Word Count:Abstract Word Count: 299
Manuscript Word Count: 3,194

Title: Electronic health record biobank cohort recapitulates an association between the *MUC5B* promoter polymorphism and ARDS in critically ill adults.

Author List: V. Eric Kerchberger, MD, MS[1,2]; J. Brennan McNeil, BS[1,3]; Neil Zheng, MD[2,4]; Diana Chang, PhD[5]; Carrie Rosenberger, PhD[5]; Angela J. Rogers, MD[6]; Julie A. Bastarache, MD[1,6,7]; QiPing Feng, PhD[1]; Wei-Qi Wei MD, PhD[2]; Lorraine B. Ware, MD[1,7]

Author Affiliations:
[1] Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, USA
[2] Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, USA
[3] Quillen College of Medicine, East Tennessee State University, Johnson City, Tennessee, USA
[4] Brigham and Women's Hospital, Boston, Massachusetts, USA
[5] Genentech Inc., South San Francisco, California, USA
[6] Department of Medicine, Stanford University, Palo Alto, California, USA
[7] Department of Cell and Developmental Biology, Vanderbilt University, Nashville, TN
[8] Department of Pathology, Microbiology and Immunology, Vanderbilt University Medical Center, Nashville, Tennessee

Institution: All work was performed at Vanderbilt University Medical Center, Nashville, Tennessee.

Corresponding Author: V. Eric Kerchberger, MD MS. Division of Allergy, Pulmonary, and Critical Care Medicine, Vanderbilt University Medical Center. 1161 21st Ave South, T-1218, Medical Center North, Nashville, TN 37232. Email: vern.e.kerchberger@vumc.org

Conflicts of Interest: Dr. Ware reports consulting fees from Novartis, Arrowhead, Akebia, and Global Blood Therapeutics, stock ownership in Virtuoso Surgical, and institutional research contracts with Genentech, Inc. and Bluejay Diagnostics The remaining authors report no conflicts of interest with the reported study.

Financial Support: Research reported in this publication was supported by the National Institutes of Health under award numbers: NIH K01HL157755 (VEK), NIH R01HL158906 and R01HL164937 (LBW), NIH R01 HL126671 (JAB), NIH R35 HL150783 (JAB), NIH R01 HL171809 and R56 AG084550 (QF and WQW) and research contract funds from Genentech, Inc. The project publication described was also supported by CTSA award No. UL1TR002243 from the National Center for Advancing Translational Sciences. Its contents are solely the responsibility of the authors and do not necessarily represent official views of the National Center for Advancing Translational Sciences or the National Institutes of Health. This research was also supported by Courtney's Race for the Acute Respiratory Distress Syndrome Cure and the Courtney Charneco Family.

Prior Presentation: Portions of this study were previously presented at the American Thoracic Society 2022 International Conference (May 17, 2022, San Francisco, California) and the Sixth International ARDS Conference (June 19, 2023, Dublin, Ireland).

Author Contributions: VEK had full access to all of the data and takes responsibility for the integrity of the analyses. VEK and LBW designed the study. VEK, JBN, NZ, DC, CR, AJR, QF, WQW, and LBW contributed to data acquisition. VEK performed the data analyses. VEK, WQW, JAB, and LBW contributed to data interpretation. VEK drafted the initital manuscript. All authors participated in critical revision of the manuscript and provided intellectual content; and all authors approved the final version of the manuscript.

MeSH Terms: Respiratory Distress Syndrome, Acute/genetics; Genetic variation; *MUC5B* protein, human; Mucin-5B; Electronic Health Records

ABSTRACT

Background: Large population-based DNA biobanks linked to electronic health records (EHRs) may provide novel opportunities to identify genetic drivers of ARDS.

Research Question: Can we develop an EHR-based algorithm to identify ARDS in a biobank database, and can this validate a previously reported ARDS genetic risk factor?

Study Design and Methods: We analyzed two parallel genotyped cohorts: a prospective biomarker cohort of critically ill adults (VALID), and a retrospective cohort of hospitalized participants enrolled in a de-identified EHR biobank (BioVU). ARDS was identified by clinician-investigator review in VALID and an EHR algorithm in BioVU (EHR-ARDS). We tested the association between the *MUC5B* promoter polymorphism rs35705950 with development of ARDS, and assessed if age modified this genetic association in each cohort.

Results: In VALID, 2,795 patients were included, age was 55 [43, 66] (median [IQR]) years, and 718 (25.7%) developed ARDS. In BioVU, 9,025 hospitalized participants were included, age was 60 [48, 70] years, and 1,056 (11.7%) developed EHR-ARDS. We observed a significant age-related interaction effect on ARDS in VALID: among older patients, rs35705950 was associated with increased ARDS risk (OR: 1.44; 95%CI 1.08-1.92; p=0.012) whereas among younger patients this effect was absent (OR: 0.84; 95%CI: 0.62-1.14; p=0.26). In BioVU, rs35705950 was associated with increased risk for EHR-ARDS among all participants (OR: 1.20; 95%CI: 1.00-1.43, p=0.043) and this did not vary by age. The polymorphism was also associated worse oxygenation in mechanically ventilated BioVU participants, but had no association with oxygenation in VALID.

Interpretation: The *MUC5B* promoter polymorphism was associated with ARDS in two cohorts of at-risk adults. Although age-related effect modification was observed only in VALID, BioVU identified a consistent association between *MUC5B* and ARDS risk regardless of age, and a novel association with oxygenation impairment. Our study highlights the potential for EHR biobanks to enable precision-medicine ARDS studies.

KEY WORDS

Acute Respiratory Distress Syndrome; Electronic Health Records; Genetic risk factors; *MUC5B* gene; interaction analysis

ABBREVIATIONS

ARDS (Acute Respiratory Distress Syndrome)

EHR (Electronic Health Record)

MUC5B (Mucin 5B)

NPV (negative predictive value)

PPV (positive predictive value)

ILD (Interstitial lung disease)

VALID (Validating Acute Lung Injury biomarkers for Diagnosis)

INTRODUCTION

The acute respiratory distress syndrome (ARDS) is an acute inflammatory lung disorder that develops in the setting of a local or systemic insult such as pneumonia, sepsis, or massive trauma.[1] ARDS occurs in over 20% of mechanically ventilated critically ill adults, and leads to prolonged hospitalization, high costs, and high mortality risk. Understanding genetic drivers of ARDS among at-risk individuals could improve our ability to define disease subphenotypes, predict disease risk, and identify new therapeutic targets.[2]

