Severe Maternal Morbidity: Screening and Long-Term Consequences

by

Abigail Rebecca Cartus

BS, University of Pittsburgh, 2010

MPH, University of Texas Health Science Center, 2016

Submitted to the Graduate Faculty of the

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2020

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Abigail Rebecca Cartus

It was defended on

November 23, 2020

and approved by

Katherine P. Himes, Assistant Professor, Obstetrics, Gynecology, and Reproductive Services, Magee-Womens Research Institute

Marian Jarlenski, Associate Professor, Health Policy and Management

Ashley I. Naimi, Assistant Professor, Epidemiology, Emory University

Thesis Advisor/Dissertation Director: Lisa M. Bodnar, Professor, Epidemiology

Severe Maternal Morbidity: Screening and Long-Term Consequences

Abigail Rebecca Cartus, PhD

University of Pittsburgh, 2020

Severe maternal morbidity (SMM) is an important population indicator of maternal health and health care quality. SMM is often used as a proxy for maternal death in epidemiologic and quality improvement research. This is because it is much more common than maternal death (which is rare in absolute numbers), but shares the same risk factors and etiologies. However, there are two unresolved questions in severe maternal morbidity research, which this dissertation addresses. First, the long-term health consequences of SMM are not well understood. I investigated the association between SMM during the perinatal period and risk of adverse cardiovascular events (heart failure, ischemic heart disease, stroke/transient ischemic attack, and a composite of these three outcomes plus atrial fibrillation) up to 2 years postpartum among deliveries covered by Pennsylvania Medicaid, 2016-2018. I found that SMM is associated with increased risk of adverse cardiovascular events and that elevated risk persists past the traditional end of the postpartum period at 42 days post-delivery. Second, available methods to quantify SMM at the hospital or population level have serious limitations, *e.g.*, identifying a large number of false-positive cases or requiring labor-intensive medical record abstraction, that I attempted to address using ensemble machine learning. To this end, I examined the impact of undersampling, one technique for remedying outcome class imbalance (where non-events outnumber events by a factor of 2:1 or more), on the predictive performance of ensemble machine learning algorithms (SuperLearner). We found that, in a simulated setting with moderate class imbalance, undersampling does not markedly improve the predictive performance of either logistic

regression or SuperLearner. We then attempted to use SuperLearner as an alternative to existing screening criteria or medical record review to identify true-positive SMM from a sample of deliveries at Magee-Womens Hospital, 2013-2017. Our SuperLearner algorithms performed better than existing SMM screening criteria on some predictive performance metrics and worse on others, indicating that the choice of SMM screening method involves tradeoffs. This work contributes to improved understanding of maternal health in the United States and points to several future directions for SMM research.

Table of Contents

List of Tables

List of Figures

1.0 Chapter 1: Proposal

The rate of maternal deaths in the United States has been steadily increasing since the late 1980s, such that the US now has the highest maternal mortality rate of any high-income country. However, because maternal deaths are rare even in settings with comparatively high mortality rates, health authorities including the World Health Organization (WHO) recommend research into severe maternal morbidity (SMM) events, 50 times more numerous than deaths, to better understand and ultimately prevent maternal mortality. Though it is well established that SMM shares similar causes, risk factors, and temporal trends with maternal mortality, important gaps in our knowledge of SMM must be addressed to advance maternal health research. Specifically, it is difficult to accurately identify SMM cases using existing methods, and it is not known whether surviving a SMM confers a higher risk for longer-term health impacts. While understanding and reducing SMM is critical to achieving reductions in maternal death and improvements in maternal health, research is needed to fill gaps in our current understanding.

The long-term goal of this project is to improve maternal health research and practice by better understanding SMM. In order to do this, we propose to address two major research needs. First, SMM is frequently studied by identifying cases and non-cases from administrative data using screening criteria. However, these criteria reflect clinical judgment and historical practice rather than empirical evidence and as such frequently fail to accurately designate case status. As a consequence, existing screening methods may generate inaccurate SMM prevalence estimates, which may in turn lead to incorrect inferences about the population risk factors for SMM and the areas where resources should be applied to hospital quality improvement. Second, it is also

important to determine the postpartum burden of SMM. Little research to date has established the long-term risk of adverse outcomes after surviving a SMM; however, such knowledge is critical to understand the total burden of SMM. Our overall objectives are to improve screening for SMM in administrative data and to determine the cardiovascular sequelae of SMM events after delivery. To this end, our specific aims are:

Specific Aim 1. To develop a screening algorithm with high sensitivity and specificity to identify SMM cases from administrative data.

The American College of Obstetricians and Gynecologists recommends a "screen and review" approach to SMM identification, whereby deliveries that screen positive for SMM qualify for medical record review. Although medical record review is the most reliable way to identify SMM cases, deliveries are too numerous to review comprehensively. Consequently, screening is a critical tool for SMM research, but existing screening criteria used to identify SMM demonstrate variable but overall modest positive predictive value. We will use a subset of the Magee Obstetric Maternal and Infant (MOMI) Database (2010-2011, n = 685) to train an ensemble machine learning algorithm to identify SMM cases, then validate the algorithm on a temporally distinct subset of the same database (2013-2017, n = 498). MOMI is a rich administrative database containing maternal demographic, behavioral, and medical characteristics, labor and delivery information, and fetal/infant outcomes. We hypothesize that machine learning methods will perform screening for SMM better than extant criteria, which involve few variables and have not been extensively validated.

<u>Specific Aim 2.</u> To determine the postpartum consequences of SMM in a Medicaid population.

Although the postpartum period is increasingly recognized as a critical component of maternal health, little is known about whether US women who suffer a SMM but do not die have more adverse outcomes in the postpartum period than women who do not. Using data from Pennsylvania Medicaid (2016-2018, N = 137,140 deliveries), we will assess the risk of several adverse cardiovascular outcomes (heart failure, ischemic heart disease, and stroke/transient ischemic attack) following deliveries with SMM compared to deliveries without SMM over the 24 months postpartum. We hypothesize that risk of adverse cardiovascular events will be higher among women with an SMM at any point during pregnancy, labor, delivery, and postpartum.

Successful completion of this work will result in an innovative method for screening for SMM in administrative data and estimates of the long-term impacts of SMM in a Medicaid population. These findings will inform intervention and management strategies to reduce the burden of SMM and maternal death in the US.

1.1 Significance

In this proposal, we will address some major gaps in our knowledge of SMM. First, although surveillance of SMM in the US relies heavily on screening administrative databases for SMM cases, a standard definition of SMM does not exist and existing screening criteria are imprecise. We propose to use novel machine learning methods to exploit the information available in hospital databases and to develop a scalable, adaptable method to more accurately screen for SMM. Second, we will characterize risk of adverse cardiovascular events following SMM in a US

population, as few published studies examine the health profiles of women who experience SMM. In so doing, we will contribute to the development of a more complete picture of the burden of SMM for both patients and society.

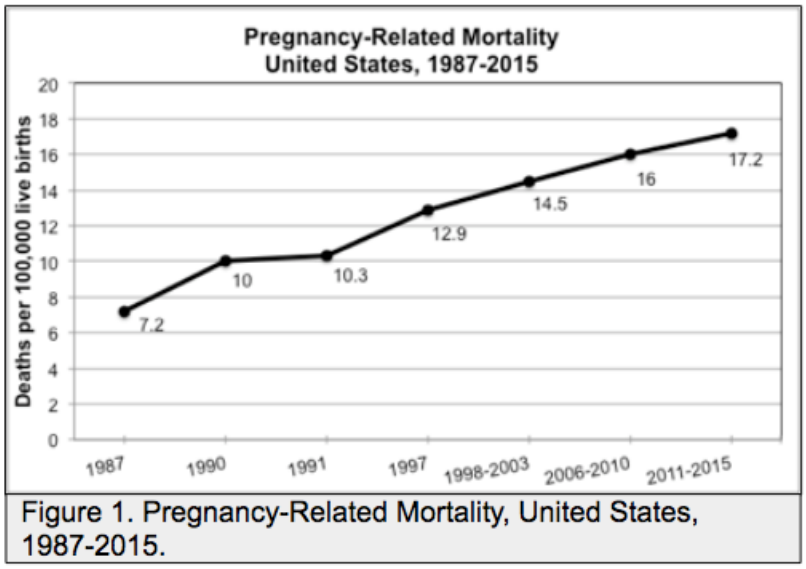

Figure 1.A. Pregnancy-related mortality in the United States, 1987-2015

Maternal mortality, an urgent public health priority, is difficult to research. The rate of maternal death in the United States far exceeds the maternal mortality rates of other industrialized countries.[1-5] For the first time since 2007, the United States recently reported a national maternal mortality rate for the year 2018: 17.4/100,000 live births.[6] This is high compared to other high-income countries: nearly six times higher than the maternal mortality rate of Finland and three times higher than that of Canada.[7] Despite the comparatively high rate, maternal mortality is still a rare outcome in the United States, with Centers for Disease Control (CDC) estimates indicating approximately 700 maternal deaths annually.[8] National maternal mortality estimates are derived from the Pregnancy Mortality Surveillance System (PMSS), a supplemental data system using

vital statistics (death certificates) to collect information about pregnancy-related death.[9] The apparent increase in the US maternal mortality rate shown in Fig. 1[10-14] is evidently attributable to improved ascertainment via the "pregnancy checkbox" on the US standard death certificate.[15,16] Uptake of the death certificate containing the pregnancy checkbox has been uneven, with different states adopting the death certificate at different times, which has made discernment of a temporal trend difficult. While it is not resolved whether the pregnancy checkbox results in improved ascertainment or more false positives, it is likely that the maternal mortality figures reported in previous years were artificially low.[6]

Severe maternal morbidity is a surrogate for maternal mortality. Because maternal deaths are so rare in absolute numbers, it is now widely recommended that SMM be monitored in addition to maternal mortality in order to identify areas of maternal care that require improvement. SMM includes: a.) specific conditions (such as eclampsia or HELLP syndrome, disseminated intravascular coagulation, and sepsis), b.) events (such as acute myocardial infarction and aneurysm), and c.) procedures (including blood transfusion, hysterectomy, and ventilation) that lie on the continuum from uncomplicated delivery to death and which may result in death if not for luck or medical intervention.[17] The causes of SMM and death overlap, but cases of SMM are much more numerous than deaths: for every one maternal death, approximately 50-100 women experience an SMM.[3] The emergence of SMM as an important clinical endpoint, population health target, and health care quality indicator reflects both the similarities between SMM and maternal mortality and the much higher prevalence of SMM relative to mortality.

The epidemiology of SMM is similar to the epidemiology of maternal mortality. This similarity reflects the common causes of each phenomenon. Whether or not the maternal mortality rate has been increasing, the rate of SMM has been steadily increasing in the United States since

the late 1980s.[5] The overall rate of SMM increased from 1013 per 100,000 live births in 2006 to 1466 per 100,000 live births in 2015.[4] Over the same period, the pregnancy mortality rate increased from 14.5 to 17.2 per 100,000 live births.[8,13,14] Similar racial/ethnic disparities in the rates of maternal mortality and SMM are also evident. Black/African American women are 3.3 times more likely to die as a result of pregnancy, labor, and delivery than white women[8] and are more than twice as likely to experience an SMM as white women.[4,5,18-20] Risk factors for SMM and maternal death are largely the same, and include older and younger maternal age, maternal obesity, mode of delivery, diabetes, and hypertension.[21-25] Finally, as with maternal deaths, over half of severe maternal morbidities are considered preventable.[8,26,27]

1.1.1 Improving methods for SMM case identification for research

Approaches to defining and identifying SMM may be intended for quality improvement or research purposes, and all rely on some screening criteria to identify true SMM cases. We will briefly review the evolution and current state of each of two dominant approaches to SMM identification, the World Health Organization (WHO) maternal near miss tool and the so-called "CDC criteria" in the United States, and highlight some practical challenges to SMM research arising from the disjointed development of SMM research practices across space and time. These challenges include the absence of reliable indicators of severity in readily available data, the absence of "gold standard" screening criteria for SMM, the heterogeneity of SMM, and the necessity of dichotomizing a continuous spectrum of morbidities into "severe" and "not severe" categories.

1.1.1.1 Development and validation of the World Health Organization Near Miss Tool

Low and declining rates of maternal death in industrialized countries through the latter half of the 20th century prompted interest in SMM audits to guide quality improvement in maternal care.[28-32] The concept of a "near miss" for death[30] laid the foundation for the development of subsequent research into SMM.

A 2004 World Health Organization (WHO) systematic review synthesized 30 studies of near miss morbidity and classified them according to reliance on organ system dysfunction-based criteria, management-based criteria (emergency hysterectomy, ICU admission), or disease-specific criteria (presence of a specific condition) to identify SMM. On the basis of the review findings, the investigators concluded that organ system dysfunction-based criteria for identifying SMM are the most flexible and accurate.[33] Findings from this systematic review informed development of a "WHO standard" definition of near miss SMM (2009)[34] ("a woman who nearly died but survived a complication that occurred during pregnancy, childbirth, or within 42 days of termination of pregnancy") and a set of criteria for identifying near miss cases, known as the WHO maternal near miss tool (Table 1).[34] The criteria, developed to be feasible and yield comparable estimates of SMM prevalence across settings, follow the categories developed in the systematic review and consist of disease-based (clinical), organ system dysfunction-based (laboratory), and management-based indicators[34] (Table 1).

Table 1. The WHO Maternal Near-Miss Criteria

Clinical criteria	
Acute cyanosis	Gasping
Loss of consciousness lasting ≥12 hours	Respiratory rate >40 or <6/ min
Loss of consciousness AND absence of pulse/hearbeat	Shock
Stroke	Oliguria non-responsive to fluids or diuretics

Uncontrollable fit/total paralysis	Clotting failure
Jaundice in the presence of preeclampsia	Loss of consciousness lasting ≥12 hours
Laboratory-based criteria	
Oxygen saturation <90% for ≥60 minutes	pH<7.1
PaO_2/FiO_2 <200 mmHg	Lactate>5
Creatinine ≥300 µmol/L or ≥3.5 mg/dL	Acute thrombocytopenia (<50,000 platelets)
Bilirubin >100 µmol/L or >6.0 mg/dL	Loss of consciousness AND the presence of glucose and ketoacids in urine
Management-based criteria	
Use of continuous vasoactive drugs	Intubation and ventilation for ≥60 minutes not related to anesthesia
Hysterectomy following infection or hemorrhage	Dialysis for acute renal failure
Transfusion of ≥5 units red cells	Cardio-pulmonary resuscitation

Results of multiple validation studies of the maternal near miss tool suggest that changes to the structure of the tool might be warranted. Two pilot studies of the tool confirmed that near miss indicators predict maternal death[35] and organ failure,[36] and a multicenter validation study found that a summary score derived from the tool accurately predicted maternal death.[37,38] However, three validation studies conducted in low-income settings[37,39,40] found that the WHO maternal near miss tool organ system dysfunction-based criteria are too restrictive and underestimate the number of true cases. Another validation study conducted in a high-income country[41] also found that the organ system dysfunction-based criteria underestimated the number of severe acute maternal morbidity cases.

1.1.1.2 Development of US screening criteria for SMM

The WHO maternal near miss tool has not gained widespread use in North America, possibly because the criteria use information from medical records that may not be available in the claims and billing databases most frequently used to study SMM in the US. In 2016, the American College of Obstetricians and Gynecologists (ACOG) released a consensus document recommending a two-step "screen and review" process for identifying SMM.[42] In the absence of

a gold standard for identifying SMM, chart review is the most reliable way to identify cases. However, since cases of SMM are much more numerous and harder to identify than cases of maternal death, ACOG recommended screening for massive transfusion (≥ 4 units of blood) or ICU admission to identify potential SMM cases that merit full review.[42] Although similar to the WHO disease- and management-based criteria, the ACOG recommendation reflects years of research in North America that has relied on screening for the presence of specific morbidities combined with indicators of severity.

Table 2. Centers for Disease Control and Prevention SMM Indicators

SMM indicators	
Acute myocardial infarction	Pulmonary edema/acute heart failure
Aneurysm	Severe anesthesia complications
Acute renal failure	Sepsis
Adult respiratory distress syndrome	Shock
Amniotic fluid embolism	Sickle cell disease with crisis
Cardiac arrest/ventricular fibrillation	Air or thrombotic embolism
Conversion of cardiac rhythm	Blood products transfusion
Disseminated intravascular coagulation	Hysterectomy
Eclampsia	Temporary tracheostomy
Heart failure/arrest during procedure or surgery	Ventilation
Puerperal cerebrovascular disorders	

Much US research into SMM has used administrative and billing data rather than hospital data, precluding the use of some information that is predictive of SMM such as transfusion volume. This lack of information indicating severity complicated early contributions to maternal morbidity research with claims data, which used antenatal hospitalization[43,44] or any/all complications during the delivery hospitalization[45] as proxies for maternal morbidity. All estimated very high prevalence of maternal morbidity as a consequence of not capturing the *severity* of maternal morbidity.[43-45] In a 2005 study, Wen *et al*. screened Canadian hospital discharges from 1991-2001 using a list of ICD billing codes for diagnoses and Canadian billing codes for procedures in an explicit attempt to identify *severe* maternal morbidity.[46] They reported a rate of 438 severe maternal morbidities

per 100,000 deliveries,[46] much lower than the morbidity rates reported in studies using hospitalization or delivery complications as proxies for SMM.[43-45]

The use of billing codes to identify instances of SMM informed the development of the current CDC criteria (Table 3). Following Wen (2005) and a conceptual model of SMM,[17,46] CDC investigators (2008) devised a list of ICD codes for diagnoses and procedures indicative of severe morbidity (Table 3).[3] They used these ICD codes supplemented with information about length of stay to screen National Hospital Discharge Survey data for SMM, and reported a rate of 510 severe morbidities per 100,000 delivery hospitalizations in the period from 1991 to 2003.[3] In a subsequent paper, Callaghan *et al.* (2012) updated the list of ICD codes for SMM originally proposed in 2008, and, again supplementing these codes with information about length of stay in an attempt to identify women with true severe morbidities, reported prevalences of 1290 severe maternal morbidities per 100,000 delivery hospitalizations and 290 severe maternal morbidities per 100,000 postpartum hospitalizations in 1998-1999.[1] This CDC criteria were updated again in 2015 in anticipation of the transition to ICD-10, resulting in the list of 21 procedures and diagnoses in use today. These screening criteria were not validated by the investigators who developed them, nor in subsequent studies using them to identify SMM.[2,22,23,42,47-49]

Though criteria for identifying SMM have developed differently in the United States than in the rest of the world, there are some important parallels. The inclusion of ICU admission in many screening criteria for SMM reflects the importance of management-based criteria in the WHO maternal near miss tool, and the CDC criteria are similarly analogous to the disease- and management-based criteria. Prolonged postpartum length of stay was explicitly introduced into US screening criteria as an indicator of severity, to augment the billing codes' sensitivity for *severe* morbidity as a way to compensate for some of the inaccuracy of the billing codes.

1.1.1.3 Validation of US screening criteria for SMM

There have been two notable validation studies of the most widely-used US screening criteria. Main *et al*. (2016) conducted a comprehensive validation study of 4 SMM identification criteria: the CDC list of diagnosis and procedure codes, prolonged postpartum length of stay defined as > 3 standard deviations above the mean length of stay for delivery hospitalization in the population (not stratified by mode of delivery), any maternal ICU admission, and any blood transfusion.[50] This study first used the screening criteria to identify likely SMM cases; these candidate cases then underwent full chart review to determine whether they accorded with the SMM "gold standard" (a set of clinical criteria chosen by investigators). Due to this screening-review study design, sensitivity and specificity of the identification criteria should not be reported and only their positive and negative predictive values can be properly interpreted. The 4 criteria exhibited a wide range of positive predictive values, alone and in combination, ranging from 0.38 for any of the CDC list or prolonged postpartum length of stay to 0.88 for massive transfusion (≥4 units) alone.[50] Although negative predictive values were all high, in the range of 0.99-1.00 for all criteria, without reviewing screen-negatives it is unknown how many true cases were incorrectly classified as negatives. This suggests that the probability that a screen-positive SMM case is a true case is highly variable based on these criteria, while the probability that a screen-negative case is a true non-case is unknown.

A more recent study validated three identification criteria used to screen for SMM: the CDC list, ICU admission, or prolonged postpartum length of stay.[51] Using a similar methodology to Main *et al.* (2016),[50] these authors first screened an administrative database (the Magee Obstetric Maternal and Infant Database, or MOMI) for cases, then conducted chart review on screen-positives and a randomly selected equal number of screen-negatives. Again, due to this

design they could only report positive and negative predictive values. They found that 166 of 349 cases that screened positive based on presence of any of the screening criteria were true cases, yielding a positive predictive value of 0.48.[51] They also reviewed screen-negative cases and identified two false-positives (2/349 screen-negative cases), yielding a negative predictive value of 0.99.[51] Taken together, this study and the validation study performed by Main *et al.* (2016)[50] suggest that commonly-used criteria used to identify SMM have difficulty identifying true cases. Furthermore, review of screen-negative cases is important; the rarity of SMM means that even a low false-negative proportion could substantially affect prevalence estimates derived from screening.[51]

1.1.1.4 Applying novel methods to SMM screening.

Important contributions to this field have established that ICU admission and massive blood transfusion are associated with severity[52] of maternal complications and that laboratory indicators of organ system dysfunction are associated with death.[37] However, the empirical basis for determining which ICD codes, diagnoses, procedures, and other characteristics are associated with the severity of maternal morbidity remains sparse. Available SMM identification tools are based on accumulated clinical judgment, and few have undergone formal statistical analysis of their association with severe morbidity or death. As such, these criteria serve the critical purpose of guiding resource allocation for quality improvement, but most are not optimized for routine surveillance or research. Our proposal is responsive to these considerations; we aim to create a broadly applicable method that can accurately identify SMM events in a variety of different data settings.

Some recent literature has applied novel machine learning methods to SMM identification. One group screened delivery records from a large academic hospital using the ACOG/SMFM

criteria, then chart-reviewed screen-positives only, assuming that all screen-negatives were true negatives. These authors then used cross-validated penalized regression models to attempt to predict true-positive and true-negative SMM based on a large number of predictors available in the electronic health record.[53]More of the recent literature on SMM involves risk prediction rather than accurate case identification. Another group used a targeted causal inference approach incorporating ensemble machine learning to quantify the adjusted risk ratios for screen-positive SMM for a wide range of obstetric comorbidities. Their goal was to expand an existing obstetric comorbidity index[54] for use in administrative data to be able to identify women at high risk of screen-positive SMM among all deliveries in California over a two-year period.[55] Another validation study of the original obstetric comorbidity index[56] and a prospective clinical risk prediction model for maternal ICU admission[57] have also been published recently. To our knowledge, our work will be the first to attempt to leverage ensemble machine learning methods to identify true-positive SMM.

Although different resource levels, different case mixes, different case-fatality rates for SMM conditions, and different types of data collected all make establishment of a standard SMM identification procedure difficult,[58] it may be possible to exploit the diversity of information available to accurately identify SMM and expand the number of variables that are used in SMM identification. Traditional regression-based methods have important limitations relevant to this task: they are parametric, imposing functional forms on the relationship between predictors and outcome, and they place constraints on the number of variables that can be effectively evaluated in a single model. A regression-based approach to SMM identification would necessitate specifying a causal structure and model adapted to the specific case mix and other characteristics of each setting. By contrast, machine learning methods offer a promising opportunity to more

easily and accurately identify SMM cases in a variety of settings. These flexible, data-adaptive methods can be applied to any dataset where a number of true case designations are available to "train" a screening algorithm. By incorporating a wider range of potentially predictive information, these methods can fill the need for a scalable, accurate approach to SMM screening. Better identification of SMM cases can, in turn, lead to more accurate SMM prevalence estimates.

1.1.2 The long-term impacts of SMM are not fully understood.

SMM and maternal mortality research has historically focused on the experiences of women who survive a SMM, including the natural history and progression of the morbid condition, the sequence of events leading to the morbidity, the hospital and care procedures involved, and the preventability of the morbidity, and the sequence of events leading to the morbidity.[2,5,21,23,25,42,48,59-63] However, it is also possible that SMM may have longer-term effects or consequences following delivery. A greater burden of complications and a greater need for health care services postpartum are not only critical components of maternal health, but may also confer a greater risk of maternal death following delivery. The postpartum period, including "late" SMM and maternal death, has been the subject of increased focus in maternal health research in recent years, with a consensus bundle[42] developed specifically to improve postpartum care and health outcomes. Our proposal will integrate knowledge of the postpartum consequences of SMM with established knowledge of the antepartum and intrapartum consequences of SMM to better understand the total burden of SMM in the United States.

1.1.2.1 Literature on postpartum health impacts of SMM from low- and middle-income countries.

Many studies of the impacts of maternal morbidity on postnatal health involve less severe or "indirect" complications and morbidities, such as perineal injury, depressive symptoms, and incontinence. 84 of 136 studies included in a 2017 scoping review of the effects of maternal morbidity on postnatal health involved such indirect morbidities.[64] Most of the remaining studies in this review examining "direct" severe or near miss maternal morbidity were conducted in low- or middle-income countries.[64] Storeng *et al.* (2012) evaluated the mortality rate among women with near miss mortality as compared to women with uncomplicated delivery in Burkina Faso and found that among 337 women with near miss morbidity (out of 1014 total participants), 5.3% died within four years postpartum as compared to 0.9% of the women without a near miss morbidity.[65] In a study of 176 women in Morocco, prevalences of both physical and psychological symptoms, especially depression, were higher at 8 months postpartum among women who experienced near miss morbidity (n = 76) than among women who did not (n = 100).[66] In a series of studies published in a Malaysian cohort of 145 women with SMM and 187 without, Norhayati *et al.* reported that SMM was associated with lower self-reported general health[67] and greater functional impairment,[68] but not impaired sexual function[69] or depressive symptoms,[70] at 1 and 6 months postpartum. In a larger Brazilian cohort (n = 368), Silveira *et al.* (2016) described increased functional impairment (*e.g.* difficulty with mobility or household tasks) between 1 and 5 years postpartum among women who experienced a SMM as compared to women who experienced an uncomplicated delivery.[71] Finally, as part of the Cohort of Severe Maternal Morbidity – Multidimensional Evaluation of Long Term Repercussions of SMM (COMMAG) study, Ferreira *et al.* (2020) identified women who experienced SMM according to the WHO near miss criteria

and a random sample of women who did not experience SMM from a tertiary public hospital in Campinas, Brazil. These women were contacted – some up to 5 years after delivery – by telephone to complete the SF-36 questionnaire as well as a questionnaire assessing symptoms of post-traumatic stress disorder (PTSD). The women who experienced SMM had worse self-rated health at the time of assessment than women without.

While the studies briefly reviewed here employed cohort designs with comparison groups, sample sizes are small. Other studies in this field take a detailed approach to examining the consequences of near miss morbidities for women, sometimes incorporating in-depth ethnographic interviews, but do not employ comparison groups and sample sizes are frequently small.[72-74] Consequently, while these studies contribute to a holistic understanding of SMM in its proper social context, they can be difficult to evaluate as epidemiologic literature. A majority of the studies in this area, regardless of design, tend to focus on quality of life and self-reported health as postpartum outcomes; an issue affecting all such studies is the use of questionnaires and other tools to measure these outcomes that are not validated for pregnant or postpartum populations.[64]

1.1.2.2 Studies on postpartum morbidity following SMM in high-income countries.

Few studies have assessed the impact of SMM on postnatal health in high-income settings. In a UK cohort of 1670 women (331 with severe morbidity—hemorrhage, preeclampsia, sepsis, or uterine rupture—and 1339 without) Waterstone *et al.* (2003) reported that, compared to controls, SMM cases scored worse on the Short Form 36 (SF-36) questionnaire assessing general health and reported more sexual difficulties at 6 months postpartum.[75] As is common with studies of the health consequences of SMM, at time of writing the SF-36 was not validated for a general obstetric or postpartum population.[76] Furuta *et al.* (2014) reported an association between SMM and symptoms of post-traumatic stress disorder (PTSD) in a UK cohort.[77] Using disease-based

(*e.g.* hemorrhage, preeclampsia) and management-based (*e.g.* ICU admission) criteria to identify SMM cases, these authors surveyed 1823 women at 6-8 weeks postpartum and found that women with SMM were more likely than women with uncomplicated delivery to report two symptoms of PTSD, intrusion (aOR [95% CI] = 2.11 [1.17, 3.78]) and avoidance (aOR [95% CI] = 3.28 [2.01-5.36]). Finally, Lewkowkitz *et al.* (2019) used the CDC SMM identification criteria in a cohort of Florida deliveries and concluded that SMM was associated with increased risk of severe psychiatric morbidity (a composite) and increased risk of substance used disorder after delivery within one year of hospital discharge, particularly in the 4 months following delivery.[78] The studies conducted in high-income countries reviewed here all use different definitions of SMM, and none account for timing of SMM relative to delivery.

