Gastroenterology 2020;159:1129–1131

BRIEF COMMUNICATIONS

Famotidine Use Is Associated With Improved Clinical Outcomes in Hospitalized COVID-19 Patients: A Propensity Score Matched Retrospective Cohort Study

Daniel E. Freedberg,[1] Joseph Conigliaro,[2,3] Timothy C. Wang,[1] Kevin J. Tracey,[4] Michael V. Callahan,[5,6] and Julian A. Abrams,[1] on behalf of the Famotidine Research Group

[1]*Division of Digestive and Liver Diseases, Columbia University Irving Medical Center-New York Presbyterian Hospital, New York, New York;* [2]*Division of General Internal Medicine, Department of Medicine, Northwell Health, Manhasset, New York;* [3]*Zucker School of Medicine at Hofstra/Northwell, Hempstead, New York;* [4]*Feinstein Institutes for Medical Research, Northwell Health, Manhasset, New York;* [5]*Division of Infectious Diseases, Massachusetts General Hospital, Boston, Massachusetts; and* [6]*Office of the Assistant Secretary for Public Health Preparedness and Response, U.S. Department of Health and Human Services, Washington, DC*

See Covering the Cover synopsis on page 803.

Keywords: Coronavirus 2019; SARS-CoV-2; Famotidine; Histamine-2 Receptor Antagonists.

Coronavirus Disease 2019 (COVID-19) caused 2 million cases and more than 150,000 deaths worldwide as of mid-April 2020.[1] Clinical trials are under way to assess the efficacy of a variety of antiviral drugs; however, many of these drugs have toxicities and thus far no drug has been proven to improve outcomes in patients with COVID-19.

Famotidine is a histamine-2 receptor antagonist that suppresses gastric acid production. In vitro, famotidine inhibits human immunodeficiency virus replication.[2] Recently, Wu et al.[3] used computational methods to predict structures of proteins encoded by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genome and identified famotidine as one of the drugs most likely to inhibit the 3-chymotrypsin-like protease ($3CL^{pro}$), which processes proteins essential for viral replication.[4] We hypothesized that famotidine would be associated with improved clinical outcomes among hospitalized patients with COVID-19. To explore this, we performed a retrospective cohort study at a single academic center located at the epicenter of the COVID-19 pandemic in the United States.

Methods

Complete methods are available in the Supplementary Materials. In brief, adults were eligible for the study if they were admitted to our institution from February 25, 2020, to April 13, 2020, and tested positive for SARS-CoV-2 within no more than 72 hours following admission. Patients were excluded if they died or were intubated within 48 hours following hospital admission. The primary exposure was use of famotidine (any dose, form of administration, or duration), classified as present if famotidine was received within 24 hours of hospital admission and otherwise as absent. The primary outcome was a composite of death or endotracheal intubation from hospital day 2 to day 30 (intubation-free survival). This follow-up period avoided immortal time bias because the exposure was classified based on the 24-hour period after hospitalization and the at-risk period began on hospital day 2. Cox proportional hazards modeling was performed on the full cohort, and a matched subset was examined with propensity scoring matching to balance baseline characteristics based on use of famotidine.

Results

Population and Use of Famotidine

A total of 1620 patients met criteria for analysis, including 84 patients (5.1%) who received famotidine within 24 hours of hospital admission. Home use of famotidine was documented on admission medication reconciliation in 15% of those who used famotidine while hospitalized compared with 1% of those who did not ($P < .01$). Twenty-eight percent of all famotidine doses were intravenous; 47% were 20 mg, 35% were 40 mg, and 17% were 10 mg. Famotidine users received a median 5.8 days of drug for a total median dose of 136 mg (63–233 mg). There were minimal differences comparing patients who used famotidine with those who did not, and balance between the groups was further improved after propensity score matching (Supplementary Table 1).

Death or Intubation

A total of 142 (8.8%) patients were intubated and 238 (15%) died; 340 (21%) patients met the composite study outcome. In crude analysis, use of famotidine was significantly associated with reduced risk for the composite outcome of death or intubation (Figure 1*A*, log-rank $P < .01$). This association was driven primarily by the relationship between famotidine and death (Figure 1*B*, log-rank

Abbreviations used in this paper: **CI, confidence interval; COVID-19, Coronavirus Disease 2019; PPI, proton pump inhibitor; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.**

Most current article

0016-5085/$36.00
https://doi.org/10.1053/j.gastro.2020.05.053

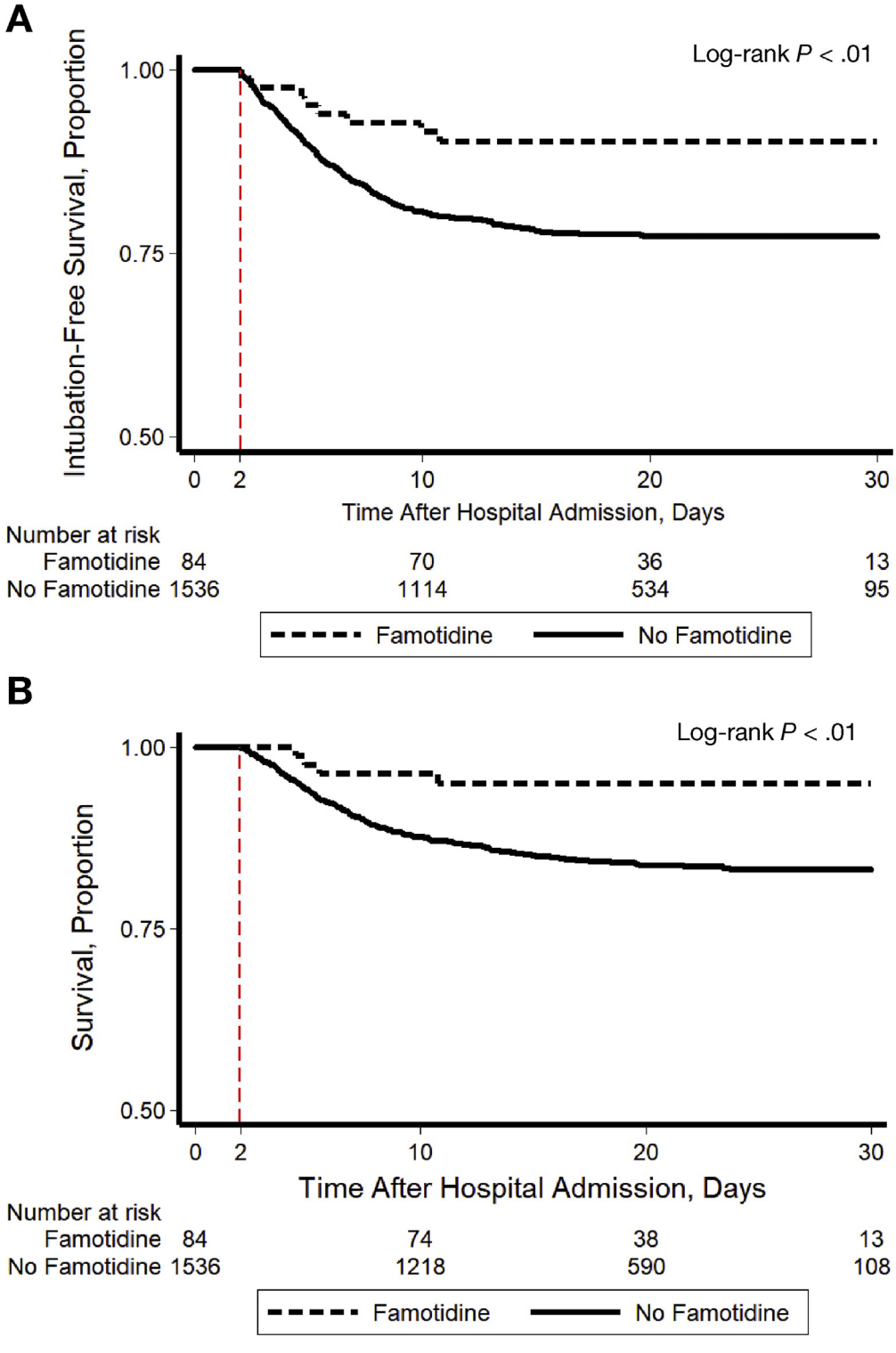


Figure 1. Kaplan-Meier plot showing (*A*) intubation-free survival and (*B*) survival through a maximum of 30 days after hospital admission, stratified by use of famotidine. Patients were included in the study if they survived without intubation for 2 days following hospital admission. Use of famotidine was classified as present if it was received within the first 24 hours following hospital admission (any dose, form of administration, or duration) and otherwise as absent. The at-risk time began on hospital day 2 (indicated with a *dashed red line*) and patients were followed until hospital day 30. This study design avoided potential for immortal time bias because the exposure was classified before the start of the at-risk period.

$P < .01$) and when those who died before intubation were excluded, there was no association between use of famotidine and intubation (log-rank $P = .40$). After adjusting for baseline patient characteristics, use of famotidine remained independently associated with risk for death or intubation (Supplementary Table 2, adjusted hazard ratio 0.42, 95% confidence interval [CI] 0.21–0.85) and this remained unchanged after propensity score matching to further balance the covariables (hazard ratio 0.43, 95% CI 0.21–0.88).

Additional Analyses

Use of proton pump inhibitors (PPIs) was analyzed because PPIs are also gastric acid suppression medications with similar indications as famotidine. There was a no protective effect

associated with use of PPIs (adjusted hazard ratio 1.34, 95% CI 1.06–1.69). Next, 784 patients without COVID-19 who were hospitalized during the same study period were analyzed; among these patients, famotidine was not associated with reduced risk for death or intubation (24 deaths or intubations, log-rank $P = .70$). The maximum plasma ferritin value during the hospitalization was assessed to address the hypothesis that, by blocking viral replication, famotidine reduces cytokine storm during COVID-19. Median ferritin was 708 ng/mL (interquartile range 370–1152) among users of famotidine vs 846 ng/mL (interquartile range 406–1552) among nonusers (rank-sum $P = .03$).

Conclusions

This retrospective study found that, in patients hospitalized with COVID-19, famotidine use was associated with a reduced risk of clinical deterioration leading to intubation or death. The study was premised on the assumption that use of famotidine represented a continuation of home use, but documentation of why famotidine was given was poor. The results were specific for famotidine (no protective association was seen for PPIs) and also specific for COVID-19 (no protective association in patients without COVID-19). A lower peak ferritin value was observed among users of famotidine, supporting the hypothesis that use of famotidine may decrease cytokine release in the setting of SARS-CoV-2 infection. A randomized controlled trial is currently under way to determine whether famotidine can improve clinical outcomes in hospitalized patients with COVID-19 (NCT04370262).

Famotidine has not previously been studied in patients for antiviral effects, and there are limited relevant prior data. An untargeted computer modeling analysis identified famotidine as one of the highest-ranked matches for drugs predicted to bind $3CL^{pro}$,[3] a SARS-CoV-2 protease that generates nonstructure proteins critical to viral replication.[4] In the 1990s, histamine-2 receptor antagonists including famotidine were shown to inhibit human immunodeficiency virus replication without affecting lymphocyte viability in vitro.[2,5,6]

There are limitations to the study. It was observational, and we cannot exclude the possibility of unmeasured confounders or hidden bias that account for the association between famotidine use and improved outcomes. No samples were gathered, and mechanism cannot be directly assessed. Finally, this was a single-center study, which may limit generalizability of the findings.

In sum, in patients hospitalized with COVID-19 and not initially intubated, famotidine use was associated with a 2-fold reduction in clinical deterioration leading to intubation or death. These findings are observational and should not be interpreted to mean that famotidine has a protective effect against COVID-19. Randomized controlled trials are under way.

Supplementary Material

Note: To access the supplementary material accompanying this article, visit the online version of *Gastroenterology* at www.gastrojournal.org, and at https://doi.org/10.1053/j.gastro.2020.05.053.

References

1. Gottschlich MM, DeLegge MH, Guenter P. American Society for Parenteral and Enteral Nutrition. Silver Spring, MD: American Society for Parenteral and Enteral Nutrition, 2007.
2. Bourinbaiar AS, Fruhstorfer EC. Life Sci 1996;59:PL365–370.
3. Wu C, Liu Y, Yang Y, et al. Acta Pharm Sin B 2020; 10:766–788.
4. Anand K, Ziebuhr J, Wadhwani P, et al. Science 2003; 300:1763–1767.
5. Chen X, Deng H, Churchill MJ, et al. Cell Stem Cell 2017; 21:747–760.e7.
6. Li X, Zhang C, Liu L, Gu M. FASEB J 2020;34:6008–6016.

Received May 4, 2020. Accepted May 14, 2020.

Correspondence
Address correspondence to: Daniel Freedberg, MD, MS, Division of Digestive and Liver Diseases, Columbia University Irving Medical Center, 630 West 168th Street, P&S 3-401, New York, NY 10032. e-mail: def2004@cumc.columbia.edu; or Julian Abrams, MD, MS, Division of Digestive and Liver Diseases, Columbia University Irving Medical Center, 630 West 168th Street, P&S 3-401, New York, NY 10032. e-mail: ja660@cumc.columbia.edu.

Acknowledgments
The authors thank Dr Michael Wigler and Dr Richard Axel for useful suggestions.

Appendix
Additional members of the Famotidine Research Group
Magdalena E. Sobieszczyk, MD, MPH (Division of Infectious Diseases, Columbia University Irving Medical Center-New York Presbyterian Hospital, New York, New York), David D. Markowitz, MD (Division of Digestive and Liver Diseases, Columbia University Irving Medical Center-New York Presbyterian Hospital, New York, New York), Aakriti Gupta, MD, MS (Division of Cardiology, Columbia University Irving Medical Center-New York Presbyterian Hospital, New York, New York), Max R. O'Donnell, MD, MPH (Division of Pulmonary, Allergy, and Critical Care Medicine, Columbia University Irving Medical Center-New York Presbyterian Hospital, New York, New York), Jianhua Li, MD (Department of Medicine, Columbia University Irving Medical Center-New York Presbyterian Hospital, New York, New York), David A. Tuveson, MD, PhD (Cancer Center, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York), Zhezhen Jin, PhD (Department of Biostatistics, Columbia University Mailman School of Public Health, New York, New York), William C. Turner, MD (Department of Medicine, Columbia University Irving Medical Center-New York Presbyterian Hospital, New York, New York), and Donald W. Landry, MD, PhD (Department of Medicine, Columbia University Irving Medical Center-New York Presbyterian Hospital, New York, New York)

CRediT Authorship Contributions
Daniel E. Freedberg, MD, MS (Conceptualization: Equal; Data curation: Equal; Formal analysis: Equal; Investigation: Equal; Methodology: Equal; Project administration: Equal; Validation: Equal; Writing – original draft: Equal; Writing – review & editing: Equal). Joseph Conigliaro, MD, MPH (Conceptualization: Equal). Timothy C. Wang, MD (Conceptualization: Equal). Kevin J. Tracey, MD (Conceptualization: Equal). Michael V. Callahan, MD (Conceptualization: Equal). Julian A. Abrams, MD, MS (Conceptualization: Equal; Funding acquisition: Equal; Investigation: Equal; Methodology: Equal; Project administration: Equal; Writing – original draft: Equal; Writing – review & editing: Equal).

Conflict of interest
The authors disclose no conflicts.

Supplementary Methods

Population

Adults aged 18 years or older were eligible for the study if they were admitted to Columbia University Irving Medical Center or its affiliate the Allen Pavilion from February 25, 2020, to April 13, 2020, and tested positive for SARS-CoV-2 by nasopharyngeal polymerase chain reaction at presentation or within no more than 72 hours following admission. This 72-hour window was selected because, during the earliest phase of the SARS-CoV-2 pandemic, testing availability was limited and could take up to 72 hours for a result. Patients were excluded if they survived less than 48 hours following hospital admission or if they required urgent or semi-urgent intubation within 48 hours of hospital admission. This study was approved by the institutional review board of the Columbia University Irving Medical Center.

Exposure

The primary exposure was use of famotidine, classified as present if famotidine was received within 24 hours of hospital admission and otherwise classified as absent. Famotidine use was ascertained directly from electronic medical order entry records and could be intravenous or oral, at any dose or duration. Home use of famotidine was examined to understand the reason underlying in-hospital use of famotidine and was classified based on electronic medication reconciliation performed at the time of hospital admission.

Primary Outcome

The primary outcome was a composite of death or endotracheal intubation within 30 days of hospital admission (intubation-free survival). Mortality data were ascertained from the electronic medical record (EMR), which interfaces with the social security death index. Endotracheal intubation was ascertained from EMR documentation of need for mechanical ventilation. The rationale for the combined primary outcome was 2-fold: (1) many patients who deteriorated clinically died without being intubated, often due to transition to palliative care; (2) hospitalization stays for intubated patients with COVID-19 have been very long, and many intubated patients with COVID-19 at the time of the analyses may ultimately not survive.

Covariables

Based on emerging reports of risk factors for COVID-19, the following covariables were selected for inclusion in the analysis: preexisting diabetes, hypertension, coronary artery disease, heart failure, end-stage renal disease or chronic kidney disease, and chronic pulmonary disorders, all classified based on the presence of corresponding International Classification of Diseases, 10th Revision codes at the time of hospital admission; obesity, classified based on body mass index; and age, classified as $<$50 years, 50 to 65 years, and $>$65 years. To assess severity of COVID-19, the first recorded form of supplemental oxygen after triage was captured and was classified as room air, nasal cannula oxygen, or non-rebreather/similar. Use of PPIs was classified in the same manner as use of famotidine so that PPIs could be evaluated to test whether any effects of famotidine might be related to acid suppression. The maximum value of plasma ferritin was obtained during the study period for each patient to use as a surrogate for the extent of cytokine storm (normal laboratory range 13.0 to 150.0 ng/mL).

Statistical Approach

Categorical variables were compared across exposure groups using χ^2 tests. Full and reduced Cox proportional hazards models were constructed within the complete cohort, with patients followed from the time of hospital admission until the first of the following events: death, intubation, 30 days of follow-up, or the close of the study on April 20, 2020. Because patients were excluded if they died or were intubated before hospital day 2, this effectively meant that patients were followed from day 2 to day 30. This design was selected to avoid immortal time bias (ie, because the exposure was classified based on the 24-hour period after hospitalization and the at-risk period did not begin until hospital day 2). Cox proportional hazards modeling was performed on the full cohort, and a matched subset was examined with propensity scoring matching to balance baseline characteristics based on use of famotidine.

This provided the opportunity for a minimum of 7 days of follow-up time for all patients in the study. The proportional hazards assumption was verified by visual inspection of time-to-event data and by testing for a nonzero slope in the Schoenfeld residuals (11). The full Cox model included all baseline variables. For the reduced model, variables were dropped stepwise unless they had a significant independent relationship with the composite outcome or unless they altered the β-coefficient representing famotidine by at least 10%. Propensity score matching was then performed to balance the baseline characteristics of patients with respect to use of famotidine with a 5:1 nearest-neighbor matching strategy and a caliper of 0.2. The primary analysis was conducted as a time-to-event model within the propensity score–matched cohort, using the same approach. All analyses were performed using STATA statistical software (version 14; StataCorp, College Station, TX) at the $\alpha = 0.05$ level of significance.

Additional Analyses

Several sensitivity analyses were performed. First, use of PPIs was compared with no PPIs within the complete (unmatched) cohort after excluding those who used famotidine. The purpose of this analysis was to test whether unmeasured patient characteristics related to use of acid suppression rather than famotidine were associated with improved outcomes in COVID-19. Second, an additional study cohort was built including records from patients who tested negative for SARS-CoV-2 during the study period. Within this cohort, use of famotidine was compared with no famotidine to test whether unmeasured patient characteristics related to use of famotidine were associated with improved outcomes regardless of reason for hospitalization (ie, to test whether the observed association with famotidine was specific for patients with COVID-19).

Supplemental Table 1. Patient Characteristics at the Time of Hospital Admission for COVID-19, Stratified by Use of Famotidine

Characteristics	Complete cohort			After propensity score matching		
	Famotidine (n = 84), n (%)	No famotidine (n = 1536), n (%)	*P* value	Famotidine (n = 84), n (%)	No famotidine (n = 420), n (%)	*P* value
Age (*y*)			.39			.51
<50	13 (15)	320 (21)		13 (15)	57 (14)	
50–65	31 (37)	483 (31)		31 (37)	184 (44)	
>65	40 (48)	733 (48)		40 (48)	179 (43)	
Female sex	39 (46)	864 (56)	.63	39 (46)	208 (50)	.60
Race/ethnicity			.20			.90
Hispanic	25 (30)	601 (39)		25 (30)	127 (30)	
White, non-hispanic	19 (23)	336 (22)		19 (23)	82 (20)	
Black, non-hispanic	18 (21)	322 (21)		18 (21)	102 (24)	
Other	22 (26)	277 (18)		22 (26)	109 (26)	
BMI, kg/m^2			.17			.97
<25.0	15 (18)	295 (19)		15 (18)	66 (16)	
25.0–29.9 (overweight)	30 (36)	388 (25)		30 (36)	157 (37)	
≥30 (obese)	22 (26)	434 (28)		22 (26)	110 (26)	
Not recorded	17 (20)	419 (27)		17 (20)	87 (21)	
Comorbidities						
Diabetes	24 (29)	311 (20)	.07	24 (29)	106 (25)	.52
Hypertension	29 (35)	428 (28)	.19	29 (35)	124 (30)	.36
CAD	9 (11)	109 (7)	.21	9 (11)	37 (9)	.58
Heart failure	7 (8)	85 (6)	.28	7 (8)	26 (6)	.47
ESRD or CKD	11 (13)	130 (8)	.14	11 (13)	47 (11)	.62
Chronic pulmonary disorders	2 (2)	120 (8)	.07	2 (2)	6 (11)	.52
Initial oxygen requirement			.39			.85
Room air	25 (30)	378 (25)		25 (30)	116 (28)	
Nasal canula	38 (45)	678 (44)		38 (45)	187 (44)	
Non-rebreather or similar	21 (25)	480 (31)		21 (25)	117 (28)	

BMI, body mass index; CAD, coronary artery disease; CKD, chronic kidney disease; ESRD, end-stage renal disease.

Supplemental Table 2. Final Cox Proportional Hazards Model of risk factors for death or Intubation Among Patients With COVID-19

Characteristics	Death or intubation/n at risk (%)	Hazard ratio (95% CI)	
		Full model	Final model
Famotidine			
No	332/1536 (22)	Reference	Reference
Yes	8/84 (10)	0.43 (0.21–0.86)	0.42 (0.21–0.85)
Age (*y*)			
<50	19/333 (5.7)	Reference	Reference
50–65	75/514 (15)	2.94 (1.77–4.89)	3.03 (1.83–5.03)
>65	246/773 (32)	7.51 (4.66–12.1)	7.68 (4.79–12.3)
Sex			
Male	197/909 (22)	Reference	—
Female	143/711 (20)	1.11 (0.89–1.38)	
Race/ethnicity			
Hispanic	129/626 (21)	Reference	—
White, non-Hispanic	84/355 (24)	0.99 (0.75–1.31)	
Black, non-Hispanic	59/340 (17)	0.82 (0.60–1.13)	
Other	68/299 (23)	1.14 (0.85–1.53)	
Body mass index, kg/m^2			
<25.0	86/310 (28)	Reference	—
25.0–29.9 (overweight)	92/418 (22)	0.88 (0.65–1.18)	
≥30 (obese)	89/456 (20)	0.97 (0.72–1.31)	
Not recorded	73/436 (17)	0.67 (0.49–0.92)	
Comorbidities			
Diabetes	72/335 (21)	1.02 (0.75–1.37)	—
Hypertension	94/457 (21)	0.72 (0.54–0.97)	0.74 (0.58–0.94)
CAD	24/118 (20)	0.77 (0.49–1.21)	—
Heart failure	24/92 (26)	1.06 (0.67–1.67)	—
ESRD or CKD	33/141 (23)	1.16 (0.77–1.75)	—
Chronic pulmonary disorders	29/122 (24)	1.29 (0.87–1.93)	—
Initial oxygen requirement			
Room air	52/403 (13)	Reference	—
Nasal canula	155/716 (22)	1.60 (1.17–2.19)	1.63 (1.19–2.24)
Non-rebreather	133/501 (27)	2.48 (1.79–3.44)	2.39 (1.73–3.29)

CAD, coronary artery disease; CKD, chronic kidney disease; ESRD, end-stage renal disease.

