


# Innovations in suicide prevention research (INSPIRE): a protocol for a population-based case–control study

Shabbar I Ranapurwala <sup>1,2</sup>, Vanessa E Miller,<sup>2</sup> Timothy S Carey,<sup>3,4</sup> Bradley N Gaynes,<sup>5</sup> Alexander P Keil,<sup>1</sup> Catherine Vinita Fitch <sup>1,2</sup>, Monica E Swilley-Martinez,<sup>1,2</sup> Andrew L Kavee,<sup>3</sup> Toska Cooper,<sup>2</sup> Samantha Dorris,<sup>2</sup> David B Goldston,<sup>6</sup> Lewis J Peiper <sup>7</sup>, Brian W Pence<sup>1,2</sup>

<sup>1</sup>Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

<sup>2</sup>Injury Prevention Research Center, University of North Carolina, Chapel Hill, North Carolina, USA

<sup>3</sup>Cecil G Sheps Center for Health Services Research, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

<sup>4</sup>Department of Medicine, University of North Carolina at Chapel Hill School of Medicine, Chapel Hill, North Carolina, USA

<sup>5</sup>Department of Psychiatry, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

<sup>6</sup>Department of Psychiatry and Behavioral Sciences, Duke University School of Medicine, Durham, North Carolina, USA

<sup>7</sup>Division of Adult Correction – Prisons, North Carolina Department of Public Safety, Raleigh, North Carolina, USA

## Correspondence to

Dr Shabbar I Ranapurwala, Epidemiology, University of North Carolina at Chapel Hill Department of Epidemiology, Chapel Hill, NC 27510, USA; sirana@email.unc.edu

Received 11 April 2022

Accepted 28 May 2022



© Author(s) (or their employer(s)) 2022. No commercial re-use. See rights and permissions. Published by BMJ.

**To cite:** Ranapurwala SI, Miller VE, Carey TS, *et al.* *Inj Prev* Epub ahead of print: [please include Day Month Year]. doi:10.1136/injuryprev-2022-044609

## ABSTRACT

**Background** Suicide deaths have been increasing for the past 20 years in the USA resulting in 45 979 deaths in 2020, a 29% increase since 1999. Lack of data linkage between entities with potential to implement large suicide prevention initiatives (health insurers, health institutions and corrections) is a barrier to developing an integrated framework for suicide prevention.

**Objectives** Data linkage between death records and several large administrative datasets to (1) estimate associations between risk factors and suicide outcomes, (2) develop predictive algorithms and (3) establish long-term data linkage workflow to ensure ongoing suicide surveillance.

**Methods** We will combine six data sources from North Carolina, the 10th most populous state in the USA, from 2006 onward, including death certificate records, violent deaths reporting system, large private health insurance claims data, Medicaid claims data, University of North Carolina electronic health records and data on justice involved individuals released from incarceration. We will determine the incidence of death from suicide, suicide attempts and ideation in the four subpopulations to establish benchmarks. We will use a nested case–control design with incidence density-matched population-based controls to (1) identify short-term and long-term risk factors associated with suicide attempts and mortality and (2) develop machine learning-based predictive algorithms to identify individuals at risk of suicide deaths.

**Discussion** We will address gaps from prior studies by establishing an in-depth linked suicide surveillance system integrating multiple large, comprehensive databases that permit establishment of benchmarks, identification of predictors, evaluation of prevention efforts and establishment of long-term surveillance workflow protocols.

## INTRODUCTION

Suicide mortality rates continue to climb in the USA, despite a decreasing global trend. In 2020, there were 45 979 lives lost to suicide in the USA, at an age-adjusted rate of 13.5 per 100 000 population,<sup>1</sup> representing a 33% increase since 1999, and making suicide the 10th leading cause of death nationwide.<sup>1</sup> The suicide trends in North Carolina (NC)—the setting of this research—are similar to those of the nation overall, with increases from about 13 suicide deaths to 16 suicide deaths per 100 000 population from 2012 to 2018 (or 1527 lives lost in 2017).<sup>2,3</sup>

## WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Our current understanding of and response to suicide deaths are severely impacted by (1) lack of large, longitudinal, population-based studies, (2) limited internal validity from studies using specialised populations, (3) a general focus on reporting associations for individual factors without combining factors in a fashion that may be able to help develop tools for clinicians and (4) lack of an ongoing surveillance system to assess emerging trends, establish benchmarks and evaluate ongoing interventions.

## WHAT THIS STUDY ADDS

⇒ Our study addresses these gaps by establishing an in-depth linked suicide surveillance system integrating multiple large, comprehensive databases from the healthcare system, public insurer, private insurer and corrections perspectives in North Carolina (NC), the 10th most populous state in the USA.

## HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE AND/OR POLICY

⇒ Our project will be the first to link these large state-level and healthcare system-level databases to establish an ongoing suicide surveillance system to identify short-term and long-term predictors of suicide and to inform and evaluate suicide prevention efforts in NC and across the USA. We will also develop tools to identify patients at a high risk for suicide in a clinic-based setting and help adequately link them to mental health treatment resources.