1.1.2.3 Cardiovascular sequelae of SMM may be particularly important.

SMM is a heterogeneous group, but hypertensive and cardiovascular SMM complications are emerging as important drivers of SMM and maternal death over the past 10-15 years. A recent Report from Nine Maternal Mortality Review Commissions estimated that 50% of maternal deaths are caused by hemorrhage, infection, cardiovascular conditions, and cardiomyopathy.[79] Hypertensive disorders of pregnancy are major contributors to SMM. One retrospective study found that, compared to no hypertension in pregnancy, preeclampsia with severe features is strongly associated with SMM at delivery (aOR (95% CI) 5.4 (3.9-7.3)).[62] Another study reported that eclampsia, compared to no hypertension, was strongly associated with risk of SMM (aOR (95% CI) 13 (7.7-20)).[80] A French team developed risk prediction models for cardiovascular SMM specifically; preeclampsia, chronic hypertension, and gestational hypertension were among the most important predictive factors for SMM.[81] The CDC further estimates that approximately one-third of maternal deaths are attributable to cardiovascular conditions and records a 25% increase

in the number of women beginning pregnancy with cardiovascular disease from 2003-2012.[8] Similarly, an analysis of discharge data from the National Inpatient Sample from 2004-2011 found an increase in peripartum cardiomyopathy over this period.[82]

While the full picture of SMM timing is not well understood, evidence is emerging that cardiovascular complications are a particular concern postpartum. For example, in a cohort study of California births from 2008-2012, 17% of women with SMM at their postpartum hospital readmission had pulmonary edema or acute heart failure.[83] It is also clear that some hypertensive and cardiovascular pregnancy complications increase cardiovascular risk later in life. Compared to normotensive women, women with a hypertensive disorder of pregnancy have twice the risk of cardiovascular disease later in life.[84] In a Danish registry-based cohort study of still or live births from 1995-2012, rates of hypertension were 3-10 times higher among women with hypertensive disorders of pregnancy compared to normotensive women over the first 10 years postpartum, and approximately twice as high 20 or more years postpartum.[85] Similarly, an earlier Danish registry-based cohort study (1978-2007) found that, compared to normotensive women, women with severe preeclampsia had substantially elevated risk of thromboembolism up to 30 years postpartum: aHR (95% CI): 1.91 (1.35, 2.70).[86]

Although the contributions of specific SMM conditions to morbidity, mortality, and additional health consequences or complications postpartum remain to be elucidated, there is a large and diverse body of evidence suggesting that cardiovascular complications postpartum warrant special attention.

1.1.2.4 Methodological inconsistencies in studies of SMM impact on postnatal health.

Most studies directly assessing the health impacts of SMM were conducted in low- and middle-income settings, which limits the generalizability of their findings to a high-income

country context. The dearth of relevant studies conducted in high-income countries like the US may be at least partially attributable to the increasing focus on postpartum and "late" severe maternal morbidities in these countries; instead of conceptualizing symptoms or issues arising after 42 days postpartum as postnatal morbidity, some work in the US has extended the concept of SMM to accommodate these morbidities.[1,4] Regardless, studies conducted in low- and middle-income countries still inform US research in critical ways. Many identify lack of social support and financial stress as major contributors to persistent postpartum health effects of SMM.[65,66] Such a conclusion might be applicable to the United States, where women experience high rates of "churn" in insurance coverage around labor and delivery[87] and where there is no national paid family leave policy.[88] Many studies, especially those with an ethnographic or anthropological focus, also identify poor communication between patients and clinicians as a critical contributor to SMM and mortality, which is applicable in the United States as well.[89]

However, there are significant limitations to this literature, which complicate any effort to use existing literature to make inferences about the likely postpartum health consequences of SMM in the United States. Our proposal will address this gap by comprehensively characterizing the postpartum health consequences of SMM in a Medicaid population, which in the United States covers about half of all deliveries.[90]

1.2 Innovation

Our proposal is innovative in both its methods and scope of research. Currently, screening using the CDC criteria is the standard method for identifying SMM. This screening is rarely followed by chart review to ascertain true case status. Our proposal applies an innovative method

to SMM case identification; this novel methodological approach addresses long-standing issues with SMM screening by leveraging more data that may be predictive of severe maternal morbidity and establishing the empirical basis of SMM case identification more firmly. In terms of scope, SMM is currently understood as a proxy for maternal death. As a consequence, most research into the risk and burden of SMM concerns the delivery hospitalization encounter and only occasionally postpartum readmission encounters following delivery. Our work broadens the scope of inquiry by capturing SMM events throughout pregnancy, delivery, and the postpartum period and by extending the period of risk relevant to SMM for two year after delivery, thus bringing the long-term consequences of SMM into the purview of SMM research.

1.3 Approach

1.3.1 Overview of project plan

We have the unique opportunity to use existing data to fill important gaps in knowledge about SMM case identification and postpartum consequences of SMM. We will improve screening for SMM by developing an ensemble machine learning algorithm to identify SMM in administrative data (the Magee Obstetric Maternal and Infant Database, or MOMI). This represents an important step towards the development of a more robust empirical basis for SMM screening. We will also build on established knowledge of SMM by comprehensively characterizing a number of postpartum sequelae. In order to determine the long-term consequences of SMM, we will use data from Pennsylvania Medicaid to assess risk of adverse cardiovascular events over the first two years postpartum among who had an SMM during delivery as compared

to women who did not. Overall, this project will develop a new method to screen for SMM and determine the impact of SMM on health outcomes postpartum.

1.3.2 Data sources and populations

1.3.2.1 Pennsylvania Medicaid

Pennsylvania Medicaid data include inpatient, outpatient, pharmacy, and health care provider data for all enrolled individuals. We will create an analytic dataset covering deliveries in 2016-2018 (inclusive) (Fig. 1, left). The study period covers 280 days prior to the delivery date and up to 2 years of follow-up for each delivery. Women with a delivery on or after July 7, 2016 will be included along with information from 280 days prior to delivery (corresponding to the first date of ICD-10 code use in 2015). Women with a delivery prior to July 7, 2016 will be excluded to avoid introducing bias from left truncation.

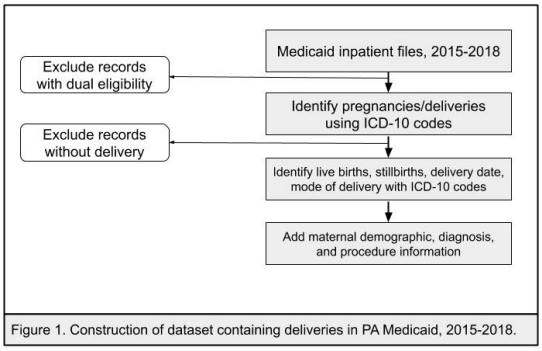

Figure 1.B. Construction of dataset containing deliveries in PA Medicaid, 2015-2018

Exclusion criteria are dual (Medicare and Medicaid) eligibility, male gender, and absence of a delivery during the study period. After excluding records with male gender and dual eligibility,

we will identify pregnancies/deliveries using ICD-10 diagnosis codes from inpatient, outpatient, or professional files and PROC_CODES from inpatient, outpatient, and professional files. Next, we will exclude women without deliveries based on ICD codes for, *e.g.*, elective abortion. We will then identify live births and stillbirths, with associated delivery dates, using ICD codes. Cesarean deliveries will also be identified from inpatient files using ICD codes. Figure 1 shows this process with a flow chart. Maternal demographic variables will be identified from enrollment files, and information needed to determine SMM status, health conditions, and health care utilization variables will be obtained from inpatient files.

Table 3 shows some characteristics of the women in the analytic sample (2016-2018). In the full sample, the majority of deliveries were to women who are ages 20-34, non-Hispanic white, eligible for Medicaid based on income, and who delivered vaginally. Compared to women without SMM, women with SMM were more likely to be 35 years old or older, non-Hispanic Black, eligible for Medicaid based on delivery status, to deliver via c-section, and to have gestational diabetes and preeclampsia.

Table 3. Weighted characteristics of the study sample, Pennsylvania Medicaid, 2016-2018, N = 139,531 deliveries

	Full sample N = 139,531	Severe maternal morbidity N = 5832	No severe maternal morbidity N = 133,699
		% or mean (SD)	
Maternal age category			
< 20	6.9	7.1	6.9
20-34	81	77	81
≥ 35	12	16	12
Maternal race/ethnicity			
Non-Hispanic White	46	43	47
Non-Hispanic Black	24	31	24

Hispanic	20	18	20
Other	9	8	9
Eligibility category			
Disability	4.3	7.6	4.1
Expansion/income	40	41	40
Pregnancy	56	52	56
Gestational diabetes	7.7	11	7.5
Preeclampsia	2.1	7.5	1.9
Mode of delivery			
Vaginal delivery	71	52	72
Cesarean delivery	29	48	28

1.3.2.2 The Magee Obstetric Maternal and Infant (MOMI) Database

The Magee Obstetric Maternal and Infant (MOMI) Database contains detailed information on all deliveries at Magee-Womens Hospital of UPMC in Pittsburgh, Pennsylvania. This hospital has approximately 10,000 deliveries per year and serves a 4-million-person catchment area. The MOMI database contains >300 variables related to deliveries at Magee-Womens Hospital, including information from admissions records, medical record abstraction, the birth record, ultrasound, and other ancillary systems. All deliveries of singleton infants born in any of the study years will be eligible for inclusion.

Table 4 shows characteristics of both the training cohort (MOMI 2010-2011, N = 19,266) and the test cohort (MOMI 2013-2017, N = 47,067). Distributions of maternal characteristics are similar between the training and test cohorts; only prevalence of preexisting diabetes or hypertension is slightly higher in the test cohort (5.3%) than in the training cohort (3.7%).

Table 4. Characteristics of liveborn singleton deliveries at Magee-Womens Hospital, 2010-2011 and 2013-2017

	Eligible training cohort,	Eligible test cohort, 2013-2017

	2010-2011 (N =19,266)	(N = 47,067)
Maternal race, %		
NH White	75	70
NH Black	20	21
Other, declined, or unspecified	5.4	8.9
Maternal age, mean (sd)	29 (5.9)	29 (5.5)
Maternal age ≥ 35, %	16	17
Married, %	56	55
Maternal education, %		
Less than high school	8.3	6.8
High school graduate	22	22
Some college	15	13
4-year college graduate	56	59
Type of insurance, %		
Private	63	62
Public nor none	37	38
Nulliparous, %	48	44
Smoked during pregnancy, %	14	12
Preexisting diabetes or hypertension, %	3.7	5.3
Gestational age, weeks, mean (sd)	39 (2.4)	39 (2.4)
Preterm birth < 37 weeks, %	10	10
Mode of delivery, %		
Vaginal	72	71
Cesarean	28	29
Birthweight, grams, mean (sd)	3277 (615)	3253 (622)

1.3.3 SMM definitions and measurement

1.3.3.1 Pennsylvania Medicaid

In Pennsylvania Medicaid, we will use the same three screening criteria to identify SMM. Following other published literature on SMM using Pennsylvania Medicaid,[91] we will use the CDC list of ICD codes corresponding to SMM to identify cases. Since ICU admission is ascertained using revenue codes and is not reliable unless supplemented by postpartum length of stay information, any delivery with 1.) any of the ICD codes indicating SMM or 2.) prolonged

postpartum length of stay (>3 standard deviations from the mean by delivery type) *and* ICU admission will be designated a SMM case.

Just over 4% of deliveries in Pennsylvania Medicaid had a SMM event each year per the CDC criteria alone. This is considerably higher than national estimates of SMM prevalence derived from the National Inpatient Sample (approximately 1.3%). The higher prevalence of SMM events observed in this dataset is likely the result of the data structure. The National Inpatient Sample only records information from the delivery hospitalization, whereas the Medicaid data contain claims for events occurring at any time during an individual's period of enrollment – including before pregnancy, during pregnancy, during delivery, and postpartum. We calculated the prevalence of SMM conditions in the 2016 National Inpatient Sample (1.7%) and compared this to the prevalence of SMM only during the delivery date and the 1 day preceding and following delivery in Pennsylvania Medicaid in 2016 (1.5%). The comparability of these estimates indicates that the Pennsylvania Medicaid data do not systematically over- or under-estimate SMM prevalence compared to other national databases.

Our ability to access information not only during the immediate time interval around delivery but through pregnancy and up to one year postpartum is a strength of our approach. However, there are also limitations inherent in using this data set to study SMM. Though the prevalence of SMM during delivery is comparable to that of other data sources, this approach still relies on screening to determine prevalence estimates and is thus vulnerable to issues arising from using billing codes to determine severe maternal morbidity. Since information on transfusion volume is not available in claims data, we will assess robustness of our estimates to this substantial misclassification of cases by conducting sensitivity analyses excluding cases whose only SMM indicator is blood transfusion (unknown quantity).

1.3.3.2 The Magee Obstetric Maternal and Infant (MOMI) Database

Both the MOMI training set (2010-2011) and the MOMI test set (2013-2017) were constructed using screening followed by chart review. Each of these datasets contains records from individuals whose medical charts were abstracted to determine true SMM status. To identify records for chart review, we screened the MOMI cohorts (2010-2011 and 2013-2017) using: 1) any of the CDC list of diagnosis and procedure codes, 2) ICU admission, or 3) prolonged postpartum length of stay (>3 standard deviations above the mean length of stay for delivery type). Any delivery with any of these criteria was considered a screen-positive case eligible for chart review, while any delivery not meeting these criteria was considered a screen-negative case.

To construct the MOMI training set (2010-2011), the full MOMI cohort (2010-2011, n = 19,266) was screened with the above criteria. All 336 screen-positive cases as well as a random sample of 349 screen-negative cases then underwent chart review, yielding 171 true positive and 506 true negative records (Table 5). To construct the MOMI test set (2013-2017), the full MOMI cohort (2013-2017, n = 47,067) was screened with the same criteria, yielding 1500 screen-positive cases. Across each year, 250 screen-positives (50 per year) and 258 screen-negatives were sampled at random to undergo chart review, yielding 160 true-positive cases and 337 true-negative cases in the test set.

A true gold standard definition for SMM case identification has neither been developed nor uniformly adopted in the US. However, the records in both the training and test sets were reviewed using the Gold Standard Severe Maternal Morbidity Case Review Guidelines outlined in the 2016 American College of Obstetricians and Gynecologists/Society for Maternal-Fetal Medicine Obstetric Care Consensus document.[42,51] These guidelines propose a process for reviewing SMM events involving chart abstraction to identify the documented type of SMM event,

detailed case synopsis, sequence of morbidity, analysis of patient, provider, and system factors contributing to the morbidity as well as the preventability of the SMM outcome.

The prevalence of SMM in MOMI (2010-2011) is 0.094/100,000 deliveries based on the screening definition and 0.045/100,000 deliveries based on the "gold standard" chart review definition. This discrepancy may be due to the inaccuracy of some billing codes. Specifically, the billing codes for blood transfusion are known to yield a large number of false positives. Blood transfusion alone (without any other SMM indicator) is the primary driver of the increase of SMM in the United States, growing from 789/100,000 delivery hospitalizations in 2006 to 1211/100,000 delivery hospitalizations in 2015. (Fig. 2).[4] However, billing codes for blood transfusion do not include volume of blood transfused, and evidence indicates that the inclusion of blood transfusion codes in the definition of SMM results in a large number of false positives.[50]

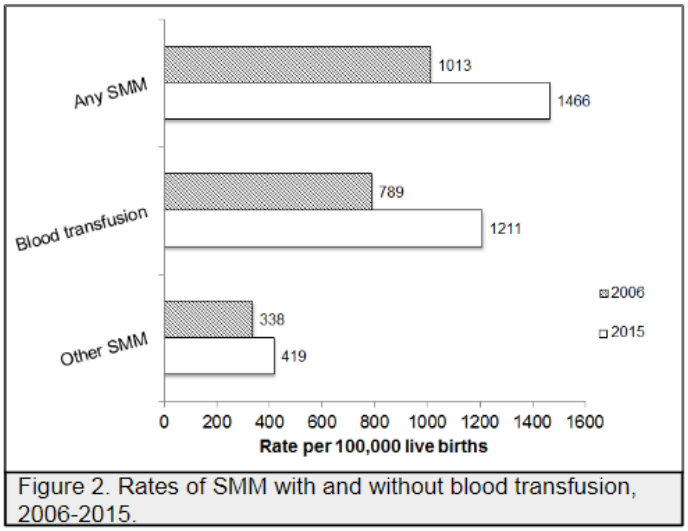

Figure 2. Rates of SMM with and without blood transfusion, 2006-2015.

Figure 1.C. Rates of SMM with and without blood transfusion, 2006-2015

The MOMI database contains only variables measured during the delivery hospitalization. Consequently, we will not be able to observe SMM events that occur before delivery (during pregnancy) or after delivery in the postpartum period. This is likely to result in an underestimate of the prevalence of SMM in this setting. Limitation to the delivery hospitalization may also introduce bias if women included as non-cases in the training or test datasets had an SMM event or condition prior to or following delivery.

1.3.4 Consequences of SMM

1.3.4.1 Adverse cardiovascular events

The health outcomes following delivery that we will examine are: atrial fibrillation, heart failure, ischemic heart disease, and stroke/transient ischemic attack (Table 6). These outcomes were chosen in consultation with an obstetrician and identified from Pennsylvania Medicaid patient and enrollment files using ICD-10 diagnosis and procedure codes provided by the Centers for Medicare and Medicaid Services Chronic Conditions Warehouse.

Any new diagnosis of any of these conditions recorded after delivery will be considered a relevant outcome for the purposes of this analysis. New diagnoses of these adverse cardiovascular events will be indexed to the date of delivery for both cases and controls. Any condition with a diagnosis date preceding delivery will not be considered a new diagnosis regardless of recurrence of the same diagnosis at a date post-delivery.

This approach to defining the outcome is subject to several limitations. First, some diagnoses may be incorrectly classified as new because it is not possible to ascertain whether these diagnoses existed prior to Medicaid enrollment or the beginning of the study period. We will be unable to assess any diagnoses that are not recorded in Medicaid or in our analytic data, which

may introduce outcome misclassification. This misclassification could be differential if women with SMM and women without SMM have different profiles of contact with the health care system before or after delivery. Non-differential misclassification of the outcome may be introduced if, regardless of SMM status, women are more likely to receive a diagnosis of a preexisting condition for the first time after delivery due to increased contact with health care providers during pregnancy, delivery, and the postpartum period. Since the adverse cardiovascular events we will study are generally acute episodes requiring hospitalization, we hope to minimize the impact of this bias.

1.3.5 Statistical analyses

1.3.5.1 Specific Aim 1: Determining the long-term consequences of SMM

To assess adverse cardiovascular events following SMM, we will first construct inverse probability of censoring weights to account for non-random loss to follow-up in our Medicaid cohort. We will then construct a pooled logistic regression model, weighted by inverse probability of censoring, for each of the adverse cardiovascular event outcomes. We will use the parametric g-formula in a time-fixed setting (marginal standardization) to estimate, based on the parameters from the pooled logistic regression models, risk of each adverse cardiovascular event under counterfactual hypothetical scenarios in which every woman in the sample had SMM and no woman in the sample had SMM.

Because of the potential for differential misclassification of the outcome if women with SMM events have more medical encounters and therefore receive new diagnoses at a higher rate, we will assess rates of new diagnosis of three genetic conditions (cystic fibrosis, hereditary hemorrhagic telangiectasia, and ornithine transcarbamoylase deficiency) before and after delivery

for women with and without SMM at any time during the study period. If rates are substantially different before and after delivery and/or between women with SMM and women without, we will conduct a quantitative bias analysis to determine the degree to which this misclassification affects the new diagnosis risk estimates.

For all models, we will also conduct sensitivity analyses. Since cases involving blood transfusion comprise the plurality of SMM cases and since blood transfusion is not an accurate indicator of SMM, we will repeat all analyses excluding any individuals whose only SMM indicator is an ICD-10 code for blood transfusion. We will determine how robust the estimates derived from all models described previously are to the exclusion of these individuals.

Our approach has important limitations. One limitation is the possibility for exposure and outcome misclassification. Use of a screening definition for SMM will result in exposure misclassification via creation of a large number of false positives. We will conduct sensitivity analyses excluding SMM cases whose only SMM indicator is blood transfusion to assess the robustness of our results to this misclassification. The outcome may also be misclassified if women received a diagnosis of any adverse cardiovascular event prior to enrollment in Medicaid, or if women both with and without SMM (or women with SMM in particular) had more contact with health care providers during the puerperium. We will not adjust for multiple deliveries to the same person and instead will treat the unit of analysis as deliveries, which may introduce bias if many women in the dataset have multiple deliveries over the study period. Finally, the data we will use are administrative data designed for billing, not research purposes.

1.3.5.2 Specific Aim 2: Developing a screening algorithm for SMM

To develop a screening algorithm for SMM, we will use the ensemble machine learning algorithm SuperLearner. Briefly, SuperLearner combines predictions from a library of user-

specified machine learning algorithms, then weights the predictions from each algorithm to create a final, weighted "ensemble" prediction algorithm. To implement SuperLearner, a library of candidate algorithms must be specified and supplied to the program. The data must then be split into training and test sets,[92] and each algorithm in the library is fit on the training set. We will use a temporal data split, wherein the training set is the 2010-2011 MOMI subset and the test set is the 2013-2017 MOMI subset. The estimated fits developed on the training set are used to generate predictions in the test data set,[92,93] and predictions from each algorithm are then "stacked"[93] and regressed against the true outcome. This regression determines the combination of weights that minimizes prediction error according to a user-specified function.[93] The result is a final SuperLearner algorithm, which will (theoretically) perform at least as well as the "best" candidate algorithm included in the library.[94]

We will perform a simulation study to evaluate sample size and class imbalance considerations. Two features of the MOMI training and test data sets (Table 5) that may impact SuperLearner performance are sample size and class imbalance (ratio of controls to cases). The machine learning literature is unclear on whether these issues categorically impact prediction algorithm performance; while some algorithms are thought to be more sensitive to imbalance and others less sensitive, most evidence suggests that the impact of sample size and degree of imbalance on machine learning algorithm performance is context-specific and depends on the structure of the training and test data.[95]

Table 5. Algorithms included in SuperLearner library for simulation.

Algorithm	Tuning parameters
K-nearest-neighbors	• k = 5 • Weights = null (kernel KNN), Epanechnikov

Extreme gradient boosting (xgboost)	• Number of trees = 200, 500, 1000 • Maximum depth = 4, 5, 6 • Shrinkage parameter = 0.01, 0.001, 0.0001
Random forests	• 2500 trees
Support vector machines	• Default
Penalized regression (glmnet)	• Default

To investigate these issues, we will conduct a simulation study with completely synthetic data. We will perform 2000 Monte Carlo simulations, each time generating a dataset from a logit model with N = 1000, 20 covariates (10 categorical and 10 continuous), and outcome prevalence ranging from 15%-50%. For half of the simulated datasets, we will perform simple downsampling to balance the number of cases and controls: randomly sampling a number of non-cases equal to the number of cases to create a perfectly balanced dataset. We will leave the other 1000 simulated datasets unbalanced. For all datasets (downsampled and unbalanced), we will use SuperLearner (Table 8 lists algorithms and tuning parameters to be included) and logistic regression with data splitting, separately, to predict the outcome and examine predictive performance (sensitivity, specificity, positive and negative predictive value, overall accuracy, and area under the receiver operating curve).

We will train a SuperLearner algorithm to screen for SMM in a hospital data set (MOMI). First, we will determine which variables in the MOMI data set to include in the training and test data sets according to availability in all years (2010-2011 and 2013-2017) and known or suspected association with true SMM status.

In the training set, we will train a SuperLearner algorithm with a rich library of base learners. In the absence of practical guidance about optimal algorithm selection generally and in small sample settings specifically, this choice is intended to include as many diverse algorithms as

possible since the true data-generating mechanism is unknown to us and to maximize the likelihood of obtaining the "oracle property" in SuperLearner.[96] We will include the screening criteria used to identify screen-positive and screen-negative cases (the CDC list, ICU admission, or prolonged postpartum length of stay) as one of the algorithms in the SuperLearner library, and evaluate characteristics of the SuperLearner including the discrete SuperLearner (the single candidate algorithm that performs the best) and the weight coefficients of the component algorithms. We will then apply the algorithm to the test data set (2013-2017) to generate predictions, which will be used to generate metrics of predictive performance (area under the receiver operating curve, positive and negative predictive values, overall accuracy).

This approach has important limitations. As already mentioned, SuperLearner performance in small samples is not well-characterized, nor are guidelines for power or sample size calculations for ensemble learning well-developed.[97] Although our simulation was performed in completely synthetic data that may not approximate the characteristics of the MOMI subset, our preliminary simulation data indicated that an N of 400 yields acceptable performance, and that the SuperLearner tolerates moderate class imbalance relatively well. Furthermore, the objective of our analysis is limited to prediction (not inference), and as such appropriate power to construct derivatives of the standard error for inferential purposes is not a critical consideration. Another limitation is that guidance is lacking on optimal algorithm selection for prediction generally and in small-sample settings more specifically, which we have attempted to address by including a diverse SuperLearner algorithm library. Finally, our result may be difficult to interpret and not transportable to other data sets, a common critique of machine learning applications in clinical science.[98] We will attempt to increase the transparency and generalizability of our result by specifying all algorithms and tuning parameters *a priori*, reporting their weights in the final

SuperLearner ensemble, and by choosing variables for inclusion in the training set that are widely available.

1.3.6 Overall impact

The overall impact of this work is to improve understanding of SMM in the United States. The proposed work leverages rich data sources and innovative analytic approaches to better characterize the burden of SMM, improve SMM surveillance and research, and inform strategies to reduce the overall burden of SMM. First, screening for SMM case identification forms the core of SMM research in the US. Our proposal will address well-known issues with existing screening criteria and test a novel method to perform SMM case identification. This will result in more accurate prevalence estimates of severe maternal morbidity and consequently more accurate inferences about the burden, population risk factors, and quality improvement implications of SMM. Second, postnatal morbidity and other consequences of SMM have not been extensively investigated. Postnatal morbidity is not only important an endpoint in its own right, but may also be related to risk of indirect or late maternal mortality. Thus, our proposal responds to the critical need to add extensive understanding of the postpartum consequences of SMM to the already well-developed knowledge base about the pre-delivery risk factors for SMM and maternal death.

2.0 Chapter 2: Adverse cardiovascular events following severe maternal morbidity

2.1 Introduction

The maternal mortality rate, defined as the number of deaths related to pregnancy or delivery per 100,000 live births, is higher in the United States (17.4 per 100,000)[99] than any other comparably wealthy country.[79] However, maternal mortality signals only the "tip of the iceberg" of a much larger burden of poor maternal health in the US.[42] This burden of poor maternal health is partially reflected in the high annual incidence of severe and potentially life-threatening complications of pregnancy, labor, and delivery collectively referred to as severe maternal morbidity (SMM).[42,100] The incidence of SMM during delivery hospitalizations increased 45% from 1010/100,000 in 2006 to 1470/100,000 in 2015.[4]

Over the past 30 years, cardiovascular conditions (*e.g.* cardiomyopathy, hypertensive disorders) have eclipsed hemorrhage and sepsis as the leading causes of maternal mortality and SMM.[5,19] Cardiovascular conditions may be exacerbated by the "stress test" of pregnancy,[101,102] and some pregnancy complications (such as preeclampsia) have been associated with cardiovascular disease later in life.[103] While this suggests that SMM may have long-term effects after delivery, little is currently known about the health consequences of SMM in the postpartum period and beyond.

Currently, national organizations such as the American College of Obstetricians and Gynecologists (ACOG) are focused on optimizing postpartum care as part of the continuum of maternal health.[42] Because severe complications or sequelae may occur after delivery even as many US women lose access to peripartum medical care,[87] the postpartum period may be a

particularly high-risk time for women who had a SMM. One US study concluded that women who experienced a SMM during delivery are more likely to be readmitted over the first year postpartum.[104] Some evidence from primarily small cohorts in low- and middle-income countries have also documented increased risk of functional impairment and poor quality of life following SMM.[64] However, no prior studies have estimated the association between SMM and cardiovascular outcomes after delivery. Therefore, the objective of this study was to determine the extent to which severe maternal morbidity in pregnancy or postpartum is associated with increased risk of adverse cardiovascular events in the two years after delivery.