British Journal of Educational Psychology (2012), 82, 647–671

The British Psychological Society

www.wileyonlinelibrary.com

Reading development in a tracked school system: A longitudinal study over 3 years using propensity score matching

Jan Retelsdorf[1]*, Michael Becker[2,3], Olaf Köller[1] and Jens Möller[4]

[1]Leibniz Institute for Science and Mathematics Education, Kiel, Germany
[2]University of Potsdam, Germany
[3]Max Planck Institute for Human Development, Germany
[4]Christian-Albrechts-University of Kiel, Germany

Background. Assigning students to different school tracks on the basis of their achievement levels is a widely used strategy that aims at giving students the best possible learning opportunity. There is, however, a growing body of literature that questions such positive effects of tracking.

Aims. This study compared the developmental trajectories of reading comprehension and decoding speed between students at academic track schools that typically prepare students for university entrance and students at non-academic track schools that usually prepare students for vocational education.

Sample. In a longitudinal design with three occasions of data collection, the authors drew on a sample of $N = 1{,}508$ 5th graders (age at T1 about 11 years, age at T3 about 14 years) from 60 schools in Germany. The academic track sample comprised $n = 568$ students; the non-academic track sample comprised $n = 940$ students.

Method. Achievement measures were obtained by standardized tests of reading comprehension and decoding speed. Students at the different tracks were closely matched using propensity scores. To compare students' growth trajectories between the different school tracks, we applied multi-group latent growth curve models.

Results. Comparable results were recorded for the complete (unmatched) sample and for the matched pairs. In all cases, students at the different tracks displayed a similar growth in reading comprehension, whereas larger growth rates for students at academic track schools were recorded for decoding speed.

Conclusions. Our findings contribute to an increasing body of literature suggesting that tracking might have undesired side effects.

*Correspondence should be addressed to Jan Retelsdorf, Leibniz Institute for Science and Mathematics Education, Kiel, Olshausenstraße 62, 24118 Kiel, Germany (e-mail: jretelsdorf@ipn.uni-kiel.de).

DOI:10.1111/j.2044-8279.2011.02051.x

Reading skills are mainly developed during preschool and elementary school years. Even though these fundamentals affect later comprehension in adolescence (Cunningham & Stanovich, 1997), reading development does not come to an end on finishing elementary school. In fact, fostering reading comprehension is still an important task of secondary schools, since many students lack sufficient proficiency as readers even at the age of 15 (Kirsch *et al*., 2002). During secondary school, tracking – the assignment of students to different school tracks on the basis of their achievement levels – is a particular feature of school in many countries, which might affect the development of reading skills during these years. However, research on reading skill development in a tracked school system is scarce. This study aimed at comparing the developmental trajectories of reading comprehension and decoding speed between students at different school tracks. First, we will review some literature on tracking before we present the aims of our research in detail.

Tracking

Although there are certain differences, the educational systems of most industrialized countries use ability grouping in one way or another (LeTendre, Hofer, & Shimizu, 2003). The most important rationale underlying grouping students with regard to their achievement level is that all students are supposed to learn best, when they are in a homogeneous group of students with comparable abilities (e.g., Oakes, 1987; Pallas, Entwisle, Alexander, & Stluka, 1994). According to this assumption, teaching commensurate with individual requirements is said to be much easier and more effective in groups displaying homogenous capabilities and teachers can more easily provide appropriate learning opportunities (e.g., Kulik & Kulik, 1992). Thus, the aim of tracking is to give students the best possible learning opportunity. Apart from this, tracking has been criticized because students in lower track schools are at a disadvantage to those in higher track schools with a resultant lowering of achievement and motivation (e.g., Lucas, 1999; Oakes, 1987) or even an anti-school culture (Van de Gaer, Pustjens, Van Damme, & De Munter, 2006). However, there is also some indication that tracking enhances students' achievement at all tracks (Mulkey, Catsambis, Steelman, & Crain, 2005) or that students at lower track schools might even benefit from tracking with regard to a positive development of the self-concept (Ireson, Hallam, & Plewis, 2001; Liu, Wang, & Parkins, 2005; Marsh, 1987) and lower amounts of burnout (Salmela-Aro, Kiuru, & Nurmi, 2008). Since there are many different forms of tracking, following we will describe the particular kind of tracking that is relevant in our study before we will discuss its possible effects on achievement development.

The extent of tracking varies across countries and educational systems. Referring to Trautwein, Lüdtke, Marsh, Köller, and Baumert (2006) there are three main features describing forms and intensity of tracking: the institutional level, the role of achievement, and the impact on future academic careers. First, tracking strategies differ with regard to the institutional level. Hereby, we can distinguish between forms of within-class ability grouping (mainly in the early grades), course-level grouping, which is common in secondary school, and grouping at school level. According to the latter, on the one hand there is some form of implicit tracking depending on the catchment areas of schools and, on the other hand, there is explicit tracking involving different school types characterized by achievement levels and specific curricula. The second feature, the role of achievement, deals with the decision criteria on placement in a certain track. This decision can be influenced by students' achievement (achievement grouping) or

other factors such as parents' socio-economic status (SES) and educational aspirations (opt-in tracking, Trautwein *et al*., 2006). Third, the impact of tracking on future academic careers can differ to some extent. In some educational systems, placing students in a lower track reduces their chances of attending university and obtaining a degree (e.g., Japan), whereas in other systems, this association is less rigid (e.g., United States).

In this study, we investigated the effects of tracking on reading achievement in the German educational system where we are dealing with explicit grouping at school level. After elementary school, students in Germany are assigned to different types of school that either place a focus on students' gaining qualifications that would enable them to begin a vocational apprenticeship (non-academic track schools) or prepare them for university entrance (academic track schools) so that tracking does affect future academic careers to a certain extent. Based on students' mean achievement level, teachers recommend an appropriate school track for each student at the end of elementary school. This recommendation, however, is not obligatory, and the final decision regarding which school track a child is to be placed in is ultimately that of the parents. Thus, in Germany the decision criterion for a particular track is a hybrid of achievement grouping and opt-in tracking. All in all, the German educational system is said to be the most strictly stratified school system of the Western industrialized countries (cf. Trautwein *et al*., 2006).

Academic and non-academic tracks differ with regard to composition, such as mean achievement levels or parents' SES, and institutional factors such as the curriculum (e.g., LeTendre *et al*., 2003). These compositional and institutional track differences can have manifold consequences that might affect students' learning and achievement. For example, the composition of students can create an environment, in which particular values and norms are predominant. Students internalize these norms and values (e.g., Barth, Dunlap, Dane, Lochman, & Wells, 2004) and thus develop attitudes towards learning that might have either positive or negative effects on achievement. Moreover, there is some evidence that teachers' beliefs, knowledge, and instructional practices differ between tracks. Hallam and Ireseon (2003) found that teachers differ in their beliefs about ability grouping. Furthermore, in a recent study Baumert *et al*. (2010) observed significant differences between teachers from academic and non-academic tracks in content knowledge and pedagogical content knowledge. These differences in teachers' beliefs and their knowledge might influence their use of instructional practices and in the long run students' achievement. Indeed, there is some research showing that teachers at higher track schools provide higher levels of problem solving and cognitive activating instruction, whereas in lower track schools exercises in class, memorization, and disciplining students are emphasized (Kunter & Baumert, 2006; Oakes, 1985; Raudenbush, Rowan, & Cheong, 1993; Retelsdorf, Butler, Streblow, & Schiefele, 2010; Van Houtte, 2004). All in all, these track differences in compositional and institutional conditions might be responsible for achievement gaps between academic and non-academic track schools.

Despite the long-term political and scientific debate on explicit tracking (e.g., Oakes, 1985), empirical studies on its particular effects on reading achievement are scarce. One of the rare longitudinal studies dealing with the effects of explicit tracking on reading achievement is from Maughan and Rutter (1987). They found that reading scores at age of 14 were higher in grammar schools than at non-selective schools when controlling for differences in intake scores. The sample size in that study, however, was rather small ($N = 160$). Another investigation of track differences in reading was a previous study with the sample of the present study, in which Retelsdorf and Möller (2008) researched

the consequences of explicit school-level tracking in Germany just at the beginning of secondary school with two available waves of data. On applying latent difference score analyses, large differences in the initial level between different school types were found but no significant differences for the development of reading comprehension were recorded. The effect sizes of reading comprehension growth, however, indicated that, by trend, academic track students increased more (growth at academic track: $d = 0.82$; average growth on non-academic tracks: $d = 0.61$). In addition, academic track students displayed significantly lower decreases in reading motivation than non-academic track students. Thus, the authors speculated that the achievement gap might widen during forthcoming school years.

In contrast to research on reading development in tracked school systems, there seems to be strong empirical support for widening achievement gaps in the mathematics domain (e.g., Argys, Rees, & Brewer, 1996; Becker, Lüdtke, Trautwein, & Baumert, 2006; Hoffer, 1992; Köller & Baumert, 2001). These studies found disadvantages for students attending non-academic tracks. Moreover, Becker (2009) recently found that students in academic track schools benefited with regard to psychometric intelligence when compared to students in non-academic track schools.

As such, one might well conclude that the type of school affects students' achievement growth. Drawing such causal inferences, however, is usually inappropriate when based on data derived from observational studies that lacks additional assumptions and adjustment, respectively. Thus, we are unable to decide whether a widening achievement gap between different school tracks is a result of the type of school or if it is quite simply the result of previously existing differences among students.

The present investigation

The purpose of our study was to investigate the effect of explicitly different school tracks in Germany on students' reading development over a 3-year period. Therefore, we extended a previous study (Retelsdorf & Möller, 2008) drawing on the same sample. The previous study, however, comprised of only two occasions of data collection and did not address the problem of causal inference. The results of that previous study indicated that the reading achievement gap between different school tracks might widen as the students grow older, because a trend favouring academic track students was observed. Moreover, academic track students benefited from lower motivational decreases. Thus, the authors conjectured that the achievement gap might widen over a longer period of time. Furthermore, Retelsdorf and Möller (2008) investigated only reading comprehension as an achievement measure of the domain of reading. However, in research investigating individual Matthew effects (Stanovich, 1986) in reading – which describe reading development as a cumulative process, where those with a high initial reading level also gain higher growth than those with a lower initial level – it was observed that various reading skills might develop quite differently (Bast & Reitsma, 1998; Parrila, Aunola, Leskinen, Nurmi, & Kirby, 2005). In fact, there is some evidence that reading skills might even develop in a rather compensatory manner, that is, children with lower initial levels catch up (Aarnoutse, Leeuwe, Voeten, & Oud, 2001; Aunola, Leskinen, Onatsu-Arvilommi, & Nurmi, 2002; Baumert, Nagy, & Lehmann, in press). Describing the research on a cumulative versus compensatory individual reading development in detail is far beyond the scope of this paper. Some particular findings, however, might reveal that there is a great variability in the development of various aspects of reading,

which emphasizes the importance of the investigation of different reading skills. For example, Parrila *et al*. (2005) found rather consistent support for the compensatory model across different reading measures in a Canadian sample, whereas in a Finnish sample their results were comparable to Bast and Reitsma (1998). These authors observed that a compensatory development was applicable for sentence comprehension, while a cumulative development was the result for word recognition. To test the developmental trajectories of different reading skills, an additional test covering decoding speed was included in this study to investigate exactly that. Since the development of reading skills is not necessarily linear, another limitation of Retelsdorf and Möller (2008) was that their study only comprised of two points of assessment and thus did not allow for a modelling of non-linear developmental trajectories. To address this problem and to test the authors' assumption that the achievement gap might widen in the long run, we used data from three points of measurement in the present study.

For both aspects of reading, we first applied latent growth curve models (LGCMs) for the whole sample to describe the developmental trajectories within the different school tracks. To enhance the validity of the conclusion that possible differences are due to type of school, we then repeated our analyses for a matched sample obtained by applying propensity score matching (e.g., Rosenbaum & Rubin, 1983). This approach allows for a careful drawing of causal conclusions from non-experimental data. The idea of propensity score matching is to approximate the effect of randomization by modelling the mechanism of group assignment (in our case, the assignment to different school tracks) and, thus, to eliminate the correlation between this assignment and the outcome. This procedure, if successfully applied, leads to groups of individuals with the same probability of belonging to either academic or non-academic track schools. Thus, propensity score matching is a question of modelling group assignment rather than the outcome (e.g., Schafer & Kang, 2008). Therefore, it is important to have comprehensive background information that is connected to the group assignment. In this study, we used covariates, which have been associated with the preference or choice of a particular school track in former research (Arnold, Bos, Richert, & Stubbe, 2007; Maaz, Trautwein, Lüdtke, & Baumert, 2008; Schnabel *et al*., 2002). Our set of covariates comprised demographics (sex, age), social background indicators (Highest International Socio-Economic Index of Occupational Status [HISEI], parents' educational degree, ethnic background), student's achievement (elementary school grades and T1 achievement test scores), school career recommendation, preschool/kindergarten time, and parents' educational aspirations.

With regards to findings showing that ability grouping does matter (e.g., Becker, 2009; Ireson & Hallam, 2001; Lucas, 1999; Oakes, 1985), we expected to observe an effect of school track on reading development. With regards to the observation that – on the individual level – different reading skills develop in a different manner in terms of the cumulative versus compensatory model, it was not that clear whether comparable tracking effects would result for both reading skills. As aforementioned, Bast and Reitsma (1998) as well as Parrila *et al*. (2005) found a cumulative development according to more basic reading skills such as word recognition but not according to comprehension. Extrapolating from these results on the individual level, the development of reading skills might also result in varying track differences for each of the two components of reading. Thus, it seemed plausible to carefully expect that tracking would particularly result in different levels of decoding speed, whereas the expected effect of tracking on reading comprehension was rather ambiguous.

Method

Sample

The initial sample comprised $N = 1{,}508$ 5th graders (49% girls; age at T1: $M = 10.88$ years, $SD = .56$) from 60 schools that were drawn so as to be representative of the federal state of Schleswig-Holstein, Germany. The majority of our sample ($n = 940$ students, 62%) attended non-academic track schools comprising 13 lower track (Hauptschule), 22 middle track (Realschule), and 4 comprehensive schools (Gesamtschule). The academic track sample consisted of $n = 568$ high-track students (Gymnasium) from 21 schools. Data were collected by trained research students and took place as group tests carried out in class during regular lessons. In our study, achievement tests and questionnaires were administered at the beginning of 5th grade (T1), at the end of 6th grade (T2), and at the beginning of 8th grade (T3). Hence, T1 took place right after students were assigned to different tracks (up to 4th grade there is no tracking in the German school system). The time intervals between each measurement point were approximately 18 months. Each phase of data collection took place within a time slot of 14 days.

Measures

Reading comprehension

Age-appropriate reading tests from the German PIRLS study (Progress in International Reading Literacy Study, Bos *et al*., 2005) and the German large-scale study 'Aspects of Students' Initial Level and Development at Schools in Hamburg' (Lehmann, Gänsfuß, & Peek, 1999) were used. These achievement tests have been well developed in the context of large-scale studies. The students' task was to read several texts and answer questions on them. The questions focused mainly on students' skills in forming a broad and general understanding of the texts and retrieving information from them. The questions mainly comprised of multiple-choice items, but some open-format questions have also been included. Since some items were scored polytomously, the item parameters were estimated by applying the partial credit model. A common scale for the varying reading tests at the three testing occasions was obtained by means of test linking using an anchor item design (Kolen & Brennan, 2004). Using ConQuest (Wu, Adams, & Wilson, 1998), weighted likelihood estimates (WLEs) were estimated as subjects' ability scores. The WLE reliabilities of the reading tests on all occasions were sufficient ($\geq .80$).

Decoding speed

To measure decoding speed, we used a test in which the students had 2 minutes of time to read a 740-word fairytale containing a great deal of numerals (e.g., 'seven', 'twenty-two') and which has been developed in accordance with the German PISA decoding speed test (Schneider, Schlagmüller, & Ennemoser, 2007). The students' task was to underline these numerals. Thus, this test places a particular emphasis on the mere decoding ability since text comprehension is not needed in recognizing the target words. As the text was too long to finish within the 2-minute time limit, the number of read words marked by the students indicated the speed of decoding (measured as number of read words per 2 minutes). In an additional data collection, a satisfying correlation of $r = .54$ was recorded between this test and the PISA-decoding speed test, which also assesses comprehension with one and the same reading task. The decoding speed test was repeated on each

occasion where data were collected. Referring to Schneider, Schlagmüller *et al*. (2007), tests measuring the speed of reading provide valuable information about basic reading skills such as decoding.

Reasoning

The subtest 'Figure Analogies' from the 'Cognitive Abilities Test for grades 4 to 12' (Heller & Perleth, 2000) was used at T1 to test children's reasoning skills as an indicator of intelligence. WLEs have been estimated as ability scores. The test's WLE reliability was satisfactory (.87).

School grades

The grades from the final report card of elementary school were collected for four school subjects: German as the first language, mathematics, science, and physical education. The German grading system includes grades from 1 (outstanding) to 6 (fail). Thus, lower grades indicate better performance.

Socio-economic status

Several aspects of the families' SES were assessed by means of a parent questionnaire. First, the 'HISEI' (Ganzeboom & Treiman, 1996) was derived from the parents' occupation. This index ranges between 16 and 90 and indicates the socio-economic background that is usually associated with a particular occupational area with higher scores indicating higher incomes. As an overall family indicator, we used the higher value obtained, irrespective of whether it belonged either to the mother or father. Second, we asked the students' parents to detail the highest level of educational achievement they had attained (graduation plus apprenticeship). As per Baumert, Watermann, and Schümer (2003), their responses were coded from 1 (lower track graduation without any apprenticeship) to 7 (university degree).

Ethnic background

We captured students' ethnic background by asking if their mother and father were born in Germany. Ethnic background was dummy coded (0 = *at least one parent was not born in Germany*, 1 = *both parents were born in Germany*).

School career recommendation and parents' educational aspirations

In Germany, teachers recommend an appropriate school track for each individual student at the end of elementary school. This recommendation, however, is not obligatory. Nevertheless, it serves as a strong indicator of the school track that parents chose since many of them follow the teachers' recommendations. Therefore, we asked the parents to report the recommendation given for their children with higher values indicating a higher school track. Parents were also asked to respond to a range of vocational aspirations regarding their children's careers. This range started with 1 (*apprenticeship*) and continued through to 5 (*university studies*).

Preschool/kindergarten time and age at school entrance

Parents were asked if and how long their children had attended preschool or kindergarten. They rated their answers from 0 ('not at all') to 5 ('more than 2 years'). Finally,

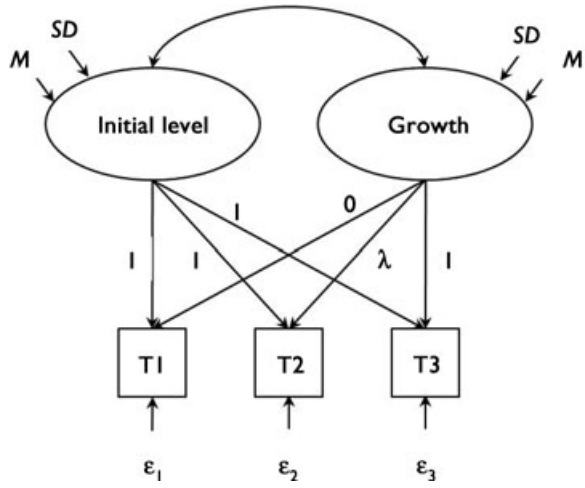


Figure 1. Basic latent growth curve model.

the parents had to indicate their children's age on school entrance on a 4-point scale anchored 1 ('5 years or younger'), 2 ('6 years'), 3 ('7 years'), and 4 ('8 years or older'). Usually, in Germany children enter school at the age of six.

Analytical issues

Missing data

Missing data is a common problem in longitudinal studies. One recommended solution to handle this problem is multiple imputation (e.g., Graham, 2009; Schafer & Graham, 2002). In our present study, on average about 13.5% of the data were missing per variable. We used the STATA implementation (ICE, Royston, 2004) of the MICE program (Van Buuren & Oudshoorn, 1999) to create $m = 5$ complete datasets. All information available has been used to obtain a good imputation model including squares and interaction terms of core variables as these often lead to better solutions and were also used for the estimation of propensity scores (see below). The matching procedure and all subsequent analyses were then conducted five times and the results were combined in accordance with Rubin (1987).

Latent growth curve modelling

When dealing with reading development, an important issue next to the amount of growth is the shape of growth. To analyse whether students at different school tracks differ in this area, we applied LGCMs (e.g., Duncan, Duncan, & Strycker, 2006) in addition to simple mean comparisons of T2 and T3 achievement scores. The idea of LGCM is to describe trajectories over time in terms of initial level (intercept) and growth (slope). In our study, LGCMs were specified by means of structural equation modelling using the software M*plus* 5.2 (Muthén & Muthén, 2008).

In Figure 1, the intercept factor describes the initial level. Likewise, the growth factor reflects the overall change. In this study, positive values for the slope mean indicate increases in reading comprehension and decoding speed. The loadings on the growth factor to T1 and T3 were constrained to 0 and 1, respectively. The loading to T2 (λ) was freely estimated. This enabled the model to capture any shape of non-linear

change. Moreover, λ reflects the proportion of change between T1 and T2 relative to the total change between T1 and T3 (e.g., Bollen & Curran, 2006). As we were interested in group differences due to school track, multi-group LGCMs were applied. This kind of analysis allows for testing of a number of invariance hypotheses. With regard to our research questions, the equalities of shape, initial level, and growth between academic and non-academic track schools were of particular interest. To test the significance of the group differences, the relevant parameters were constrained to be equal between the two groups and chi-square difference tests were conducted. Put concisely, we first tested if the initial level differed between tracks. Next, the differences in the shape of growth (i.e., the free loading λ) were investigated, and, finally, the group difference of the slope factor was tested to obtain the difference in absolute change. To account for the multiple imputed datasets, all analyses were conducted five times with the results being combined at the end. For the model comparisons, we merged the chi-square statistics using Allison's (2001) formula. The result of this procedure is a test statistic that is approximately *F*-distributed.[1]

Hierarchical data structure

For all mean comparisons with the complete (unmatched) sample, we used the M*plus* option 'type = complex' to obtain corrected standard errors for the hierarchical data structure (students in schools). Ignoring such a hierarchical data structure might lead to incorrect standard errors and, thus, to biased significance tests (Hox, 2002).

Propensity score matching

When comparing different treatment conditions (in our case different school tracks), a sufficient condition for analysing treatment effects would be that the treatment assignment is uncorrelated with the outcome (Morgan & Winship, 2007). The most common way to achieve this prerequisite is by randomization. With regard to school track differences, however, it is neither feasible nor ethically sound to randomly assign students to the particular treatment groups. Thus, we cannot decide if a widening achievement gap between students on the different school tracks is a result of the type of school or if it is quite simply the result of previously existing differences among the students.

One prevalent strategy for a statistical adjustment of this confounding issue is the use of analysis of covariance or multiple regression. These strategies, however, underlie several limitations that are difficult to overcome in research practice. First, the number of covariates that can be used is rather limited. Therefore, important pre-existing differences between two groups may be disregarded. Second, the results from these kinds of analyses depend on the pre-specified form (typically linear). Finally, this method does not ensure that two groups are comparable, that is, a lack of covariate overlap remains undetected and, thus, the persons in the two groups are actually too different to compare.

Therefore, the use of strategies such as propensity score matching to estimate causal effects with non-experimental data has been proposed (e.g., Rosenbaum & Rubin, 1983) and has recently begun to enter educational psychology research (e.g., Schneider,

[1] *The numerator degrees of freedom (df) are obtained by the difference of df between the two compared models ($\Delta df = 1$ in all our analyses). The denominator degrees of freedom (ddf) were approximated by applying the according formula including the number of imputations (m), the numerator df, and the square roots of the chi-square statistics over the m datasets (see Allison, 2001 for more details).*

Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007). The general idea of these matching methods is to model the treatment assignment process directly and to create subgroups which match in their likelihood to either belong to treatment or control group – a situation which experimental research achieves through randomization. Propensity score matching includes two fundamental steps: the estimation of the propensity score and the matching procedure.

