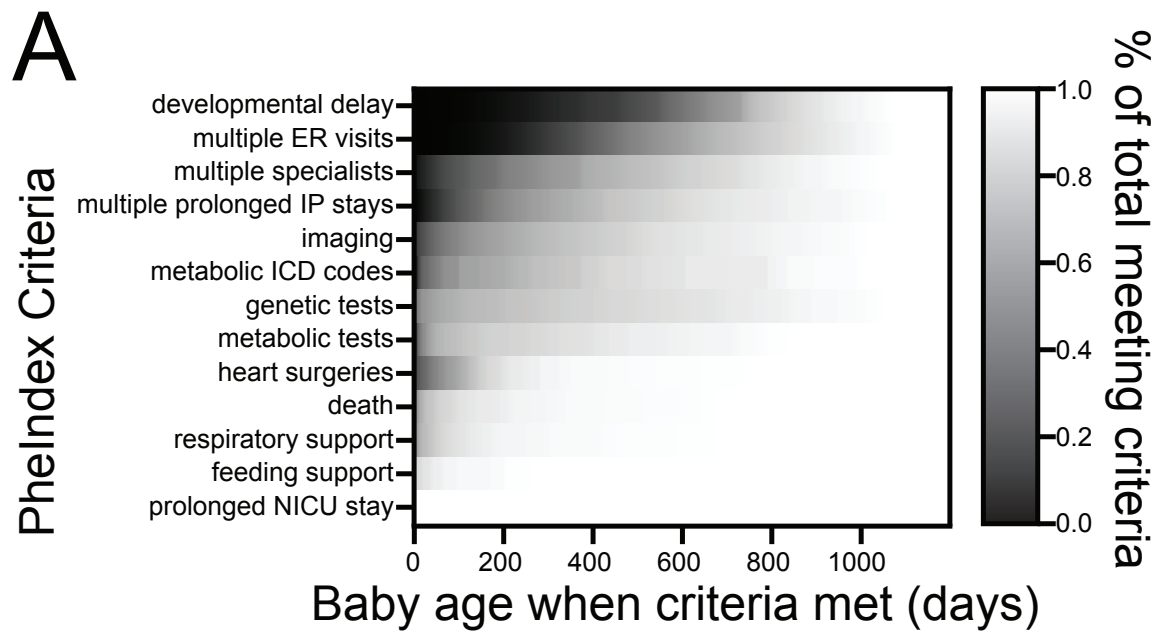
(C) Percentage of children labelled as “positive” for each CS testing cohort. Error bars show 95% confidence interval. (D-F) Same as (A-C) but only including children with at least two years of follow-up.

Figure 4. Kaplan-Meier curves stratified by carrier screening panel size, Carrier screening, small panel / CS-S (blue) vs Carrier screening, large panel / CS-L (orange).

(A) Shows that children whose mothers received CS-S met the multiple specialist criterion in a greater proportion than those whose mothers received CS-L at three years of follow-up.

(B) Shows that children whose mothers received CS-S met the multiple ER visits criterion in a greater proportion than those whose mothers received CS-L at three years of follow-up. Shaded areas denote 95% confidence interval.

Figure 1



medRxiv preprint doi: <https://doi.org/10.1101/2023.01.27.23285056>; this version posted January 31, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY-NC-ND 4.0 International license](https://creativecommons.org/licenses/by-nc-nd/4.0/).