2.2 Methods

We conducted a retrospective longitudinal cohort study using Pennsylvania Medicaid administrative claims and encounters data for all enrolled individuals from 10/01/2015 to 12/31/2018. The study was approved by the University of Pittsburgh Institutional Review Board (IRB) #STUDY19100253. Women were eligible for inclusion in this study if they were enrolled in Pennsylvania Medicaid from 2015 to 2018 and delivered at 20 to 42 weeks' gestation. We identified 138,564 eligible deliveries that occurred from 07/07/2016 to 12/31/2018 (Figure 1) using an algorithm previously developed by our group to identify deliveries from Medicaid files.[105-107] There were 18,022 women with more than one delivery in this time period. Of their 27,461 repeated deliveries, we excluded 1085 deliveries (4%) with <28 days to the next pregnancy, which we considered implausible. Finally, 495 deliveries (<1% of total) had an outcome event (acute myocardial infarction, heart failure, and stroke/transient ischemic attack) from conception through ≤ 42 days postpartum, but no documentation of any of the SMM diagnoses included in the Centers

for Disease Control (CDC) definition. To adhere as closely as possible to the CDC definition of severe maternal morbidity [4,100] and attempt to capture incident outcomes after 42 days postpartum, we excluded these records. The final analytic sample consisted of 137,140 deliveries to 128,686 women.

We linked maternal delivery records to child records using a family identification number included in the Medicaid enrollment data (97% linkage). Child records include a variable indicating the gestational age (an integer from 20 to 36 weeks) for infants who are born preterm. Infants born at term (≥37 weeks) have a missing value for this variable. For term deliveries (91%) and deliveries that we could not link to a child record (3%), we assumed a 40-week (280-day) gestation. We varied the length of gestation from 38 to 40 weeks for term deliveries to evaluate the sensitivity of our results to this assumption.

The exposure of interest was the presence of any severe maternal morbidity. We defined severe maternal morbidity as either: any of the diagnosis or procedure codes for 21 SMM indicators as outlined by the CDC (Table 9)[3,100] or intensive care unit (ICU) admission occurring from conception to 42 days after delivery (Table 9).[108]

We studied five cardiovascular outcomes occurring during the first 2 postpartum years (43 days to 773 days): atrial fibrillation, heart failure, ischemic heart disease (including acute myocardial infarction), stroke/transient ischemic attack, and a composite outcome of any of these events. These outcomes were identified using ICD-10 code algorithms provided by the Centers for Medicare and Medicaid Services Chronic Conditions Warehouse (Table 9).

Maternal age, race/ethnicity (non-Hispanic white, non-Hispanic Black, Hispanic, and other race/ethnicity), and the six Medicaid eligibility categories[109] were obtained from enrollment files. We categorized eligibility category due to disability status, low household income, or pregnancy-

related eligibility. The hierarchy for assigning eligibility for records with multiple eligibility indicators was disability, income/expansion, and then pregnancy. Mode of delivery (vaginal or cesarean) was ascertained from inpatient files using diagnosis codes for infants (Table 11). We used ICD-10 code algorithms developed from inpatient, outpatient, and professional files[105-107] to identify parity, preexisting conditions (HIV infection, hepatitis C infection, asthma, and obesity), pregnancy complications (gestational diabetes, preeclampsia, and thyroid disease in pregnancy), tobacco use, and non-tobacco substance use disorders (including alcohol, cannabis, cocaine, hallucinogens, opioids, and sedatives).

The unit of analysis for this study was a delivery. Follow-up time was counted from 43 days postpartum until the woman experienced a cardiovascular events, was lost to follow-up, or until the end of the study on 12/31/2018. Follow-up time for a given delivery was considered administratively right-censored if it occurred on or after 1/1/2018 (resulting in less than a full year of follow-up). Women were considered lost to follow-up if they were not right-censored but disenrolled from Medicaid before either 12/31/2018 or 365 days after delivery (23% of deliveries) to ensure at least one year of follow-up. For women with more than one delivery in the study period, censoring occurred at the start of the next pregnancy.

Statistical analysis

To minimize potential bias from non-random loss to follow-up during the postpartum period, we constructed stabilized inverse probability of censoring weights[110] that we applied to all analyses. We used logistic regression to predict loss to follow-up using (Medicaid eligibility category,[111] maternal preexisting conditions, preeclampsia, gestational diabetes, parity, tobacco use, race/ethnicity, and other substance use disorder. The mean stabilized weight was 1.0 (standard deviation: 0.14; range: 0.79 to 1.6; Figure 2.C) All analyses were weighted.

We generated adjusted risk curves for each cardiovascular outcome and to calculate effect measures throughout follow-up.[112] Atrial fibrillation is exceedingly rare in this sample, so our analysis of this outcome was descriptive. We used g-computation, which in the case of a time-fixed exposure such as SMM, is equivalent to marginal standardization to the entire study population.[113] G-computation requires specification of exposure regimens usually corresponding to an intervention on the exposure of interest, even if such an intervention is hypothetical. We estimated the average treatment effect, which compares the risk of cardiovascular events in the hypothetical scenario where every delivery in the data set had a SMM (henceforth "SMM" or "under SMM" to refer to risks calculated under this hypothetical exposure scenario) to the effect in the hypothetical scenario where no deliveries had a SMM (henceforth "no SMM" or "under no SMM"). This average treatment effect quantifies the effect of the largest alterations of the exposure possible in this population.

We obtained population-average risks of each outcome by adapting an algorithm[114-116] to implement the parametric g-formula. Our adaptation of this algorithm consisted of four steps. First, we fit separate binomial pooled logistic models[117] for each outcome to the original data. Time was modeled using a quadratic term to allow for flexible, non-linear relations. These models were adjusted for confounders chosen via causal diagrams.[118] Selected confounders were maternal race/ethnicity, age, Medicaid eligibility category, tobacco use, substance use disorder (including alcohol), asthma, HIV infection, hepatitis C infection, obesity, preeclampsia, gestational diabetes, thyroid disease in pregnancy, mode of delivery, and parity.

Second, we constructed two simulated data sets under each intervention (SMM and no SMM). We created two copies of the original data set, setting the exposure to correspond to the desired hypothetical intervention in each: all deliveries with SMM or all deliveries with no SMM.

Third, we used the parameters from each pooled logistic model fit to the original data in Step 1 to predict, in each simulated data set, the survival probabilities for each individual under each intervention, given their observed covariate values.

Fourth, under each intervention, we used the complement of the survival probabilities from Step 3 to calculate the risks of each outcome at each month of follow-up. We calculated risk differences and generated 95% confidence intervals using nonparametric bootstrapping with 200 resamples.

We also performed three sensitivity analyses. We repeated our analyses after excluding deliveries whose sole SMM indicator was blood transfusion because its associated ICD-10 code is frequently misclassified.[2,51,119] We also repeated analyses varying gestational length of all term deliveries to 38 weeks (266 days) or 42 weeks (294 days).

2.3 Results

More than half (56%) of the sample qualified for Medicaid based on pregnancy eligibility. Mean maternal age in the sample was 27 (SD: 5.7) years. Nearly half of the deliveries were to non-Hispanic white mothers, and over half of the deliveries were to primiparous women. Maternal asthma, obesity, tobacco use disorder, and drug use in pregnancy were common. Gestational diabetes affected 7.7% of deliveries. Preeclampsia was present in 2.1% of deliveries. Nearly one-third of the births in the sample were delivered via cesarean section (Table 6).

The cumulative incidence of SMM was 4.2%. Most deliveries with SMM (85%) had only one SMM indicator; 10% had two indicators, and 5% had three or more. The most common indicators among deliveries with any SMM event were blood transfusion (21%), eclampsia (17%), and sepsis

(14%) (Table 12). Cardiovascular complications accounted for nearly 50% of SMM events. SMM events were evenly split according to timing: during pregnancy (33%), at labor and delivery (35%), or in the first 42 days postpartum (32%).

Deliveries with SMM occurred more frequently to non-Hispanic Black mothers and women covered by disability eligibility for Medicaid compared with deliveries lacking an SMM event. Prevalence of maternal asthma, obesity, tobacco use, drug use in pregnancy, preeclampsia, and gestational diabetes was higher than among deliveries without SMM than among deliveries without SMM. Maternal age and parity did not vary by SMM (Table 6).

Cumulative incidences of adverse cardiovascular events at the end of the follow-up period were highest for stroke or transient ischemic attack and ischemic heart disease, followed by heart failure and atrial fibrillation (Table 7). Approximately 25/1000 deliveries had any one of these cardiovascular events. At the end of follow-up, each adverse cardiovascular outcome occurred 3 to 10 times as frequently among deliveries with SMM than among deliveries without.

The risk of cardiovascular events increased with each month of follow-up (Figure 2). Under the hypothetical exposure scenario where we set all deliveries to have a SMM, the risk of each outcome increased more quickly than under the hypothetical intervention where we "prevented" SMM (set every delivery to have no SMM). (Table 13). For instance, from 1 month to 26 months of follow-up, the risk of ischemic heart disease per 1000 deliveries in the SMM scenario increased from 0.80 to 13, while risk of ischemic heart disease in the no SMM scenario increased from 0.30 to 5.0.

Comparing the SMM scenario with the no SMM scenario, there was an excess risk of heart failure, ischemic heart disease, and stroke/transient ischemic attack associated with SMM that increased throughout most of the follow-up period. From the first month of follow-up to the last

month of follow-up, the risk differences per 1000 deliveries comparing SMM to no SMM increased from 0.5 to 6.4 for ischemic heart disease, 1.0 to 8.2 for stroke/transient ischemic attack, 1.2 to 12 for heart failure, and 2.7 to 28 for any cardiovascular event (Table 8). As follow-up progressed, the risk differences comparing SMM to no SMM increased to a greater magnitude for heart failure than for ischemic heart disease or stroke/transient ischemic attack (Figure 2.D).

Sensitivity analysis excluding blood transfusion from the definition of SMM yielded risk differences that were not meaningfully different from the primary analysis (Table 14). Our results were also similar after varying the gestational age assumed for term pregnancies to 38 weeks (Table 15) or 42 weeks (Table 16).

2.4 Discussion

In this large cohort of Pennsylvania Medicaid deliveries, we found that women who had a pregnancy with SMM were at substantially increased risk of heart failure, ischemic heart disease, and stroke ortransient ischemic attack up to 2 years postpartum. SMM was most strongly associated with risk of heart failure, with the risk difference comparing SMM to no SMM rising steadily over follow-up. Ischemic heart disease and stroke ortransient ischemic attack were also associated with SMM, but the increase in risk began to flatten around 20 months of follow-up. While these cardiovascular events are rare in reproductive-age women, the average treatment effect of SMM on risk of each cardiovascular event is pronounced.

We are unaware of research relating SMM to postpartum cardiovascular events. Our results are consistent with previous work on the non-cardiovascular postpartum health consequences of SMM[5,21,22,25,42,48,61] A UK cohort study found SMM to be associated with 2-3 times the odds of

post-traumatic stress disorder at 6-8 weeks postpartum.[77] A recent Brazilian study reported that women with SMM had more reproductive and general self-rated health after delivery than women without.[120] and one US study, notable for using the CDC screening definition of SMM, reported that SMM was associated with more than double the risk of readmission for an inpatient stay at 6 and 12 months postpartum.[104] Conclusions from most studies on the postpartum consequences of SMM may not be generalizable to US populations: most were conducted in low- and middle-income countries, the definition of SMM commonly used to identify cases in the US is not widely used outside North America,[121] and many are limited by study design issues (small sample sizes and and lack of a control group being common issues).[64] Most importantly, none of these studies examine cardiovascular disease risk after SMM, although this is important for US reproductive-aged women, whose leading modifiable cause of death is cardiovascular disease.[122,123]

Our findings also broadly agree with work connecting pregnancy complications to elevated cardiovascular risk from delivery through mid-life.[84,85,124-129] Because hypertensive disorders of pregnancy are major contributors to SMM,[62] and cardiovascular conditions (including cerebrovascular disorders, pulmonary edema, and acute heart failure) accounted for over 50% of the SMM events in our sample, it is possible that cardiovascular conditions and eclampsia may be primarily responsible for our findings. However, we did not have adequate sample size to evaluate individual SMM events or groups of SMM events in relation to the outcome, and so are unable to make inferences about specific disease processes or biological mechanisms. Although our results cannot determine whether pregnancy complications initiate pathophysiological processes generating cardiovascular risk or whether they primarily exacerbate preexisting cardiovascular risk,[84,130] they support the idea that development of SMM may identify women who would benefit from individualized follow-up care after delivery to manage cardiovascular risk.[129,131]

Our results should be considered in the context of several important limitations. First, we used the CDC screening definition of SMM, which has a high negative predictive value and a low positive predictive value (44-48%).[51,132] The gold standard definition[42] requires medical record review to evaluate specific clinical criteria. Lack of access to medical records prevented us from using the gold standard definition or conducting an internal validation study. However, our sensitivity analysis showed no meaningful differences when we excluded blood transfusion (a commonly misclassified indicator) from our definition. Our reliance on ICD-10 codes to define the outcomes and covariates has the potential to introduce misclassification bias. While ICD-10 codes for adverse cardiovascular events have reasonably high positive predictive values (> 90%),[133-135] it is likely that confounders including obesity are subject to measurement error.[136,137] In the absence of comprehensive validation studies of the ICD-10 codes we used, we cannot rule out the potential for misclassification bias..

Selection bias may be a problem because we could not capture information about women when they are not enrolled in Medicaid. We used inverse probability of censoring weighting to address potential selection bias resulting from nonrandom loss to follow-up, but this technique does not address potential confounding by, *e.g.*, maternal health history before the start of the study or enrollment in Medicaid. Finally, some of the outcomes were not truly incident since SMM diagnoses and outcome diagnoses can be identical in some cases (in the case of, *e.g.*, some types of ischemic heart disease). Caution should thus be exercised in interpreting these results.

Our results could be interpreted causally only if the causal identifiability assumptions of conditional exchangeability, positivity, and consistency hold.[138] The consistency assumption poses an intractable problem to the causal interpretation of our results[139] because SMM is a heterogeneous construct with no unifying biological or physiological feature; any two SMM cases

can have different diagnoses, clinical courses, and prognoses. Although severe maternal morbidity has clinical utility as a concept, we cannot interpret our findings causally. However, this study provides important evidence about the long-term risks of SMM and opens the door to future causal work in this area.

The present work has important implications for maternal care after delivery. It is well-known that the transition from postpartum care to well-woman care represents a discontinuity in maternal health care provision.[140] Our results demonstrate that women with an SMM are at high risk for adverse postpartum cardiovascular events, which suggests that both women's delivery history and multidisciplinary clinical expertise can be leveraged to craft optimal policies to guide this care transition for women with obstetric morbidities. Women who suffer a SMM constitute a high-risk group that would likely benefit from proposed policy changes[141] to streamline and individualize postpartum and well-woman care, both with regard to adverse cardiovascular events but likely other possible sequelae as well. Although the absolute number of women experiencing adverse cardiovascular events after delivery is small, SMM is an important risk factor for adverse cardiovascular events beyond the traditional 42-day postpartum period and is associated with substantially increased risk of these events which are serious, costly, and potentially impact long-term health.

2.5 Tables and figures

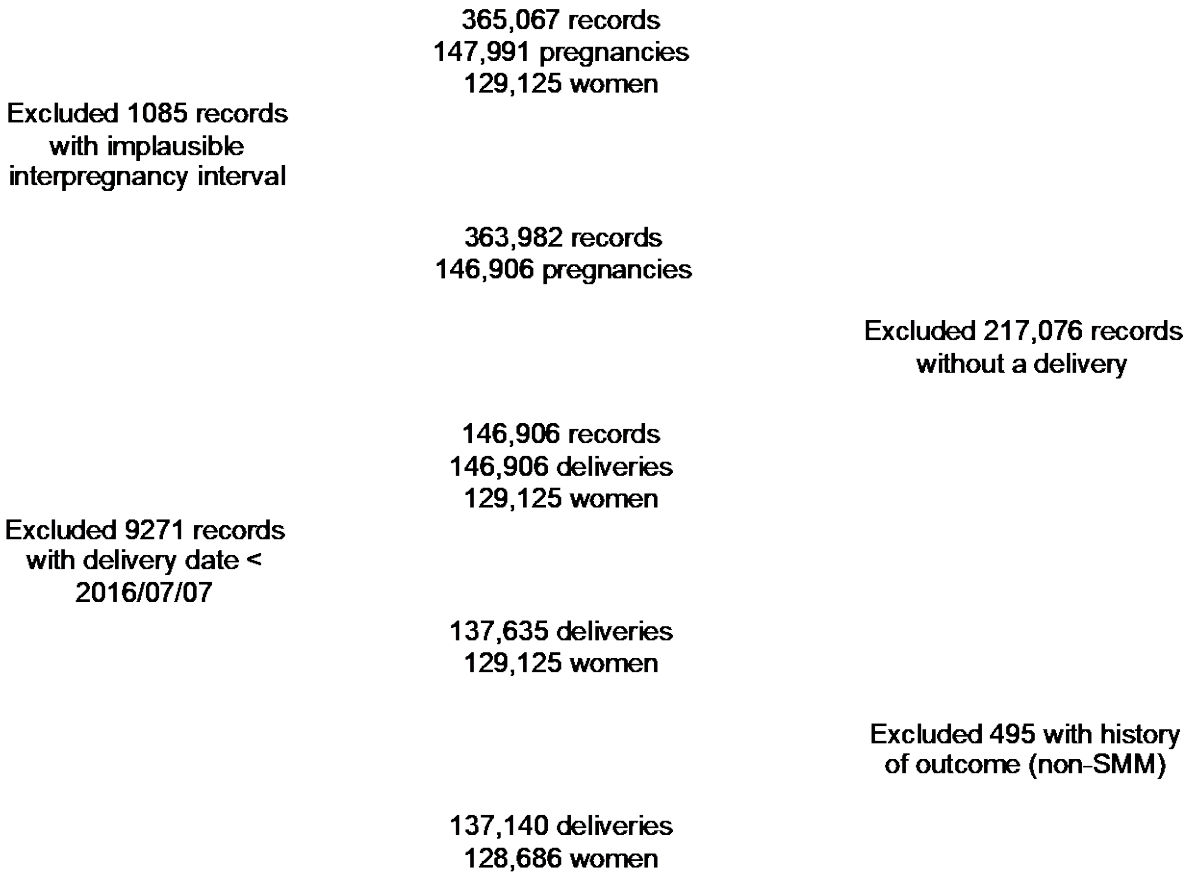

Figure 2.A. Study sample selection flow chart, Pennsylvania Medicaid, 2016-2018

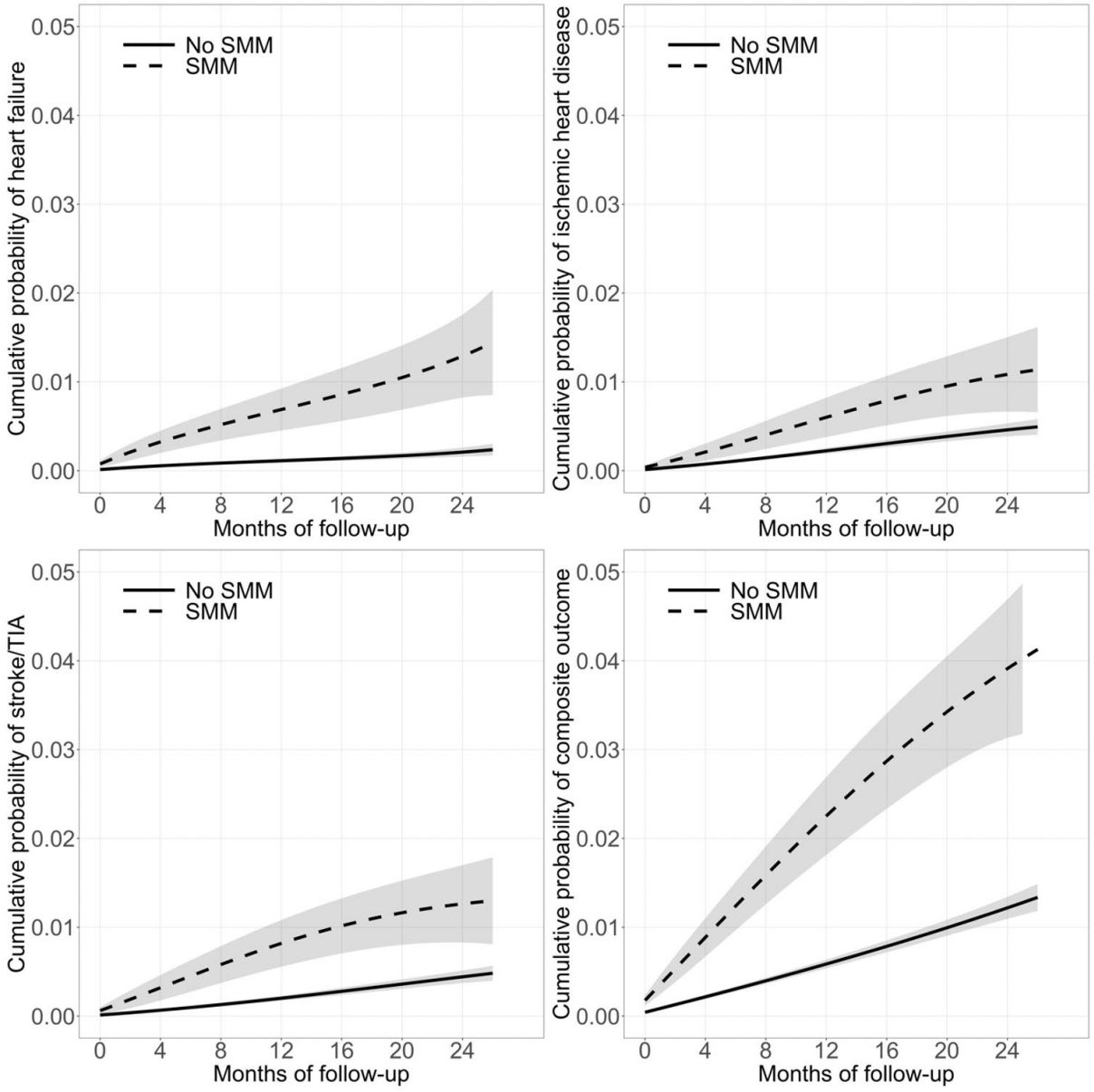

Figure 2.B. Adjusted cumulative distribution functions for cardiovascular events for SMM and no SMM hypothetical intervention scenarios, Pennsylvania Medicaid, 2016-2018, N = 139,531 deliveries.

Table 6. Weighted characteristics of the study sample, Pennsylvania Medicaid, 2016-2018, N = 139,531 deliveries.

	Full sample N = 139,531	Severe maternal morbidity N = 5832	No severe maternal morbidity N = 133,699
		% or mean (SD)	
Maternal age	27 (5.7)	28 (6.1)	27 (5.7)
Maternal age category			
< 20	6.9	7.1	6.9
20-34	81	77	81
≥ 35	12	16	12
Maternal race/ethnicity			
Non-Hispanic White	46	43	47
Non-Hispanic Black	24	31	24
Hispanic	20	18	20
Other	9	8	9
Eligibility category			
Disability	4.3	7.6	4.1
Expansion/income	40	41	40
Pregnancy	56	52	56
Parity			
1	57	56	57
2	30	29	30
3 or more	13	15	13
Preexisting asthma	11	17	10
Preexisting hepatitis C infection	1.9	3.5	1.9
Preexisting HIV infection	0.23	0.42	0.22
Preexisting obesity	20	29	20
Preexisting thyroid condition	4.8	9.3	4.6
Substance use disorder: tobacco	36	42	35
Other substance use disorder	3.5	4.5	3.4
Gestational diabetes	7.7	11	7.5
Preeclampsia	2.1	7.5	1.9
Gestational age at delivery (days)	277 (12)	271 (21)	277 (12)
Mode of delivery			
Vaginal delivery	71	52	72
Cesarean delivery	29	48	28
Birth status			
Stillbirth	0.95	2.6	0.88
Live birth	99	97	99

Table 7. Unadjusted cumulative incidence of adverse cardiovascular events per 1,000 deliveries following deliveries with severe maternal morbidity vs. deliveries without severe maternal morbidity, Pennsylvania Medicaid, 2016-2018, N = 104,888 deliveries not lost to follow-up.

Severe maternal morbidity	Population at risk, N	Cardiovascular events at the end of follow-up, N	Cumulative incidence of cardiovascular event per 1,000 deliveries
Heart failure			
Severe maternal morbidity	4596	46	10
No severe maternal morbidity	100,292	102	1.0
Total	104,888	148	1.4
Ischemic heart disease			
Severe maternal morbidity	4596	30	6.5
No severe maternal morbidity	100,292	195	1.9
Total	104,888	225	2.1
Stroke or transient ischemic attack			
Severe maternal morbidity	4596	35	7.6
No severe maternal morbidity	100,292	184	1.8
Total	104,888	219	2.1
Any cardiovascular event			
Severe maternal morbidity	4596	116	25
No severe maternal morbidity	100,292	527	5.3
Total	104,888	643	6.1
Atrial fibrillation			
Severe maternal morbidity	4596	<10[1]	1.5
No severe maternal morbidity	100,292	43	0.43
Total	104,888	50	0.48

Table 8. Excess risk of cardiovascular events per 1,000 live births for severe maternal morbidity vs. no severe maternal morbidity, Pennsylvania Medicaid, 2016-2018.

	Risk difference (95% CI) per 1,000 live births at each month of follow-up[1]				
	1 month	6 months	12 months	18 months	26 months
Heart failure					
Severe maternal morbidity	1.2 (0.50, 1.9)	3.4 (2.0, 5.1)	5.8 (3.4, 8.2)	8.0 (4.7, 11)	12 (6.2, 18)
No severe maternal morbidity	Ref	Ref	Ref	Ref	Ref
Ischemic heart disease					
Severe maternal morbidity	0.5 (0.06, 1.0)	2.0 (0.7, 3.2)	3.8 (1.5, 6.0)	5.3 (2.2, 8.3)	6.4 (1.7, 11)
No severe maternal morbidity	Ref	Ref	Ref	Ref	Ref
Stroke/transient ischemic attack					
Severe maternal morbidity	1.0 (0.3, 1.7)	3.5 (1.8, 5.3)	6.2 (3.5, 8.8)	7.8 (4.4, 11)	8.2 (3.2, 13)
No severe maternal morbidity	Ref	Ref	Ref	Ref	Ref
Any cardiovascular event					
Severe maternal morbidity	2.7 (1.6, 3.8)	9.3 (6.5, 12)	17 (12, 21)	23 (17, 28)	28 (19, 37)
No severe maternal morbidity	Ref	Ref	Ref	Ref	Ref

The figure below is inserted so that there is an item in the sample List of Figures.

2.6 Supplementary tables and figures

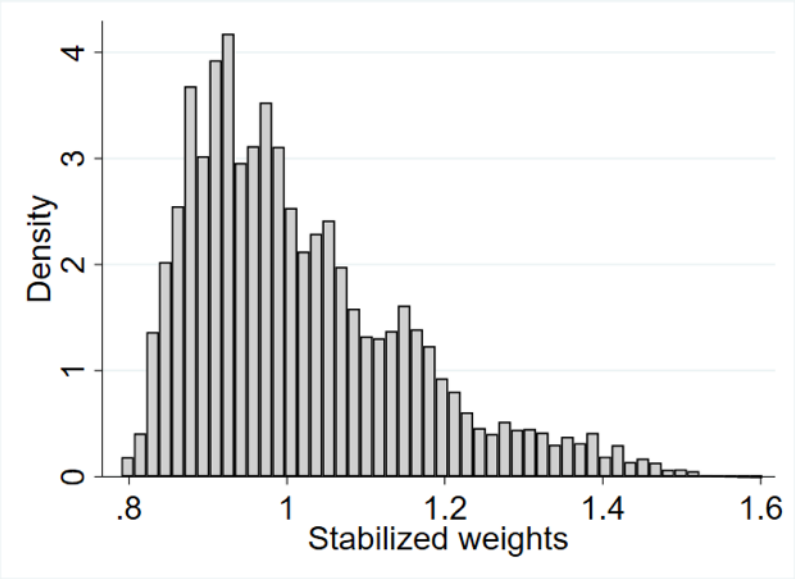

Figure 2.C. Histogram of stabilized inverse probability of censoring weights.