In NC, suicide is the third leading cause of death among those aged 5–44 years.<sup>3</sup> Suicides mortality rates vary by gender, race, and age.<sup>1</sup>

Substantial research has explored association of single predictors with suicide mortality risk.<sup>4–8</sup> These predictors can be separated into distal factors (ie, genetics, personality, fetal exposures, adverse childhood experiences or mental health and substance use disorders) and proximal stressors (ie, financial insecurity, occupational stress, legal problems or health problems).<sup>6</sup> Of these, mental health and substance use disorders are commonly understood to be the strongest predictors of suicide death and suicidal behaviours.<sup>6,8,9</sup> For example, one review on suicide deaths reported a standardised

mortality ratio (SMR) for borderline personality disorder of 45.1 (95% CI 29.0 to 61.3) and an SMR for opioid use of 13.5 (95% CI 10.5 to 17.2).<sup>5</sup> The relationships between physical health disorders and suicide deaths were recently evaluated in a linked US population of health maintenance organisation healthcare systems (mainly Kaiser Permanente), with traumatic brain injury (adjusted OR: 8.80; 95% CI 7.37 to 10.50) representing one of the strongest physical disorder risk factors.<sup>10</sup> Other studies have also highlighted many such risk factors<sup>11</sup> and their association with self-harm and suicide.<sup>12–13</sup> However, association of single risk factors does little to aid clinicians in intervening at the right time to prevent suicide in their patients. For example, 54% of people who died of suicides did not suffer from a mental health condition,<sup>14</sup> which, even if was due to lack of information on mental health disorders, suggests that the strong SMRs related to mental health conditions are not enough for preventing the majority of suicide deaths on their own.<sup>15</sup> New research considering development of predictive algorithms that can incorporate a multivariable approach in clinic-based settings is urgently needed.<sup>15</sup>

Prior research shows that healthcare utilisation increases immediately before death from suicide.<sup>16–21</sup> Hence, the healthcare system may offer a good opportunity for suicide prevention. Prior studies from outside the USA have evaluated the associations between healthcare utilisation and suicide, using large, longitudinal, population-based record linkage. However, the populations represented in these studies are exposed to very different health systems than are found in the USA.<sup>16–21</sup> Care access patterns among US patients suffering from mental health disorders and exhibiting suicidal ideation may be different.

While some US-based large linkage studies have shed light on important risk factors, they lack ongoing linkage and surveillance, and are non-representative of the US population. Some active US-based studies have linked individuals determined to have demonstrated self-harm behaviours to mortality.<sup>13 22–24</sup> However, this focus on self-harm may miss many individuals at risk of suicide death, as many individuals who die from suicide do not present with self-harm during their healthcare interactions prior to death. Other rigorous US-based linkage studies from South Carolina and Kentucky have linkages between violent death data and statewide healthcare data, which have been used to describe healthcare utilisation.<sup>25 26</sup> The Mental Health Research Network (MHRN)<sup>27</sup> has taken this a step further, using a case–control design for 2674 suicides in eight health maintenance organization (HMO) systems.<sup>28</sup> However, the MHRN has little representation of Medicaid or self-pay individuals.

Despite these advances, the US evidence base for health exposure-based suicide prediction remains limited in number and subject to internal and external validity limitations. Because national registries do not exist, US studies very frequently focus on suicide mortality risk factors for specific vulnerable populations, including youths,<sup>29 30</sup> those  $\geq 65$  years of age,<sup>31</sup> prisoners,<sup>32</sup> formerly incarcerated individuals,<sup>33 34</sup> veterans,<sup>35–41</sup> patients with specific diagnoses<sup>42</sup> or patients who screen as high risk for death from suicide.<sup>12 22</sup> Even the HMO-based MHRN, which strives for a more population-based sample, does not represent uninsured patients, those with Medicaid or individuals with criminal justice involvement, which are highly vulnerable populations. This research is also subject to internal validity concerns, because the models frequently only adjust for age and sex, without addressing confounding between health exposures and suicide.

Overall, the gaps in our current understanding of and response to suicide deaths are severely impacted by (a) lack of

large, longitudinal, population-based studies, (b) limited internal validity from studies using specialised populations, (c) a general focus on reporting associations for individual factors without combining factors in a fashion that may be able to help develop tools for clinicians and (d) lack of an ongoing surveillance system to assess emerging trends, establish benchmarks and evaluate ongoing interventions.

Our study addresses these gaps by establishing an in-depth linked suicide surveillance system integrating multiple large, comprehensive databases that permit establishment of benchmarks, identification of predictors, evaluation of prevention efforts and establishment of long-term surveillance workflow protocols from the healthcare system, public insurer, private insurer and corrections department perspectives in NC, the 10th most populous state in the USA.

The specific objectives of our study are:

1. Define the incidence of death from suicide in (a) a large healthcare system, (b) a large privately insured population, (c) a state Medicaid population and (d) justice involved individuals released from incarceration in a NC prison, so as to establish benchmarks for each of these entities to evaluate future prevention initiatives. We will accomplish this aim by developing a surveillance system through individual linkage of the state death registry and NC Violent Deaths Reporting System (NCVDRS) to healthcare, private and Medicaid claims, and NC corrections data.
2. Identify demographic, clinical and short-term and long-term care access patterns that predict risk of death from suicide, suicide attempts and suicidal ideation in four study populations, addressing limitations of prior research through large samples, machine learning and density-matched population-based case–control designs.
3. Establish a long-term workflow protocol for each of these four systems to regularly update linkages to prospectively maintain surveillance, monitor benchmarks and evaluate future prevention initiatives.

## METHODS AND ANALYSIS

### Study design

In this study, we will link multiple large data sources from 2006 onward to accomplish the four aims of the study. The data come from the 10th most populous state in the USA, NC. We will link mortality information from two comprehensive statewide systems, the NC death certificate data and the NCVDRS, to four different target subpopulations representing possible suicide prevention intervention points: (1) electronic health records of a large healthcare system serving residents of all 100 counties of the state, (2) claims records of a large single private health insurance carrier in the state, (3) claims records of the state Medicaid programme and (4) justice individuals released from the NC prisons following incarceration.

