Published in final edited form as:

Pediatr Res. 2020 October ; 88(4): 653–660. doi:10.1038/s41390-020-0796-7.

Prediction of Short-Term Neonatal Complications in Preterm Infants Using Exome-Wide Genetic Variation and Gestational Age: A Pilot Study

William C L. Stewart[2,5,6], **Komla M. Gnona**[1,2,4], **Peter White**[3,5], **Ben Kelly**[3,5], **Mark Klebanoff**[1,5], **Irina A. Buhimschi**[1,7], **Leif D. Nelin**[1,5]

[1]Center for Perinatal Research, Columbus, OH

[2]Battelle Center for Mathematical Medicine, Columbus, OH

[3]Institute of Genomic Medicine at The Abigail Wexner Research Institute at Nationwide Children's Hospital, Columbus, OH

[4]Biophysics Graduate Program, The Ohio State University, Columbus, OH

[5]Pediatrics, The Ohio State University, Columbus, OH

[6]Department of Statistics, The Ohio State University, Columbus, OH

[7]Department of Obstetrics and Gynecology, University of Illinois, Chicago, IL

Abstract

Background—Preterm birth is the leading cause of mortality and morbidity in young children with over a million deaths per year worldwide arising from neonatal complications (NC). NC are moderately heritable although the genetic causes are largely unknown. Therefore, we investigated the impact of accumulated genetic variation (burden) on NC in Non-Hispanic White (NHW) and Non-Hispanic Black (NHB) preterm infants.

Methods—We sequenced 182 exomes from infants with gestational ages from 26 to 31 weeks. These infants were cared for in the same time period and hospital environment. Eighty-one preterm infants did not develop NC, whereas 101 developed at least one severe complication. We measured the effect of burden at the single-gene and exome-wide levels, and derived a polygenic risk score (PRS) from the top 10 genes to predict NC.

Results—Burden across the exome was associated with NC in NHW (p=0.05) preterm infants suggesting that multiple genes influence susceptibility. In a *post hoc* analysis, we find that PRS

Corresponding Author: Komla Gnona, 575 Children's Crossroad, Columbus, OH 43215, Phone number: 614-722-2210, komla.gnona@nationwidechildrens.org.

Author Contributions: WCLS, KMG, PW, BK, MK, IAB and LDN made substantial contributions to conception and design, acquisition of data, or analysis and interpretation of data. WCLS and KMG drafted the article, and PW, BK, MK, IAB and LDN revised it critically for important intellectual content. All authors gave final approval of the version to be published.

Statement of Originality: This work is original, not previously published, and has not been submitted for publication or consideration elsewhere.

Disclosure Statement: The authors have nothing to disclose.

alone predicts NC (AUC=0.67) and that PRS is uncorrelated with GA ($\hat{\rho} = 0.05$; p=0.53). When PRS and GA at birth are combined, the AUC is 0.87.

Conclusion—Our results support the hypothesis that genetic burden influences NC in NHW preterm infants.

INTRODUCTION

Preterm birth is the leading cause of mortality and morbidity in young children (age < 5 years) worldwide, with more than a million deaths per year arising from neonatal complications (NC) (1–3). Therefore, to improve the health outcomes of preterm infants in the era of precision medicine and to reduce medical costs, we need to develop better predictive models for NC. One attractive approach is to combine well-established risk factors (e.g. gestational age at birth) with genetic risk factors (e.g. exonic mutations) into a single predictor of NC.

Based on evidence from twin studies, we know that genes play a role in individual neonatal complications like bronchopulmonary dysplasia (BPD) and retinopathy of prematurity (ROP). In particular, Bhandari et al (4) estimated the heritability of BPD at 53% in a multicenter twin study using logistic regression with mixed effects; and in a retrospective twin study of ROP, Bizzaro et al estimated the heritability at 70% (5). However, pinpointing the specific genes that account for the impact of sex and race on susceptibility for NC has been much more difficult.

Epidemiological studies demonstrate that NC may also be influenced by gestational age (GA), sex, and race. For example, Draper et al(6) showed that GA is negatively correlated with risk for NC. In addition, Trembath A et al. (7) and Peelen et al. (8) showed that preterm males have increased risk for NC relative to preterm females. Finally, Loftin et al (9) and Ryan et al (10) showed that conditional on GA, Non-Hispanic White infants have increased risk relative to Non-Hispanic Black infants.

The objective of this study is to test the hypothesis that, among preterm infants, the accumulation of genetic variation across coding regions of the genome (i.e., the exome) influences risk for NC. As a corollary, we sought to determine if any observed sex and race disparities in NC relate to the burden as defined by the gene-specific accumulation of minor alleles found by whole exome sequencing (WES). Our study was partially motivated by several promising examples in the study of complex traits for adult diseases and morbidities (e.g. schizophrenia, Parkinson’s disease, and obesity), where investigators have found associations with the accumulation of minor alleles (11–13). Lastly, in a *post hoc* analysis, we compared the predictive power of a burden-based polygenic risk score (PRS) and a composite biomarker that combines PRS and GA into a single predictor of NC. Overall, we (1) demonstrate that NC are influenced by the accumulation of minor alleles found by WES in Non-Hispanic White preterm infants (2) did not detect an effect of burden on NC in Non-Hispanic Black preterm infants (3) confirm the effects of previously reported traditional risk factors (e.g. GA) and show that the impact of minor allele accumulation is independent of GA (at least within the GA range studied) and (4) show that susceptibility to NC can be accurately predicted by a composite biomarker that combines GA and PRS into a single predictor of NC.

METHODS

Study design, patient population and samples.

This study was approved by the Institutional Review Board at Nationwide Children's Hospital. The study utilized the Perinatal Research Repository (PRR), which is a data and biospecimen repository of preterm neonates admitted to the neonatal intensive care unit (NICU) at Nationwide Children's Hospital (NCH). Parents of infants eligible for inclusion in the PRR (i.e. infants <37 completed weeks of gestation) were approached to provide written informed consent for their participation and for the participation of their child. Once consent was obtained, blood or buccal swabs were obtained for DNA extraction. Samples were processed for DNA extraction at the Nationwide Children Hospital Biopathology Center and stored at −80°C until analysis. Clinical data was abstracted from the electronic medical record upon death or discharge by research personnel who were not directly involved in the present study.

Of the infants enrolled in PRR, eligible preterm infants for this study were singletons with gestational age at birth between 26 and 31 weeks, inclusive. Infants with known chromosomal abnormalities or congenital anomalies were *a priori* excluded. The gestational age range was chosen on the rationale that for infants with gestational age less than 26 weeks, extreme immaturity could potentially overwhelm any genetic component. Conversely, severe NC are far less common in infants born at 32 weeks gestation or more. Because both outcomes tend to reduce statistical power, restricting gestational age to the range of 26 to 31 weeks would likely provide the greatest power to detect genetic factors influencing NC.

Clinical phenotype of relevant neonatal complications (NC) and study groups

Among eligible infants with samples and data available as of December 2015 we defined as *"susceptible"* (SUS) the infants who were diagnosed at death or discharge with at least one of the of the following severe short-term neonatal outcomes.