Based on background information relevant for the group assignment, the probability of group membership (the propensity score) is estimated, typically by applying logistic regression analysis. In this research, the probability of attending academic versus non-academic track schools (i.e., the propensity score) was estimated by means of logistic regression using STATA 10. In accordance with Rosenbaum and Rubin (1985), we defined the logit of the propensity score rather than the probability itself due to better distributional properties.

Treatment and control group individuals were then matched using the obtained propensity score. For this matching procedure, we pursued two strategies. First, we used the presumably most straightforward approach by matching one child on the academic track with one child on the non-academic track via nearest neighbour matching method with caliper. This approach requires a random order of the subjects in the treated (academic) and control (non-academic) group. Then, the first treated subject is selected and the control subject within a tolerance region (caliper) is sought. We chose a caliper size of $C = .05 \times SD$ (propensity score logit). As this approach often involves quite a large sample reduction, we also applied nearest neighbour matching with caliper and replacement as a second approach. In this procedure, control subjects are allowed to be used as a match for several treatment subjects and thus, a larger proportion of the treatment group is preserved. This approach, however, results in different sample sizes for treatments and controls and thus, we used frequency weights for all subsequent analyses.

Both matching strategies were implemented using the STATA module PSMATCH2 (Leuven & Sianesi, 2003). Thereby, we used bootstrapping to obtain more reliable standard errors for all subsequent analyses (e.g., Caliendo & Kopeinig, 2005). This correction is preferable since classical standard errors underestimate the variance due to the error in the estimation of the propensity score and the additional variance induced in the matching process. As the informational status of standard errors is also challenged by the fact that propensity score matching can change sample sizes and lead to progressive significance testing only due to changes in sample size, other indicators of group differences are recommended, which are not sensitive to the sample size, for example, standardized mean differences (Imai, King, & Stuart, 2008).

When applying propensity score matching, ideally, the correlation between treatment assignment and outcome is removed and, therefore, treatment effects can be estimated (for an extensive description, see Morgan & Winship, 2007). Propensity score matching also requires the assumption that selection into treatment is based on observable characteristics, as does regression analyses. Yet, it has been found to be more appropriate regarding the aforementioned limitations and to be more robust against misspecifications (Drake, 1993; Zhao, 2008).

In terms of school track differences, propensity score matching involves perceiving school tracks as treatment conditions (in our study: treatment = academic track; control = non-academic track). The aim of matching is then to eliminate initial differences between students in the two conditions in order to investigate how comparable individuals develop in different conditions. Thus, if matching succeeds, this approach

allows the tentative conclusion that divergent developmental trajectories are rather more a result of features of different school tracks than they are due to previously existing differences. The success of the matching procedure is tested by the check of balance between the two groups (whether matching homogenized the groups successfully) and the inspection of the area of common support (whether comparable individuals are available for the whole sample or only in some parts of the propensity score distributions).

Essential for the success is that the prediction of the treatment assignment is sufficient, that is, that predictors of assignment and outcome can be assumed to be equally distributed between groups and, therefore, assignment can be assumed to be random regarding these variables (Augurzky & Schmidt, 2001; Morgan & Winship, 2007). The predictors of assignment have to be selected by theoretical assumptions and/or results from previous research. With regard to school track assignment background characteristics such as previous achievement or SES have been proven to explain substantial proportions of assignment variance (see above).

Subsequent to the matching procedure, we estimated the treatment effect only for those students who stem from the population that actually attended the treatment (average treatment effect for the treated), but not for the whole population. This approach seemed to be more appropriate because it only requires that students are similar in their expected baseline but can differ in their expected treatment effects and does not require full identification of the treatment effect over the whole population (for further details, see Morgan & Winship, 2007).

With regard to LGCM for the matched samples, we did not expect a difference in the mean initial level because the T1 scores were part of the matching procedure. Therefore, we tested if the initial levels significantly differed between the two groups. If this test resulted in non-significance, as it should due to the matching procedure, the initial levels for both groups were constrained to be equal to obtain a more parsimonious model.

Results

Track differences before matching

Reading comprehension

To estimate track differences for the unmatched sample, we analysed differences in the according parameters of the LGCM: initial level, shape of growth (λ), and amount of growth. As presented in Table 1, there were large differences in the initial level of reading comprehension ($F(1,91) = 242.04$, $p < .001$, $d = 1.29$). The shape of growth did not significantly differ between school tracks ($F(1,82) = 2.07$, ns) and was $\lambda = .64$, meaning that about 65% of the overall growth in reading comprehension took place until T2 indicating that comprehension growth was slightly negatively accelerated. This model with different initial levels and an equal shape of growth fitted the data well for reading comprehension: $F(4,28) = 0.51$, ns, Comparative Fit Index (*CFI*) = 1.00, Root Mean Square Error of Approximation (*RMSEA*) = .007, Standardized Root Mean Square Residual (*SRMR*) = .012. Finally, the mean growth was constrained to be equal across tracks; no significant difference resulted ($F(1,34) = 1.22$, ns; see Table 1).

Decoding speed

The same procedure was then applied for decoding speed (see Table 1). Again, the difference in the initial level was striking ($F(1,402) = 69.82$, $p < .001$, $d = 0.83$).

Table 1. Means, standard errors, effect sizes, and invariance tests for initial level and growth factors of reading comprehension and decoding speed for unmatched sample ($n_{\text{academic}} = 568$; $n_{\text{non-academic}} = 940$)

		Initial level				Growth			
		M	*SE*[a]	*d*	$\Delta\chi^2$	*M*	*SE*[a]	*d*	$\Delta\chi^2$
Reading comprehension	Non-academic	−0.41	0.08	1.29	$F(1,91) = 242.04$ $p < .001$	0.77	0.04	0.04	$F(1,34) = 1.22$ ns
	Academic	0.59	0.05			0.75	0.05		
Decoding speed	Non-academic	276.43	5.23	0.83	$F(1,402) = 69.82$ $p < .001$	143.72	6.43	0.49	$F(1,43) = 6.94$ $p < .05$
	Academic	332.45	6.35			168.18	7.72		

Note. Applying the chi-square difference test ($\Delta\chi^2$), significant differences indicate a worse fit for models with equality constraints. For multiple imputation, the combined chi-square statistic is approximately *F*-distributed (Allison, 2001).[a]Corrected standard errors for nested data.

Moreover, a significant difference in the shape of growth was observed ($F(1,57) = 4.02$, $p < .05$). At non-academic track schools, growth was nearly linear ($\lambda = .54$), whereas about two-thirds of growth have taken place until T2 at academic track schools ($\lambda = .64$). The model with different initial levels and different shapes of growth fitted the data well for decoding speed: $F(3,444) = 1.10$, ns, $CFI = .998$, $RMSEA = .014$, $SRMR = .037$. Finally, the difference in the growth factor yielded significance ($F(1,43) = 6.94$, $p < .05$, $d = 0.49$) indicating a higher growth rate for academic track students than for non-academic track students (see Table 1).

In summary, we found substantial differences between academic and non-academic track school students in the initial levels of decoding speed and reading comprehension. Moreover, there was a difference between tracks in students' development in decoding speed, favouring academic track school students. To strengthen the conclusion that this difference was in fact due to school track, we then applied propensity score matching to account for the non-randomized treatment (school track) assignment. Even though there were no differences between both tracks in the growth factors of reading comprehension, we also tested the differences for reading comprehension for the matched sample to provide a more complete picture of our results.

Propensity score matching

To obtain a good set of background variables for computing the propensity score, we considered a total of 16 variables that are usually associated with the preference or choice of a particular school track (Arnold *et al.*, 2007; Schnabel *et al.*, 2002). This set of variables was comprised of demographics (sex, age), social background indicators (HISEI, parents' level of educational degree, ethnic background), student's achievement (elementary school grades and T1 achievement test scores), school track recommendation, time spent at preschool/kindergarten, and parents' educational aspirations for their children. Moreover, squares of all metric variables and interactions have been added to the logistic regression model to improve balance even on higher order moments. Prior to estimating the propensity scores by logistic regression, we tested the background variables for significant differences between the academic and non-academic track (see Table 2). As can be seen from Table 2, there was an average standardized bias[2] of 77.7%, indicating

[2] *The standardized bias is an indicator that assesses the absolute difference between treatment and control group in sample means divided by an estimate of the pooled standard deviation (Rosenbaum & Rubin, 1985).*

Table 2. Background variable differences between non-academic and academic track before matching ($N = 1{,}508$)

	Non-academic track ($n = 568$)		Academic track ($n = 940$)				
Background variable	*M*	*SE*	*M*	*SE*	*t*	*p*	% bias
School career recommendation	1.75	0.02	2.81	0.02	−30.77	.000	−157.2
Sex (0 = boys)	0.49	0.02	0.50	0.02	−0.46	.648	−2.0
Age at T1	10.99	0.02	10.71	0.02	9.63	.000	44.9
HISEI	44.87	0.49	56.73	0.68	−14.37	.000	−60.9
Parents' highest educational degree	3.71	0.05	5.27	0.07	−18.29	.000	−79.7
Ethnic background	0.81	0.01	0.88	0.01	−3.13	.002	−14.3
Preschool/kindergarten time	4.14	0.05	4.51	0.04	−5.56	.000	−26.5
Age at school entrance	2.24	0.02	2.19	0.02	1.85	.065	8.1
Parents' educational aspirations	2.63	0.05	4.47	0.04	−25.80	.000	−124.2
German grade[a]	3.86	0.03	4.89	0.02	−27.34	.000	−130.3
Mathematics grade[a]	3.93	0.03	4.91	0.03	−22.99	.000	−109.1
Science grade[a]	4.24	0.03	5.13	0.02	−23.37	.000	−110.9
Physical education grade[a]	4.94	0.02	5.19	0.03	−6.74	.000	−29.9
Reasoning	−0.48	0.05	0.61	0.06	−13.42	.000	−58.6
Decoding speed T1	276.20	2.89	332.63	3.68	−12.03	.000	−52.3
Reading comprehension T1	−0.40	0.03	0.58	0.03	−19.79	.000	−93.3
Propensity score (logit)	−4.28	0.10	2.46	0.09	−44.87	.000	−218.1
Total bias (Mean [|% bias|])					77.7		

Note. HISEI, Highest International Socio-Economic Index of Occupational Status.
[a]Grades are recoded (larger scores = better grade).

a large total bias between the two groups. Although this is not the primary criterion in evaluating the success of the propensity model, the logistic regression equation led to a quite good prediction of group membership (Nagelkerke pseudo-R^2 index of .79).

The area of common support is the range where treatment and comparison group overlap. In the case of propensity score matching, this region is defined by the propensity score and indicates for which sub-sample the estimated treatment effects apply. As pictured in Figure 2, the distributions of the propensity score were quite different between academic and non-academic track students. The proportion of academic track students in the area of common support (i.e., within the two vertical lines), however, was relatively high (478 out of 568 or 84%), whereas only about 62% (582 out of 940) from the non-academic track was in this area.

Our matching procedure without replacement resulted in a sample of 139 matched pairs. Table 3 shows that the standardized bias after matching – that should be below 3–5% (e.g., Caliendo & Kopeinig, 2005) – was 2.2%. Applying matching with replacement demanded the additional consideration of Mahalanobis distances for the interactions between reasoning × decoding speed, and school career recommendation × reading comprehension to reach a satisfying balance. This matching approach resulted in a

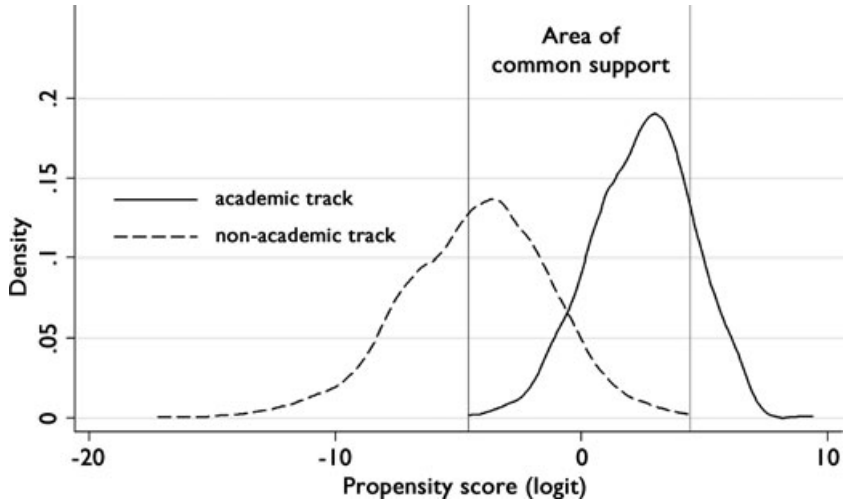


Figure 2. Area of common support.

sample of $n = 297$ academic track students matched to $n = 111$ non-academic track students. For all subsequent analyses, we used frequency weights to account for the fact that the latter ones were repeatedly used as matches. Despite there still being some considerable bias for single covariates (see Table 3), the overall standardized bias was sufficiently reduced (4.4%). When comparing students within the area of common support that were used for matching with those that were not used for matching, the bias between both groups also was low (<5% for matching with and without replacement). Thus, the results of the subsequent analyses may be taken for representative for students within the area of common support.

Track differences after matching[3]

Decoding speed

To test the effect of ability grouping, we first compared the means of decoding speed at T2 and T3. Due to the propensity score procedure we did not control baseline differences for these analyses. By and large, findings did not differ between the two matching procedures. As presented in Table 4, the mean differences for decoding speed at T2 and T3 were significant for both matching procedures resulting in higher scores for academic track students.

We then applied multi-group LGCM. In the first step, we tested differences between the matched samples of both tracks in the initial levels of decoding speed. These differences were not significant for both matching approaches ($p \geq .525$) again indicating the good balance between both groups. As a second step, we tested the differences in the shape of growth (λ in Figure 1) between the two groups. Therefore, we tested an LGCM, in which the loading was constrained to be equal between the treatment group (academic track students) and the control group (non-academic track students), against a model,

[3] *We also tested track differences in reading comprehension for the matched samples. As was true for the full (unmatched) sample, no significant differences in the shape and the amount of growth were observed. Full results are available from the first author on request.*

Table 3. Background variable differences between non-academic and academic track after matching

	Matching without replacement ($n = 139$ pairs)						Matching with replacement ($n = 297$ pairs)[c]					
	Non-academic		Academic				Non-academic		Academic			
Background variable	*M*	*SE*[b]	*M*	*SE*[b]	*t*	%	*M*	*SE*[b]	*M*	*SE*[b]	*t*	%
School career recommendation	2.44	0.06	2.45	0.04	−0.07	−0.7	2.67	0.04	2.70	0.03	−0.64	−4.5
Sex (0 = boys)	0.54	0.04	0.53	0.04	0.25	2.4	0.54	0.03	0.55	0.03	−0.19	−1.2
Age at T1	10.74	0.04	10.75	0.04	−0.18	−1.8	10.73	0.02	10.69	0.02	1.08	7.0
HISEI	52.77	1.29	53.12	1.25	−0.20	−1.9	55.04	0.86	55.19	0.92	−0.12	−0.8
Parents' highest educational degree	4.90	0.14	4.83	0.14	0.31	3.1	5.06	0.10	4.92	0.09	1.03	6.9
Ethnic background	0.85	0.03	0.83	0.03	0.46	4.5	0.87	0.02	0.89	0.02	−0.66	−4.4
Preschool/kindergarten time	4.38	0.09	4.38	0.10	−0.02	−0.2	4.51	0.07	4.41	0.06	1.08	7.4
Age at school entrance	2.18	0.04	2.17	0.04	0.08	0.8	2.18	0.03	2.17	0.03	0.25	1.7
Parents' educational aspirations	3.92	0.10	3.87	0.12	0.30	2.8	4.26	0.06	4.17	0.07	1.07	7.0
German grade[a]	4.52	0.05	4.52	0.05	0.03	0.3	2.26	0.03	2.30	0.03	−0.84	−5.6
Mathematics grade[a]	4.52	0.06	4.49	0.06	0.26	2.5	2.28	0.04	2.26	0.04	0.33	2.2
Science grade[a]	4.89	0.05	4.85	0.05	0.54	5.2	1.96	0.04	1.87	0.03	1.86	12.7
Physical education grade[a]	5.12	0.06	5.12	0.06	−0.01	−0.1	1.82	0.04	1.81	0.04	0.15	1.0
Reasoning	0.22	0.13	0.19	0.13	0.16	1.6	0.66	0.08	0.66	0.09	−0.01	−0.0
Decoding speed T1	300.17	7.17	297.67	6.97	0.25	2.5	306.01	4.01	304.26	4.16	0.30	2.0
Reading comprehension T1	0.20	0.07	0.17	0.07	0.23	2.3	0.48	0.04	0.46	0.04	0.48	3.2
Propensity score (logit)	0.02	0.14	−0.05	0.14	0.40	4.0	1.38	0.10	1.22	0.10	1.07	7.2
Total bias (Mean [|%bias|])						2.2						4.4

Note. HISEI, Highest International Socio-Economic Index of Occupational Status.
[a]Grades are recoded (larger scores = better grade).
[b]Bootstrapped standard errors.
[c]Frequency weights accounted for subjects serving as repeated matches.

Table 4. Means, standard errors, and effect sizes for the treatment effects for decoding speed for the matched samples

		Non-academic track		Academic track			
Matching		*M*	*SE*[b]	*M*	*SE*[b]	*t*	*d*
Without replacement	T2	376.70	8.21	407.50	7.52	−2.77**	0.33
(*n* = 139 pairs)	T3	441.05	7.98	463.63	8.15	−1.98*	0.24
With replacement	T2	397.23	5.43	415.10	5.41	−2.33*	0.19
(*n* = 297 pairs)[a]	T3	461.27	6.64	479.08	5.61	−2.05*	0.17

Note. [a]Frequency weights accounted for subjects serving as repeated matches.
[b]Bootstrapped standard errors.
*$p < .05$; **$p < .01$ (two-tailed).

Table 5. Means, standard errors, effect sizes, and invariance tests for the growth factors of decoding speed for the matched samples

Matching		*M*	*SE*[b]	*d*	$\Delta\chi^2$
Without replacement	Non-academic	144.58	7.55	0.40	$F(1,538) = 3.90$
(*n* = 139 pairs)	Academic	161.61	7.23		$p < .05$
With replacement	Non-academic	156.92	5.54	0.22	$F(1,51) = 4.09$
(*n* = 297 pairs)[a]	Academic	170.59	4.96		$p < .05$

Note. Applying the chi-square difference test ($\Delta\chi^2$), significant differences indicate a worse fit for models with equality constraints. For multiple imputation, the combined chi-square statistic is approximately *F*-distributed (Allison, 2001).
[a]Frequency weights accounted for subjects serving as repeated matches.
[b]Bootstrapped standard errors.

in which this loading was allowed to vary between both groups. For both matching approaches, this test yielded significance (matching without replacement: $F(1,194) = 6.16$, $p < .05$; matching with replacement: $F(1,27) = 4.31$, $p < .05$) indicating that the shape of growth was different between both groups. Students at academic track schools showed a higher growth rate until T2 ($\lambda = .62$) compared to students at non-academic tracks ($\lambda = .54$). The same was true for matching without replacement ($\lambda_{\text{academic track}} = .68$; $\lambda_{\text{non-academic track}} = .53$). Thus, for the comparison of the slope means, we started with a model with equal initial levels for academic and non-academic track students but with different shapes of growth. The overall fit for the model was good (matching without replacement: $F(4,59) = 0.45$, ns, *CFI* = .999, *RMSEA* = .009, *SRMR* = .054; matching with replacement: $F(4,16) = 1.80$, ns, *CFI* = .987, *RMSEA* = .077, *SRMR* = .065). We then constrained the slope means to be equal across the two groups resulting in a significant decrease of model fit (see Table 5). Thus, larger growth rates for students at the academic track than for students at the non-academic track were recorded.

Discussion

A core feature of Germany's secondary school system is explicit between-school tracking. The idea of tracking is that it should help to provide an educational setting which is appropriate for fostering students with regard to their individual skill level. However, there is a growing body of literature that questions such positive effects of tracking. As reading skills are an important prerequisite not only for success in an academic context but also in daily life, tracking effects on reading development are of particular interest to politicians, practitioners, and scientists. In this study, we used multi-group LGCMs to compare the developmental trajectories of reading comprehension and decoding speed between students on academic and non-academic tracks during the first 3 years of secondary school. To strengthen the conclusion that different developments are actually a result of school track, we applied propensity score matching. Drawing on differences between students on different tracks in mathematics achievement (e.g., Argys *et al*., 1996; Becker *et al*., 2006), we expected that students at academic track schools attain larger reading growth rates than students at non-academic track schools. However, according to the ambiguous results for studies on a cumulative versus compensatory individual development of reading, it was not obvious if tracking effects would have similar results for different reading skills. With regard to findings from earlier studies (Bast & Reitsma, 1998; Parrila *et al*., 2005), it seemed plausible to expect larger track differences for decoding speed than for reading comprehension.

The results for the analyses before and after matching were quite similar as regards to the growth factors: no significant track differences between the slope means for reading comprehension and a significantly higher slope mean for academic track students' decoding speed than for their non-academic counterparts. Regarding the initial level, however, the findings for the complete sample were naturally quite different from the results for the matched sample. At the beginning of secondary school, academic track students achieve largely higher levels of reading comprehension as well as decoding speed than students in non-academic track schools. These initial level differences disappeared when propensity score matching was applied, indicating a good balance between both tracks. An additional finding of our study indicated a slowing down in reading growth. Students in our sample gained more than two-thirds of overall reading comprehension growth during the first 18 months of our study (until T2). With regard to decoding speed, this was only true for students on academic tracks while non-academic track students' development was nearly linear. It seems as if a student's full learning potential for reading comprehension – and similarly so for decoding speed in academic track schools – might be released at some point during their secondary school years. The slowing down in development indicates that students in our study may approach this point. Non-academic track students, however, still seem to have some more room to improve their decoding skills.

Our hypothesis proposing benefits for academic track school students was supported with regard to decoding speed. Academic track students' larger growth rate might depend on various facts. First, teachers in different school tracks vary in their use of instructional practices. In academic track schools, for example, critical thinking, cognitive activating tasks and problem solving are emphasized, whereas in non-academic track schools, exercises and repetition are stressed (Kunter & Baumert, 2006; Raudenbush *et al*., 1993; Retelsdorf *et al*., 2010; Van Houtte, 2004). Moreover, at the expense of academic goals controlling students' behaviour is a considerable task for teachers at non-academic track schools. Altogether, teachers in higher track schools

seem to promote learning to a larger degree than teachers in lower track schools do. Basic reading skills such as decoding abilities, however, are not the focus of secondary school teachers – within neither the academic track nor the non-academic track. Thus, other explanations such as classroom or school composition might be more compelling. This idea involves the assumption that students' achievement development depends on aggregated student characteristics per school (Hattie, 2002; Thrupp, Lauder, & Robinson, 2002). Students at schools with, for example, high mean achievement levels, high mean SES, and low ethnic heterogeneity ought to benefit compared to students at schools with an adverse mixture of characteristics. This might be due to achievement-related norms and values that develop subject to the student composition. A third explanation concerns routine. Decoding speed involves high automaticity, which in turn requires great amounts of practice. Academic track students' decoding speed might grow faster because the overall preoccupation with written texts is, in general, more pronounced in academic track schools (across all school subjects) and thus, decoding might automatize to a higher extent at the academic track. As students were tested using the same test several times, we cannot rule out training effects. The time span between the three measurement occasions, however, was quite long (18 months). Moreover, even though it might be a training effect, this effect was more pronounced at academic track schools than at non-academic track schools again indicating a beneficial learning environment.

With regard to the development of reading comprehension, we did not find significant differences between academic and non-academic track students. Thus, the assumption Retelsdorf and Möller (2008) made was not supported. These authors speculated that the achievement gap between different school tracks might widen over time, since they found some diverging trend that, however, failed to show statistical significance. The absence of a tracking effect for reading comprehension might be due to the levels of comprehension our test required. As the students' task was to form a broad and general understanding of the texts and to retrieve particular information from the texts, we used tasks that rather required surface or propositional representation for comprehension (cf. Kintsch, 1998). Maybe, applying tests that demand higher level processes from students such as inferential skills or integration into background knowledge, increased benefits for academic track students might also emerge for reading comprehension.