Completion of these linkages will permit us to (1) establish incidence benchmarks of suicide deaths over multiple years and across subgroups for each of these four actors, (2) complete a comprehensive assessment of short-term and long-term predictors of suicide mortality risk in each population addressing sample size/power and methodological limitations of prior research and (3) establish a long-term workflow protocol and regularly update the linkages to maintain a prospective suicide surveillance system for the four NC-based populations. The study then will follow a retrospective cohort design for accomplishing objective 1, a nested case–control design for objective 2 and establish standard operating procedures or protocols for ongoing surveillance.

## Data sets and linkage

This study will link six data sources containing information on mortality, healthcare delivery, public and private health insurance and correctional services. We will measure suicide deaths by drawing on (1) NC death certificate records and (2) the NCVDRS. We will link these data to four different target populations of interest representing (3) a large healthcare system treating patients from all 100 counties (the University of NC healthcare system electronic health records (University of NC (UNC) healthcare system electronic health records (EHR)), (4) a large private insurance provider in the state, (5) the NC Medicaid population and (6) individuals recently released from a state prison (NC Division of Prisons (NC DOP)).

NC death certificate records are public records, housed at the NC Department of Health and Human Services (DHHS). NC DHHS provides death record data with identifiers directly to researchers with approved data use agreements. We will use death certificates from 2006 onward. Updated death records are released annually approximately 6 months after year's end.

NCVDRS is a state-based surveillance system that harvests information from law enforcement, coroners and medical examiners, vital statistics and crime laboratories to develop detailed records of violent deaths in each state. In addition to demographics and date and cause of death, VDRS data also include substance use and mental health disorders and treatment, and recent incarceration history prior to death, education and circumstantial information of the violent death, including factors such as loss of job, relationship, property or finances, or intimate partner violence. NCVDRS is housed at the NC DHHS Injury and Violence Prevention Branch. We will use NCVDRS from 2006 onward and receive updated data pulls annually.

Private health insurance claims data from a large private health insurance provider in NC that insures >3 million people in NC (30% of the NC population) with substantial geographic, socioeconomic and demographic diversity. These data are maintained by the UNC Sheps Center from 2006 onward, updated semiannually. Although analysis datasets provided to researchers have identifiers removed, the central dataset maintained by the Sheps Center includes all identifiers to enable internal Sheps programmers acting as 'honest brokers' to complete linkages to other data sources and provide researchers with de-identified, linked datasets.

NC Medicaid claims that data are overseen by the Division of Health Benefits (DHB) in the NC DHHS and made available to UNC researchers through the UNC Sheps Center's Carolina Cost and Quality Initiative. The linkage will be conducted by the DHB and will then provide the Sheps Center with the linked Medicaid IDs, which the Sheps Center programmers use to pull the population of interest and incidence density-matched control populations. NC Medicaid covers >2 million people in NC each year (~20% of NC population).

UNC electronic health records (UNC EHR): the UNC Health System provides medical care to >1 million patients (10% of NC population) annually across all 100 NC counties. The UNC EHR is a repository of this EHR data from 2006 to present, maintained by the NC Translational and Clinical Sciences Institute and updated daily. The UNC EHR contains full identifiers, which internal programmers serving as honest brokers can use to link to other data sources to provide linked, de-identified datasets to researchers. We will access data from 2006 onwards and update it annually.

NC DOP data in NC are considered public data and the data collected for this study include all incarceration release

data consisting of approximately 20 000–25 000 releases from incarceration in NC per year. The data include personal identifiers such as last, first, middle and maiden names, date of birth, gender, race, ethnicity, marital status, occupation and socioeconomic status, all dates of prison entry and exit and cause for incarceration. In addition, we will use non-public information from the NC DOP data, like type of confinement, participation in mental health and substance use disorder treatment and education programmes while in prison. We will access data from 2006 onward to be collected and linked on an annual basis.

Overall, the four subpopulations in this study represent about 35%–40% of NC population. Since the NC Medicaid and the private health insurer data mainly include under 65-year-old population, those 65 years or older may be under-represented in this study. However, the UNC EHR data and NC DOP datasets do contain 65 years and older population and would be representative of all age groups presented in those systems. We will first link NCVDRS with NC death certificates data using deterministic linkage based on date of death, death certificate, first letter of last name, age and sex. The four cohorts, NC Medicaid, UNC EHR, NC private health insurance and NC DOP, will be linked to mortality records, and with each other using deterministic linkage based on last and first names, date of birth and sex. Formal documentation of the linkage process used to establish the surveillance system will be prepared so that the entire system can be continuously updated in order to maintain surveillance and evaluate suicide prevention initiatives.

## Patient and public involvement

We will work with a clinical tailoring advisory board who will provide input into the development and tailoring of the clinical decision tools to ensure their usefulness and interpretability in clinical settings. The advisory board will include patients with mental health disorders, patient advocates, physicians who treat patients with mental health conditions and other mental health professionals.

## Outcomes

(1) Suicidal ideation identified using international classification of diseases (ICD) versions 9 (V62.84) and 10 (R45.851) codes—counted as only one encounter per person per year, (2) suicide attempts and/or self-harm or self-inflicted injury identified using ICD 9 (E950–E958) and ICD 10 (T14.91, T36–T71, X71–X84 and Z91.51) codes in UNC EHR, private health insurance and Medicaid—counted using initial encounters only and (3) suicide deaths identified using ICD 10 codes (U03, X60–X70, X71–X83 and Y87.0) via linkage of NC death records and NCVDRS with UNC EHR, private health insurance, Medicaid and NC DOP data.