As the risk for ARDS is influenced by many genes, each exerting only a modest contribution to an individual's overall risk, large numbers of cases and at-risk control patients are necessary to design appropriately powered studies.[2–6] However, prospectively identifying and genotyping thousands of critically ill patients and phenotyping them specifically for ARDS is prohibitively costly and time consuming. Therefore, innovative sampling approaches are needed to design adequately powered genetic studies in the ICU. Large population-based DNA biobanks pairing clinical data from the electronic health record (EHR) with genomic information have improved our understanding of genetic contributors to many chronic diseases.[7] However, leveraging EHR biobanks to study ARDS requires novel methods because of its syndromic definition.[8–11]

We previously reported that a gain-of-function polymorphism (rs35705950) in the gel-forming mucin gene *MUC5B* promoter was associated with increased risk for ARDS among critically ill adults aged 50 years or older.[12] The original study included 1,534 adults (including 903 aged ≥50 years) enrolled from a prospective observational ICU cohort and genotyped using a dedicated polymerase-chain reaction (PCR) assay. In this study, we have expanded genotyping of this prospective ICU cohort to nearly 3,000 patients on a single genome-wide array platform (which

includes patients in the original study), increasing our power to detect genetic associations with ARDS. Then to validate the association between rs35705950 and ARDS, we developed a parallel cohort from our institution's de-identified EHR biobank by employing a rule-based ARDS classifier based on a combination of diagnostic billing codes, laboratory test results, and text from unstructured radiography reports. We analyzed both cohorts in a derivation-validation approach to test the association between ARDS risk and rs35705950, and hypothesized that the *MUC5B* polymorphism would exert differential risk for developing ARDS based on age.

MATERIALS AND METHODS

Patient Cohorts

A study schematic is provided in **Figure 1**. The derivation cohort came from the Validating Acute Lung Injury biomarkers for Diagnosis (VALID) study, a prospective observational cohort of critically ill adults at risk for ARDS, admitted to either the surgical, medical, trauma, or cardiovascular intensive care units (ICUs) at Vanderbilt University Medical Center (VUMC) from January 2006 to December 2020. All patients were enrolled on the morning of the second ICU day. Other inclusion criteria, enrollment and consent procedures for VALID have been previously described.[9,13] Two expert physician investigators manually reviewed clinical records and chest radiographs to adjudicate ARDS status according to the Berlin Definition,[14] and we defined ARDS cases as those with expert-adjudicated ARDS on any of the first four ICU days.

The validation cohort was identified from BioVU, VUMC's DNA-biobanking program linked to a de-identified EHR (**Figure 1**).[7,15] Additional information on the BioVU enrollment process and the de-identified EHR are provided in the **Supplementary Methods**. We identified BioVU participants age ≥18 years admitted to the hospital from 2002 to 2019 with at least one

diagnostic billing code for an ARDS risk factor during the first 7 days of hospital admission (**e-Table 1**). We included both ICU and non-ICU participants as many patients with ARDS risk factors who ultimately do not develop the syndrome are managed outside of the ICU setting. We then developed a computable ARDS algorithm (EHR-ARDS) to capture the four criteria of the Berlin ARDS Definition using structured EHR data.[14] EHR-ARDS cases were identified by presence of (i) a diagnostic billing code for acute hypoxemic respiratory failure (**e-Table 2**), (ii) procedural codes for mechanical ventilation (**e-Table 3**), (iii) PaO2:FiO2 ratio ≤300, (iv) presence of a chest x-ray report containing terms consistent with bilateral opacities as identified using a set of regular expressions that we designed *a priori* for the study (**e-Table 4**), and (v) more days with diagnosis codes for ARDS risk factors than days with codes for congestive heart failure (**e-Table 5**). At-risk controls included all participants who were admitted with an ARDS risk factor diagnosis code but did not meet all EHR-ARDS criteria during the first 7 days of hospitalization. For participants with more than one qualifying admission, we used the first hospitalization only. Patients/participants with missing genotypes or diagnosis codes for interstitial lung diseases (ILDs), lung transplant, or bone marrow transplant were excluded from both cohorts (**e-Table 6**). The Vanderbilt Institutional Review Board (Nashville, TN) reviewed and approved the study protocols for both the VALID (IRB #051065) and BioVU cohorts (IRB # 202530).

EHR-ARDS Classification Performance

We assessed the EHR-ARDS classifier's performance in both cohorts using standard metrics including sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F_1 score, and Cohen's κ-statistic.[16] In VALID, we assessed performance of EHR-ARDS against investigator-adjudicated Berlin ARDS among all genotyped patients included in the

study. In the BioVU cohort, EHR-ARDS classifier performance was compared to ARDS status for 125 randomly selected participants (50 EHR-ARDS, 75 at-risk control participants) by manual adjudication of Berlin ARDS status using clinical notes and radiography reports available in the de-identified EHR's user interface.[15]

MUC5B Promoter Genotyping

VALID patients were genotyped using the Illumina Global Screening Array and BioVU participants were genotyped using the Illumina Expanded Multi-Ethnic Global Array. We ascertained *MUC5B* promoter (rs35705950) genotype in both cohorts by imputation using the Michigan Imputation Server with the Haplotype Reference Consortium version 1.1 reference panel.[17,18] Additional genotyping information is provided in the **Supplementary Methods**.

Statistical Analysis

Continuous variables are presented as median and interquartile ranges, categorical variables as frequencies and proportions, and differences between groups as mean differences and 95% confidence intervals (CI). We assessed group-wise differences using the Mann-Whitney *U* test for continuous outcomes and Pearson's χ^2-test for categorical outcomes. We tested the association between the *MUC5B* promoter polymorphism and outcomes using logistic regression for ARDS in VALID or EHR-ARDS in BioVU, linear regression for lowest PaO_2:FiO_2 ratio, and ordinal regression for Berlin ARDS severity categories.[14] In-hospital survival was assessed using the method of Kaplan and Meier. As we used imputed genotypes for rs35705950, genotype was treated as a continuous variable ranging from 0.0 to 2.0 to represent the predicted number of alterative (T) alleles and to account for uncertainty in the imputed gene dosage. We selected model covariates using a causal directed acyclic graph (**e-Figure 1**),[19] which included age, sex,

race, ARDS risk factors (aspiration, pancreatitis, pneumonia, sepsis, shock, or trauma), and cardiovascular comorbidities (heart failure, acute myocardial infarction, or other coronary artery disease). Based on the age-dependent effect of the promoter polymorphism observed among VALID patients in the previous study,[12] we assessed the associations between genotype and ARDS/EHR-ARDS both without any interaction terms and with including a genotype × age interaction term in the regression model. The combined significance of the *MUC5B* genotype and the genotype × age interaction term was also assessed using an F-test on both parameters. A p-value of less than 0.05 was considered statistically significant as all analyses were considered exploratory. All regression models employed heteroscedasticity-consistent standard errors using the "HC3" covariance matrix estimators.[20,21] We performed all analyses using R version 4.3.2 (Vienna, Austria) with the R packages *interactions* and *sandwich.*[21,22]