Our tentative assumption that tracking effects might be larger for decoding speed than for reading comprehension was supported. This result fits to the equivocal findings of research investigating the development of individual differences in reading. For example, drawing on a Finnish sample, Parrila *et al*. (2005) found that comprehension developed rather compensatory, whereas a Matthew effect was observed for decoding speed. The ambiguous findings in our study might be due to the particular tests we applied. As previously stated, the reading comprehension test required relatively low levels of comprehension. These functional reading tasks are practiced by everyone in daily life so that students at different school tracks might develop very similarly. By contrast, decoding speed needs high automaticity – in particular due to the speed test character – this is only reached by extensive practice (Lundberg, 2002). As mentioned above, basic reading skills such as decoding speed, however, are neither the focus of academic nor non-academic track teachers as this skill is rather an educational objective of elementary school. However, in academic track schools, students are generally exposed to larger amounts of text throughout the spectrum of school subjects than their non-academic track counterparts. Thus, it seems plausible that basic reading skills might improve faster at academic track schools than at non-academic track schools due to the increased amount of practice.

Regardless of the explanation of our results, there are certain concerns that must be taken into account when extrapolating our findings. First, we only estimated the average treatment effect for the treated, and the results were only representative for the proportion of students within the area of common support. Within this area, however, the differences between students included in the matched samples and those not included were small. The amount of academic track students in this area was quite satisfying, though only slightly more than half of the non-academic track students lay in this area. However, this could even strengthen the conclusion that school track does matter since even positively selected non-academic students did not achieve equal growth in decoding speed when viewed against comparable students at higher track schools. Second, when extrapolating our results one should keep in mind that, in general, students' achievement level at non-academic tracks is well below that of the level at academic tracks as it becomes apparent when one views the results for the unmatched (complete) sample. As aforementioned, according to the growth factors, these analyses yielded similar results as for the analyses for the matched sample did.

Compared with the widening achievement gap in the mathematics domain (e.g., Becker *et al*., 2006), the differences in reading development in our study overall were somewhat smaller. This might be due to the very different role these competencies play in the curriculum of secondary schools. While there is no explicit curriculum for reading in Germany's secondary schools, the acquisition of mathematical competencies depends much more on the curriculum and, thus, school's learning environment. Even though tracking effects for reading seem to be smaller than in other domains, and even though our results are somewhat equivocal, this study contributes to research that questions positive effects of tracking. Indeed, the developmental trajectories for reading comprehension are comparable within academic and non-academic track schools. However, one should keep in mind that there is still a considerable difference between students at both tracks at the entrance of secondary school, which remains stable over time (see results for the unmatched samples). In fact, students in non-academic track schools hardly reach the level of reading comprehension at grade eight that their counterparts in academic track schools have already reached at the beginning of grade 5. Moreover, for decoding speed, it turned out that even for students with a very similar composition of characteristics at the entry to secondary school – in terms of prior achievement and socio-demographic background – academic track school students gain more in decoding speed. This result is in line with previous studies that confirmed that students achieve better in higher ability groups than in lower ability groups (e.g., Fuligni, Eccles, & Barber, 1995; Hoffer, 1992; Pallas *et al*., 1994; van Houtte, 2004; Wiliam & Bartholomew, 2004). Our results, however, allow only very cautious interpretations of tracking effects, since we only investigated the domain of reading, whereas previous research has shown that effects of ability grouping are not necessarily universal (cf. Baumert, Becker, Neumann, & Nikolova, 2010). Thus, achievement data from different domains and additional non-achievement-related outcome measures should be taken into account in future studies investigating tracking effects. Moreover, propensity score matching provides a promising approach for the comparison of similar students in different treatments. However, our analyses do not allow comparing extremely high- or low-performing students. Accordingly, the question if these students particularly benefit from tracking remains unanswered. Finally, with our current study we cannot decide whether these differences go back to instructional, compositional, or institutional causes. All in all, we may, nevertheless, conclude that – within particular domains – tracking does have undesired side effects, which may cumulate during adolescents'

further development and affect their educational outcomes, and thus will increase the spread of achievement within one age cohort. Such track-specific developmental trajectories become particularly problematic, when students at a particular track will not reach minimum standards of achievement anymore.

Limitations and strengths

One benefit of using propensity score matching is the enhancement of causal inferences (e.g., West & Thoemmes, 2008). This analytical approach offers vast opportunities for the investigation of schooling effects. In practice, related research questions are always limited to non-experimental data since we cannot randomly assign students to particular schools or school tracks. Propensity score matching allows researchers to take into account the assignment procedure and offers a way to directly address the issue of selection bias for the estimation of treatment (i.e., school) effects with observational data. In this study, our matching procedure has reduced the total selection bias comprehensively ($bias_{before\ matching} = 78\%$ vs. $bias_{after\ matching} = 2\%$ and accordingly 4%). Additionally, gaining similar results for track differences in the growth of both components of reading without matching as well as applying two different propensity score matching approaches, gives our findings a certain robustness. It is still conceivable, however, that the differences in decoding speed growth between academic and non-academic track students depend on selection bias of unobservable background variables (cf. Winship & Morgan, 1999). This is a common problem of methods relying on the control of observed characteristics as regression or propensity score matching, which we cannot overcome with our data. In order to rule this problem out, studies investigating the effects of tracking should draw on methods controlling also for non-observables, for example, on longitudinal samples including several phases of data collection before track assignment and several thereafter (cf. Becker, 2009).

Moreover, our study did not investigate the particular characteristics of tracking that are responsible for the divergent trajectories of decoding speed. In addition to institutional aspects such as curriculum differences or instructional styles another explanation for tracking effects is classroom or school composition. Despite quite a long history of research on such compositional effects of about 30 years, there is still no consensus on the size of unique and shared variance of compositional and institutional differences (see above). Thus, future studies should account for these desiderata.

A final concern involves the number of phases of data collection. With only three occasions of measurement, the number of alternative growth shapes that can be investigated is quite limited. As the trajectories for decoding speed bend down at T2 to a higher extent for the academic track students than for the non-academic students, whereas the difference at T3 continues to remain significant, we cannot conclude if students at non-academic track schools might catch up later. A related limitation is the time span between the particular waves of data collection. The fact that there were intervals of approximately 18 months between each phase of data collection also meant that we were not able to analyse short-time effects.

In summation, we found a clear effect of school track on the development of decoding speed, whereas reading comprehension growth seems not to be directly related to school track. Thus, our results suggest that school track does make a difference, but not for every particular reading skill. Despite some limitations of our study, our study suggests that propensity score matching might be a viable way for studying school track differences when self-selection into a treatment is present.

Acknowledgements

The research reported in this paper is part of the project 'Self-concept, Motivation, and Literacy: Development of Student Reading Behavior' directed by Jens Möller (Christian-Albrechts-University of Kiel). The project was funded by the German Research Foundation (DFG; Mo 648/15–1/15–3). Furthermore, we would like to thank Declan Donaghey for language editing.

References

Aarnoutse, C., Leeuwe, J. V., Voeten, M., & Oud, H. (2001). Development of decoding, reading comprehension, vocabulary and spelling during the elementary school years. *Reading and Writing*, *14*, 61–89. doi:10.1023/A:1008128417862

Allison, P. D. (2001). *Missing data*. Thousand Oaks, CA: Sage Publications, Inc.

Argys, L. M., Rees, D. I., & Brewer, D. J. (1996). Detracking America's schools: Equity at zero cost? *Journal of Policy Analysis and Management*, *15*, 623–645. doi:10.1002/(SICI)1520-6688(199623)15:4<623::AID-PAM7>3.0.CO;2-J

Arnold, K.-H., Bos, W., Richert, P., & Stubbe, T. C. (2007). Schullaufbahnpräfenzen am Ende der vierten Klassenstufe [School type preferences at the end of the 4th grade]. In W. Bos, S. Hornberg, K.-H. Arnold, G. Faust, L. Fried, E.-M. Lankes, . . . R. Valtin (Eds.), *IGLU 2006. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (pp. 271–297). Münster, Germany: Waxmann.

Augurzky, B., & Schmidt, C. M. (2001). The propensity score: A means to an end. *IZA Discussion Paper*, *271*. Retrieved from: http://ftp.iza.org/dp271.pdf

Aunola, K., Leskinen, E., Onatsu-Arvilommi, T., & Nurmi, J.-E. (2002). Three methods for studying developmental change: A case of reading skills and self-concept. *British Journal of Educational Psychology*, *72*, 343–364. doi:10.1348/000709902320634447

Barth, J. M., Dunlap, S. T., Dane, H., Lochman, J. E., & Wells, K. C. (2004). Classroom environment influences on aggression, peer relations, and academic focus. *Journal of School Psychology*, *42*, 115–133. doi:10.1016/j.jsp.2003.11.004

Bast, J., & Reitsma, P. (1998). Analyzing the development of individual differences in terms of Matthew effects in reading: Results from a Dutch longitudinal study. *Developmental Psychology*, *34*, 1373–1399. doi:10.1037/0012-1649.34.6.1373

Baumert, J., Becker, M., Neumann, M., & Nikolova, R. (2010). Besondere Förderung von Kernkompetenzen an Spezialgymnasien? Der Frühübergang in grundständige Gymnasien in Berlin [Do academic tracks with specific curricular profiles accelerate the development of achievement in reading, mathematics, and English literacy? Early transition to the academic track of secondary schooling in Berlin]. *Zeitschrift für Pädagogische Psychologie*, *24*, 5–22. doi:10.1024/1010-0652/a000001

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., . . . Tsai, Y.-M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, *47*, 133–180. doi:10.3102/0002831209345157

Baumert, J., Nagy, G., & Lehmann, R. H. (in press). Cumulative advantages and the emergence of social and ethnic inequality: Matthew effects in reading and mathematics development within elementary schools? *Child Development*.

Baumert, J., Watermann, R., & Schümer, G. (2003). Disparitäten der Bildungsbeteiligung und des Kompetenzerwerbs. Ein institutionelles und individuelles Mediationsmodell [Disparities in educational participation and attainment: An institutional and individual mediation model]. *Zeitschrift für Erziehungswissenschaft*, *6*, 46–72. doi:10.1007/s11618-003-0004-7

Becker, M. (2009). *Kognitive Leistungsentwicklung in differenziellen Lernumwelten: Effekte des gegliederten Sekundarschulsystems in Deutschland* [*Cognitive development in differential learning environments: Effects of the tracked secondary school system in Germany*]. Berlin: Max-Planck-Institut für Bildungsforschung.

Becker, M., Lüdtke, O., Trautwein, U., & Baumert, J. (2006). Leistungszuwachs in Mathematik. Evidenz für einen Scherenеffekt im mehrgliedrigen Schulsystem? [Achievement gains in mathematics: Evidence for differential achievement trajectories in a tracked school system?]. *Zeitschrift für Pädagogische Psychologie*, *20*, 233–242. doi:10.1024/1010-0652.20.4.233

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models. A structural equation perspective*. Hoboken, NJ: John Wiley & Sons, Inc. doi:10.1002/0471746096

Bos, W., Lankes, E.-M., Prenzel, M., Schwippert, K., Valtin, R., Voss, A., & Walther, G. (Eds.). (2005). *IGLU. Skalenhandbuch zur Dokumentation der Erhebungsinstrumente* [*Scale handbook of the German PIRLS study*]. Münster, Germany: Waxmann.

Caliendo, M., & Kopeinig, S. (2005). Some practical guidance for the implementation of propensity score matching. *IZA Discussion Paper*, *1588*. Retrieved from http://ftp.iza.org/dp1588.pdf

Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology*, *33*, 934–945. doi:10.1037/0012-1649.33.6.934

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, *49*, 1231–1236.

Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2006). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fuligni, A. J., Eccles, J. S., & Barber, B. L. (1995). The long-term effects of seventh-grade ability grouping in mathematics. *Journal of Early Adolescence*, *15*, 58–89. doi:10.1177/0272431695015001005

Ganzeboom, H. B. G., & Treiman, D. J. (1996). Internationally comparable measures of occupational status for the 1988 international standard classification of occupations. *Social Science Research*, *25*, 201–239. doi:10.1006/ssre.1996.0010

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549–576. doi:10.1146/annurev.psych.58.110405.085530

Hallam, S., & Ireson, J. (2003). Secondary school teachers' attitudes towards and beliefs about ability grouping. *British Journal of Educational Psychology*, *73*, 343–356. doi:10.1348/000709903322275876

Hattie, J. A. C. (2002). Classroom composition and peer effects. *International Journal of Educational Research*, *37*, 449–481. doi:10.1016/S0883-0355(03)00015-6

Heller, K. A., & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision* [*Cognitive Abilities Test (CogAT; Thorndike, L. & Hagen, E., 1954–1986) – German adapted version/author*]. Göttingen: Beltz.

Hoffer, T. B. (1992). Middle school ability grouping and student achievement in science and mathematics. *Educational Evaluation and Policy Analysis*, *14*, 205–227. doi:10.3102/01623737014003205

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings among experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *171*, 481–502. doi:10.1111/j.1467-985X.2007.00527.x

Ireson, J., & Hallam, S. (2001). *Ability grouping in education*. London: Chapman.

Ireson, J., Hallam, S., & Plewis, I. (2001). Ability grouping in secondary schools: Effects on pupils' self-concepts. *British Journal of Educational Psychology*, *71*, 315–326. doi:10.1348/000709901158541

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, UK: Cambridge University Press.

Kirsch, I. S., de Jong, J., Lafontaine, D., McQueen, J., Mendelovits, J., & Monseur, C. (2002). *Reading for change: Performance and engagement across countries. Results from PISA 2000*. Paris: OECD Publishing.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. Berlin, Germany: Springer.

Köller, O., & Baumert, J. (2001). Leistungsgruppierungen in der Sekundarstufe I – Ihre Konsequenzen für die Mathematikleistung und das mathematische Selbstkonzept der Begabung [Ability grouping at secondary level 1. Consequences for mathematics achievement and the self-concept of mathematical ability]. *Zeitschrift für Pädagogische Psychologie*, *15*, 99–110. doi:10.1024//1010-0652.15.2.99

Kulik, J. A., & Kulik, C.-l. C. (1992). Meta-analytic findings on grouping programs. *Gifted Child Quarterly*, *36*, 73–77. doi:10.1177/001698629203600204

Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research*, *9*, 231–251. doi:10.1007/s10984-006-9015-7

Lehmann, R. H., Gänsfuß, R., & Peek, R. (1999). *Aspekte der Lernausgangslage und der Lernentwicklung von Schülerinnen und Schülern an Hamburger Schulen – Klassenstufe 7. Bericht über die Untersuchung im September 1998* [*Aspects of students' initial level and development at schools in Hamburg – grade 7. Report on the study in September 1998*]. Hamburg, Germany: Landesinstitut für Lehrerbildung und Schulentwicklung.

LeTendre, G. K., Hofer, B. K., & Shimizu, H. (2003). What is tracking? Cultural expectations in the United States, Germany, and Japan. *American Educational Research Journal*, *40*, 43–89. doi:10.3102/00028312040001043

Leuven, E., & Sianesi, B. (2003). PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing (Version 3.0.0). Retrieved from http://ideas.repec.org/c/boc/bocode/s432001.html

Liu, W. C., Wang, C. K. J., & Parkins, E. J. (2005). A longitudinal study of students' academic self-concept in a streamed setting: The singapore context. *British Journal of Educational Psychology*, *75*, 567–586. doi:10.1348/000709905×42239

Lucas, S. R. (1999). *Tracking inequality. Stratification and mobility in American high schools*. New York: Teachers College Press.

Lundberg, I. (2002). The child's route into reading and what can go wrong. *Dyslexia*, *8*, 1–13. doi:10.1002/dys.204

Maaz, K., Trautwein, U., Lüdtke, O., & Baumert, J. (2008). Educational transitions and differential learning environments: How explicit between-school tracking contributes to social inequality in educational outcomes. *Child Development Perspectives*, *2*, 99–106. doi:10.1111/j.1750-8606.2008.00048.x

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology*, *79*, 280–295. doi:10.1037/0022-0663.79.3.280

Maughan, B., & Rutter, M. (1987). Pupils' progress in selective and nonselective schools. *School Leadership & Management*, *7*, 50–68. doi:10.1080/0260136870070110

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge: Cambridge University Press.

Mulkey, L. M., Catsambis, S., Steelman, L. C., & Crain, R. L. (2005). The long-term effects of ability grouping in mathematics: A national investigation. *Social Psychology of Education*, *8*, 137–177. doi:10.1007/s11218-005-4014-6

Muthén, L. K., & Muthén, B. O. (2008). *Mplus* (Version 5.2) [Computer software]. Los Angeles, CA: Muthén & Muthén.

Oakes, J. (1985). *Keeping track: How schools structure inequality*. New Haven, CT: Yale University Press.

Oakes, J. (1987). Tracking in secondary schools: A contextual perspective. *Educational Psychologist*, *22*, 129–153. doi:10.1207/s15326985ep2202_3

Pallas, A. M., Entwisle, D. R., Alexander, K. L., & Stluka, M. F. (1994). Ability-group effects: Instructional, social, or institutional? *Sociology of Education*, *67*, 27–46. doi:10.2307/2112748

Parrila, R., Aunola, K., Leskinen, E., Nurmi, J.-E., & Kirby, J. R. (2005). Development of individual differences in reading: Results from longitudinal studies in English and Finnish. *Journal of Educational Psychology*, *97*, 299–319. doi:10.1037/0022–0663.97.3.299

Raudenbush, S. W., Rowan, B., & Cheong, Y. F. (1993). Higher order instructional goals in secondary schools: Class, teacher, and school influences. *American Educational Research Journal*, *21*, 523–553. doi:10.3102/00028312030003523

Retelsdorf, J., Butler, R., Streblow, L., & Schiefele, U. (2010). Teachers' goal orientations for teaching: Associations with instructional practices, interest in teaching, and burnout. *Learning and Instruction*, *20*, 30–46. doi:10.1016/j.learninstruc.2009.01.001

Retelsdorf, J., & Möller, J. (2008). Entwicklungen von Lesekompetenz und Lesemotivation: Schereneffekte in der Sekundarstufe? [Developments of reading literacy and reading motivation: Achievement gaps in secondary school?]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *40*, 179–188. doi:10.1026/0049–8637.40.4.179

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*, 33–38. doi:10.2307/2683903

Royston, P. (2004). Multiple imputation of missing values. *Stata Journal*, *4*, 227–241.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: J. Wiley & Sons.

Salmela-Aro, K., Kiuru, N., & Nurmi, J.-E. (2008). The role of educational track in adolescents' school burnout: A longitudinal study. *British Journal of Educational Psychology*, *78*, 663–689. doi:10.1348/000709908×281628

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177. doi:10.1037//1082–989X.7.2.147

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*, *13*, 279–313. doi:10.1037/a0014268

Schnabel, K. U., Alfeld, C., Patterson, F. D., Eccles, J. S., Köller, O., & Baumert, J. (2002). Parental influence on students' educational choices in the United States and Germany: Different ramifications – same effect? *Journal of Vocational Behavior*, *60*, 178–198. doi:10.1006/jvbe.2001.1863

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects using experimental and observational designs: A think tank white paper*. Washington, DC: American Educational Research Association.

Schneider, W., Schlagmüller, M., & Ennemoser, M. (2007). *LGVT 6–12. Lesegeschwindigkeits- und -verständnistest für die Klassen 6–12* [*Reading speed and comprehension test for grades 6 to 12*]. Göttingen: Hofgrefe.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*, 360–406. doi:10.1598/RRQ.21.4.1

Thrupp, M., Lauder, H., & Robinson, T. (2002). School composition and peer effects. *International Journal of Educational Research*, *37*, 483–504. doi:10.1016/S0883–0355(03)00016–8

Trautwein, U., Lüdtke, O., Marsh, H. W., Köller, O., & Baumert, J. (2006). Tracking, grading, and student motivation: Using group composition and status to predict self-concept and interest in ninth-grade mathematics. *Journal of Educational Psychology*, *98*, 788–806. doi:10.1037/0022–0663.98.4.788

Van Buuren, S., & Oudshoorn, K. (1999). *Flexible imputation by MICE. Report TNO-PG 99.054*. Retrieved from http://www.multiple-imputation.com

Van de Gaer, E., Pustjens, H., Van Damme, J., & De Munter, A. (2006). Tracking and the effects of school-related attitudes on the language achievement of boys and girls. *British Journal of Sociology of Education*, *27*, 293–309. doi:10.1080/01425690600750478

Van Houtte, M. (2004). Tracking effects on school achievement: A quantitative explanation in terms of the academic culture of school staff. *American Journal of Education*, *110*, 354–388. doi:10.1086/422790

West, S. G., & Thoemmes, F. (2008). Equating groups. In P. Alasuutari, L. Bickman, & J. Brannen (Eds.), *Handbook of social research methods* (pp. 414–430). London: Sage.

Wiliam, D., & Bartholomew, H. (2004). It's not which school but which set you're in that matters: The influence of ability grouping practices on student progress in mathematics. *British Educational Research Journal*, *30*, 279–293. doi:10.1080/0141192042000195245

Winship, C., & Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology*, *25*, 659–706. doi:10.1146/annurev.soc.25.1.659

Wu, M. L., Adams, R. J., & Wilson, M. (1998). *ConQuest: Generalized item response modeling software* [Computer Software]. Melbourne: Australian Council for Educational Research.

Zhao, Z. (2008). Sensitivity of propensity score methods to the specifications. *Economics Letters*, *98*, 309–319. doi:10.1016/j.econlet.2007.05.010

Received 11 August 2010; revised version received 9 August 2011

Emulating a Novel Clinical Trial Using Existing Observational Data

Predicting Results of the PreVent Study

Andrew J. Admon[1,2], John P. Donnelly[2,3,4], Jonathan D. Casey[5], David R. Janz[6], Derek W. Russell[7], Aaron M. Joffe[8], Derek J. Vonderhaar[9,10], Kevin M. Dischert[9], Susan B. Stempek[11], James M. Dargin[11], Todd W. Rice[5], Theodore J. Iwashyna[1,2,4,12‡], and Matthew W. Semler[5]; on behalf of the Pragmatic Critical Care Research Group*

[1]Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, [2]Institute for Healthcare Policy and Innovation, [3]Department of Learning Health Sciences, and [12]Institute for Social Research, University of Michigan, Ann Arbor, Michigan; [4]Veterans Affairs Center for Clinical Management Research, Health Services Research and Development Center of Innovation, Ann Arbor, Michigan; [5]Division of Pulmonary, Allergy, and Critical Care Medicine, Vanderbilt University Medical Center, Nashville, Tennessee; [6]Section of Pulmonary/Critical Care & Allergy/Immunology, Louisiana State University School of Medicine, New Orleans, Louisiana; [7]Division of Pulmonary, Allergy, & Critical Care Medicine, University of Alabama at Birmingham, Birmingham, Alabama; [8]Department of Anesthesiology and Pain Medicine, University of Washington, Seattle, Washington; [9]Department of Pulmonary and Critical Care Medicine, Ochsner Health System New Orleans, New Orleans, Louisiana; [10]Department of Medicine, Section of Emergency Medicine, Louisiana State University School of Medicine–New Orleans, New Orleans, Louisiana; and [11]Department of Medicine, Division of Pulmonary and Critical Care Medicine, Lahey Hospital and Medical Center, Burlington, Massachusetts

ORCID ID: 0000-0002-7432-3764 (A.J.A.).

Abstract

Rationale: "Target trial emulation" has been proposed as an observational method to answer comparative effectiveness questions, but it has rarely been attempted concurrently with a randomized clinical trial (RCT).

Objectives: We tested the hypothesis that blinded analysts applying target trial emulation to existing observational data could predict the results of an RCT.