## Predictors

The candidate predictors (table 1) include demographic variables such as age, sex, race, body mass index, mental health disorders (eg, anxiety disorder, depression and post-traumatic stress disorder), treatment for mental health disorders, substance use disorders (eg, opioids and alcohol), treatment for substance use disorders, medical history (eg, chronic pain and disability), access to care (eg, hospitalisations, frequency of visits and emergency department visits), medication history (opioids, anxiolytics and antidepressants), social history (alcohol, smoking and drug use) and circumstances (loss of relationship, employment and violence). We will follow two analytical approaches, first to

**Table 1** Available data sources and information

Available variables	Proposed study datasets					
	UNC EHR	NC private health insurance	NC Medicaid	NC DPS	NCVDRS	Death records
Demographics (age, sex and race)	Yes	Age and sex	Yes	Yes	Yes	Yes
Geographical information (ZIP codes)	Yes	Yes	Yes	Yes	Yes	Yes
Diagnoses (ICD codes)	Yes	Yes	Yes	In-prison	–	–
Suicide attempts (ICD codes)	Yes	Yes	Yes	In-prison	–	–
Deaths (date and cause: ICD 10)	In-hospital	In-hospital	In-hospital	In-prison	Yes	Yes
Circumstances (eg, violence or loss of job, relationship, etc)	–	–	–	–	Yes	–
Medications, dosage, duration	Yes	Yes	Yes	In-prison	–	–
Inpatient/outpatient	Yes	Yes	Yes	–	–	–
Suicide screening	Yes	–	–	In-prison	–	–
SUD (diagnosis and treatment)	Yes	Yes	Yes	In-prison	Yes	–
MHD (diagnosis and treatment)	Yes	Yes	Yes	In-prison	Yes	–
Social history (smoking/alcohol/drug use)	Yes	–	–	Yes	Yes	–
Laboratory results	Yes	–	–	–	–	–
Patient reported outcomes	Yes	–	–	–	–	–
Triage and chart notes	Yes	–	–	–	–	–

ICD, International classification of diseases; MHD, Mental health disorders; NC, North Carolina; NCVDRS, North Carolina Violent Deaths Reporting System; SUD, Substance use disorders; UNC EHR, University of North Carolina electronic health records.

estimate associations between risk factors and suicide outcomes, and second to develop predictive algorithms.

**Planned analytic approach: objective 1**

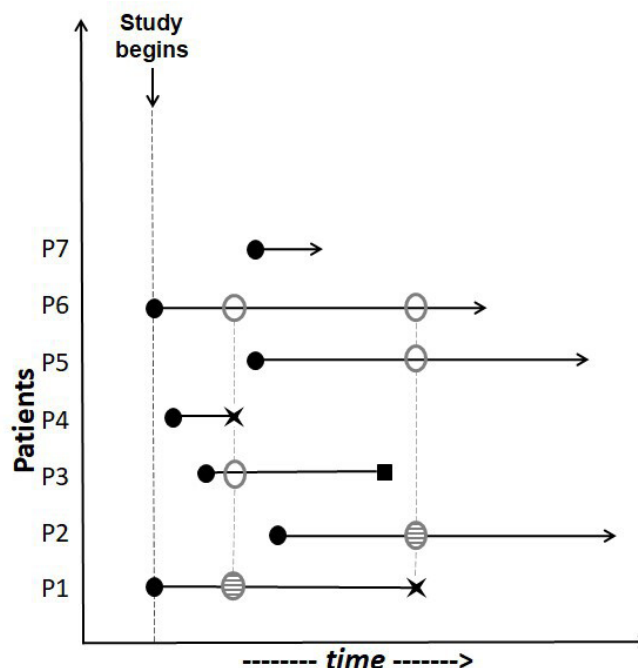
We will define rates of suicide deaths, suicide attempts (combined with self-harm or self-inflicted injury) and suicidal ideation by dividing the number of these outcomes by the total number of person-years of at-risk population from which the outcomes arise. These rates will be defined in each of the four study cohorts and by age, sex, race, ethnicity and other socio-demographic information.

**Planned analytic approach: objective 2**

We will conduct a nested case-control study, where each case will be defined as someone who presents with suicidal ideation, attempt or death. We will match each case with 10 randomly selected age-matched and sex-matched controls within 1 month of the incidence of the outcome such that the controls represent the denominator (person-time) from which the cases arose (figure 1). Incidence density-matched control selection overcomes important selection bias issues that arise with convenience or high-risk control groups common in the suicide literature. The cases and controls thus obtained will be used to (1) identify potential risk factors associated with suicide ideations, attempts and mortality and (2) develop a set of predictive algorithms to identify individuals at risk of suicide death over the next week (proximal risk prediction), month, 6 months and year (distal risk).

To estimate the association of short-term and long-term risk factors with suicide outcomes, we will conduct a set of conditional logistic regression analyses, while controlling for potential confounders. We will consider both bivariable (ie, descriptive) and multivariable associations of risk factors with outcomes. In multivariable models, covariates will be identified using directed acyclic graphs (DAG). When a particular risk factor (eg, history of suicide attempt) is of primary interest, the set of covariates will be tailored based on the exposure outcome relationship being examined. For example, in prior research, to estimate the

association between exposure to restrictive housing in prison and suicide mortality after incarceration release,<sup>34</sup> we controlled for a DAG-identified set of variables that included time-varying age, number of prior incarcerations, drug-related convictions, violence-related convictions, mental health treatment recommendation, mental health treatment received, number of days served in the most recent sentence and time-fixed sex. In addition, we will measure modification of such exposure outcome



**Figure 1** Incidence density control sampling: a nested case-control study. Black cross: suicide death/cases; black square: other cause of death; black circle: follow-up begins; black arrow: lost to follow or censored; grey circle: eligible controls; and shaded grey circle: randomly selected control.