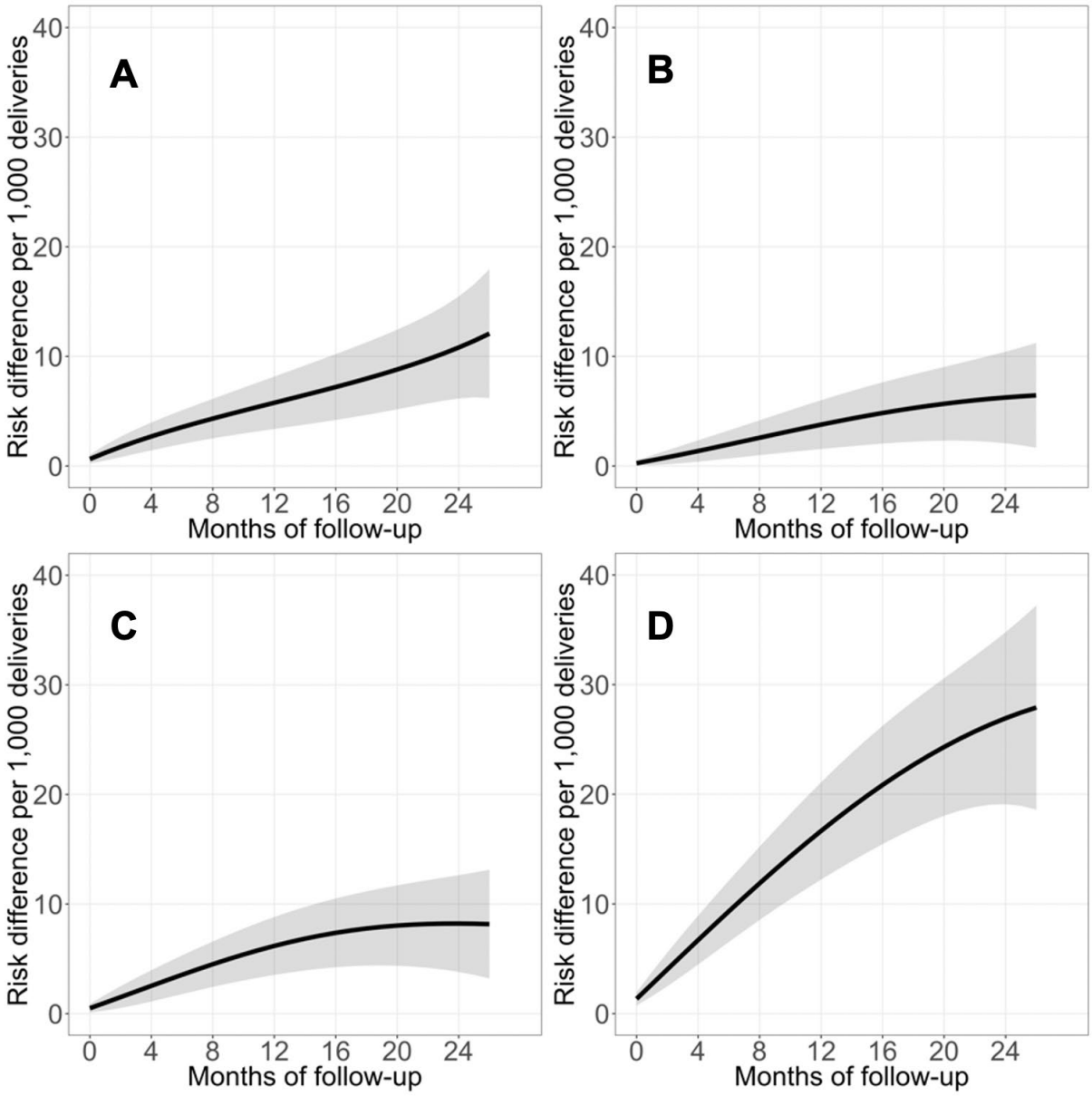

Figure 2.D. Adjusted risk differences per 1,000 live births for cardiovascular events under SMM and no SMM (referent), Pennsylvania Medicaid, 2016-2018, N = 139,531 deliveries.

Table 9. ICD-10 CM codes for SMM diagnoses and procedures.

Indicator	Diagnosis or procedure code
Acute myocardial infarction	I21.xx, I22.x
Aneurysm	I71.xx*, I79.0*
Acute renal failure	N17.x, O90.4

Adult respiratory distress syndrome	J80, J95.1, J95.2, J95.3, J95.82x, J96.0x, J96.2x R09.2
Amniotic fluid embolism	O88.1x
Cardiac arrest/ventricular fibrillation	I46.x, I49.0x
Conversion of cardiac rhythm	5A2204Z, 5A12012
Disseminated intravascular coagulation	D65, D68.8, D68.9, O72.3
Eclampsia	O15.X, O14.22
Heart failure/arrest during procedure/surgery	I97.12x, I97.13x, I97.710, I97.711
Puerperal cerebrovascular disorders	I60.xx- I68.xx, O22.51, O22.52, O22.53, I97.81x, I97.82x, O87.3 I62.9 – included but should not be captured if this is not a valid code.
Pulmonary edema/acute heart failure	J81.0, I50.1, I50.20, I50.21, I50.23, I50.30, I50.31, I50.33, I50.40, I50.41, I50.43, I50.9
Severe anesthesia complications	O74.0 , O74.1 , O74.2, O74.3, O89.0x, O89.1, O89.2
Sepsis	O85, O86.04, T80.211A, T81.4XXA, T81.44xx; or R65.20; or A40.x, A41.x, A32.7
Shock	O75.1, R57.x, R65.21, T78.2XXA, T88.2 XXA, T88.6 XXA, T81.10XA , T81.11XA, T81.19XA
Sickle cell disease with crisis	D57.0x, D57.21x, D57.41x, D57.81x
Air and thrombotic embolism	I26.x, O88.0x, O88.2x, O88.3x, O88.8x
Blood products transfusion	30233H1, 30233L1, 30233K1, 30233M1, 30233N1, 30233P1, 30233R1, 30233T1,30233H0, 30233L0, 30233K0, 30233M0, 30233N0, 30233P0, 30233R0, 30233T0,30230H1, 30230L1, 30230K1, 30230M1, 30230N1, 30230P1, 30230R1, 30230T1,30230H0, 30230L0, 30230K0, 30230M0, 30230N0, 30230P0, 30230R0, 30230T0, 30240H1, 30240L1, 30240K1, 30240M1, 30240N1, 30240P1, 30240R1, 30240T1,30240H0, 30240L0, 30240K0, 30240M0, 30240N0, 30240P0, 30240R0, 30240T0, 30243H1, 30243L1, 30243K1, 30243M1, 30243N1, 30243P1, 30243R1, 30243T1, 30243H0, 30243L0, 30243K0, 30243M0, 30243N0, 30243P0, 30243R0, 30243T0, 30250H1, 30250L1, 30250K1, 30250M1, 30250N1, 30250P1, 30250R1, 30250T1, 30250H0, 30250L0, 30250K0, 30250M0, 30250N0, 30250P0, 30250R0, 30250T0, 30253H1, 30253L1, 30253K1, 30253M1, 30253N1, 30253P1, 30253R1, 30253T1, 30253H0, 30253L0, 30253K0, 30253M0, 30253N0, 30253P0, 30253R0, 30253T0, 30260H1, 30260L1, 30260K1, 30260M1, 30260N1, 30260P1, 30260R1, 30260T1, 30260H0, 30260L0, 30260K0, 30260M0, 30260N0, 30260P0, 30260R0, 30260T0, 30263H1, 30263L1, 30263K1, 30263M1, 30263N1, 30263P1, 30263R1, 30263T1, 30263H0, 30263L0, 30263K0, 30263M0, 30263N0, 30263P0, 30263R0, 30263T0’
Hysterectomy	0UT90ZZ, 0UT94ZZ, 0UT97ZZ, 0UT98ZZ, 0UT9FZZ
Temporary tracheostomy	0B110Z, 0B110F, 0B113, 0B114

Ventilation	5A1935Z, 5A1945Z, 5A1955Z

Table 10. ICD-10 CM codes for adverse cardiovascular outcomes.

Outcome event	ICD-10 codes
Atrial fibrillation	I48.0, I48.1, I48.2, I48.91 (ONLY first or second DX on the claim)
Heart failure	I09.81, I11.0, I13.0, I13.2, I50.1, I50.20, I50.21, I50.22, I50.23, I50.30, I50.31, I50.32, I50.33, I50.40, I50.41, I50.42, I50.43, I50.810, I50.811, I50.812, I50.813, I50.814, I50.82, I50.83, I50.84, I50.89, I50.9 (any DX on the claim)
Ischemic heart disease	I20.0, I20.1, I20.8, I20.9, I21.01, I21.02, I21.09, I21.11, I21.19, I21.21, I21.29, I21.3, I21.4, I21.A1, I21.A9, I22.0, I22.1, I22.2, I22.8, I22.9, I23.0, I23.1, I23.2, I23.3, I23.4, I23.5, I23.6, I23.7, I23.8, I24.0, I24.1, I24.8, I24.9, I25.10, I25.110, I25.111, I25.118, I25.119, I25.2, I25.3, I25.41, I25.42, I25.5, I25.6, I25.700, I25.701, I25.708, I25.709, I25.710, I25.711, I25.718, I25.719, I25.720, I25.721, I25.728, I25.729, I25.730, I25.731, I25.738, I25.739, I25.750, I25.751, I25.758, I25.759, I25.760, I25.761, I25.768, I25.769, I25.790, I25.791, I25.798, I25.799, I25.810, I25.811, I25.812, I25.82, I25.83, I25.84, I25.89, I25.9 (any DX on the claim)
Stroke/transient ischemic attack	G45.0, G45.1, G45.2, G45.8, G45.9, G46.0, G46.1, G46.2, G46.3, G46.4, G46.5, G46.6, G46.7, G46.8, G97.31, G97.32, I60.00, I60.01, I60.02, I60.10, I60.11, I60.12, I60.20, I60.21, I60.22, I60.30, I60.31, I60.32, I60.4, I60.50, I60.51, I60.52, I60.6, I60.7, I60.8, I60.9, I61.0, I61.1, I61.2, I61.3, I61.4, I61.5, I61.6, I61.8, I61.9, I63.00, I63.02, I63.011, I63.012, I63.013, I63.019, I63.02, I63.031, I63.032, I63.039, I63.09, I63.10, I63.111, I63.112, I63.119, I63.12, I63.131, I63.132, I63.139, I63.19, I63.20, I63.211, I63.212, I63.213, I63.219, I63.22, I63.231, I63.232, I63.233, I63.239, I63.29, I63.30, I63.311, I63.312, I63.313, I63.319, I63.321, I63.322, I63.323, I63.329, I63.331, I63.332, I63.333, I63.339, I63.341, I63.342, I63.343, I63.349, I63.39, I63.40, I63.411, I63.412, I63.413, I63.419, I63.421, I63.422, I63.423, I63.429, I63.431, I63.432, I63.433, I63.439, I63.441, I63.442, I63.443, I63.449, I63.49, I63.50, I63.511, I63.512, I63.513, I63.519, I63.521, I63.522, I63.523, I63.529, I63.531, I63.532, I63.533, I63.539, I63.541, I63.542, I63.543, I63.549, I63.59, I63.6, I63.8, I63.9, I66.01, I66.02, I66.03, I66.09, I66.11, I66.12, I66.13, I66.19, I66.21, I66.22, I66.23, I66.29, I66.3, I66.8, I66.9, I67.841, I67.848,

	I67.89, I97.810, I97.811, I97.820, I97.821 (any DX on the claim)

Table 11. Identification of cesarean deliveries.

Step	Type of code	Codes
1.Identify cesarean delivery	DRG	0370, 0371, 0540
2.For any missing, identify cesarean delivery from infant codes	ICD-10	Z3801, Z3831, Z3862, Z3864, Z3869
3.Set mode of delivery to "vaginal" for any remaining missing		

Table 12. Frequency of Centers for Disease Control severe maternal morbidity indicators among deliveries with severe maternal morbidity, Pennsylvania Medicaid, 2016-2018.

Severe maternal morbidity indicator	N (%)
Acute myocardial infarction	96 (1.6)
Acute renal failure	456 (7.8)
Acute respiratory distress syndrome	479 (8.2)
Air or thrombotic embolism	489 (8.4)
Amniotic fluid embolism	56 (1.0)
Aneurysm	29 (0.50)
Blood transfusion	1229 (21)
Cardiac arrest/ventricular fibrillation	78 (1.3)
Conversion of cardiac rhythm	17 (0.29)
Disseminated Intravascular Coagulation	509 (8.7)
Eclampsia	1016 (17)
Heart failure or arrest during procedure or surgery	<10 (0.17)[1]
Hysterectomy	109 (1.9)
Intensive care unit (ICU) admission	1188 (20)
Puerperal cerebrovascular disease	434 (7,4)
Pulmonary edema/acute heart failure	508 (8.7)
Sepsis	811 (14)
Severe anesthesia complications	22 (0.38)
Shock	226 (3.9)
Sickle cell with crisis	41 (0.70)
Temporary tracheostomy	<10 (0.17)[1]

Ventilation	137 (2.3)
Total[2]	5832 (100)

Table 13. Risk per 1000 deliveries of adverse cardiovascular events across follow-up for severe maternal morbidity vs. no severe maternal morbidity, Pennsylvania Medicaid, 2016-2018.

Severe maternal morbidity	Risk at 1 month	Risk at 6 months	Risk at 12 months	Risk at 18 months	Risk at 26 months
Heart failure					
Yes	1.5 (0.76, 2.2)	4.3 (2.7, 5.8)	6.9 (4.5, 9.3)	9.5 (6.2, 13)	14 (8.5, 20)
No	0.26 (0.18, 0.33)	0.72 (0.56, 0.88)	1.1 (0.90, 1.4)	1.5 (1.2, 1.9)	2.4 (1.7, 3.0)
Ischemic heart disease					
Yes	0.79 (0.35, 1.2)	3.1 (1.8, 4.3)	6.0 (3.8, 8.2)	8.7 (5.7, 12)	11 (6.6, 16)
No	0.28 (0.21, 0.35)	1.1 (0.91, 1.3)	2.2 (1.9, 2.6)	3.5 (3.0, 4.0)	4.9 (4.1, 5.8)
Stroke/transient ischemic attack					
Yes	1.2 (0.52, 2.0)	4.5 (2.7, 6.3)	8.2 (5.6, 11)	11 (7.6, 14)	13 (8.1, 18)
No	0.25 (0.17, 0.32)	0.97 (0.79, 1.1)	2.0 (1.7, 2.3)	3.2 (2.7, 3.7)	4.8 (4.0, 8.2)
Any cardiovascular event					
Yes	3.5 (2.5, 4.6)	12 (9.7, 15)	23 (18, 27)	32 (26, 37)	41 (32, 50)
No	0.86 (0.73, 1.0)	3.1 (2.8, 3.4)	5.9 (5.4, 6.4)	8.9 (8.1, 9.7)	13 (12, 15)

Table 14. Excess risk of cardiovascular events per 1,000 live births for severe maternal morbidity vs. no severe maternal morbidity, excluding deliveries whose only SMM indicator was blood transfusion, Pennsylvania Medicaid, 2016-2018.

	Risk difference (95% CI) per 1,000 live births at each month of follow-up[1]				
Severe maternal morbidity	1 month	6 months	12 months	18 months	26 months
Heart failure					
Yes	1.4 (0.62, 2.3)	4.2 (2.3, 6.0)	6.9 (4.1, 9.7)	9.4 (5.6, 13)	14 (7.4, 21)
No	Ref	Ref	Ref	Ref	Ref
Ischemic heart disease					
Yes	0.63 (0.11, 1.1)	2.4 (0.94, 3.9)	4.7 (2.1, 7.2)	6.5 (3.0, 10)	8.0 (2.6, 13)
No	Ref	Ref	Ref	Ref	Ref
Stroke/transient ischemic attack					
Yes	1.1 (0.30, 2.0)	4.0 (1.9, 6.1)	6.9 (3.9, 10)	8.9 (4.9, 13)	9.7 (4.0, 15)
No	Ref	Ref	Ref	Ref	Ref
Any cardiovascular event					
Yes	3.2 (1.9, 4.5)	11 (7.7, 14)	20 (15, 25)	27 (20, 33)	34 (23, 44)
No	Ref	Ref	Ref	Ref	Ref

Table 15. Excess risk of cardiovascular events per 1,000 live births for severe maternal morbidity vs. no severe maternal morbidity, assuming term gestation of 38 weeks, Pennsylvania Medicaid, 2016-2018.

Severe maternal morbidity	Risk difference (95% CI) per 1,000 live births at each month of follow-up[1]				
	1 month	6 months	12 months	18 months	26 months
Heart failure					
Yes	1.4 (0.57, 2.3)	4.0 (2.1, 5.6)	6.2 (3.5, 8.9)	8.4 (4.8, 12)	13 (6.3, 19)
No	Ref	Ref	Ref	Ref	Ref
Ischemic heart disease					
Yes	0.34 (-0.18, 0.87)	1.6 (-0.002, 3.2)	3.5 (0.83, 6.2)	5.2 (1.7, 8.8)	6.4 (1.4, 11)
No	Ref	Ref	Ref	Ref	Ref
Stroke/transient ischemic attack					
Yes	1.1 (0.33, 1.9)	4.2 (2.1, 6.2)	7.5 (4.0, 11)	9.3 (5.0, 14)	9.6 (4.0, 15)
No	Ref	Ref	Ref	Ref	Ref
Any cardiovascular event					
Yes	2.9 (1.6, 4.3)	10 (6.7, 14)	18 (12, 23)	24 (17, 31)	29 (18, 40)
No	Ref	Ref	Ref	Ref	Ref

Table 16. Excess risk of cardiovascular events per 1,000 live births for severe maternal morbidity vs. no severe maternal morbidity, assuming term gestation of 42 weeks, Pennsylvania Medicaid, 2016-2018.

Severe maternal morbidity	Risk difference (95% CI) per 1,000 live births at each month of follow-up[1]				
	1 month	6 months	12 months	18 months	26 months
Heart failure					
Yes	1.4 (0.58, 2.3)	4.1 (2.3, 5.8)	6.4 (3.7, 9.1)	8.7 (5.1, 12)	13 (6.3, 19)
No	Ref	Ref	Ref	Ref	Ref
Ischemic heart disease					
Yes	0.34 (-0.11, 0.78)	1.6 (0.07, 3.1)	3.5 (0.62, 6.3)	5.1 (1.2, 9.0)	6.2 (0.79, 12)
No	Ref	Ref	Ref	Ref	Ref
Stroke/transient ischemic attack					
Yes	1.1 (0.31, 1.9)	4.1 (2.0, 6.3))	7.3 (3.7, 11)	9.2 (4.6, 14)	9.5 (4.0, 15)
No	Ref	Ref	Ref	Ref	Ref
Any cardiovascular event					
Yes	2.9 (1.7, 4.2)	10 (7.0, 13)	19 (14, 24)	25 (18, 32)	29 (18, 39)
No	Ref	Ref	Ref	Ref	Ref

3.0 Chapter 3: The impact of undersampling on the predictive performance of logistic regression and machine learning algorithms: A simulation study.

3.1 Introduction

Machine learning techniques may improve risk prediction and disease screening. Class imbalance (ratio of non-cases to cases > 1) routinely occurs in epidemiologic data and may degrade the predictive performance of machine learning algorithms.[95,142-144] Of the many techniques developed to address class imbalance,[145,146] here we investigated simple undersampling. This method is straightforward and accessible, but evidence on its performance is mixed and practical guidance is needed. Using simulated data, we assessed the predictive performance of the ensemble machine learning algorithm SuperLearner and logistic regression in imbalanced and undersampled data to investigate whether undersampling alters predictive accuracy.

3.2 Methods

3.2.1 Data-generating mechanism

We used Monte Carlo simulation with 4 groups of 1000 Monte Carlo samples each. We simulated each Monte Carlo sample to have a sample size of 1000, 10 independent standard normal covariates generated from a random uniform distribution, and 10 independent dichotomous covariates generated from a binomial distribution. A dichotomous outcome was simulated from a

logistic regression model conditional on all 20 covariates. Parameters were chosen to lie between -1 and 1 for the continuous variables, and the outcome prevalence was set to lie between 0.15 and 0.50.

3.2.2 Study design

In 2 of the 4 groups of Monte Carlo samples, we left all samples unbalanced. In the remaining 2 groups, we performed undersampling to balance each sample by randomly selecting a number of non-cases equal to the number of cases. To avoid overfitting, we split each Monte Carlo sample into training (70%) and testing (30%) sets with similar outcome prevalences.[147] We generated predicted probabilities on 1000 undersampled and 1000 unbalanced samples parametrically via logistic regression and nonparametrically via stacking (SuperLearner).[144] SuperLearner was implemented with 10-fold cross-validation and a library of 5 algorithms with default tuning parameters: extreme gradient boosting, random forests, kernel k-nearest neighbors, kernel support vector machines, and penalized regression (LASSO). Logistic regression was implemented as a generalized linear model with binomial variance and a logit link function. Average performance metrics (sensitivity, specificity, positive and negative predictive value, and overall accuracy) were evaluated across all 1000 Monte Carlo samples in each group using a classification threshold close to the outcome prevalence of 0.2 for unbalanced groups and 0.5 for undersampled groups. Areas under the receiver operating curve for each sample were generated using the roc() function in the "pROC" package.[148] All analyses were conducted using R version 3.6.1.

3.3 Results

Figure 3.A shows the receiver operating characteristic (ROC) curves for all 1000 Monte Carlo samples in each group and average predictive performance metrics. Areas under the curve across all Monte Carlo samples were similar for all groups. Performance metrics were higher for logistic regression than SuperLearner regardless of data preprocessing method except sensitivity and positive predictive value, which were higher for SuperLearner than logistic regression. Undersampling did not have a substantial impact on logistic regression performance; however, undersampling improved SuperLearner accuracy, specificity, and positive predictive value and worsened SuperLearner sensitivity and negative predictive value. Repeating the analysis with a lower outcome prevalence (2%-10%) did not substantially affect the results.

3.4 Discussion

We observed generally more accurate predictive performance with logistic regression than with SuperLearner regardless of data preprocessing method. This is expected because we simulated our data from a logistic model. However, SuperLearner performed nearly as well on average as the true data-generating mechanism although logistic regression was intentionally excluded from the SuperLearner library. In our simulations, undersampling did not dramatically improve predictive performance, suggesting that ensemble machine learning can achieve adequate performance in similar settings with moderate class imbalance. These results provide some insight on the optimal use of machine learning for predicting imbalanced outcomes.

3.5 Tables and figures

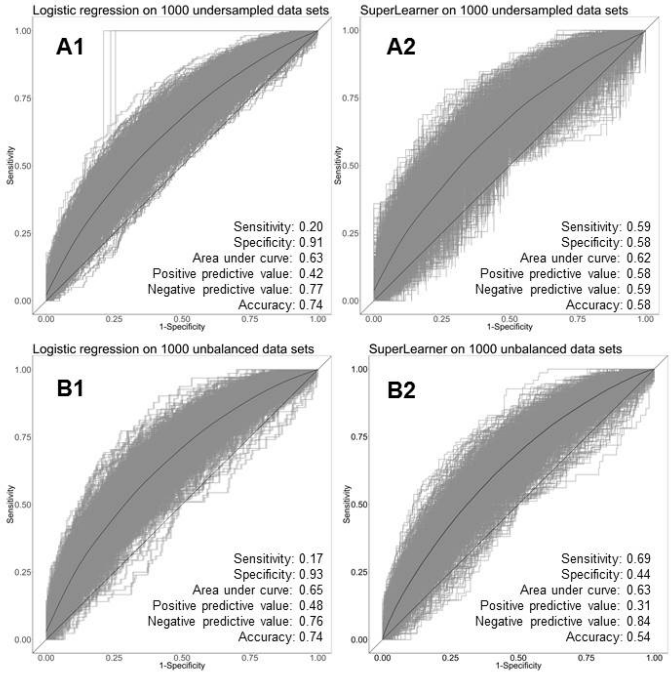

Figure 3.A Receiver operating characteristics (ROC) curves and predictive performance of each simulated data set, N = 1000.

Table 17. Average performance metrics of SuperLearner and logistic regression over 1000 Monte Carlo samples by data preprocessing method.

	Undersampled		Unbalanced	
	SuperLearner	Logistic regression	SuperLearner	Logistic regression
Area under the receiver operating curve	0.62	0.63	0.63	0.65
Sensitivity	0.59	0.20	0.69	0.17
Specificity	0.58	0.91	0.44	0.93
Positive predictive value	0.58	0.42	0.31	0.48
Negative predictive value	0.59	0.77	0.84	0.76
Accuracy	0.58	0.74	0.54	0. 74

4.0 Chapter 4: Ensemble machine learning for severe maternal morbidity identification

4.1 Introduction

The United States' maternal mortality rate is the highest of any comparably high-income country, but the small absolute number of deaths each year and inconsistencies in reporting pose serious challenges to maternal mortality research. Severe maternal morbidity (SMM) is a broad designation for severe adverse complications in peripartum period that do not result in death.[3] Because SMM and maternal mortality share risk factors and etiologies,[19] SMM is a proxy for maternal mortality. Further, SMM itself is important to study because it is costly and associated with poor maternal health outcomes.[119,149] Accurate measurement of SMM is important for research and for guiding maternal care quality improvement efforts. Evidence suggests that available tools used to classify SMM may not accurately quantify population or hospital-level prevalence of SMM.

There is no global consensus on what constitutes a SMM and accordingly, its definitions vary.[3,33,46,150] In 2016, the American College of Obstetricians and Gynecologists/Society for Maternal-Fetal Medicine (ACOG/SMFM) issued guidelines for severe maternal morbidity identification at the hospital level.[42] This "screen and review" procedure involves identifying putative SMM cases by screening with a few simple criteria (transfusion of > 3 units of blood products or ICU admission), then performing medical record review on screen-positive cases according to a more extensive set of clinical criteria. In North America, the most commonly-used SMM definition for studies using administrative or claims databases is one developed by the Centers for Disease Control and Prevention (CDC),[1-3] consisting of 21 diagnoses and procedures

with corresponding International Classification of Diseases (ICD) codes. This list of diagnoses and procedures is often combined with other nonspecific indicators of severe complications, such as intensive care unit (ICU) admission, prolonged postpartum length of stay (PPLOS), and/or transfusion of > 3 units of blood products.[2] Although the validity of these SMM identification criteria (henceforth "screening criteria") have not been extensively evaluated, low positive predictive value is a concern.[51,132]

The ACOG/SMFM screen-and-review guidelines are intended to improve ascertainment of true SMM cases at the hospital level, unlikely to be scaled up without dedication of extensive resources. Machine learning methods may provide an alternative approach to identifying cases of SMM. Reflecting the increasing popularity of machine learning methods for predictive modeling in epidemiology,[8,92,151] two recent papers sought to identify SMM cases from administrative and hospital data using these methods. One developed an expanded version of an existing obstetric comorbidity index[54] to predict screen-positive SMM across administrative data settings using techniques that incorporate machine learning.[55] The other aimed to identify true-positive (medical record-reviewed) SMM cases using cross-validated regression modeling in a machine learning framework with a large number of predictors derived from electronic health records.[53] To our knowledge, researchers have not attempted to improve the identification of true SMM cases by leveraging ensemble machine learning techniques, which "stack" multiple machine learning algorithms to generate optimal predictions[92] Our objective was to retrospectively identify true SMM cases and non-cases using the ensemble machine learning algorithm SuperLearner[144] in a rich hospital database, and to compare the predictive performances of these algorithms and screening criteria.

4.2 Methods

4.2.1 Data source

Data were obtained from the Magee Obstetric Maternal and Infant (MOMI) database, a detailed perinatal database of deliveries at Magee-Womens Hospital, University of Pittsburgh Medical Center, Pittsburgh, PA. MOMI is populated using data from medical records, billing, outpatient encounters, admitting services, ultrasound, and other ancillary systems. This study was approved under expedited review by the Institutional Review Board at the University of Pittsburgh (IRB #STUDY19030089).

Live, singleton deliveries occurring at Magee-Womens Hospital from 2010-2011 (N = 19,266) and from 2013-2017 (N = 47,067) were eligible for inclusion. We constructed two subcohorts for training and testing our SuperLearner ensemble algorithms by sampling 685 records from 2010-2011 and 498 records from 2013-2017. To sample records from 2010-2011 deliveries, we applied the screening criteria (described below and in Table 21) to the cohort and selected all screen-positive deliveries (n = 336) and a random sample of screen-negative deliveries (n = 349). A similar process was followed to sample records from 2013-2017, but in this case random samples of screen-positive and screen negative deliveries were drawn for each year for a total of n = 250 and n = 258. All sampled records underwent medical record review, as described below. A schematic diagram detailing the construction of each subcohort is shown in Figure 4.A.

4.2.2 Severe maternal morbidity

We screened deliveries in both cohorts using the screening criteria outlined in Table 21. Deliveries screened positive for SMM if they had any of the CDC list of 21 diagnosis or procedure codes for SMM identification,[1] maternal intensive care unit (ICU) admission, or PPLOS (> 3 standard deviations above the mean length of stay by mode of delivery).[51] Screen-negative deliveries were those with no screening or identification criteria for SMM.