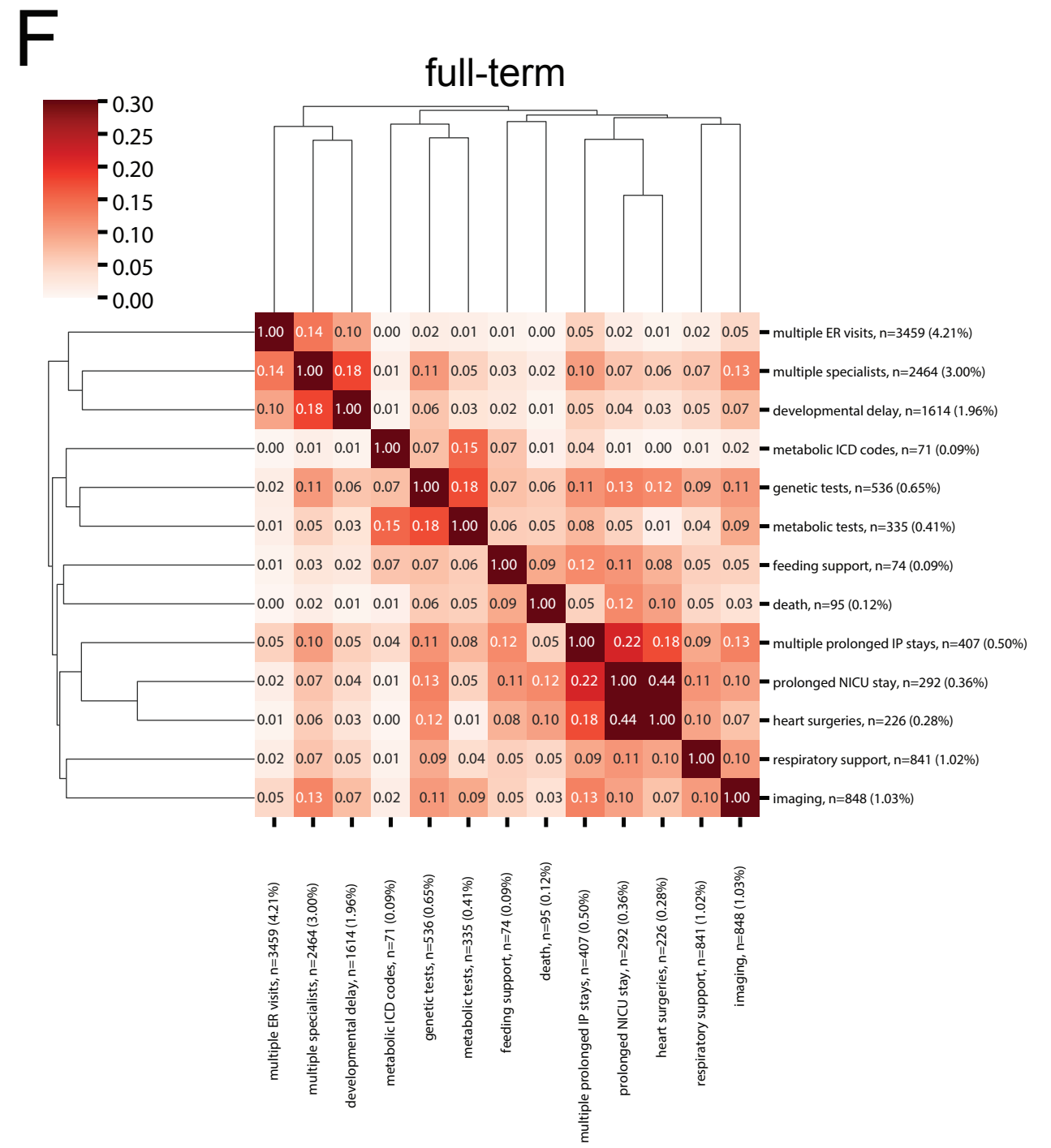
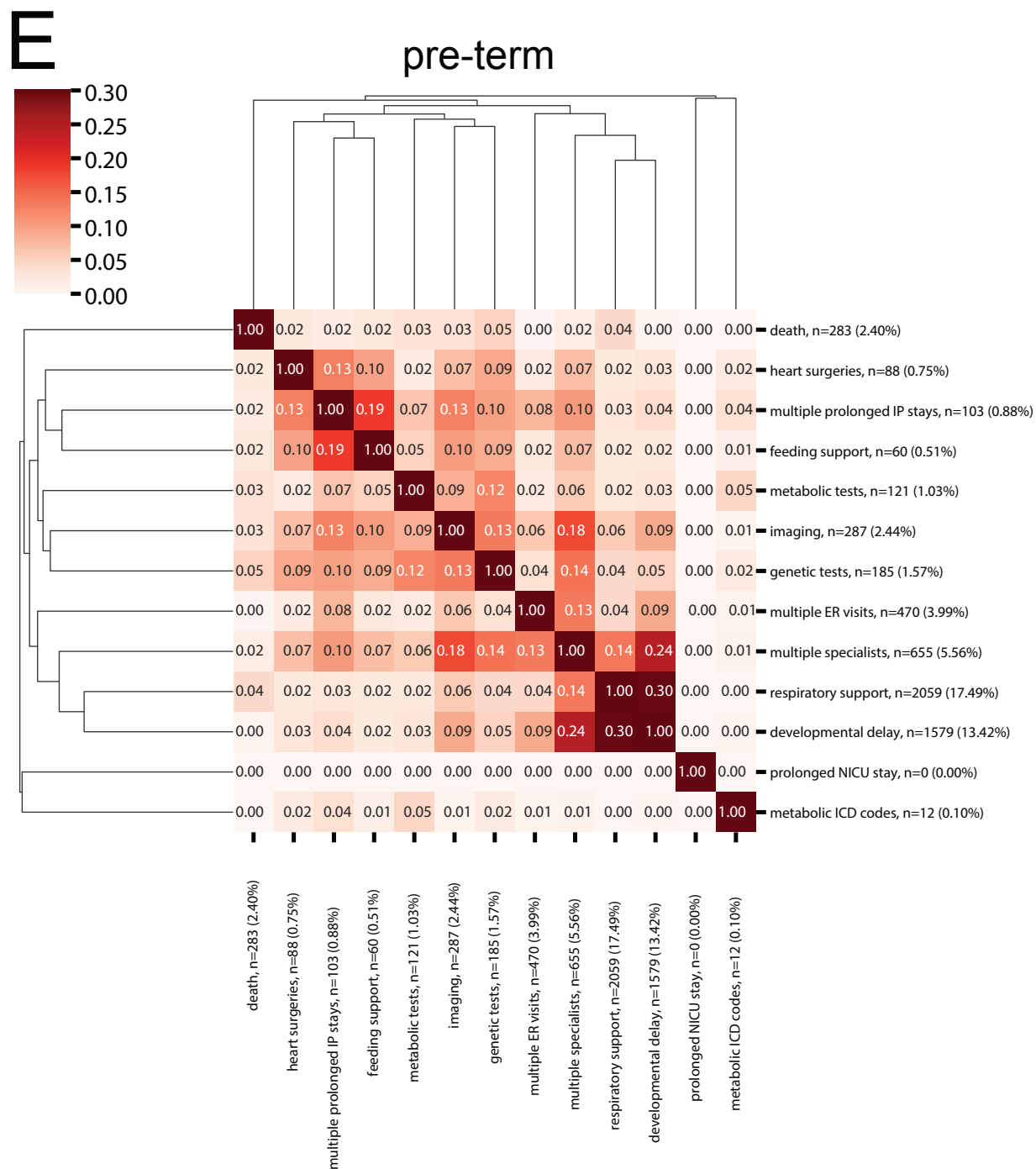