Methods: PreVent (Preventing Hypoxemia with Manual Ventilation during Endotracheal Intubation) was a multicenter RCT examining the effects of positive-pressure ventilation during tracheal intubation on oxygen saturation and severe hypoxemia. Analysts unaware of PreVent's results used patient-level data from three previous trials evaluating airway management interventions to emulate PreVent's eligibility criteria, randomization procedure, and statistical analysis. After PreVent's release, results of this blinded observational analysis were compared with those of the RCT. Difference-in-differences estimates for comparison of treatment effects between the observational analysis and the PreVent trial are reported on the absolute scale.

Results: Using observational data, we were able to emulate PreVent's randomization procedure to produce balanced groups for comparison. The lowest oxygen saturation during intubation was higher in the positive-pressure ventilation group than the no positive-pressure ventilation group in the observational analysis ($n = 360$; mean difference = 1.8%; 95% confidence interval [CI] = −1.0 to 4.6) and in the PreVent trial ($n = 401$; mean difference = 3.9%; 95% CI = 1.4 to 6.4), though the observational analysis could not exclude no difference. Difference-in-differences estimates comparing treatment effects showed reasonable agreement for lowest oxygen saturation between the observational analysis and the PreVent trial (mean difference = −2.1%; 95% CI = −5.9 to 1.7). Positive-pressure ventilation resulted in lower rates of severe hypoxemia in both the observational analysis (risk ratio = 0.60; 95% CI = 0.38 to 0.93) and in the PreVent trial (risk ratio = 0.48; 95% CI = 0.30 to 0.77). The absolute reduction in the incidence of severe hypoxemia with positive-pressure ventilation was similar in the observational analysis (9.4%) and the PreVent trial (12.0%), though the difference between these estimates had wide CIs (mean difference = 2.5%; 95% CI = −8.0 to 13.6%).

Conclusions: Applying target trial emulation methods to existing observational data for the evaluation of a novel intervention produced results similar to those of a randomized trial. These findings support the use of target trial emulation for comparative effectiveness research.

Keywords: clinical trials; intubation; epidemiology; causal inference; target trial emulation

(Received in original form March 18, 2019; accepted in final form April 29, 2019)

*A complete list of members of the Pragmatic Critical Care Research Group Investigators may be found in the online supplement.

‡T.J.I. is a Section Editor of *AnnalsATS*. His participation complies with American Thoracic Society requirements for recusal from review and decisions for authored works.

Supported by U.S. National Heart, Lung, and Blood Institute grants T32HL007749 (A.J.A.), K12HL138039 (J.P.D.), T32HL087738 (J.D.C.), and K23HL143053 (M.W.S.), and by U.S. Department of Veterans Affairs Health Services Research and Development grant 17-045 (T.J.I.).

Ann Am Thorac Soc Vol 16, No 8, pp 998–1007, Aug 2019

DOI: 10.1513/AnnalsATS.201903-241OC
Internet address: www.atsjournals.org

Randomized clinical trials (RCTs) represent the strongest evidence for comparing the effectiveness of two interventions because randomization of large populations ensures that other factors that influence the outcome are equally distributed between the treatment and control groups. However, RCTs are costly, slow, and often impractical to generate evidence for many important questions (1, 2). Observational studies may fill this gap, though confounding and bias are persistent risks (3–5).

Applying modern epidemiological methods to detailed data from completed clinical trials may reduce the likelihood of misleading results in observational comparative effectiveness studies (6, 7). "Target trial emulation" is a method in which investigators explicitly mimic an idealized clinical trial using observational techniques (8–10). Because data from prior clinical trials typically include detailed, accurately collected clinical information, they may allow investigators to better control for relevant confounders and emulate the design features and "randomization" of an idealized RCT (11, 12). These methods may be particularly helpful for conducting observational studies of therapies in critical care where clinician judgment is a strong determinant of treatment received (13). Few target trial emulations have employed detailed data from completed clinical trials, and even fewer have occurred concurrently with the clinical trial they seek to emulate, rendering direct comparison of findings between the approaches difficult.

In this study, we combined data from several completed clinical trials examining unrelated airway management interventions to estimate the effects of a novel intervention: positive-pressure ventilation during tracheal intubation of critically ill adults. Whether positive-pressure ventilation should be provided between induction and laryngoscopy during tracheal intubation has been debated for decades (14). Positive-pressure ventilation has been hypothesized to prevent hypoxemia (a contributor to cardiac arrest and death), but has also been hypothesized to increase the risk of oropharyngeal or gastric aspiration. Because an operator's decision regarding whether to administer positive-pressure ventilation may be based on perceived risk of periprocedural hypoxemia and aspiration, observational studies of bag mask ventilation are highly prone to confounding. To overcome this limitation, we applied causal methods and target trial emulation to predict the results of the PreVent (Preventing Hypoxemia with Manual Ventilation during Endotracheal Intubation) trial before its release (15). To ensure a fair evaluation, investigators performing the observational analysis (A.J.A., J.P.D., and T.J.I.) were kept unaware of the results of the PreVent trial, and registered both their analytical plan and the results of their observational analyses before the PreVent trial results were published (16). We hypothesized that the effect of positive-pressure ventilation on lowest oxygen saturation and severe hypoxemia would not differ between the observational study cohort and the randomized trial.

Methods

Target trial emulation is an observational research method that incorporates design features from idealized randomized trials (i.e., target trials) to improve the quality and interpretability of observational research (8–10, 12). Specifically, this approach requires careful design of inclusion and exclusion criteria similar to those that would have been used in the target trial and consideration of the mode of treatment allocation and the timing and intensity of the exposure. Investigators recreate these trial design features in observational data and use statistical methods that capitalize on natural variability in provider practice to emulate random treatment assignment.

On February 7, 2019, investigators from the PreVent trial provided the PreVent study protocol and deidentified patient-level data from three prior randomized trials of unrelated airway management interventions to a team of observational analysts (A.J.A., J.P.D., and T.J.I.) who did not know the results of the PreVent study (17). The analysts registered an analytical plan, statistical code, and results of their observational analysis (*see* the online supplement) with an honest broker on February 17, 2019, before the results of the PreVent trial were released (16). This approach allowed examination of whether causal methods applied to rich observational data yield similar results to a clinical trial, while minimizing the risk of bias that could be introduced by observational analysts knowing the results of the gold standard trial. An overview of the design decisions made in the observational analysts to emulate the randomized trial is shown in Table 1.

Description of the Target Trial

The observational analysis was intended to emulate PreVent, a multicenter, parallel-group, unblinded, pragmatic, randomized trial comparing positive-pressure ventilation with a bag-mask device to no positive-pressure ventilation between induction and laryngoscopy during tracheal intubation of critically ill adults. Briefly, PreVent

Author Contributions: A.J.A. had full access to all of the data in the study and takes responsibility for the integrity of the data and accuracy of the data analysis; study concept and design—A.J.A., J.P.D., J.D.C., T.J.I., and M.W.S.; acquisition, analysis, or interpretation of the data—A.J.A., J.P.D., J.D.C., D.R.J., D.W.R., A.M.J., D.J.V., K.M.D., S.B.S., J.M.D., T.W.R., T.J.I., and M.W.S.; drafting of the manuscript—A.J.A., J.P.D., J.D.C., T.J.I., and M.W.S.; critical revision of the manuscript for important intellectual content—A.J.A., J.P.D., J.D.C., D.R.J., D.W.R., A.M.J., D.J.V., K.M.D., S.B.S., J.M.D., T.W.R., T.J.I., and M.W.S.; statistical analysis—A.J.A. and J.P.D.; obtained funding—M.W.S.; administrative, technical, or material support—J.D.C., D.R.J., D.W.R., A.M.J., D.J.V., K.M.D., S.B.S., J.M.D., T.W.R., and M.W.S.; study supervision—M.W.S.

The content is the responsibility of the authors alone and does not necessarily reflect the views or policies of the Department of Veterans Affairs or the U.S. Government.

Correspondence and requests for reprints should be addressed to Andrew J. Admon, M.D., M.P.H., University of Michigan, 2800 Plymouth Road., Bldg. 16-169c, Ann Arbor, MI 48109. E-mail: ajadmon@umich.edu.

This article has an online supplement, which is accessible from this issue's table of contents at www.atsjournals.org.

Table 1. Comparison of target trial and observational analysis

Characteristic	Target Trial (PreVent)	Observational Analysis
Eligibility criteria	Included Admitted in a unit participating in PreVent Intubation planned 18 years old or older Excluded Pregnant Incarcerated Need for tracheal intubation too emergent Clinician deemed positive-pressure ventilation to be required (hypoxemia, severe acidemia, respiratory arrest) or contraindicated (active emesis, hematemesis, hemoptysis)	Included Admitted in a unit participating in Check-UP, FELLOW, or PrePARE Intubation planned 18 years old or older Any operator Excluded Pregnant Incarcerated Need for tracheal intubation too emergent Clinician deemed positive-pressure ventilation to be required (hypoxemia, severe acidemia, respiratory arrest) Sp_{O_2} at baseline <90% (49 people)
Treatment Strategies	-Intervention: positive-pressure ventilation with a bag-mask device (modified RSI) -Control: No positive-pressure ventilation except after a failed attempt or for Sp_{O_2} <90% (classic RSI)	-Exposure: positive-pressure ventilation with a bag-mask device or noninvasive ventilator (modified RSI) -Control: No positive-pressure ventilation except after a failed attempt or for Sp_{O_2} <90% (classic RSI)
Assignment procedures	Randomization in a 1:1 ratio to positive-pressure ventilation or no positive-pressure ventilation using permuted blocks of two, four, and six stratified by study site.	1. Exposure known for 125 patients (PrePARE patients) 2. For others, exposure imputed based on the following rules: a. Patients intubated on the first attempt without receiving bag-mask or noninvasive ventilation were assigned control group b. Patients who received bag-mask or noninvasive ventilation and were intubated on the first attempt were assigned intervention group, unless they had a prolonged intubation (top decile of intubation length) c. Patients who received bag-mask or noninvasive ventilation and required more than one intubation attempt or had a prolonged intubation were assigned using a predictive model based on known recipients of exposure
Follow-up period for primary outcome	From induction until 2 min after tracheal intubation	From induction until 2 min after tracheal intubation
Causal contrasts of interest	Lowest oxygen saturation between induction and 2 min after tracheal intubation	Lowest oxygen saturation between Induction and 2 min after tracheal intubation
	Other outcomes: Severe hypoxemia (Sp_{O_2} < 80%) Lowest Sp_{O_2}, highest FI_{O_2}, and highest PEEP between 6 and 24 h Operator reported aspiration New opacity on chest radiography within 48 h	Other outcomes: Severe hypoxemia (Sp_{O_2} < 80%) Lowest Sp_{O_2}, highest FI_{O_2}, and highest PEEP between 6 and 24 h Operator-reported aspiration
Analysis plan	Primary analysis: Mann-Whitney *U* test	Primary analysis: Mann-Whitney *U* test on propensity score matched population

Definition of abbreviations: Check-UP = Checklists and Upright Positioning in endotracheal intubation of critically ill patients; FELLOW = Facilitating EndotracheaL Intubation by Laryngoscopy Technique and Apneic Oxygenation within the Intensive Care Unit; FI_{O_2} = fraction of inspired oxygen; PEEP = positve end-expiratory pressure; PrePARE = Preventing Cardiovascular Collapse with Administration of Fluid Resuscitation before Tracheal Intubation; PreVent = Preventing Hypoxemia with Manual Ventilation during Endotracheal Intubation; RSI = rapid sequence intubation; Sp_{O_2} = oxygen saturation as measured by pulse oximetry.

(registered with www.clinicaltrials.gov [NCT03026322]) enrolled 401 patients between March 15, 2017 and May 6, 2018, with complete details of the methods and results reported February 18, 2019 (15).

Data Sources

Data for the observational analysis were obtained from three completed randomized trials evaluating airway management interventions among patients similar to PreVent's intended study population. The Check-UP (Checklists and Upright Positioning in endotracheal intubation of critically ill patients) study, was a randomized,

multicenter, pragmatic, two-by-two factorial trial comparing *1*) the ramped position with the sniffing position and *2*) the use of a written preintubation checklist during endotracheal intubation of critically ill adults (18). The FELLOW (Facilitating EndotracheaL Intubation by Laryngoscopy Technique and Apneic Oxygenation within the Intensive Care Unit) study was a randomized, open-label, parallel-group, pragmatic, two-by-two factorial trial comparing apneic oxygenation with usual care and direct laryngoscopy with video laryngoscopy among critically ill adults (19). The PrePARE (Preventing Cardiovascular Collapse with Administration of Fluid Resuscitation before Tracheal Intubation) trial compared the effects of a fluid bolus administered before induction versus no preinduction fluid bolus on the incidence of cardiovascular collapse during tracheal intubation (registered with www.clinicaltrials.gov [NCT03026777]). Although patients could be coenrolled in PrePARE and PreVent, the only data from the PrePARE trial used in the current analysis were those of patients at centers not actively enrolling in PreVent. Despite being comprised of data from randomized trials, the cohort derived from these data was referred to as "observational" because the specific intervention of interest—positive-pressure ventilation between induction and laryngoscopy—was allocated based on clinician choice rather than by randomization.

Eligibility Criteria

To derive the cohort for the observational analysis, we included adults (≥18 yr of age) enrolled in Check-UP (Checklists and Upright Positioning in endotracheal intubation of critically ill patients), FELLOW, or PrePARE (Table 1). We excluded patients known to have met exclusion criteria for PreVent (e.g., active emesis, brisk upper gastrointestinal bleeding, or unstable facial fractures). This led to the exclusion of 43 patients from PrePARE, the only trial in which these data were collected. We also excluded patients with an oxygen saturation at induction of less than 90% as a technique for discriminating positive-pressure ventilation used for rescue (which was allowed in both groups in PreVent after a failed attempt or for treatment of oxygen saturations <90%) from *preventive* positive-pressure ventilation (*see* Assignment Procedures).

Treatment Strategies

The PreVent trial compared positive-pressure ventilation with a bag-mask device between induction and laryngoscopy (sometimes referred to as modified rapid sequence intubation [RSI]) to no positive-pressure ventilation between induction and laryngoscopy except as treatment for hypoxemia or after a failed intubation attempted (conventional RSI). Although positive-pressure ventilation between induction and laryngoscopy may be delivered with either a bag-mask device or a noninvasive ventilator, to limit imprecision in the delivery of ventilation, all ventilation in the PreVent trial was required to be delivered using a bag-mask device. Because clinicians could deliver positive-pressure ventilation during intubation using either a bag-mask device or noninvasive ventilation in the observational cohort, we included both modalities of positive-pressure ventilation in the treatment group in the observational analysis.

Assignment Procedures

We sought to emulate the PreVent trial's randomization to positive-pressure ventilation at induction using data from our observational cohort. This was done in two steps: first by imputing treatment status at induction among patients for whom timing of treatment was uncertain, and second by using coarsened exact matching to identify comparable groups.

Our exposure of interest was positive-pressure ventilation beginning at induction, which was prospectively recorded in the PreVent and PrePARE trials, but the Check-UP and FELLOW trials captured only use of positive-pressure ventilation at *any time* from induction to intubation. This included patients who received positive-pressure ventilation at induction (the intervention studied in PreVent) and those who received rescue ventilation for treatment of hypoxemia or between successive intubation attempts (which was permitted in *either* arm in PreVent).

As a result, timing of positive-pressure ventilation at induction was known only for patients in the PrePARE trial, and in patients in Check-UP and FELLOW who *never* received positive-pressure ventilation between induction and intubation (and so, received conventional RSI). For patients in Check-UP and FELLOW who received positive-pressure ventilation at some point from induction to intubation, we followed a series of rules to determine whether it was started at induction (e.g., modified RSI, the intervention studied in the PreVent trial) or later in the intubation procedure for rescue. First, anyone with a baseline oxygen saturation of less than 90% was excluded from the observational analysis, as these patients would have qualified for rescue positive-pressure ventilation regardless of randomization arm in PreVent. Second, patients who received positive-pressure ventilation and were intubated on the first attempt were assumed to have received it at induction, unless they had a prolonged intubation (in the top 10% of procedure durations). Finally, in the remaining patients (those who received positive-pressure ventilation and either required more than one intubation attempt or had a prolonged intubation), we used single imputation with a model based on those for whom timing of positive-pressure ventilation was known, assigning patients with predicted probabilities of less than 0.5 to the no positive-pressure ventilation at induction (e.g., conventional RSI or control) group. In a sensitivity analysis, we used multiple imputation with 100 imputed datasets.

Next, to mimic the randomization performed in the PreVent trial, we generated propensity scores for receiving positive-pressure ventilation using the independent variables of study site, intervention group in the original completed trial, age, body mass index (BMI), race, sex, presence of sepsis, presence of chronic obstructive pulmonary disease exacerbation, indication for intubation (hypoxia or hypercarbia), highest fraction of inspired oxygen in the prior 6 hours, and baseline oxygen saturation (20). These variables were selected because they were thought to influence an operator's decision to apply positive-pressure ventilation at induction (21). Using these scores, we matched patients in a K:K ratio with coarsened exact matching and verified covariate balance across matched groups.

Follow-Up Period

In both the observational study and the PreVent trial, oxygen saturations were measured from induction through

2 minutes after tracheal intubation. Secondary and safety outcomes were similarly measured for up to 24 hours after intubation.

Causal Contrasts of Interest

The primary outcome in both the observational analysis and the PreVent trial was the lowest arterial oxygen saturation between induction and two minutes after tracheal intubation. This outcome was captured in the same manner in all three trials (Check-UP, FELLOW, and PrePARE) comprising our observational cohort. Our secondary outcome, also matching PreVent, was the proportion of patients with severe hypoxemia, defined as oxygen saturation less than 80%. We also examined each of the PreVent exploratory outcomes (oxygen saturation <70% and 90%, decrease in saturation from induction to nadir, ventilator-free days, intensive care unit–free days, and in-hospital death) and safety outcomes (operator-reported aspiration during intubation and cardiac arrest within 1 h of intubation).

Analysis Plan

To arrive at results that were directly comparable between the observational analysis and PreVent, we applied the statistical analysis plan developed for PreVent to analyses of all outcomes in the observational cohort. We first compared baseline patient characteristics between this observational study and PreVent. After ensuring covariate balance, we used the Mann-Whitney *U* test to compare distributions of nadir oxygen saturation between our treatment and control groups and a series of regression analyses to estimate mean differences or risk ratios (RRs) with 95% confidence intervals (CIs) for each outcome. To more closely emulate the analyses reported in PreVent, linear models used for effect estimation included only one independent variable (treatment assignment). We also estimated differences in effect estimates (difference-in-differences [DIDs]) for outcomes between the observational study and PreVent using bootstrapping with 10,000 replicates and bias-corrected 95% CIs. All tests were two-tailed. Analyses were conducted using Stata 14.2 (StataCorp LLC). This secondary analysis of deidentified data was determined nonhuman subject research by the Vanderbilt Institutional Review Board (no. 160158).

Results

Derivation of the Observational Cohort

The initial cohort for the observational analysis included 567 total patients, including 292 from Check-UP, 150 from FELLOW, and 125 from PrePARE (Figure 1). Among these patients, 348 received positive-pressure ventilation between induction and intubation. Before any adjustment, the median lowest oxygen saturation was 92% among patients who received positive-pressure ventilation compared with 95% among patients who did not receive positive-pressure ventilation (mean difference = −3.23; 95% CI = −5.59 to −0.88; $P = 0.02$).

From these 567 patients, 105 were excluded for baseline oxygen saturation of less than 90%, contraindications to enrollment in PreVent, or missing data. Of the remaining 462 patients, treatment status was known in 374 patients and assigned by imputation in 88 patients. The imputation model for positive-pressure ventilation at induction had adequate discrimination (c-statistic = 0.80) and fit (pseudo-$R^2 = 0.21$). Propensity scores using the updated treatment variable were created and 360 patients were matched (180 in both the treatment and control groups). These groups were well balanced with regard to median age, sex, median BMI, median APACHE (Acute Physiology and Chronic Health Evaluation) II scores, indications for intubation (hypoxemic respiratory failure, hypoxic respiratory failure, or altered mental status), and other variables. (Table 2 and Figure E1 in the online supplement).

Primary Outcome

For the 360 matched patients in the observational target trial emulation, the median lowest oxygen saturation was numerically higher in the positive-pressure ventilation group (94%; interquartile range [IQR] = 86–98%) compared with the no positive-pressure ventilation group (93%; IQR = 82–99%) (mean difference = 1.8 percentage points; IQR = −1.0 to 4.6), though this difference was neither clinically nor statistically significant ($P = 0.76$; Figure 2A).

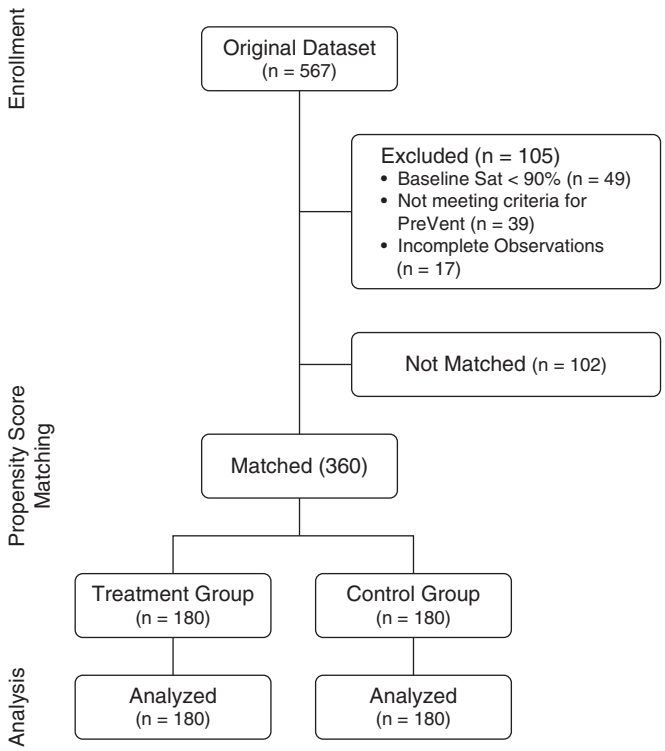


Figure 1. CONSORT (Consolidated Standards of Reporting Trials) diagram. PreVent = Preventing Cardiovascular Collapse with Administration of Fluid Resuscitation before Tracheal Intubation; Sat = saturation.

Table 2. Patient characteristics at baseline—observational study and PreVent Trial

Patient Characteristic	Observational Study		PreVent	
	Positive-Pressure Ventilation (*n* = 180)	No Ventilation (*n* = 180)	Bag Mask Ventilation (*n* = 199)	No Ventilation (*n* = 202)
Age, median (IQR), yr	60 (49–68)	56 (48–68)	59 (45–67)	60 (48–68)
Male sex, *n* (%)	106 (58.9)	102 (56.7)	118 (59.3)	108 (53.5)
White race, *n* (%)	140 (77.8)	140 (77.8)	141 (71.9)	134 (66.3)
Body mass index, median (IQR), kg/m^2	28.1 (23.4–33.4)	27.4 (23.5–32.2)	27.1 (22.7–32.3)	27.6 (23.4–34.2)
APACHE II score, median (IQR)	20.5 (16–26)	21 (16–25)	22 (16–29)	22 (16–28)
Vasopressors, *n* (%)	27 (15.0)	42 (23.3)	35 (17.6)	40 (19.9)
Active medical conditions, *n* (%)				
Sepsis or septic shock	98 (54.4)	97 (53.9)	98 (49.3)	97 (48.0)
Gastrointestinal bleeding	21 (11.7)	33 (18.3)	28 (14.1)	16 (7.9)
Indications for intubation, *n* (%)				
Hypoxemic respiratory failure	103 (57.2)	101 (56.1)	117 (58.8)	116 (57.4)
Hypercarbic respiratory failure	23 (12.8)	25 (13.9)	39 (19.6)	55 (27.2)
Airway protection for decreased level of consciousness	60 (33.3)	60 (33.3)	74 (37.2)	76 (37.6)
BiPAP in prior 6 h, *n* (%)	71 (39.4)	37 (20.6)	44 (22.1)	57 (28.2)
Highest F_{IO_2} in prior 6 h, median (IQR)	0.4 (0.3–0.7)	0.4 (0.3–0.8)	0.4 (0.3–1.0)	0.5 (0.3–1.0)
Lowest oxygen saturation in prior 6 h, median (IQR), %	91 (88–94)	92 (89–95)	91 (87–95)	92 (88–95)

Definition of abbreviations: APACHE II = Acute Physiology and Chronic Health Evaluation II; BiPAP = bilevel positive airway pressure; F_{IO_2} = fraction of inspired oxygen; IQR = interquartile range; PreVent = Preventing Hypoxemia with Manual Ventilation during Endotracheal Intubation.