- BPD was defined as a requirement for oxygen and/or positive airway pressure at 36 weeks post-menstrual age.(14)
- NEC was defined as Bell's Stage 2 or greater(15)
- ROP was defined as stage 2 or greater according to the ICROP (16),
- Severe IVH was defined as grade 3 or greater according to the Papile classification.(17)

We defined *"resilient"* (RES) as those infants that had none of the NC listed above. For each group, we recorded the following: gestational age at birth, the birthweight, Apgar scores (at 1min and 5min), race, sex, delivery route, any exposure to antenatal steroids, whether surfactant was given, and maternal characteristics initiating birth.

We used Wilcoxon rank sum tests to compare quantitative variables and chi-squared tests to compare categorical variables. A *p*-value <0.05 was considered significant.

Sequencing

WES was performed on each infant using the SureSelectXT Target Enrichment System for Illumina Paired End Sequencing Protocol (Agilent Technologies, CA). DNA libraries were captured and enriched for exons using the SureSelect Clinical Research Exome version 1 kit (Agilent). Paired-end 96 base pair reads were generated for exome-enriched libraries sequenced across eight Illumina HiSeq 2500 runs. Samples were sequenced to an average of 76X depth of coverage, with a minimum depth of 50X targeted region coverage.

Following sequencing, primary data analysis consisted of using Illumina's Real-Time Analysis software to perform base calling and quality scoring from the raw intensity files. The resulting base call format files were then converted and demultiplexed using Illumina's bcl2fastq2 Conversion Software into the standard FASTQ file format appropriate for secondary analysis.

Secondary analysis was performed using Churchill, a pipeline developed in house for the discovery of human genetic variation that implements a best practices workflow for variant discovery and genotyping(18). Churchill utilizes the Burrows-Wheeler Aligner to align sequence data to the GRCh37/hg19 reference genome. Duplicate sequence reads were removed using PicardTools (version 1.104). Local realignment was performed on the aligned sequence data using the Genome Analysis Toolkit (version 3.3-0) Churchill's own deterministic implementation of base quality score recalibration was used. The GATK's HaplotypeCaller was used to call variants. All analysis was performed by uploading FASTQ files to GenomeNext LLC, which automated execution of the Churchill pipeline for the entire dataset. Resulting VCF files were downloaded from GenomeNext for subsequent analysis. We used GRCh37/hg19 from the University of California at Santa Cruz database (19–21) and the 1000 Genome project phase 3(22) for reference human genome annotation.

Assessment of genetic burden

Burden can be tested on multiple levels (i.e. at the level of individual genes, the whole-exome, and across a selected set of selected genes).

1. Single-gene Burden—We defined single-gene burden as the total count of minor alleles across a given region which includes 7.5 kilobases upstream and downstream flanking sequences. Gene regions were further filtered to remove genes with excessive amounts of missing data (i.e. genes with more than 90% missing sequence data), variants with extremely low reads (i.e. number of reads less than one), and variants with extremely high reads (i.e. number of reads exceeds 2.5 standard deviations from the median). These analyses were conducted for each infant and for all genes in the human genome.

However, because our sequence data are organized around the count of alternate alleles at polymorphic sites, and because alternate alleles are not necessarily minor alleles, computing the single-gene burden is not trivial. Specifically, we scored the number of minor alleles at each site as follows:

- 0 if the site is homozygous for the alternate allele and the alternate allele is major. (i.e., no minor alleles are present)
- 1 if the site is heterozygous (site carries exactly 1 minor allele)
- 2 if the site is homozygous for the alternate allele and the alternate allele is minor (i.e., both alleles present are minor)

Note in our data, sites which are homozygous for the reference allele are not recorded. That said, if we index infants by *i*, genes by *j*, and polymorphic sites by *k*, then the single-gene burden of the j^{th} gene in the i^{th} infant is $B_{ij} \equiv \sum_k M_{ijk}$. For each gene, we use a logistic regression (with GA as covariate) to test single-gene burden for association to NC; SUS is coded as 1 (i.e. high risk) and RES is coded as 0 (i.e. low risk). To correct for the number of multiple tests we implemented a Bonferroni procedure.

2. Exome-wide Burden—In contrast to our single-gene test (described above), we also assessed the joint effect of *all* genes by looking for an excess of low *p*-values (i.e. p<0.05) among the observed burden *p*-values. (Note, burden *p*-values were corrected for heteroscedasticity (23, 24)). Because burden is discrete, the distribution of burden *p*-values is not uniform under the null hypothesis of “no association”. Therefore, to determine the statistical significance of the observed burden *p*-values, we performed a permutation test(25, 26). Specifically, we first computed the *observed* area under the curve (AUC) from the cumulative distribution of the burden *p*-values. Then, we permuted preterm infant status (e.g. RES and SUS) 10,000 times to obtain a permutation distribution for AUC. Because a large *observed* AUC is evidence for an excess of low *p*-values, we used the proportion of permutations with AUC larger than (or equal to) the *observed* AUC to estimate the permutation *p*-value.

3. Burden-based Polygenic Risk Scores—Using the results of our whole-exome association study (WEAS) of NC in Non-Hispanic Whites, we derived a polygenic risk score based on burden. Specifically, for the i^{th} infant we computed the polygenic risk score (PRS) (27) as $\sum_i \hat{\beta}_{ik} B_{ik}$ where for the k^{th} gene, $\hat{\beta}_k$ is the estimated coefficient for burden and B_{ik} is the observed burden in the k^{th} infant. The summation is taken over the burden of 10 genes showing the strongest evidence for association. We then used the “ROCR” package in R (28) to compute the Receiver Operating Characteristic (ROC) curve (29) and the corresponding AUC for PRS alone. For comparison, we also computed the ROC curve and AUC for GA alone and for our proposed composite biomarker, where the estimated coefficients from a logistic regression were used to combine PRS and GA into a single predictor of NC. Then, we averaged the AUC over a 10-fold cross-validation procedure (30) to mitigate the negative effects of over-fitting. Finally, we used the average AUC and a 95% confidence interval to compare the predictive power of each predictor.

RESULTS

Clinical characteristics of the study groups

The workflow of the subjects with inclusion and exclusion criteria is summarized in Figure 1. There were 287 eligible babies of which 94 fulfilled SUS criteria and 125 were classified as RES, after excluding for multiple births. Of these 219 infants, high quality exome sequencing data was successfully generated on 209 newborns (SUS: n=90, RES: n=119). In anticipation of future meta-analyses, we did not exclude eligible candidates based on race or sex. However due to analytical limitations of understanding the clinical significance of burden in admixed populations with small sample sizes we restricted our analyses to Non-Hispanic White (n=131) and Non-Hispanic Black (n=51) infants. Two infants in our study died as a result of their complications. Table 1 summarizes the clinical outcomes, and Table 2 summarizes the clinical characteristics of the SUS and RES groups limited to Non-Hispanic Whites and Non-Hispanic Blacks.