**Table 2** Algorithms used in super learner, including source R package and tuning parameters

Algorithm	R package	Tuning parameters
Bagging of regression trees	ipred	Complexity parameter
Bayesian additive regression trees	bartMachine	Number of trees
Elastic net (including Ridge and LASSO)	glmnet	Penalty term and mixing parameter
Elastic net including first order interactions	glmnet	Penalty term and mixing parameter
Generalised additive model	mgcv	Spline df
Gradient boosting of regression trees	xgboost	Number of trees
Least angle regression	lars	None
Multivariate adaptive regression splines	plspline	None
Neural et	nnet, h2o	Number of hidden layers and nodes per layer
Random forest	ranger	Number of trees
Stepwise logistic regression	stats	None
Support vector machine	kernlab	None

relationships by demographic factors, especially age, sex, race and ethnicity.

To develop a predictive algorithm, we will use super learner.<sup>43</sup> We will develop a predictive algorithm for each of our two outcomes (suicide attempts and suicide deaths) within each time frame, focusing in particular on predicting outcomes within the next week. Super learner is a name for a modern implementation of ‘stacking’.<sup>44 45</sup> Informally, super learner is a machine learning algorithm that minimises discrepancies between predictions (where low discrepancies mean high predictive accuracy) and observed outcomes by forming a weighted average of candidate algorithms or candidate learners. The candidate learners are standard regression and machine learning algorithms (eg, generalised additive models and random forests). The algorithm uses cross-validation to target predictive accuracy in novel data, which yields algorithms that can be trained in one dataset and provide accurate predictions when deployed in new settings, such as predicting clinical outcomes following a patient visit.<sup>46</sup> This approach has been shown to be asymptotically as accurate as the best possible prediction algorithm that is tested and thus can serve as a gold standard for prediction purposes.<sup>43</sup> We will use an approach to fit the super learner, as described by van der Laan and colleagues in 2007<sup>43</sup> and simplified by Naimi and Balzer in 2018,<sup>47</sup> and adapted for SAS and cross-platform use by Keil *et al.*<sup>48 49</sup> Table 2 shows the different algorithms used in super learner.

We will assess super learner model performance using the 40% validation sample data on measures of: (1) calibration—observed versus predicted risks, also thought of as model fit or internal validity of the model,<sup>50 51</sup> and (2) discrimination<sup>50 51</sup>—the ability to distinguish patients with outcome from patients without, accounting for right-censored data. Discrimination also helps in determining the sensitivity of the model, that is the minimum and maximum predicted probabilities of the suicide attempt.<sup>51</sup> For calibration, we will plot the observed and predicted suicide attempts and deaths in the testing sample data, estimate the Hosmer-Lemeshow ‘goodness of fit’ statistic,<sup>52</sup> and evaluate the calibration slope and report 95% CIs.<sup>53</sup> For discrimination, we will calculate the time-dependent concordance or c-statistic,<sup>54</sup> which estimates the area under the receiver operating characteristics curve, analogous to Harrell’s c-index<sup>55</sup> or

the concordance probability estimate for right-censored data.<sup>56</sup> CIs for the calibration and concordance will be estimated based on bootstrapping with 200 replicates.

For external validation, the algorithms developed within the UNC EHR will be tested in private health insurance, Medicaid and correctional databases. Furthermore, we will examine potential algorithmic racial bias in our predictive algorithms for black and other non-white Americans (including Hispanic Americans) compared with white Americans to see if there are other factors that better predict these populations that may be related to systematic racism or represent historic disadvantage, for example, due to lack of access to healthcare, insurance, lower income, housing, schooling or other factors. This algorithmic racial bias examination will only be possible in UNC EHR, NC Medicaid and NC DOP data since race variable is not available in private health insurance data.

### Sample size

The NC death records and NCVDRS identify ~16 000 confirmed suicide deaths from 2006 to 2018, or >1200 suicide deaths per year. Once all data sources are combined through 2020, the number of suicide deaths in each target population will be ~1800 in the UNC EHR patient population from 2006 to 2020, ~3600 in private health insurance from 2006 to 2020, ~2400 in NC Medicaid from 2011 to 2020 and ~800 among DOP releases from 2000 to 2020. Further from 2010 to 2019, the UNC EHR had 5948 suicide attempts. At this rate, there will be many more suicide attempts in the privately insured population and NC Medicaid.

These data, along with 10 controls per case, will provide a sample size of >90 000 patients for predictive models that model suicide mortality risk, and a sample size of >270 000 patients for predictive models that model the risk of suicide attempts.

As our study aims mainly to focus on benchmarking incidence rate and prediction rather than hypothesis testing, formal considerations of power for hypothesis testing do not apply. Considerations of precision also do not apply, as incidence rate estimates from comprehensive surveillance systems are conventionally reported as observed values without CIs.

### Missing data

Missing data are often informative of predictions. For example, missing data on history of drug use often indicate that a patient does not have a history of drug use and is, therefore, at a lower risk of suicidal thoughts or action. Such missingness indicates that missingness itself is important to prediction and the missingness is likely not amenable to schemes that assume missingness at random, such as multiple imputation. While certain algorithms can accommodate missingness as a separate category (eg, tree-based methods like random forest can use missingness as a unique predictor value with which to classify individuals), super learner model includes other algorithms that cannot. To address this issue, we will create missingness indicator variables (1=missing, 0=not missing) for each predictor under consideration, and we will impute missing values for variables at their medians (continuous variables) or modes (categorical variables). The missingness indicator and the original variable will then both be considered predictors for super learner model.

### Planned analytic approach: objective 3

We will generate both an overall workflow for maintaining the linkage of all six data sources, and individual workflows tailored to each of the four entities (UNC EHR, private insurer,

NC Medicaid and NC DOP) detailing the linkage of that entity's data with the outcome data sources (NC death records and NCVDRS) and calculation of suicide death end points.

