DEVELOPMENT AND IMPLEMENTATION OF A PREDICTIVE MODEL OF DOCUMENTATION STATUS TO EXAMINE FACTORS RELATED TO HEALTH OUTCOMES FOR UNDOCUMENTED IMMIGRANTS

By

Patrick Rivers

____

A Dissertation Submitted to the Faculty of the

MEL & ENID ZUCKERMAN COLLEGE OF PUBLIC HEALTH

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

HEALTH BEHAVIOR HEALTH PROMOTION

In the Graduate College

THE UNIVERSITY OF ARIZONA

2023

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by: Patrick Rivers

titled: Development and Implementation of a Predictive Model of Documentation Status to Examine Factors Related to Health Outcomes for Undocumented Immigrants

and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

David Marrero (Jul 26, 2023 18:01 PDT)
David Marrero — Date: Jul 26, 2023

Kristen Pogreba Brown — Date: Jul 26, 2023

David O. Garcia — Date: Jul 27, 2023

Scott Carvajal — Date: Jul 28, 2023

Xiaoxiao Sun — Date: Jul 27, 2023

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

David Marrero (Jul 26, 2023 18:01 PDT)
David Marrero — Date: Jul 26, 2023
Health Behavior and Health Promotion

ACKNOWLEDGEMENTS

I owe a tremendous debt of gratitude to my advisor, Dr. David Marrero: without your encouragement and support I would not have finished this manuscript. Deep thanks as well to the other members of my committee – Drs. Scott Carvajal, David Garcia, Kristen Pogreba-Brown, and Xiaoxiao Sun. Thank you for your thoughtful and engaged comments, and your patience.

Thank you to my parents for your unending encouragement in all of my academic endeavors, and family and friends – in Tucson and elsewhere – for all things big and small that made possible this endeavor.

Dedication

To all those excluded from the research process and accessing the services they need.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Background: The United States (US) is home to more than 8.6 million undocumented Hispanic immigrants. These individuals face significant barriers to accessing health services and are a subpopulation on whom relatively little research has been done. Undocumented Hispanics are faced with unique vulnerabilities, and research on them should be done thoughtfully and with care. One proposed method of investigating health outcomes for this subpopulation, and the factors that influence these outcomes, is utilizing predictive modeling for documentation status.

Objectives: This dissertation is composed of three studies that aim to: 1) Synthesize the literature to assess health outcomes of undocumented immigrants with at least one chronic condition; 2) Create and internally validate a predictive model of documentation status utilizing two data sets by comparing demographic and health characteristics of individuals alongside their known documentation status; and 3) Utilize the predictive model on a prospective cohort of essential workers – in which documentation status is unknown – to examine healthcare utilization for the treatment of SARS-CoV-2 infections by predicted undocumented status.

Methods: Aim one is explored through a scoping review of the literature related to chronic disease outcomes for undocumented Hispanic immigrants compared to documented immigrants or US-born citizens, following the Preferred Reporting Items for Systematic reviews and Meta-Analyses checklist for scoping reviews (PRISMA-ScR). A systematic search was conducted on October 10, 2019 in five databases (PubMed, MEDLINE, Web of Science, Embase, and CINAHL), with supplementary searches up to June 24, 2023. Studies were examined by two reviewers, and those found to be quantitative studies conducted in the US, comparing health

outcomes for undocumented Hispanic immigrants aged 18+ compared to either or both documented Hispanic immigrants or US-born citizens were included.

Aim two is examined by the development and internal validation of three multivariable prediction models of Hispanic immigrant documentation status utilizing data from two previous studies on health outcomes of individuals living near the Arizona-Mexico border. For the predictive models, multiple imputation by chained equations (MICE), random forest machine learning, and support vector machine learning algorithms were employed, with each model assessed for its accuracy, precision, recall, and Matthews correlation coefficient (MCC).

Aim three is investigated through the implementation of the best-performing prediction model in a longitudinal cohort of essential workers, and utilizing multivariable logistic regression models to ascertain if there were differences between predicted undocumented immigrants and the rest of the cohort, and to identify factors associated with healthcare utilization for the treatment of SARS-CoV-2 infection.

Results: Of 275 records identified through database search, 26 full-text articles were assessed, and eight articles were found to examine health outcomes for undocumented immigrants with at least one chronic condition. There was conflicting evidence as to whether undocumented immigrants have poorer health outcomes compared to documented immigrants or US-born citizens.

For aim two, a combined sample of 473 individuals was used to build our prediction models. The model using MICE had the lowest rates of accuracy (64.3%), precision (0.81), recall (0.74), and MCC (-0.07). The random forests model was fared slightly better in all metrics with an accuracy of 73.1%, precision of 0.82, recall of 0.86, and MCC of 0.002, and the support vector machine was the strongest in all categories, with an accuracy of 90.7%, precision of 0.90, recall of 0.99, and MCC of 0.66.

For aim three, predicted undocumented status of 1,074 Hispanic individuals from the longitudinal cohort was assigned using the support vector machine algorithm. Across 466 SARS-CoV-2 infections that occurred in this group, there was no statistically significant difference between healthcare utilization by predicted undocumented immigrants compared to the rest of the cohort, and the only factors associated with healthcare utilization were number of illness symptoms, number of days spent in bed, and illness duration.

Conclusions: There are significant research gaps on undocumented immigrants, their health outcomes, and the factors that influence these health outcomes. One of the predictive models for documentation status developed in Aim 2 shows promise and warrants additional research. This work aims to make some incremental progress towards understanding this subpopulation better, the unique challenges they face, and their healthcare behaviors and needs.

CHAPTER 1: INTRODUCTION

More foreign-born individuals reside in the United States (US) than any nation in the world.[1] A total of 43.7 million immigrants live in the country, with 11.1 million of them being undocumented – meaning they lack U.S. citizenship or lawful temporary or permanent residency.[2] Of these undocumented immigrants, 8.6 million, or two-thirds, are Hispanic.[3] Over the last decade, international migration has risen, despite efforts by many countries to tighten their borders.[4] In the US, strategies including increased border patrol, detention, and deportation have been used in attempts to restrict immigrants crossing the border.[5,6]

For individuals who do make it into the US, policies have been put in place to limit access to health services, e.g. excluding undocumented immigrants from accessing health insurance under the Affordable Care Act.[7] Undocumented immigrants face additional barriers to access health care services, including: experiences of discrimination, marginalization, and stigma; confusion about rules and regulations; and mistrust of the health care system.[8,9]

As a result, undocumented immigrants in the US are less likely to have a regular source medical care outside of the emergency room, and less frequently use health services.[10-12] This less frequent utilization of health services endangers not only their health, but also that of the general public, particularly in instances of infectious diseases.[13,14] While immigrants overall exhibit relatively favorable health,[15,16] this advantage seems to deteriorate as they gain more experience in the "host" country, and as they become more acculturated.[17,18]

Research on undocumented Hispanic individuals faces a number of unique challenges. First is a lack of data which includes immigration status in nationally representative survey data that is often used in secondary data analyses to examine health science questions.[19] Researchers may omit questions on documentation status due to the possibility of putting undocumented respondents at risk if identifying data were to be requested by government agencies.[20,21] Even in instances where researchers might wish to gather information on documentation status, Hispanic individuals overall are more likely to refuse to participate in research studies than members of other racial and ethnic groups,[22] or refuse to answer questions about documentation status even when researchers may believe they are following best practices for research.[23]

While Common Rule guidelines for human subjects research, set forth by the US Department of Health and Human Services, have not specially designated immigrant populations as "vulnerable", some have argued that they may fall into this designation for a variety of reasons, and researchers therefore owe a particular duty of care to this subpopulation. In addition to the health-related vulnerabilities outlined above, immigrants, particularly undocumented immigrants have high rates of poverty,[24] may have low levels of English proficiency,[25,26] and may put themselves or their loved ones in peril by answering questions about their documentation status,[27] particularly in the climate of the last decade where high rates of deportations and punitive state and local legislations have been common.[28] Extra precautions therefore, above and beyond the usual protections to confidentiality, should be taken when it comes to researching undocumented immigrants.

One way this has been implemented in other disciplines, particularly economics,[19,29] is utilizing predictive modeling for documentation status. By assigning predicted documentation status to a data set where true documentation status is not known – particularly when doing secondary data analysis on deidentified data sets – researchers can examine outcomes without putting undocumented individuals at additional risk.[29] This dissertation seeks to utilize the foundations for predictive modeling from other disciplines to build and validate a predictive model of documentation status using variables found in health research, and implement it in a data set where documentation status is not known in order to examine healthcare utilization by individuals in response to SARS-CoV-2 infection.

Theoretical Framework

This dissertation is guided by the Fundamental Cause Theory. This theory aims to explain why health inequalities exist under conditions that should eliminate or reduce them.[30] Namely, how do stark inequalities in overall health and mortality exist between individuals of different socioeconomic statuses during modern times when the major diseases and risk factors that seemingly accounted for differences in earlier periods of history (e.g. infections like typhoid, tuberculosis, and diphtheria caused in large part due to poor sanitation and overcrowding) have been largely eradicated in the modern developed world.[31]

According to the original authors, a fundamental social cause of health inequality has four features. First, it influences multiple disease outcomes. Second, it affects these outcomes through multiple risk factors. Third, it involves access to resources that can be used to avoid risks or minimize the consequences of disease once it occurs. And fourth, the association between a fundamental cause and health is reproduced over time.[30] The most enduring examples include socioeconomic status and racism. Some researchers have argued that documentation status should be considered a distinct fundamental cause of health, even though some other fundamental causes like the two examples listed above are deeply intertwined with it.[32-35] There are a number of reasons for this, including clear social and economic disadvantages that are directly due to an individual's undocumented status, like: limited occupational mobility and higher likelihood of being employed in sectors with higher occupational health risks;[36] less access to employment benefits;[37] limited access to healthcare and poorer experiences with care;[38] and being subject to restrictive immigration laws and law enforcement which generate fear and stress, and undermine the economic wellbeing of immigrants.[39]

Approach

This dissertation follows a three-paper model with three aims, each investigated as one paper, which are presented as chapters in the dissertation. Chapter 2 is a scoping review of the literature, synthesizing findings from studies examining chronic disease outcomes for undocumented Hispanic immigrants compared to documented immigrants or US-born citizens. The scoping review assessed each study's location, purpose, data source, sample characteristics, health outcomes, and limitations, with a discussion as to potential reasons the findings from these studies are mixed, and the need for future studies to address the research gap.

Chapter 3 explores building and validating predictive models of documentation status for Hispanic immigrants given a number of sociodemographic, health behavior, and disease status indicators. It lays out challenges and ethical considerations in conducting research on undocumented immigrants and seeks to build a robust predictive model that could assign predicted documentation status to individuals in cohorts where documentation status is not known. Three models were built and evaluated, under the belief that such a tool would allow for the examination of differences between predicted undocumented individuals and others in a cohort across a wide range of health outcomes.

As initially envisioned at the beginning of 2020, chapter 4 would have utilized a de-identified data set of patients treated at Banner University Medical Center – Tucson between November 1, 2013 to September 30, 2017 and assigned predicted documentation status to each individual. Then, the effect of predicted documentation status on hemoglobin A1c levels of patients at diagnosis of diabetes, and approximately one year after diagnosis would have been examined. However, the onset of the COVID-19 pandemic led to delays in the completion of this dissertation, and staffing shortages at the University of Arizona Clinical Data Warehouse, from which the patient data would have come, meant that the timeline for getting access to the required data was uncertain. In consultation with the dissertation committee, it was decided that an alternate data source should be utilized.

The replacement was data from the Arizona Healthcare Emergency Response and Other Essential Workers Surveillance Study (AZ HEROES), a longitudinal cohort study of over 4,200 essential workers in Arizona. The purpose of AZ HEROES was not related to chronic diseases, but rather describing the estimating incidence and symptom presentation of SARS-CoV-2 infection, establishing molecular and immunologic characteristics of SARS-CoV-2 infection, and examining post-vaccine immunologic response. Instead of focusing on clinical outcomes of SARS-CoV-2 infection, the decision was made to examine one of the factors most important for determining the outcome of a SARS-CoV-2 infection: healthcare utilization. Chapter 4, then, focuses on assigning predicted documentation status to participants in the AZ HEROES study using the strongest predictive model from chapter 3, and examining if there were differences in healthcare utilization between predicted undocumented immigrants and the rest of the cohort, as well as to identify other factors that might be associated with healthcare utilization for the treatment of SARS-CoV-2 infection.

SPECIFIC AIMS

Aim 1: Synthesize the literature to assess chronic disease outcomes for undocumented Hispanic immigrants compared to documented immigrants or US-born citizens.

Aim 2: Create and internally validate a predictive model of documentation status through two data sets comparing demographic and health characteristics of individuals alongside their known documentation status.

Aim 3: Utilize the predictive model on a prospective cohort of essential workers – in which documentation status is unknown – to examine healthcare utilization for the treatment of SARS-CoV-2 infections by predicted undocumented immigrants.

CHAPTER 2: IMMIGRATION STATUS AND CHRONIC DISEASE OUTCOMES – A SCOPING REVIEW

Abstract

Undocumented immigrants face significant barriers to accessing regular medical care in the US. This is a concern especially for individuals with chronic conditions. This scoping review provides an overview of the current knowledge on chronic health outcomes for undocumented immigrants compared to documented immigrants or US-born citizens. We have used a review methodology recommended for summarizing existing research evidence, and to identify knowledge gaps. The review process was conducted in accordance with the PRISMA-ScR guidelines, and the selection of studies was based on pre-defined criteria. Eight articles were included. Key information such as population details, and study details, outcomes, and limitations are presented. There was conflicting evidence as to whether undocumented immigrants have poorer health outcomes. Several possible explanations for this were described, but future, rigorous studies are recommended to address this gap and better understand the health of this vulnerable population.

Background

The US hosts more foreign-born individuals than any nation in the world.[1] Of the 43.7 million immigrants living in the country, 11.1 million are undocumented[2] Undocumented immigrants often face a number of barriers that impact their ability or willingness to access health care services, including: English language proficiency; restricted mobility; experiences of marginalization and stigma; and mistrust of the health care system.[8] In addition, undocumented immigrants are ineligible for U.S. medical insurance and health care coverage options,[40] and are less likely to have a primary care provider,[41] a regular source of medical care,[42] or access to and use of health services.[38] Undocumented immigrants are also often unwilling to seek out medical care due to fear of deportation.[43,44] Moreover, although immigrants and, to a lesser extent, US.-born Hispanics in the US exhibit relatively favorable health,[15,16] this advantage seems to deteriorate as immigrants gain more experience in the "host" country, and as they become more acculturated.[17,18]

All of these challenges are particularly concerning for individuals with a chronic health condition, which requires ongoing management and supervision from both the individual and their health care team to achieve optimal outcomes and reduce the risk of developing a wide array of serious and costly health complications.[45] Given that undocumented status has been shown to reduce healthcare utilization in Hispanic individuals with chronic conditions,[46] it is likely that undocumented individuals do not receive the same level of medical care to manage their condition as documented immigrants or U.S. citizens. This could lead to poorer disease management and health outcomes which can have significant social and economic consequences.[47]

Rationale and Objective

While previous research has uncovered a number of barriers to care for undocumented immigrants, the health burden of chronic diseases on undocumented immigrants in the US remains almost entirely unexplored. No review of any kind has been published that explores and summarizes evidence on chronic disease health outcomes of undocumented immigrants compared to documented immigrants and U.S.-born citizens. Due to the paucity of research on health outcomes of undocumented immigrants with chronic diseases we conducted a scoping review of studies that have examined chronic disease outcomes for undocumented Hispanic immigrants compared to documented immigrants or US-born citizens.

Methods

The current study was guided by the PRISMA-ScR guidelines as well as methods outlined by the Joanna Briggs Institute for scoping reviews, and includes the following stages: 1) identifying the research questions, 2) identifying relevant studies, 3) study selection, 4) extracting and charting the data, and 5) collating, summarizing and reporting the results.[48-50]

Search Strategy

A systematic literature search was conducted on October 10, 2019 in PubMed, MEDLINE, Web of Science, Embase, and CINAHL, with supplementary searches up to June 24, 2023. No restrictions were placed on article publication date, though only those published in English were considered. The following Medical Subject Headings (MeSH) and keywords were used in the search: ("emigrants and immigrants"[mesh] OR migrant OR immigrant OR emigrant OR immigration OR "foreign-born") AND ("legal status" OR undocumented OR documented OR authorized OR unauthorized OR illegal) AND ("diabetes mellitus"[mesh] OR diabetes OR A1C OR chronic illness OR chronic disease OR chronic condition) AND ("United States" OR "US" OR "U.S.").

References in included papers were searched for additional articles that should be added.

Eligibility Criteria

Following best practices recommendations for scoping reviews,[51] we utilized the PCC strategy for our eligibility criteria, outlined in **Table 1**.