As anticipated, the SUS group had a much lower birthweight (p<0.001) and gestational age (p<0.001) at birth. The SUS group also had more (p<0.05) male newborns than did the RES group. In addition, the Apgar scores at 1min and 5min also were statistically different (p<0.009 and p<0.03 respectively) between the two groups of infants with SUS group having somewhat lower scores. There was no significant difference in delivery route (p<0.38), in antenatal steroids (p<0.20), or surfactant administration (p<0.96) between the two groups.

Confirmation of known risk factors in our study groups

Several authors have previously reported associations between NC and gestational age (31), between NC and sex (1, 32), and between NC and race (32, 33). Therefore, we sought to confirm these findings in our sample of 182 preterm infants. First, we found that gestational age (GA) is associated with NC in Non-Hispanic Whites [OR_{GA} = 0.43; 95% CI: (0.31, 0.56)], and in Non-Hispanic Blacks [OR_{GA} = 0.42; 95% CI: (0.24, 0.63)]. Second, we tested for an association between sex and NC with gestational age as a covariate. Although the evidence for association was not statistically significant, the effect of male sex may be stronger in Non-Hispanic Blacks [OR_{Sex} = 3.6; 95% CI: (0.84,17.3)] than in Non-Hispanic Whites [OR_{Sex} = 1.9; 95% CI: (0.77,4.83)].

Within in each race, male preterm infants appear to have higher risk of NC than female preterm infants. Third, we found suggestive evidence for an association between race and NC with gestational age as a covariate (OR_{Race} = 2.2; 95% CI: (0.98, 5.22)], implying that Non-Hispanic White preterm infants may be more susceptible to NC than Non-Hispanic Black preterm infants.

Exome-wide burden associates with NC in Non-Hispanic White infants

Using the human genome reference annotations (GRCh37/hg19) from the University of California at Santa Cruz database (19–21) we identified a total of 27939 gene regions. After filtering of the WES data, we retained a total of 23,854 gene regions for our analysis of data from Non-Hispanic White infants and 20,232 gene regions for analysis of Non-Hispanic Black infants. On average, each infant provided sequence data at about 273,168 variants.

In our study, burden was not statistically significant for any gene at the exome-wide level in either cohort (i.e. Non-Hispanic White and Black preterm infants) (Figure 2). However, after permuting infant status (i.e. SUS and RES) 10,000 times and using AUC as our test statistic, we did find an excess of low p-values across all genes in Non-Hispanic Whites (p=0.05;

Figure 3). The excess of low *p*-values suggests that burden influences NC through multiples genes of small effect, especially since no single gene was statistically significant at the exome-wide level.

Evaluating various predictors of NC

We compared the ROC curves of PRS alone and our composite biomarker (PRS+GA) in terms of AUC (Figure 4). Recall that our composite biomarker combines gestational age and PRS into a single predictor of NC, where PRS is computed from the top ten genes (Table 3) of the exome-wide analysis for Non-Hispanic White infants. However, this comparison does not account for over-fitting. To compare the predictive power of PRS alone and PRS+GA without over-fitting, we averaged the AUC of each predictor in a 10-fold cross-validation procedure (see Methods). The average AUC for PRS alone was 0.67 (p<0.003), and increased to 0.87 (p<0.001) when GA was combined with PRS (Table 4). Note that the average AUC based on PRS+GA is significantly larger than the average AUC based on PRS alone (p=0.0012). Moreover, for each gene in the PRS-gene set, the average burden was higher among SUS preterm infants than RES preterm infants (data not shown). This suggests that large values of PRS are associated with increased risk of NC. Interestingly, PRS and GA were not correlated [$\hat{\rho} = 0.05$; p=0.53]. This may explain why the predictive power of our composite biomarker, which combines PRS and GA into a single predictor of NC, is so high compared to the predictive power of PRS alone.

DISCUSSION

Preterm infants are at increased risk for NC (neonatal complications), and as such, predicting the health outcomes of preterm infants could have a tremendous impact on their outcomes, and furthermore could be extremely useful in the development of preventative therapies by identifying high-risk populations. In this study, we confirm that gestational age is an important predictor of NC; and we find suggestive evidence for an effect of race. Our data also indicates that male sex may play a stronger role in Non-Hispanic Blacks than Non-Hispanic Whites. Furthermore, we show that the combination of traditional risk factors and genetic risk factors can substantially improve prediction. In what follows, we discuss each of these findings, their implications, and any relevant limitations of our study.

1. Genetic Predictors of Neonatal Complications

We demonstrate that in aggregate (i.e. across genes) the accumulation of variants found by whole-exome sequencing is positively correlated with NC, but as with most whole-exome association studies, there are potential limitations. First, there’s always the question of how *should* one summarize variation across potentially overlapping, nested, and alternatively spliced genomic regions. Here, we took the simplest possible approach and assessed burden at the gene-level, where the start and end positions of each gene were given by the UCSC GRCh38/hg38 annotation file (see Data Description for more details). Second, because exons are highly conserved, exonic variation in human populations is typically low(34). As such, WEAS often require large sample sizes to detect an association(35). Because it is often difficult to collect large samples of preterm infants (36), most WEAS studies of preterm infants are underpowered. For example, given the size of the genetic effects that we see in our Non-Hispanic White infants, our Non-Hispanic Black sample (*n*=51) is probably too small to detect an association at either the *single-gene* level, or at the *exome-wide* level.

Similarly, in Non-Hispanic Whites, we investigated the role of intronic variants by comparing genes with "little or no" evidence for association (i.e. *p*-values > 0.1) to genes with "suggestive" evidence for association (i.e. *p*-values < 0.1). In genes with "little or no" evidence, 87% of all variants were intronic, whereas, in genes with "suggestive" evidence 89% of all variants were intronic. While intronic variation could play an important role in the genetics of NC, we would need many more preterm infants to ensure that the observed increase of 2% is, in fact, statistically significant.

In a *post hoc* analysis, we tested our newly discovered PRS-gene set, which contains 10 genes showing the largest evidence for association. There's almost always some difficultly in deciding *exactly* how many genes to include in a PRS risk score(27, 37). Nevertheless, we decided to use the top ten genes because the degree of over-fitting seemed acceptably small (data not shown). To further mitigate the potentially negative effects of over-fitting, we chose to implement a 10-fold cross validation procedure. Interestingly, our polygenic risk score does not appear to be correlated with GA, which likely stems from the fact that our logistic regression included GA as a covariate, but may also suggest that NC and prematurity have different genetic etiologies.

2. Race and Sex as Predictors of NC

To date, the strongest predictors for NC are gestational age [MANUCK, 2016], birth weight (38, 39), sex (40) and race (10). While the evidence for an association between race and NC is *only* suggestive in our data (*p*=0.06), when we imputed the percentage of African ancestry in each preterm infant from the available exome sequence data, we found that imputed African ancestry was negatively correlated with risk for NC (*p*=0.049; data not shown). This suggests that preterm infants of African descent may be less likely to develop NC. Furthermore, our results (and the results of Morse et al (32)) suggest that the effect of sex may depend on race. Lastly, our PRS gene set predicted NC poorly in Non-Hispanic Black infants. Here, the PRS was computed from the PRS gene set (and the corresponding regression coefficients) identified in Non-Hispanic Whites, and from the observed burden of Non-Hispanic Black infants. Although the relatively small number of Non-Hispanic Black infants in our study (*n*=51) could explain our inability to detect an association, another possible explanation is that NC in Non-Hispanic Blacks and Non-Hispanic Whites are influenced by different genes.