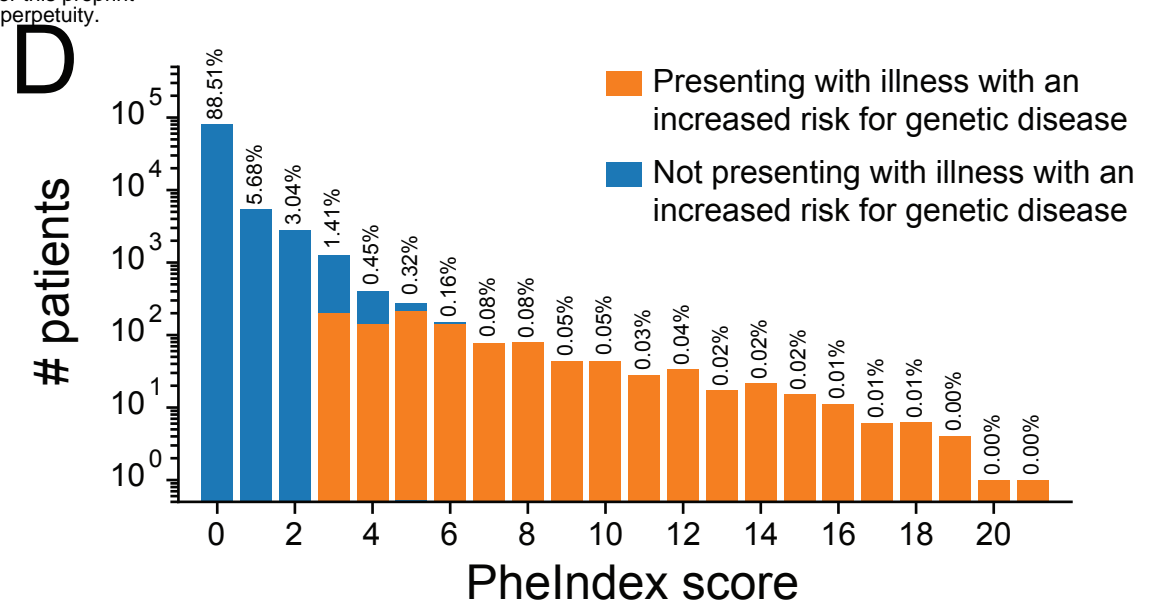