Each workflow will comprise the following elements: (1) the owners of each data source, (2) permissions required for each data source, with contact information and renewal frequency, (3) data dictionaries for all data sources defining linking fields and any fields needed for outcome definition or calculation of rates, (4) a data linkage workflow diagram indicating which data elements are transferred and where the linkage occurs, including any needed details about an honest broker and finder files, (5) specific logic for the linkage, including which fields are compared between each data source and how issues such as variation in name spellings are handled, (6) statistical code to complete the linkage, (7) definition of the calculation of suicide death rates, including definition of the numerator (codes to include for confirmed and possible suicide deaths) and the denominator (defining the population at risk for a given period of time) and (8) statistical code to compute the suicide death rate.

### Ethics and dissemination

This study is a secondary data analysis of existing records that involves no new data collection. We will link six data sources containing information on mortality, healthcare delivery, public and private health insurance, and correctional services. All these data sources are owned by different entities. We have completed data use agreements with each data entity and the study receives oversight from the institutional review board (IRB) at UNC, the data stewards at the NC DHHS and the NC DOP Research Review Committee. Individual-level linkage of all datasets will be completed by the UNC Sheps Centre serving as an 'honest broker' using an IRB-approved and HIPAA-compliant protocol.

Only aggregate data will be shared and published and numbers below five will be censored. All linked data will be stored either on the UNC secured research workspace or the Sheps Center secured servers.

The final super learner model will comprise a set of trained machine learning algorithms as well as a weight for each algorithm. The trained algorithms will be used to make predictions in a clinical setting, thereby providing tools for clinicians treating patients who may be likely to attempt suicide or die of a suicide. These algorithms can be combined with the 'shiny' package for R to create a graphical interface in which clinicians can enter patient characteristics predictions will be generated nearly instantly.<sup>51</sup> Once completed and validated, we will share the suicide (attempt and death) prediction tools with the stakeholders representing all four subpopulations so that they can be used to prevent suicide deaths. However, we recognise the potential for unintended consequences of such tools, including stigmatisation, increased treatment coercion and psychological harm to patients. Therefore, strategies for deployment will be carefully considered in consultation with our clinical tailoring advisory board, which will include stakeholders representing all target subpopulations, including patient advocates to identify mitigation strategies like clinician guidelines for use of the tool.

### DISCUSSION

Our study will be the first to develop an ongoing US-based suicide surveillance system that includes a large healthcare system, private and public insurer, and correctional cohorts along with statewide mortality and violent death reporting system data. Prior linkage studies for suicide research in the USA have been in specialised private healthcare management organisations or with

limited publicly insured data, which lack representativeness to the USA and exclude vulnerable populations.<sup>13 22–24 27 28</sup>

We will use an advanced ensemble machine learning algorithm to build prediction tool for clinicians that will help them to identify individuals with a high risk of self-harm behaviours in a clinic or in a one-on-one setting, thereby helping them to target potential interventions and prevent suicide deaths. Prior research tries to identify individual risk factors; however, risk, hazard or ORs of such factors are not particularly helpful for a clinician on a day-to-day basis in decision-making regarding mental health treatment referrals and targeting interventions.

We will use an incidence density-matched nested case-control study design to increase internal validity of our study and reduce computational time to run the machine learning algorithm with many short-term and long-term predictors of suicide deaths. While incidence density-matched nested case-control studies are well known, their application to study suicide outcomes in four linked large population-based cohorts makes it unique.

Lastly, we will streamline the linkage processes and develop documentation and protocols for addition of new data to allow ongoing surveillance and evaluation of suicide prevention initiatives in NC. As a demonstration of the capability of the surveillance system in evaluating ongoing interventions, we will evaluate the impact of the UNC hospitals' recent healthcare system policy change implementing routine suicide screenings on all inpatients. In July 2019, with the goal of reducing suicide attempts and mortality, the UNC Hospitals, part of the UNC healthcare system, implemented mandatory screening of all inpatients with the Columbia-Suicide Severity Risk Screenings (C-SSRS). Inpatients identified as at risk for suicide are further assessed with the Suicide Assessment Five-step Evaluation and Triage (SAFE-T). While the UNC Hospitals implemented this suicide screening and assessment policy in inpatient and emergency department patients, it was not implemented in the other 11 hospitals that are a part of the UNC healthcare system. Furthermore, our linkage with private health insurer data will allow us to compare privately insured patients at UNC Hospitals with uninsured or publicly insured patients in the state. The suicide surveillance system with multiple linked data sources will facilitate the use of multiple control groups to robustly estimate the effect of UNC Hospitals' C-SSRS and SAFE-T implementation on suicide attempts and mortality using controlled interrupted time series methods. This demonstration project will illustrate the utility of our surveillance system for evaluating other policy changes that may be implemented in any of our four target subpopulations. The data linkage among the four target subpopulations allows access to a more complete picture for the linked cases and controls. These integrated datasets and resulting predictive analyses will provide a framework for understanding how multiple risk and protective factors, that usually cannot be accessed from a single data system, predict self-harm and suicide-related patient outcomes.

**Contributors** All authors made significant contributions and reviewed and edited multiple drafts of the manuscript and provided approval for the final submitted version. In addition, SR and BWP conceptualised the study, secured funding, wrote the original draft of the study and this protocol, secured data use agreements and provide supervision to all aspects of the study. VM developed manuscript outline, edited the original draft of this protocol manuscript and helps with data collection and analysis. SD provides project and funding management and data collection support. TSC reviewed and edited the original draft of the protocol and supported data collection from electronic health records. BNG conceptualised the study, reviewed and edited the original study proposal and provides clinical and scientific expertise. APK provides machine learning expertise and wrote the original draft of machine learning methods for the study and this protocol. CVF conducts data analysis and linkage. DG provides clinical and scientific expertise and reviewed,

edited and wrote the original study protocol. ALK provides data linkage and curation expertise. MS-M conducts data cleaning, curation and analysis. TC provides project and funding management and data collection support. LP supports data collection and liaisons the work with North Carolina Department of Public Safety.