3. Study Limitations

To the best of our knowledge, this is the first-ever whole-exome sequencing study of NC, and as such, there are limitations that we describe below. First, although we have "lumped" several complications together into a single trait (i.e., NC), the genetic heterogeneity in our sample of preterm infants is mitigated (to some degree) by the fact that more than 90% of our SUS infants have BPD (see Table 1). Second, our study is in some sense a study of extreme phenotypes because we have RES (resilient) infants that are very preterm (i.e., GA between 26 and 31 weeks); and relative to a standard case-control design, using extreme phenotypes in this way increases our power to detect genetic factors associated with NC.

4. Candidate Genes and Improved Biomarkers for NC

Among the 10 genes composing our PRS score for neonatal complications, the top score was assigned to RANBP2 (Ran-binding protein-2) a protein located on the cytoplasmic surface of the nuclear pore complex that plays a role in intracellular trafficking(41). Although additional studies need to investigate the mechanistic role of RANBP2 in the pathogenesis of NC, there is some biologically plausibility based on what is known so far about this protein. First, the mouse RANBP2 knockout is embryonic lethal suggesting RANBP2 plays an important role in development. And studies on conditional knockout mice linked the cause of lethality to inadequate nuclear import, although whether this affected a broad spectrum of proteins or a small subset remains to be determined(42). Second, several RANBP2 mutations in children are known to cause acute necrotizing encephalopathy (ANE), a disorder where previously normal children develop encephalopathy in response to a common viral infection. Therefore, the possible involvement of RANBP2 in a broader adaptive response such as the ones required by premature newborns in the context of a NICU environment needs to be studied (43).

Another gene among the top 10 with direct biological plausibility is GUCY1A3. This gene encodes for the α1 subunit of the soluble guanylyl cyclase (sGC) enzyme, and it is important in the nitric oxide/cGMP signaling pathway—a pathway that regulates sensitivity to nitric oxide. Note that, nitric oxide is an established mediator of newborn lung development; and when inhaled, it has been proposed to exert therapeutic benefit to prevent BPD. While randomized clinical trials have yielded conflicting results on the protective effect of inhaled nitric oxide in the general population, there is the suggestion of a possible subgroup benefit in non-white infants. Furthermore, a polymorphism in GUCY1A3 has been previously associated with decreased risk for pulmonary hypertension in a high-altitude population(44).

Although we limited our candidate gene discussion to RANBP2 and GUCY1A3, future mechanistic investigations of NC should be extended to other genes as well, especially because genetic burden could have additive effects in genes with seemingly unrelated function. Furthermore with regard to BPD specifically, it would be interesting to determine if the accumulation of minor alleles in GUCY1A3 modulates the therapeutic response to inhaled nitric oxide(45).

Interestingly, while the burden-based PRS predicts NC in Non-Hispanic Whites, it is a poor predictor of GA. This implies (and our AUC analyses confirm) that our composite predictor of NC—which combines information from both PRS *and* GA—should perform better than a predictor based on PRS alone. Overall, we believe that our composite predictor could facilitate the design of individualized treatments for preterm infants that, in turn, could substantially improve health outcomes and reduce hospitalization costs.

CONCLUSIONS

This work demonstrates clearly, our ability to predict NC among Non-Hispanic White preterm infants using information from both, genetic risk factors (e.g. burden) and traditional risk factors (e.g. gestational age). Provided that future studies of NC continue to collect genetic data on preterm infants from under-represented populations—where the chance of preterm birth is higher—then more efficacious composite biomarkers could be developed and implemented for members of these extremely vulnerable and historically under studied populations.

Acknowledgments

WES was performed by the Biomedical Genomics Core in The Research Institute at Nationwide Children's Hospital. We thank the participants and their families for their involvement in this study, the Ohio Perinatal Research Network (OPRN) for recruiting preterm infants, and Dr. Will C. Ray for his help editing the manuscript. We are also grateful to Andrew M Corris, Senior Licensing Associate, who provided professional advice concerning the patentability of various aspects of this research.

Statement of Financial Support: This work was supported by the Abigail Wexner Research Institute at Nationwide Children's Hospital and the Center for Clinical and Translational Science CTSA Grant UL1TR002733.

References:

1. Glass HC. et al. 2015 Outcomes for extremely premature infants. Anesth Analg 120:1337–1351. [PubMed: 25988638]
2. Liu L et al. 2016 Global, regional, and national causes of under-5 mortality in 2000-15: an updated systematic analysis with implications for the Sustainable Development Goals. Lancet 388:3027–3035. [PubMed: 27839855]
3. Mathews TJ, MacDorman MF 2011 Infant mortality statistics from the 2007 period linked birth/infant death data set. Natl Vital Stat Rep 59:1–30.
4. Bhandari V et al. 2006 Familial and genetic susceptibility to major neonatal morbidities in preterm twins. Pediatrics 117:1901–1906. [PubMed: 16740829]
5. Bizzarro MJ. et al. 2006 Genetic susceptibility to retinopathy of prematurity. Pediatrics 118:1858–1863. [PubMed: 17079555]
6. Draper ES, Manktelow B, Field DJ, James D 1999 Prediction of survival for preterm births by weight and gestational age: retrospective population based study. BMJ 319:1093–1097. [PubMed: 10531097]
7. Trembath A, Laughon MM 2012 Predictors of bronchopulmonary dysplasia. Clin Perinatol 39:585–601. [PubMed: 22954271]
8. Peelen MJ. et al. 2016 Impact of fetal gender on the risk of preterm birth, a national cohort study. Acta Obstet Gynecol Scand 95:1034–1041. [PubMed: 27216473]
9. Loftin R, Chen A, Evans A, DeFranco E 2012 Racial differences in gestational age-specific neonatal morbidity: further evidence for different gestational lengths. Am J Obstet Gynecol 206:259.e251–256. [PubMed: 22265090]
10. Ryan RM. et al. 2019 Black Race Is Associated with a Lower Risk of Bronchopulmonary Dysplasia. J Pediatr.
11. He P et al. 2017 Accumulation of minor alleles and risk prediction in schizophrenia. Sci Rep 7:11661. [PubMed: 28916820]
12. Still CD. et al. 2011 High allelic burden of four obesity SNPs is associated with poorer weight loss outcomes following gastric bypass surgery. Obesity (Silver Spring) 19:1676–1683. [PubMed: 21311511]
13. Zhu Z et al. 2015 Enrichment of Minor Alleles of Common SNPs and Improved Risk Prediction for Parkinson's Disease. PLoS One 10:e0133421. [PubMed: 26207627]