We defined true SMM status according to the ACOG/SMFM Obstetric Care Consensus document guidelines (Table 22). After screening and sampling, both screen-positive and screen-negative deliveries underwent detailed medical record abstraction using these guidelines to adjudicate whether each delivery was a true case or non-case. The protocol for jurying each case has been described in detail previously;[51] briefly, an experienced medical record abstractor performed standardized chart abstractions into a custom data entry system designed to collect all clinical data necessary to jury cases according to the ACOG/SMFM guidelines. A maternal-fetal medicine specialist (KPH) reviewed every 10th case and the false-negative cases and provided feedback to the research analyst to improve case ascertainment. This so-called true case designation is the outcome we used our SuperLearner ensemble algorithms to predict.

4.2.3 Predictors of SMM

Predictors were chosen in consultation with a maternal-fetal medicine specialist (KPH) with the goal of maximizing the predictive accuracy of our ensemble machine learning algorithm. First, we identified a set of variables *a priori* as candidates for use in prediction (Table 23). Because SMM is associated with adverse birth outcomes[42] and because the goal of this prediction

algorithm is to retroactively identify true SMM cases from medical record data available after delivery, we included intrapartum and infant characteristics in the predictor set. These candidate variables fell into four broad categories: maternal demographic and behavioral variables (*e.g.* race/ethnicity, insurance type, age, education), maternal health history variables (*e.g.* chronic hypertension or diabetes, anemia, renal disease, depression), labor and delivery variables (*e.g.* mode of delivery, whether labor was induced, whether general anesthesia was administered for delivery), and fetal and infant characteristics (*e.g.* 1- and 5-minute APGAR scores, preterm delivery, birthweight, respiratory distress). These data were ascertained from medical record coding and abstraction, electronic birth records, and other ancillary data systems.

4.2.4 Statistical analysis

A critical tenet of predictive modeling, the "firewall principle," is that any predictive algorithm be evaluated on data not used to train or develop it.[152] To accomplish this separation, we used two temporally distinct subcohorts of the MOMI data for training our ensemble algorithms (2010-2011) and testing their performance (2013-2017). No data that were used to train any of our algorithms were included in the testing process.

We used the recipes() package for R (version 4.0.2) to apply the same data preprocessing and feature engineering blueprint to both the training (n = 685) and test (n = 498) subcohorts according to the following steps. First, we assessed missingness of data in the subcohorts. We excluded 3 predictors with >20% missing (maternal prepregnancy weight, smoking, and type of labor onset) from the predictor set. Second, for any remaining predictors with missing data, we performed mode imputation of all categorical predictors and mean imputation of all continuous predictors. Third, we transformed, centered, and scaled all continuous predictors and created

indicator variables for all categorical predictors. We added 0.01 to zero values of NICU length of stay to enable log transformation; Box-Cox transformations were applied to all other continuous predictors. We also prepared a version of this dataset after converting all continuous predictors to cubic splines for use with generalized additive models, one of the individual base learners in our SuperLearner ensemble algorithms in the main analysis.

We developed 7 SuperLearner ensemble algorithms[144] on the training data in order to predict true SMM case status in the test data. Each ensemble algorithm used different predictor sets and/or incorporated different elements and specifications of the screening criteria as base learners in the ensembles. We used a rich library of different base learner algorithms and tuning parameters (Table 24) for each of the ensemble algorithms we evaluated. These ensemble algorithms, 5 of which are presented in the main analyses and 2 of which are presented in supplementary material, are described in Table 18. The screening criteria alone were not an ensemble algorithm, did not require base learners, and made use of only of the CDC diagnosis and procedure codes, ICU admission, and PPLOS. Ensemble 1 used all available predictors but only one base learner, a generalized linear model with the screening criteria as independent variables. Ensemble 1 is an adaptation of the screening criteria that can be used to assess receiver operating characteristic (ROC) and precision-recall performance. Ensembles 2-4 used all base learners presented in Table 24 with different predictor sets (MOMI predictors only, MOMI predictors plus CDC diagnoses/procedures, and MOMI predictors plus CDC diagnoses/procedures plus ICU admission and PPLOS). Ensemble 5 incorporated the screening criteria as an additional base learner in the ensemble algorithm. Finally, Ensembles 6 and 7 are visualized in Table 18 but yielded results similar to the other ensemble algorithms; consequently, results for Ensembles 6 and 7 are shown in the Appendix. Ensemble 6 left the CDC list of diagnoses and procedures out of the

predictor set, and Ensemble 7 used the full predictor set with a linear model including the screening criteria predictors as independent variables as a base learner.

To generate predicted probabilities of true-positive SMM, we used the SuperLearner ensemble algorithm with 10-fold cross validation and the rank loss function,[144] which minimizes the complement of the area under the receiver operating characteristic (ROC) curve (1-AUC).

For the main analysis, each of these 7 SuperLearner ensemble algorithms (referred to hereafter as "ensemble algorithms" or "ensembles") was fit using a variable selection algorithm applied together with each of 3 base learners based on linear models: Bayesian generalized linear models, generalized linear models, and generalized additive models. This variable selection algorithm selected, for inclusion in each of these base learner models, the top 20 predictors most highly correlated with the outcome. We also performed sensitivity analyses, repeating all analyses but conducting no variable selection.

Predictive performance of the ensemble algorithms was evaluated using receiver operating characteristic (ROC) curves and area under the ROC curve (AUC), precision-recall curves and area under the precision-recall curve (PRC), and 6 measures of predictive accuracy: overall accuracy (total number correctly classified/total), balanced accuracy (arithmetic mean of sensitivity and specificity), positive predictive value (proportion of true-positives among records classified as positive), negative predictive value (proportion of true-negatives among records classified as negative), and detection rate (total true-positives/total). Predicted probabilities ≥ 0.5 from the ensembles were classified as 1 (SMM case) and those < 0.5 were classified as 0 (SMM non-case). The screening criteria, which do not yield predicted probabilities of the outcome but rather classify records as cases or non-cases based on the presence of any of the qualifying conditions, were also used to generate measures of predictive accuracy. We also evaluated

predictive accuracy when the classification threshold was lowered to 0.3. Finally, we used the R package tlverse() (function sl3())[153] to generate variable importance measures using a binomial likelihood loss function. This function generates, for each covariate, an importance measure that is based on the difference in the loss when that covariate is omitted from the SuperLearner fit versus when the covariate is included. We obtained variable importance measures using a modified version of Ensemble 3 excluding *k*-nearest-neighbors from the set of base learners.

4.3 Results

The full cohorts from which the training and test subcohorts were drawn were comparable in terms of descriptive characteristics (Table 25). Medical record validation of screen-positive and screen-negative deliveries resulted in a negative predictive value of 0.99 in both the training and test datasets; positive predictive values were 0.51 in the training set and 0.64 in the test set (Tables 26-28). The true-positive SMM cases (n=171 in the training n=160 in the test sets) and true-negative SMM deliveries (n=506 in the training and n=337 in the test sets) were used for the remaining analyses.

In both the training and test sets, SMM true-negatives were more likely than true positives to be non-Hispanic white, married, privately insured, normal weight, and to deliver vaginally (Table 19). True-negatives were less likely than true-positives to have preexisting diabetes or hypertension and to deliver preterm, and had higher birthweight.

We evaluated discriminatory ability of each algorithm using ROC curves and precision-recall plots (4.B). Areas under the ROC curve were high for all ensemble algorithms, including Ensemble 1 (an approximation of the screening criteria). The highest-performing models were

Ensembles 4 and 5 (AUC = 0.90) and the lowest-performing were Ensembles 2 and (AUC = 0.83 and AUC = 0.84, respectively). Similarly, Ensembles 4 and 5 exhibited the highest area under the precision-recall curve (PRC = 0.79 and PRC = 0.80, respectively) while Ensembles 2 and 3 exhibited the lowest (PRC = 0.69). These results suggest that addition of ICU admission and PPLOS to the predictor set had the most pronounced effect on improving the performance of the ensemble algorithms.

We also examined predictive performance measures based on predicted classifications. These measures of predictive accuracy (Table 20) were generally similar among the ensemble algorithms; the predictive accuracy of the screening criteria, however, was meaningfully different from that of the ensemble algorithms. With a classification threshold of 0.5, the screening criteria had the highest overall accuracy (0.82), balanced accuracy (0.86), negative predictive value (0.99), and detection rate (0.32), but the lowest positive predictive value (0.64). Compared to the screening criteria, the ensemble algorithms exhibited similar performance in terms of overall accuracy (0.76-0.79), but markedly lower balanced accuracy (0.64-0.69), negative predictive value (0.75-0.78), and detection rate (0.11-0.14). The only measure of predictive performance on which the ensemble algorithms all outperformed the screening criteria was positive predictive value (0.78-0.86). The highest-performing ensemble algorithm in terms of positive predictive value was Ensemble 5, including the screening criteria as a base learner in the library (0.86); this ensemble algorithm also had high accuracy, balanced accuracy, negative predictive value, and detection rate compared to the other ensemble algorithms. The overall and balanced accuracy values indicate that ensemble algorithms and the screening criteria perform moderately well in terms of the total number of correct classifications made, whether positive or negative. However, the detection rates

suggest that the ensemble algorithms achieve higher positive predictive value relative to the screening criteria by classifying fewer records as positive, whether true-positive or false-positive.

When we lowered the classification threshold to 0.30, we observed a similar pattern in the predictive accuracy performance of the ensemble algorithms. The screening criteria are threshold-invariant and have the same values for each of the predictive performance metrics in Table 20 regardless of the threshold. With a lower threshold, overall accuracy of the ensemble algorithms improved slightly (0.75-0.82), as did balanced accuracy (0.69-0.80). Negative predictive values (0.80-0.87) and detection rates (0.17-0.24) also improved. Positive predictive values of the ensembles were lower with a lower classification threshold, however, ranging from 0.67-0.78. Ensemble 5 again exhibited the highest positive predictive value (0.78). .These results indicate that more records were classified as both true positives and false positives when we lowered the classification threshold.

The most "important" variables were ICU admission, PPLOS, CDC diagnosis codes for hysterectomy, severe preeclampsia or eclampsia derived from the medical record (as opposed to the CDC billing diagnosis code for eclampsia), and CDC procedure codes for blood transfusion 4.C. Unsurprisingly, when comparing the most important predictors between true positives and true negatives in both the training and test data sets, there were striking differences in ICU admission (50% and 5.3%, respectively, in the training cohort), PPLOS (26% and 2.9%, respectively, in the training cohort), and other highly important predictors (Table 29).

We performed sensitivity analyses to evaluate the performance of Ensembles 6 and 7. The performance of Ensembles 6 and 7 was similar to the performance of other ensemble algorithms. Both exhibited high AUC performance (0.88 and 0.89, respectively) and PRC performance (0.79 and 0.75, respectively) (Figure 4.D). At a classification threshold of 0.5, Ensembles 6 and 7

exhibited high overall accuracy (0.76-0.79) and positive predictive value (0.82-0.89) relative to the screening criteria, but lower balanced accuracy (0.65-0.69), negative predictive value (0.75-0.77), and detection rate (0.11-0.13) (Table 30). Similar performance to the other algorithms was also evident when the classification threshold was lowered to 0.3: relative to Ensembles 6 and 7 at the higher classification threshold, overall accuracy (0.82-0.83), balanced accuracy (0.77-0.79), negative predictive value (0.84-0.85), and detection rate (0.20-0.22) were higher but positive predictive value (0.84-0.85) was lower.

We also conducted sensitivity analyses not performing variable selection procedures. When we did not perform variable selection, predictive performance of the ensembles was not meaningfully different from the performance of the ensembles incorporating variable selection (Table 31). This was true at both classification thresholds. Areas under the ROC and precision-recall curves were also not meaningfully different from those presented in the main analysis when variable selection algorithms were not applied (Figures 4.E and 4.F).

4.4 Discussion

In two temporally distinct cohorts from the same academic maternity hospital, we demonstrated that our SuperLearner ensemble algorithms generally exhibited higher positive predictive value and lower negative predictive value for SMM identification than existing screening criteria. However, their true-positive detection rate was considerably lower. All of the ensemble algorithms exhibited similar performance; comparison of AUC, precision-recall, and predictive accuracy performance across Ensembles 1-5 suggests that only the inclusion of the ICU admission and PPLOS screening criteria markedly improved the predictive performance of the

ensemble algorithms. ICU admission and PPLOS were also among the variables with the highest variable importance rankings.

Overall accuracy for the ensemble algorithms and the screening criteria was comparable and fairly high, but was achieved differently. The screening criteria achieved good overall accuracy by maximizing negative predictive value and misclassifying some true-negative cases as positive, while the ensemble algorithms achieved good overall accuracy by maximizing positive predictive value and misclassifying some true-positive cases as negative. Finally, ICU admission and PPLOS were the components of the screening criteria that most improved prediction in the ensemble learners, and were also the most highly ranked in terms of variable importance. Effect sizes (risk or odds ratios) of >50 are necessary to use variables for screening;[154,155] although we did not quantify effect sizes, differences in prevalence of the most "important" variables in our predictor set between true-positives and true-negatives are striking (Table 29). While some of the other variables we included in our predictor set (*e.g.,* preterm delivery, mode of delivery) appear to be moderately associated with SMM, the magnitude of these associations is likely not large enough to reliably discriminate between true-positive and true-negative SMM.

To our knowledge, only one published paper has attempted to predict true-positive SMM using machine learning techniques.[53] These investigators used electronic health record data from deliveries at a large academic medical center to screen for SMM based on ICU admission and/or blood transfusion >3 units and reviewed the medical records of screen-positives. They then used cross-validated penalized regression models to predict true-positive SMM using a large number of diagnosis and procedure variables available in the electronic health record. These authors found that their algorithms outperformed the CDC diagnosis and procedure codes alone in terms of area under the ROC curve, sensitivity, and positive predictive value; AUC and sensitivity were high

(AUC = 0.79 for the CDC diagnoses/procedures and 0.94 for the best-performing algorithm; sensitivity = 0.61 for the CDC diagnoses/procedures and 0.77 for the best-performing algorithm) but positive predictive value was low (0.22 and 0.35 for the CDC diagnoses/procedures the best-performing algorithm, respectively). The results of this analysis broadly agree with our own findings, although we used a different machine learning framework (ensemble learning), reviewed screen-negatives in addition to screen-positives to construct our training and test sets, and evaluated a wider range of performance measures. Ensemble machine learning is generally more accurate than individual machine learning algorithms, but may require more preparation and may be less interpretable.[80]

Our ensemble algorithms were built to improve quantification of true-positive SMM from hospital delivery data in a labor-efficient way, not to predict SMM risk among women during antenatal care. However, several recent papers have developed or assessed tools to prospectively predict risk of severe maternal morbidity in pregnant women. One team used a targeted causal inference technique to identify maternal comorbidities that were most predictive of screen-positive SMM among all hospitalizations for live births in California.[55] While the ultimate goal of this paper—to expand on an existing obstetric comorbidity index to identify women at high risk of screen-positive SMM during the delivery hospitalization – was different from ours, the causal inference technique these investigators employed used ensemble machine learning to identify many comorbidities that are highly predictive of SMM. Among the most predictive they found were placenta accreta spectrum disorder, pulmonary hypertension, and chronic renal disease. In the MOMI data, the variable for placenta accreta spectrum disorders was not one of the most "important" variables, and the prevalence of this comorbidity was not as dramatically different between true-positives and true-negatives in either the training or test data. Other groups have

assessed the ability of the original version of the obstetric comorbidity index[54] to identify women at high risk of true-positive SMM[56] and developed a predictive model for risk of maternal intensive care unit admission using prenatal risk factors.[57]

Our work has several important limitations which should inform interpretation of the results. First, we were not able to report sensitivity or specificity because we sampled from our cohort using screening criteria.[156] We therefore do not have an appropriate or interpretable denominator for either measure (true-positive or true-negative SMM in the full cohorts) to be able to evaluate the ability of screening criteria to identify true-positives and true-negatives. We instead reported positive and negative predictive value and other accuracy measures. This means that, giving how our training and test subcohorts were sampled, we can estimate the probability that a case is a true-positive given that it screened positive with either the screening criteria or an ensemble algorithm, but we can't estimate the probability that a case will screen positive given that it is a true-positive. Second, there is some evidence that the degree of imbalance in outcome events (most commonly when non-events outnumber events) can affect the predictive performance of machine learning algorithms, including ensemble algorithms.[95,142,157] We did not take any steps to remedy outcome class imbalance analytically. Since true-negatives outnumbered true-positives in both our training and test sets, and since predicted probabilities of true-positive SMM were low, it is possible that our ensemble algorithms achieved high overall accuracy simply by classifying most records as negative. The high positive predictive values and low true-positive detection rates of the ensemble algorithms also suggest that these algorithms maximized accuracy by classifying most records as negative. Third, the generalizability and transportability of our ensemble algorithms to other settings are limited. Since the algorithms were trained in MOMI data, our conclusions may be generalizable only to other years of the MOMI data for which the same

predictors are available. Overfitting, a common generalizability issue, is a possibility; although we minimized this possibility by using distinct training and test sets and by using 10-fold cross-validation to train our algorithms.[147] Another, more insidious threat to generalizability is so-called contextual bias of the training data. For example, the MOMI data were derived from an academic maternity hospital; our algorithms may thus not be generalizable to data collected in a community clinic or from a claims database. This type of bias is difficult to detect and is not corrected by, *e.g.*, methods to minimize overfitting.[158] Finally, in the absence of clear guidance in the literature on sample size recommendations for classification tasks with ensemble machine learning, we relied on prior simulations and some partially applicable reported results[159,160] to plan our sample sizes. However, it is possible that our sample sizes are inappropriately small given the number and quality of predictors we used.[161]

Analytically, future directions from this work include incorporating cost-sensitive learning procedures into our ensemble algorithms to differentially penalize false-positive or false-negative classifications, and evaluating the effect of techniques to remedy class imbalance on their predictive performance. This work also has clinical, epidemiologic, and public health implications. In terms of clinical practice, our ensemble algorithms should not be used as risk prediction tools. In terms of epidemiologic research, our ensemble algorithms should not be used to estimate prevalence of true-positive SMM. However, since the positive predictive value of our ensemble algorithms was high, this approach might be useful to identify a sample of true-positive SMM cases for research in settings where medical record review may not be feasible. Finally, in terms of public health and policy implications, our results suggest that accurate identification of SMM remains a challenge, and likely will remain a challenge in the absence of a universal definition of SMM or national obstetric surveillance systems in the US.

Here, we demonstrated that ensemble machine learning for SMM identification does not globally improve SMM ascertainment relative to existing screening criteria. While using ensemble machine learning improves some performance metrics, it comes with important tradeoffs.

4.5 Tables and Figures

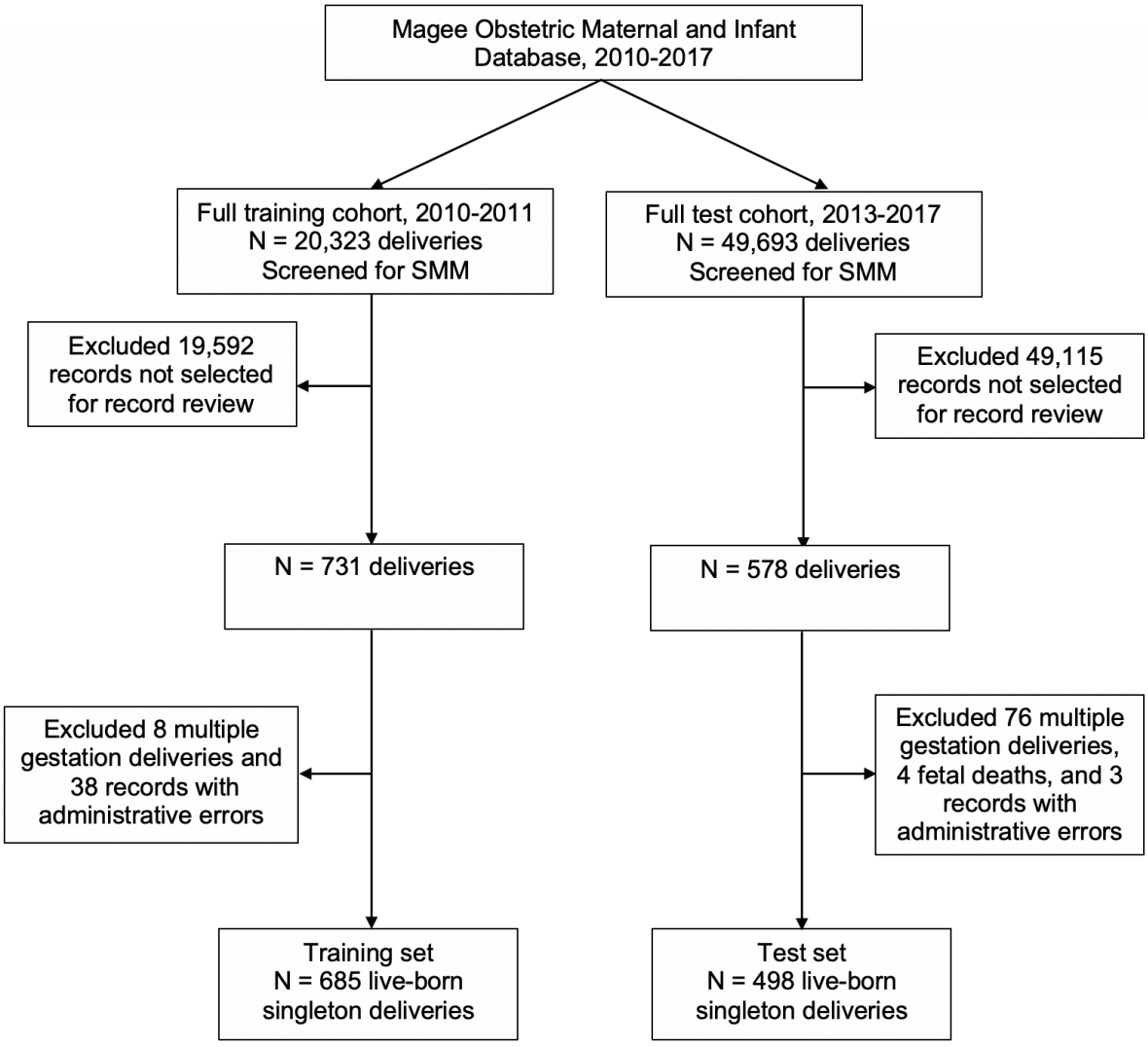

Figure 4.A Study selection flow chart, Magee Obstetric Maternal and Infant Database, 2010-2017.

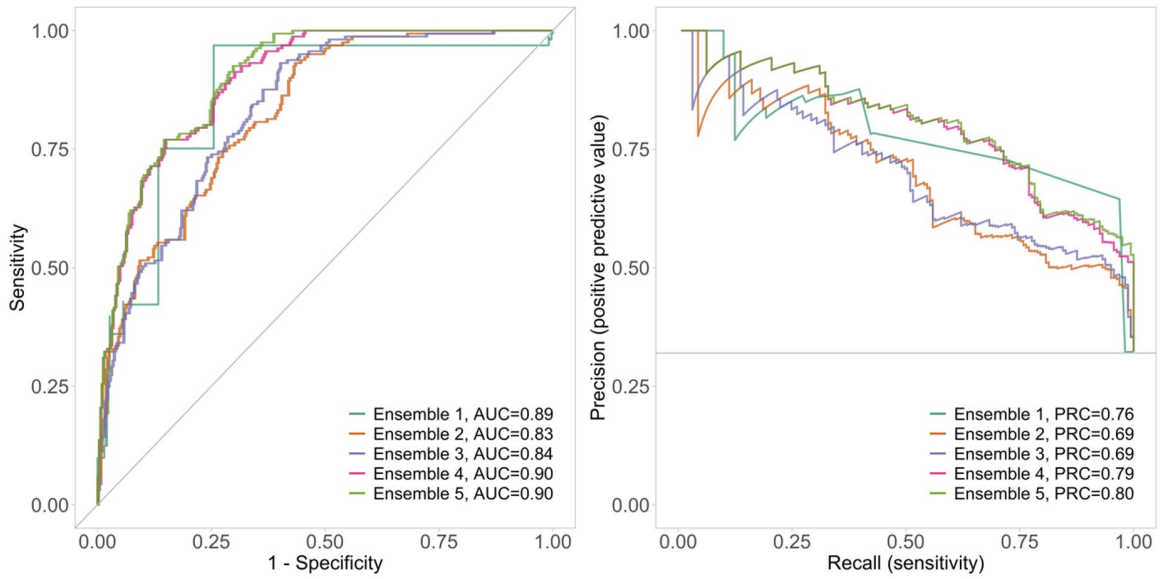

Figure 4.B. Receiver-operating characteristic (ROC) 1 and precision-recall curves for 5 ensemble algorithms, Magee Obstetric Maternal and Infant Database, 2013-2017, N = 498.

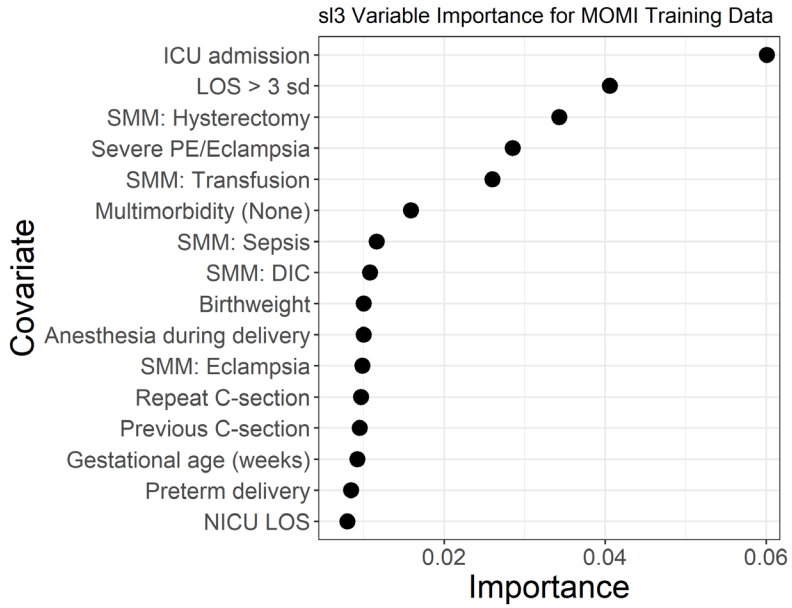

Figure 4.C. Variable importance ranking of predictors for SuperLearner ensemble algorithms, Magee Obstetric Maternal and Infant Database, 2010-2011, N = 685.

Table 18. Predictors and base learners included in each ensemble algorithm designed to predict true-positive SMM, Magee Obstetric Maternal and Infant Database, 2013-2017, N = 498.

	Predictors			Base learners		
		Components of the screening criteria				
	MOMI predictors	CDC diagnosis and procedure codes	ICU admission and PPLOS	SuperLearner library	Screening criteria	Linear model with screening criteria variables
Screening criteria		X	X			
Ensemble 1						X
Ensemble 2	X			X		
Ensemble 3	X	X		X		
Ensemble 4	X	X	X	X		
Ensemble 5	X	X	X	X	X	
Ensemble 6	X		X	X		
Ensemble 7	X	X	X	X		X

Table 19. Characteristics of true-positive severe maternal morbidity cases and true-negative cases in the training (2010-2011) and test (2013-2017) subcohorts, Magee Obstetric Maternal and Infant Database.

	Training subcohort, 2010-2011 (n = 685)		Test subcohort, 2013-2017 (n = 498)	
	True positives (n = 171)	True negatives (n = 506)	True positives (n = 160)	True negatives (n = 337)
Maternal race, %				
NH White	72	78	60	70
NH Black	25	17	30	23
Other	3.5	4.2	10	7.1
Maternal age, mean (sd)	29 (6.2)	29 (6.1)	29 (5.7)	29 (5.3)
Maternal age ≥ 35, %	21	17	19	15
Married, %	47	55	41	54
Maternal education, %				
Less than high school	16	9.0	8.3	8.0
High school graduate	30	21	29	21
Some college	17	15	9.7	13
4-year college graduate	38	54	53	58
Type of insurance, %				
Private	46	62	48	60
Public nor none	54	38	52	40

Nulliparous, %	44	54	47	45
Smoked during pregnancy, %	18	14	12	14
Preexisting diabetes or hypertension, %	18	5.1	22	8.6
Body mass index category, %				
Underweight	3.4	6.6	4.3	3.1
Normal weight	42	50	30	49
Overweight	24	24	18	23
Obese	31	19	48	26
Gestational age, weeks, mean (sd)	35 (4.0)	38 (3.0)	35 (4.6)	38 (2.5)
Preterm birth < 37 weeks, %	56	12	48	14
Mode of delivery, %				
Vaginal	27	61	41	69
Cesarean	73	39	59	31
Birthweight, grams, mean (sd)	2652 (918)	3226 (674)	2526 (1013)	3185 (641)

Table 20. Measures of predictive accuracy for the screening criteria and each ensemble, with variable selection, under two different classification thresholds, Magee Obstetric Maternal and Infant Database, 2013-2017, N = 498.