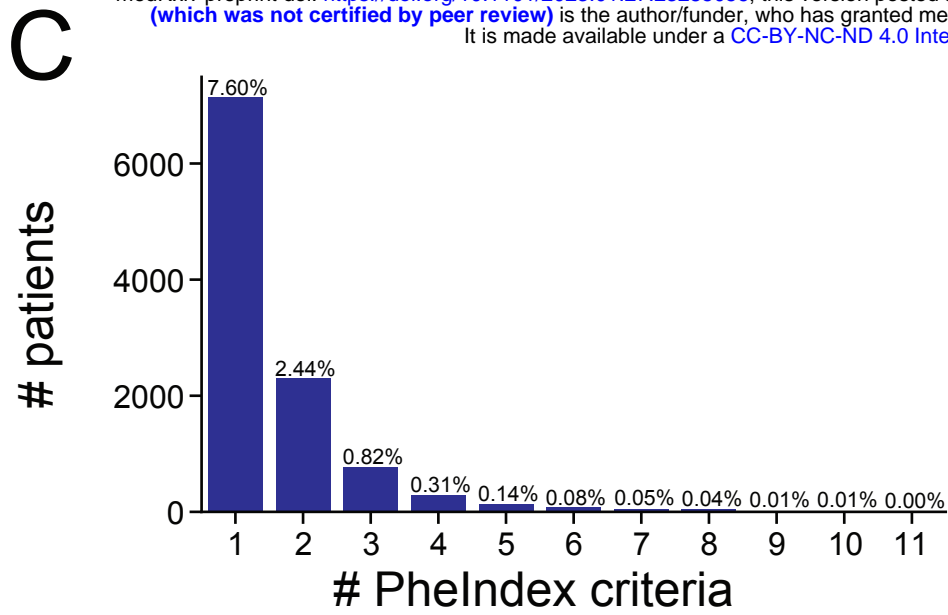
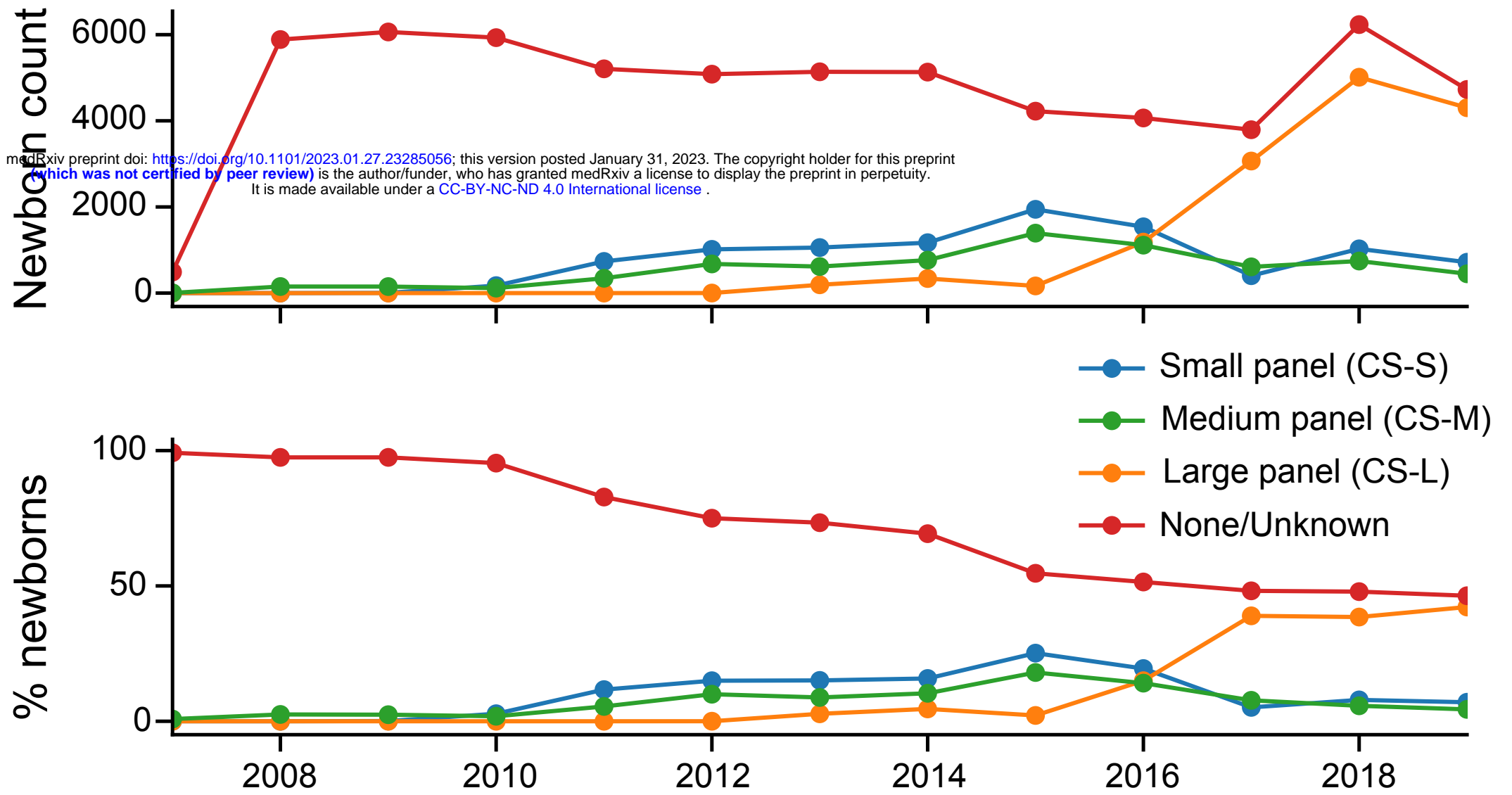