14. Jobe AH, Bancalari E 2001 Bronchopulmonary dysplasia. Am J Respir Crit Care Med 163:1723–1729. [PubMed: 11401896]
15. Bell MJ. et al. 1979 Epidemiologic and bacteriologic evaluation of neonatal necrotizing enterocolitis. J Pediatr Surg 14:1–4. [PubMed: 370356]
16. International Committee for the Classification of Retinopathy of P 2005 The International Classification of Retinopathy of Prematurity revisited. Arch Ophthalmol 123:991–999. [PubMed: 16009843]
17. Papile LA, Burstein J, Burstein R, Koffler H 1978 Incidence and evolution of subependymal and intraventricular hemorrhage: a study of infants with birth weights less than 1,500 gm. J Pediatr 92:529–534. [PubMed: 305471]
18. Kelly BJ. et al. 2015 Churchill: an ultra-fast, deterministic, highly scalable and balanced parallelization strategy for the discovery of human genetic variation in clinical and population-scale genomics. Genome Biol 16:6. [PubMed: 25600152]
19. Haeussler M et al. 2019 The UCSC Genome Browser database: 2019 update. Nucleic Acids Res 47:D853–D858. [PubMed: 30407534]
20. Rosenbloom KR. et al. 2013 ENCODE data in the UCSC Genome Browser: year 5 update. Nucleic Acids Res 41:D56–63. [PubMed: 23193274]
21. Lander ES. et al. 2001 Initial sequencing and analysis of the human genome. Nature 409:860–921. [PubMed: 11237011]
22. Genomes Project C. et al. 2015 A global reference for human genetic variation. Nature 526:68–74. [PubMed: 26432245]
23. Barton SJ. et al. 2013 Correction of unexpected distributions of P values from analysis of whole genome arrays by rectifying violation of statistical assumptions. BMC Genomics 14:161. [PubMed: 23496791]
24. Mackinnon JG, White H 1985 Some Heteroskedasticity-Consistent Covariance-Matrix Estimators with Improved Finite-Sample Properties. Journal of Econometrics 29:305–325.
25. Ludbrook J, Dudley H 1998 Why permutation tests are superior to t and F tests in biomedical research. American Statistician 52:127–132.
26. Bush WS, Moore JH 2012 Chapter 11: Genome-wide association studies. PLoS Comput Biol 8:e1002822. [PubMed: 23300413]
27. Dudbridge F 2013 Power and predictive accuracy of polygenic risk scores. PLoS Genet 9:e1003348. [PubMed: 23555274]
28. Sing T, Sander O, Beerenwinkel N, Lengauer T 2005 ROCR: visualizing classifier performance in R. Bioinformatics 21:3940–3941. [PubMed: 16096348]
29. Hajian-Tilaki K 2013 Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Caspian J Intern Med 4:627–635. [PubMed: 24009950]
30. Stone M 1974 Cross-Validatory Choice and Assessment of Statistical Predictions. Journal of the Royal Statistical Society Series B-Statistical Methodology 36:111–147.
31. Ward RM, Beachy JC 2003 Neonatal complications following preterm birth. BJOG 110 Suppl 20:8–16.
32. Morse SB. et al. 2006 Racial and gender differences in the viability of extremely low birth weight infants: a population-based study. Pediatrics 117:e106–112. [PubMed: 16396844]
33. Schieve LA, Handler A 1996 Preterm delivery and perinatal death among black and white infants in a Chicago-area perinatal registry. Obstet Gynecol 88:356–363. [PubMed: 8752239]
34. Irimia M et al. 2008 Widespread evolutionary conservation of alternatively spliced exons in Caenorhabditis. Mol Biol Evol 25:375–382. [PubMed: 18048400]
35. Hong EP, Park JW 2012 Sample size and statistical power calculation in genetic association studies. Genomics Inform 10:117–122. [PubMed: 23105939]
36. Torgerson DG. et al. 2018 Ancestry and Genetic Associations with Bronchopulmonary Dysplasia in Preterm Infants. Am J Physiol Lung Cell Mol Physiol.
37. Choi SW, Heng Mak TS, O’Reilly PF 2018 A guide to performing Polygenic Risk Score analyses. bioRxiv:416545.

38. Gooden M, Younger N, Trotman H 2014 What is the best predictor of mortality in a very low birth weight infant population with a high mortality rate in a medical setting with limited resources? Am J Perinatol 31:441–446. [PubMed: 23945903]

39. Abolfotouh MA, Al Saif S, Altwaijri WA, Al Rowaily MA 2018 Prospective study of early and late outcomes of extremely low birthweight in Central Saudi Arabia. BMC Pediatr 18:280. [PubMed: 30134865]

40. Zisk JL. et al. 2011 Do premature female infants really do better than their male counterparts? Am J Perinatol 28:241–246. [PubMed: 21046537]

41. Yokoyama N et al. 1995 A giant nucleopore protein that binds Ran/TC4. Nature 376:184–188. [PubMed: 7603572]

42. Hamada M et al. 2011 Ran-dependent docking of importin-beta to RanBP2/Nup358 filaments is essential for protein import and cell viability. J Cell Biol 194:597–612. [PubMed: 21859863]

43. Neilson DE. et al. 2009 Infection-triggered familial or recurrent cases of acute necrotizing encephalopathy caused by mutations in a component of the nuclear pore, RANBP2. Am J Hum Genet 84:44–51. [PubMed: 19118815]

44. Wilkins MR. et al. 2014 alpha1-A680T variant in GUCY1A3 as a candidate conferring protection from pulmonary hypertension among Kyrgyz highlanders. Circ Cardiovasc Genet 7:920–929. [PubMed: 25373139]

45. Hasan SU. et al. 2017 Effect of Inhaled Nitric Oxide on Survival Without Bronchopulmonary Dysplasia in Preterm Infants: A Randomized Clinical Trial. JAMA Pediatr 171:1081–1089. [PubMed: 28973344]