Table 1. Scoping review inclusion and exclusion criteria

PCC Strategy	Inclusion criteria	Exclusion criteria
P – Population	Undocumented immigrants, 18+ years old, living in the U.S., focus on Hispanic individuals (at least 80% of study participants)	Studies focused on children and/or adolescents (<18 years of age)
C - Concept	Quantitative studies with individuals with at least one chronic disease (type 2 diabetes, heart disease, obesity, or chronic kidney disease) reporting outcomes compared to either or both documented immigrants/U.S.-born citizens	Descriptive articles such as reviews, lectures, conference proceedings, or qualitative-only studies.
C - Context	Studies conducted in the United States	Studies not published in the English language

Study Selection and Data Extraction

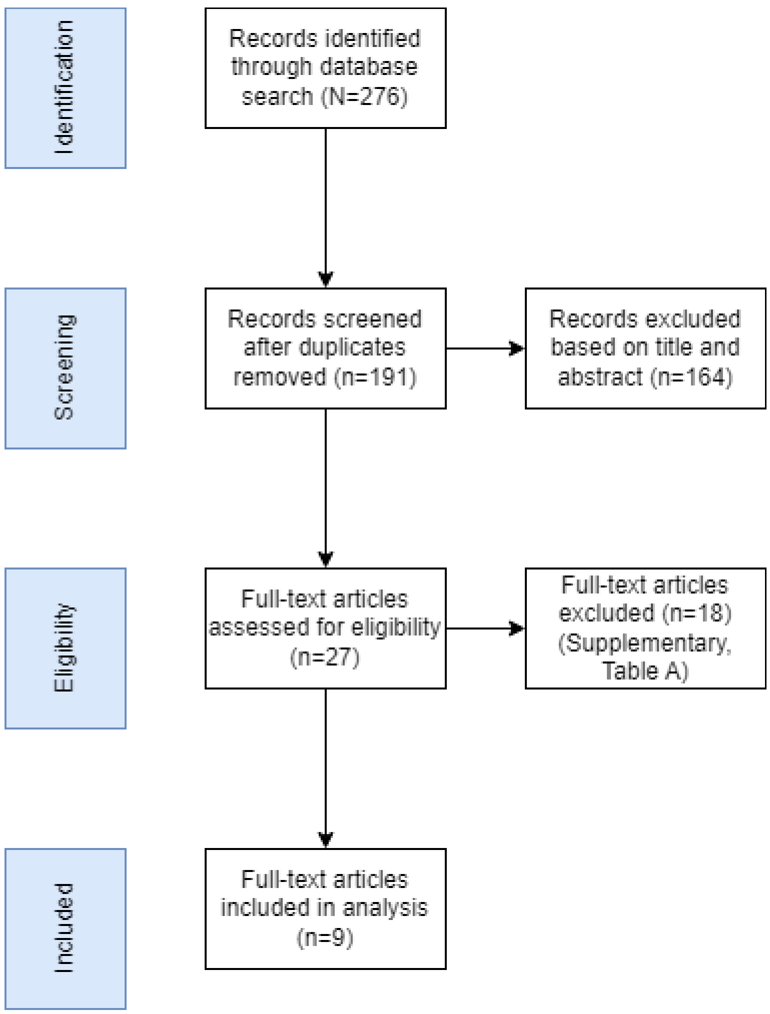
Two researchers independently screened titles and abstracts, and a third checked for discrepancies and resolved any conflicts. This process was repeated for full text screening with each author fulfilling the same role. Reasons for exclusion during full text screening are detailed in **Table 5**. A Research Electronic Data Capture (REDCap) database was developed *a priori* specifically for data extraction for this project. Two researchers independently extracted the following data: title; first author; year of publication; geographic location; sample size; age range; gender breakdown; aim of the study; country(ies) of origin of undocumented study participants; method of determining documentation status; comparator; information about primary and secondary outcomes (list of outcomes, results and significance); major limitations; and references to obtain. Data extraction reports were checked by a third researcher for discrepancies with conflicts resolved by all three authors. In accordance with the scoping review methodology, a formal assessment of the quality of individual articles was not included, though quality was addressed in the discussion. No articles were excluded based on methodological quality.

Results

Search Results, Inclusion, and Exclusion

Figure 1 presents a flow chart of the search results and selection of eligible publications. The literature search generated a total of 276 articles, 191 after the removal of duplicates. A total of 164 articles were removed after title and abstract screening if there was enough information to demonstrate that they did not meet the inclusion criteria or that they fulfilled one or more of the exclusion criteria. This left 27 full-text articles to be assessed, of which 9 met the eligibility criteria and were included in this review.

Figure 1. PRISMA flow diagram of literature search and selection for scoping review



Study Characteristics

Study characteristics were gathered, including first author, year of publication, geographic location, aims/purpose, and data source (**Table 2**). All nine studies were published after 2000, with one published in 2004 and the remaining seven published 2011-2022. Five studies examined outcomes for patients with End Stage Renal Disease, one on farmworkers, and only three that might be considered the “general population.” Four studies were centered on a specific hospital or group of local hospitals, four utilized national survey data, and one used survey data with information from two different states. A total of 59,103 individuals were included in analysis across the nine studies.

Table 2. Study characteristics of studies included in scoping review

Authors	Title	Year	Geographic Location	Aims/Purpose	Data Source
Coritsidis et al.[52]	The Initiation of Dialysis in Undocumented Aliens: The Impact on a Public Hospital System	2004	New York City, NY	Compare health outcomes of undocumented individuals with ESRD to American citizens	Medical records from Nassau University Medical Center, New York NY; and Jacobi Medical Center, New York NY, 1993-2001, with follow-up interview surveys administered by physicians
Do et al.[46]	Healthcare Utilization Among Hispanic Immigrants with Diabetes: Investigating the Effect of US Documentation Status	2014	US (national survey)	1) To determine if documented and undocumented Hispanic immigrants with diabetes differ by health insurance status and having a usual place of care, 2) To investigate the effect that documentation status has on health care utilization among diabetic Hispanic immigrants.	Pew Hispanic Center and Robert Johnson Wood Foundation's 2007 Hispanic Healthcare Survey
Hamilton et al.[53]	Immigrant Legal Status and Health: Legal Status Disparities in Chronic Conditions and Musculoskeletal Pain Among Mexican-Born Farm Workers in the United States	2019	US (national survey)	To determine whether a legal status disparity exists in chronic conditions and musculoskeletal pain among Mexican-born immigrants employed as farm workers in the US	Farmworkers interviewed as part of the National Agricultural Workers Survey, 2000-2015
Iten et al.[54]	Undocumented Immigration Status and Diabetes Care Among Mexican Immigrants In Two Immigration "Sanctuary" Areas	2014	San Francisco, CA; Chicago, IL	To investigate the relationship between immigration status and the patient experience of health care, diabetes self-management, and clinical outcomes among Mexican immigrants with diabetes receiving health care in two immigration sanctuary cities	Immigration, Culture and Health Care study with respondents from San Francisco, CA and Chicago, IL, 2008-2009
Linden et al.[55]	Kidney Transplantation in Undocumented Immigrants With ESRD: A Policy Whose Time Has Come?	2012	New York City, NY	To examine the demographic and health characteristics of undocumented immigrants with ESRD, as well as transplant attitudes and options	Medical records from Elmhurst Hospital, New York NY, 2010

Madden et al.[56]	Dialysis Hospitalization Inequities by Hispanic Ethnicity and Immigration Status	2017	Albuquerque, New Mexico	To examine the likelihood of being admitted to an emergency room for dialysis among undocumented immigrants, compared with other patients	Medical records from patients admitted via the emergency department at the University of New Mexico Hospital. 2013-2016
Raghavan et al.[57]	Descriptive Analysis of Undocumented Residents with ESRD in a Public Hospital System	2011	Houston, TX	To evaluate the demographics of the patient population receiving hemodialysis in two Houston public hospitals and an affiliated clinic	Medical records of patients receiving hemodialysis in the Harris County Health district, 2010
Ruhnke et al.[58]	A healthy migrant effect? Estimating health outcomes of the undocumented immigrant population in the United States using machine learning	2021	US (national survey)	To evaluate whether an immigrant health advantage persists among undocumented immigrants in the US	Pooled cross-sections of the National Health Interview Survey, 2000-2018
Shen et al.[59]	Association of Citizenship Status With Kidney Transplantation in Medicaid Patients	2018	US (national survey)	To assess if undocumented immigrants would have similar outcomes to US citizens following kidney transplant if equally insured.	Medical records of adult Medicaid patients in the US Renal Data System who received a kidney transplant, 1990-2011

Study Population Characteristics

Sample sizes of the various studies ranged from 127 to 434,985 individuals (**Table 3**). Studies with smaller numbers of participants, particularly undocumented individuals, tended to be those where data was collected from an individual hospital system. The larger samples generally came from studies conducting secondary data analysis of national survey data. One study did not report any information on the ages of participants. For the remaining eight studies, the mean age of participants in the studies ranged from mid-30s to mid-50s. Two studies did not break out ages of undocumented individuals versus the comparison group, but for the six that did, undocumented individuals were approximately ten years younger than the comparison group in four of them. Sex of participants was about two-thirds male (four studies), relatively equal between male and female (four studies), or not disclosed (one study). The comparison group was both documented immigrants and US-born citizens in six studies, and only documented immigrants in the remaining three studies. Finally, four studies used self-report questions to determine undocumented status, two studies used administrative data (lack of social security number), one counted use of an interpreter as a proxy for undocumented status, one used imputation methods, and one did not disclose the method of ascertaining documentation status.

Table 3. Participant characteristics of studies included in scoping review

Authors	Title	Sample	Age Range or Mean Age	Gender Breakdown	Comparator	Country(ies) of Origin	Measurement of legal status
Coritsidis et al.[52]	The Initiation of Dialysis in Undocumented Aliens: The Impact on a Public Hospital System	Undocumented immigrant patients (n=55) compared to American citizens (n=223), both with ESRD	Undocumented immigrants = 45.9 years; American citizens = 54.3 years	Undocumented immigrants = 62% male; American citizens = 65% male	Documented Immigrants and US-born Citizens	Undocumented immigrants were primarily of Hispanic origin (58%), mostly Central American from El Salvador	Not disclosed in article
Do et al.[46]	Healthcare Utilization Among Hispanic Immigrants with Diabetes: Investigating the Effect of US Documentation Status	Hispanic immigrants 18 years of age or older, undocumented (n=105), documented (n=422)	51.7 years	Male = 49.9%, female = 50.1%	Documented Immigrants	Mexico = 75.6%, Other = 24.4%	Self-report
Hamilton et al.[53]	Immigrant Legal Status and Health: Legal Status Disparities in Chronic Conditions and Musculoskeletal Pain Among Mexican-Born Farm Workers in the United States	Mexican-born respondents under the age of 65 (n = 17,462, breakdown by immigration status not disclosed)	Undocumented immigrants = 31.8 years, documented immigrants = 42.2 years	Undocumented immigrants = 21.4% female; documented immigrants = 27.0% female	Documented Immigrants	Mexico	Self-report
Iten et al.[54]	Undocumented Immigration Status and Diabetes Care Among Mexican Immigrants In Two Immigration "Sanctuary" Areas	Individuals with diabetes who received care in nine free-standing or hospital-based safety-net clinics: undocumented individuals (n=111) compared to the rest of the patient population (n=300)	Undocumented immigrants = 49.8 years; documented immigrants = 55.7 years; US born = 55.7 years	Undocumented immigrants = 51.6% female; documented immigrants = 50.0% female; US born = 54.0% female	Documented Immigrants and US-born Citizens	Mexico	Self-report by elimination

Linden et al.[55]	Kidney Transplantation in Undocumented Immigrants With ESRD: A Policy Whose Time Has Come?	All outpatient dialysis patients from March until May 2010 undocumented individuals (n=45) compared to the rest of the patient population (n=82)	43.8 years	71% male	Documented Immigrants	Mexico and Central America = 58% South America = 29% Asia = 13%	No social security number
Madden et al.[56]	Dialysis Hospitalization Inequities by Hispanic Ethnicity and Immigration Status	Undocumented immigrant patients (n=759) admitted via the emergency department and receiving dialysis compared to random sample of patients (n=3,345) admitted via the emergency department and not receiving dialysis	Not reported	Not reported	Documented Immigrants and US-born Citizens	Not reported	Use of an interpreter
Raghavan et al.[57]	Descriptive Analysis of Undocumented Residents with ESRD in a Public Hospital System	Undocumented immigrants receiving hemodialysis in Houston (total n=186, surveyed n=48) compared to all Hispanics on hemodialysis in Texas (n=13,283)	Undocumented immigrants surveyed = 43.4 years; undocumented immigrants in Houston = 42.6 years; Hispanic individuals in Texas = 58.9 years	Undocumented immigrants surveyed = 35.0% female; undocumented immigrants in Houston = 37.5% female; Hispanic individuals in Texas = 48% female	Documented Immigrants and US-born Citizens	Not reported	No social security number
Ruhnke et al.[58]	A healthy migrant effect? Estimating health outcomes of the undocumented immigrant population in the United States using machine learning	434,985 individuals responding to the National Health Interview Survey, 2000-2018, of which 12,368 were undocumented	Undocumented immigrants = 34.2% 35-49 years, documented immigrants = 40.4% 35-49 years, US-born = 34.1% 35-49 years	Undocumented immigrants = 47.6% female; documented immigrants = 52.0% female; US born = 51.2% female	Documented Immigrants and US-born Citizens	Reported only by region (Asia = 19%, Africa = 3%, Europe = 6%, Mexico & Central America = 61.2%, Middle East = 2%, South America = 7%)	Multi-stage imputation using data from the Survey of Income and Program Participation
Shen et al.[59]	Association of Citizenship Status With Kidney Transplantation in Medicaid Patients	Nonresident aliens (n=346) compared to permanent residents (n=1,489) and US citizens (n=8,660)	Undocumented immigrants = 34.5 years; documented immigrants = 46.0 years; US citizens = 42.0 years	Undocumented immigrants = 50.4% male; documented immigrants = 56.1% male; US citizens = 50.4% male	Documented Immigrants and US-born Citizens	Not reported	Self-report

Study Findings

The nine studies presented twenty-one different physical health outcome comparisons between undocumented individuals and documented immigrants or US citizens. On nine of these measures (43%), undocumented individuals fared worse than their comparison group (**Table 4**). On nine outcomes (43%), undocumented individuals fared better than the comparison group. On the final three outcomes (14%), there was no difference found between undocumented individuals and the comparison group. In other words, in over half of the studied outcomes, researchers found no difference for undocumented individuals compared to documented immigrants or US citizens, or better health outcomes for undocumented individuals.

Table 4. Findings and limitations of studies included in scoping review

Authors	Title	Primary Health outcome(s)	0 = no difference, + = undocumented better, - = undocumented worse	Primary Health Outcome Details	Major Limitations
Coritsidis et al.[52]	The Initiation of Dialysis in Undocumented Aliens: The Impact on a Public Hospital System	Mean Arterial Pressure Creatinine levels Calculated glomerular filtration rates Magnesium Alkaline phosphatase Albumin	- - - - - -	Undocumented immigrants vs. American citizens for all results: MAP (mm Hg): 119.9 vs. 108.9, p=0.001 Creatinine (mg/dL): 11.90 vs. 9.92, p=0.03 GFR (mL/min/1.73 m2): 5.53 vs. 6.29, p=0.03 Magnesium (mg/dL): 2.43 vs. 2.21, p=0.008 Alkaline phosphatase: 136 vs. 111, p=0.03 Albumin (g/dL): 3.26 vs. 3.0, p=0.04	• Limited generalizability • Cross-sectional data
Do et al.[46]	Healthcare Utilization Among Hispanic Immigrants with Diabetes: Investigating the Effect of US Documentation Status	Time since last visit Self-reported excellent or good health status	- -	Undocumented immigrants vs. documented immigrants for all results: Time since last visit 6 months or less: 12.4% vs. 69.2% Excellent or good health status: AOR = 0.3, 95% CI 0.18, 0.80	• Cross-sectional study data • Single item measure used to assess health care utilization. May be incomplete. • Low response rate: potential sampling bias

Hamilton et al.[53]	Immigrant Legal Status and Health: Legal Status Disparities in Chronic Conditions and Musculoskeletal Pain Among Mexican-Born Farm Workers in the United States	Lifetime, diagnosed chronic conditions Musculoskeletal pain in the last year	+ +	Lifetime diagnosis of at least one chronic condition: 6.4 % of unauthorized immigrants vs. 16.8 % authorized immigrants Musculoskeletal pain: 15% of unauthorized immigrants vs. 21% of authorized immigrants	• Unauthorized immigrants have less access to health care, less opportunity to have chronic condition diagnosed. • Information is collected on the worksite, unauthorized immigrants may have greater fear of retaliation and less likely to report a health problem • Limited generalizability
Iten et al.[54]	Undocumented Immigration Status and Diabetes Care Among Mexican Immigrants In Two Immigration "Sanctuary" Areas	Hemoglobin A1C Systolic blood pressure Low-density lipoprotein	0 0 0	Undocumented immigrants vs. documented immigrants vs. US-born citizens for all results: Hemoglobin A1C (%): 8.45 vs. 7.98 vs. 8.7 Systolic blood pressure (mmHg): 126.07 vs. 130.2 vs. 130.46 Low-density lipoprotein (mg/dL): 88.29 vs. 89.70 vs. 94.28	• Limited generalizability • Cross-sectional data • Small sample size (and low power for some analyses)
Linden et al.[55]	Kidney Transplantation in Undocumented Immigrants With ESRD: A Policy Whose Time Has Come?	Coronary heart disease Diabetes	+ +	Undocumented immigrants vs. documented immigrants for all results: Coronary heart disease: 7% vs. 33% Diabetes mellitus: 40% vs. 68%	• Limited generalizability • Younger undocumented population
Madden et al.[56]	Dialysis Hospitalization Inequities by Hispanic Ethnicity and Immigration Status	Likelihood of receiving emergent vs. scheduled dialysis during first emergency department visit	-	AOR = 3.49, 95% CI 1.03, 11.81	• Interpreter use as proxy for immigration status (if provider is bilingual, may not need interpreter) • Length of stay imprecise - Limited generalizability
Raghavan et al.[57]	Descriptive Analysis of Undocumented Residents with ESRD in a Public Hospital System	Diabetes	+	Undocumented immigrants vs. US-born citizens for all results: Diabetes: 35% vs. 68%	• Limited generalizability • Cross-sectional data • Younger undocumented population
Ruhnke et al.[58]	A healthy migrant effect? Estimating health outcomes of the undocumented immigrant population in the United States using machine learning	Fair/poor self-rated health Any chronic conditions Obesity	+ + +	Undocumented immigrants had 26% lower odds of reporting poor or fair health, 23% lower odds of having been diagnosed with any chronic conditions and 8% lower odds of being obese	• Cannot be verified that individuals assigned "undocumented" truly are, leading to possible mis-specification. • Cross-sectional data
Shen et al.[59]	Association of Citizenship Status With Kidney Transplantation in Medicaid Patients	All-cause transplant failure	+	Undocumented immigrants compared to US-born citizens for all results: All-cause transplant failure: Adjusted HR 0.56, 95% CI 0.41, 0.76	• Limited generalizability • Cross-sectional data • Younger undocumented population

The nine measures on which undocumented individuals fared worse than documented immigrants or US citizens came from three studies. Coritsidis et al. found that undocumented individuals had: higher creatinine levels, a measure of kidney function with higher levels meaning more impaired kidney function (11.9mg/dL vs. 9.9mg/dL, p=0.03); higher blood pressure (mean arterial pressure of 119.9mmHg vs. 108.9mmHg, p=0.001); lower glomerular filtration rates, a test of kidney function with lower levels meaning worse filtration by the kidneys (5.5mL/min/1.73m^2 vs. 6.3mL/min/1.73m^2); higher magnesium levels (2.43mg/dL vs. 2.21mg/dL, p=0.008); higher alkaline phosphatase levels (136 vs. 111, p=0.03); and higher albumin levels (3.26g/dL vs. 3.0g/dL, p=0.04). It should be noted, however, that none of the levels for either group in the last three outcomes (magnesium, alkaline phosphatase, and albumin) are high enough to be of any real medical concern. Do et al. found that undocumented immigrants with diabetes had worse healthcare utilization (12.4% utilizing health services in the last six months vs. 69.2%) and less likelihood of self-reporting a health status of excellent or good health (AOR = 0.3, 95% CI 0.18, 0.80). Finally, Madden et al. found undocumented individuals at much higher odds of being placed on emergent vs. scheduled dialysis upon first admittance to the emergency department, meaning they were in particularly poor condition upon arrival (AOR = 3.49, 95% CI 1.03, 11.81).