	Classification threshold = 0.50					Classification threshold = 0.30				
	Accuracy	Balanced accuracy[1]	Positive predictive value	Negative predictive value	Detection rate[2]	Accuracy	Balanced accuracy[1]	Positive predictive value	Negative predictive value	Detection rate[2]
Screening criteria	0.82	0.86	0.64	0.99	0.32	0.82	0.86	0.64	0.99	0.32
Ensemble 1	0.78	0.69	0.78	0.78	0.14	0.82	0.80	0.72	0.87	0.24
Ensemble 2	0.76	0.65	0.79	0.76	0.11	0.77	0.70	0.67	0.80	0.17
Ensemble 3	0.76	0.64	0.81	0.75	0.13	0.75	0.69	0.65	0.80	0.17
Ensemble 4	0.78	0.68	0.85	0.77	0.12	0.82	0.77	0.78	0.84	0.20
Ensemble 5	0.79	0.69	0.86	0.77	0.13	0.82	0.77	0.78	0.84	0.20

4.6 Supplemental tables and figures.

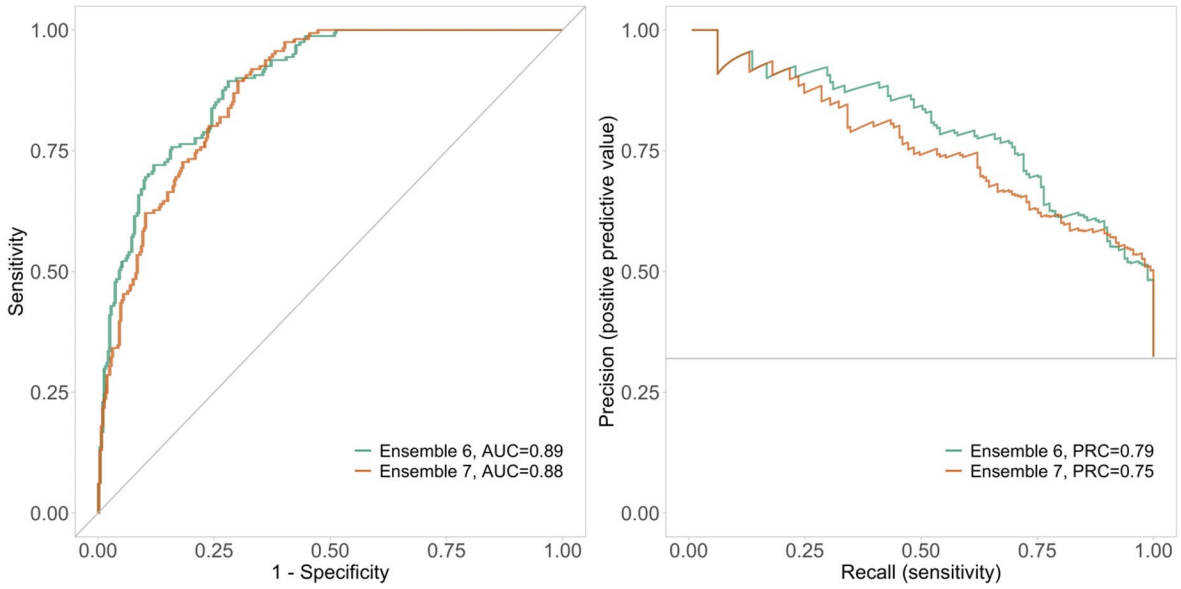

Figure 4.D. Receiver-operating characteristic (ROC) 1 and precision-recall curves for Ensembles 6 and 7, with variable selection, Magee Obstetric Maternal and Infant Database, 2013-2017, N = 498.

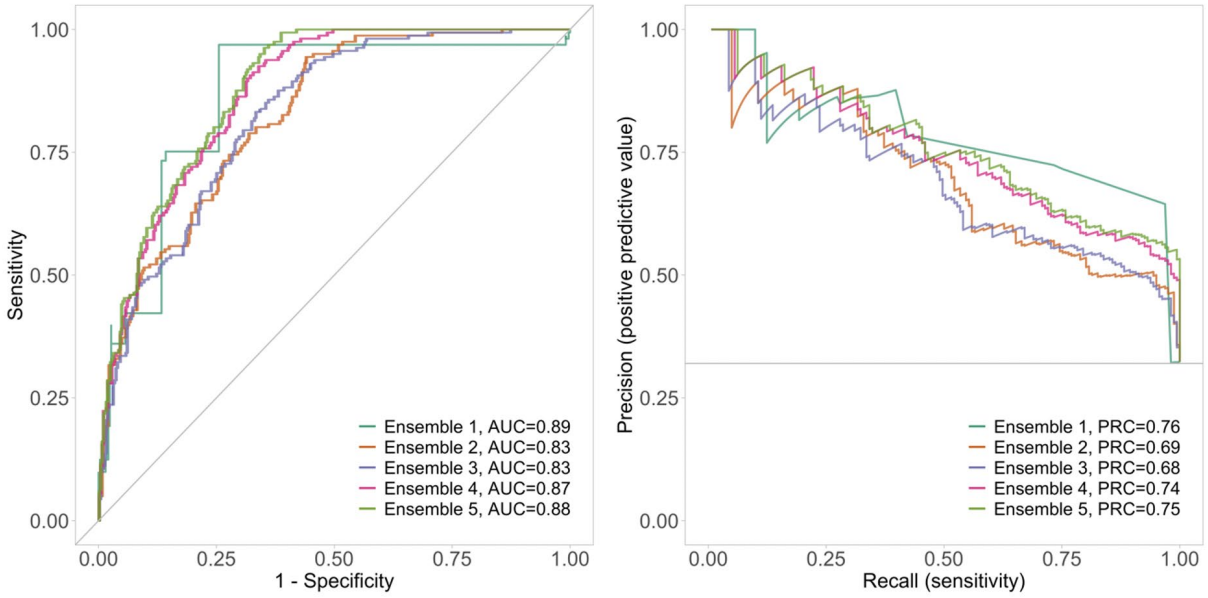

Figure 4.E. Receiver-operating characteristic (ROC) 1 and precision-recall curves for 5 ensemble algorithms, Magee Obstetric Maternal and Infant Database, 2013-2017, N = 498.

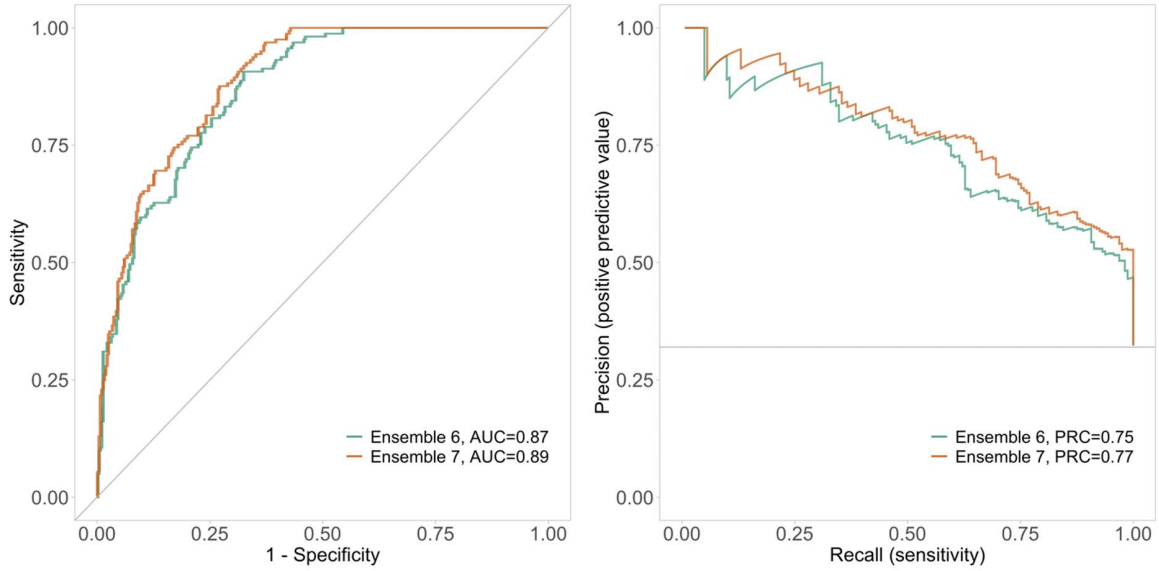

Figure 4.F. Receiver-operating characteristic (ROC) 1 and precision-recall curves for Ensembles 6 and 7, with variable selection, Magee Obstetric Maternal and Infant Database, 2013-2017, N = 498.

Table 21. Centers for Disease Control ICD-10 CM codes for identifying severe maternal morbidity (SMM) diagnoses and procedures.

Indicator	Diagnosis or procedure code
Acute myocardial infarction	I21.xx, I22.x
Aneurysm	I71.xx*, I79.0*
Acute renal failure	N17.x, O90.4
Adult respiratory distress syndrome	J80, J95.1, J95.2, J95.3, J95.82x, J96.0x, J96.2x R09.2
Amniotic fluid embolism	O88.1x
Cardiac arrest/ventricular fibrillation	I46.x, I49.0x
Conversion of cardiac rhythm	5A2204Z, 5A12012
Disseminated intravascular coagulation	D65, D68.8, D68.9, O72.3
Eclampsia	O15.X, O14.22
Heart failure/arrest during procedure/surgery	I97.12x, I97.13x, I97.710, I97.711
Puerperal cerebrovascular disorders	I60.xx- I68.xx, O22.51, O22.52, O22.53, I97.81x, I97.82x, O87.3 I62.9 – included but should not be captured if this is not a valid code.

Pulmonary edema/acute heart failure	J81.0, I50.1, I50.20, I50.21, I50.23, I50.30, I50.31, I50.33, I50.40, I50.41, I50.43, I50.9
Severe anesthesia complications	O74.0 , O74.1 , O74.2, O74.3, O89.0x, O89.1, O89.2
Sepsis	O85, O86.04, T80.211A, T81.4XXA, T81.44xx; or R65.20; or A40.x, A41.x, A32.7
Shock	O75.1, R57.x, R65.21, T78.2XXA, T88.2 XXA, T88.6 XXA, T81.10XA , T81.11XA, T81.19XA
Sickle cell disease with crisis	D57.0x, D57.21x, D57.41x, D57.81x
Air and thrombotic embolism	I26.x, O88.0x, O88.2x, O88.3x, O88.8x
Blood products transfusion	30233H1, 30233L1, 30233K1, 30233M1, 30233N1, 30233P1, 30233R1, 30233T1,30233H0, 30233L0, 30233K0, 30233M0, 30233N0, 30233P0, 30233R0, 30233T0,30230H1, 30230L1, 30230K1, 30230M1, 30230N1, 30230P1, 30230R1, 30230T1,30230H0, 30230L0, 30230K0, 30230M0, 30230N0, 30230P0, 30230R0, 30230T0, 30240H1, 30240L1, 30240K1, 30240M1, 30240N1, 30240P1, 30240R1, 30240T1,30240H0, 30240L0, 30240K0, 30240M0, 30240N0, 30240P0, 30240R0, 30240T0, 30243H1, 30243L1, 30243K1, 30243M1, 30243N1, 30243P1, 30243R1, 30243T1, 30243H0, 30243L0, 30243K0, 30243M0, 30243N0, 30243P0, 30243R0, 30243T0, 30250H1, 30250L1, 30250K1, 30250M1, 30250N1, 30250P1, 30250R1, 30250T1, 30250H0, 30250L0, 30250K0, 30250M0, 30250N0, 30250P0, 30250R0, 30250T0, 30253H1, 30253L1, 30253K1, 30253M1, 30253N1, 30253P1, 30253R1, 30253T1, 30253H0, 30253L0, 30253K0, 30253M0, 30253N0, 30253P0, 30253R0, 30253T0, 30260H1, 30260L1, 30260K1, 30260M1, 30260N1, 30260P1, 30260R1, 30260T1, 30260H0, 30260L0, 30260K0, 30260M0, 30260N0, 30260P0, 30260R0, 30260T0, 30263H1, 30263L1, 30263K1, 30263M1, 30263N1, 30263P1, 30263R1, 30263T1, 30263H0, 30263L0, 30263K0, 30263M0, 30263N0, 30263P0, 30263R0, 30263T0’
Hysterectomy	0UT90ZZ, 0UT94ZZ, 0UT97ZZ, 0UT98ZZ, 0UT9FZZ
Temporary tracheostomy	0B110Z, 0B110F, 0B113, 0B114
Ventilation	5A1935Z, 5A1945Z, 5A1955Z

Table 22. American College of Obstetricians and Gynecologists/Society for Maternal-Fetal Medicine guidelines for determining true-positive severe maternal morbidity status.

Severe maternal morbidity	Not severe maternal morbidity (insufficient evidence if this is the only criterion)
Hemorrhage	
Obstetric hemorrhage with ≥ 4 units of red blood cells transfused	Obstetric hemorrhage with 2-3 units of red blood cells transfused ALONE

Obstetric hemorrhage with 2 units of red blood cells and 2 units of fresh frozen plasma transfused (without other procedures or complications) if not judged to be overexuberant transfusion	Obstetric hemorrhage with 2 units of red blood cells and 2 units of fresh frozen plasma transfused AND judged to be "overexuberant"
Obstetric hemorrhage with < 4 units of blood products transfused and evidence of pulmonary congestion that requires >1 dose of furosemide	Obstetric hemorrhage with <4 units of blood products transfused and evidence of pulmonary edema requiring only 1 dose of furosemide
Obstetric hemorrhage with return to operating room for any major procedure (excludes dilation Any emergency/unplanned peripartum hysterectomy, regardless of number of unites transfused (includes all placenta accretas)	Planned peripartum hysterectomy for cancer/neoplasia
Obstetric hemorrhage with uterine artery embolization, regardless of number of units transfused	
Obstetric hemorrhage with uterine balloon or uterine compression suture placed and 2-3 units of blood products transfused	Obstetric hemorrhage with uterine balloon or uterine compression suture placed and ≤1 unit of blood products transfused
Obstetric hemorrhage admitted to intensive care unit for invasive monitoring or treatment (either medication or procedure; not just observed overnight)	Any obstetric hemorrhage that went to the intensive care unit for observation only without further treatment
Hypertension/neurologic	
Eclamptic seizure(s) or epileptic seizures that were "status"	
Continuous infusion (intravenous drip) of an antihypertensive medication	
Nonresponsiveness or loss of vision, permanent or temporary (but not momentary), documented in physician's progress notes	
Stroke, coma, intracranial hemorrhage	
Preeclampsia with difficult-to-control severe hypertension (> 160 systolic blood pressure or >110 diastolic blood pressure) that requires multiple intravenous doses, persistent ≥48 hours after delivery, or both	Chronic hypertension that drifts up to severe range and needs postoperative medication dose alteration: preeclampsia blood pressure control with an oral medication ≥48 hours after delivery
Liver or subcapsular hematoma or severe liver injury admitted to the intensive care unit (bilirubin >6 or liver enzymes >600)	Abnormal liver function requiring extra prolonged postpartum length of stay but not in an intensive care unit
Multiple coagulation abnormalities or severe hemolysis, elevated liver enzymes, and low platelet count (HELLP) syndrome	Severe thrombocytopenia (<50,000) alone that does not require a transfusion or intensive care unit admission
Renal	
Diagnosis of acute tubular necrosis or treatment with renal dialysis	Oliguria treated with intravenous fluids (no intensive care unit admission)

Oliguria treated with multiple doses of Lasix	Oliguria treated with 1 dose of intravenous fluids (no intensive care unit admission)
Creatinine ≥2.0 in a woman without preexisting renal disease OR a doubling of the baseline creatinine in a woman with preexisting renal disease	
Sepsis	
Infection with hypotension with multiple liters of intravenous fluid or pressors used (septic shock)	Fever >38.5°C with elevated lactate alone without hypotension
Infection with pulmonary complications such as pulmonary edema or acute respiratory distress syndrome	Fever >38.5°C with presumed choriometritis/endometritis with elevated pulse but no other cardiovascular signs and normal lactate
	Positive blood culture without other evidence of significant systemic illness
Pulmonary	
Diagnosis of acute respiratory distress syndrome, pulmonary edema, or postoperative pneumonia	Administration of oxygen without a pulmonary diagnosis
Use of a ventilator (with either intubation or noninvasive technique)	
Deep vein thrombosis or pulmonary embolism	
Cardiac	
Preexisting cardiac disease (congenital or acquired) with intensive care unit admission for treatment	Preexisting cardiac disease (congenital or acquired) with intensive care unit admission for observation only
Peripartum cardiomyopathy	Preexisting cardiac disease (congenital or acquired) without intensive care unit admission for observation only
Arrhythmia requiring >1 dose of intravenous medication but not intensive care unit admission	Arrhythmia requiring 1 dose of intravenous medication but no intensive care unit admission
Arrhythmia that requires intensive care unit with further treatments	Arrhythmia that requires intensive care unit observation but no extra treatments
Intensive Care Unit/Invasive Monitoring	
Any intensive care unit admission that includes treatment or diagnostic or therapeutic procedure	Intensive care unit admission for observation of hypertension that does NOT require intravenous medications
Central line or pulmonary catheter used to monitor a complication	Intensive care unit admission for observation after general anesthesia
Surgical, Bladder, and Bowel Complications	
Bowel or bladder injury during surgery beyond minor serosal tear	
Small-bowel obstruction, with or without surgery during pregnancy/postpartum period	
Prolonged ileus for ≥4 days	Postoperative ileus that resolved without surgery in ≤3 days
Anesthesia Complications	

Total spinal anesthesia	Failed spinal anesthesia that requires general anesthesia
Aspiration pneumonia	Spinal headache treated with a blood patch
Epidural hematoma	

Table 23. List of candidate predictors for ensemble algorithms, Magee Obstetric Maternal and Infant Database, 2010-2011.

Maternal demographic and behavioral variables	Maternal health variables	Labor and delivery variables	Fetal and infant variables
Age	Pregnancy-related hypertensive disorders	Type of labor onset	Fetal arrhythmia
Education	Chronic hypertension	Delivery type	Infant sex
Race/ethnicity	Depression	Repeat cesarean	NICU length of stay
Marital status	Adult respiratory distress syndrome	Vaginal birth after cesarean	Birthweight
Insurance status	Asthma	Premature rupture of membranes	Growth restriction
Smoking during pregnancy	Pneumonia	Vacuum-assisted delivery	Fetal death
Height	Diabetes	Chorioamnionitis	Respiratory distress
Weight	Anemia	Uterine rupture	Placental pathology
	Kidney disease	Preterm delivery	Infant status at discharge
	Postpartum complications	Prolonged membrane rupture	1-minute APGAR score
	Maternal status at discharge	Previous cesarean delivery	5-minute APGAR score
	Outcome of prior pregnancy	General anesthesia administered	Fetal presentation
	Maternal arrhythmia	Vaginal laceration	Gestational age at delivery
	History of infertility	Labor induction	
	Number of abortions	Type of labor induction	
	Parity	Placenta previa	
	Gravidity		

Table 24. Base learners and tuning parameters included in each SuperLearner ensemble algorithm.

Learner	Tuning parameters
Bayesian generalized linear models	
Random forests	• Number of trees: 500, 2500 • Variables for splitting at each node: 2, 3, 4 • Sampling with replacement: TRUE, FALSE
Mean	
Generalized linear models	
Generalized additive models	
Penalized regression	• Alpha: 0.0, 0.2, 0.4, 0.6, 0.8, 1.0
k-nearest-neighbors	• *k*: 2, 3, 4, 5

Table 25. Characteristics of liveborn singleton deliveries at Magee-Womens Hospital, 2010-2011 and 2013-2017.

	Eligible training cohort, 2010-2011 (N =19,266)	Eligible test cohort, 2013-2017 (N = 47,067)
Maternal race, %		
NH White	75	70
NH Black	20	21
Other, declined, or unspecified	5.4	8.9
Maternal age, mean (sd)	29 (5.9)	29 (5.5)
Maternal age ≥ 35, %	16	17
Married, %	56	55
Maternal education, %		
Less than high school	8.3	6.8
High school graduate	22	22
Some college	15	13
4-year college graduate	56	59
Type of insurance, %		
Private	63	62
Public nor none	37	38
Nulliparous, %	48	44
Smoked during pregnancy, %	14	12
Preexisting diabetes or hypertension, %	3.7	5.3
Body mass index category, %		
Underweight	4.8	4.0

Normal weight	53	48
Overweight	22	24
Obese	19.4	24.1
Gestational age, weeks, mean (sd)	39 (2.4)	39 (2.4)
Preterm birth < 37 weeks, %	10	10
Mode of delivery, %		
Vaginal	72	71
Cesarean	28	29
Birthweight, grams, mean (sd)	3277 (615)	3253 (622)

Table 26. Validation of screening criteria in the training subcohort, Magee Obstetric Maternal and Infant Database, 2010-2011, N = 685).

	Gold-standard definition		
Screening result	SMM	No SMM	Total
SMM	173	163	336
No SMM	1	348	349
Total	174	511	685

Table 27. Validation of screening criteria in the test subcohort, Magee Obstetric Maternal and Infant Database, 2010-2011, N = 498.

	Gold-standard definition		Total
Screening result	SMM	No SMM	
SMM	160	90	250
No SMM	1	247	248
Total	161	337	498

Table 28. Positive and negative predictive values of the screening criteria in the training and test subcohorts, Magee Obstetric Maternal and Infant Database.

	Training set	Test set
Positive predictive value	0.51	0.64
Negative predictive value	0.99	0.99

Table 29. Distribution of variables most highly-ranked in terms of variable importance by true-positive and true-negative SMM status in the training and test subcohorts, Magee Obstetric Maternal and Infant Database, 2010-2011 (N = 685) and 2013-2017 (N = 498).

	Training subcohort, 2010-2011 (n = 685)		Test subcohort, 2013-2017 (n = 498)	
	True positives (n = 171)	True negatives (n = 506)	True positives (n = 160)	True negatives (n = 337)
Birthweight, grams, mean (sd)	2652 (918)	3226 (674)	2526 (1013)	3185 (641)
ICU admission, %	50	5.3	63	8.0
Prolonged length of stay by delivery type, %	26	2.2	16	3.6
SMM: hysterectomy, %	23	0.79	6.3	0.0
Severe preeclampsia/Eclampsia, %	26	3.5	24	2.1
SMM: transfusion, %	47	20	44	12
Multimorbidity (2+), %	32	1.2	18	0.59
SMM: sepsis, %	6.6	0.0	5.6	0.30
SMM: DIC, %	9.0	0.0	2.5	1.5
General anesthesia during delivery, %	38	4.0	8.8	0.64
SMM: Eclampsia, %	7.2	0.59	1.25	0.0
Repeat c-section, %	26	13	23	15
Previous c-section, %	27	15	23	14
NICU length of stay, days, mean (sd)	11 (16)	2.6 (11)	15 (25)	3.2 (11)

Table 30. Measures of predictive accuracy for Ensembles 6 and 7, with variable selection, under two different classification thresholds, Magee Obstetric Maternal and Infant Database, 2013-2017, N = 498.

	Classification threshold = 0.5					Classification threshold = 0.3				
	Accuracy	Balanced accuracy[1]	Positive predictive value	Negative predictive value	Detection rate[2]	Accuracy	Balanced accuracy[1]	Positive predictive value	Negative predictive value	Detection rate[2]
Ensemble 6	0.79	0.69	0.89	0.77	0.13	0.83	0.79	0.77	0.85	0.22
Ensemble 7	0.76	0.65	0.82	0.75	0.11	0.82	0.77	0.78	0.84	0.20

Table 31. Measures of predictive accuracy for the screening criteria and each ensemble, with no variable selection, under two different classification thresholds, Magee Obstetric Maternal and Infant Database, 2013-2017, N = 498.

	Classification threshold = 0.5					Classification threshold = 0.3				
	Accuracy	Balanced accuracy[1]	Positive predictive value	Negative predictive value	Detection rate[2]	Accuracy	Balanced accuracy[1]	Positive predictive value	Negative predictive value	Detection rate[2]
Ensemble 1	0.76	0.65	0.79	0.75	0.11	0.76	0.71	0.68	0.80	0.17
Ensemble 2	0.76	0.64	0.80	0.75	0.10	0.76	0.69	0.65	0.79	0.17
Ensemble 3	0.76	0.65	0.82	0.75	0.11	0.79	0.73	0.73	0.81	0.18
Ensemble 4	0.76	0.65	0.80	0.75	0.11	0.79	0.73	0.73	0.81	0.18
Ensemble 5	0.78	0.69	0.78	0.78	0.14	0.82	0.80	0.72	0.87	0.24
Ensemble 6	0.77	0.66	0.81	0.76	0.12	0.80	0.75	0.75	0.82	0.19
Ensemble 7	0.77	0.67	0.84	0.76	0.12	0.79	0.73	0.73	0.81	0.18
Screening criteria	0.82	0.86	0.64	0.99	0.32	0.82	0.86	0.64	0.99	0.32

5.0 Chapter 5: Conclusion

The broad of objective of this work was to improve both understanding and measurement of severe maternal morbidity (SMM). As maternal health in the US continues to be an issue of urgent concern, these findings can be used to inform both research strategies and public policies to better quantify, treat, manage, and ultimately prevent SMM. This chapter will summarize key findings from this work, discuss strengths and limitations of each project to better contextualize the major findings, consider public health implications of the work, and discuss possible future directions for further research.

5.1 Summary of findings

The first key finding from this work was that SMM is associated with increased risk of severe adverse cardiovascular events up to 2 years postpartum. By following women enrolled in Pennsylvania Medicaid through their pregnancies, deliveries, and postpartum enrollment in the program, we were able to calculate the average treatment effect of SMM during the peripartum period on risk of ischemic heart disease, stroke/transient ischemic attack, heart failure, and a composite of these three conditions plus atrial fibrillation up to 2 years after delivery. These average treatment effect estimates are large, indicating that SMM is associated with greatly increased risk. For example, the risk differences per 1,000 live births for SMM vs. no SMM for the composite outcome were 2.7 (1.6, 3.8) at 1 month postpartum, 9.3 (6.5, 12) at 6 months postpartum, and 17 (12, 21) at 12 months postpartum. Modeling risk at each month of follow-up,

we demonstrated that this high risk persists after the traditional end of the postpartum period at 42 days post-delivery. Taken together, our results strongly suggest that women with SMM constitute a high-risk group for cardiovascular complications in the postpartum period. These women may require more individualized and extensively coordinated postpartum care. Our findings also illustrate the broader burden of poor maternal health surrounding SMM that extends beyond the postpartum period. Although most SMM research treats SMM as an outcome, the vast majority of people who experience an SMM survive. Our work is some of the only work conducted in a US population to investigate postpartum consequences of SMM. Most SMM research is also conducted in hospital discharge databases or other administrative databases where all data are collected during the hospitalization for delivery. Our results, however, shed light on aspects of the population burden of SMM in the US that can't be investigated using discharge databases. Our findings, demonstrating that women with SMM remain at elevated risk of these severe, costly, and potentially life-threatening events long after delivery, support recommendations that target comprehensive, high-quality postpartum care for every birthing person in the United States.

Increasing use of machine learning in the biomedical sciences, including epidemiology, means that epidemiologists have to consider issues specific to machine learning, such as outcome class imbalance. Outcome class imbalance refers to situations, common in epidemiologic data, where non-events outnumber events by some factor: 2:1, 5:1, 10:1, or more. The second key finding from this work was that moderate imbalance in outcome classes (where, for example, non-events outnumber events 5:1) does not greatly affect SuperLearner ensemble performance in terms of area under the receiver operating characteristic (ROC) curve, and further that downsampling an imbalanced data set to achieve balance between outcome classes does not deliver meaningful improvements in predictive performance. Performance of an unbalanced vs. balanced (via

downsampling) SuperLearner algorithm was similar in terms of overall accuracy (0.54 and 0.57, respectively) and area under the ROC curve (0.63 and 0.62, respectively); sensitivity was reduced by downsampling while positive predictive value was improved. We also explored the effects of selecting different thresholds of predicted probability of the outcome above which a prediction is classified as a positive outcome case; in the absence of any data-driven method to choose optimal thresholds, exploring thresholds based on the distribution of predicted probabilities can help investigators choose the one that best suits their objective. Most epidemiology papers report only the area under the ROC curve, sensitivity and specificity; however, these measures are not sensitive to the prevalence of the outcome and can be misleading when outcome classes are imbalanced. Thus, guidance on how to deal with class imbalance is needed, and our findings have implications for epidemiologists who are interested in incorporating machine learning techniques into their work.