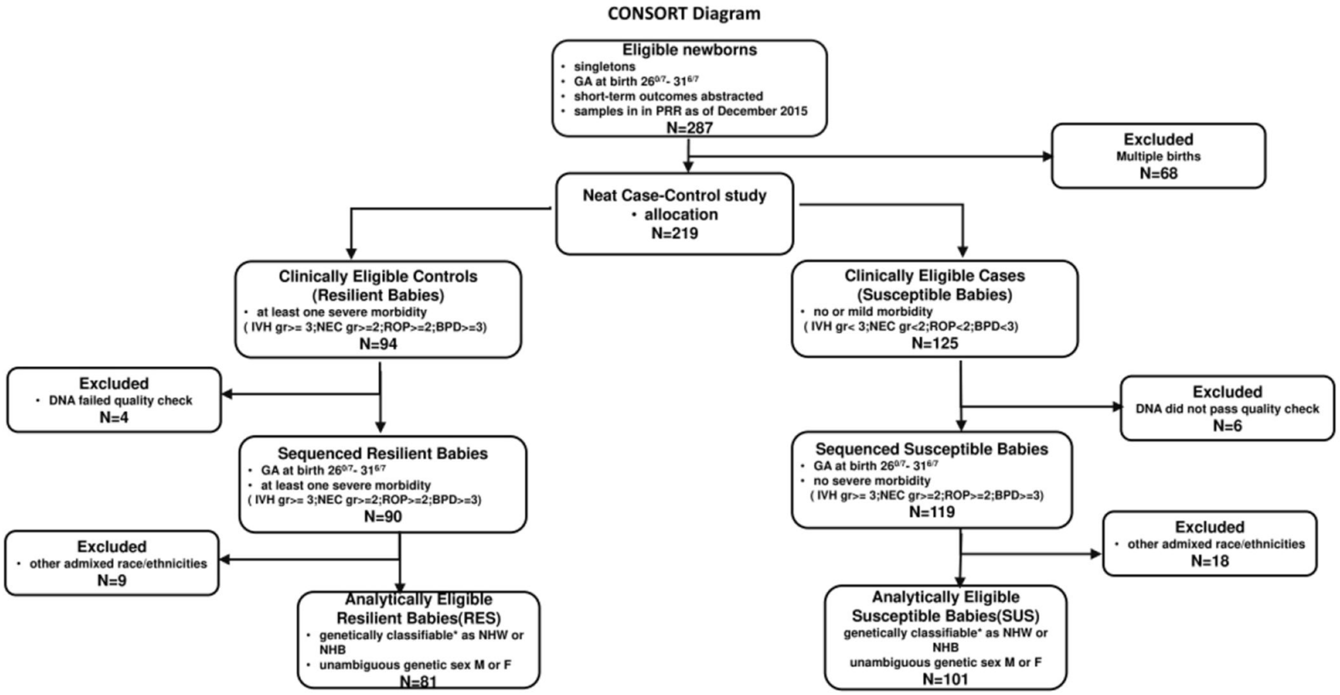
CONSORT Diagram
Eligible newborns
• singletons
• GA at birth $26^{0/7}$- $31^{6/7}$
• short-term outcomes abstracted
• samples in in PRR as of December 2015
N=287
Excluded
Multiple births
N=68
Neat Case-Control study
• allocation
N=219
Clinically Eligible Controls
(Resilient Babies)
• at least one severe morbidity
(IVH gr>= 3;NEC gr>=2;ROP>=2;BPD>=3)
N=94
Clinically Eligible Cases
(Susceptible Babies)
• no or mild morbidity
(IVH gr< 3;NEC gr<2;ROP<2;BPD<3)
N=125
Excluded
• DNA failed quality check
N=4
Excluded
DNA did not pass quality check
N=6
Sequenced Resilient Babies
• GA at birth $26^{0/7}$- $31^{6/7}$
• at least one severe morbidity
(IVH gr>= 3;NEC gr>=2;ROP>=2;BPD>=3)
N=90
Sequenced Susceptible Babies
• GA at birth $26^{0/7}$- $31^{6/7}$
• no severe morbidity
(IVH gr>= 3;NEC gr>=2;ROP>=2;BPD>=3)
N=119
Excluded
• other admixed race/ethnicities
N=9
Excluded
• other admixed race/ethnicities
N=18
Analytically Eligible
Resilient Babies(RES)
• genetically classifiable* as NHW or NHB
• unambiguous genetic sex M or F
N=81
Analytically Eligible
Susceptible Babies(SUS)
genetically classifiable* as NHW or NHB
unambiguous genetic sex M or F
N=101

Figure 1:

Flow diagram of our whole-exome association study (WEAS) for NC. From a total sample of 287 preterm infants, only 182 were retained for the final analysis. Exclusion criteria included: DNA control check, multiple births and unknown race of the preterm infants.

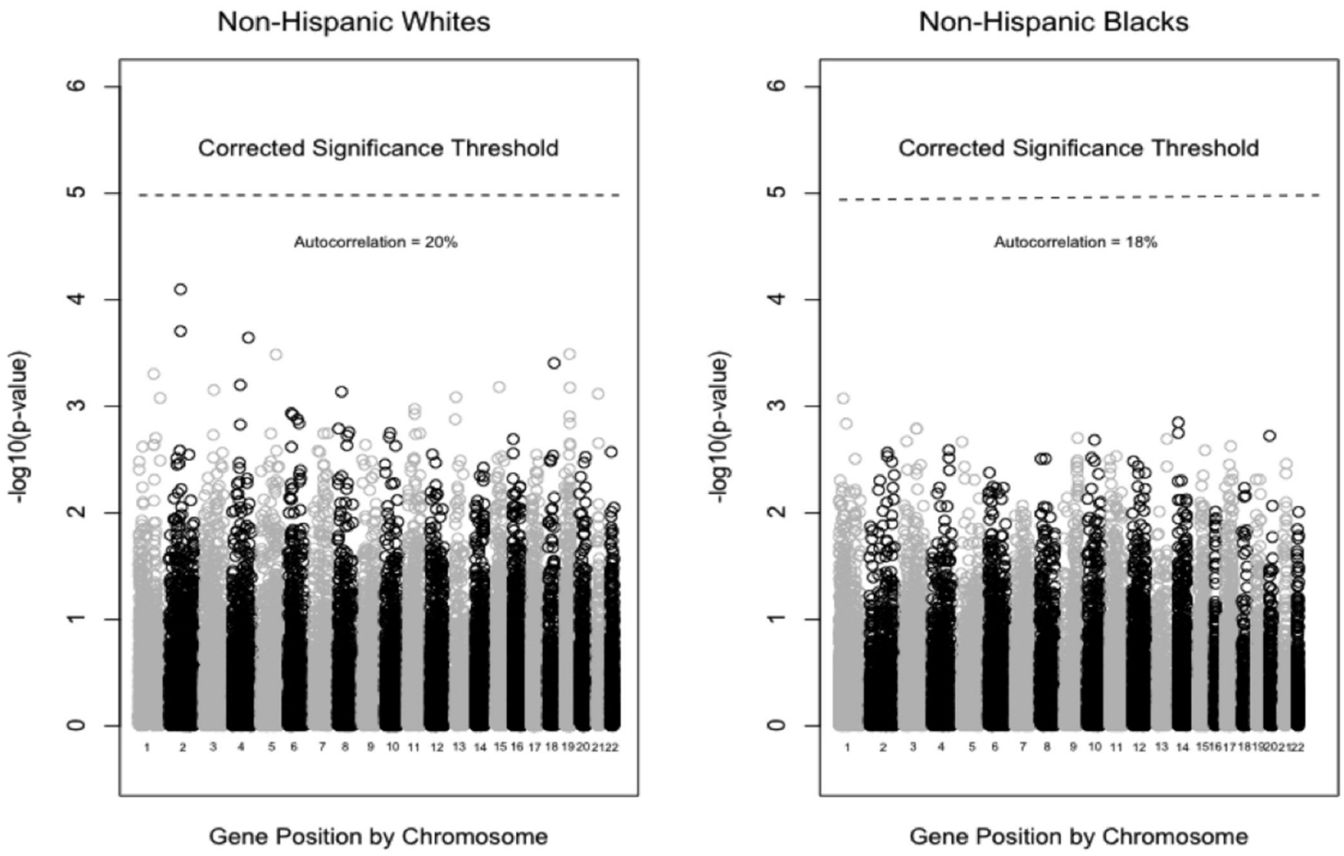

Figure 2.