The three measures on which there was no difference for undocumented individuals compared to documented individuals or US-born citizens came from one study. Iten et al. found no statistical difference for undocumented individuals in: hemoglobin A1c, a measure of blood sugar (8.4% compared to 8.7% in U.S.-born citizens, p=0.41, and 7.98% in documented immigrants, p=0.06); systolic blood pressure (126.1mmHg compared to 130.5 in U.S.-born citizens, p=0.07, and 130.2 in documented immigrants, p=0.06); and low density lipoprotein cholesterol (88.3mg/dL compared to 94.3mg/dL in U.S.-born citizens, p=0.2, and 89.7mg/dL in documented immigrants, p=0.8).

Finally, the nine measures on which undocumented immigrants fared better than documented immigrants were found in three studies. Hamilton et al. found better outcomes for undocumented immigrants in the percent of individuals who have had a diagnosed chronic condition in their lifetime (6.4% compared to 16.8%, p<0.001), and prevalence of individuals reporting musculoskeletal pain in the last year (15.1% compared to 21.2%, p<0.001). Linden et al. found lower rates, in undocumented individuals, of coronary heart disease (7% compared to 33%) and diabetes mellitus (40% compared to 68%). Raghavan et al. also found lower rates of diabetes in undocumented individuals (35% compared to 68%). Ruhnke et al. found that Undocumented immigrants had 26% lower odds of reporting poor or fair health (p < 0.001), 23% lower odds of having been diagnosed with any chronic conditions (p < 0.001), and 8% lower odds of being obese (p < 0.01). Finally, Shen et al. found lower rates of all-cause transplant failure for undocumented individuals (adjusted HR 0.56, 95% CI 0.41, 0.76).

Discussion

In this first scoping review on the chronic health outcomes of undocumented immigrants compared to documented immigrants or U.S.-born citizens, we found relatively few articles that have studied this question. The increasing number in the last several years is somewhat encouraging, even if the overall number remains low.

Evidence for differential health outcomes based on immigration status was mixed, which was in contrast to the relatively clear background literature on barriers to undocumented individuals' accessing care, and lack of insurance and regular source of care. Three studies, comprising 43% of the 21 health outcomes examined, found evidence of poorer health outcomes for undocumented individuals. Five studies, examining nine outcomes, found evidence of better health outcomes for undocumented individuals. The final study, examining three outcomes, found no difference. There are several possibilities why the studies included in this review do not align with background research findings.

First are the methodological limitations of the nine studies. Eight of the studies used proxies or indirect methods of identifying documentation status that might not be entirely reliable, while one study did not disclose their method of identifying undocumented immigrants. Most studies also relied on small samples, especially the undocumented immigrant population – with the associated analytic problems of small sample sizes and potential bias they carry. Of particular concern was that of representativeness or generalizability. Five of the studies were focused on individuals with End Stage Renal Disease, and one on farmworkers. Of the final three that might be considered from the general public, one of them was a convenience sample from two "sanctuary cities", places that have statutes limiting the reporting of the documentation status of individuals seeking health services. It is possible that individuals included in most of these studies differ, in meaningful ways, from the larger population and their outcomes are not generalizable to most other undocumented individuals. Undocumented immigrants were also, on average, approximately ten years younger than the comparator group in the six studies that broke out age characteristics for the comparison groups, which seems a likely contributor to a healthier state.

A second potential reason that the studies did not find stronger support for poorer health outcomes in undocumented immigrants compared to documented immigrants or US citizens is linked to the idea of immigrant health selectivity. This refers to how immigrants tend to differ from non-migrants because of the challenges associated with migration, and how immigrants, even if undocumented, might have, on average, better health than the average US-born citizen, or comparable health characteristics as documented immigrants to which they are compared in individual studies. This is because the costs of migration select on good health as well as on other characteristics that might be associated with good health, such as self-efficacy.[60] Individuals who migrate tend to have differences in resources, networks, and opportunities than those who do not migrate or are unable to. In essence, individuals who migrate might be healthier than those who do not. As with the methodological limitations detailed above, this again might lead to a skewed sample that may not be generalizable to a larger population, which is an issue when only one study (Hamilton et al.) controls in any way for possible positive health selectivity.

Finally, only two of the nine studies account for length of time spent in the US since immigration. Undocumented immigrants have spent less time, on average, than documented immigrants,[61] let alone US citizens. As acculturation and longer time in the US is associated with worse health outcomes – likely as a result of stress, cumulative disadvantage, and adoption of unhealthy behaviors common in the country of residence[62] – it is possible that the undocumented immigrants in several of the studies are more recently arrived, and are therefore likely healthier compared to an earlier-arrived, documented immigrant population or US citizen.

Strengths and Limitations of this Review

The present scoping review provides an overview of the published knowledge of health outcomes for undocumented immigrants with chronic conditions compared to documented immigrants or US-born citizens. The main strength of this review is its methodological approach, which utilized the systematic framework recommended in the PRISMA checklist for scoping reviews (PRISMA-ScR).[50] This methodology is useful for topics such as this with relatively few studies and to identify gaps in knowledge.

Although a thorough search for literature in a number of databases as well as hand-searches was conducted, there is a possibility that all studies relevant for the review scope were not identified. We did not contact the authors of any studies to provide answers to questions we may have had in the review process. We also decided to limit our inclusion to quantitative studies, as qualitative studies on the subject seemed to be patient experiences and limited to one, or less than five, individuals per study. This might be a limitation to the number and breadth of results, but this decision was made to ensure that our review was not a review of case studies, which requires a separate methodology.[63]

Conclusions

Undocumented individuals face significant barriers to accessing the same level of health care for their chronic conditions compared to documented immigrants and US-born citizens. It is unclear from the published literature if this translates to poorer health outcomes. There are few studies that have addressed this question, however, and there has been little structure about what health outcomes to assess, what populations to include, and what methodology should be used. Given the paucity of research on the topic, and the numerous methodological limitations of the studies included in this article, further research is warranted. This seems particularly true when weighed against the high human cost of chronic diseases, and the more than ten million undocumented immigrants in the US.

Table 5. Excluded papers read in full-text, and reason for exclusion from scoping review

First Author, Year	Title	Reason for Exclusion
Balakrishnan, 2019[64]	The Undocumented Elderly: Coverage Gaps and Low Health Care Use	Commentary/Narrative
Bohan, 2005[65]	Illegal immigrants with ESRD: The impact on U.S. and Canadian health care facilities	Full text not available
Cantu, 2013[66]	New estimates of racial/ethnic differences in life expectancy with chronic morbidity and functional loss: evidence from the National Health Interview Survey	Wrong patient population
Carrillo, 2011[67]	A Regional Health Collaborative Formed By New York-Presbyterian Aims To Improve The Health Of A Largely Hispanic Community	Wrong patient population
Cervantes, 2019[68]	The Status of Provision of Standard Outpatient Dialysis for US Undocumented Immigrants with ESKD	Commentary/Narrative
Cervantes, 2018[69]	Circumstances of Death among Undocumented Immigrants Who Rely on Emergency-Only Hemodialysis	Wrong comparator
Cervantes, 2018[70]	Association of Emergency-Only vs Standard Hemodialysis With Mortality and Health Care Use Among Undocumented Immigrants With End-stage Renal Disease	Wrong comparator
De Jesus Diaz-Perez, 2004[71]	A program to improve access to health care among Mexican immigrants in rural Colorado	Wrong patient population
Garcini, 2018[72]	Health-related quality of life among Mexican-origin Latinos: the role of immigration legal status	Wrong outcomes

Hogan, 2017[73]	Emergent Dialysis and its Impact on Quality of Life in Undocumented Patients with End-Stage Renal Disease	Wrong study design
Hsueh, 2018[74]	Associations between immigrant status and pharmacological treatments for diabetes in U.S. adults	Wrong patient population
Kang, 2015[75]	Perceptions of affordability of care and access to information about health insurance among immigrants after Massachusetts health reform	Full text not available
Lee, 2018[76]	The prevalence and impact of diabetes mellitus among undocumented immigrants in an indigent care program in Riverside, California	Full text not available
Ortega, 2015[77]	When politics trumps health: Undocumented Latino immigrants and US health care	Commentary/Narrative
Peota, 2017[78]	Care in the shadows: The challenges for undocumented immigrants	Commentary/Narrative
Ross, 2019[79]	Association between immigration status and anxiety, depression, and use of anxiolytic and antidepressant medications in the Hispanic Community Health Study/Study of Latinos	Wrong concept
Sher, 2017[80]	Healthcare outcomes in undocumented immigrants undergoing two emergency dialysis approaches	Wrong comparator
Thomsen, 2015[81]	Caring for the Uninsured, Undocumented Population: Challenges of Chronic Disease Management and Access to Care	Commentary/Narrative

CHAPTER 3: DEVELOPMENT AND INTERNAL VALIDATION OF THREE MULTIVARIABLE PREDICTION MODELS OF HISPANIC IMMIGRANT DOCUMENTATION STATUS UTILIZING TWO SAMPLES IN ARIZONA

Abstract

Research on undocumented populations is sparse, both because documentation status is rarely ascertained in studies, and because undocumented individuals are less likely overall to participate in research studies. A better understanding of the health of these individuals is important, though, and one method that has been proposed to examine outcomes for this group without putting them at additional risk is through predictive modelling.

Utilizing combined data from two population-based samples in Arizona in which documentation status were collected, three prediction models of documentation status were created and internally validated. Models created used multiple imputation by chained equations (MICE), and two machine learning algorithms: random forests machine learning and support vector machine. Predicted documentation status was modelled from demographic, health behaviors, and disease status indicators commonly collected in public health research studies. The performance of each model was assessed in terms of accuracy, precision, recall, and Matthews correlation coefficient.

The combined sample consisted of 473 individuals, of whom 85 were undocumented. The sample included 209 men and 263 women, aged 20 to 83 years of age at the time of their participation in the respective studies. The MICE model had the lowest rates of accuracy (64.3%), precision (0.81), recall (0.74), and MCC (-0.07). The random forests model was slightly better with an accuracy of 73.1%, precision of 0.82, recall of 0.86, and MCC of 0.002. The

support vector machine was the strongest in all categories, with an accuracy of 90.7%, precision of 0.90, recall of 0.99, and MCC of 0.66.

Utilizing predictive models could protect the safety of vulnerable individuals as well as make possible investigation into potentially undocumented individuals in settings where it would be impractical or impossible to gather information on documentation status first-hand. The results of the current analysis demonstrate some promise and further investigation, particularly the support vector machine model.

Background

Of the 11.3 million undocumented immigrants in the US, more than 8.6 million (76%) are Hispanic.[82] Undocumented immigrants are often forced to forego medical care due to lack of insurance,[7] experiences of discrimination or mistrust of the health care system,[8,9] or fear of deportation,[43] and receive more infrequent and worse care than documented immigrants or U.S. citizens.[10-12] Fears related to their immigration status also makes undocumented individuals less likely to participate in research studies,[29] making them a challenging group to study directly.

Research on undocumented populations has been hindered by the lack of data on immigrant legal status in the types of large-scale, nationally-representative survey data generally used to answer many health and economic questions.[19] One reason that studies omit questions on documentation status is the possibility of putting undocumented respondents at risk if they respond honestly, and data including direct or indirect identifiers are somehow released to or requested by government agencies.[20,21] For this reason, proxy measures of immigration status have been recommended for use in research studies attempting to better understand immigrant health. Preliminary research has shown that predictive modeling is an effective approach to address these concerns without putting undocumented individuals at additional risk.[29]

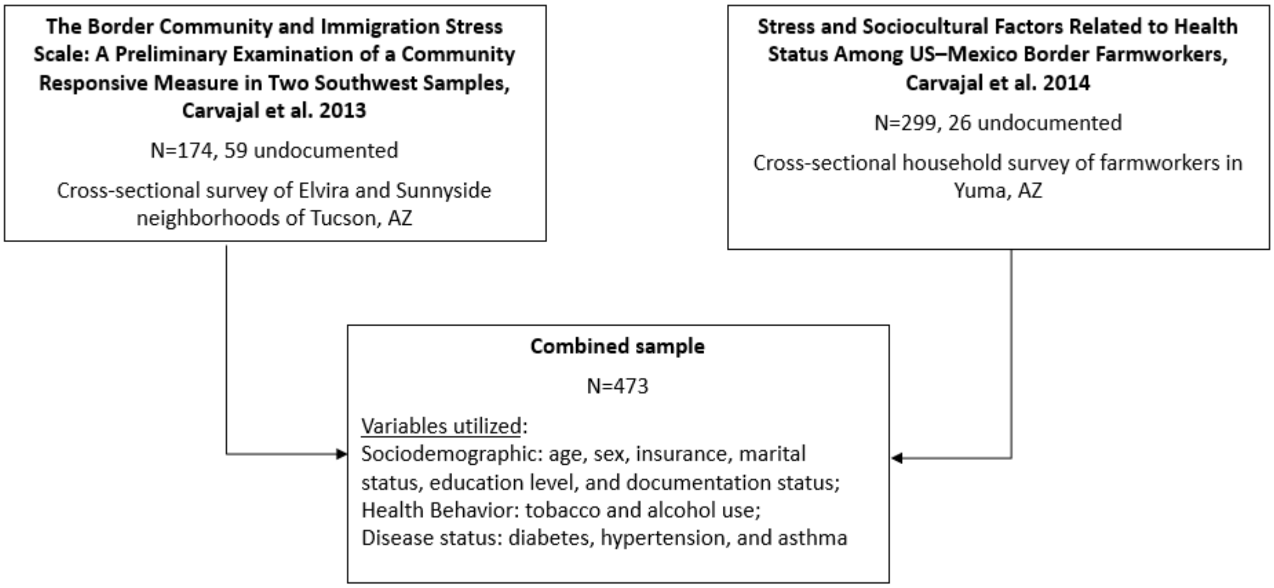
Our approach utilized a methodology developed for previous predictive models of immigrant documentation status that have been used to answer economic questions (multiple imputation by chained equations),[19,29] and two machine learning methods (random forest and support vector machine). The setting of the study groups used for this prediction model (Southwestern US) positions it well for better understanding Hispanic undocumented immigrants. The objectives of this study are to develop and internally validate three predictive models of documentation status for Hispanic immigrants in the Southwestern US.

Methods

Source of Data

Data for this project comes from two population-based samples from 2013 to 2014 in Tucson and Yuma, AZ, and includes 473 total individuals.[83,84] Participants were recruited from two majority Hispanic neighborhoods (Tucson study)[12] or farmworker communities (Yuma study).[11] Participants in the Tucson study were selected from community residents, with those in the Yuma study drawn from randomly selected households within specific census blocks from low-income communities. A range of demographic, health behaviors, and disease status indicators were collected in each cross-sectional study examining the effects of stress on mental and physical health outcomes in border communities (**Figure 2**). In both studies, participants self-reported their documentation status. All participants from the Tucson and Yuma studies were used in the creation of the prediction models.

Figure 2. Study sample pooling for prediction models



Measures

Individual documentation status, a binary variable of documented or undocumented, was the outcome of our prediction mode. While documentation status was known for the entire sample, it was randomly assigned as “missing” for 50 percent of the sample, with the known values of documentation status for this 50 percent forming the basis of comparison between the predicted values and the observed values to gauge the models’ performance.

Predictors taken from each study and used in the prediction models included demographic, health behavior, and disease status indicators. Age was recorded as a continuous variable and was measured in years. Categorical variables included were: sex (male or female); insurance status (insured or uninsured); overall health (excellent, very good, good, fair, or poor); marital status (single, married, or other); education level (never attended, completed grade 8 or less, completed grades 9-11, high school diploma or equivalent, completed some college, and college graduate); tobacco use (current smoker, former smoker, never smoker); as well as previous diagnosis (yes/no) of each of the following conditions: hypertension, asthma, and diabetes.