Finally, the third major finding of this work was that ensemble machine learning (SuperLearner) can improve some aspects of true (adjudicated by medical record review) SMM case ascertainment, but that existing screening criteria for SMM also perform well by comparison. For example, the screening criteria had a negative predictive value of 0.99, a positive predictive value of 0.64, an overall accuracy of 0.82, and a detection rate of 0.32. By contrast, one of the best-performing SuperLearner ensemble algorithms had a negative predictive value of 0.77, a positive predictive value of 0.86, an overall accuracy of 0.79, and a detection rate of 0.13. This ensemble algorithm included all components of the screening criteria in the predictor set and additionally incorporated the screening criteria as a base learner. The practical utility of these findings will depend on context and objective of researchers who might be interested in using machine learning tools to identify SMM cases in their data. Our results illustrate an important

tradeoff between the extremely high negative predictive value achieved by the screening criteria and the much-improved positive predictive value achieved with ensemble machine learning. However, since higher positive predictive value with machine learning is achieved by classifying few cases as true positives overall, an ensemble machine learning approach might not be optimal for population-level surveillance of severe maternal morbidity. Such an approach is still less labor-intensive than comprehensive medical record review, and could be a quick and easy way to identify some true-positive cases in a given data setting without relying on screening criteria, or if screening criteria are not available. Importantly, we found that both the screening criteria and ensemble machine learning algorithms have difficulty identifying true-positive SMM. This could be because the definition of SMM is imprecise, or because we did not have predictors in our data with strong enough associations with the outcome to reliably distinguish true-positives from true-negatives.[155]

5.2 Strengths and limitations

The findings outlined above each have important strengths and limitations. These should be used to contextualize and critically evaluate the findings from this work and their public health relevance. Our finding of increased cardiovascular risk postpartum following SMM should be evaluated in light of 6 major limitations. First, we relied on a screening definition of SMM (the CDC list of diagnosis and procedure codes or any ICU admission) to define the exposure, although evidence is emerging that the screening definition identifies many false-positive cases (ref). This high false-positive rate is most commonly attributed to the inclusion of blood transfusion procedure codes in the screening definition of SMM as a proxy for obstetric hemorrhage. We addressed this limitation with sensitivity analyses excluding blood transfusion; though our results

did not meaningfully change, the point estimates were slightly larger. This is to be expected, since some individuals with transfusion as their only SMM indicator are likely not SMM cases; excluding them from the "exposed" group thus increases the magnitude of the effect estimate. Nonspecificity of the blood transfusion procedure codes is widely recognized, and effect estimates for SMM and non-transfusion SMM are commonly reported together, as we did.

Second, the screening criteria were based on ICD codes and intended to be used in administrative data sets, but administrative data were not designed for research purposes. Therefore, they may have a number of data quality issues such as miscoding and under-ascertainment of important variables. We were unable to use other definitions of SMM due to the constraints of the Medicaid claims data that we used for this analysis; however, conducting this work using Medicaid claims data allowed us to count SMM over the entire peripartum period in our exposure definition. Similarly, we did not have information on length of gestation in this data. We addressed this limitation by assuming 40 weeks of gestation for all term deliveries. We used sensitivity analyses to assess the impact of varying the length of gestation assumed for term deliveries to 38 weeks and 42 weeks; doing so only slightly affected our effect estimates. Use of Medicaid claims data also limited the generalizability of our results. Medicaid is only one payer, and Medicaid populations tend to have lower socioeconomic status than the total obstetric population in the US. While our results may not be generalizable to the general obstetric population in the US, 40% of all U.S. deliveries covered by Medicaid.

Third, in addition to the exposure definition, the outcome and confounder definitions used in the analysis were also based on ICD codes. Consequently, our study is potentially subject to information biases, the effects of which are difficult to predict. In this case, under-ascertainment rather than coding errors are likely responsible; ICD coding has differential accuracy depending

on the condition. For example, obesity is almost certainly under-ascertained in our analyses; the prevalence in our sample was approximately 20%, while the prevalence of obesity in reproductive-age women in the United States is 45%.[162] ICD-10 codes for obesity have high positive predictive value, but low sensitivity – the codes are accurate when obesity is recorded, but it is frequently not recorded.[137] Obesity is just one example; many intrapartum diagnoses and procedures are under-ascertained in delivery discharge records compared to the medical record.[163] It is thus possible that there is pervasive under-ascertainment of exposure, outcome, and covariates in our analysis and that such under-ascertainment is differential by exposure or outcome status. It is difficult to estimate how these potential biases might affect our results; this is another important limitation of using administrative data for population health research.

Fourth, the heterogeneity of SMM complicates interpretation of our findings. Though it is unlikely that all SMM conditions or procedures equally predispose individuals to adverse cardiovascular events postpartum, even in our large Medicaid cohort we did not have sufficient sample size to examine individual SMM conditions in relation to the outcomes. This heterogeneity also limits our ability to make causal inferences about the relationship between SMM and adverse cardiovascular events postpartum; the consistency assumption for causal identification is *de facto* violated because SMM is a group of many different conditions and procedures.[138,139] However, the associational measures we calculated are informative although our results cannot be interpreted causally.

Fifth, there is potential for unmeasured confounding to bias our results. In particular, confounding by history of adverse cardiovascular events or prior SMM is a threat to the validity of our findings. It is possible that some people in our study had adverse cardiovascular events or SMM events prior to either their enrollment in Medicaid or their inclusion in the study cohort. We

attempted to address this by restricting the analysis only to those individuals with incident outcomes after delivery (*i.e.* with no history of any of the outcomes prior to the start of follow-up). However, it is still possible that there are some prior outcome events that we could not detect because they occurred prior to the start of the study period or prior to an individual's enrollment in Medicaid. Bias resulting from residual confounding is also a possibility; we attempted to minimize the impact of residual confounding by choosing covariates for adjustment carefully using theory-based causal diagrams.

Sixth and finally, loss to follow-up is not random our Medicaid cohort or in Medicaid claims data generally. People may be covered by Medicaid under a variety of eligibility categories, including through the expansion of Medicaid under the Patient Protection and Affordable Care Act (PPACA), due to a disability, or due to pregnancy. These patterns of eligibility contribute to known differential patterns in loss to follow-up in Medicaid populations.[111] Non-random loss to follow-up can introduce substantial selection bias. In our case, our results would be underestimates if the factors that predispose women to SMM also predispose them to disenrollment from Medicaid, or overestimates if the women most likely to have SMM are also differentially likely to remain enrolled. We are confident that we addressed this issue by constructing and applying inverse probability of censoring weights.[110]

Findings from the simulation analysis to explore class imbalance are not affected by unmeasured confounding, selection bias, or information biases, because the data for the analysis were simulated from a logistic model according to our specifications. However, these findings are not without important limitations. First, we optimized our ensemble algorithms for area under the receiver operating characteristic (ROC) curve (AUC); we could have evaluated other performance metrics since AUC can give a misleadingly optimistic picture of classifier performance, especially

if the outcome classes are imbalanced.[164] Second, we only evaluated downsampling as a means to balance the outcome classes, though there are other ways of creating balance between outcome classes such as oversampling and synthetic minority class oversampling (SMOTE). Downsampling is somewhat controversial because it discards data that could otherwise be used for prediction. However, downsampling is intuitive and easy to implement and consequently is potentially attractive to epidemiologists without extensive machine learning or predictive modeling expertise. Third, we simulated moderate class imbalance (5:1) and so these findings may not apply to settings with extreme class imbalance such as SMM or cancer screening, where negative screens might outnumber positive screens by a factor of 10 or more.[147,165,166] However, the degree of class imbalance we simulated does approximate the outcome class distribution of many conditions in many settings, such as diabetes in a population-based cohort.[167] This work thus provides some practical guidance for epidemiologists.

The strengths and limitations of our project to build an ensemble algorithm to predict SMM reflect its goals: accurate prediction rather than etiologic inference. First, because of the sampling schemes we used to construct them, our training and test data have artificially high prevalence of SMM that does not reflect the true rarity of SMM in population-based cohorts. Thus, outcome class imbalance might substantially affect the predictive performance of our algorithms should they be applied to "real," un-sampled data (*e.g.*, to the MOMI test cohorts of all deliveries in 2013-2017). Although our simulation analyses indicated that class imbalance would not pose serious problems for SuperLearner, it is possible that in this real world, unsimulated scenario, class imbalance might have affected SuperLearner performance. Also because of the sampling scheme, we cannot report sensitivity and specificity; though we optimized our ensemble algorithms for area under the ROC curve, derivatives of sensitivity and specificity do not have an interpretation when

applied to data that were sampled based on a screening definition.[156] We did evaluate many different measures of predictive performance that take prevalence of the outcome into account.

Second, though confounding and causal structure are not relevant to our goal of optimizing prediction, there were some nontrivial differences between the training and test data sets. Similar to the class-imbalance scenario outlined above, this might cause issues with generalizability of our algorithms from the training data to the test data – performance on the test data might not be optimal. So-called "contextual biases" in the training set could further limit the generalizability of these ensemble algorithms. For example, these ensemble algorithms were trained using medical record review data from a high-resource academic obstetric hospital. It is therefore not guaranteed that these algorithms will exhibit acceptable performance in data gathered in a lower-resource setting where, for example, medical record abstractors or certain kinds of diagnostic tests are not available. Furthermore, if the patient population at Magee-Womens Hospital in 2010-2011 differed from the patient population at Magee-Womens Hospital in 2013-2017, or differ in how and whether their data is entered into MOMI, performance of the ensemble algorithms in the test set could be poor. It is well known that this type of contextual bias is difficult, if not impossible, to detect and that methods to improve algorithm generalizability and external validation performance (like cross-validation) cannot correct for it. Similarly, ensemble algorithms in particular can be difficult to interpret; how and why predictions are made is not clear ("black box"). This most severely limits the utility of risk prediction algorithms in clinical settings; however, it has implications for our work as well. We evaluated variable importance measures to assess which predictors contributed most to predictive performance of our algorithms, but variable importance measures do not "unbox" the algorithms, nor do they convey any contextual information about the relationship between a given predictor and the outcome.

Third, there is generally not firm guidance on sample size requirements for ensemble machine learning. We chose our sample size based only on the availability of medical record review and guidance from prior simulations. The current consensus in epidemiology is that the optimal sample size for a machine learning classification task is context- and data-dependent.[161] Optimal sample size also depends on how "noisy" the training and test data are, *i.e.*, if they contain a large number of redundant predictors or predictors that are not associated with the outcome. It is possible that we did not adequately reduce the number of predictors available to our ensemble algorithms. The similarity between our results with variable selection algorithms versus without suggests that this was not a major limitation in our study, but it is possible that test set performance could be improved by exploring other so-called feature reduction strategies.

5.3 Public health implications

Severe maternal morbidity is many times more common than maternal death, with over 60,000 women affected in the United States every year; the vast majority of women affected by SMM survive. Our work contends that the true population burden of SMM includes not only SMM events themselves, but also any sequelae following from SMM. Although SMM is heterogeneous and although we did not have sufficient statistical power to evaluate associations between individual SMM conditions and risk of adverse cardiovascular events postpartum, hypertensive and cardiovascular SMM (*e.g.*, eclampsia, puerperal cerebrovascular disorders) were among the most common in our sample. Hypertensive and cardiovascular SMM are also some of the most common nationwide[62] and contribute disproportionately to more severe outcomes.[168] It is thus plausible that in our sample, hypertensive and cardiovascular SMM may have been more strongly

associated with increased risk of adverse cardiovascular events postpartum than other SMM events like sepsis or blood transfusion. Consequently, one implication of our findings for both clinical and public health practice is that better identification and treatment of women with hypertension or cardiovascular disease risk (both prior to and during pregnancy) could contribute to primary prevention of SMM and improvement in SMM outcomes.

Our findings also suggest that people who survive SMM represent a high-risk group postpartum, at least in terms of risk of adverse cardiovascular events. Recognition of people who survived SMM as a high-risk group could be used to better tailor the transition from obstetric and postpartum care to longer-term well-woman care. This transition, when many women lose health insurance or change providers, is a known point of vulnerability for postpartum women and in our health care delivery system. Our results suggest that information from a woman's medical history, including her delivery history, can be used to design individualized care coordination postpartum. The potential to implement such changes in care delivery is complicated by medical specialization, health finance, geographical access to care, and a number of other factors.

This work has implications for public health and social policy as well as for individual-level clinical care. Our results support expanding Medicaid eligibility nationwide and extending pregnancy-related Medicaid coverage for a longer period postpartum. Currently, Medicaid (which covers nearly half of all deliveries in the United States) extends postpartum coverage for at least 60 days after delivery, with some states that expanded Medicaid under PPACA extending postpartum coverage for longer. Better and more continuous postpartum care would benefit all pregnant people, regardless of payer. Beyond Medicaid specifically, the policy implications of this work complement and support the American College of Obstetricians and Gynecologists's (ACOG) recommendations for optimizing postpartum care.[169] Specifically, our findings support

recommendations to change reimbursement policies for all payers (not just Medicaid) to support more comprehensive and continuous postpartum care, and to implement paid family leave policies at all levels of government to support pregnant and especially postpartum people. Finally, though the ACOG recommendations stop short of this, our findings support implementation of a federal single-payer health care system to replace the fragmented, inefficient, and frequently inaccessible patchwork of private and public insurance coverage available to pregnant and postpartum people today.

There is increasing interest in predictive modeling in the biomedical sciences. Flexible, nonparametric techniques for predictive modeling, including machine learning, are an increasingly popular option; these techniques are especially appealing for use in high-dimensional data, where the number of available covariates or predictors may be even greater than the sample size. Although the data sets available to epidemiologists are increasingly high-dimensional, predictive modeling principles are not widely taught as part of epidemiology curricula. As epidemiologists seek to integrate machine learning into predictive modeling studies or into doubly-robust estimation procedures for causal inference, outcome class imbalance is likely to be a common issue. This is important from a clinical perspective: if, for example, severe class imbalance is not remedied, a risk-prediction algorithm to identify women at increased risk of developing SMM could incorrectly predict every woman to be low-risk, which could lead to errors of clinical judgment or practice. There are population health implications to the use of machine learning in health care settings as well, such as when algorithms determine allocation of social or medical services or are used to "adjust" for race/ethnicity in common clinical calculations like the estimated glomerular filtration rate (eGFR). Although machine learning is a technique and not an epistemological framework, the kinds of applications for which machine learning is used (and the

types of inquiry that machine learning facilitates) also have important implications for public health practice. Individual-level risk prediction is possible with machine learning, but perhaps not always advisable or always in line with a coherent public health goal. Technical guidance, coupled with conceptual guidance, can help epidemiologists use machine learning tools responsibly and effectively.

Accurate identification of SMM is an ongoing challenge in public health, particularly for epidemiologic research and surveillance. The results of our analysis to use ensemble machine learning to identify true-positive SMM does not have implications at the level of clinical practice, but rather at the level of clinical and epidemiologic research. That is, our results should not be used to guide patient care or to identify pregnant women at high risk of SMM. Rather, we demonstrated that ensemble machine learning is a possible alternative to existing screening criteria. We did not demonstrate a clear benefit to using ensemble machine learning instead of existing screening criteria, however. For many applications, the screening criteria perform well, particularly in terms of negative predictive value (screen-negatives are likely to be true-negatives) and true-positive detection rate. The ensemble algorithms we built demonstrated higher positive predictive value than the screening criteria (greater likelihood that a case classified as positive by the algorithm is a true-positive), but a low detection rate. A researcher who is not interested in building an ensemble algorithm in their own data would be well-served by the screening criteria, as would researchers whose priority is identifying most cases of true-positive SMM even if the false-positive rate is high. The public health implications of this work depend somewhat on the goal of the investigator, but until more sophisticated algorithms can be trained our tool should not be used to quantify SMM for, *e.g.*, care quality improvement initiatives.

These results also touch on some policy implications. First, our results highlight the challenges of machine learning algorithm transportability. A nationwide obstetric surveillance system, similar to UKOSS in the UK, would improve many aspects of perinatal public health research and practice. Such a system would facilitate SMM identification from the hospital level to the national level and ensure accuracy and consistency of national estimates of SMM prevalence. A national surveillance system would also permit design of a single algorithm for true-positive SMM identification. Second, a clear and universal definition of SMM would also improve researchers' ability to identify SMM. The need for such a definition is the crux of an active debate in the literature, with some arguing for more globally comparable definitions and some arguing in favor of locally-generated definitions responsive to local capacity.[150] Until a resolution is reached, it is likely that quantifying SMM will remain imprecise; again, this is especially true in the US because SMM is measured and studied using several partially overlapping definitions in multiple data systems that capture SMM information, few if any of which are interoperable.

5.4 Future directions

Little research has been conducted in US populations to investigate the long-term consequences of SMM. This is partially because of the data sources and structures available to study SMM in the US. Consequently, the true burden of SMM is not well understood in its totality, since most studies assess SMM as an endpoint rather than an exposure and because most studies restrict exposure and outcome ascertainment to the delivery hospitalization. However, several future directions are feasible given currently existing data. Cardiovascular conditions account for or contribute to a large proportion of severe maternal morbidities and maternal deaths nationwide.

SMM, particularly cardiovascular SMM, and postpartum adverse cardiovascular events may fall on the same spectrum of poor cardiovascular health throughout the life course. Future research could address this by examining cardiovascular SMM specifically in relation to adverse cardiovascular events. National health registry data would facilitate this type of research, but in the United States such data is neither widely collected nor readily available. An ideal study could enroll women prior to conception, comprehensively assess history of cardiovascular morbidity at enrollment, and follow women prospectively through pregnancy and beyond to identify cardiovascular SMM and cardiovascular sequelae after delivery. Similarly, a cohort study in administrative data like Medicaid could use several strategies to comprehensively assess pre-pregnancy morbidity.

Another future direction for research in this area is identifying opportunities to shift the population burden of cardiovascular morbidity. The US has high levels of comorbidities among reproductive-age women compared to other wealthy countries. This comorbidity burden is not borne equally within the US, as striking racial/ethnic disparities in both pre-pregnancy comorbidities and SMM attest. It is plausible that a number of policy determinants shape the distribution of cardiovascular morbidity in the US; poor nutrition, precarious work, environmental pollution, lack of health insurance, and COVID-19 may all contribute to high levels of cardiovascular risk at young ages in the US. Observational studies, including so-called natural experiments, could be used to examine the impact of policies that might improve cardiovascular health at the population level – for example, comparing rates of cardiovascular morbidity, cardiovascular SMM, and/or adverse cardiovascular events postpartum before and after implementation of a policy to reduce air pollution in a residential area.

Adverse cardiovascular events are very serious rare events. Other potentially important sequelae of SMM might be more difficult to study. For example, psychiatric morbidity and substance use disorders following SMM are potentially important sources of population morbidity associated with SMM. Disaggregating SMM by type of morbidity might be particularly informative; future studies that aim to determine which SMM complications are most responsible for any associations between SMM and psychiatric or substance use disorders should ensure adequate sample size to make inferences about individual morbidities. Another challenge for work that seeks to determine the association between SMM and psychiatric disorders or symptoms postpartum is identification of incident psychiatric morbidity from observational or claims data. Bias may be introduced through confounding by history of the outcome, other unmeasured confounding, or differential ascertainment of psychiatric (*e.g.*, depression) symptoms during pregnancy and potentially among women with SMM as compared to women with uncomplicated deliveries. Similar considerations apply to any studies of SMM and substance use disorder; an additional measurement challenge involves accurate identification of substance use disorders or associated morbidities from administrative or claims data.

The postpartum impacts of SMM are not limited just to health conditions. SMM is expensive, sometimes requiring intensive medical intervention, and as such health care utilization outcomes are also important to understand. Claims databases can be used to study utilization outcomes like postpartum emergency department visits, hospital readmission, and uptake of the postpartum visit; a limitation of such research questions is reliance on the screening criteria to identify the exposed group (women with SMM).

Different analytic approaches could extend this work as well. In our work, we generated the average treatment effect comparing the hypothetical counterfactual scenarios “everyone in the

population had SMM" and "no one in the population had SMM." However, an estimand with a natural course comparison might be more informative to clinicians and policymakers. The natural course refers to a factual outcome summary generated from a hazard or pooled logistic regression model, essentially corresponding to "no intervention." Consequently, future work might use a natural course comparison to quantify the effect of preventing some or all cases of SMM, relative to the status quo. This would be informative to public health scientists and policymakers, and also encourages a paradigm shift to consider primary prevention of SMM in addition to identification of high-risk women in a clinical setting.

There are few general guidelines for machine learning in an imbalanced data set for epidemiology or other biomedical research, and a lack of consensus in the machine learning literature on whether class imbalance can be expected to categorically affect machine learning algorithm performance. Consequently, the future directions from our simulation work are almost limitless. In terms of analytic strategy, we could evaluate a wider range performance metrics, explore other mechanisms to remedy class imbalance, and construct different ensembles with different algorithms or tuning parameters. Future methodological research could also include developing and demonstrating a way to choose optimal classification thresholds for predictions from ensemble models. To our knowledge, there is no "data-driven" way to choose such thresholds.

Lastly, a critical future direction for any kind of algorithmic or predictive modeling work in epidemiology is developing understanding of the epistemological and real-world consequences of predictive analytic techniques and use of algorithms for statistical learning, classification, and decision-making in the biomedical sciences. Such theoretical and conceptual understanding must complement simple technical guidance.

Our work to develop an ensemble algorithm for SMM identification was rate-limited by the necessity of performing chart review to identify true-positive and true-negative cases. This is likely to be an important obstacle to investigators aiming to build tools to efficiently identify SMM for research or surveillance. The generalizability of our algorithms is also limited to other years of the same data source that we used to build and test them. One possible future direction for this work could be a nationwide collaborative effort to pool chart-reviewed cases from many sources, or to undertake a large chart review initiative to create a large, nationally representative data set of true SMM cases and non-cases and associated demographic and clinical information. Such a data source could be used to develop an algorithm for SMM identification using predictors likely to be available in most data settings, that could be used nationally to estimate the prevalence of SMM. Such work would not be a replacement for an obstetric health surveillance registry or a universal definition of SMM. However, in the absence of coordinated national infrastructure to facilitate SMM identification, such an effort could assist researchers and promote accurate national estimates of SMM incidence.

As previously mentioned, another potential extension of this work is to further externally validate our ensemble algorithms in other data. Because we trained our ensemble algorithms on MOMI data in 2010-2011, an easy next step would be to test our ensemble algorithms in other years of MOMI data to evaluate their performance. Our algorithms could also be evaluated in any other data with the same predictors that were used to train them. Alternatively, we could repeat the same analysis but including only predictors available across several different external validation sets. A limitation of this approach is the need for chart review (and the variables necessary to conduct chart review according to the ACOG/SMFM guidelines) to ascertain true SMM status in any external validation data set.

Although our simulations indicated that moderate class imbalance would have minimal impact on predictive performance, it is plausible that class imbalance might have affected the predictive performance of our ensemble algorithm to identify SMM. A natural next step would be to apply some type of class imbalance correction to our training and test data. Options include downsampling (randomly sampling a number of non-cases equal to the number of cases), oversampling (randomly selecting and duplicating a number of cases equal to the number of non-cases), and synthetic minority oversampling technique or SMOTE (a data augmentation method that simulates additional cases). These techniques have various drawbacks and benefits; downsampling is easy to implement but discards data, simple oversampling does not discard data but may lead to overfitting, and SMOTE may perform better than random downsampling or oversampling but is more technically challenging to implement.

Since practical guidance for machine learning in epidemiology is still limited, there are a number of potential future directions from this work that apply different analytic strategies and choices. First, performance might be improved by applying cost-sensitive learning techniques to the ensemble to differentially penalize false negatives or false positives. Our ensemble algorithm penalized both equally, but preferentially penalizing (for example) false positives might maximize predictive accuracy in terms of positive predictive value or detection rate. Second, investigating the performance of these algorithms with different feature selection techniques could be informative in terms of optimizing predictive performance as well as identifying appropriate sample size. Finally, exploring the impact of different transformations of predictor variables, explicitly modeling interactions between predictors, and incorporating different base learner algorithms into the SuperLearner ensemble could yield insights into best practices for adapting ensemble machine learning methods for obstetric research.

Bibliography

1. Callaghan WM, Creanga AA, Kuklina EV. Severe maternal morbidity among delivery and postpartum hospitalizations in the United States. *Obstet Gynecol.* 2012;120(5):1029-1036.
2. Callaghan WM, Grobman WA, Kilpatrick SJ, Main EK, D'Alton M. Facility-based identification of women with severe maternal morbidity: it is time to start. *Obstet Gynecol.* 2014;123(5):978-981.
3. Callaghan WM, Mackay AP, Berg CJ. Identification of severe maternal morbidity during delivery hospitalizations, United States, 1991-2003. *Am J Obstet Gynecol.* 2008;199(2):133.e131-138.
4. Fingar KF (IMB Watson Health) HMA, Heslin KC (AHRQ), Moore JE (Institute for Medical Innovation). Trends and Disparities in Delivery Hospitalizations Involving Severe Maternal Morbidity, 2006-2015. *HCUP Statistical Brief.* 2018;243.
5. Creanga AA, Bateman BT, Kuklina EV, Callaghan WM. Racial and ethnic disparities in severe maternal morbidity: a multistate analysis, 2008-2010. *Am J Obstet Gynecol.* 2014;210(5):435.e431-438.
6. Hoyert DL, Minino AM. Maternal Mortality in the United States: Changes in Coding, Publication, and Data Release, 2018. *Natl Vital Stat Rep.* 2020;69(2):1-18.
7. Collaborators GBDMM. Global, regional, and national levels of maternal mortality, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet.* 2016;388(10053):1775-1812.

8. Petersen EE, Davis NL, Goodman D, et al. Vital Signs: Pregnancy-Related Deaths, United States, 2011-2015, and Strategies for Prevention, 13 States, 2013-2017. *MMWR Morb Mortal Wkly Rep.* 2019;68(18):423-429.

9. MacDorman MF, Declercq E, Cabral H, Morton C. Recent Increases in the U.S. Maternal Mortality Rate: Disentangling Trends From Measurement Issues. *Obstet Gynecol.* 2016;128(3):447-455.

10. Berg CJ, Atrash HK, Koonin LM, Tucker M. Pregnancy-related mortality in the United States, 1987-1990. *Obstet Gynecol.* 1996;88(2):161-167.

11. Berg CJ, Chang J, Callaghan WM, Whitehead SJ. Pregnancy-related mortality in the United States, 1991-1997. *Obstet Gynecol.* 2003;101(2):289-296.

12. Berg CJ, Callaghan WM, Syverson C, Henderson Z. Pregnancy-related mortality in the United States, 1998 to 2005. *Obstet Gynecol.* 2010;116(6):1302-1309.

13. Creanga AA, Berg CJ, Syverson C, Seed K, Bruce FC, Callaghan WM. Pregnancy-related mortality in the United States, 2006-2010. *Obstet Gynecol.* 2015;125(1):5-12.

14. Creanga AA, Syverson C, Seed K, Callaghan WM. Pregnancy-Related Mortality in the United States, 2011-2013. *Obstet Gynecol.* 2017;130(2):366-373.

15. Joseph KS, Lisonkova S, Muraca GM, et al. Factors Underlying the Temporal Increase in Maternal Mortality in the United States. *Obstet Gynecol.* 2017;129(1):91-100.

16. Davis NL, Hoyert DL, Goodman DA, Hirai AH, Callaghan WM. Contribution of maternal age and pregnancy checkbox on maternal mortality ratios in the United States, 1978-2012. *Am J Obstet Gynecol.* 2017;217(3):352 e351-352 e357.

17. Geller SE, Rosenberg D, Cox SM, Kilpatrick S. Defining a conceptual framework for near-miss maternal morbidity. *J Am Med Womens Assoc (1972).* 2002;57(3):135-139.

18. Admon LK, Winkelman TNA, Zivin K, Terplan M, Mhyre JM, Dalton VK. Racial and Ethnic Disparities in the Incidence of Severe Maternal Morbidity in the United States, 2012-2015. *Obstet Gynecol.* 2018;132(5):1158-1166.

19. Hirshberg A, Srinivas SK. Epidemiology of maternal morbidity and mortality. *Seminars in perinatology.* 2017;41(6):332-337.

20. Holdt Somer SJ, Sinkey RG, Bryant AS. Epidemiology of racial/ethnic disparities in severe maternal morbidity and mortality. *Seminars in perinatology.* 2017;41(5):258-265.