Secondary, Safety, and Exploratory Outcomes

Among the 360 matched patients in the observational cohort, fewer patients experienced the secondary outcome of severe hypoxemia in the positive-pressure ventilation group (13.9%) compared with the no positive-pressure ventilation group (23.3%) (RR = 0.6; 95% CI = 0.4 to 0.9; Table 3 and Figure 2B). The observational cohort did not have adequate statistical power to exclude clinically meaningful differences in the safety and exploratory outcomes. For example, rates of operator-reported aspiration during intubation were 1.7% in the positive-pressure ventilation group and 0.6% in the no positive-pressure ventilation group (RR = 3.0; 95% CI = 0.3 to 28.6) and rates of death before hospital discharge were 40% in both treatment and control groups (RR = 1.0; 95% CI = 0.7 to 1.3).

Results of the Observational Analysis Compared with the PreVent Trial

In comparing baseline characteristics of the positive-pressure ventilation group in the observational analysis (the group at highest risk for imbalances in disease severity) to the positive-pressure ventilation group in the PreVent trial, we found that median age (60 yr, IQR = 49–68 vs. 59, IQR = 45 to 67), median BMI (28.1, IQR = 32.4 to 33.5 vs. 27.1, IQR = 22.7 to 32.3), and median APACHE II score (20.5, IQR = 16 to 26 vs. 22.0, IQR = 16 to 29) were similar between the observational analysis and PreVent. There were differences in the rates of noninvasive ventilation in the prior 6 hours before intubation (39.4% vs. 22.1% in the observational analysis vs. PreVent, respectively), likely a result of our inclusion of noninvasive ventilation in the definition of positive-pressure ventilation during induction.

Similar to the findings of the observational analysis, the primary outcome of lowest oxygen saturation during intubation was higher for patients in the positive-pressure ventilation group of the PreVent trial (96%, IQR = 89 to 99%) compared with the no positive-pressure ventilation group of the PreVent trial (93%, IQR = 81 to 98%) (mean difference = 3.9 percentage points, IQR = 1.4 to 6.4). Unlike in the observational analysis, the PreVent trial's difference in the primary outcome between groups met statistical significance ($P = 0.01$). Similar to the findings of the observational analysis, rates of severe hypoxemia in the PreVent trial were lower in the positive-pressure ventilation group (10.9%) than in the no positive-pressure ventilation group (22.8%) (RR = 0.5; 95% CI = 0.3 to 0.8). The differences between positive-pressure ventilation and no positive-pressure ventilation in safety and exploratory outcomes in the observational analysis and PreVent trial are displayed in Table 3.

In our DIDs analyses directly comparing the differences between the intervention and control groups across both studies, we found that the effect of positive-pressure ventilation on the primary outcome of lowest oxygen saturation was similar between the observational analysis and the PreVent trial (DID = −2.1%; 95% CI = −5.9 to 1.7%; Table 4). We also found that the effect of positive-pressure ventilation on rates of severe hypoxemia was similar between the observational analysis and the PreVent trial (DID = 2.5%; 95% CI = −8.1% to 13.6%). There were no substantial differences for any of the other exploratory or safety outcomes between studies, though we could not exclude clinically important differences for several exploratory outcomes including median ventilator-free days (DID = −1.1; 95% CI = −4.5 to 2.2), median intensive care unit–free days (−0.7; 95% CI = −3.7 to 2.4), or death before hospital discharge (0.0%; 95% CI = −13.7 to 13.8).

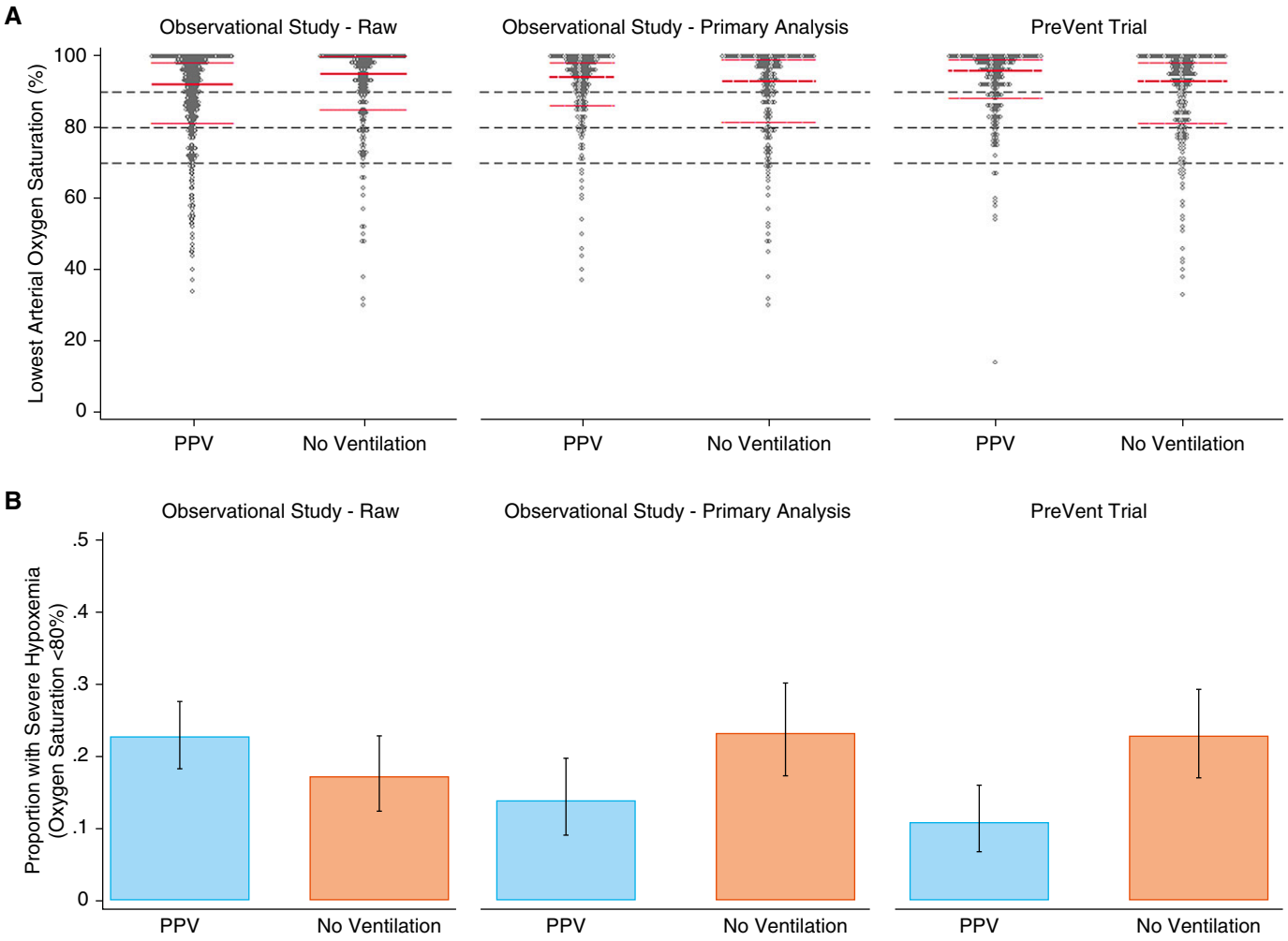


Figure 2. Lowest arterial oxygen saturation and severe hypoxemia in treatment versus control groups. (*A*) Dot plots showing the distributions of lowest oxygen saturation by group. Bold line represents the median; thinner lines represent the 25th and 75th percentile. (*B*) Bar charts showing the proportion with severe hypoxemia (lowest oxygen saturation <80%) by group. Error bars represent 95% confidence interval limits. PPV = positive pressure ventilation; PreVent = Preventing Hypoxemia with Manual Ventilation during Endotracheal Intubation.

Discussion

This study found that blinded application of target trial emulation analysis to observational data accurately predicted the results of a critical care clinical trial. The effects of positive-pressure ventilation on lowest oxygen saturation and severe hypoxemia during tracheal intubation of critically ill adults were similar in our analysis of observational data and in the subsequently published PreVent randomized trial. Our analysis of observational data, however, was unable to exclude the possibility of no difference between groups in the primary outcome of lowest oxygen saturation, and neither the observational cohort nor the clinical trial provided adequate power to assess rare safety or exploratory outcomes. To our knowledge, this is the first study to directly compare results of a blinded target trial emulation using observational data to the realized target trial. Collectively, the similarities between our results and those of the PreVent trial suggest that target trial emulation using existing observational data may offer a powerful tool for comparative effectiveness research using rich, clinical data.

Observational studies of critical care interventions face unique challenges due to confounding by indication, whereby providers allocate a treatment nonrandomly in a way that is dependent on other variables (e.g., clinical trajectory or severity of illness), which themselves influence the outcome of interest (22). In the absence of effective statistical adjustment (using propensity score matching, regression modeling, inverse probability of treatment weighting, or other techniques), confounding by indication can produce misleading results in observational comparative effectiveness studies (23). The influence of these biases was evident in our unadjusted analysis, where positive-pressure ventilation was associated with a significantly lower oxygen saturation than no positive-pressure ventilation—the opposite finding of our adjusted analysis and the PreVent trial. Despite this imbalance in the raw data, we were able to arrive at closely balanced intervention and control groups using

Table 3. Target trial outcomes—observational cohort and PreVent Trial

Outcomes	Observational Study			PreVent Trial		
	Positive-Pressure Ventilation (*n* = 180)	No Ventilation (*n* = 180)	RR or Mean Difference (95% CI)	Bag Mask Ventilation (*n* = 199)	No Ventilation (*n* = 202)	RR or Mean Difference (95% CI)
Primary outcome						
Lowest oxygen saturation, median (IQR), %	94 (86 to 98)	93 (81.5 to 99)	1.8 (−1.0 to 4.6)	96 (88 to 99)	93 (81 to 98)	3.9 (1.4 to 6.4)
Secondary outcome						
Lowest oxygen saturation <80%, *n* (%)	25/180 (13.9)	42/180 (23.3)	0.60 (0.4 to 0.9)	21/193 (10.9)	45/197 (22.8)	0.48 (0.3 to 0.8)
Exploratory oxygen saturation outcomes						
Lowest oxygen saturation <90%, *n* (%)	59/180 (32.8)	66/180 (36.7)	0.89 (0.7 to 1.2)	57/193 (29.5)	79/197 (40.1)	0.74 (0.6 to 1.0)
Lowest oxygen saturation <70%, *n* (%)	12/180 (6.7)	19/180 (10.6)	0.63 (0.3 to 1.3)	8/193 (4.2)	20/197 (10.2)	0.41 (0.2 to 0.9)
Decrease in oxygen saturation, median (IQR), %	3 (0 to 11)	3 (0 to 13)	−1.9 (−4.5 to 0.8)	1 (0 to 7)	5 (0 to 14)	−4.5 (−6.8 to −2.2)
Exploratory safety outcomes						
Operator-reported aspiration, *n* (%)	3/180 (1.7)	1/180 (0.6)	3.00 (0.3 to 28.6)	5/198 (2.5)	8/202 (4.0)	0.64 (0.2 to 1.9)
Cardiac arrest 1 h after intubation, *n* (%)	4/180 (2.2)	2/180 (1.1)	2.00 (0.4 to 10.8)	3/199 (1.5)	4/202 (2.0)	0.76 (0.2 to 3.4)
Exploratory clinical outcomes						
Ventilator-free days, median (IQR)	15 (0 to 25)	15 (0 to 25)	−0.5 (−3.0 to 1.9)	19 (0 to 25)	17.5 (0 to 25)	0.6 (−1.7 to 2.9)
ICU-free days, median (IQR)	10 (0 to 23)	10 (0 to 22)	0.1 (−2.2 to 2.4)	16 (0 to 22)	13.5 (0 to 22)	0.8 (−1.3 to 2.9)
Died before hospital discharge, *n* (%)	72/180 (40.0)	72/180 (40.0)	1.00 (0.8 to 1.3)	71/199 (35.7)	72/202 (35.6)	1.00 (0.8 to 1.3)

Definition of abbreviations: CI = confidence interval; ICU = intensive care unit; IQR = interquartile range; PreVent = Preventing Hypoxemia with Manual Ventilation during Endotracheal Intubation; RR = risk ratio.

propensity score and coarsened exact matching. This was dependent on the availability of several important confounders (e.g., APACHE II score, reason for intubation, oxygen saturation at induction, and BMI) in our dataset (11, 20). In general, clinical data are better suited to target trial analysis than administrative claims or other datasets with less complete capture of relevant confounders (24).

Other groups have used observational data to explicitly emulate randomized experiments. Hernán and colleagues (8) analyzed data from the Nurses' Health Study to emulate an intention-to-treat analysis of an idealized randomized trial of postmenopausal hormone therapy on coronary heart disease. By emulating design features of an idealized RCT, the authors show that differences in results between observational and randomized studies of hormone replacement therapy may be due to discrepancies in the included patient population and other design factors, and not to unmeasured confounding. A criticism of prior target trial emulations, however, is that they are conducted with knowledge of the resultant randomized experiment. By emulating PreVent *before* its results were known to the analysts, we eliminated the risk that prior knowledge of study results would influence our analytical approach even inadvertently.

Target trial emulation also draws attention to the timing of eligibility and treatment assignment, two standard design features of randomized experiments that can lead to bias when mismatched in observational analyses (12). García-Albéniz and colleagues (9) applied target trial emulation to evaluate the effects of emulated "randomization" to screening colonoscopy on subsequent colorectal cancer risk. By determining both eligibility and assigning treatment simultaneously at the start of a study, and by analyzing patients who "break protocol" with their initially assigned group, the authors show that they are able to arrive at more credible effect estimates. We similarly assigned treatment at the "start" of induction, using imputation to ascertain intervention status when this was not known, before subsequently emulating the randomization process using coarsened exact matching.

Our study does have several limitations. A major limitation is that criteria for

Table 4. Difference-in-differences estimates comparing the observational study and PreVent

Outcome	Treatment Effect		Difference-in-Differences (95% CI)
	Observational Study	PreVent Trial	
Primary outcome			
Lowest oxygen saturation, %	1.8	3.9	−2.1 (−5.9 to 1.7)
Secondary outcome			
Lowest oxygen saturation <80%	−9.4	−12.0	2.5 (−8.1 to 13.6)
Exploratory oxygen saturation outcomes			
Lowest oxygen saturation <90%	−3.9	−10.6	6.7 (−7.3 to 20.4)
Lowest oxygen saturation <70%	−3.9	−6.0	2.1 (−5.4 to 10.1)
Decrease in oxygen saturation, %	−1.9	−4.5	2.6 (−0.8 to 6.2)
Exploratory safety outcomes			
Operator-reported aspiration	1.1	−1.4	2.6 (−1.5 to 6.8)
Cardiac arrest 1 h after intubation	1.1	−0.5	1.6 (−2.0 to 5.3)
Exploratory clinical outcomes			
Ventilator-free days, median	−0.5	0.6	−1.1 (−4.5 to 2.2)
ICU-free days, median	0.1	0.8	−0.7 (−3.7 to 2.4)
Died before hospital discharge	0.0	0.0	0.0 (−13.7 to 13.8)

Definition of abbreviations: CI = confidence interval; ICU = intensive care unit; PreVent = Preventing Hypoxemia with Manual Ventilation during Endotracheal Intubation.

formally assessing the agreement between the observational analysis and the PreVent trial were not prespecified as part of our analytical plan. Authoritative guidance for comparing target trial emulations with realized target trials does not yet exist; we suggest using a prespecified minimal clinically important difference in outcomes and testing agreement with prespecified statistical tests, such as a DIDs estimator (25). In light of this limitation, however, the findings of this analysis should be considered exploratory. Second, we included positive-pressure ventilation delivered via noninvasive ventilation in our treatment group, though the PreVent trial required that positive pressure be delivered via bag-mask device for all patients randomized to positive-pressure ventilation. Although this decision was made with the intention of approximating the causal question in PreVent without excluding the large group of patients receiving noninvasive ventilation in the prior trials we analyzed, there could be differences in positive-pressure ventilation delivered by noninvasive ventilation versus bag-mask device that render this a sufficiently different intervention. Third, although we used both content knowledge and published evidence to generate a causal model and controlled for all relevant confounders in that model, this approach does not eliminate the possibility that unmeasured confounding may have influenced our findings. Finally, our observational analysis could not exclude there being no effect of positive-pressure ventilation on the primary outcome, median lowest oxygen saturation, although the direction and magnitude of our result was similar to findings from PreVent. Potential reasons for this may be our more restrictive inclusion criteria (including only patients with oxygen saturations of ⩾90%), effects of the other interventions (e.g., apneic oxygenation) tested in the trials that comprised our observational cohort, or our smaller effective sample size after matching. Nevertheless, our findings yielded a similar interpretation, that positive-pressure ventilation applied starting at induction reduced rates of severe hypoxemia.

Despite these limitations, our findings suggest several potential applications for target trial emulations. When RCTs are not available or are impractical, rigorously conducted target trial emulation using rich clinical data may yield stronger evidence to support clinical decisions than other types of observational research (9, 12). In the case where an RCT is planned, target trial emulation may inform study design, including sample size calculations, outcome selection, and selection of inclusion/exclusion criteria. Specifically, when selecting outcomes for clinical trials, investigators often attempt to balance a number of factors, including relevance to patients, providers, and other stakeholders, measurement characteristics, and statistical efficiency (26). This latter consideration typically favors the selection of continuous outcomes.

With regard to PreVent, target trial emulation suggested that the intervention may have had a greater effect on the outcome of severe hypoxemia than on the outcome of lowest oxygen saturation. This finding might have led investigators to consider severe hypoxemia as a dichotomous primary outcome. Based on the observational analysis, the relatively high baseline of severe hypoxemia in the control group coupled with the large difference in outcome rates between the intervention and control groups would have yielded an estimated sample size of 359 patients per group to achieve 90% power for detecting a similar difference at the α level of 0.05 (27). Though this is greater than the sample size ultimately used for the continuous outcome, and would have led to a lengthier trial, it may have prompted investigators to favor severe hypoxemia given its clinical relevance. Finally, clinical trialists interested in evaluating several closely related interventions may consider building in prespecified observational studies into prospective randomized trials. These methods might generate stronger evidence than would be possible with other types of observational data, with a significant gain

in efficiency over the cost or time involved in conducting separate clinical trials. For example, our observational analysis was designed and conducted in 11 days, whereas the PreVent trial itself required more than 2 years.

In summary, we used preexisting clinical data to evaluate a novel intervention and arrived at results comparable to those of a contemporaneous randomized experiment. This suggests that target trial emulation using modern causal methods and rich clinical trial data can provide informative results for comparative effectiveness research. These results also support the routine use of positive-pressure ventilation during tracheal intubation of critically ill adults to prevent severe hypoxemia. ■

Author disclosures are available with the text of this article at www.atsjournals.org.

References

1 Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med* 2000;342:1887–1892.
2 Vincent J-L. We should abandon randomized controlled trials in the intensive care unit. *Crit Care Med* 2010;38(10 suppl):S534–S538.
3 Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med* 2000;342:1878–1886.
4 Shikata S, Nakayama T, Noguchi Y, Taji Y, Yamagishi H. Comparison of effects in randomized controlled trials with observational studies in digestive surgery. *Ann Surg* 2006;244:668–676.
5 Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JPA. Agreement of treatment effects for mortality from routinely collected data and subsequent randomized trials: meta-epidemiological survey. *BMJ* 2016;352:i493.
6 Moodie EEM, Stephens DA. Using directed acyclic graphs to detect limitations of traditional regression in longitudinal studies. *Int J Public Health* 2010;55:701–703.
7 Pearl J. An introduction to causal inference. *Int J Biostat* 2010;6:7.
8 Hernán MA, Alonso A, Logan R, Grodstein F, Michels KB, Willett WC, *et al*. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology* 2008;19:766–779.
9 García-Albéniz X, Hsu J, Hernán MA. The value of explicitly emulating a target trial when using real world evidence: an application to colorectal cancer screening. *Eur J Epidemiol* 2017;32:495–500.
10 Labrecque JA, Swanson SA. Target trial emulation: teaching epidemiology and beyond. *Eur J Epidemiol* 2017;32:473–475.
11 Dahabreh IJ, Sheldrick RC, Paulus JK, Chung M, Varvarigou V, Jafri H, *et al*. Do observational studies using propensity score methods agree with randomized trials? A systematic comparison of studies on acute coronary syndromes. *Eur Heart J* 2012;33:1893–1901.
12 Hernán MA, Sauer BC, Hernández-Díaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *J Clin Epidemiol* 2016;79:70–75.
13 Walkey AJ. Expanding the dimensions of effectiveness research in sepsis. *Am J Respir Crit Care Med* 2014;190:970–971.
14 El-Orbany M, Connolly LA. Rapid sequence induction and intubation: current controversy. *Anesth Analg* 2010;110:1318–1325.
15 Casey JD, Janz DR, Russell DW, Vonderhaar DJ, Joffe AM, Dischert KM, *et al*. Bag-mask ventilation during tracheal intubation of critically ill adults. *N Engl J Med* 2019;380:811–821.
16 Admon AJ, Donnely JP, Iwashyna, TJ. Emulating a novel target trial using previously collected clinical trial data: an application to positive pressure ventilation during tracheal intubation. Charlottesville, VA: The Center for Open Science; 18 February 2019 [accessed 2019 Feb 18]. Available from: http://osf.io/p8cz3.
17 Casey JD, Janz DR, Russell DW, Vonderhaar DJ, Joffe AM, Dischert KM, *et al*.; PreVent Investigators and the Pragmatic Critical Care Research Group. Manual ventilation to prevent hypoxaemia during endotracheal intubation of critically ill adults: protocol and statistical analysis plan for a multicentre randomised trial. *BMJ Open* 2018;8:e022139.
18 Semler MW, Janz DR, Russell DW, Casey JD, Lentz RJ, Zouk AN, *et al*.; Check-UP Investigators(*); Pragmatic Critical Care Research Group. A multicenter, randomized trial of ramped position vs sniffing position during endotracheal intubation of critically ill adults. *Chest* 2017;152:712–722.
19 Semler MW, Janz DR, Lentz RJ, Matthews DT, Norman BC, Assad TR, *et al*.; FELLOW Investigators; Pragmatic Critical Care Research Group. Randomized trial of apneic oxygenation during endotracheal intubation of the critically ill. *Am J Respir Crit Care Med* 2016;193:273–280.
20 Garrido MM, Kelley AS, Paris J, Roza K, Meier DE, Morrison RS, *et al*. Methods for constructing and assessing propensity scores. *Health Serv Res* 2014;49:1701–1720.
21 McKown AC, Casey JD, Russell DW, Joffe AM, Janz DR, Rice TW, *et al*. Risk factors for and prediction of hypoxemia during tracheal intubation of critically ill adults. *Ann Am Thorac Soc* 2018;15:1320–1327.
22 Sjoding MW, Luo K, Miller MA, Iwashyna TJ. When do confounding by indication and inadequate risk adjustment bias critical care studies? A simulation study. *Crit Care* 2015;19:195.
23 Wunsch H, Linde-Zwirble WT, Angus DC. Methods to adjust for bias and confounding in critical care health services research involving observational data. *J Crit Care* 2006;21:1–7.
24 Poses RM, Smith WR, McClish DK, Anthony M. Controlling for confounding by indication for treatment. Are administrative data equivalent to clinical data? *Med Care* 1995;33(4 suppl):AS36–AS46.
25 Chan LS. Minimal clinically important difference (MCID)—adding meaning to statistical inference. *Am J Public Health* 2013;103:e24–e25.
26 Iwashyna TJ, McPeake J. Choosing outcomes for clinical trials: a pragmatic perspective. *Curr Opin Crit Care* 2018;24:428–433.
27 Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. *Lancet* 2005;365:1348–1353.