Figure 2

A



B

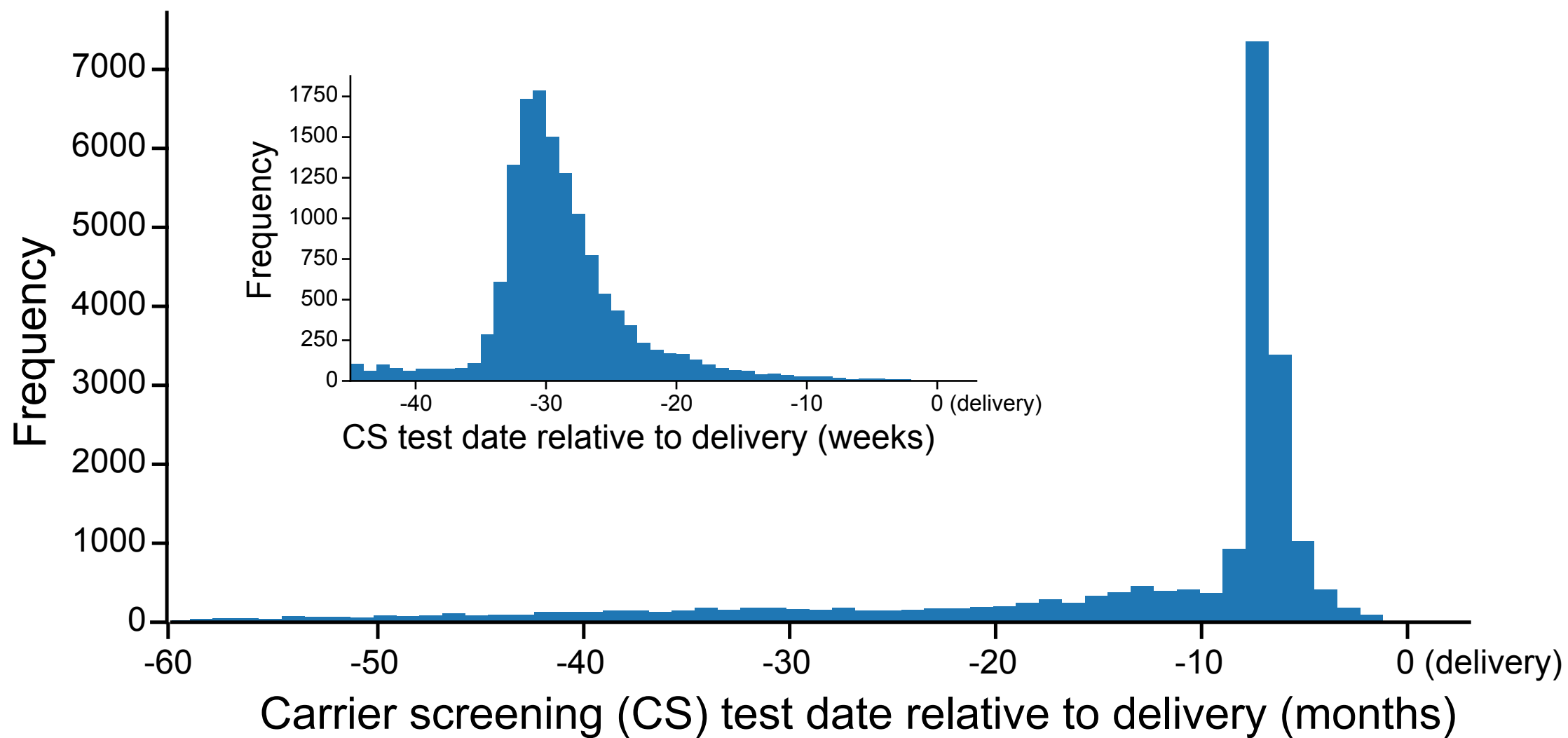
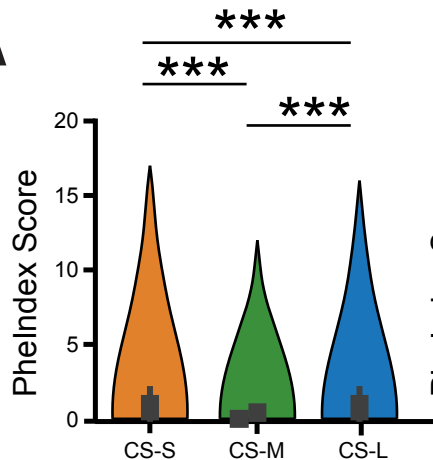


Figure 3

All children

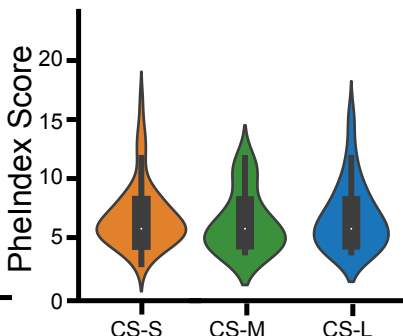
A

>1 year follow-up



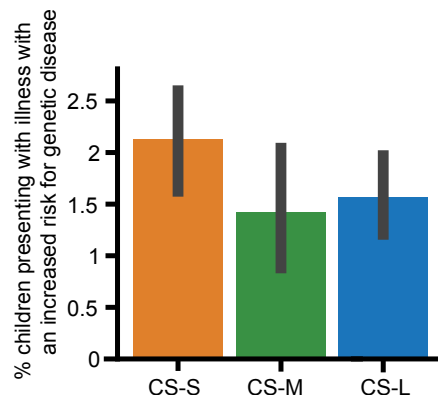
Children classified to have illness with an increased risk for genetic disease