**Funding** This work is supported by funding from the National Institute of Health's National Institute of Mental Health (grant number: R01MH124752) (BWP and SR, Multiple Principle Investigators).

**Competing interests** None declared.

**Patient and public involvement** Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

**Patient consent for publication** Not applicable.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** Data may be obtained from a third party and are not publicly available. The data used in this study are not publicly available but can be obtained upon request from the entities noted under the 'Data sets and linkage' section in the Methods section of this manuscript.

#### ORCID iDs

Shabbir I Ranapurwala <http://orcid.org/0000-0002-3944-3912>

Catherine Vinita Fitch <http://orcid.org/0000-0002-4410-4608>

Lewis J Peiper <http://orcid.org/0000-0003-1871-2295>

#### REFERENCES

- Hedegaard H, Curtin SC, Warner M. Suicide mortality in the United States, 1999–2017. *NCHS Data Brief* 2018;330:1–8.
- North Carolina Vital Statistics 2017. American health rankings, 2018. Available: <https://www.americashealthrankings.org/explore/annual/measure/Suicide/state/NC>
- North Carolina violent death reporting system (NC-VDRS) data dashboard. suicide trends. Available: [https://dashboards.ncdhs.gov/t/DPH/views/NCVDRSDashboard/NC-VDRSDashboard?%3AshowAppBanner=false&%3Adisplay\\_count=n&%3AshowVizHome=n&%3Aorigin=viz\\_share\\_link&%3AisGuestRedirectFromVizportal=y&%3Aembed=y](https://dashboards.ncdhs.gov/t/DPH/views/NCVDRSDashboard/NC-VDRSDashboard?%3AshowAppBanner=false&%3Adisplay_count=n&%3AshowVizHome=n&%3Aorigin=viz_share_link&%3AisGuestRedirectFromVizportal=y&%3Aembed=y) [Accessed 10 Mar 2022].
- Hamza CA, Stewart SL, Willoughby T. Examining the link between nonsuicidal self-injury and suicidal behavior: a review of the literature and an integrated model. *Clin Psychol Rev* 2012;32:482–95.
- Chesney E, Goodwin GM, Fazel S. Risks of all-cause and suicide mortality in mental disorders: a meta-review. *World Psychiatry* 2014;13:153–60.
- Hawton K, van Heeringen K. Suicide. *Lancet* 2009;373:1372–81.
- Bachmann S. Epidemiology of suicide and the psychiatric perspective. *Int J Environ Res Public Health* 2018;15. doi:10.3390/ijerph15071425. [Epub ahead of print: 06 07 2018].
- Nock MK, Borges G, Bromet EJ, et al. Suicide and suicidal behavior. *Epidemiol Rev* 2008;30:133–54.
- Borges G, Angst J, Nock MK, et al. Risk factors for the incidence and persistence of suicide-related outcomes: a 10-year follow-up study using the National comorbidity surveys. *J Affect Disord* 2008;105:25–33.
- Ahmedani BK, Peterson EL, Hu Y, et al. Major physical health conditions and risk of suicide. *Am J Prev Med* 2017;53:308–15.
- Sivaraman JJC, Greene SB, Naumann RB, et al. Association between medical diagnoses and suicide in a Medicaid beneficiary population, North Carolina 2014–2017. *Epidemiology* 2022;33:237–45.
- Olfson M, Wall M, Wang S, et al. Suicide after deliberate self-harm in adolescents and young adults. *Pediatrics* 2018;141. doi:10.1542/peds.2017-3517. [Epub ahead of print: 19 03 2018].
- Olfson M, Wall M, Wang S, et al. Suicide following deliberate self-harm. *Am J Psychiatry* 2017;174:765–74.
- CDC. Suicide rising across the US. VitalSigns, 2018. Available: <https://www.cdc.gov/vitalsigns/suicide/index.html> [Accessed 1 Feb 2020].
- Fox KR, Huang X, Guzmán EM, et al. Interventions for suicide and self-injury: a meta-analysis of randomized controlled trials across nearly 50 years of research. *Psychol Bull* 2020;146:1117–45.
- Qin P, Nordentoft M. Suicide risk in relation to psychiatric hospitalization: evidence based on longitudinal registers. *Arch Gen Psychiatry* 2005;62:427–32.
- Ho T-P. The suicide risk of discharged psychiatric patients. *J Clin Psychiatry* 2003;64:702–7.
- Schou Pedersen H, Fenger-Grøn M, Bech BH, et al. Frequency of health care utilization in the year prior to completed suicide: a Danish nationwide matched comparative study. *PLoS One* 2019;14:e0214605.
- Hunt IM, Kapur N, Robinson J, et al. Suicide within 12 months of mental health service contact in different age and diagnostic groups: national clinical survey. *Br J Psychiatry* 2006;188:135–42.
- Meehan J, Kapur N, Hunt IM, et al. Suicide in mental health in-patients and within 3 months of discharge. National clinical survey. *Br J Psychiatry* 2006;188:129–34.
- Morrison KB, Laing L. Adults' use of health services in the year before death by suicide in Alberta. *Health Rep* 2011;22:15–22.
- Arias SA, Miller I, Camargo CA, et al. Factors associated with suicide outcomes 12 months after screening positive for suicide risk in the emergency department. *Psychiatr Serv* 2016;67:206–13.
- Bridge JA, Marcus SC, Olfson M. Outpatient care of young people after emergency treatment of deliberate self-harm. *J Am Acad Child Adolesc Psychiatry* 2012;51:213–22.