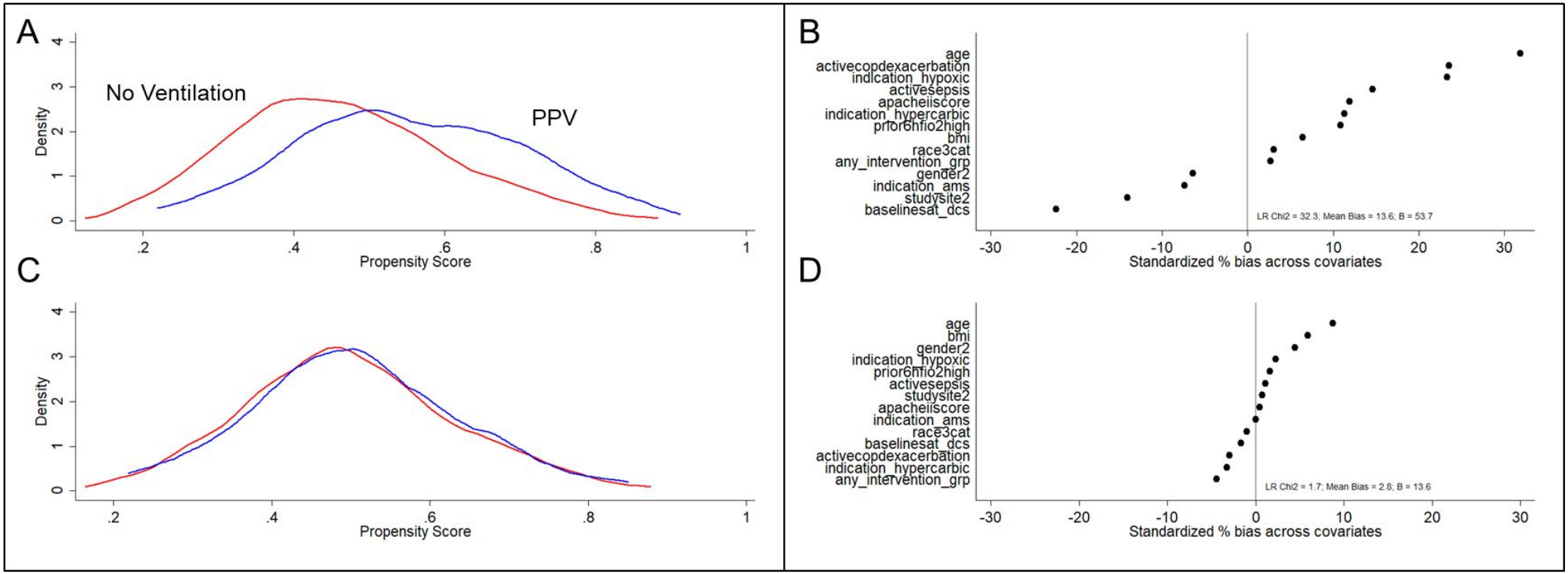
Online Supplement

Emulating a Novel Clinical Trial using Existing Observational Data: Predicting Results of the PreVent Study

Andrew J. Admon, MD, MPH, John P. Donnelly, PhD, Jonathan D. Casey, MD, David. R. Janz, MD, MSc, Derek W. Russell, MD, Aaron M. Joffe, DO, Derek J. Vonderhaar, MD, Kevin M. Dischert, MD, Susan B. Stempek, MMSc, PA, James M. Dargin, MD, Todd W. Rice, MD, MSc, Theodore J. Iwashyna, MD, PhD, and Matthew W. Semler, MD, MSc for the Pragmatic Critical Care Research Group.

Table of Contents

Supplementary Figure 1 - Propensity score distributions and balance of covariates between matched groups. (A) Densities of propensity score by treatment and control group before matching. (B) Standardized percent bias before matching. (C) Densities of propensity score by treatment and control group after matching. (D) Standardized percent bias after matching.



List of Investigators for the Pragmatic Critical Care Research Group

Vanderbilt University Medical Center, Nashville, TN – Ryan M. Brown, Jonathan D. Casey*, Todd W. Rice*, Wesley H. Self, Matthew W. Semler*; Lahey Hospital & Medical Center, Burlington, MA – James Dargin*, Susan Stempek*, Joanne Wozniak; Louisiana State University School of Medicine, New Orleans, LA – Bennett P. deBoisblanc, David R. Janz*, Yasin A. Khan; Ochsner Health System, New Orleans, LA – Kevin M. Dischert*, Derek J. Vonderhaar*; University of Alabama, Birmingham, AL – Swati Gulati, Derek W. Russell*, William S. Stigler, Aline N. Zouk; University of Washington, Seattle, WA – Itay Bentov, Aaron M. Joffe*

*Denotes members of the Writing Committee.

Correspondences with Dr. David Lederer (Honest Broker)

Initial submission from analysts to honest broker:

From: Admon, Andrew
Date and Time: February 17, 2019 11:58 PM, ET
Subject: pre-registering a prospective Target Trial

Hi Dr. Lederer,

Thank you again for your willingness to serve as an 'honest broker'. In the attached document is a short introduction, our complete analysis plan, and our results with a brief interpretation. We've also included all of our statistical code in the appendix. We're looking forward to seeing the results of the actual target trial (PreVent) tomorrow!

Regards,
Andy Admon

—
Andrew J. Admon, MD, MPH
Fellow, Pulmonary and Critical Care Medicine
Department of Internal Medicine
University of Michigan

Response from honest broker:

From: Lederer, David
Date and Time: February 18, 2019 6:20 AM, ET
Subject: pre-registering a prospective Target Trial

Hi Andy,

I am confirming that I received your report including the complete analysis plan and results. Received February 18, 2019. 6:20am Eastern Time.

Best
Dave

Pre-Registered Description of Analytical Plan and Results (Registered 2/18/2019)

Introduction

Well-conducted randomized, controlled trials (RCTs) represent the strongest level of evidentiary support for or against medical interventions. Unfortunately, RCTs are costly, lengthy, and, for some interventions, impractical, leaving us limited in our ability to apply this gold standard to a number of important clinical questions. Target trial emulation using observational data may fill this important gap, though confounding by indication, whereby patients selectively receive the intervention due to factors also associated with the study's outcome, may lead to misleading results.(1–4)

Modern causal methods applied to rich, clinical trial data may reduce this important source of confounding. Specifically, clinical trial data often include detailed, accurately collected information obtained with the intent of informing trial interpretation. In the context of critical care trials, this often includes several patient-level (i.e., detailed physiologic data), provider-level (i.e., experience and specialty), and unit-level (i.e., medical versus surgical) variables not often available in other datasets. By capitalizing on these data, investigators may extend the impact of a clinical trial by accurately emulating related target trials to answer additional clinical questions.

In this study, we use existing clinical trial data to estimate the effects of positive pressure ventilation (PPV) added to conventional rapid sequence intubation on procedural hypoxia. To do so, we analyze data collected from three prior randomized trials to emulate the recently completed Preventing Hypoxemia with Manual Ventilation during Endotracheal Intubation (PreVent) trial. The PreVent study protocol was provided to the analysts (AJA and JPD) by trial investigators (JDC and MWS) for the purpose of target trial emulation before the results of the trial were published. The analysts subsequently registered an analytical plan, statistical code,

and results without knowledge of the target trial's findings in order to evaluate whether causal methods could be applied to the observational data to 1) emulate the randomization procedure of the target trial and 2) arrive at similar conclusions.

Methods

Description of the Target Trial

The target trial is the PreVent trial, a multi-center, parallel-group, unblinded, pragmatic randomized trial comparing bag-mask ventilation to no ventilation between induction and laryngoscopy during endotracheal intubation among critically ill adults (Citation Pending). Complete details about the target trial are reported elsewhere.

Data Source

Data were obtained from three randomized trials evaluating interventions related to endotracheal intubation in similar patient populations and study settings to the PreVent trial. The Checklists and Upright Positioning in endotracheal intubation of critically ill patients (Check-UP) study was a randomized, multicenter, pragmatic two-by-two factorial trial comparing 1) the ramped position with the sniffing position and 2) the use of an a written pre-intubation checklist during endotracheal intubation of critically ill adults.(5) The Facilitating Endotracheal intubation by Laryngoscopy technique and apneic Oxygenation Within the intensive care unit (FELLOW) study was a randomized, open-label, parallel group, pragmatic two-by-two factorial trial comparing apneic oxygenation with usual care and direct laryngoscopy with video laryngoscopy among critically ill adults.(6) The PrePARE trial compared administration of a fluid bolus prior to induction to no administration of a fluid bolus with regard to the incidence of cardiovascular collapse during tracheal intubation of critically ill adults. Although patients could be co-enrolled in the PrePARE and PreVent trials, the only data from the PrePARE trial used in the current analysis was data from centers not actively enrolling in the PreVent trial. The cohort derived

from this dataset is referred to as the observational cohort because it was observational with respect to the primary intervention, which was not allocated by randomization.

Eligibility Criteria

We included adult (aged >= 18) patients enrolled in Check-UP, FELLOW, or PrePARE. (**Table 1**). Patients enrolled in PreVent were not included in this analysis. We excluded pregnant or incarcerated patients, patients for whom the need for tracheal intubation was deemed too urgent for randomization, or cases where bag-valve mask ventilation was deemed necessary or contraindicated (known and true for 43 patients). For purposes of the observational study, we also excluded patients with a baseline oxygen saturation below 90% by pulse oximetry. This was done in order to more accurately separate use of PPV for rescue (done in those with oxygen saturations below 90% in PreVent) from those treated with *preventive* PPV, at the risk of biasing our observational results towards the null (see: *Assignment Procedures,* below).

Treatment Strategies

The PreVent trial compared modified RSI (defined as RSI with positive pressure ventilation via bag-valve mask between induction and laryngoscopy) to standard RSI (defined as RSI without positive pressure between induction and laryngoscopy). To limit imprecision in the exposure of interest (positive pressure ventilation) stemming from differences in pressure settings, all PPV in the PreVent trial was delivered using non-invasive ventilation. Because the decision to deliver PPV during intubation in the observational cohort could have led to either bag valve mask or non-invasive ventilation, we included both of these modalities in our treated group. Patients receiving neither modality of PPV were included in the untreated, or control group.

Assignment Procedures

We sought to emulate the PreVent trial's randomization to PPV at induction using data from our observational cohort.

First, although the exposure of interest was PPV beginning at induction, Check-UP and FELLOW captured only use of positive pressure ventilation at any time from induction to intubation. This included patients who received PPV at induction (e.g., the intervention arm in PreVent) and those who received PPV for rescue during prolonged intubations, in between successive intubation attempts, or when oxygen saturation dropped below 90% (either the intervention or control arms in PreVent). As a result, although we knew intubation status for subjects who never received PPV between induction and intubation, we followed a series of rules to recreate treatment status for patients who received PPV at an uncertain point during intubation. First, anyone with a baseline oxygen saturation of <90% was excluded, as these patients would have qualified for PPV regardless of randomization arm in PreVent. Second, subjects who never received PPV between induction and intubation were assigned to the control group. Third, those subjects without a prolonged intubation (defined as either > 1 attempt or in the top 10% of procedure durations) who received PPV were assumed to have received this at or close to the time of induction, and so were assigned to the treatment arm. Finally, intubation status among the 371 patients for whom exposure status was known more confidently was used to predict exposure status at intubation among the remaining 91 patients using a logistic model. This model had good discrimination (c-statistic: 0.80) and fit (pseudo R^2: 0.21) and resulted in reclassifying 27 people from the treatment to control groups.

Next, to mimic the randomization performed in the PreVent trial, we generated propensity scores using study site, intervention group, age, BMI, race, gender, presence of sepsis, presence of chronic obstructive pulmonary disease exacerbation, indication for intubation (hypoxia or hypercarbia), highest fraction of inspired oxygen in the prior six hours, and baseline

oxygen saturation. These variables were selected because they were thought to influence both an operator's decision to apply positive pressure ventilation and a subject's lowest oxygen saturation (i.e., they are possible confounders).(7) We matched patients in a 1:1 ratio using coarsened exact matching and verified that covariates were balanced across matched groups. Of 462 patients in the analytical sample, 360 patients were matched (180 in both the treatment and control groups).

Follow-up Period

Follow-up periods were similar between this observational study and the PreVent trial. Specifically, oxygen saturations were measured from induction through two minutes after tracheal intubation. Secondary and safety outcomes were measured for up to 24 hours after intubation.

Causal Contrasts of Interest

The primary outcome in both this observational study and in PreVent was nadir oxygen saturation. This was the primary outcome in all three trials (Check-UP, FELLOW, and PrePARE) comprising our analytical cohort. Secondary outcomes included proportion of patients with severe hypoxemia (below 80%),

Analysis Plan

We first evaluated our propensity matching procedure by evaluating balance of baseline patient characteristics ("Table 1" variables from PreVent) across our treatment and control groups using two sample t-tests. After assessing balance of these covariates across groups, we next sought to apply the statistical analysis plan used in the PreVent trial in analyzing primary and all secondary/safety outcomes. As a result, we used the Mann-Whitney U test to compare distributions of nadir oxygen saturation between our treatment and control groups and a series

of regression analyses to estimate mean differences or risk ratios for each outcome. Because propensity score matching was employed to adjust for confounding, linear models used for effect estimation included only one independent variable (treatment assignment). All tests were two-tailed. Analyses were conducted using Stata 14.2 (StataCorp LLC, College Station, Texas).

Brief Interpretation of Results

Using propensity score matching, we were able to arrive at two balanced groups for comparison that differed in their use of PPV at induction (**Figure 1, Table 2**). The median oxygen saturation, our primary outcome, was slightly higher in the intervention (PPV) group, though we could not exclude there being no difference. (**Figure 2**). There were fewer cases of severe hypoxia, defined as an arterial oxygen saturation of <80% by pulse oximeter, in the intervention group with no substantial differences in either of the safety outcomes (**Table 3**).

References

1. Kyriacou DN, Lewis RJ. Confounding by Indication in Clinical Research. *JAMA* 2016;316:1818.
2. Hernán MA, Sauer BC, Hernández-Díaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *J Clin Epidemiol* 2016;doi:10.1016/j.jclinepi.2016.04.014.
3. Cain LE, Saag MS, Petersen M, May MT, Ingle SM, Logan R, Robins JM, Abgrall S, Shepherd BE, Deeks SG, Gill MJ, Touloumi G, Vourli G, Dabis F, Vandenhende MA, Reiss P, van Sighem A, Samji H, Hogg RS, Rybniker J, Sabin CA, Jose S, del Amo J, Moreno S, Rodríguez B, Cozzi-Lepri A, Boswell SL, Stephan C, Pérez-Hoyos S, *et al.* Using observational data to emulate a randomized trial of dynamic treatmentswitching strategies: An application to antiretroviral therapy. *Int J Epidemiol* 2016;doi:10.1093/ije/dyv295.
4. Hernán MA, Alonso A, Logan R, Grodstein F, Michels KB, Willett WC, Manson JE, Robins JM. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology* 2008;19:766–79.
5. Janz DR, Semler MW, Joffe AM, Casey JD, Lentz RJ, deBoisblanc BP, Khan YA, Santanilla JI, Bentov I, Rice TW, Check-UP Investigators* GP, Pragmatic Critical Care Research Group JL, Vonderhaar DJ, Lapinel NC, Samant SD, Paccione R, Dischert K, Majid-Moosa A, Crespo J, Fashho MB, Matthews DT, Berg JZ, Assad TR, McKown AC, Huerta LE, Kocurek EG, Halliday SJ, Kerchberger VE, Merrick CM, *et al.* A Multicenter Randomized Trial of a Checklist for Endotracheal Intubation of Critically Ill Adults. *Chest* 2018;153:816–824.
6. Semler MW, Janz DR, Lentz RJ, Matthews DT, Norman BC, Assad TR, Keriwala RD, Ferrell BA, Noto MJ, McKown AC, Kocurek EG, Warren MA, Huerta LE, Rice TW,

FELLOW Investigators, Pragmatic Critical Care Research Group. Randomized Trial of Apneic Oxygenation during Endotracheal Intubation of the Critically Ill. *Am J Respir Crit Care Med* 2016;193:273–80.

7. McKown AC, Casey JD, Russell DW, Joffe AM, Janz DR, Rice TW, Semler MW. Risk factors for and prediction of hypoxemia during tracheal intubation of Critically Ill Adults. *Ann Am Thorac Soc* 2018;15:1320–1327.

Pre-Registered Comparison of Target Trial and Observational Study

Characteristic	Target Trial	Observational Study

Eligibility Criteria	• Included
	  ○ Admitted in a unit participating in PreVent
	  ○ Intubation planned
	  ○ 18 years old or older
	  ○ Qualified operator
	• Excluded
	  ○ Pregnant
	  ○ Incarcerated
	  ○ Need for tracheal intubation too emergent
	  ○ Clinician deemed patient to need BVM (hypoxemia, severe acidemia, respiratory arrest)
	  ○ BVM Contraindicated or intubation too urgent
		• Included
		  ○ Admitted in a unit participating in Check-UP, FELLOW, or PreVent
		  ○ Intubation planned
		  ○ 18 years old or older
		  ○ Any operator
		• Excluded
		  ○ Pregnant
		  ○ Incarcerated
		  ○ Need for tracheal intubation too emergent
		  ○ Clinician deemed patient to need BVM (hypoxemia, severe acidemia, respiratory arrest)
		  ○ BVM contraindicated or intubation too urgent, when known
		  ○ SaO2 at baseline < 90% [49 people]

Treatment Strategies	• Modified RSI [PPV via BVM at induction] • Classic RSI [control] with BVM Rescue (if SaO2 < 90%) and in between attempts	• Modified RSI [PPV via BVM or NIV at induction] • Classic RSI [control] with BVM Rescue (if SaO2 < 90%) and in between attempts

Assignment Procedures	Randomization 1:1 ratio to BVM using permuted blocks of two, four, and six stratified by study site.	1. Exposure known for 125 people
		2. For others, exposure imputed based on the following rules:
		  a. BVM or BiPAP during first attempt without prolonged intubation -> assigned intervention group
		  b. No BVM or BiPAP during first attempt without prolonged intubation -> assigned control group
		  c. For those with prolonged intubation (Top decile of intubation length or >1 attempt) assigned using a predictive model based on known recipients of exposure

Follow-up Period	During RSI to two minutes after tube placement	During RSI to two minutes after tube placement
Causal Contrasts of Interest	Lowest oxygen saturation between induction and two minutes after tracheal intubation Other outcomes: • Severe hypoxemia (SaO2 < 80%)	Lowest oxygen saturation between induction and two minutes after tracheal intubation Other outcomes: • Severe hypoxemia (SaO2 < 80%)

	• Lowest SaO2, highest FiO2, and highest PEEP between 6-24 hours • Operator reported aspiration • New opacity within 48 hours	• Lowest SaO2, highest FiO2, and highest PEEP between 6-24 hours • Operator reported aspiration
Analysis Plan	Primary Analysis: Mann-Whitney U Test	Primary Analysis: Mann-Whitney U Test on propensity score matched population

Pre-Registered Results

Table 2: Patient Characteristics at Baseline – Observational Cohort

Patient Characteristic	**Positive Pressure Ventilation (n=180)**	**No Ventilation (n=180)**	**p-value**
Age, median [IQR], years	60 (49-68)	56 (48-68)	0.41
Male sex, No. (%)	106 (58.9)	102 (56.7)	0.67
White race, No. (%)	140 (77.8)	140 (77.8)	0.97
Body mass index, median [IQR], kg/m^2	28.1 (23.4-33.4)	27.4 (23.5-32.2)	0.57
APACHE II score, median [IQR]	20.5 (16-26)	21 (16-25)	0.96
Vasopressors, No. (%)	27 (15.0)	42 (23.3)	0.05
Active medical conditions, No. (%)			
Sepsis or septic shock	98 (54.4)	97 (53.9)	0.92
Gastrointestinal bleeding	21 (11.7)	33 (18.3)	0.08
Indications for intubation, No. (%)			
Hypoxemic respiratory failure	103 (57.2)	101 (56.1)	0.83
Hypercarbic respiratory failure	23 (12.8)	25 (13.9)	0.76
Airway protection for decreased level of consciousness	60 (33.3)	60 (33.3)	1.0
BiPAP in prior 6 hours, No. (%)	71 (39.4)	37 (20.6)	<0.001
Highest FiO_2 in prior 6 hours, median [IQR]	0.4 (0.3-0.7)	0.4 (0.3-0.8)	0.88
Lowest oxygen saturation in prior 6 hours, median [IQR], %	91 (88-94)	92 (89-95)	0.27

Table 3: Outcomes after Tracheal Intubation – Observational Cohort

Outcomes	Positive Pressure Ventilation (n=180)	No Ventilation (n=180)	Relative Risk or Mean Difference (95% Confidence Intervals)
Primary Outcome			
Lowest oxygen saturation, median [IQR], %	94 (86-98)	93 (81.5-99)	1.81 (-0.97, 4.59)
Secondary Outcome			
Lowest oxygen saturation < 80%, No. (%)	25 (13.9)	42 (23.3)	0.60 (0.38-0.93)
Exploratory Oxygen Saturation Outcomes			
Lowest oxygen saturation < 90%, No. (%)	59 (32.8)	66 (36.7)	0.89 (0.67-1.19)
Lowest oxygen saturation < 70%, No. (%)	12 (6.7)	19 (10.6)	0.63 (0.32-1.26)
Decrease in oxygen saturation, median [IQR], %	3 (0-11)	3 (0-13)	-1.86 (-4.50, 0.78)
Exploratory Safety Outcomes			
Operator-reported aspiration, No. (%)	3 (1.7)	1 (0.6)	3.00 (0.32-28.57)
Cardiac arrest within 1 hour after intubation, No. (%)	4 (2.2)	2 (1.1)	2.00 (0.37-10.78)
Exploratory Clinical Outcomes			
Ventilator-free days, median [IQR]	15 (0-25)	15 (0-25)	-0.53 (-3.01, 1.95)
ICU-free days, median [IQR]	10 (0-23)	10 (0-22)	0.12 (-2.17, 2.40)
Died before hospital discharge, No. (%)	72 (40.0)	72 (40.0)	1.00 (0.78-1.29)