RESULTS

Study Populations

Demographic and clinical characteristics of both cohorts are shown in **Table 1** and flow diagrams illustrating patients meeting inclusion and exclusion criteria are shown in **e-Figure 2**. Compared to VALID, the BioVU cohort was older (median [IQR] age: 60 [48, 70] years vs 55 [43, 66] years, p<0.001), had a more even sex distribution (male sex: 49.2% vs 60.1%, p<0.001), had lower rates of most ARDS risk factor diagnoses, and had lower rates of mechanical ventilation (20.7% vs 76.1%, p<0.001). Hospital length of stay and inpatient mortality were also lower in the BioVU cohort, reflecting inclusion of participants hospitalized in both ICU and non-ICU settings. *MUC5B* promoter polymorphism T-allele frequencies were similar between cohorts (VALID: 10.5% BioVU: 9.6%, p-value: 0.15, **e-Figure 3**). We observed

high agreement in VALID between promoter genotype determined by imputation and by polymerase chain reaction (**e-Table 7**, Cohen's κ: 0.98).

EHR-ARDS Classifier Performance

Performance of the EHR-ARDS classifier was acceptable. In VALID, the classifier had a sensitivity 0.86, specificity 0.70, PPV 0.49, NPV 0.93, and Cohen's κ 0.45, indicating moderate agreement with investigator-adjudicated ARDS (**Table 2**). We observed similar performance compared with investigator-adjudicated ARDS among 125 randomly selected participants in the BioVU cohort, with an observed sensitivity 0.94, specificity 0.81, PPV 0.66, NPV 0.97, and Cohen's κ 0.67 (**Table 2**). Among BioVU participants with EHR-ARDS, we observed a clear association with mortality based on Berlin severity criteria: mild EHR-ARDS participants had a 30-day in-hospital survival of 87% (95% CI: 0.82 to 0.93), while survival was 73% (68% to 80%) moderate EHR-ARDS and 66% (60% to 72%) for severe ARDS (**e-Figure 4**).

Elevated ARDS Risk among Older VALID Patients with *MUC5B* Promoter Polymorphism

Similar to our previous report from VALID,[12] we found no significant association between rs35705950 and ARDS in the overall cohort when not accounting for any age-related effect modification (OR per T allele: 1.11; 95% CI: 0.90-1.36; p-value: 0.35). In contrast, under the interaction analysis we observed a significant age × genotype interaction effect such that the effect of rs35705950 on ARDS increased among older patients (OR for interaction term [per 10 years increased age]: 1.17, 95% CI: 1.04-1.33, p-value: 0.015, joint rs35705950 and interaction term p-value: 0.026, **Figure 2**). For a patient aged 70 years (mean + 1 standard deviation for the cohort), each T-allele was associated with an odds ratio of 1.44 for ARDS (95% CI 1.08-1.92; p-value: 0.012), whereas for a patient age 37 years (mean - 1 standard deviation), the effect was

substantially attenuated (OR per T-allele: 0.84; 95% CI: 0.62-1.14; p-value: 0.26). Subgroup analyses without the interaction term revealed more pronounced associations among older patients (age ≥ 50) compared to younger patients, females compared to males, and among patients who died during hospitalization compared to survivors, although the 95% confidence intervals did not exclude 1.0 for any analyzed subgroup (**e-Figure 5**).

Elevated Risk of EHR-ARDS in BioVU Participants with *MUC5B* Promoter Polymorphism

In the BioVU cohort, we observed a significant association between rs35705950 and EHR-ARDS risk even without accounting for any interaction effects (OR per T-allele: 1.20; 95% CI: 1.00-1.43; p-value: 0.043). In contrast to VALID, the effect of the polymorphism on EHR-ARDS was consistent across all ages (OR for interaction term [per 10 years increased age]: 1.02, 95% CI: 0.91-1.13; p-value: 0.78; joint rs35705950 and interaction term p-value: 0.12, **Figure 2**). Subgroup analyses revealed that the association between rs35705950 and EHR-ARDS was generally consistent across all ARDS risk factor subgroups, although the association only reached our pre-specified level of statistical significance among the subgroup of participants hospitalized with shock (**Figure 3**). Similar to VALID, the association was also stronger among older patients (age ≥ 50) compared to younger patients, females compared to males, and among patients who died during the hospitalization compared to survivors (**Figure 3**).

Impaired Oxygenation among Mechanically Ventilated BioVU Patients with *MUC5B* Promoter Polymorphism

Among BioVU participants who received invasive mechanical ventilation (regardless of EHR-ARDS status), rs35705950 variant carriers had more severe oxygenation impairment, with a mean -8 mmHg (95% CI: -15 to -1 mmHg) lower PaO_2/FiO_2 ratio per T-allele (p-value: 0.034, **Figure 4**). We observed a similar direction of effect among BioVU participants with EHR-ARDS

(mean -6 mmHg per T-allele, 95% CI: -14 to +2, p-value: 0.14, **Figure 4**) which corresponded to increased odds for being in a more severe ARDS category by the Berlin Definition (OR: 1.23, 95% CI: 0.94 to 1.63, p-value: 0.14), although these associations did not meet our pre-specified level of statistical significance. We did not observe any such association between rs35705950 and oxygenation impairment in the VALID cohort (**e-Figure 6**).

DISCUSSION

Summary of Findings

In this proof-of-concept study, we tested the utility of our de-identified longitudinal EHR to recapitulate a previously-observed genetic association with the *MUC5B* promoter polymorphism rs35705950 and ARDS risk in hospitalized adults. While our prior analysis in VALID reported an association between rs35705950 and ARDS among patients aged ≥50 years, our current study extends these findings in the VALID cohort by demonstrating a significant age × genotype interaction effect in nearly twice as many patients as the previous study cohort.[12] These findings were additionally robust to adjustment for the presence of several ARDS risk factors and potentially confounding cardiac diagnoses. Complementary to our findings in VALID, in our de-identified BioVU cohort we found an association between rs35705950 and EHR-ARDS risk, although this association was generally consistent across all age ranges. We also observed a novel genotype-specific association with oxygenation impairment in BioVU. Each T-allele was associated with a mean 8 mmHg lower P_aO_2:F_iO_2 ratio among BioVU participants receiving invasive mechanical ventilation, and there was a directionally similar effect among those meeting our EHR-ARDS definition. Finally, our EHR-ARDS algorithm had similar levels of agreement with investigator-adjudicated ARDS as clinician documentation[9] or inter-observer agreement between multiple clinicians,[23] supporting the clinical validity of this EHR phenotype. The

performance of our algorithm differed between cohorts, with a higher specificity and PPV in BioVU compared to VALID; this may reflect differences between cohorts in baseline demographics, severity of illness, and rates of ARDS risk factor diagnoses.