All predictors were self-reported by Tucson and Yuma study participants during interviews with trained study staff.

Statistical Analysis

Three prediction models were created and evaluated to determine the best approach for identifying documentation status. In the first model, documentation status was predicted using combined sample multiple imputation (CSMI), a method utilized in economics literature that aims to impute the immigration status of noncitizens in surveys lacking questions about immigration status.[29,85-87] Documentation status and all predictors listed above from each study were combined into one sample. Documentation status was randomly assigned as “missing” for 50 percent of the sample. Multiple imputation by chained equations (MICE) was then used to predict documentation status for the 50 percent of the sample for which it was “missing”, utilizing all predictors listed above. MICE specifies a suitable conditional imputation model for each incomplete variable and iteratively imputes until convergence.[88] It has shown large efficiency gains and substantial reduction in omitted variable bias compared to other imputation techniques and has been the preferred method of imputing the immigration status of noncitizens into surveys lacking questions on immigration status.[86,87]

The next two models utilized machine learning algorithms. In both, the data was split in two, with one half acting as the training data set to develop the algorithms, and the other half acting as the testing data set to refine the algorithms’ performance. In the first of these algorithms, we utilized random forests, which is a non-linear model aggregating the results of a series of tree structure classifiers, or decision trees, in which each tree depends on the values of a random vector sampled independently and with the same distribution for all trees.[89,90] Random forests generally demonstrate high levels of accuracy, manage outliers and noise well, and are resistant to overfitting. In this model, we used 500 trees, with the node-size set to default, and 20 variables tried at each split.[91] The final model, and the second machine learning algorithm, was a support vector machine, which is a highly robust prediction method in which input vectors are non-linearly mapped to a high dimension linear decision surface.[92] It demonstrates high generalization and flexibility, and is also resistant to over-fitting.[93]

Performance

All models were assessed in terms of accuracy, precision, recall, and Matthews correlation coefficient. Accuracy is the rate at which the classifier is correct, or the level of agreement between predictions from the model and observed outcomes, in this case the predicted documentation status.[94] Accuracy is defined as the sum of the number of true positives and true negatives, divided by the total number of observations, and ranges from 0 percent to 100 percent with higher numbers indicating better accuracy. While it can trend towards overoptimism on the ability of the models to classify the majority class in an unbalanced data set (which ours was, with the number of undocumented individuals smaller than the number of documented individuals), it was included here, following traditional reporting methods.

Recall, also known as the true positive rate or sensitivity, corresponds to the number of true positives divided by the sum of true positives and false negatives (in this case undocumented status). Values range from 0 to 1, with higher values demonstrating better recall. Precision is the number of true positives divided by the sum of true positives and false positives (also undocumented status). Similar to recall, values range from 0 to 1, with higher values demonstrating better precision.

Matthews correlation coefficient is a special case of the phi coefficient.[95,96] It is a more reliable statistical rate than the F_1 score, which is commonly used, and only produces a high score with all of the following: high proportions of true positives and true negatives, and low proportions of false negatives and false positives.[96] It is resistant to the overoptimism problems of accuracy and is generally viewed as a means of overcoming class imbalance issues.[19] It ranges from -1 (complete misclassification) to 1 (perfect classification).

All analyses were completed in R (version 4.1.2).[97]

Results

Baseline characteristics

The combined sample consisted of 473 individuals, of whom 85 were undocumented. The sample included 209 men and 263 women, aged 20 to 83 years of age at the time of their participation in the respective studies. Participants in the Tucson study had a higher proportion of undocumented individuals (33.9% to 8.7%, $p < 0.001$), and females (67.8% to 48.5%, $p < 0.001$) compared to individuals in the Yuma study. Tucson participants were also more likely to be single (21.3% to 11.8%, $p < 0.001$) were more educated ($p < 0.001$), reported higher good or better overall health (75.9% to 62.2%, p < 0.001) and a larger proportion that had never smoked (70.6% to 21.4%, p < 0.001) compared to individuals in the Yuma study. Rates between the two studies were similar for age, insurance status, and previous diagnosis of each of diabetes, hypertension, and asthma (**Table 6**).

Table 6. Descriptive statistics, stratified by study group, of key study characteristics of participants in two Southwest studies examining the effect of stress on mental and physical health outcomes

Characteristic	**Carvajal, et al. 2013 (Tucson, AZ) N = 174**	**Carvajal, et al. 2014 (Yuma, AZ) N = 299**	**p-value**
Undocumented, n (%)	59 (33.9)	26 (8.7)	< 0.001
Female, n (%)	118 (67.8)	145 (48.5)	< 0.001
Age, mean (SD)	46.2 (15.0)	44.3 (11.1)	0.06
Marital Status			
Single, n (%)	37 (21.3)	35 (11.8)	< 0.001
Married, n (%)	108 (62.1)	239 (80.5)	
Other, n (%)	29 (16.7)	23 (7.7)	
Education Level			< 0.001
Never attended, n (%)	3 (1.8)	3 (1.0)	
Grades 1-8, n (%)	21 (12.5)	246 (82.3)	
Grades 9-11, n (%)	46 (27.4)	47 (15.7)	
High school diploma or GED, n (%)	46 (27.4)	3 (1.0)	
Some college, n (%)	36 (21.4)	0 (0)	
College degree or above, n (%)	16 (9.5)	0 (0)	
Insurance			0.07
Yes, n (%)	106 (62.0)	160 (53.5)	
No, n (%)	65 (38.0)	139 (46.5)	
Overall health			< 0.001
Excellent	19 (11.2)	14 (4.7)	
Very good	40 (23.5)	36 (12.0)	
Good	70 (41.2)	136 (45.5)	
Fair	31 (18.2)	109 (36.5)	
Poor	10 (5.8)	4 (1.3)	
Tobacco use			< 0.001
Never smoker, n (%)	120 (70.6)	64 (21.4)	
Former smoker, n (%)	23 (13.5)	45 (15.1)	

Current smoker, n (%)	27 (15.9)	190 (63.6)	
Previous diagnosis of:			
Diabetes, n (%)	32 (18.8)	48 (16.1)	0.51
Hypertension, n (%)	40 (24.1)	64 (21.4)	0.69
Asthma, n (%)	16 (9.3)	16 (5.4)	0.11

Accuracy, Precision, Recall, and Matthews Correlation Coefficient

Table 7 details the accuracy, precision, recall, and Matthews correlation coefficient (MCC) of the three models. The model using MICE had the lowest rates of accuracy (64.3%), precision (0.81), recall (0.74), and MCC (-0.07). The random forests model was slightly better in all metrics with an accuracy of 73.1%, precision of 0.82, recall of 0.86, and MCC of 0.002. The support vector machine was the strongest in all categories, with an accuracy of 90.7%, precision of 0.90, recall of 0.99, and MCC of 0.66. Both the MICE and random forests models had relatively high false positive rates (19.5 and 13.9%, respectively) and very high false negative rates (86.8% and 85.9%, respectively), which led to very low MCCs. A similar rate of false positives (19.4%) combined with a very low false negative rate (0.5%) led to a much higher MCC for the support vector machine model.

Table 7. Internal validation evaluation metrics for three models predicting undocumented status from two Southwest studies examining the effect of stress on mental and physical health outcomes

	Accuracy	Precision	Recall	Matthews Correlation Coefficient
MICE[a]	64.3%	0.81	0.74	-0.07
Random Forests	73.1%	0.82	0.86	0.002
Support Vector Machine	90.7%	0.9	0.99	0.66

a. MICE = Multiple Imputation by Chained Equations

Discussion

The predictive models utilizing MICE and random forests fared poorly in all evaluation metrics. The support vector machine model, though, exhibited high levels of accuracy, precision, recall, and MCC, showing a strong association between predicted and actual documentation status for this group of Hispanic individuals. The high MCC in particular demonstrates the strength of the support vector machine model. While the MICE and random forests models do not appear to produce results worth further investigation, at least within this population, the support vector machine model demonstrated enough promise to justify its use in a new population.

These models were developed to aid in further research into Hispanic undocumented populations. Given the sensitive nature of documentation status, it may not be feasible or advisable for researchers to gather that information regularly or directly from participants. While there is a burgeoning literature on this population and the development of predictive models of documentation status within other disciplines, it remains almost entirely undeveloped within health sciences. With the challenges of collecting documentation status, the ability to utilize a proxy created via predictive modeling is useful.

Because of their vulnerable status, better understanding this population is important, especially if it is possible to do so via means that do not put individuals at direct risk or require identifying information. As these predictive models are built utilizing basic demographic, health behavior, and health outcome variables readily available in population and community research projects, they would be widely adaptable and could be utilized to help researchers gain a better understanding of the health status or health outcomes of a range of diseases experienced by potentially undocumented Hispanic individuals.

Strengths and Limitations

The current prediction model complements and builds upon established prediction models for undocumented status used in economics and demonstrates the feasibility of incorporating health behavior and health outcome variables in models related to documentation status. While the model used in economics (MICE) fared relatively poorly, the usage of machine learning, particularly support vector machine algorithms showed significant promise. One benefit of the current study is the assessment of the model's performance via precision, recall, and Matthews correlation coefficient, which is not something typically done with CSMI within economics literature as documentation status in one of the samples is always unknown. Finally, the current analysis builds from predictive modeling related to documentations status from a different discipline and found that while MICE may be useful and accurate in economics, the integration of health behavior and health status variables in addition to ones typically used previously appears to have had deleterious effects on its applicability to health data.

This study is subject to several limitations. First, each model uses split-sample validation, in which the data set was divided into two, and half of it was used to create the prediction model, and the other half to validate it. This has efficiency losses because it does not use all available data for the model development.[98]. Secondly, these models were created using samples from southern Arizona. These specific models, and the predictors used to create it, might have limited generalizability to different populations, and future researchers would likely need to validate it for their own populations. Finally, these models were not able to be externally validated, and would benefit from further validation with an unrelated data set in which documentation status and all predictor variables are known.

The support vector machine algorithm in particular shows promise and seems a fruitful area of further research. While the rate of false positives remained relatively high, the effect of this in a real-world analysis may have minimal impact on the overall results. Because documented individuals are, relatively, healthier than undocumented individuals, incorrectly assigning an undocumented status to an individual who is not undocumented will likely only attenuate the effect of the object of study. This could lead to higher p-values and make it more difficult to determine statistical significance, but attenuating an effect towards the null is preferable to moving it away from the null in the opposite direction, which could happen with false negatives. The extremely low rate of false negatives, therefore, is encouraging, and overall support vector machine algorithms appear to warrant further investigation.

Conclusions

The expansion of predictive modeling of documentation status into health sciences presents a unique method for understanding the profile and potential needs of undocumented individuals. Utilizing predictive models could protect the safety of vulnerable individuals as well as make possible investigation into potentially undocumented individuals in settings where it would be impractical or impossible to gather information on documentation status first-hand. The results of the current analysis demonstrate some promise and further investigation, and application, of these models appear warranted.

CHAPTER 4: HEALTHCARE UTILIZATION FOR THE TREATMENT OF SARS-COV-2 INFECTIONS BY PREDICTED UNDOCUMENTED IMMIGRANTS IN A COHORT OF ESSENTIAL WORKERS IN ARIZONA

Abstract

Background: Demands on health systems due to COVID-19 are substantial, and undocumented immigrants face particular vulnerabilities in the pandemic that warrant examination. Drivers of healthcare utilization are not well defined in non-severe SARS-CoV-2 infections. Among a prospective cohort of frontline workers from July 2020 to February 2023, we examined differences in healthcare utilization during SARS-CoV-2 infection for predicted undocumented immigrants as well as overall factors associated with utilization.

Methods: Predicted documentation status was assigned by a support vector machine model. Weekly specimens tested via real-time reverse transcriptase polymerase chain reaction analysis. Participants reported sociodemographic, health status information, and illness experience information. Primary outcome was healthcare utilization during SARS-CoV-2 infection. Predictors included sociodemographic characteristics, baseline health status, and measures of illness severity. Multivariable logistic regression was utilized to generate odds ratios for predictors of healthcare utilization.

Results: Between July 2020 and February 2023, a total of 466 SARS-CoV-2 infections occurred in 1,074 Hispanic AZ HEROES participants, 51 (10.9%) of which were in participants predicted to be undocumented: 292 (62.7%) individuals were between 40-65 years old; 263 (56.4%) were female; 115 (24.7%) individuals sought medical care for treatment of their infection.

In adjusted and unadjusted models, predicted documentation status was not associated with whether a participant sought medical care for their infection. Number of days spent in bed (OR 1.39, 95% CI 1.06 – 1.97) and illness duration (OR 1.02, 95% CI 1.01 – 1.07) were positively associated with healthcare utilization during the Wuhan/Alpha predominance. During the Omicron predominance, number of illness symptoms (OR 1.17, 95% CI 1.04 – 1.33) and illness duration (OR 1.01, 95% CI 1.00 – 1.03) were positively associated with healthcare utilization.

Conclusion: Findings of the current analysis that differential utilization of healthcare services by documentation status were not present in our cohort and should be explored in additional cohorts. Of factors associated with healthcare utilization for SARS-CoV-2 infection (symptom severity and duration), practices and therapeutics aimed at decreasing these factors would be most helpful in easing the burden on health systems.

Background:

The COVID-19 pandemic has placed an unprecedented demand on the US health care system and continues to be a substantial burden at the system and individual level. Since August 2020, 6.15 million people in the US were admitted to hospitals for COVID-19.[99] While the risk of death from the Omicron (B.1.1.529) variant is lower than previous variants, the overall number of cases, emergency department visits, and hospital admissions is higher.[100,101] Healthcare utilization and COVID-19 outcomes are not uniform, and a better understanding of individuals who may require care following SARS-CoV-2 infection is important,[102] particularly among vulnerable groups.

An estimated 8.6 million undocumented immigrants do not have health insurance in the US,[103,104] and are therefore more likely to not have primary care providers and rely on emergency departments for their care.[105] Early reports during the pandemic indicated that many undocumented individuals were unable or unwilling to seek care for SARS-CoV-2 infection due to lack of health insurance.[106] Differences by age, race, ethnicity, gender, underlying medical conditions, and insurance status have also all been associated with rates of delaying medical care and overall healthcare utilization once medical care is sought.[107-109] It is unclear, however, if these findings have persisted throughout the pandemic, or in settings where longitudinal follow-up is possible.

Some limited studies found that SARS-CoV-2 testing rates were lower, and rates of infection and severe morbidity due to SARS-CoV-2 infection were both higher among foreign-born individuals compared to native US citizens.[110-112] Building on previous literature related to other diseases,[113-116] it has been posited that limited access to healthcare services is one of, if not the most important factors in lower health outcomes, and these this might be exacerbated during the COVID-19 pandemic.[117] However, research testing this hypothesized relationship has not been conducted. Given the lack of published studies, the urgency of better understanding the COVID-19 pandemic and factors that influence outcomes from SARS-CoV-2 infection – especially for a group as numerous and vulnerable as undocumented immigrants – implementing the support vector machine predictive model from the previous chapter in a data set of a COVID-19 longitudinal cohort study is warranted.

The study chosen was the Arizona Healthcare Emergency Response and Other Essential Workers Surveillance (AZ HEROES) Study – a longitudinal cohort study of over 4,200 essential workers in Arizona.[118] The primary objective of this analysis was to examine differences in healthcare utilization for predicted undocumented Hispanic essential workers who experienced SARS-CoV-2 infection in this cohort. In an exploratory analysis, we examined predictors of healthcare utilization and differences by predicted documentation status, sociodemographic characteristics, health status, and illness experience.

Methods

Study design and population

The AZ HEROES study began in July 2020 as a longitudinal, prospective cohort funded by the Centers for Disease Control and Prevention (CDC), recruiting over 4200 essential workers in Arizona.[118] Data collected through April 5, 2023, were included in this analysis. Enrollment groups included healthcare personnel, first responders, and other essential workers (e.g., individuals working in education, retail, food service, hospitality, infrastructure, manufacturing, utility, and transportation) across the state of Arizona who worked a minimum of 20 hours per week and had regular close-contact exposure to coworkers or the public. All study participants provided written informed consent, and the study protocol was reviewed and approved by the University of Arizona IRB and the CDC.

Data Collection

Upon enrollment, participants reported sociodemographic information (including gender, age, race/ethnicity, education, and occupation), health information (including medical conditions, height and weight, tobacco use, and daily medications use), and SARS-CoV-2 infection history. Each week, participants completed text message surveys and reported onset of COVID-like symptoms – including fever, chills, cough, shortness of breath, sore throat, diarrhea, muscle aches, or a change in smell or taste.

Participants also submitted self-collected mid-turbinate nasal swabs each week of study participation. If participants experienced an onset of illness with COVID-like symptoms outside their regular swab collection, they completed an additional illness swab. Both weekly and illness swabs were tested by Marshfield Clinical Research Laboratories (Marshfield, WI, USA) via quantitative reverse-transcriptase-polymerase-chain-reaction (RT-PCR) assays to detect the presence of SARS-CoV-2.[119] Whole-genome sequencing, to determine variant of infection, was completed on specimens positive for SARS-CoV-2 with a cycle threshold value <30 (the number of cycles required for the fluorescent signal to cross the threshold or exceed background level). For specimens positive for SARS-CoV-2 but ineligible for sequencing – those with a cycle threshold ≥30 – variant was assigned using the date of infection and which variant accounted for more than 50% of infections according to CDC data.[120]

Participants who reported symptoms during weekly surveillance, submitted an illness swab, or submitted any sample that was positive for SARS-CoV-2 were prompted to complete an acute illness survey, an illness update survey (if the illness lasts longer than one week), and a final illness survey once symptoms resolved. The illness surveys gathered information on illness symptoms, severity (e.g., number of hours of work missed and number of days spent at least half in bed), medical utilization, and illness duration.