- Bridge JA, Olfson M, Fontanella CA, et al. Emergency department recognition of mental disorders and short-term risk of repeat self-harm among young people enrolled in Medicaid. *Suicide Life Threat Behav* 2018;48:652–60.
- Cerel J, Singleton MD, Brown MM, et al. Emergency department visits prior to suicide and homicide: linking statewide surveillance systems. *Crisis* 2016;37:5–12.
- Weis MA, Bradberry C, Carter LP, et al. An exploration of human services system contacts prior to suicide in South Carolina: an expansion of the South Carolina violent death reporting system. *Inj Prev* 2006;12 Suppl 2:ii17–21.
- Rossom RC, Simon GE, Beck A, et al. Facilitating action for suicide prevention by learning health care systems. *Psychiatr Serv* 2016;67:830–2.
- Ahmedani BK, Westphal J, Autio K, et al. Variation in patterns of health care before suicide: a population case-control study. *Prev Med* 2019;127:105796.
- Sheftall AH, Asti L, Horowitz LM, et al. Suicide in elementary school-aged children and early adolescents. *Pediatrics* 2016;138. doi:10.1542/peds.2016-0436. [Epub ahead of print: 19 09 2016].
- Roche AM, Giner L, Zalsman G. Suicide in early childhood: a brief review. *Int J Adolesc Med Health* 2005;17:221–4.
- Turvey CL, Conwell Y, Jones MP, et al. Risk factors for late-life suicide: a prospective, community-based study. *Am J Geriatr Psychiatry* 2002;10:398–406.
- Blaauw E, Kerkhof AJFM, Hayes LM. Demographic, criminal, and psychiatric factors related to inmate suicide. *Suicide Life Threat Behav* 2005;35:63–75.
- Rosen DL, Schoenbach VJ, Wohl DA. All-Cause and cause-specific mortality among men released from state prison, 1980–2005. *Am J Public Health* 2008;98:2278–84.
- Brinkley-Rubinstein L, Sivaraman J, Rosen DL, et al. Association of restrictive housing during incarceration with mortality after release. *JAMA Netw Open* 2019;2:e1912516.
- Bohnert KM, Ilgen MA, Louzon S, et al. Substance use disorders and the risk of suicide mortality among men and women in the US veterans health administration. *Addiction* 2017;112:1193–201.
- Copeland LA, Finley EP, Bollinger MJ, et al. Comorbidity correlates of death among new veterans of Iraq and Afghanistan deployment. *Med Care* 2016;54:1078–81.
- Zivin K, Yosef M, Miller EM, et al. Associations between depression and all-cause and cause-specific risk of death: a retrospective cohort study in the Veterans health administration. *J Psychosom Res* 2015;78:324–31.
- Conner KR, Bohnert AS, McCarthy JF, et al. Mental disorder comorbidity and suicide among 2.96 million men receiving care in the Veterans health administration health system. *J Abnorm Psychol* 2013;122:256–63.
- Conner KR, Bossarte RM, He H, et al. Posttraumatic stress disorder and suicide in 5.9 million individuals receiving care in the Veterans health administration health system. *J Affect Disord* 2014;166:1–5.
- Basham C, Denneson LM, Millet L, et al. Characteristics and Va health care utilization of U.S. veterans who completed suicide in Oregon between 2000 and 2005. *Suicide Life Threat Behav* 2011;41:287–96.
- Shepardson RL, Kosiba JD, Bernstein LI, et al. Suicide risk among veteran primary care patients with current anxiety symptoms. *Fam Pract* 2019;36:91–5.
- Jones JE, Hermann BP, Barry JJ, et al. Rates and risk factors for suicide, suicidal ideation, and suicide attempts in chronic epilepsy. *Epilepsy Behav* 2003;4 Suppl 3:S31–8.
- van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol* 2007;6:Article25.
- Wolpert DH. Stacked generalization. *Neural Networks* 1992;5:241–59.
- Breiman L. Stacked regressions. *Mach Learn* 1996;24:49–64.
- Rittenhouse KJ, Vwalika B, Keil A, et al. Improving preterm newborn identification in low-resource settings with machine learning. *PLoS One* 2019;14:e0198919.
- Naimi AI, Balzer LB. Stacked generalization: an introduction to super learning. *Eur J Epidemiol* 2018;33:459–64.
- Keil AP, Westreich D, Edwards JK. Super learning in the SAS system. Preprint submitted to computer methods and programs in biomedicine, 2019. Available: <https://arxiv.org/abs/1805.08058> [Accessed 8 Jul 2020].
- Keil AP. SuperLearnerMacro. Available: <https://cirl-unc.github.io/SuperLearnerMacro/> [Accessed 8 Jul 2020].
- Steyerberg EW, van der Ploeg T, Van Calster B. Risk prediction with machine learning and regression methods. *Biom J* 2014;56:601–6.
- Ranapurwala SI, Cavanaugh JE, Young T, et al. Public health application of predictive modeling: an example from farm vehicle crashes. *Inj Epidemiol* 2019;6:31.
- Hosmer DW, Hosmer T, Le Cessie S, et al. A comparison of goodness-of-fit tests for the logistic regression model. *Stat Med* 1997;16:965–80.
- Miller ME, Langefeld CD, Tierney WM, et al. Validation of probabilistic predictions. *Med Decis Making* 1993;13:49–58.

- 54 Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. *Biometrics* 2005;61:92–105.
- 55 Harrell F. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. New York: Springer Books, 2001.
- 56 Gönen M, Heller G. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika* 2005;92:965–70.