Relationship to prior literature

The gel-forming mucin *MUC5B* is highly expressed in bronchial epithelium and has important roles in mucociliary clearance and control of bacterial infection and inflammation.[24,25] Originally identified as a risk factor for interstitial lung disease,[26,27] the promoter polymorphism rs35705950 lies 3 kilobases upstream of the transcription start site and is associated with transcriptional over-expression of *MUC5B* mRNA in distal airways and respiratory bronchioles.[26,28] Potential mechanisms for the association between rs35705950 and ARDS may include increased risk for ARDS among patients with preclinical or undiagnosed interstitial lung disease, or *MUC5B* over-expression leading to mucociliary dysfunction and increased lung inflammation during acute illness.[29,30] This study is the second to report an association between rs35705950 and ARDS risk in a general population of critically ill adults, although there is some overlap as both studies included patients from the VALID cohort. Our findings contrast those of a multinational genome-wide meta-analysis which reported a lower risk for hospitalization for rs35705950 T-allele carriers in patients with COVID-19,[31] although potential explanations may include virus-specific effects of airway mucins, selection bias due to differences in self-isolation behaviors during the COVID-19 pandemic, or survivor bias among variant carriers who never developed interstitial lung disease.[32]

Prior efforts to identify ARDS from the EHR often focused on the role(s) of specific EHR data elements such as clinician documentation,[9,33] chest radiograph report text,[8,34–37] or radiographic

images,[10] with a focus on clinical applications such as implementing evidenced-based ARDS care or clinical trial recruitment. In contrast, our study integrated multiple data sources available in a widely-used EHR common data model including diagnosis codes, laboratory studies, and free-text chest radiograph reports.[15,38] Sathe et. al. evaluated the performance of a conceptually similar EHR-ARDS classifier among adults hospitalized with COVID-19, and reported high agreement between their EHR-ARDS classifier and clinician-adjudicated ARDS (κ=0.85) in 175 manually reviewed patients.[11] Li et. al. evaluated an EHR-ARDS classifier using laboratory values and chest radiograph report text across seven medical centers in a large multi-state health system, also reporting very high performance of their algorithm (sensitivity and specificity both above 90%) against manual adjudication for 150 patients.[39] Out study had over 10-fold more patients with manually-adjudicated ARDS status than either previous study, therefore our results may provide more precise estimates of the real-world performance for a rules-based EHR-ARDS classifier.

Strengths

Our VALID cohort is well-established in the critical care literature and provided nearly 2800 patients with extensive investigator-adjudicated ARDS phenotyping data.[9,12,13,40] Similarly, BioVU is one of the largest institutional EHR biobanks in the United States and has yielded important insights into genetic drivers underlying a range of medical conditions.[7,41,42] Both cohorts captured a broad swath of ARDS including patients with medical, surgical, and trauma admissions as well as those with both direct and indirect causes of lung injury. Additionally, rather than *a priori* excluding patients with heart failure, our EHR-ARDS classifier algorithm specifically allowed for patients with concomitant cardiogenic and noncardiogenic causes of pulmonary edema since cardiac comorbidities and volume overload are common in adults with

ARDS.[8,43,44] Whereas many prior ARDS genetics studies were limited only to ICU patients,[4–6,12,45,46] we specifically included non-ICU patients in BioVU to mitigate selection bias as many patients with ARDS risk factors who ultimately do not develop the syndrome are managed outside of the ICU. The stronger association between the *MUC5B* promoter polymorphism and EHR-ARDS among the subgroup of mechanically ventilated BioVU participants suggests that our overall findings were not driven by non-critically ill participants. Although we used imputed *MUC5B* promoter genotypes, we observed very high correlation between genotypes determined by imputation and genotypes determined by polymerase chain reaction which were available from the prior study in VALID,[12] so the likelihood of genotype misclassification in the BioVU cohort is low.

Limitations

This study has some limitations. Although our use of two cohorts supports the internal validity of our findings, both the VALID and BioVU cohorts came from a single center. Therefore, the generalizability of our findings to patient populations outside our center remains to be assessed. Overlap between the cohorts is possible as institutional protocols to prevent re-identification of BioVU participants precluded us from specifically excluding VALID patients from the BioVU cohort. Over this study's time frame our center admitted over 105,000 adults to these ICUs, most of whom (>85%) have never submitted DNA samples to BioVU. So although there may be some patients enrolled in both cohorts, the amount of overlap is likely low. We excluded patients with pre-existing ILD based on presence of relevant diagnosis codes prior to hospital discharge. While manual review of chest imaging might be more sensitive, particularly for preclinical ILD,[27] the large numbers and de-identified nature of BioVU made manual review impractical. We used a directed acyclic graph to inform model design and identify potential confounding variables,

however as with all observational cohort studies these analyses may be subject to residual confounding. Finally, our EHR-ARDS classifier used a simple rules-based algorithm which demonstrated acceptable sensitivity but only moderate PPV compared to investigator-adjudicated ARDS. Additional refinements to the EHR-ARDS classifier incorporating additional data elements, more intensive natural language processing of chest radiograph reports, or machine learning algorithms may improve performance in future studies.[47]

CONCLUSION

The *MUC5B* promoter polymorphism was associated with an increased risk for ARDS among two parallel cohorts of at-risk hospitalized adults enrolled at a single academic medical center. Although age-related effect modification was only observed in the prospective VALID cohort, the de-identified BioVU biobank cohort identified a more consistent association between this polymorphism and EHR-ARDS risk across all ages, and a novel association with more impaired oxygenation among mechanically ventilated participants. Our study illustrates the complementary roles of these differing cohort enrollment strategies and highlights the potential for population-based EHR biobanks to enable future investigation into the genetic determinants of critical illness at large scales.

REFERENCES

1. Matthay MA, Arabi Y, Arroliga AC, et al. A New Global Definition of Acute Respiratory Distress Syndrome. *Am J Respir Crit Care Med* 2024;209(1):37–47.

2. Reilly JP, Christie JD, Meyer NJ. Fifty Years of Research in ARDS. Genomic Contributions and Opportunities. *Am J Respir Crit Care Med* 2017;196(9):1113–1121.

3. Christie JD, Wurfel MM, Feng R, et al. Genome Wide Association Identifies PPFIA1 as a Candidate Gene for Acute Lung Injury Risk Following Major Trauma. *PLOS ONE* 2012;7(1):e28268.