Beginning in December 2020, participants were prompted to report uptake of the COVID-19 vaccines. Participants received a survey at least every eight weeks until reporting a completed two-dose series. Beginning in September 2021, participants who had completed their initial series at least eight weeks prior began receiving a survey to report COVID-19 boosters. Participants received the survey at least every eight weeks through April 2023 to be able to report additional booster shots, with participants completing their initial series during this time also asked about boosters. Vaccination was verified by participant-provided vaccination cards and the Arizona State Immunization Information System, as available.

Predictors

Participants who did not test positive for SARS-CoV-2 during the study period were excluded from analysis, as were non-Hispanic participants as the support vector machine prediction model was validated only for a cohort of Hispanic individuals. The primary outcome was healthcare utilization during the course of a SARS-CoV-2 infection. Healthcare utilization was a binary variable defined as accessing medical care from an emergency department, hospital, outpatient clinic, telehealth, or other medical care setting, or not – as self-reported by participants. Healthcare utilization was derived using answers to all surveys completed by participants during their illness.

The SARS-CoV-2 infection was defined as the period between onset of symptoms or positive SARS-CoV-2 test result (whichever came first) until the cessation of symptoms and a self-reported health of 90% normal health or better. The first SARS-CoV-2 infection was defined as those from participants who entered the study reporting no prior SARS-CoV-2 infection, and the first in-study positive. SARS-CoV-2 reinfections included all in-study positive infections for a participant who entered the study having had a prior SARS-CoV-2 infection, or all in-study positive infections after the first for participants entering the study having no prior SARS-CoV-2 infection. Infections were deemed reinfections if at least 90 days had elapsed since the date of last positivity for any variant, or at least 45 days if the infections were two different variants.

The main independent variable of interest was predicted undocumented status: a binary variable of predicted undocumented or predicted documented/US-born assigned to each participant utilizing a machine learning algorithm known as support vector machine.[121] Using the support vector machine model from chapter 3, predicted documentation status was assigned using predictors including gender, insurance status, overall health, marital status, education level, tobacco use, and previous diagnosis of each of hypertension, asthma, and type 2 diabetes.

Predictors utilized in the support vector machine model were not used again in regression models analyzing healthcare utilization to avoid potential overfitting. Additional independent variables included in regression models included whether the infection was the participant's first or a reinfection, vaccination status at time of infection, SARS-CoV-2 variant of infection (presumed or confirmed), health characteristics (number of chronic conditions, body mass index (BMI), and number of daily medications), and illness characteristics (number of symptoms, work missed, days spent in bed, and illness duration). These variables were chosen a priori based on previous literature[104-106] or their believed theoretical association with healthcare utilization.

Statistical analysis

Independent variables listed previously were stratified by healthcare utilization and predicted documentation status. Pearson's chi-squared tests and Welch two sample t-test (an unequal variance t-test) were used to determine unadjusted differences between healthcare utilization groups with statistical significance based on p-values < 0.05. For the main analysis, multivariable logistic regression models with robust standard errors[122] were used to ascertain if there were differences between individuals predicted to be undocumented compared to the rest of the sample, and to identify factors associated with healthcare utilization. Multivariable models were stratified by variant to account for differences in the timing of variant predominance and adjusted for all independent variables of interest. Statistical significance for the regression models was based on 95% confidence intervals (95% CI). All statistical analyses were completed using R, version 4.1.2 (2021, R Foundation for Statistical Computing, Vienna, Austria).

Results

Participant characteristics

Between July 2020 and April 2023, a total of 466 SARS-CoV-2 infections occurred in 1,074 Hispanic AZ HEROES participants, 51 (10.9%) of which were in participants predicted to be undocumented (**Table 8**). Over 78% (n = 364) of infections were from the Omicron variant, and it was the participant's first infection in nearly two-thirds of cases (n = 291, 62.4%). In over one-third (n = 184, 39.5%) of the infections, the participant was booster-vaccinated, approximately one quarter were unvaccinated or vaccinated with their initial series only (n = 131, 28.1% and n = 133, 28.5% respectively), and only 4% (n = 18) were partially vaccinated (received only one dose of a non-Johnson & Johnson vaccine). Sixty-three percent (n = 292) were 40-65 years of age and 56% (n = 263) were female. Approximately half of the cohort did not report taking any daily medications (n = 229, 49.1%).

Across all SARS-CoV-2 infections, the average illness duration was 22.9 days (SD 22.1), and participants missed an average of 46.4 hours of work (SD 45.5) and spent 2.8 days in bed (SD 3.6), with nearly half (n = 202, 43.3%) experiencing 6-10 illness symptoms.

Factors associated with healthcare utilization

Of all participants experiencing a SARS-CoV-2 infection during the study, 115 (24.7%) reported seeking medical care, and 351 (75.3%) reported they did not (**Table 8**). In unadjusted analyses, participants who did or did not seek medical care were not statistically significantly different based on their predicted documentation status by any characteristic.

Table 8. Characteristics of 466 SARS-CoV-2 infections during enrollment in AZ HEROES stratified by medical utilization and predicted documentation status

Characteristic	Total (N = 466)		Did not seek medical care (n = 351, 75.3%)					Sought medical care (n = 115, 24.7%)				
			Predicted non-undocumented (n = 314, 89.5%)		Predicted undocumented (n = 37, 10.5%)			Predicted non-undocumented (n = 101, 87.8%)		Predicted undocumented (n = 14, 12.2%)		
	No. or Mean	Col % or SD[a]	No. or Mean	Row % or SD	No. or Mean	Row % or SD	p-value[b]	No. or Mean	Row % or SD	No. or Mean	Row % or SD	p-value
Variant of infection							0.17					0.09
Wuhan/Alpha[c]	69	14.8%	42	60.9%	1	1.4%		26	37.7%	0	0.0%	
Delta	34	7.3%	17	50.0%	2	5.9%		13	38.2%	2	5.9%	
Omicron	364	78.1%	255	70.1%	34	9.3%		63	17.3%	12	3.3%	
Infection number							0.34					0.18
First infection	291	62.4%	195	67.0%	20	6.9%		69	23.7%	7	2.4%	
Reinfection	175	37.6%	119	68.0%	17	9.7%		32	18.3%	7	4.0%	
Vaccination status at infection							0.87					0.26
Unvaccinated	131	28.1%	86	65.6%	8	6.1%		35	26.7%	2	1.5%	
Partially vaccinated with initial series	18	3.9%	11	61.1%	1	5.6%		6	33.3%	0	0.0%	
Completed initial series	133	28.5%	90	67.7%	11	8.3%		26	19.5%	6	4.5%	
Booster vaccinated	184	39.5%	127	69.0%	17	9.2%		34	18.5%	6	3.3%	
Age							0.52					0.39
18-39 years	168	36.1%	113	67.3%	10	6.0%		41	24.4%	4	2.4%	
40-65 years	292	62.7%	196	67.1%	26	8.9%		60	20.5%	10	3.4%	
65+ years	6	1.3%	5	83.3%	1	16.7%		0	0.0%	0	0.0%	
Gender							0.41					0.83
Female	263	56.4%	168	63.9%	24	9.1%		62	23.6%	9	3.4%	
Male	202	43.3%	145	71.8%	13	6.4%		39	19.3%	5	2.5%	
Missing	1	0.2%	1	100.0%	0	0.0%		0	0.0%	0	0.0%	
# of chronic conditions							0.96					0.93

0	284	60.9%	203	71.5%	26	9.2%		48	16.9%	7	2.5%	
1	118	25.3%	78	66.1%	8	6.8%		28	23.7%	4	3.4%	
2	29	6.2%	12	41.4%	1	3.4%		15	51.7%	1	3.4%	
3	9	1.9%	4	44.4%	0	0.0%		4	44.4%	1	11.1%	
4+	1	0.2%	1	100.0%	0	0.0%		0	0.0%	0	0.0%	
Missing	25	5.4%	16	64.0%	2	8.0%		6	24.0%	1	4.0%	
Body mass index							0.85					0.43
Less than 18.5	3	0.6%	3	100.0%	0	0.0%		0	0.0%	0	0.0%	
18.5 - 24.9	99	21.2%	68	68.7%	9	9.1%		19	19.2%	3	3.0%	
25 - 29.9	175	37.6%	125	71.4%	14	8.0%		34	19.4%	2	1.1%	
30 - 39.9	128	27.5%	84	65.6%	9	7.0%		31	24.2%	4	3.1%	
Greater than 40	29	6.2%	15	51.7%	1	3.4%		10	34.5%	3	10.3%	
Missing	32	6.9%	19	59.4%	4	12.5%		7	21.9%	2	6.3%	
# of daily medications							0.11					0.60
0	229	49.1%	159	69.4%	24	10.5%		38	16.6%	8	3.5%	
1	97	20.8%	68	70.1%	6	6.2%		21	21.6%	2	2.1%	
2	53	11.4%	32	60.4%	1	1.9%		19	35.8%	1	1.9%	
3	22	4.7%	12	54.5%	4	18.2%		6	27.3%	0	0.0%	
4	20	4.3%	17	85.0%	0	0.0%		2	10.0%	1	5.0%	
5+	20	4.3%	10	50.0%	0	0.0%		9	45.0%	1	5.0%	
Missing	25	5.4%	16	64.0%	2	8.0%		6	24.0%	1	4.0%	
# of illness symptoms							0.58					0.10
0	58	12.4%	46	79.3%	4	6.9%		8	13.8%	0	0.0%	
1-5	118	25.3%	91	77.1%	15	12.7%		12	10.2%	0	0.0%	
6-10	202	43.3%	136	67.3%	14	6.9%		47	23.3%	5	2.5%	
10+	81	17.4%	34	42.0%	4	4.9%		34	42.0%	9	11.1%	
Missing	7	1.5%	7	100.0%	0	0.0%		0	0.0%	0	0.0%	
# of hours of work missed	46.4	45.5	38.9	28.2	38.4	28.2	0.95	63.1	67.9	44.5	38.9	0.24
# of days spent at least half in bed	2.8	3.6	2.0	2.8	1.9	1.4	0.75	4.9	4.8	2.9	4.1	0.11
Illness duration (# of days)	22.9	22.1	20.8	21.7	18.7	15.4	0.50	30.2	22.5	29.3	29.5	0.91

a. SD = Standard deviation

b. p-values based on Pearson's chi-squared tests or Welch two sample t-test

c. Wuhan and Alpha combined due to very few infections during Alpha variant predominance

In adjusted models, predicted documentation status again did not have a statistically significant effect on whether a participant sought medical care for their infection. The only factors associated with healthcare utilization from any variant of infection were the number of illness symptoms, number of days spent at least half in bed, and illness duration (**Table 9**).

Table 9. Factors associated with healthcare utilization in SARS-CoV-2 infections during enrollment in AZ HEROES, July 2020 to February 2023, by variant of infection[a]

Characteristic	**Wuhan/Alpha infection n = 69**		**Omicron infection n = 363**	
	OR[b]	**95% CI[c]**	**OR**	**95% CI**
Predicted undocumented status	0.00	0.0001 - 379.84	1.43	0.52 - 3.7
Reinfection	4.54	0.13 - 357.21	1.46	0.68 - 3.18
Vaccination status at infection	1.35	0.23 - 6.88	1.03	0.74 - 1.47
# of chronic conditions	4.59	0.86 - 36.94	2.05	0.89 - 4.00
Body mass index	0.81	0.65 - 95.02	1.04	0.99 - 1.09
# of daily medications	1.25	0.63 - 2.56	0.91	0.65 - 1.26
# of illness symptoms	1.20	0.93 - 1.6	**1.17***	**1.04 - 1.33**
# of hours of work missed	2.56	0.18 - 55.22	1.50	0.62 - 3.87
# of days spent at least half in bed	**1.39***	**1.06 - 1.97**	1.07	0.95 - 1.21
Illness duration (# of days)	**1.02***	**1.01 - 1.07**	**1.01***	**1.00 - 1.03**

a. Odds ratios for Delta infections unable to be calculated due to limited observations and model convergence issues
b. OR = Odds Ratio, calculated with multivariable logistic regression models with robust standard errors. Models adjusted for all listed characteristics
c. CI = Confidence Interval
* statistically significant

Participants who tested positive for SARS-CoV-2 during the Wuhan/Alpha variant predominance were 39% more likely to seek care for each day they spent half in bed (OR 1.39, 95% CI 1.06 – 1.97), and for each additional day that their illness lasted, they were 2% more likely to seek care (OR 1.02, 95% CI 1.01 – 1.07). In infections from the Omicron variant, participants were 17% more likely to seek medical care for each additional illness symptom (OR 1.17, 95% CI 1.04 – 1.33) and 1% more likely to seek care for each additional day that their illness lasted (OR 1.01, 95% CI 1.00 – 1.03).

Discussion

This study examined healthcare utilization during SARS-CoV-2 infections in a cohort of Hispanic essential workers in Arizona from July 2020 through April 2023, based on predicted documentation status. The prospective nature of the study allowed for a unique comparison of healthcare utilization throughout the course of the pandemic in workers across an entire state, and to examine drivers of utilization. While overall healthcare utilization, and the subsequent strain on healthcare resources, has been greater during the COVID-19 pandemic than other recent pandemics,[123] the overall healthcare utilization rate amongst this cohort was relatively small. This is likely due to the overall healthy worker make-up of study participants and the high rates of vaccination in the cohort. While previous literature found differences in healthcare utilization rates by undocumented individuals in the early stages of the pandemic,[106] the current analysis did not. It is unclear if this is due to potential misclassification of individuals in our cohort, changes in healthcare utilization as the pandemic went on, or some other factor.

During SARS-CoV-2 infection, participants infected with the Omicron variant had the lowest rate of healthcare utilization of all variants, which has not been the case with the general population.[101,102] This too is likely due to the healthy worker make-up of the cohort and higher than average rates of initial series of COVID-19 vaccination and booster vaccinations. And in adjusted models for SARS-CoV-2 infections, we found no difference in healthcare utilization by predicted documentation status, whether the infection was a first or reinfection, vaccination status, sociodemographic characteristics, or participant baseline health characteristics by any variant of infection. Because of the novelty of the current study, it is impossible to contextualize this within the larger public health literature, though it remains to be seen if these findings might be replicated by future analyses of different cohorts.

The main drivers of healthcare utilization found in our analysis of infections were related to the severity and duration of illness. Increasing number of days spent at least half in bed, and duration of illness were associated with higher medical utilization if the infection came from the Wuhan/Alpha variants, and increasing number of symptoms and illness duration if the infection was due to the Omicron variant. Despite the relatively modest effect sizes of the ORs, particularly illness duration, their cumulative impact on healthcare can be contextualized by the incremental risk posed from single symptoms, number of days spent in bed, and number of days of illness. Given the wide distribution of these variables within the cohort, they represent a meaningful difference.

This study is subject to several limitations. First, the true documentation status of participants is not known. It is possible that differences in healthcare utilization by documentation status are present in this cohort but were not found in this analysis because participants may have been misclassified. While outside the scope of the current dissertation, validating the support vector machine model on an external validation data set would be advisable. Second, the number of infections from the Wuhan, Alpha, and Delta variants were lower than from the Omicron variant. This, combined with the low percentage of infections from predicted undocumented individuals, created sparse data issues leading to model convergence problems for the Delta variant, and very wide confidence intervals for a number of variables in the Wuhan/Alpha variant adjusted model. Third, because participants are all from the state of Arizona, there may be limited generalizability, though there was wide distribution by several sociodemographic and health characteristics. Fourth, the study population was of overall good health, and individuals willing to complete weekly nasal swabs and participate in a long prospective cohort may not be generalizable to the overall population. Fifth, all healthcare utilization was self-reported. It is possible that there were some individuals who did not report care that they received and may have been misclassified, though each participant had at least two opportunities to report any medical care they may have received. Lastly, the final illness survey was administered upon the cessation of symptoms. It is possible that some participants may have experienced a relapse in health and may have received medical care following their final illness survey.

Conclusions

Integrating a previously internally validated predictive model of documentation status into the AZ HEROES cohort allowed for a unique examination of the drivers of healthcare utilization throughout the different variants of the COVID-19 pandemic, and if rates of healthcare utilization differed by documentation status. Our findings suggest that differential utilization of healthcare services by documentation status were not present in our cohort, though these findings should be accepted with caution given that this was the first implementation of our predictive model in a data set where true documentation status is not known, and sample size issues for non-Omicron infections. Further investigation in a data set with more participants is warranted.

Public health implications

Our findings also suggest that efforts to ease the strain on healthcare systems might best be focused on decreasing symptom number and severity. Continued efforts to encourage vaccination would seem helpful, as would further research into the development of antiviral treatments that show promise in reducing healthcare utilization when administered to SARS-CoV-2-positive patients at risk of severe disease but are experiencing mild to moderate symptoms at the time of administration.[124]

CHAPTER 5: OVERALL CONCLUSIONS AND FUTURE DIRECTIONS

Dissertation Conclusions

The scoping review described in Chapter 2 demonstrated that there have been few studies related to chronic health outcomes of undocumented immigrants compared to documented immigrants or US-born citizens. Evidence for differential health outcomes based on immigration status was mixed, which contrasted with the relatively clear background literature on barriers to undocumented individuals' accessing care, and lack of insurance and regular source of care. All eight studies included in the scoping review had significant methodological limitations, however. Most relied on indirect methods of determining documentation status, had small sample sizes, and had limited generalizability. It was well established, though, that undocumented immigrants face significant barriers to accessing medical care. Further research is warranted given the paucity of studies on the topic.