B



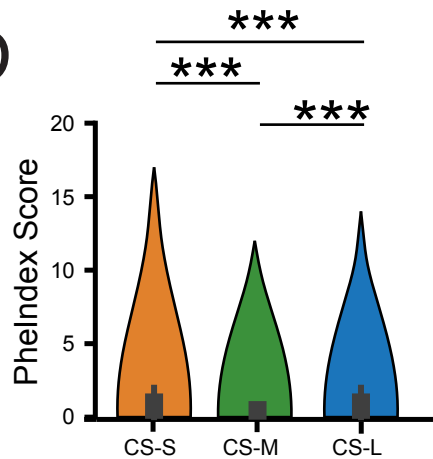
All children

C

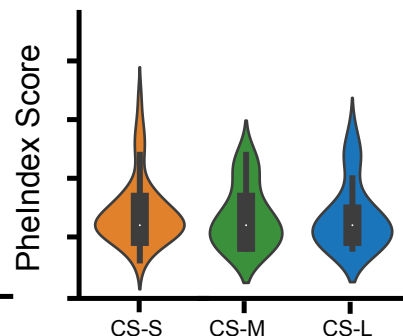


D

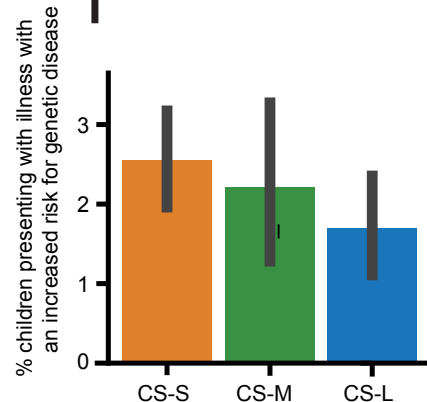
>2 years follow-up



E



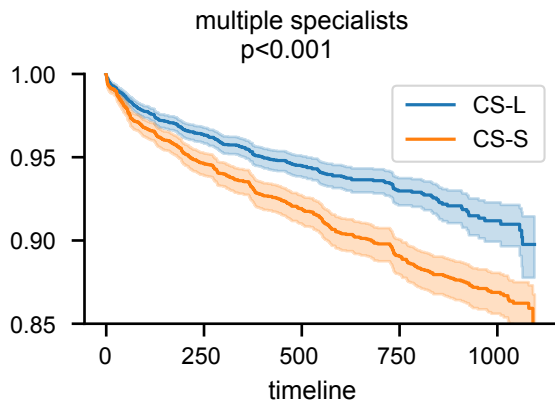
F



CS-S: small panel; CS-M: medium panel; CS-L: large panel

Figure 4

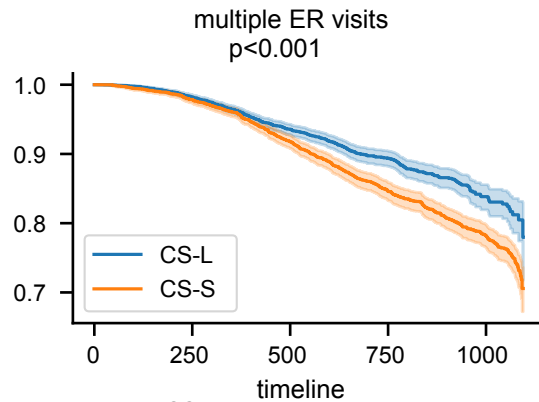
A



	Large panel (CS-L)				
At risk	8270	4574	2774	1369	464
Censored	6022	9493	11221	12594	13482
Events	5	230	302	334	351

	Small panel (CS-S)				
At risk	5759	3493	2674	1968	970
Censored	4028	6054	6783	7416	8374
Events	1	241	331	404	444

B



	Large panel (CS-L)				
At risk	8275	4697	2790	1350	444
Censored	6022	9502	11232	12577	13428
Events	0	98	275	370	425

	Small panel (CS-S)				
At risk	5760	3636	2705	1911	930
Censored	4028	6065	6796	7402	8266
Events	0	87	287	475	592