4. Bime C, Pouladi N, Sammani S, et al. Genome-Wide Association Study in African Americans with Acute Respiratory Distress Syndrome Identifies the Selectin P Ligand Gene as a Risk Factor. *Am J Respir Crit Care Med* 2018;197(11):1421–1432.

5. Reilly JP, Wang F, Jones TK, et al. Plasma Angiopoietin-2 as a Potential Causal Marker in Sepsis-Associated ARDS Development: Evidence from Mendelian Randomization and Mediation Analysis. *Intensive Care Med* 2018;44(11):1849–1858.

6. Guillen-Guio B, Lorenzo-Salazar JM, Ma S-F, et al. Sepsis-associated acute respiratory distress syndrome in individuals of European ancestry: a genome-wide association study. *The Lancet Respiratory Medicine* 2020;8(3):258–266.

7. Ritchie MD, Denny JC, Crawford DC, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am J Hum Genet* 2010;86(4):560–572.

8. McKown AC, Brown RM, Ware LB, Wanderer JP. External Validity of Electronic Sniffers for Automated Recognition of Acute Respiratory Distress Syndrome. *J Intensive Care Med* 2017;0885066617720159.

9. Kerchberger VE, Brown RM, Semler MW, et al. Impact of Clinician Recognition of Acute Respiratory Distress Syndrome on Evidenced-Based Interventions in the Medical ICU. *Crit Care Explor* 2021;3(7):e0457.

10. Sjoding MW, Taylor D, Motyka J, et al. Deep learning to detect acute respiratory distress syndrome on chest radiographs: a retrospective study with external validation. *The Lancet Digital Health* 2021;3(6):e340–e348.

11. Sathe NA, Xian S, Mabrey FL, et al. Evaluating construct validity of computable acute respiratory distress syndrome definitions in adults hospitalized with COVID-19: an electronic health records based approach. *BMC Pulm Med* 2023;23(1):292.

12. Rogers AJ, Solus JF, Hunninghake GM, et al. MUC5B Promoter Polymorphism and Development of Acute Respiratory Distress Syndrome. *Am J Respir Crit Care Med* 2018;198(10):1342–1345.

13. O'Neal HR, Koyama T, Koehler EAS, et al. Prehospital Statin and Aspirin Use and the Prevalence of Severe Sepsis and ALI/ARDS. *Crit Care Med* 2011;39(6):1343–1350.

14. The ARDS Definition Task Force. Acute Respiratory Distress Syndrome: The Berlin Definition. *JAMA* 2012;307(23):2526–2533.

15. Danciu I, Cowan JD, Basford M, et al. Secondary Use of Clinical Data: the Vanderbilt Approach. *J Biomed Inform* 2014;52:28–35.

16. Cohen J. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 1960;20(1):37–46.

17. Das S, Forer L, Schönherr S, et al. Next-generation genotype imputation service and methods. *Nat Genet* 2016;48(10):1284–1287.

18. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016;48(10):1279–1283.

19. Etminan M, Collins GS, Mansournia MA. Using Causal Diagrams to Improve the Design and Interpretation of Medical Research. *CHEST* 2020;158(1):S21–S28.

20. Long JS, Ervin LH. Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model. *The American Statistician* 2000;54(3):217–224.

21. Zeileis A, Köll S, Graham N. Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in R. *Journal of Statistical Software* 2020;95:1–36.

22. Long JA. interactions: Comprehensive, User-Friendly Toolkit for Probing Interactions [Internet]. 2019. Available from: https://cran.r-project.org/package=interactions

23. Sjoding MW, Hofer TP, Co I, Courey A, Cooke CR, Iwashyna TJ. Interobserver Reliability of the Berlin ARDS Definition and Strategies to Improve the Reliability of ARDS Diagnosis. *Chest* 2018;153(2):361–367.

24. Roy MG, Livraghi-Butrico A, Fletcher AA, et al. Muc5b is required for airway defence. *Nature* 2014;505(7483):412–416.

25. Costain G, Liu Z, Mennella V, et al. Hereditary Mucin Deficiency Caused by Biallelic Loss of Function of MUC5B. *Am J Respir Crit Care Med* 2022;205(7):761–768.

26. Seibold MA, Wise AL, Speer MC, et al. A Common MUC5B Promoter Polymorphism and Pulmonary Fibrosis. *New England Journal of Medicine* 2011;364(16):1503–1512.

27. Hunninghake GM, Hatabu H, Okajima Y, et al. MUC5B Promoter Polymorphism and Interstitial Lung Abnormalities. *N Engl J Med* 2013;368(23):2192–2200.

28. Nakano Y, Yang IV, Walts AD, et al. MUC5B Promoter Variant rs35705950 Affects MUC5B Expression in the Distal Airways in Idiopathic Pulmonary Fibrosis. *Am J Respir Crit Care Med* 2016;193(4):464–466.

29. Putman RK, Hunninghake GM, Dieffenbach PB, et al. Interstitial Lung Abnormalities Are Associated with Acute Respiratory Distress Syndrome. *Am J Respir Crit Care Med* 2017;195(1):138–141.

30. Hancock LA, Hennessy CE, Solomon GM, et al. Muc5b overexpression causes mucociliary dysfunction and enhances lung fibrosis in mice. *Nat Commun* 2018;9(1):5363.

31. Pathak GA, Karjalainen J, Stevens C, et al. A first update on mapping the human genetic architecture of COVID-19. *Nature* 2022;608(7921):E1–E10.

32. Fadista J, Kraven LM, Karjalainen J, et al. Shared genetic etiology between idiopathic pulmonary fibrosis and COVID-19 severity. *EBioMedicine* 2021;65:103277.

33. Schwede M, Lee RY, Zhuo H, et al. Clinician Recognition of the Acute Respiratory Distress Syndrome: Risk Factors for Under-Recognition and Trends Over Time. *Crit Care Med* 2020;48(6):830–837.

34. Azzam HC, Khalsa SS, Urbani R, et al. Validation Study of an Automated Electronic Acute Lung Injury Screening Tool. *J Am Med Inform Assoc* 2009;16(4):503–508.

35. Herasevich V, Yilmaz M, Khan H, Hubmayr RD, Gajic O. Validation of an electronic surveillance system for acute lung injury. *Intensive Care Med* 2009;35(6):1018–1023.