Chapter 3 examined three predictive models of documentation status to overcome many of the logistical challenges and ethical considerations related to studying undocumented populations directly. All predictive models were evaluated on several metrics as a part of a formal internal validation process, and while two of the models fared poorly in all evaluation metrics, the support vector machine model demonstrated a strong association between predicted and actual documentation status for this group of Hispanic individuals and passed all four metrics of evaluation. There appeared to be enough promise from this model to justify its implementation in a new population.

Chapter 4 assigned predicted documentation status to individuals in a longitudinal cohort of essential workers in Arizona using the support vector machine model, and then examined if rates of healthcare utilization differed by documentation status via multivariable logistic regression models. Findings suggested that there was no statistically significant association between predicted documentation status and healthcare utilization, and the only factors associated with utilization were illness-related variables. This was the first examination of its kind, and the findings should be interpreted with caution. Further investigation, especially in data sets with more participants and in which documentation status is collected (to externally validate the prediction model) would be useful. So too would implementation of the predictive model in a data set from a healthcare facility's electronic health record system, preferably one where patients were treated over time for chronic conditions. This would not only fulfil the original vision of this dissertation project but would also give insight into health outcomes for conditions where repeat contact with providers is important for successful management.

Implications for Public Health and Future Directions

The methods developed for the current dissertation have implications for public health policy and practice. First, if the support vector machine model could be externally validated, it would go some way towards affirming its usefulness and demonstrating its utility in a variety of settings. A model like this, because of the variables on which it is based, could be implemented not only in data sets from public health research, but also the clinical setting or in the community. A better understanding of potential differences in undocumented immigrants compared to documented immigrants or US-born individuals in the prevalence of certain diseases, as well as their progression or severity, could be useful in a number of different circumstances.

It could help inform policy decisions regarding the cost of health services for different populations and resource allocation at different levels of government. It could also be useful for community organizations and agencies conducting community outreach and health education efforts with immigrant communities by providing a clearer picture of the health challenges and needs of local populations. If used to predict health outcomes for predicted undocumented immigrants, it could further be used to aid health service entities in developing care models based on the overall number of undocumented immigrants in their service area, as well as assist in forecasting provider and support staff needs and allocations. While true representation within health research for undocumented immigrants does not lie solely in machine learning and predictive models, this is a tool that can be utilized to make some small progress towards understanding this population better, the unique challenges they face, and planning for their needs – in order to be able to act towards their benefit from a place of knowledge.

APPENDIX A: MANUSCRIPT 1 AS SUBMITTED TO BMC PUBLIC HEALTH

TITLE: IMMIGRATION STATUS AND CHRONIC DISEASE OUTCOMES – A SCOPING REVIEW

Author listing: Patrick Rivers, MPP[1]; Kelly N.B. Palmer, PhD[1]; Abidemi Okechukwu, MPH[1]; D. Jean McClelland, MLS[2]; David O. Garcia, PhD[1]; Xiaoxiao Sun, PhD[1]; Kristen Pogreba-Brown, PhD, MPH[1]; Scott C. Carvajal, PhD[1]; David G. Marrero, PhD[1]

(1) University of Arizona, Mel and Enid Zuckerman College of Public Health (2) University of Arizona, Health Sciences Library

Corresponding author: Patrick Rivers, Department of Health Promotion Sciences, University of Arizona, 655 N. Alvernon Way, Tucson, AZ 85712, 406-461-0270, privers@arizona.edu

Abstract

Background: Undocumented immigrants face significant barriers to accessing regular medical care in the U.S. This is a concern especially for individuals with chronic conditions. This scoping review provides an overview of the current knowledge on chronic health outcomes for undocumented immigrants compared to documented immigrants or US-born citizens.

Methods: The review process was conducted in accordance with the PRISMA-ScR guidelines, and the selection of studies was based on pre-defined criteria.

Results: Nine articles were included. Key information such as population details, and study details, outcomes, and limitations are presented. There was conflicting evidence as to whether undocumented immigrants have poorer health outcomes, with undocumented immigrants faring better on nine (43%) measures, faring worse on nine (43%), and with no difference found on three (14%) measures when compared to documented immigrants or US-born citizens.

Conclusion: Undocumented individuals face significant barriers to accessing the same level of health care for their chronic conditions, but it is unclear if this translates to poorer health outcomes. Future, rigorous studies are recommended to address this gap and better understand the health of this vulnerable population.

KEYWORDS: Undocumented Immigrants, Chronic Disease, Scoping Review, Public Health

Background

The United States (U.S.) hosts more foreign-born individuals than any nation in the world.[1] Of the 43.7 million immigrants living in the country, 11.1 million are undocumented[2] Undocumented immigrants often face a number of barriers that impact their ability or willingness to access health care services, including: English language proficiency; restricted mobility; experiences of marginalization and stigma; and mistrust of the health care system.[8] In addition, undocumented immigrants are ineligible for U.S. medical insurance and health care coverage options,[40] and are less likely to have a primary care provider,[41] a regular source of medical care,[42] or access to and use of health services.[38] Undocumented immigrants are also often unwilling to seek out medical care due to fear of deportation.[43,44] Moreover, although immigrants and, to a lesser extent, U.S.-born Hispanics in the United States exhibit relatively favorable health,[15,16] this advantage seems to deteriorate as immigrants gain more experience in the "host" country, and as they become more acculturated.[17,18]

All of these challenges are particularly concerning for individuals with a chronic health condition, which requires ongoing management and supervision from both the individual and their health care team to achieve optimal outcomes and reduce the risk of developing a wide array of serious and costly health complications.[45] Given that undocumented status has been shown to reduce healthcare utilization in Hispanic individuals with chronic conditions,[46] it is likely that undocumented individuals do not receive the same level of medical care to manage their condition as documented immigrants or U.S. citizens. This could lead to poorer disease management and health outcomes which can have significant social and economic consequences.[47]

Rationale and Objective

While previous research has uncovered a number of barriers to care for undocumented immigrants, the health burden of chronic diseases on undocumented immigrants in the United States remains almost entirely unexplored. No review of any kind has been published that explores and summarizes evidence on chronic disease health outcomes of undocumented immigrants compared to documented immigrants and U.S.-born citizens. Due to the paucity of research on health outcomes of undocumented immigrants with chronic diseases we conducted a scoping review of studies that have examined chronic disease outcomes for undocumented Hispanic immigrants compared to documented immigrants or US-born citizens.

Methods

The current study was guided by the PRISMA-ScR guidelines as well as methods outlined by the Joanna Briggs Institute for scoping reviews, and includes the following stages: 1) identifying the research questions, 2) identifying relevant studies, 3) study selection, 4) extracting and charting the data, and 5) collating, summarizing and reporting the results.[48-50]

Search Strategy

A systematic literature search was conducted on October 10, 2019 in PubMed, MEDLINE, Web of Science, Embase, and CINAHL, with supplementary searches up to June 24, 2023. No restrictions were placed on article publication date, though only those published in English were considered. The following Medical Subject Headings (MeSH) and keywords were used in the search: (“emigrants and immigrants”[mesh] OR migrant OR immigrant OR emigrant OR immigration OR "foreign-born") AND ("legal status" OR undocumented OR documented OR authorized OR unauthorized OR illegal) AND ("diabetes mellitus"[mesh] OR diabetes OR A1C OR chronic illness OR chronic disease OR chronic condition) AND ("United States" OR "US" OR "U.S."). PR and JM were responsible for creating the search strategy and conducting the literature search. References in included papers were searched for additional articles that should be added.

Eligibility Criteria

Following best practices recommendations for scoping reviews,[51] we utilized the PCC strategy for our eligibility criteria, outlined in **Table 1**.

Table 1. Scoping review inclusion and exclusion criteria

PCC Strategy	Inclusion criteria	Exclusion criteria
P – Population	Undocumented immigrants, 18+ years old, living in the U.S., focus on Hispanic individuals (at least 80% of study participants)	Studies focused on children and/or adolescents (<18 years of age)
C - Concept	Quantitative studies with individuals with at least one chronic disease (type 2 diabetes, heart disease, obesity, or chronic kidney disease) reporting outcomes compared to either or both documented immigrants/U.S.-born citizens	Descriptive articles such as reviews, lectures, conference proceedings, or qualitative-only studies.
C - Context	Studies conducted in the United States	Studies not published in the English language

Study Selection and Data Extraction

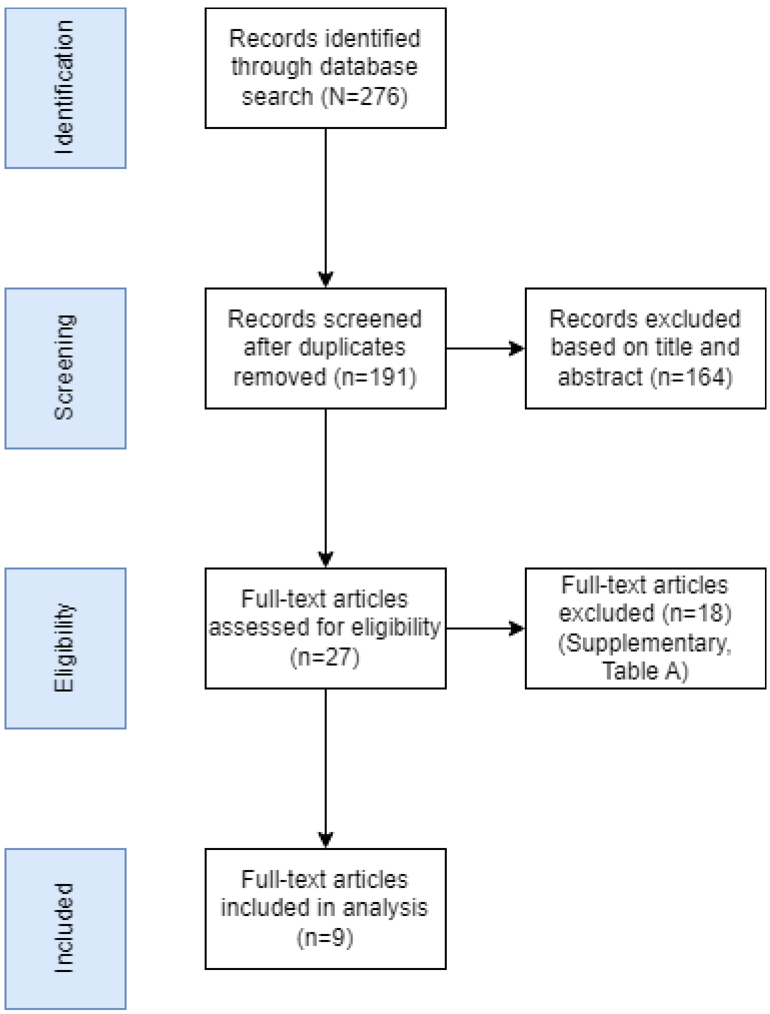
Two researchers (PR and KP) independently screened titles and abstracts, and a third (AO) checked for discrepancies and resolved any conflicts. This process was repeated for full text screening with each author fulfilling the same role. Reasons for exclusion during full text screening are detailed in **Table 5**. A Research Electronic Data Capture (REDCap) database was developed *a priori* specifically for data extraction for this project. PR and KP independently extracted the following data: title; first author; year of publication; geographic location; sample size; age range; gender breakdown; aim of the study; country(ies) of origin of undocumented study participants; method of determining documentation status; comparator; information about primary and secondary outcomes (list of outcomes, results and significance); major limitations; and references to obtain. Data extraction reports were checked by AO for discrepancies with conflicts resolved by all three authors. In accordance with the scoping review methodology, a formal assessment of the quality of individual articles was not included, though quality was addressed in the discussion. No articles were excluded based on methodological quality.

Results

Search Results, Inclusion, and Exclusion

Figure 1 presents a flow chart of the search results and selection of eligible publications. The literature search generated a total of 276 articles, 191 after the removal of duplicates. A total of 164 articles were removed after title and abstract screening if there was enough information to demonstrate that they did not meet the inclusion criteria or that they fulfilled one or more of the exclusion criteria. This left 27 full-text articles to be assessed, of which 9 met the eligibility criteria and were included in this review.

Figure 1. PRISMA flow diagram of literature search and selection for scoping review



Study Characteristics

Study characteristics were gathered, including first author, year of publication, geographic location, aims/purpose, and data source (**Table 2**). All nine studies were published after 2000, with one published in 2004 and the remaining seven published 2011-2022. Five studies examined outcomes for patients with End Stage Renal Disease, one on farmworkers, and only three that might be considered the “general population.” Four studies were centered on a specific hospital or group of local hospitals, four utilized national survey data, and one used survey data with information from two different states. A total of 59,103 individuals were included in analysis across the nine studies.

Table 2. Study characteristics of studies included in scoping review

Authors	Title	Year	Geographic Location	Aims/Purpose	Data Source
Coritsidis et al.[52]	The Initiation of Dialysis in Undocumented Aliens: The Impact on a Public Hospital System	2004	New York City, NY	Compare health outcomes of undocumented individuals with ESRD to American citizens	Medical records from Nassau University Medical Center, New York NY; and Jacobi Medical Center, New York NY, 1993-2001, with follow-up interview surveys administered by physicians
Do et al.[46]	Healthcare Utilization Among Hispanic Immigrants with Diabetes: Investigating the Effect of US Documentation Status	2014	US (national survey)	1) To determine if documented and undocumented Hispanic immigrants with diabetes differ by health insurance status and having a usual place of care, 2) To investigate the effect that documentation status has on health care utilization among diabetic Hispanic immigrants.	Pew Hispanic Center and Robert Johnson Wood Foundation's 2007 Hispanic Healthcare Survey
Hamilton et al.[53]	Immigrant Legal Status and Health: Legal Status Disparities in Chronic Conditions and Musculoskeletal Pain Among Mexican-Born Farm Workers in the United States	2019	US (national survey)	To determine whether a legal status disparity exists in chronic conditions and musculoskeletal pain among Mexican-born immigrants employed as farm workers in the US	Farmworkers interviewed as part of the National Agricultural Workers Survey, 2000-2015
Iten et al.[54]	Undocumented Immigration Status and Diabetes Care Among Mexican Immigrants In Two Immigration "Sanctuary" Areas	2014	San Francisco, CA; Chicago, IL	To investigate the relationship between immigration status and the patient experience of health care, diabetes self-management, and clinical outcomes among Mexican immigrants with diabetes receiving health care in two immigration sanctuary cities	Immigration, Culture and Health Care study with respondents from San Francisco, CA and Chicago, IL, 2008-2009
Linden et al.[55]	Kidney Transplantation in Undocumented Immigrants With ESRD: A Policy Whose Time Has Come?	2012	New York City, NY	To examine the demographic and health characteristics of undocumented immigrants with ESRD, as well as transplant attitudes and options	Medical records from Elmhurst Hospital, New York NY, 2010

Madden et al.[56]	Dialysis Hospitalization Inequities by Hispanic Ethnicity and Immigration Status	2017	Albuquerque, New Mexico	To examine the likelihood of being admitted to an emergency room for dialysis among undocumented immigrants, compared with other patients	Medical records from patients admitted via the emergency department at the University of New Mexico Hospital. 2013-2016
Raghavan et al.[57]	Descriptive Analysis of Undocumented Residents with ESRD in a Public Hospital System	2011	Houston, TX	To evaluate the demographics of the patient population receiving hemodialysis in two Houston public hospitals and an affiliated clinic	Medical records of patients receiving hemodialysis in the Harris County Health district, 2010
Ruhnke et al.[58]	A healthy migrant effect? Estimating health outcomes of the undocumented immigrant population in the United States using machine learning	2021	US (national survey)	To evaluate whether an immigrant health advantage persists among undocumented immigrants in the US	Pooled cross-sections of the National Health Interview Survey, 2000-2018
Shen et al.[59]	Association of Citizenship Status With Kidney Transplantation in Medicaid Patients	2018	US (national survey)	To assess if undocumented immigrants would have similar outcomes to US citizens following kidney transplant if equally insured.	Medical records of adult Medicaid patients in the US Renal Data System who received a kidney transplant, 1990-2011

Study Population Characteristics

Sample sizes of the various studies ranged from 127 to 434,985 individuals (**Table 3**). Studies with smaller numbers of participants, particularly undocumented individuals, tended to be those where data was collected from an individual hospital system. The larger samples generally came from studies conducting secondary data analysis of national survey data. One study did not report any information on the ages of participants. For the remaining eight studies, the mean age of participants in the studies ranged from mid-30s to mid-50s. Two studies did not break out ages of undocumented individuals versus the comparison group, but for the six that did, undocumented individuals were approximately ten years younger than the comparison group in four of them. Gender of participants was about two-thirds male (four studies), relatively equal between male and female (four studies), or not disclosed (one study). The comparison group was both documented immigrants and U.S.-born citizens in six studies, and only documented immigrants in the remaining three studies. Finally, four studies used self-report questions to determine undocumented status, two studies used administrative data (lack of social security number), one counted use of an interpreter as a proxy for undocumented status, one used imputation methods, and one did not disclose the method of ascertaining documentation status.