Table 1: Description and scoring for the 13 *PheIndex* criteria

Description	Scoring
<i>Prolonged stay in the neonatal intensive care unit (NICU) for term babies.</i> Full-term newborns who were admitted to the NICU and stayed for ≥ 4 days.	Major; score = 3
<i>Prolonged or multiple hospitalizations after discharged from birth.</i> Hospitalization is defined as inpatient stays with a duration ≥ 48 hours. We included hospital stays where the calculated gestational age is older than 35 weeks and exclude the first newborn encounter if earlier than 35-week gestation. To meet this criterion, the patient must have either at least one prolonged hospitalization (≥ 14 days) or at least two hospitalizations (≥ 48 hours duration) for full-term or ≥ 3 hospitalizations (≥ 48 hours) for pre-term babies.	Major; score = 3
<i>Visits or consults with multiple specialists other than general pediatricians.</i> Twenty types of specialists were considered: Medical Genetics, Neurosurgery, Pediatric Allergy and Immunology, Pediatric Cardiology, Pediatric Dermatology, Pediatric Endocrinology, Pediatric GI/Pediatric Liver, Pediatric Hematology/Oncology, Pediatric Nephrology, Pediatric Neurology, Pediatric Ophthalmology, Pediatric Orthopedics, Pediatric Otolaryngology, Pediatric Pulmonology, Pediatric Rheumatology, Pediatric Surgery, Pediatric Urology, Transplant, Plastic Surgery. We counted the types of specialists each patient visited or consulted with and not the number of individual specialist visits. Preterm babies with ≥ 4 types of specialists or full-term babies with ≥ 3 types of specialists meet this criterion. We excluded Pediatric Infectious Disease specialty visits as infections in general are primarily due to environmental and non-genetic etiologies, and our aim was to identify a patient population enriched for children with genetic disorders.	Minor; score = 1
<i>Multiple emergency room (ER) visits.</i> Full-term babies with ≥ 5 ER visits or preterm babies with ≥ 7 ER visits meet this criterion.	Minor; score = 1
<i>Feeding support (Gastrostomy tube).</i> Patients who required feeding support were identified using ICD codes (Supplemental Table 2A) and procedures with description of “nasogastric”, “gastrostomy” and “feed”, or “gastrostomy” and “tube” in the procedure name.	Minor; score = 2
<i>Respiratory support (tracheostomy and mechanical ventilation outside of surgery).</i> We used tracheostomy and ventilation (including CPAP) identified by procedure codes and diagnosis codes. If a surgical procedure was performed, the ventilatory support was required to be performed either 1 day before or 5 days after surgeries to be able to meet this criterion.	Minor; score = 2
<i>Imaging.</i> We included patients that received computed tomography (CT) or magnetic resonance imaging (MRI) with completed encounter order status or preliminary/final results available from radiological exams.	Minor; score = 1
<i>Genetic diagnostic tests.</i> We included patients who received genetic diagnostic tests such as gene sequencing or array comparative genomic hybridization regardless of test results. The records of genetic diagnostic tests were retrieved from procedure codes and labs.	Minor; score = 1
<i>Metabolic diagnostic tests.</i> We included patients who received metabolic tests such as plasma amino acids panel or urine organic acids panel, regardless of test results. The records of metabolic diagnostic tests were retrieved from procedure codes and labs.	Minor; score = 1
<i>In-hospital death.</i> Death information was retrieved from discharge location/disposition (expired, to funeral home/morgue or organ harvest) from encounter records.	Major; score = 3
<i>Developmental delay.</i> Patients with developmental delay were identified by either a specialist visit with a developmental pediatrician or at least two occurrences of related ICD codes (Supplemental Table 2B)	Minor; score = 1
<i>Diagnosis codes corresponding to metabolic diseases with ≥ 2 encounters, (major, score=3).</i> We included patients with ICD codes for metabolic diseases (Supplemental Table 2C).	Major; score = 3

Description	Scoring
<i>Heart surgeries, (major, score=3).</i> Newborns that received heart surgeries were identified by encounters related to cardiothoracic surgeries or cardiothoracic intensive care unit (CTICU).	Major; score = 3

Table 2: Demographic information of cohorts by carrier screening status.

		CS-S	CS-M	CS-L	All	p-value	Test
	#	9786	7130	14264	93154		
Demographics & socioeconomics of mothers	Delivery age, median [Q1,Q3]	32.9 [29.4,36.2]	34.3 [31.5,37.2]	33.7 [30.6,36.7]	32.5 [28.2,36.1]	<0.001	Kruskal-Wallis
	Race, n (%)						
	African-American/Black	1475 (15.1)	142 (2.0)	1609 (11.3)	9423 (10.1)	<0.001	Chi-squared
	Asian	1283 (13.1)	238 (3.3)	1437 (10.1)	6911 (7.4)		
	Caucasian/White	3538 (36.2)	6167 (86.5)	7002 (49.1)	52667 (56.5)		
	Hispanic/Latino	2422 (24.7)	226 (3.2)	2180 (15.3)	15543 (16.7)		
	Native American	34 (0.3)	10 (0.1)	39 (0.3)	201 (0.2)		
	Other	820 (8.4)	225 (3.2)	1207 (8.5)	5747 (6.2)		
Unknown	214 (2.2)	122 (1.7)	790 (5.5)	2662 (2.9)			
Health Insurance	Mother on Medicaid, n (%)	2821 (28.8)	229 (3.2)	3258 (22.8)	29219 (31.4)	<0.001	Chi-squared
	Child on Medicaid, n (%)	2461 (25.1)	105 (1.5)	2381 (16.7)	27392 (29.4)	<0.001	Chi-squared
	Child switched to Medicaid, n (%)	25 (0.3)	7 (0.1)	28 (0.2)	154 (0.2)	0.07	Chi-squared
Birth of children	Year of birth, median [Q1,Q3]	2015 [2013,2016]	2015 [2013,2017]	2018 [2017,2019]	2015 [2011,2017]	<0.001	Kruskal-Wallis
	Pre-term birth, n (%)	1299 (13.3)	844 (11.8)	1671 (11.7)	11676 (12.5)	<0.001	Chi-squared
	Birth facility, n (%)						
	Mount Sinai Hospital	8370 (85.5)	6518 (91.4)	9011 (63.2)	79350 (85.2)	<0.001	Chi-squared
	Mount Sinai West	858 (8.8)	262 (3.7)	1934 (13.6)	5916 (6.4)		
Other	558 (5.7)	350 (4.9)	3319 (23.3)	7888 (8.5)			
Record completeness	latest follow-up age (days), median [Q1,Q3]	17.0 [0.0,713.0]	0.0 [0.0,191.2]	4.0 [0.0,401.0]	16.0 [0.0,596.0]	<0.001	Kruskal-Wallis
	# of encounters, median [Q1,Q3]	2.00 [1.00,16.00]	1.00 [1.00,2.00]	2.00 [1.00,12.00]	2.00 [1.00,6.00]	<0.001	Kruskal-Wallis