36. Yetisgen-Yildiz M, Bejan C, Wurfel M. Identification of Patients with Acute Lung Injury from Free-Text Chest X-Ray Reports [Internet]. In: Cohen KB, Demner-Fushman D, Ananiadou S, Pestian J, Tsujii J, editors. Proceedings of the 2013 Workshop on Biomedical Natural Language Processing. Sofia, Bulgaria: Association for Computational Linguistics; 2013 [cited 2024 May 30]. p. 10–17.Available from: https://aclanthology.org/W13-1902

37. Afshar M, Joyce C, Oakey A, et al. A Computable Phenotype for Acute Respiratory Distress Syndrome Using Natural Language Processing and Machine Learning. *AMIA Annu Symp Proc* 2018;2018:157–165.

38. Crawford DC, Crosslin DR, Tromp G, et al. eMERGEing progress in genomics—the first seven years. *Front Genet* 2014;5:184.

39. Li H, Odeyemi YE, Weister TJ, et al. Rule-Based Cohort Definitions for Acute Respiratory Distress Syndrome: A Computable Phenotyping Strategy Based on the Berlin Definition. *Crit Care Explor* 2021;3(6):e0451.

40. Sinha P, Delucchi KL, Chen Y, et al. Latent class analysis-derived subphenotypes are generalisable to observational cohorts of acute respiratory distress syndrome: a prospective study. *Thorax* 2022;77(1):13–21.

41. Denny JC, Bastarache L, Ritchie MD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* 2013;31(12):1102–1110.

42. Liu G, Jiang L, Kerchberger VE, et al. The relationship between high density lipoprotein cholesterol and sepsis: A clinical and genetic approach. *Clin Transl Sci* 2023;16(3):489–501.

43. The Acute Respiratory Distress Syndrome Network. Comparison of Two Fluid-Management Strategies in Acute Lung Injury. *N Engl J Med* 2006;354(24):2564–2575.

44. Bellani G, Laffey JG, Pham T, et al. Epidemiology, Patterns of Care, and Mortality for Patients With Acute Respiratory Distress Syndrome in Intensive Care Units in 50 Countries. *JAMA* 2016;315(8):788–800.

45. Meyer NJ, Christie JD. Genetic Heterogeneity and Risk of Acute Respiratory Distress Syndrome. *Semin Respir Crit Care Med* 2013;34(4):459–474.

46. Jones TK, Feng R, Kerchberger VE, et al. Plasma sRAGE Acts as Genetically Regulated Causal Intermediate in Sepsis-Associated ARDS. *Am J Respir Crit Care Med* 2020;201(1):47–56.

47. Rubulotta F, Bahrami S, Marshall DC, Komorowski M. Machine Learning Tools for Acute Respiratory Distress Syndrome Detection and Prediction. *Critical Care Medicine* 2024;52(11):1768.

TABLES

Table 1: Study Population Demographics

Variable	Prospective cohort (VALID)	EHR Cohort (BioVU)
Demographics		
Number	2,795	9,025
Age (years) (median [IQR])	55 [43, 66]	60 [48, 70]
Age over 50 years (%)	1,753 (62.7)	6,413 (71.1)
Sex (%)		
Male	1,679 (60.1)	4,442 (49.2)
Female	1,116 (39.9)	4,583 (50.8)
Recorded Race (%)		
White	2,407 (86.1)	7,602 (84.2)
Black	350 (12.5)	1,170 (13.0)
Asian	5 (0.2)	70 (0.8)
Multiracial	0 (0.0)	48 (0.5)
Other/Not Recorded	33 (1.2)	135 (1.5)
Recorded Ethnicity (%)		
Non-Hispanic Non-Latino	2,769 (99.1)	8,754 (97.0)
Hispanic/Latino	21 (0.7)	157 (1.7)
Other/ Not Recorded	5 (0.2)	114 (1.3)
Hospital Characteristics		
Admitted to ICU (%)	2,795 (100.0)	3950 (43.8)
Admitting ICU (%)		
Burn	0 (0.0)	102 (2.6)
Cardiovascular	80 (2.9)	565 (14.3)
Medical	1375 (49.2)	1435 (36.3)
Neuroscience	0 (0.0)	346 (8.8)
Pediatrics	0 (0.0)	5 (0.1)
Surgical	501 (17.9)	1000 (25.3)
Trauma	839 (30.0)	497 (12.6)
Respiratory Characteristics		
Mechanical Ventilation (%)	2,128 (76.1)	1,866 (20.7)
Worst PaO_2/FiO_2 during observation (median [IQR])	161 [103, 240]	137 [87, 195]
Any PaO_2/FiO_2 Measured (%)	2,110 (75.5)	1,866 (20.7)
ARDS or EHR-ARDS (%)	718 (25.7)	1,056 (11.7)
Berlin ARDS Severity (%) [a]		
Mild (PaO_2/FiO_2 > 200 mmHg and ≤ 300 mmHg)	78 (10.8)	160 (15.1)
Moderate (PaO_2/FiO_2 > 100 mmHg and ≤ 200 mmHg)	363 (50.6)	494 (46.8)
Severe (PaO_2/FiO_2 ≤ 100 mmHg)	277 (38.6)	402 (38.1)
Hospital Outcomes		
ICU Length of Stay (days) (median [IQR]) [b]	6.0 [4.0, 11.0]	4.0 [2.0, 9.0]
Hospital Length of Stay (days) (median [IQR])	12.0 [7.0, 20.0]	5.0 [3.0, 9.0]
Inpatient Mortality (%)	456 (16.3)	537 (6.0)

Variable	Prospective cohort (VALID)	EHR Cohort (BioVU)
ARDS Risk Factors [c]		
Sepsis (%)	1,312 (46.9)	3,199 (35.4)
Pneumonia (%)	1,242 (44.4)	2,763 (30.6)
Trauma (%)	987 (35.3)	1,188 (13.2)
Shock (%)	1,913 (68.4)	4,656 (51.5)
Aspiration (%)	363 (13.0)	495 (5.5)
Pancreatitis (%)	136 (4.9)	749 (8.3)
Cardiac Comorbidity		
Heart Failure (%)	503 (18.0)	1,509 (16.7)
Acute Myocardial Infarction (%)	405 (14.5)	1,151 (12.8)
Other Coronary Artery Disease (%)	649 (23.2)	2,201 (24.4)
***MUC5B* Promoter Polymorphism Genotype (%)**		
GG	2,245 (80.3)	7,388 (81.9)
GT	513 (18.4)	1,540 (17.1)
TT	37 (1.3)	97 (1.1)

[a] Proportion of ARDS patients or EHR-ARDS participants
[b] Among patients/participants admitted to an ICU
[c] Patients/participants could have more than one ARDS risk factor