Figure 1 – CONSORT Diagram

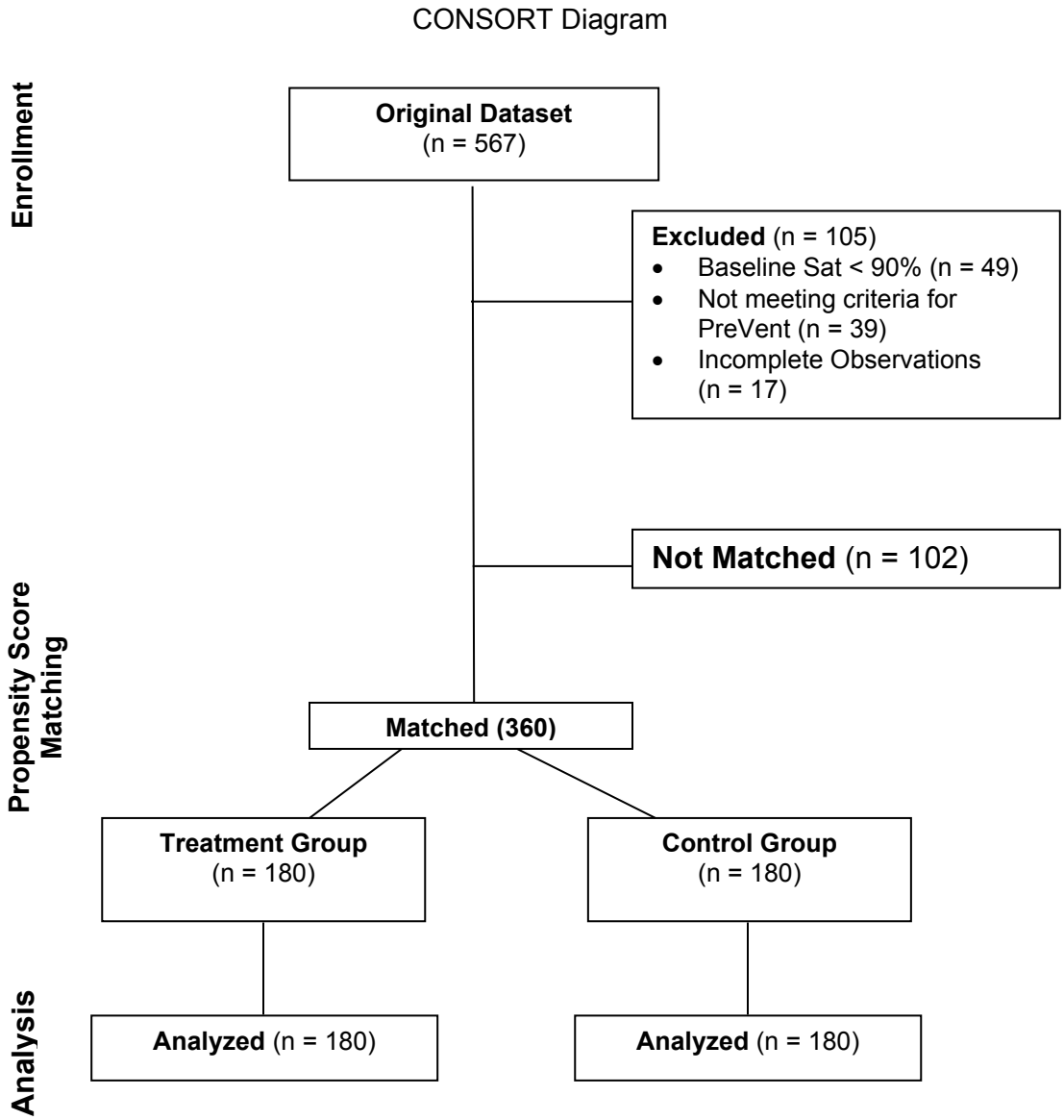
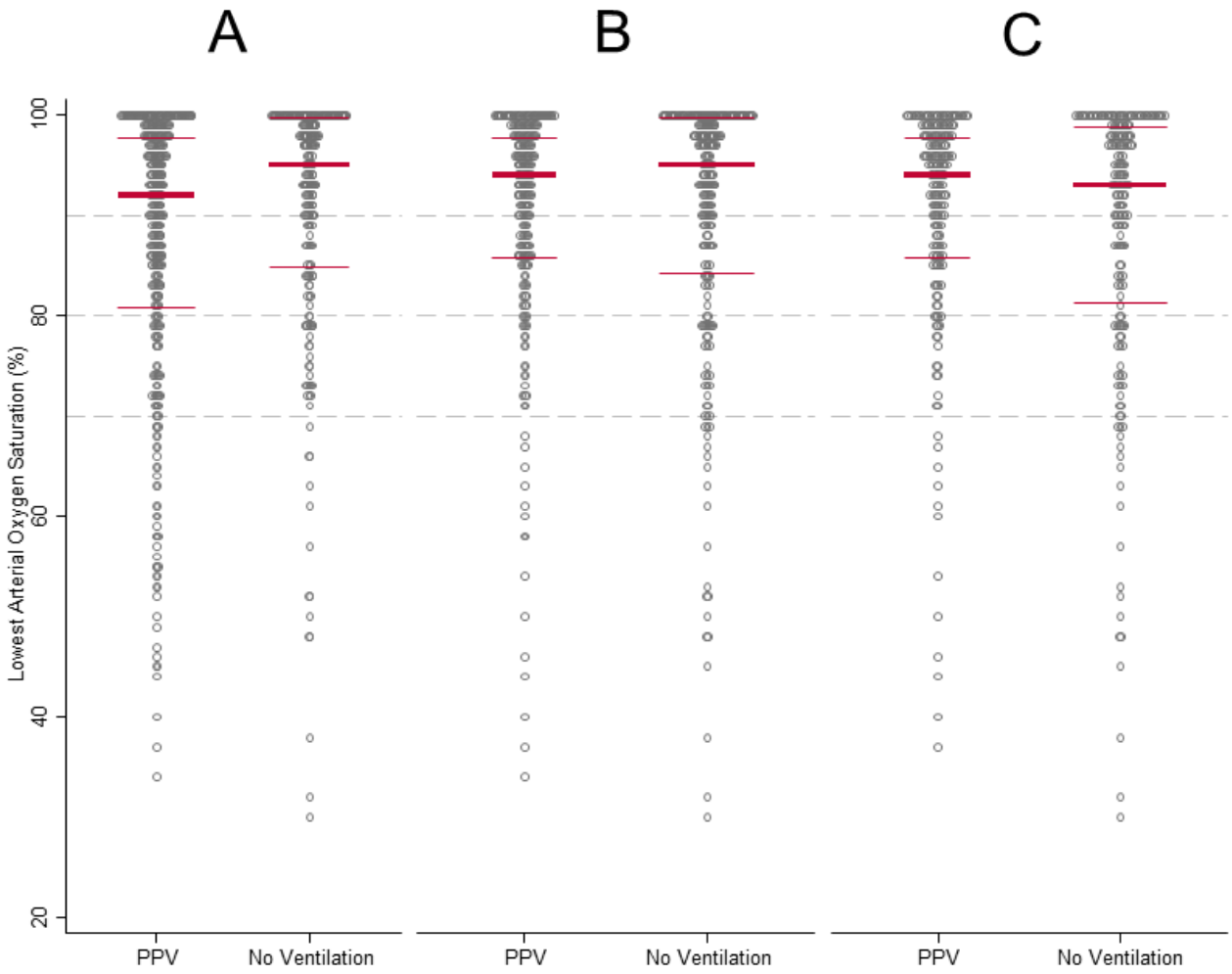


Figure 2 – Lowest arterial oxygen saturation in treatment versus control groups.



A reflects unadjusted results. B reflects results after treatment assignment procedure. C reflects results after propensity score matching.

Pre-Registered Statistical Code

```
/*************************************************************/
/*************************************************************/
/*************************************************************/
/*TARGET TRIAL EMULATION*/
/*2/17/19*/
/*VERSION 2.0 (Pre-Trial Release)*/
/*ADMON AND DONNELLY*/
/*************************************************************/
/*************************************************************/
/*************************************************************/

cd "C:\Users\joond\Desktop\UMich\Target Trial Emulation\"
set more off

version 14.2

set seed 43452

/*IMPORT*/
import delimited using Database_without_PreVent_Lincoln_ED_for_UM.csv

/*MAKE EXCLUSION BASED ON SAT AND PREVENT STATUS*/
gen baselinesat90 = .
replace baselinesat90=0 if baselinesat_dcs<90
replace baselinesat90=1 if baselinesat_dcs>=90 & baselinesat_dcs<.

gen exclude = 0
replace exclude=1 if baselinesat90==0 | prevent_nonenroll!="Center not participating in trial"

keep if exclude==0

/*GENERATE STARTING TREATMENT VARIABLE*/
/*OVERRIDE WITH BMV_INDUCTION FOR PATIENTS WHO WE KNOW DID NOT RECEIVE TREATMENT STARTING AT
INDUCTION*/

gen bvmbipap_yesno =.
replace bvmbipap_yesno=0 if ventilation_induct_laryng_none==1
replace bvmbipap_yesno=1 if ventilation_induct_laryng_none==0
replace bvmbipap_yesno=0 if bmv_induction==0

/*TAG POPULATION TO PREDICT EXPOSURE*/
/*DERIVATION POPULATION: ONE ATTEMPT WITH SHORT DURATION OR KNOWN TREATMENT OR KNOWN NO
TREATMENT*/
gen sample_tag=0
replace sample_tag=1 if (attempts_dcs==1 & duration_total_seconds<=215) | bmv_induction!=. |
bvmbipap_yesno==0

/*GENERATE OUTCOME VARIABLES*/
gen hypox80 = .
replace hypox80=0 if nadirsat_dcs>=80 & nadirsat_dcs<=.
replace hypox80=1 if nadirsat_dcs<80

gen hypox90 = .
replace hypox90=0 if nadirsat_dcs>=90 & nadirsat_dcs<=.
replace hypox90=1 if nadirsat_dcs<90

gen hypox70 = .
replace hypox70=0 if nadirsat_dcs>=70 & nadirsat_dcs<=.
replace hypox70=1 if nadirsat_dcs<70

gen sat_decrease = baselinesat_dcs - nadirsat_dcs

*aspiration_new_dcs

gen carrest1hrintub = 0
replace carrest1hrintub = 1 if cardiacarrest1hr==1 | dcscomplicationscardiacarrest==1

*vfds
*icufds
```

```
*died_in_hospital

/*CHECK VARIABLE AVAILABILITY ACROSS TRIALS*/
set more off
codebook indication_hypoxic indication_hypercarbic indication_hypercarb_hypox indication_ams
indication_airwaycompromise indication_resparrest activesepsis activesepticshock
activecardiogenicshock activedistributiveshock activeneurogenicshock activehemorrhagicshock
activehypovolemicshock activegibleeding uppergibleed lowergibleed activehepaticencephalopathy
activedelirium activecopdexacerbation activeasthmaexacerbation activecfexacerbation
activeildexacerbation activestemi activenstemi activeunstableangi activedrugoverdose
activepancreatitis campriorintubaiton lowmapsixhoursprior pressorssixhoursprior pressors_6hrs_epi
pressors_6hrs_norepi pressors_6hrs_phenyl pressors_6hrs_dopa pressors_6hrs_dobuta
pressors_6hrs_milrinone prior6hbipap prior6hfio2high prior6hspo2low prior6habg reintubation_24h
reintubation_72h preoxnone preoxnrb preoxbipap preoxbmv preoxnc preoxhfnc nmbrocuronium
nmbvecuronium nmbsuccinylcholine nmbcisatracurium anynmb induction_ativan induction_dilaudid
induction_etomidate induction_fentanyl induction_ketamine induction_lidocaine induction_propofol
induction_versed tubesize laryngoscopetype laryngoscopesize laryngoscopeblade tubetapelevel
confirmauscultation confirmetco2 confirmbronch grade_view_emr difficultystarpanel bougie_emr
rescuedevice_emr nd_proceduralist_emr complications_none_emr fail_to_intubate_emr
complications_aspiration_emr complications_equip_fail_emr complications_bleeding_emr
complications_laryngospasm_emr da_emesis_emr da_aspiration_emr da_ugib_emr da_epistaxis_emr
da_airway_mass_emr da_hn_radiation_emr da_limitedneckmobility_emr da_limitedmouthopen_emr
cardiacarrest10min enroll_prepare prepare_nonenroll prevent_nonenroll aoassigned
deviceassigneddcs baselinesat_dcs duration_total_seconds nadirsat_dcs experience_dcs
fellowtrainmonths deviceexperience_dcs bmv_ever_dcs ventilation_induct_laryng_none ao_none_dcs
ao_nc bipap_induc_laryng_dcs device_dcs grade_view_dcs attempts_dcs bougie_dcs
addition_equip_vl_dcs addition_equip_dl_dcs additional_equp_lma_dcs additional_equip_bronch
secondproceduralist_dcs aspiration_new_dcs complications_sbp80_dcs dcscomplicationscardiacarrest
dcscomplicationsesophagealintuba dcscomplicationsairwaytrauma operatorme specialty
intubationmonth monthofstudy nightintubation diedever died_1hr died_icu died_in_hospital vfds
iculos icufds

/*RECODE VARIABLES TO NUMERIC*/
foreach var of varlist trialgroup studysite gender race aoassigned position_randomized {
encode `var', gen(`var'2)
}

/*GENERATE PREDICTION AND ADJUSTMENT VARIABLES*/

gen race3cat = 3
replace race3cat = 1 if race2==4
replace race3cat = 2 if race2==2

gen difficultystarpanel2 = .
replace difficultystarpanel2 = 1 if difficultystarpanel == "Difficult"
replace difficultystarpanel2 = 2 if difficultystarpanel == "Easy"
replace difficultystarpanel2 = 3 if difficultystarpanel == "Moderate"

gen intervention_group = 5
replace intervention_group=1 if aoassigned2==1
replace intervention_group=2 if aoassigned2==2
replace intervention_group=3 if position_randomized2==1
replace intervention_group=4 if position_randomized2==2

egen studysite_ivgroup = group(studysite intervention_group)
tab studysite_ivgroup bvmbipap_yesno if sample_tag==1
*COMBINE ICUS AT SINGLE SITE DUE TO SMALL NUMBERS
replace studysite2 = 5 if studysite2==6

gen any_intervention_grp = 0
replace any_intervention_grp=1 if position_randomized2==1 | aoassigned2==1 |
fluidbolus_preintub_dcs==1

gen preoxpp_group =0
replace preoxpp_group = 1 if preoxbipap==1 | preoxbmv==1
replace preoxpp_group = . if preoxbipap==. | preoxbmv==.

/*SAVE DATA SET WITHOUT NEW TREATMENT VARIABLE*/
save emulation_cohort_premodel.dta, replace

/*USE MODEL IN PATIENTS WITH KNOWN TREATMENT STATUS TO PREDICT PROBABILITIES OUT OF SAMPLE*/
```

```
logistic bvmbipap_yesno i.studysite2 any_intervention_grp ///
age bmi apacheiiscore i.race3cat i.gender2 ///
indication_hypoxic indication_hypercarbic indication_ams ///
activesepsis activegibleeding activecopdexacerbation ///
prior6hbipap prior6hfio2high pressorssixhoursprior ///
baselinesat_dcs preoxpp_group if sample_tag==1
lroc
predict p
keep if p!=.

keep studyid p nadirsat_dcs bvmbipap_yesno sample_tag studysite2 age bmi apacheiiscore race3cat
gender2 indication_hypoxic any_intervention_grp ///
indication_hypercarbic indication_ams activesepsis activegibleeding activecopdexacerbation ///
prior6hbipap prior6hfio2high prior6hspo2low baselinesat_dcs intervention_group preoxbmv
preoxbipap bmv_induction exclude hypox80 bmv_ever_dcs preoxpp_group pressorssixhoursprior ///
hypox90 hypox70 sat_decrease aspiration_new_dcs carrest1hrintub vfds icufds died_in_hospital

/*ALTERNATIVE STRATEGY BASED ON MULTIPLE IMPUTATION*/
/*NEED VALUES FOR THIS ANALYSIS SO NEEDED TO IDENTIFY MOST COMMON VALUE ACROSS IMPUTATIONS*/
/*RESULTED IN EXACT SAME TREATMENT ASSIGNMENT*/
/*FOR SUBSEQUENT ANALYSES WILL CONSIDER POOLING*/
/*
gen bvmbipap_yesno_im = bvmbipap_yesno
replace bvmbipap_yesno_im = . if sample_tag==0
mi set flong
mi register imputed bvmbipap_yesno_im
mi register regular studysite2 any_intervention_grp /*i.intervention_group*/ ///
age bmi apacheiiscore race3cat gender2 ///
indication_hypoxic indication_hypercarbic indication_ams ///
activesepsis activegibleeding activecopdexacerbation ///
prior6hbipap prior6hfio2high /*prior6hspo2low*/ pressorssixhoursprior baselinesat_dcs
preoxpp_group
mi impute chained (logit) bvmbipap_yesno_im = i.studysite2 any_intervention_grp
/*i.intervention_group*/ ///
age bmi apacheiiscore i.race3cat i.gender2 ///
indication_hypoxic indication_hypercarbic indication_ams ///
activesepsis activegibleeding activecopdexacerbation ///
prior6hbipap prior6hfio2high /*prior6hspo2low*/ pressorssixhoursprior baselinesat_dcs
preoxpp_group, add(1000) rseed(6068)

gsort - sample_tag + studyid + _mi_m
order p, last

drop if _mi_m==0

egen total_pos = sum(bvmbipap_yesno_im), by(_mi_id)

gen bvmbipap_yesno_im2=.
replace bvmbipap_yesno_im2 = 0 if total_pos<500
replace bvmbipap_yesno_im2 = 1 if total_pos>=500

mi extract 1, clear
*/

/*GENERATE NEW TREATMENT VARIABLE BASED ON PREDICTED PROBABILITIES*/
gen bvmbipap_yesno2 = bvmbipap_yesno
replace bvmbipap_yesno2 =0 if sample_tag==0 & p<0.5

/*BASE MODEL*/
/*BASED ON PRIOR PAPER IN SAME POPULATION LOOKING AT RISK FACTORS*/
/*SOME BALANCE PROBLEMS*/
/*
logistic bvmbipap_yesno2 i.studysite2 any_intervention_grp ///
age bmi i.race3cat ///
indication_hypoxic indication_hypercarbic baselinesat_dcs
lroc
*/

/*FINAL MODEL*/
logistic bvmbipap_yesno2 i.studysite2 any_intervention_grp ///
age bmi i.race3cat i.gender2 ///
```

```
indication_hypoxic indication_hypercarbic activesepsis activecopdexacerbation prior6hfio2high ///
baselinesat_dcs
lroc

predict pa

cem pa, tr(bvmbipap_yesno2) k2k

tabstat pa if cem_matched==1, by(bvmbipap_yesno2) stat(mean sd p50 p25 p75)
pstest baselinesat_dcs studysite2 any_intervention_grp age bmi apacheiiscore race3cat gender2
indication_hypoxic indication_hypercarbic indication_ams activesepsis activecopdexacerbation
prior6hfio2high, t(bvmbipap_yesno2) raw graph
pstest baselinesat_dcs studysite2 any_intervention_grp age bmi apacheiiscore race3cat gender2
indication_hypoxic indication_hypercarbic indication_ams activesepsis activecopdexacerbation
prior6hfio2high if cem_matched==1, t(bvmbipap_yesno2) raw graph

/*ALTERNATIVE MATCHING STRATEGY BASED ON CALIPER WIDTH*/
/*SEEMS TO BE MORE UNBALANCED*/
/*
sum pa, det
gen logitpa = logit(pa)
egen sdlogitpa = sd(logitpa)
gen calip = 0.25*sdlogitpa
gen calip_alt = sqrt(sdlogitpa)/4
di calip calip_alt
psmatch2 bvmbipap_yesno2, caliper() pscore(pa) noreplacement

tabstat nadirsat_dcs pa if _weight==1, by(bvmbipap_yesno2) stat(mean sd p50 p25 p75)
ranksum nadirsat_dcs if _weight==1, by(bvmbipap_yesno2)
tab hypox80 bvmbipap_yesno2 if _weight==1, col chi2 exact

pstest baselinesat_dcs studysite2 any_intervention_grp age bmi apacheiiscore race3cat gender2
indication_hypoxic indication_hypercarbic indication_ams activesepsis activecopdexacerbation
prior6hfio2high, t(bvmbipap_yesno2) raw graph
pstest baselinesat_dcs studysite2 any_intervention_grp age bmi apacheiiscore race3cat gender2
indication_hypoxic indication_hypercarbic indication_ams activesepsis activecopdexacerbation
prior6hfio2high if _weight==1, t(bvmbipap_yesno2) raw graph
*/

/*SAVE FINAL ANALYTIC DATASET*/
save emulation_cohort_analytic.dta, replace
clear

/*ANALYSES*/
set more off
import delimited using Database_without_PreVent_Lincoln_ED_for_UM.csv, clear

gen bvmbipap_yesno =.
replace bvmbipap_yesno=0 if ventilation_induct_laryng_none==1
replace bvmbipap_yesno=1 if ventilation_induct_laryng_none==0

label define vent 0 "No Ventilation" 1 "PPV"
label values bvmbipap_yesno vent
revrs bvmbipap_yesno, replace

dotplot nadirsat_dcs, over(bvmbipap_yesno) mcolor(gs7) center median bar nogroup msize(small)
mlwidth(vvthin) mfcolor(none) name(g1, replace) ///
ytitle("Lowest Arterial Oxygen Saturation (%)", size(small)) xtitle("") ylabel(, labsize(small))
xlabel(, labsize(small)) graphregion(color(white)) bgcolor(white)  yline(90 80 70, lw(vvthin)
lcolor(gs12) lpattern(dash))
gr_edit plotregion1.plot4.style.editstyle marker(fillcolor(cranberry)) editcopy
gr_edit plotregion1.plot4.style.editstyle marker(linestyle(color(cranberry))) editcopy
gr_edit plotregion1.plot4.style.editstyle marker(size(vsmall)) editcopy
gr_edit plotregion1.plot4.style.editstyle marker(symbol(smsquare)) editcopy
gr_edit plotregion1.plot3.style.editstyle label(textstyle(color(cranberry))) editcopy
gr_edit plotregion1.plot2.style.editstyle label(textstyle(color(cranberry))) editcopy
clear

use emulation_cohort_analytic.dta

/*TABLE 1*/
```

```
tabstat age bmi apacheiiscore prior6hfio2high prior6hspo2low if cem_matched==1,
by(bvmbipap_yesno2) stat(n p50 p25 p75)
foreach var of varlist age bmi apacheiiscore prior6hfio2high prior6hspo2low {
reg `var' bvmbipap_yesno2 if cem_matched==1
}
foreach var of varlist gender race3cat activesepsis activegibleeding pressorssixhoursprior
indication_hypoxic indication_hypercarbic indication_ams prior6hbipap {
tab `var' bvmbipap_yesno2 if cem_matched==1, col chi2
}

/*TABLE 2*/
tabstat nadirsat_dcs sat_decrease vfds icufds if cem_matched==1, by(bvmbipap_yesno2) stat(n p50
p25 p75)
foreach var of varlist nadirsat_dcs sat_decrease vfds icufds {
reg `var' bvmbipap_yesno2 if cem_matched==1
}

foreach var of varlist nadirsat_dcs sat_decrease vfds icufds {
ranksum `var' if cem_matched==1, by(bvmbipap_yesno2)
}

foreach var of varlist hypox80 hypox90 hypox70 aspiration_new_dcs carrest1hrintub
died_in_hospital {
tab `var' bvmbipap_yesno2 if cem_matched==1, col chi2 exact
glm `var' bvmbipap_yesno2 if cem_matched==1, link(log) fam(bin) eform
}

/*DOTPLOT FIGURE*/
label define vent 0 "No Ventilation" 1 "PPV"
label values bvmbipap_yesno2 vent
revrs bvmbipap_yesno2, replace

dotplot nadirsat_dcs, over(bvmbipap_yesno2) mcolor(gs7) center median bar nogroup msize(small)
mlwidth(vvthin) mfcolor(none) name(g2, replace) ///
yscale(off) xtitle("") ylabel(, labsize(small)) xlabel(, labsize(small))
graphregion(color(white)) bgcolor(white) yline(90 80 70, lw(vvthin) lcolor(gs12) lpattern(dash))
gr_edit plotregion1.plot4.style.editstyle marker(fillcolor(cranberry)) editcopy
gr_edit plotregion1.plot4.style.editstyle marker(linestyle(color(cranberry))) editcopy
gr_edit plotregion1.plot4.style.editstyle marker(size(vsmall)) editcopy
gr_edit plotregion1.plot4.style.editstyle marker(symbol(smsquare)) editcopy
gr_edit plotregion1.plot3.style.editstyle label(textstyle(color(cranberry))) editcopy
gr_edit plotregion1.plot2.style.editstyle label(textstyle(color(cranberry))) editcopy
dotplot nadirsat_dcs if cem_matched==1, over(bvmbipap_yesno2) mcolor(gs7) center median bar
nogroup msize(small) mlwidth(vvthin) mfcolor(none) name(g3, replace) ///
yscale(off) xtitle("") ylabel(, labsize(small)) xlabel(, labsize(small))
graphregion(color(white)) bgcolor(white) yline(90 80 70, lw(vvthin) lcolor(gs12) lpattern(dash))
gr_edit plotregion1.plot4.style.editstyle marker(fillcolor(cranberry)) editcopy
gr_edit plotregion1.plot4.style.editstyle marker(linestyle(color(cranberry))) editcopy
gr_edit plotregion1.plot4.style.editstyle marker(size(vsmall)) editcopy
gr_edit plotregion1.plot4.style.editstyle marker(symbol(smsquare)) editcopy
gr_edit plotregion1.plot3.style.editstyle label(textstyle(color(cranberry))) editcopy
gr_edit plotregion1.plot2.style.editstyle label(textstyle(color(cranberry))) editcopy
graph combine g1 g2 g3, ycommon row(1) name(combined, replace) graphregion(color(white))
imargin(small)
graph export Trial_Emulation_Dotplot_2172019.png, replace
graph export Trial_Emulation_Dotplot_2172019.pdf, replace
```