21. Pallasmaa N, Ekblad U, Gissler M, Alanen A. The impact of maternal obesity, age, pre-eclampsia and insulin dependent diabetes on severe maternal morbidity by mode of delivery-a register-based cohort study. *Arch Gynecol Obstet.* 2015;291(2):311-318.

22. Leonard SA, Main EK, Carmichael SL. The contribution of maternal characteristics and cesarean delivery to an increasing trend of severe maternal morbidity. *BMC Pregnancy Childbirth.* 2019;19(1):16.

23. Lisonkova S, Muraca GM, Potts J, et al. Association Between Prepregnancy Body Mass Index and Severe Maternal Morbidity. *Jama.* 2017;318(18):1777-1786.

24. Schummers L, Hutcheon JA, Hacker MR, et al. Absolute Risks of Obstetric Outcomes Risks by Maternal Age at First Birth: A Population-based Cohort. *Epidemiology (Cambridge, Mass).* 2018;29(3):379-387.

25. Shamshirsaz AA, Dildy GA. Reducing Maternal Mortality and Severe Maternal Morbidity: The Role of Critical Care. *Clinical obstetrics and gynecology.* 2018;61(2):359-371.

26. Geller SE, Cox SM, Kilpatrick SJ. A descriptive model of preventability in maternal morbidity and mortality. *J Perinatol.* 2006;26(2):79-84.

27. Troiano NH, Witcher PM. Maternal Mortality and Morbidity in the United States: Classification, Causes, Preventability, and Critical Care Obstetric Implications. *The Journal of perinatal & neonatal nursing.* 2018;32(3):222-231.

28. Drife JO. Maternal "near miss" reports? *BMJ.* 1993;307(6912):1087-1088.

29. Bewley S, Creighton SB. 'Near-miss' obstetric enquiry. *J Obstet Gynaecol.* 1997;17(1):26-29.

30. Stones W, Lim W, Al-Azzawi F, Kelly M. An investigation of maternal morbidity with identification of life-threatening 'near miss' episodes. *Health Trends.* 1991;23(1):13-15.

31. Mantel GD, Buchmann E, Rees H, Pattinson RC. Severe acute maternal morbidity: a pilot study of a definition for a near-miss. *Br J Obstet Gynaecol.* 1998;105(9):985-990.

32. Department of Health; Welsh Office; Scottish Home and Health Department; Department of Health and Social Services NI. Report on Confidential Enquiries into Maternal Deaths in the United Kingdom 1991-1993. *London: HMSO, 1996.* 1996.

33. Say L, Pattinson RC, Gulmezoglu AM. WHO systematic review of maternal morbidity and mortality: the prevalence of severe acute maternal morbidity (near miss). *Reprod Health.* 2004;1(1):3.

34. Say L, Souza JP, Pattinson RC, Mortality WHOwgoM, Morbidity c. Maternal near miss--towards a standard tool for monitoring quality of maternal health care. *Best Pract Res Clin Obstet Gynaecol.* 2009;23(3):287-296.

35. Souza JP, Cecatti JG, Faundes A, et al. Maternal near miss and maternal death in the World Health Organization's 2005 global survey on maternal and perinatal health. *Bull World Health Organ.* 2010;88(2):113-119.

36. Cecatti JG, Souza JP, Oliveira Neto AF, et al. Pre-validation of the WHO organ dysfunction based criteria for identification of maternal near miss. *Reprod Health.* 2011;8:22.

37. Souza JP, Cecatti JG, Haddad SM, et al. The WHO maternal near-miss approach and the maternal severity index model (MSI): tools for assessing the management of severe maternal morbidity. *PLoS One.* 2012;7(8):e44129.

38. Souza JP, Gulmezoglu AM, Vogel J, et al. Moving beyond essential interventions for reduction of maternal mortality (the WHO Multicountry Survey on Maternal and Newborn Health): a cross-sectional study. *Lancet.* 2013;381(9879):1747-1755.

39. Nelissen E, Mduma E, Broerse J, et al. Applicability of the WHO maternal near miss criteria in a low-resource setting. *PLoS One.* 2013;8(4):e61248.

40. van den Akker T, Beltman J, Leyten J, et al. The WHO maternal near miss approach: consequences at Malawian District level. *PLoS One.* 2013;8(1):e54805.

41. Witteveen T, Bezstarosti H, de Koning I, et al. Validating the WHO maternal near miss tool: comparing high- and low-resource settings. *BMC Pregnancy Childbirth.* 2017;17(1):194.

42. American College of O, Gynecologists, the Society for Maternal-Fetal M, Kilpatrick SK, Ecker JL. Severe maternal morbidity: screening and review. *Am J Obstet Gynecol.* 2016;215(3):B17-22.

43. Bennett TA, Kotelchuck M, Cox CE, Tucker MJ, Nadeau DA. Pregnancy-associated hospitalizations in the United States in 1991 and 1992: a comprehensive view of maternal morbidity. *Am J Obstet Gynecol.* 1998;178(2):346-354.

44. Bacak SJ, Callaghan WM, Dietz PM, Crouse C. Pregnancy-associated hospitalizations in the United States, 1999-2000. *Am J Obstet Gynecol.* 2005;192(2):592-597.

45. Danel I, Berg C, Johnson CH, Atrash H. Magnitude of maternal morbidity during labor and delivery: United States, 1993-1997. *Am J Public Health.* 2003;93(4):631-634.

46. Wen SW, Huang L, Liston R, et al. Severe maternal morbidity in Canada, 1991-2001. *CMAJ.* 2005;173(7):759-764.

47. Frederiksen BN, Lillehoj CJ, Kane DJ, Goodman D, Rankin K. Evaluating Iowa Severe Maternal Morbidity Trends and Maternal Risk Factors: 2009-2014. *Maternal and child health journal.* 2017;21(9):1834-1844.

48. Gray KE, Wallace ER, Nelson KR, Reed SD, Schiff MA. Population-based study of risk factors for severe maternal morbidity. *Paediatric and perinatal epidemiology.* 2012;26(6):506-514.

49. Guglielminotti J, Landau R, Wong CA, Li G. Patient-, Hospital-, and Neighborhood-Level Factors Associated with Severe Maternal Morbidity During Childbirth: A Cross-Sectional Study in New York State 2013-2014. *Maternal and child health journal.* 2018.

50. Kilpatrick SJ, Abreo A, Gould J, Greene N, Main EK. Confirmed severe maternal morbidity is associated with high rate of preterm delivery. *Am J Obstet Gynecol.* 2016;215(2):233 e231-237.

51. Himes K.P. BLM. Validity of commonly used screening criteria for severe maternal morbidity. *Under review.* 2019.

52. Geller SE, Rosenberg D, Cox S, Brown M, Simonson L, Kilpatrick S. A scoring system identified near-miss maternal morbidity during pregnancy. *J Clin Epidemiol.* 2004;57(7):716-720.

53. Gao C, Osmundson S, Yan X, Edwards DV, Malin BA, Chen Y. Learning to Identify Severe Maternal Morbidity from Electronic Health Records. *Stud Health Technol Inform.* 2019;264:143-147.

54. Bateman BT, Mhyre JM, Hernandez-Diaz S, et al. Development of a comorbidity index for use in obstetric patients. *Obstet Gynecol.* 2013;122(5):957-965.

55. Leonard SA, Kennedy CJ, Carmichael SL, Lyell DJ, Main EK. An Expanded Obstetric Comorbidity Scoring System for Predicting Severe Maternal Morbidity. *Obstet Gynecol.* 2020;136(3):440-449.

56. Easter SR, Bateman BT, Sweeney VH, et al. A comorbidity-based screening tool to predict severe maternal morbidity at the time of delivery. *Am J Obstet Gynecol.* 2019;221(3):271 e271-271 e210.

57. Rossi RM, Hall E, Dufendach K, DeFranco EA. Predictive Model of Factors Associated With Maternal Intensive Care Unit Admission. *Obstet Gynecol.* 2019;134(2):216-224.

58. Knight M. Defining severe maternal morbidity-When is it time to stop? *Paediatric and perinatal epidemiology.* 2020;34(4):384-385.

59. Bouvier-Colle MH, Mohangoo AD, Gissler M, et al. What about the mothers? An analysis of maternal mortality and morbidity in perinatal health surveillance systems in Europe. *BJOG.* 2012;119(7):880-889; discussion 890.

60. Brandt JS, Srinivas SK, Elovitz ME, Bastek JA. Does a maternal-fetal medicine-centered labor and delivery coverage model put the 'M' back in MFM? *Am J Obstet Gynecol.* 2014;210(4):333.e331-333.e337.

61. Friedman AM, Ananth CV, Huang Y, D'Alton ME, Wright JD. Hospital delivery volume, severe obstetrical morbidity, and failure to rescue. *Am J Obstet Gynecol.* 2016;215(6):795.e791-795.e714.

62. Hitti J, Sienas L, Walker S, Benedetti TJ, Easterling T. Contribution of hypertension to severe maternal morbidity. *Am J Obstet Gynecol.* 2018.

63. Lazariu V, Nguyen T, McNutt LA, Jeffrey J, Kacica M. Severe maternal morbidity: A population-based study of an expanded measure and associated factors. *PLoS One.* 2017;12(8):e0182343.

64. Machiyama K, Hirose A, Cresswell JA, et al. Consequences of maternal morbidity on health-related functioning: a systematic scoping review. *BMJ Open.* 2017;7(6):e013903.

65. Storeng KT, Drabo S, Ganaba R, Sundby J, Calvert C, Filippi V. Mortality after near-miss obstetric complications in Burkina Faso: medical, social and health-care factors. *Bull World Health Organ.* 2012;90(6):418-425B.

66. Assarag B, Dujardin B, Essolbi A, Cherkaoui I, De Brouwere V. Consequences of severe obstetric complications on women's health in Morocco: please, listen to me! *Trop Med Int Health.* 2015;20(11):1406-1414.

67. Norhayati MN, Nik Hazlina NH, Aniza AA. Immediate and long-term relationship between severe maternal morbidity and health-related quality of life: a prospective double cohort comparison study. *BMC Public Health.* 2016;16(1):818.

68. Norhayati MN, Nik Hazlina NH, Aniza AA. Functional status of women with and without severe maternal morbidity: A prospective cohort study. *Women Birth.* 2016;29(5):443-449.

69. Norhayati MN, Azman Yacob M. Long-term postpartum effect of severe maternal morbidity on sexual function. *Int J Psychiatry Med.* 2017;52(4-6):328-344.

70. Norhayati MN, Nik Hazlina NH, Aniza AA, Asrenee AR. Severe Maternal Morbidity and Postpartum Depressive Symptomatology: A Prospective Double Cohort Comparison Study. *Res Nurs Health.* 2016;39(6):415-425.

71. Silveira C, Parpinelli MA, Pacagnella RC, et al. A cohort study of functioning and disability among women after severe maternal morbidity. *Int J Gynaecol Obstet.* 2016;134(1):87-92.

72. Poursharif B, Korst LM, Fejzo MS, MacGibbon KW, Romero R, Goodwin TM. The psychosocial burden of hyperemesis gravidarum. *J Perinatol.* 2008;28(3):176-181.

73. Adaji SE, Shittu OS, Bature SB, Nasir S, Olatunji O. Suffering in silence: pregnant women's experience of urinary incontinence in Zaria, Nigeria. *Eur J Obstet Gynecol Reprod Biol.* 2010;150(1):19-23.

74. Bangser M, Mehta M, Singer J, Daly C, Kamugumya C, Mwangomale A. Childbirth experiences of women with obstetric fistula in Tanzania and Uganda and their implications for fistula program development. *Int Urogynecol J.* 2011;22(1):91-98.

75. Waterstone M, Wolfe C, Hooper R, Bewley S. Postnatal morbidity after childbirth and severe obstetric morbidity. *BJOG.* 2003;110(2):128-133.

76. Mogos MF, August EM, Salinas-Miranda AA, Sultan DH, Salihu HM. A Systematic Review of Quality of Life Measures in Pregnant and Postpartum Mothers. *Appl Res Qual Life.* 2013;8(2):219-250.

77. Furuta M, Sandall J, Cooper D, Bick D. The relationship between severe maternal morbidity and psychological health symptoms at 6-8 weeks postpartum: a prospective cohort study in one English maternity unit. *BMC Pregnancy Childbirth.* 2014;14:133.

78. Lewkowitz AK, Rosenbloom JI, Keller M, et al. Association Between Severe Maternal Morbidity and Psychiatric Illness Within 1 Year of Hospital Discharge After Delivery. *Obstet Gynecol.* 2019;134(4):695-707.

79. *Report from nine maternal mortality review committees.* 2018.

80. Ackerman CM, Platner MH, Spatz ES, et al. Severe cardiovascular morbidity in women with hypertensive diseases during delivery hospitalization. *Am J Obstet Gynecol.* 2019;220(6):582 e581-582 e511.

81. Malhame I, Danilack VA, Raker CA, et al. Cardiovascular Severe Maternal Morbidity in Pregnant and Postpartum Women: Development and Internal Validation of Risk Prediction Models. *BJOG.* 2020.

82. Kolte D, Khera S, Aronow WS, et al. Temporal trends in incidence and outcomes of peripartum cardiomyopathy in the United States: a nationwide population-based study. *J Am Heart Assoc.* 2014;3(3):e001056.

83. Girsen AI, Sie L, Carmichael SL, et al. Rate and causes of severe maternal morbidity at readmission: California births in 2008-2012. *J Perinatol.* 2020;40(1):25-29.

84. Benschop L, Duvekot JJ, Roeters van Lennep JE. Future risk of cardiovascular disease risk factors and events in women after a hypertensive disorder of pregnancy. *Heart.* 2019;105(16):1273-1278.

85. Behrens I, Basit S, Melbye M, et al. Risk of post-pregnancy hypertension in women with a history of hypertensive disorders of pregnancy: nationwide cohort study. *BMJ.* 2017;358:j3078.

86. Lykke JA, Langhoff-Roos J, Sibai BM, Funai EF, Triche EW, Paidas MJ. Hypertensive pregnancy disorders and subsequent cardiovascular morbidity and type 2 diabetes mellitus in the mother. *Hypertension.* 2009;53(6):944-951.

87. Daw JR, Hatfield LA, Swartz K, Sommers BD. Women In The United States Experience High Rates Of Coverage 'Churn' In Months Before And After Childbirth. *Health Aff (Millwood).* 2017;36(4):598-606.

88. Bullinger LR. The Effect of Paid Family Leave on Infant and Parental Health in the United States. *J Health Econ.* 2019;66:101-116.

89. Zauderer C. Postpartum depression: how childbirth educators can help break the silence. *J Perinat Educ.* 2009;18(2):23-31.

90. Markus AR, Andres E, West KD, Garro N, Pellegrini C. Medicaid covered births, 2008 through 2010, in the context of the implementation of health reform. *Womens Health Issues.* 2013;23(5):e273-280.

91. Jarlenski M, Hutcheon JA, Bodnar LM, Simhan HN. State Medicaid Coverage of Medically Necessary Abortions and Severe Maternal Morbidity and Maternal Mortality. *Obstet Gynecol.* 2017;129(5):786-794.

92. Naimi AI, Balzer LB. Stacked generalization: an introduction to super learning. *Eur J Epidemiol.* 2018;33(5):459-464.

93. Polley EC VdLM. Super learner in prediction. 2010.

94. Van der Laan MJ PE, Hubbard AE. Super Learner. *Statistical Applications in Genetics and Molecular Biology.* 2007;6(1).

95. Sun Y, Wong, A., Kamel, M. Classification of imbalanced data: a review. *International Journal of Pattern Recognition and Artificial Intelligence.* 2011;23(4):687-719.

96. van der Laan M, Dudoit S. Unified Cross-Validation

Methodology for Selection Among Estimators and a General CrossValidated Adaptive Epsilon-Net Estimator: Finite Sample Oracle Inequalities and Examples. *Technical Report 130, Division of Biostatistics, University of California, Berkeley.* 2003.

97. Strobl C, Zeileis, A. Danger: High power! - Exploring the statistical properties of a test for random forest variable importance. 2008.

98. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol.* 2019;110:12-22.

99. Hoyert DL, Miniño A.M. *Maternal mortality in the United States: Changes in coding, publication, and data release, 2018.* . Hyattsville, MD: National Center for Health Statistics;2020.

100. CDC. Severe Maternal Morbidity in the United States. 2017.

101. Facca TA, Mastroianni-Kirsztajn G, Sabino ARP, et al. Pregnancy as an early stress test for cardiovascular and kidney disease diagnosis. *Pregnancy Hypertens.* 2018;12:169-173.

102. Chasan-Taber L. It Is Time to View Pregnancy as a Stress Test. *Journal of women's health (2002).* 2016;25(1):2-3.

103. Craici I, Wagner S, Garovic VD. Preeclampsia and future cardiovascular risk: formal risk factor or failed stress test? *Ther Adv Cardiovasc Dis.* 2008;2(4):249-259.

104. Harvey EM, Ahmed S, Manning SE, Diop H, Argani C, Strobino DM. Severe Maternal Morbidity at Delivery and Risk of Hospital Encounters Within 6 Weeks and 1 Year Postpartum. *Journal of women's health (2002).* 2018;27(2):140-147.

105. Lo-Ciganic WH, Donohue JM, Kim JY, et al. Adherence trajectories of buprenorphine therapy among pregnant women in a large state Medicaid program in the United States. *Pharmacoepidemiol Drug Saf.* 2019;28(1):80-89.

106. Jarlenski MP, Krans EE, Kim JY, et al. Five-Year Outcomes Among Medicaid-Enrolled Children With In Utero Opioid Exposure. *Health Aff (Millwood).* 2020;39(2):247-255.

107. Krans EE, Kim JY, James AE, 3rd, Kelley DK, Jarlenski M. Postpartum contraceptive use and interpregnancy interval among women with opioid use disorder. *Drug Alcohol Depend.* 2018;185:207-213.

108. Weissman GE, Hubbard RA, Kohn R, et al. Validation of an Administrative Definition of ICU Admission Using Revenue Center Codes. *Crit Care Med.* 2017;45(8):e758-e762.

109. Berg CJ, Harper MA, Atkinson SM, et al. Preventability of pregnancy-related deaths: results of a state-wide review. *Obstet Gynecol.* 2005;106(6):1228-1234.

110. Howe CJ, Cole SR, Lau B, Napravnik S, Eron JJ, Jr. Selection Bias Due to Loss to Follow Up in Cohort Studies. *Epidemiology (Cambridge, Mass).* 2016;27(1):91-97.

111. Mitra M, Parish SL, Clements KM, Cui X, Diop H. Pregnancy outcomes among women with intellectual and developmental disabilities. *Am J Prev Med.* 2015;48(3):300-308.

112. Cole SR, Hudgens MG, Brookhart MA, Westreich D. Risk. *Am J Epidemiol.* 2015;181(4):246-250.

113. Vansteelandt S, Keiding N. Invited commentary: G-computation--lost in translation? *Am J Epidemiol.* 2011;173(7):739-742.

114. Robins JM. A new approach to causal inference in mortality studies with sustained exposure periods - application to control the healthy worker survivor effect. *Math Modell.* 1986;7:1393-1512.

115. Edwards JK, Cole SR, Westreich D, et al. Age at Entry Into Care, Timing of Antiretroviral Therapy Initiation, and 10-Year Mortality Among HIV-Seropositive Adults in the United States. *Clin Infect Dis.* 2015;61(7):1189-1195.

116. Westreich D, Cole SR, Young JG, et al. The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or death. *Stat Med.* 2012;31(18):2000-2009.

117. D'Agostino RB, Lee ML, Belanger AJ, Cupples LA, Anderson K, Kannel WB. Relation of pooled logistic regression to time dependent Cox regression analysis: the Framingham Heart Study. *Stat Med.* 1990;9(12):1501-1515.

118. Textor J, Hardt J, Knuppel S. DAGitty: a graphical tool for analyzing causal diagrams. *Epidemiology (Cambridge, Mass).* 2011;22(5):745.

119. Geller SE, Koch AR, Garland CE, MacDonald EJ, Storey F, Lawton B. A global view of severe maternal morbidity: moving beyond maternal mortality. *Reprod Health.* 2018;15(Suppl 1):98.

120. Ferreira EC, Costa ML, Pacagnella RC, et al. General and reproductive health among women after an episode of severe maternal morbidity: Results from the COMMAG study. *Int J Gynaecol Obstet.* 2020.

121. England N, Madill J, Metcalfe A, et al. Monitoring maternal near miss/severe maternal morbidity: A systematic review of global practices. *PLoS One.* 2020;15(5):e0233697.

122. Garcia M, Mulvagh SL, Merz CN, Buring JE, Manson JE. Cardiovascular Disease in Women: Clinical Perspectives. *Circ Res.* 2016;118(8):1273-1293.

123. Prevention CfDCa. Leading Causes of Death - Females - All races and origins - United States, 2017. 2019; https://www.cdc.gov/women/lcod/2017/all-races-origins/index.htm.

124. Lykke JA, Langhoff-Roos J, Lockwood CJ, Triche EW, Paidas MJ. Mortality of mothers from cardiovascular and non-cardiovascular causes following pregnancy complications in first delivery. *Paediatric and perinatal epidemiology.* 2010;24(4):323-330.

125. Aye CY, Boardman H, Leeson P. Cardiac Disease after Pregnancy: A Growing Problem. *Eur Cardiol.* 2017;12(1):20-23.

126. Lane-Cordova AD, Gunderson EP, Greenland P, et al. Life-Course Reproductive History and Cardiovascular Risk Profile in Late Mid-Life: The CARDIA Study. *J Am Heart Assoc.* 2020:e014859.

127. Wu P, Haththotuwa R, Kwok CS, et al. Preeclampsia and Future Cardiovascular Health: A Systematic Review and Meta-Analysis. *Circ Cardiovasc Qual Outcomes.* 2017;10(2).

128. Brown MC, Best KE, Pearce MS, Waugh J, Robson SC, Bell R. Cardiovascular disease risk in women with pre-eclampsia: systematic review and meta-analysis. *Eur J Epidemiol.* 2013;28(1):1-19.

129. Heida KY, Franx A, van Rijn BB, et al. Earlier Age of Onset of Chronic Hypertension and Type 2 Diabetes Mellitus After a Hypertensive Disorder of Pregnancy or Gestational Diabetes Mellitus. *Hypertension.* 2015;66(6):1116-1122.

130. Osol G, Bernstein I. Preeclampsia and maternal cardiovascular disease: consequence or predisposition? *J Vasc Res.* 2014;51(4):290-304.

131. Smith GN, Louis JM, Saade GR. Pregnancy and the Postpartum Period as an Opportunity for Cardiovascular Risk Identification and Management. *Obstet Gynecol.* 2019;134(4):851-862.

132. Main EK, Abreo A, McNulty J, et al. Measuring severe maternal morbidity: validation of potential measures. *Am J Obstet Gynecol.* 2016;214(5):643 e641-643 e610.

133. Cozzolino F, Montedori A, Abraha I, et al. A diagnostic accuracy study validating cardiovascular ICD-9-CM codes in healthcare administrative databases. The Umbria Data-Value Project. *PLoS One.* 2019;14(7):e0218919.

134. Ando T, Ooba N, Mochizuki M, et al. Positive predictive value of ICD-10 codes for acute myocardial infarction in Japan: a validation study at a single center. *BMC Health Serv Res.* 2018;18(1):895.

135. Frolova N, Bakal JA, McAlister FA, et al. Assessing the use of international classification of diseases-10th revision codes from the emergency department for the identification of acute heart failure. *JACC Heart Fail.* 2015;3(5):386-391.

136. Ammann EM, Kalsekar I, Yoo A, Johnston SS. Validation of body mass index (BMI)-related ICD-9-CM and ICD-10-CM administrative diagnosis codes recorded in US claims data. *Pharmacoepidemiol Drug Saf.* 2018;27(10):1092-1100.

137. Gribsholt SB, Pedersen L, Richelsen B, Thomsen RW. Validity of ICD-10 diagnoses of overweight and obesity in Danish hospitals. *Clin Epidemiol.* 2019;11:845-854.

138. Hernan MA, Robins, J.M. *Causal Inference: What If.* Boca Raton: Chapman & Hall/CRC; 2020.

139. Hernan MA, Taubman SL. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. *Int J Obes (Lond).* 2008;32 Suppl 3:S8-14.

140. Wen T, Krenitsky NM, Clapp MA, et al. Fragmentation of postpartum readmissions in the United States. *Am J Obstet Gynecol.* 2020.

141. Gynecologists ACoOa. Optimizing postpartum care. ACOG Committee Opinion No. 736. *Obstet Gynecol.* 2018;131:e140-150.

142. Chawla NV, Japkowicz N, Kotcz A. Special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter.* 2004;6(1):1-6.

143. Naimi AI, Platt RW, Larkin JC. Machine Learning for Fetal Growth Prediction. *Epidemiology (Cambridge, Mass).* 2018;29(2):290-298.

144. *SuperLearner: Super Learner Prediction.* [computer program]. Version R package version 2.0-2.4. https://cran.r-project.org/web/packages/SuperLearner/index.html.

145. Batista G, Prati, RC, Monard, MC. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter.* 2004;6(1):20-29.

146. Chawla NV BK, Hall L, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artifi Intell Res.* 2002;16:321-357.

147. Kuhn M. Building predictive models in R using the caret package. *J Stat Soft.* 2008;28(5):1-26.

148. *pROC: an open-source package for R and S+ to analyze and compare ROC curves* [computer program]. BMC Bioinformatics2011.

149. Chen HY, Chauhan SP, Blackwell SC. Severe Maternal Morbidity and Hospital Cost among Hospitalized Deliveries in the United States. *American journal of perinatology.* 2018.

150. Schaap T, Bloemenkamp K, Deneux-Tharaux C, et al. Defining definitions: a Delphi study to develop a core outcome set for conditions of severe maternal morbidity. *BJOG.* 2019;126(3):394-401.

151. Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan MJ. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. *Lancet Respir Med.* 2015;3(1):42-52.

152. Mullainathan S. SJ. Machine learning: an applied economic approach. *Journal of Economic Perspectives.* 2017;31(2):87-106.

153. Coyle J. HNS, Malenica, I., Sofrygin, O. sl3: Modern Pipelines for Machine Learning and Super Learning. https://github.com/tlverse/sl3. 2020.

154. Wald NJ, Hackshaw AK, Frost CD. When can a risk factor be used as a worthwhile screening test? *BMJ.* 1999;319(7224):1562-1565.

155. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol.* 2004;159(9):882-890.

156. Fox MP, Lash TL, Bodnar LM. Common misconceptions about validation studies. *Int J Epidemiol.* 2020.

157. Krawczyk B. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence.* 2016;5(4):221-232.

158. Kreatsoulas C, Subramanian SV. Machine learning in social epidemiology: Learning from experience. *SSM Popul Health.* 2018;4:347-349.

159. Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH. Predicting sample size required for classification performance. *BMC Med Inform Decis Mak.* 2012;12:8.

160. Hua J, Xiong Z, Lowey J, Suh E, Dougherty ER. Optimal number of features as a function of sample size for various classification rules. *Bioinformatics.* 2005;21(8):1509-1515.

161. Wiemken TL, Kelley RR. Machine Learning in Epidemiology and Health Outcomes Research. *Annu Rev Public Health.* 2020;41:21-36.

162. Deputy NP, Dub B, Sharma AJ. Prevalence and Trends in Prepregnancy Normal Weight - 48 States, New York City, and District of Columbia, 2011-2015. *MMWR Morb Mortal Wkly Rep.* 2018;66(51-52):1402-1407.

163. Lydon-Rochelle MT, Holt VL, Nelson JC, et al. Accuracy of reporting maternal in-hospital diagnoses and intrapartum procedures in Washington State linked birth records. *Paediatric and perinatal epidemiology.* 2005;19(6):460-471.

164. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One.* 2015;10(3):e0118432.

165. Boehmke B, Greenwell, B.M. *Hands-on machine learning with R.* CRC Press; 2019.

166. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is Machine Learning? A Primer for the Epidemiologist. *Am J Epidemiol.* 2019;188(12):2222-2239.

167. Prevention. CfDCa. *National Diabetes Statistics Report, 2020.* Atlanta, GA: Centers for Disease Control and Prevention, U.S. Dept of Health and Human Services;2020.

168. Lawton BA, Jane MacDonald E, Stanley J, Daniells K, Geller SE. Preventability review of severe maternal morbidity. *Acta Obstet Gynecol Scand.* 2019;98(4):515-522.

169. American College of O, Gynecologists' Committee on Obstetric P, Association of Women's Health O, Neonatal N. Committee Opinion No. 666: Optimizing Postpartum Care. *Obstet Gynecol.* 2016;127(6):e187-192.