Manhattan plot of −log10 p-values obtained from the logistic regression of preterm infant status onto burden. After correcting for multiple tests , no single gene in Non-Hispanic White(left) and Non-Hispanic(right) preterm infants is statistically significant at the exome-wide level (dashed line).

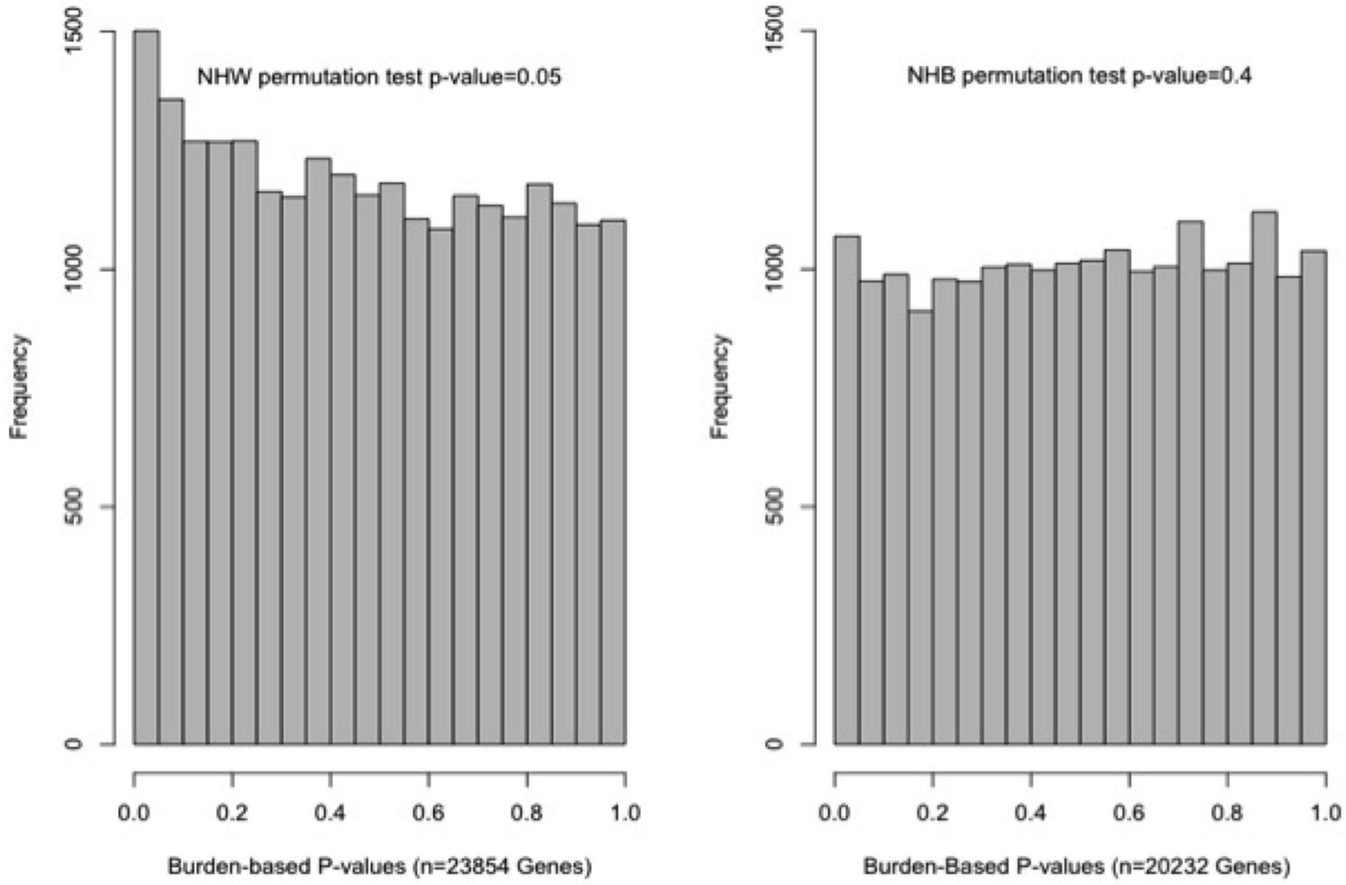

Figure 3.

The exome-wide distribution of burden-based p-values by gene in Non-Hispanic White (NHW) preterm infants (left panel), and Non-Hispanic Black (NHB) preterm infants (right panel). The distribution of p-values in NHWs (n=75 SUS, n=56 RES) shows a statistically significant excess of low with p-values (p=0.05), suggesting that genetic burden influences neonatal complications in NHWs. By contrast, the distribution of p-values in NHBs (n=25 SUS, n=26 RES) is inconclusive (p=0.4).