Table 3. Participant characteristics of studies included in scoping review

Authors	Title	Sample	Age Range or Mean Age	Gender Breakdown	Comparator	Country(ies) of Origin	Measurement of legal status
Coritsidis et al.[52]	The Initiation of Dialysis in Undocumented Aliens: The Impact on a Public Hospital System	Undocumented immigrant patients (n=55) compared to American citizens (n=223), both with ESRD	Undocumented immigrants = 45.9 years; American citizens = 54.3 years	Undocumented immigrants = 62% male; American citizens = 65% male	Documented Immigrants and US-born Citizens	Undocumented immigrants were primarily of Hispanic origin (58%), mostly Central American from El Salvador	Not disclosed in article
Do et al.[46]	Healthcare Utilization Among Hispanic Immigrants with Diabetes: Investigating the Effect of US Documentation Status	Hispanic immigrants 18 years of age or older, undocumented (n=105), documented (n=422)	51.7 years	Male = 49.9%, female = 50.1%	Documented Immigrants	Mexico = 75.6%, Other = 24.4%	Self-report
Hamilton et al.[53]	Immigrant Legal Status and Health: Legal Status Disparities in Chronic Conditions and Musculoskeletal Pain Among Mexican-Born Farm Workers in the United States	Mexican-born respondents under the age of 65 (n = 17,462, breakdown by immigration status not disclosed)	Undocumented immigrants = 31.8 years, documented immigrants = 42.2 years	Undocumented immigrants = 21.4% female; documented immigrants = 27.0% female	Documented Immigrants	Mexico	Self-report
Iten et al.[54]	Undocumented Immigration Status and Diabetes Care Among Mexican Immigrants In Two Immigration "Sanctuary" Areas	Individuals with diabetes who received care in nine free-standing or hospital-based safety-net clinics: undocumented individuals (n=111) compared to the rest of the patient population (n=300)	Undocumented immigrants = 49.8 years; documented immigrants = 55.7 years; US born = 55.7 years	Undocumented immigrants = 51.6% female; documented immigrants = 50.0% female; US born = 54.0% female	Documented Immigrants and US-born Citizens	Mexico	Self-report by elimination

Linden et al.[55]	Kidney Transplantation in Undocumented Immigrants With ESRD: A Policy Whose Time Has Come?	All outpatient dialysis patients from March until May 2010 undocumented individuals (n=45) compared to the rest of the patient population (n=82)	43.8 years	71% male	Documented Immigrants	Mexico and Central America = 58% South America = 29% Asia = 13%	No social security number
Madden et al.[56]	Dialysis Hospitalization Inequities by Hispanic Ethnicity and Immigration Status	Undocumented immigrant patients (n=759) admitted via the emergency department and receiving dialysis compared to random sample of patients (n=3,345) admitted via the emergency department and not receiving dialysis	Not reported	Not reported	Documented Immigrants and US-born Citizens	Not reported	Use of an interpreter
Raghavan et al.[57]	Descriptive Analysis of Undocumented Residents with ESRD in a Public Hospital System	Undocumented immigrants receiving hemodialysis in Houston (total n=186, surveyed n=48) compared to all Hispanics on hemodialysis in Texas (n=13,283)	Undocumented immigrants surveyed = 43.4 years; undocumented immigrants in Houston = 42.6 years; Hispanic individuals in Texas = 58.9 years	Undocumented immigrants surveyed = 35.0% female; undocumented immigrants in Houston = 37.5% female; Hispanic individuals in Texas = 48% female	Documented Immigrants and US-born Citizens	Not reported	No social security number
Ruhnke et al.[58]	A healthy migrant effect? Estimating health outcomes of the undocumented immigrant population in the United States using machine learning	434,985 individuals responding to the National Health Interview Survey, 2000-2018, of which 12,368 were undocumented	Undocumented immigrants = 34.2% 35-49 years, documented immigrants = 40.4% 35-49 years, US-born = 34.1% 35-49 years	Undocumented immigrants = 47.6% female; documented immigrants = 52.0% female; US born = 51.2% female	Documented Immigrants and US-born Citizens	Reported only by region (Asia = 19%, Africa = 3%, Europe = 6%, Mexico & Central America = 61.2%, Middle East = 2%, South America = 7%)	Multi-stage imputation using data from the Survey of Income and Program Participation
Shen et al.[59]	Association of Citizenship Status With Kidney Transplantation in Medicaid Patients	Nonresident aliens (n=346) compared to permanent residents (n=1,489) and US citizens (n=8,660)	Undocumented immigrants = 34.5 years; documented immigrants = 46.0 years; US citizens = 42.0 years	Undocumented immigrants = 50.4% male; documented immigrants = 56.1% male; US citizens = 50.4% male	Documented Immigrants and US-born Citizens	Not reported	Self-report

Study Findings

The nine studies presented twenty-one different physical health outcome comparisons between undocumented individuals and documented immigrants or U.S. citizens. On nine of these measures (43%), undocumented individuals fared worse than their comparison group (**Table 4**). On nine outcomes (43%), undocumented individuals fared better than the comparison group. On the final three outcomes (14%), there was no difference found between undocumented individuals and the comparison group. In other words, in over half of the studied outcomes, researchers found no difference for undocumented individuals compared to documented immigrants or US citizens, or better health outcomes for undocumented individuals.

Table 4. Findings and limitations of studies included in scoping review

Authors	**Title**	**Primary Health outcome(s)**	**0 = no difference, + = undocumented better, - = undocumented worse**	**Primary Health Outcome Details**	**Major Limitations**
Coritsidis et al.[52]	The Initiation of Dialysis in Undocumented Aliens: The Impact on a Public Hospital System	Mean Arterial Pressure Creatinine levels Calculated glomerular filtration rates Magnesium Alkaline phosphatase Albumin	- - - - - -	Undocumented immigrants vs. American citizens for all results: MAP (mm Hg): 119.9 vs. 108.9, p=0.001 Creatinine (mg/dL): 11.90 vs. 9.92, p=0.03 GFR (mL/min/1.73 m2): 5.53 vs. 6.29, p=0.03 Magnesium (mg/dL): 2.43 vs. 2.21, p=0.008 Alkaline phosphatase: 136 vs. 111, p=0.03 Albumin (g/dL): 3.26 vs. 3.0, p=0.04	• Limited generalizability • Cross-sectional data
Do et al.[46]	Healthcare Utilization Among Hispanic Immigrants with Diabetes: Investigating the Effect of US Documentation Status	Time since last visit Self-reported excellent or good health status	- -	Undocumented immigrants vs. documented immigrants for all results: Time since last visit 6 months or less: 12.4% vs. 69.2% Excellent or good health status: AOR = 0.3, 95% CI 0.18, 0.80	• Cross-sectional study data • Single item measure used to assess health care utilization. May be incomplete. • Low response rate: potential sampling bias

Hamilton et al.[53]	Immigrant Legal Status and Health: Legal Status Disparities in Chronic Conditions and Musculoskeletal Pain Among Mexican-Born Farm Workers in the United States	Lifetime, diagnosed chronic conditions Musculoskeletal pain in the last year	+ +	Lifetime diagnosis of at least one chronic condition: 6.4 % of unauthorized immigrants vs. 16.8 % authorized immigrants Musculoskeletal pain: 15% of unauthorized immigrants vs. 21% of authorized immigrants	• Unauthorized immigrants have less access to health care, less opportunity to have chronic condition diagnosed. • Information is collected on the worksite, unauthorized immigrants may have greater fear of retaliation and less likely to report a health problem • Limited generalizability
Iten et al.[54]	Undocumented Immigration Status and Diabetes Care Among Mexican Immigrants In Two Immigration "Sanctuary" Areas	Hemoglobin A1C Systolic blood pressure Low-density lipoprotein	0 0 0	Undocumented immigrants vs. documented immigrants vs. US-born citizens for all results: Hemoglobin A1C (%): 8.45 vs. 7.98 vs. 8.7 Systolic blood pressure (mmHg): 126.07 vs. 130.2 vs. 130.46 Low-density lipoprotein (mg/dL): 88.29 vs. 89.70 vs. 94.28	• Limited generalizability • Cross-sectional data • Small sample size (and low power for some analyses)
Linden et al.[55]	Kidney Transplantation in Undocumented Immigrants With ESRD: A Policy Whose Time Has Come?	Coronary heart disease Diabetes	+ +	Undocumented immigrants vs. documented immigrants for all results: Coronary heart disease: 7% vs. 33% Diabetes mellitus: 40% vs. 68%	• Limited generalizability • Younger undocumented population
Madden et al.[56]	Dialysis Hospitalization Inequities by Hispanic Ethnicity and Immigration Status	Likelihood of receiving emergent vs. scheduled dialysis during first emergency department visit	-	AOR = 3.49, 95% CI 1.03, 11.81	• Interpreter use as proxy for immigration status (if provider is bilingual, may not need interpreter) • Length of stay imprecise - Limited generalizability
Raghavan et al.[57]	Descriptive Analysis of Undocumented Residents with ESRD in a Public Hospital System	Diabetes	+	Undocumented immigrants vs. US-born citizens for all results: Diabetes: 35% vs. 68%	• Limited generalizability • Cross-sectional data • Younger undocumented population
Ruhnke et al.[58]	A healthy migrant effect? Estimating health outcomes of the undocumented immigrant population in the United States using machine learning	Fair/poor self-rated health Any chronic conditions Obesity	+ + +	Undocumented immigrants had 26% lower odds of reporting poor or fair health, 23% lower odds of having been diagnosed with any chronic conditions and 8% lower odds of being obese	• Cannot be verified that individuals assigned "undocumented" truly are, leading to possible mis-specification. • Cross-sectional data
Shen et al.[59]	Association of Citizenship Status With Kidney Transplantation in Medicaid Patients	All-cause transplant failure	+	Undocumented immigrants compared to US-born citizens for all results: All-cause transplant failure: Adjusted HR 0.56, 95% CI 0.41, 0.76	• Limited generalizability • Cross-sectional data • Younger undocumented population

The nine measures on which undocumented individuals fared worse than documented immigrants or US citizens came from three studies. Coritsidis et al. found that undocumented individuals had: higher creatinine levels, a measure of kidney function with higher levels meaning more impaired kidney function (11.9mg/dL vs. 9.92mg/dL, p=0.03); higher blood pressure (mean arterial pressure of 119.9mmHg vs. 108.9mmHg, p=0.001); lower glomerular filtration rates, a test of kidney function with lower levels meaning worse filtration by the kidneys (5.5mL/min/1.73m^2 vs. 6.3mL/min/1.73m^2); higher magnesium levels (2.43mg/dL vs. 2.21mg/dL, p=0.008); higher alkaline phosphatase levels (136 vs. 111, p=0.03); and higher albumin levels (3.26g/dL vs. 3.0g/dL, p=0.04). It should be noted, however, that none of the levels for either group in the last three outcomes (magnesium, alkaline phosphatase, and albumin) are high enough to be of any real medical concern. Do et al. found that undocumented immigrants with diabetes had worse healthcare utilization (12.4% utilizing health services in the last six months vs. 69.2%) and less likelihood of self-reporting a health status of excellent or good health (AOR = 0.3, 95% CI 0.18, 0.80). Finally, Madden et al. found undocumented individuals at much higher odds of being placed on emergent vs. scheduled dialysis upon first admittance to the emergency department, meaning they were in particularly poor condition upon arrival (AOR = 3.49, 95% CI 1.03, 11.81).

The three measures on which there was no difference for undocumented individuals compared to documented individuals or US-born citizens came from one study. Iten et al. found no statistical difference for undocumented individuals in: hemoglobin A1c, a measure of blood sugar (8.4% compared to 8.7% in U.S.-born citizens, p=0.41, and 7.98% in documented immigrants, p=0.06); systolic blood pressure (126.1mmHg compared to 130.5 in U.S.-born citizens, p=0.07, and 130.2 in documented immigrants, p=0.06); and low density lipoprotein cholesterol (88.3mg/dL compared to 94.3mg/dL in U.S.-born citizens, p=0.2, and 89.7mg/dL in documented immigrants, p=0.8).

Finally, the nine measures on which undocumented immigrants fared better than documented immigrants were found in three studies. Hamilton et al. found better outcomes for undocumented immigrants in the percent of individuals who have had a diagnosed chronic condition in their lifetime (6.4% compared to 16.8%, p<0.001), and prevalence of individuals reporting musculoskeletal pain in the last year (15.1% compared to 21.2%, p<0.001). Linden et al. found lower rates, in undocumented individuals, of coronary heart disease (7% compared to 33%) and diabetes mellitus (40% compared to 68%). Raghavan et al. also found lower rates of diabetes in undocumented individuals (35% compared to 68%). Ruhnke et al. found that Undocumented immigrants had 26% lower odds of reporting poor or fair health (p < 0.001), 23% lower odds of having been diagnosed with any chronic conditions (p < 0.001), and 8% lower odds of being obese (p < 0.01). Finally, Shen et al. found lower rates of all-cause transplant failure for undocumented individuals (adjusted HR 0.56, 95% CI 0.41, 0.76).

Discussion

In this first scoping review on the chronic health outcomes of undocumented immigrants compared to documented immigrants or U.S.-born citizens, we found relatively few articles that have studied this question. The increasing number in the last several years is somewhat encouraging, even if the overall number remains low.

Evidence for differential health outcomes based on immigration status was mixed, which was in contrast to the relatively clear background literature on barriers to undocumented individuals' accessing care, and lack of insurance and regular source of care. Three studies, comprising 43% of the 21 health outcomes examined, found evidence of poorer health outcomes for undocumented individuals. Five studies, examining nine outcomes, found evidence of better health outcomes for undocumented individuals. The final study, examining three outcomes, found no difference. There are several possibilities why the studies included in this review do not align with background research findings.

First is the methodological limitations of the nine studies. Eight of the studies used proxies or indirect methods of identifying documentation status that might not be entirely reliable, while one study did not disclose their method of identifying undocumented immigrants. Most studies also relied on small samples, especially the undocumented immigrant population – with the associated analytic problems of small sample sizes and potential bias they carry. Of particular concern was that of representativeness or generalizability. Five of the studies were focused on individuals with End Stage Renal Disease, and one on farmworkers. Of the final three that might be considered from the general public, one of them was a convenience sample from two "sanctuary cities", places that have statutes limiting reporting of the documentation status of individuals seeking health services. It is possible that individuals included in most of these studies differ, in meaningful ways, from the larger population and their outcomes are not generalizable to most other undocumented individuals. Undocumented immigrants were also, on average, approximately ten years younger than the comparator group in the six studies that broke out age characteristics for the comparison groups, which seems a likely contributor to a healthier state.

A second potential reason that the studies did not find stronger support for poorer health outcomes in undocumented immigrants compared to documented immigrants or U.S. citizens is linked to the idea of immigrant health selectivity. This refers to how immigrants tend to differ from non-migrants because of the challenges associated with migration, and how immigrants, even if undocumented, might have, on average, better health than the average U.S.-born citizen, or comparable health characteristics as documented immigrants to which they are compared in individual studies. This is because the costs of migration select on good health as well as on other characteristics that might be associated with good health, such as self-efficacy.[60] Individuals who migrate tend to have differences in resources, networks, and opportunities than those who do not migrate or are unable to. In essence, individuals who migrate might be healthier than those who do not. As with the methodological limitations detailed above, this again might lead to a skewed sample that may not be generalizable to a larger population, which is an issue when only one study (Hamilton et al.) controls in any way for possible positive health selectivity.

Finally, only two of the nine studies account for length of time spent in the United States since immigration. Undocumented immigrants have spent less time, on average, than documented immigrants,[61] let alone U.S. citizens. As acculturation and longer time in the United States is associated with worse health outcomes – likely as a result of stress, cumulative disadvantage, and adoption of unhealthy behaviors common in the country of residence[62] – it is possible that the undocumented immigrants in several of the studies are more recently arrived, and are therefore likely healthier compared to an earlier-arrived, documented immigrant population or US citizen.

Strengths and Limitations of this Review

The present scoping review provides an overview of the published knowledge of health outcomes for undocumented immigrants with chronic conditions compared to documented immigrants or U.S.-born citizens. The main strength of this review is its methodological approach, which utilized the systematic framework recommended in the PRISMA checklist for scoping reviews (PRISMA-ScR).[50] This methodology is useful for topics such as this with relatively few studies and to identify gaps in knowledge.

Although a thorough search for literature in a number databases as well as hand-searches was conducted, there is a possibility that all studies relevant for the review scope were not identified. We did not contact the authors of any studies to provide answers to questions we may have had in the review process. We also decided to limit our inclusion to quantitative studies, as qualitative studies on the subject seemed to be patient experiences and limited to one, or less than five, individuals per study. This might be a limitation to the number and breadth of results, but this decision was made to ensure that our review was not a review of case studies, which requires a separate methodology.[63]

Conclusions

Undocumented individuals face significant barriers to accessing the same level of health care for their chronic conditions compared to documented immigrants and US-born citizens. It is unclear from the published literature if this translates to poorer health outcomes. There are few studies that have addressed this question, however, and there has been little structure about what health outcomes to assess, what populations to include, and what methodology should be used. Given the paucity of research on the topic, and the numerous methodological limitations of the studies included in this article, further research is warranted. This seems particularly true when weighed against the high human cost of chronic diseases, and the more than ten million undocumented immigrants in the United States.