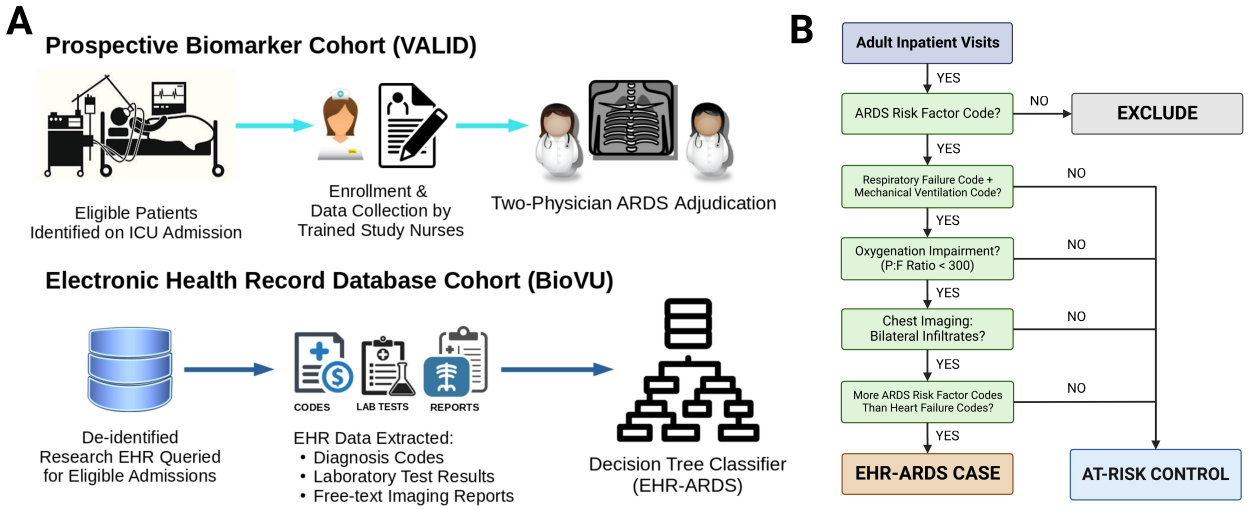
Table 2: Classification Performance of EHR-ARDS Classifier versus Clinician Review

	EHR Classifier Predictions			
	VALID		BioVU	
	EHR-ARDS	EHR At-Risk Control	EHR-ARDS	EHR At-Risk Control
ARDS by Clinician Review	616	102	33	2
Not ARDS by Clinician Review	632	1,445	17	73
Classification Metric [a]				
Sensitivity	0.86 (0.83-0.88)		0.94 (0.81-0.98)	
Specificity	0.70 (0.68-0.72)		0.81 (0.72-0.89)	
Positive Predictive Value	0.49 (0.47-0.52)		0.66 (0.52-0.78)	
Negative Predictive Value	0.93 (0.92-0.95)		0.97 (0.91-0.99)	
F_1 Score	0.63 (0.61-0.65)		0.78 (0.71-0.85)	
Cohen κ	0.45 (0.41–0.48)		0.66 (0.53-0.80)	

[a] Presented as point estimates and (95% confidence intervals). Confidence intervals were calculated using the Wilson score method for Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value, and F1 Score; and using a Wald interval for the κ-statistic.[16]