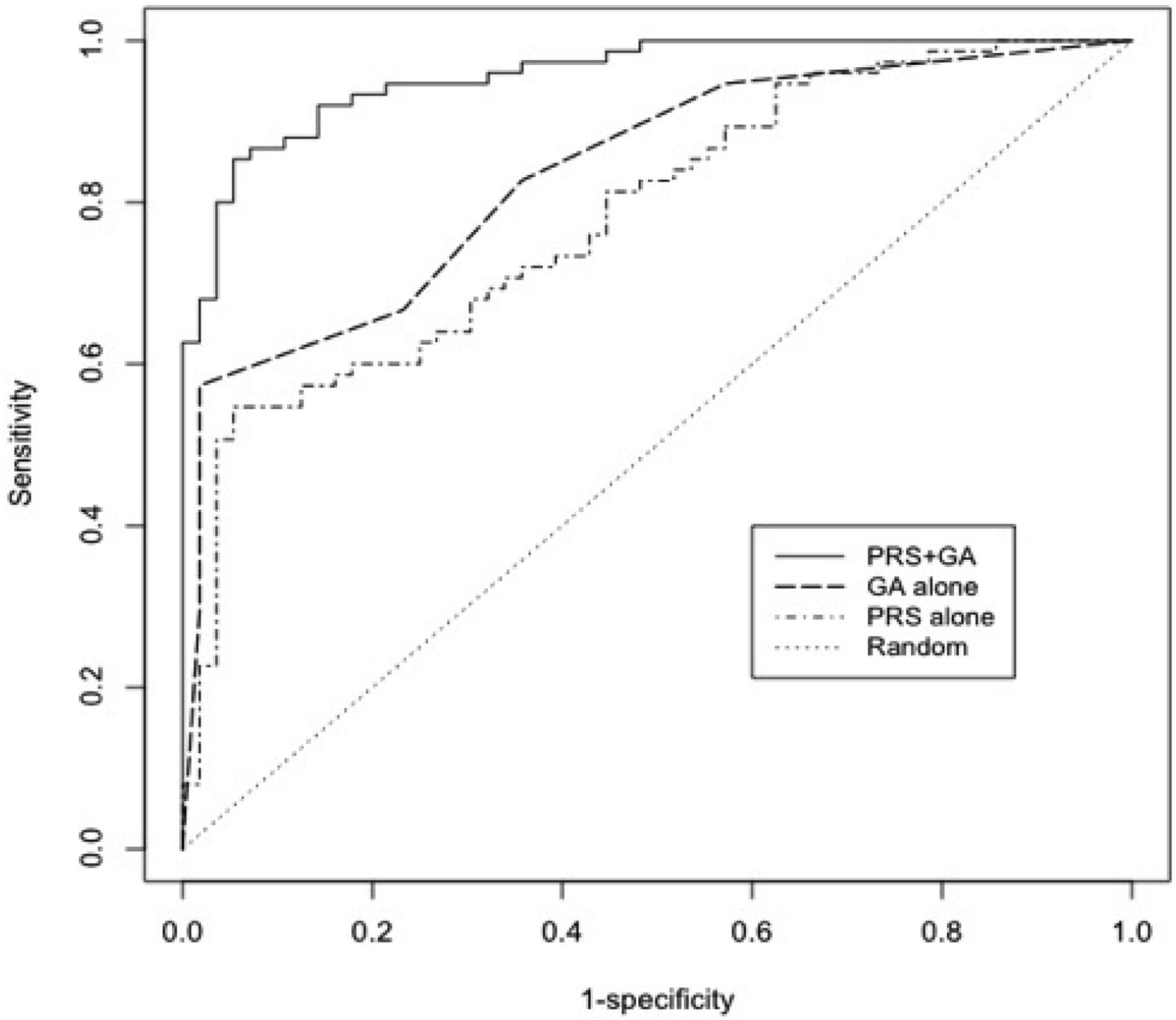

Figure 4:

ROC curves are shown for three predictors of NC: Polygenic Risk Score(PRS) +Gestational Age (GA) (solid), GA alone(dashed), PRS alone(dotdashed), and Random (dotted), with their corresponding AUC's 96%, 84%, 78%, and 50%, respectively. After correcting for over-fitting, the average AUC of PRS+GA dropped to 87%, and the average AUC of PRS alone dropped to 67%. The difference in predictive power between PRS+GA and PRS alone is significant (p=0.0012).

Table 1:

Clinical outcomes of our study group

Outcomes*	Susceptible (SUS)		Resilient (RES)	
	Infants completing evaluation (N)	Prevalence n (%)	Infants completing evaluation (N)	Prevalence n (%)
Comorbidity*				
Bronchopulmonary dysplasia(BPD)**	101	90(89.1%)	81	0(0%)
Retinopathy of prematurity(ROP)**	101	23(22.8%)	81	0(0%)
Intraventricular hemorrhage(IVH)	101	8(8%)	81	0(0%)
Necrotizing enterocolitis(NEC)	101	5(5%)	81	0(0%)

*Individual morbidities not mutually exclusive. 17 non-Hispanic White and 6 non-Hispanic Black children had multiple morbidities.

**Two susceptible infants died before BDP and ROP could be assessed; BPD is the most frequently encountered at the gestational age bracket (26-31) and Nationwide Children's Hospital is a BPD referral site, which could explain the high prevalence of BPD among SUS infants.

Table 2:

Clinical characteristics of our study group

Clinical Characteristics	**Resilient (RES) Infants N (%)**	**Susceptible (SUS) Infants N (%)**	***P*-value**
Race			
Non-Hispanic White	56 (69.1%)	75 (74.3%)	0.45
Non-Hispanic Black	25 (30.9%)	26 (25.7%)	
Sex			
Male	44 (54.3%)	70 (69.3%)	**0.04**
Female	37 (45.7%)	31 (30.7%)	
Surfactant			
Yes	30 (38.0)	38 (38.4)	0.96
No	49 (62)	61 (61.6)	
Maternal Characteristic Initiating Birth			
Spontaneous labor	25 (32.9%)	36 (37.1)	0.45
PROM	18 (23.7%)	29 (29.8)	
Induction/ Cesarean without labor	33 (43.4%)	32 (33.0)	
Antenatal Steroids			
Yes	85(57.8%)	14 (45.2%)	0.20
No	62 (42.2%)	17 (54.8%)	
Delivery Route			
Vaginal	33 (41.8%)	35 (0.35%)	0.38
Cesarean	46 (58.2%)	64 (0.65%)	
Pre-pregnancy diabetes			
No	72 (92.3%)	88 (88.9%)	0.44
Yes	6 (7.7%)	11 (11.1%)	
Hypertensive disorders of pregnancy [*]			
No	58 (75%)	73 (74.5%)	0.90
Yes	19 (25%)	25 (25.5%)	
Type of conception			
ART or Ovulation stimulation	4(5.4%)	6 (6.1%)	0.78
Unassisted	74(95.6%)	92 (93.9%)	
Gestational Age			
Gestational age at birth (days), median (IQR)	213 (202, 220)	194 (188, 204)	**<0.001**
Birth weight			
Birthweight (grams), median (IQR)	1360 (1084, 1568)	998 (864, 1168)	***<0.001***
Apgar Score			
APGAR score at 1 min. median (IQR)	6 (3.5, 8)	4 (2, 6)	**0.009**
APGAR score at 5 min. median (IQR)	8 (7, 8)	7 (6, 8)	**0.03**

[*] Gestational hypertension, preeclampsia, or HELLP (hemolysis, elevated liver enzymes, low platelet count).

Table 3:

Top 10 genes used to construct PRS

Rank	Gene ID	*P*-values	Chromosome
1	RANBP2	7.96E-05	2
2	CCDC138	1.96E-04	2
3	GUCY1A3	2.26E-04	4
4	RNU6-66P	3.23E-04	19
5	SPINK1	3.23E-04	5
6	ZCCHC2	3.92E-04	18
7	DQ579288	4.96E-04	1
8	FAM47E-STBD1	6.30E-04	4
9	RSL24D1	6.58E-04	15
10	FTL	6.65E-04	19

The top 10 genes from our exome-wide burden-based association study in Non-Hispanic Whites preterm infants; *p*-values are also shown.

Table 4:

AUC Averaged over 10 Cross-Validation Sets

	Predictive Power	**95% CI**	***p*-value**
PRS alone	0.67[*]	(0.56,0.77)	*p*<0.003
GA alone	0.84	(0.83,0.90)	*p*<0.001
PRS+GA	0.87[*]	(0.82,0.92)	*p*<0.001

[*] Summary of the ROC curve analysis for each risk factor: PRS alone (polygenic risk score, which is based on burden), GA (gestational age) alone, and the combination of PRS and GA (denoted PRS+GA). We measured the predictive power (PP) of each risk factor, where PP is define as the standard AUC (area under the ROC curve) for GA alone, but is defined as the average AUC across 10 cross-validation sets (to mitigate the negative effects of over-fitting) for PRS alone and PRS+GA. We used a non-parametric bootstrap procedure to compute 95% confidence intervals (CIs) and *p*-values for testing whether predictive power was better than chance.