Note than p-value indicates difference between all carrier screening groups.

Table 3: Number of children passing each individual *PheIndex* criteria.

PhenoIndex Criteria	n (%)
multiple ER visits	3919 (4.2)
developmental delay	3159 (3.4)
multiple specialists	3091 (3.3)
respiratory support	2838 (3.0)
imaging	1113 (1.2)
genetic tests	704 (0.8)
prolonged in-patient stays	500 (0.5)
metabolic tests	448 (0.5)
death	371 (0.4)
heart surgeries	304 (0.3)
prolonged NICU stay	279 (0.3)
feeding support	132 (0.1)
metabolic ICD codes	82 (0.1)

Table 4: Accuracy of digital phenotype algorithm compared to chart review for individual *PheIndex* criteria.

PheIndex Criteria	Accuracy
prolonged NICU stay	81%
prolonged in-patient stays	98%
multiple ER visits	94%
multiple specialists	93%
feeding support	96%
respiratory support	90%
imaging	97%
genetic tests	96%
metabolic tests	96%
death	98%
metabolic ICD codes	97%
developmental delay	93%
heart surgeries	97%

Table 5: Performance of *PheIndex* Classification against chart review.

PhenoIndex Classification		Clinical geneticist classification			Total
		Does not have genetic disease	Definitively or possibly has genetic disease	Unknown (insufficient information)	
	Negative	91 (True Negative)	9 (False Negative)	0	100
	Positive	3 (False positive)	85 (True Positive)	12	100
	Total	94	94	12	200

Table 6: Number of children passing each individual *PheIndex* criteria split by carrier screen testing status.

	CS-S	CS-M	CS-L	All	P-Value	Test
#	9786	7130	14264	93154		
multiple ER visits, n (%)	623 (6.4)	25 (0.4)	424 (3.0)	3919 (4.2)	<0.001	Chi-squared
developmental delay, n (%)	515 (5.3)	138 (1.9)	522 (3.7)	3159 (3.4)	<0.001	Chi-squared
multiple specialists, n (%)	497 (5.1)	121 (1.7)	431 (3.0)	3091 (3.3)	<0.001	Chi-squared
respiratory support, n (%)	348 (3.6)	201 (2.8)	665 (4.7)	2838 (3.0)	<0.001	Chi-squared
imaging, n (%)	132 (1.3)	43 (0.6)	162 (1.1)	1113 (1.2)	<0.001	Chi-squared
genetic tests, n (%)	75 (0.8)	31 (0.4)	95 (0.7)	704 (0.8)	0.03	Chi-squared
prolonged in-patient stays, n (%)	51 (0.5)	15 (0.2)	45 (0.3)	500 (0.5)	0.002	Chi-squared
metabolic tests, n (%)	49 (0.5)	26 (0.4)	69 (0.5)	448 (0.5)	0.38	Chi-squared
death, n (%)	44 (0.4)	10 (0.1)	41 (0.3)	371 (0.4)	0.001	Chi-squared
heart surgeries, n (%)	23 (0.2)	7 (0.1)	17 (0.1)	304 (0.3)	0.03	Chi-squared
prolonged NICU stay, n (%)	24 (0.2)	9 (0.1)	27 (0.2)	279 (0.3)	0.22	Chi-squared
feeding support, n (%)	13 (0.1)	3 (0.0)	13 (0.1)	132 (0.1)	0.16	Chi-squared
metabolic ICD codes, n (%)	3 (0.0)	4 (0.1)	11 (0.1)	82 (0.1)	0.38	Chi-squared
# of <i>PheIndex</i> criteria, mean (SD)	0.24 (0.69)	0.09 (0.41)	0.18 (0.59)	0.18 (0.63)	<0.001	One-way ANOVA
<i>PheIndex</i> Score, mean (SD)	0.31 (0.98)	0.13 (0.62)	0.24 (0.87)	0.25 (0.97)	<0.001	One-way ANOVA