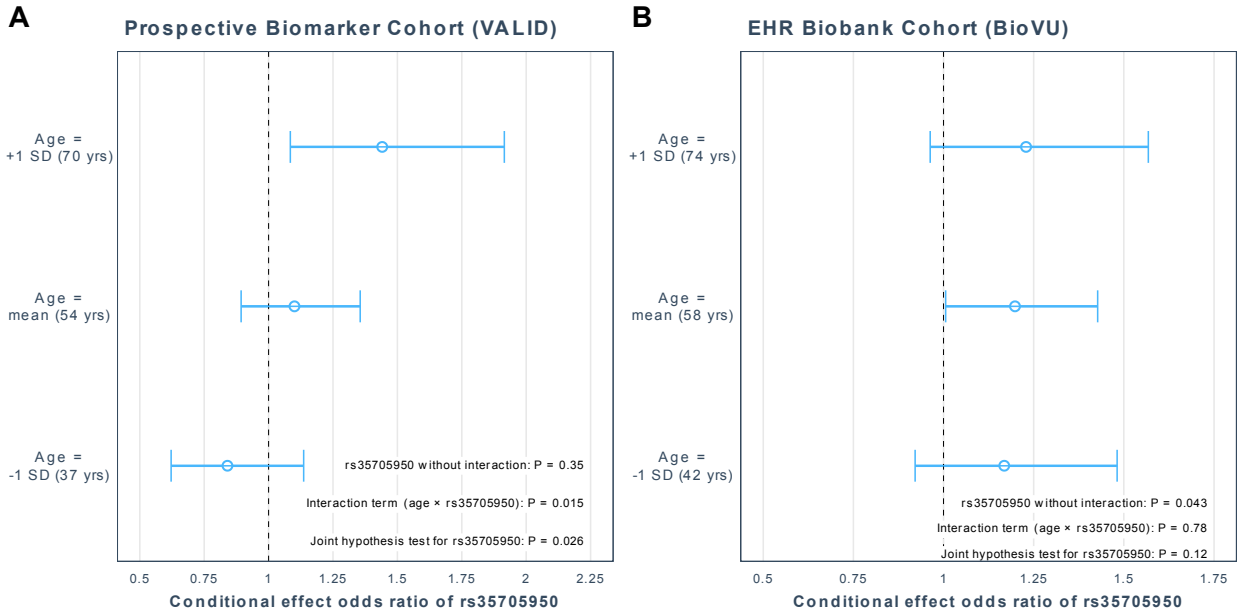
FIGURE LEGENDS

Figure 1: Schematic of Study Cohorts and EHR-ARDS Classifier



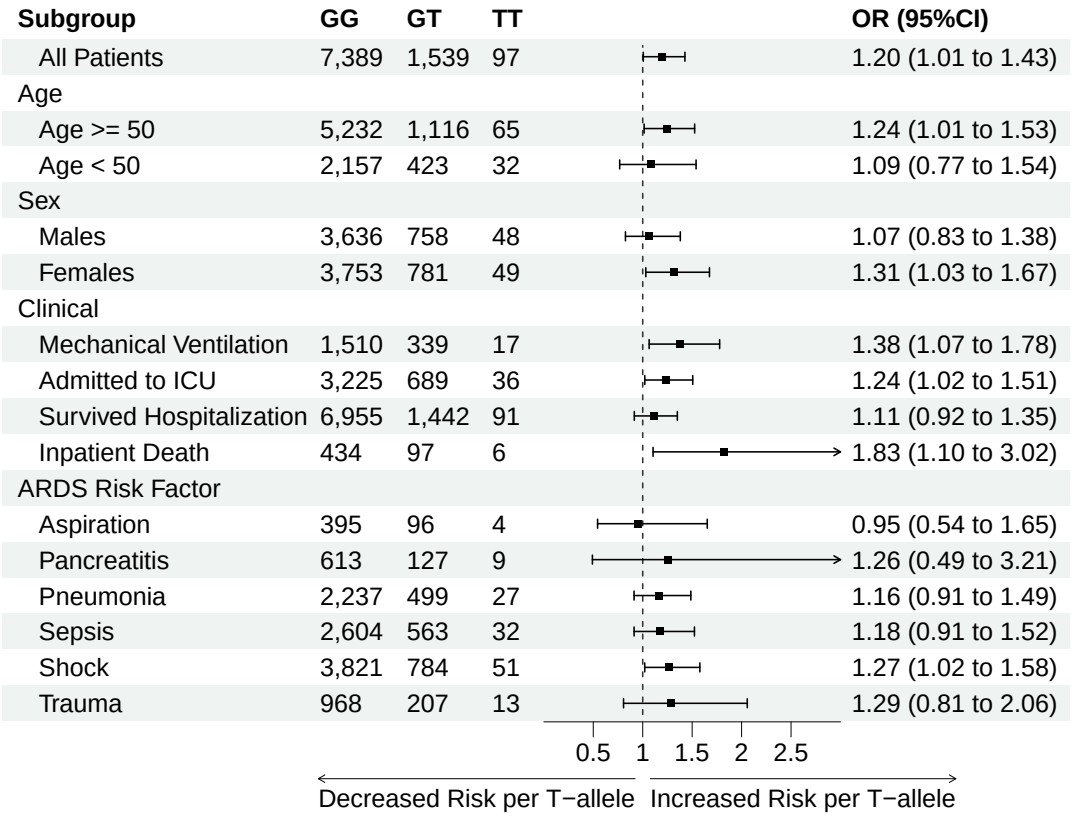
Schematic diagrams of (A) Study Cohorts and (B) EHR-ARDS Classifier.

Figure 2: *MUC5B* promoter polymorphism is associated with increased ARDS risk



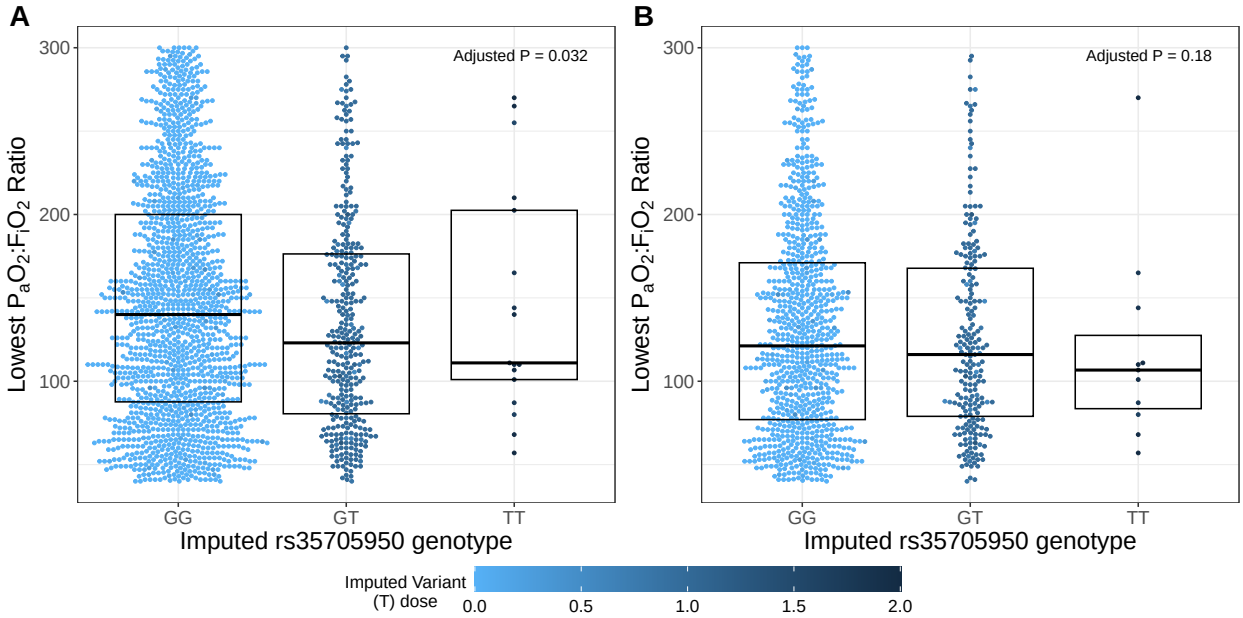
Conditional effects odds ratios for the *MUC5B* promoter polymorphism in (A) the prospective biomarker cohort (VALID), and (B) the EHR biobank cohort (BioVU). For each cohort, the y-axis indicates conditional odds ratios for (A) clinician-adjudicated ARDS or (B) EHR-ARDS per rs35705950 T-allele among younger (mean cohort age – 1 standard deviation), middle-aged (mean cohort age) and older (mean cohort age + 1 standard deviation) patients. Significant effect modification for age on rs35705950 was observed in VALID, but not in BioVU.

Figure 3: Subgroup Analysis of *MUC5B* Promoter Polymorphism and EHR-ARDS Risk in BioVU Cohort



The odds ratio and 95% confidence interval are shown overall and according to subgroup for association between the *MUC5B* promoter polymorphism variant allele and EHR-ARDS risk among participants in the BioVU cohort. All analyses were adjusted for age, sex, race (white versus non-white), presence of ARDS risk factors (exclusive of a risk factor when it is the subgroup definition), and presence of comorbid cardiac disorders (heart failure, acute myocardial infarction, or other coronary artery disease), but did not include an age × genotype interaction term.

Figure 4: *MUC5B* promoter polymorphism is associated with more severe oxygenation impairment in mechanically ventilated patients and ARDS patients in the BioVU Cohort



Oxygenation impairment as measured by lowest P_aO_2:F_iO_2 ratio during the first seven days of hospitalization according to *MUC5B* promoter polymorphism (rs35705950) genotype among (**A**) all BioVU participants who received mechanical ventilation (N = 1,866) and (**B**) BioVU participants who met EHR-ARDS classifier (N = 1,056). Boxes indicate the median and interquartile range across each genotype. Colors for each dot indicate the imputed T-allele dosage, a continuous value ranging from light blue (T-allele dosage = 0.0; GG genotype) to dark blue (T-allele dosage = 2.0; TT genotype). Imputed genotypes were assigned based on T-allele dosage ranges of 0.0 to 0.5 (GG genotype), 0.5 to 1.5 (GT genotype), or 1.5 to 2.0 (TT genotype).