Table 5. Excluded papers (n=18) read in full-text, and reason for exclusion from scoping review

First Author, Year	Title	Reason for Exclusion
Balakrishnan, 2019[64]	The Undocumented Elderly: Coverage Gaps and Low Health Care Use	Commentary/Narrative
Bohan, 2005[65]	Illegal immigrants with ESRD: The impact on U.S. and Canadian health care facilities	Full text not available
Cantu, 2013[66]	New estimates of racial/ethnic differences in life expectancy with chronic morbidity and functional loss: evidence from the National Health Interview Survey	Wrong patient population
Carrillo, 2011[67]	A Regional Health Collaborative Formed By New York-Presbyterian Aims To Improve The Health Of A Largely Hispanic Community	Wrong patient population
Cervantes, 2019[68]	The Status of Provision of Standard Outpatient Dialysis for US Undocumented Immigrants with ESKD	Commentary/Narrative
Cervantes, 2018[69]	Circumstances of Death among Undocumented Immigrants Who Rely on Emergency-Only Hemodialysis	Wrong comparator
Cervantes, 2018[70]	Association of Emergency-Only vs Standard Hemodialysis With Mortality and Health Care Use Among Undocumented Immigrants With End-stage Renal Disease	Wrong comparator
De Jesus Diaz-Perez, 2004[71]	A program to improve access to health care among Mexican immigrants in rural Colorado	Wrong patient population
Garcini, 2018[72]	Health-related quality of life among Mexican-origin Latinos: the role of immigration legal status	Wrong outcomes
Hogan, 2017[73]	Emergent Dialysis and its Impact on Quality of Life in Undocumented Patients with End-Stage Renal Disease	Wrong study design

Hsueh, 2018[74]	Associations between immigrant status and pharmacological treatments for diabetes in U.S. adults	Wrong patient population
Kang, 2015[75]	Perceptions of affordability of care and access to information about health insurance among immigrants after Massachusetts health reform	Full text not available
Lee, 2018[76]	The prevalence and impact of diabetes mellitus among undocumented immigrants in an indigent care program in Riverside, California	Full text not available
Ortega, 2015[77]	When politics trumps health: Undocumented Latino immigrants and US health care	Commentary/Narrative
Peota, 2017[78]	Care in the shadows: The challenges for undocumented immigrants	Commentary/Narrative
Ross, 2019[79]	Association between immigration status and anxiety, depression, and use of anxiolytic and antidepressant medications in the Hispanic Community Health Study/Study of Latinos	Wrong concept
Sher, 2017[80]	Healthcare outcomes in undocumented immigrants undergoing two emergency dialysis approaches	Wrong comparator
Thomsen, 2015[81]	Caring for the Uninsured, Undocumented Population: Challenges of Chronic Disease Management and Access to Care	Commentary/Narrative

Declarations

- Ethics approval and consent to participate: Not applicable.
- Consent for publication: Not applicable.
- Availability of data and materials: The datasets used for the current study are available from the corresponding author on reasonable request.
- Competing interests: The authors declare that they have no competing interests.
- Funding: No funding was used to support this research
- Authors' contributions: PR screened articles, extracted data, and designed and drafted the manuscript; KP screened articles, extracted data, and revised the manuscript; AO checked for article discrepancies and revised the manuscript; DJM assisted with search strategy and revised the manuscript; DOG, XS, KPB, SCC, and DGM contributed to the review's initial conception and revised the manuscript.

Appendix B: MANUSCRIPT 2 AS SUBMITTED TO JMIR FORMATIVE RESEARCH

Title: Development and internal validation of three multivariable prediction models of Hispanic immigrant documentation status utilizing two samples from the southwestern United States

Author listing: Patrick Rivers, MPP[1]; Xiaoxiao Sun, PhD[1]; David O. Garcia, PhD[1]; Kristen Pogreba-Brown, PhD, MPH[1]; Scott C. Carvajal, PhD[1]; David G. Marrero, PhD[1]

(1) University of Arizona, Mel and Enid Zuckerman College of Public Health

Corresponding author: Patrick Rivers, Department of Health Promotion Sciences, University of Arizona, 655 N. Alvernon Way, Tucson, AZ 85712, 406-461-0270, privers@arizona.edu

Abstract

Background: Research on undocumented populations is sparse, both because documentation status is rarely ascertained in studies, and because undocumented individuals are less likely overall to participate in research studies. A better understanding of the health of these individuals is important, though, and one method that has been proposed to examine outcomes for this group without putting them at additional risk is through predictive modelling.

Objective: Develop and internally validate a predictive model of documentation status.

Methods: Utilizing combined data from two population-based samples in Arizona in which documentation status were collected, three prediction models of documentation status were created and internally validated. Models created used multiple imputation by chained equations (MICE), and two machine learning algorithms: random forests machine learning and support vector machine. Predicted documentation status was modelled from demographic, health behavior, and disease status indicators commonly collected in public health research studies. The performance of each model was assessed in terms of accuracy, precision, recall, and Matthews correlation coefficient (MCC).

Results: The combined sample consisted of 473 individuals, of whom 85 were undocumented. The sample included 209 men and 263 women, aged 20 to 83 years of age at the time of their participation in the respective studies. The MICE model had the lowest rates of accuracy (64.3%), precision (0.81), recall (0.74), and MCC (-0.07). The random forests model was slightly better with an accuracy of 73.1%, precision of 0.82, recall of 0.86, and MCC of 0.002. The support vector machine was the strongest in all categories, with an accuracy of 90.7%, precision of 0.90, recall of 0.99, and MCC of 0.66.

Conclusions: Utilizing predictive models could protect the safety of vulnerable individuals as well as make possible investigation into potentially undocumented individuals in settings where it would be impractical or impossible to gather information on documentation status first-hand. The results of the current analysis demonstrate some promise and further investigation, particularly the support vector machine model.

Keywords: undocumented immigrants; machine learning; prediction models; public health

Background

Of the 11.3 million undocumented immigrants in the United States (US), more than 8.6 million (76%) are Hispanic.[82] Undocumented immigrants are often forced to forego medical care due to lack of insurance,[7] experiences of discrimination or mistrust of the health care system,[8,9] or fear of deportation,[43] and receive more infrequent and worse care than documented immigrants or U.S. citizens.[10-12] Fears related to their immigration status also makes undocumented individuals less likely to participate in research studies,[29] making them a challenging group to study directly.

Research on undocumented populations has been hindered by the lack of data on immigrant legal status in the types of large-scale, nationally-representative survey data generally used to answer many health and economic questions.[19] One reason that studies omit questions on documentation status is the possibility of putting undocumented respondents at risk if they respond honestly, and data including direct or indirect identifiers are somehow released to or requested by government agencies.[20,21] For this reason, proxy measures of immigration status have been recommended for use in research studies attempting to better understand immigrant health. Preliminary research has shown that predictive modeling is an effective approach to address these concerns without putting undocumented individuals at additional risk.[29]

Our approach utilized a methodology developed for previous predictive models of immigrant documentation status that have been used to answer economic questions (multiple imputation by chained equations),[19,29] and two machine learning methods (random forest and support vector machine). The setting of the study groups used for this prediction model

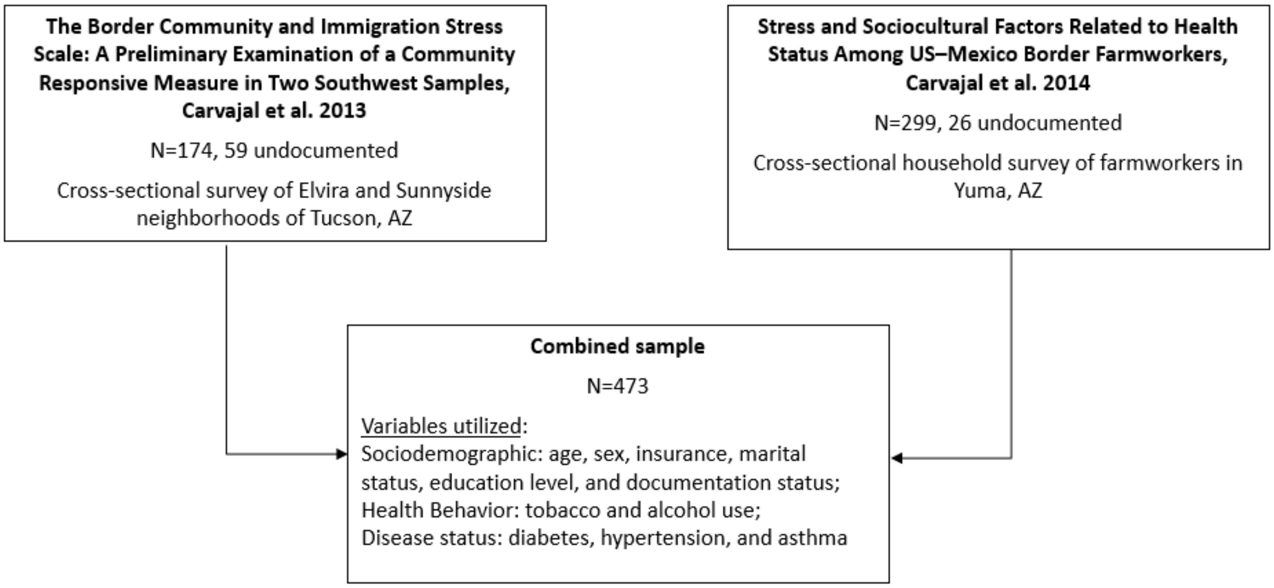
(Southwestern US) positions it well for better understanding Hispanic undocumented immigrants. The objectives of this study are to develop and internally validate three predictive models of documentation status for Hispanic immigrants in the Southwestern US.

Methods

Source of Data

Data for this project comes from two population-based samples from 2013 to 2014 in Tucson and Yuma, AZ, and includes 473 total individuals.[83,84] Participants were recruited from two majority Hispanic neighborhoods (Tucson study)[12] or farmworker communities (Yuma study).[11] Participants in the Tucson study were selected from community residents, with those in the Yuma study drawn from randomly selected households within specific census blocks from low-income communities. A range of demographic, health behavior, and disease status indicators were collected in each cross-sectional study examining the effects of stress on mental and physical health outcomes in border communities (**Figure 2**). In both studies, participants self-reported their documentation status. All participants from the Tucson and Yuma studies were used in the creation of the prediction models.

Figure 1. Study sample pooling for prediction models



Measures

Individual documentation status, a binary variable of documented or undocumented, was the outcome of our prediction mode. While documentation status was known for the entire sample, it was randomly assigned as “missing” for 50 percent of the sample, with the known values of documentation status for this 50 percent forming the basis of comparison between the predicted values and the observed values to gauge the models’ performance.

Predictors taken from each study and used in the prediction models included demographic, health behavior, and disease status indicators. Age was recorded as a continuous variable and was measured in years. Categorical variables included were: sex (male or female); insurance status (insured or uninsured); overall health (excellent, very good, good, fair, or poor); marital status (single, married, or other); education level (never attended, completed grade 8 or less, completed grades 9-11, high school diploma or equivalent, completed some college, and college graduate); tobacco use (current smoker, former smoker, never smoker); as well as previous diagnosis (yes/no) of each of the following conditions: hypertension, asthma, and diabetes.

All predictors were self-reported by Tucson and Yuma study participants during interviews with trained study staff.

Statistical Analysis

Three prediction models were created and evaluated to determine the best approach for identifying documentation status. In the first model, documentation status was predicted using combined sample multiple imputation (CSMI), a method utilized in economics literature that aims to impute the immigration status of noncitizens in surveys lacking questions about immigration status.[29,85-87] Documentation status and all predictors listed above from each study were combined into one sample. Documentation status was randomly assigned as "missing" for 50 percent of the sample. Multiple imputation by chained equations (MICE) was then used to predict documentation status for the 50 percent of the sample for which it was "missing", utilizing all predictors listed above. MICE specifies a suitable conditional imputation model for each incomplete variable and iteratively imputes until convergence.[88] It has shown large efficiency gains and substantial reduction in omitted variable bias compared to other imputation techniques and has been the preferred method of imputing the immigration status of noncitizens into surveys lacking questions on immigration status.[86,87]

The next two models utilized machine learning algorithms. In both, the data was split in two, with one half acting as the training data set to develop the algorithms, and the other half acting as the testing data set to refine the algorithms' performance. In the first of these algorithms, we utilized random forests, which is a non-linear model aggregating the results of a series of tree structure classifiers, or decision trees, in which each tree depends on the values of a random vector sampled independently and with the same distribution for all trees.[89,90] Random forests generally demonstrate high levels of accuracy, manage outliers and noise well, and are resistant to overfitting. In this model, we used 500 trees, with the node-size set to default, and 20 variables tried at each split.[91] The final model, and the second machine learning algorithm, was a support vector machine, which is a highly robust prediction method in which input vectors are non-linearly mapped to a high dimension linear decision surface.[92] It demonstrates high generalization and flexibility, and is also resistant to over-fitting.[93]

Performance

All models were assessed in terms of accuracy, precision, recall, and Matthews correlation coefficient. Accuracy is the rate at which the classifier is correct, or the level of agreement between predictions from the model and observed outcomes, in this case the predicted documentation status.[94] Accuracy is defined as the sum of the number of true positives and true negatives, divided by the total number of observations, and ranges from 0 percent to 100 percent with higher numbers indicating better accuracy. While it can trend towards overoptimism on the ability of the models to classify the majority class in an unbalanced data set (which ours was, with the number of undocumented individuals smaller than the number of documented individuals), it was included here, following traditional reporting methods.

Recall, also known as the true positive rate or sensitivity, corresponds to the number of true positives divided by the sum of true positives and false negatives (in this case undocumented status). Values range from 0 to 1, with higher values demonstrating better recall. Precision is the number of true positives divided by the sum of true positives and false positives (also undocumented status). Similar to recall, values range from 0 to 1, with higher values demonstrating better precision.

Matthews correlation coefficient is a special case of the phi coefficient.[95,96] It is a more reliable statistical rate than the F_1 score, which is commonly used, and only produces a high score with all of the following: high proportions of true positives and true negatives, and low proportions of false negatives and false positives.[96] It is resistant to the overoptimism problems of accuracy and is generally viewed as a means of overcoming class imbalance issues.[19] It ranges from -1 (complete misclassification) to 1 (perfect classification).

All analyses were completed in R (version 4.1.2).[97]

Results

Baseline characteristics

The combined sample consisted of 473 individuals, of whom 85 were undocumented. The sample included 209 men and 263 women, aged 20 to 83 years of age at the time of their participation in the respective studies. Participants in the Tucson study had a higher proportion of undocumented individuals (33.9% to 8.7%, $p < 0.001$), and females (67.8% to 48.5%, $p < 0.001$) compared to individuals in the Yuma study. Tucson participants were also more likely to be single (21.3% to 11.8%, $p < 0.001$) were more educated ($p < 0.001$), reported higher good or better overall health (75.9% to 62.2%, $p < 0.001$) and a larger proportion that had never smoked (70.6% to 21.4%, $p < 0.001$) compared to individuals in the Yuma study. Rates between the two studies were similar for age, insurance status, and previous diagnosis of each of diabetes, hypertension, and asthma (**Table 1**).

Table 1. Descriptive statistics, stratified by study group, of key study characteristics of participants in two Southwest studies examining the effect of stress on mental and physical health outcomes

Characteristic	**Carvajal, et al. 2013 (Tucson, AZ) N = 174**	**Carvajal, et al. 2014 (Yuma, AZ) N = 299**	**p-value**
Undocumented, n (%)	59 (33.9)	26 (8.7)	< 0.001
Female, n (%)	118 (67.8)	145 (48.5)	< 0.001
Age, mean (SD)	46.2 (15.0)	44.3 (11.1)	0.06
Marital Status			
Single, n (%)	37 (21.3)	35 (11.8)	< 0.001
Married, n (%)	108 (62.1)	239 (80.5)	
Other, n (%)	29 (16.7)	23 (7.7)	
Education Level			< 0.001
Never attended, n (%)	3 (1.8)	3 (1.0)	
Grades 1-8, n (%)	21 (12.5)	246 (82.3)	
Grades 9-11, n (%)	46 (27.4)	47 (15.7)	
High school diploma or GED, n (%)	46 (27.4)	3 (1.0)	
Some college, n (%)	36 (21.4)	0 (0)	
College degree or above, n (%)	16 (9.5)	0 (0)	
Insurance			0.07
Yes, n (%)	106 (62.0)	160 (53.5)	
No, n (%)	65 (38.0)	139 (46.5)	
Overall health			< 0.001
Excellent	19 (11.2)	14 (4.7)	
Very good	40 (23.5)	36 (12.0)	
Good	70 (41.2)	136 (45.5)	
Fair	31 (18.2)	109 (36.5)	
Poor	10 (5.8)	4 (1.3)	
Tobacco use			< 0.001
Never smoker, n (%)	120 (70.6)	64 (21.4)	
Former smoker, n (%)	23 (13.5)	45 (15.1)	

Current smoker, n (%)	27 (15.9)	190 (63.6)	
Previous diagnosis of:			
Diabetes, n (%)	32 (18.8)	48 (16.1)	0.51
Hypertension, n (%)	40 (24.1)	64 (21.4)	0.69
Asthma, n (%)	16 (9.3)	16 (5.4)	0.11

Accuracy, Precision, Recall, and Matthews Correlation Coefficient

Table 2 details the accuracy, precision, recall, and Matthews correlation coefficient (MCC) of the three models. The model using MICE had the lowest rates of accuracy (64.3%), precision (0.81), recall (0.74), and MCC (-0.07). The random forests model was slightly better in all metrics with an accuracy of 73.1%, precision of 0.82, recall of 0.86, and MCC of 0.002. The support vector machine was the strongest in all categories, with an accuracy of 90.7%, precision of 0.90, recall of 0.99, and MCC of 0.66. Both the MICE and random forests models had relatively high false positive rates (19.5 and 13.9%, respectively) and very high false negative rates (86.8% and 85.9%, respectively), which led to very low MCCs. A similar rate of false positives (19.4%) combined with a very low false negative rate (0.5%) led to a much higher MCC for the support vector machine model.

Table 2. Internal validation evaluation metrics for three models predicting undocumented status from two Southwest studies examining the effect of stress on mental and physical health outcomes

	Accuracy	Precision	Recall	Matthews Correlation Coefficient
MICE[a]	64.3%	0.81	0.74	-0.07
Random Forests	73.1%	0.82	0.86	0.002
Support Vector Machine	90.7%	0.9	0.99	0.66

a. MICE = Multiple Imputation by Chained Equations

Discussion

The predictive models utilizing MICE and random forests fared poorly in all evaluation metrics. The support vector machine model, though, demonstrated high levels of accuracy, precision, recall, and MCC, showing a strong association between predicted and actual documentation status for this group of Hispanic individuals. While the MICE and random forests models do not appear to produce results worth further investigation, at least within this population, the support vector machine model demonstrated enough promise to justify its use in a new population.

These models were developed to aid in further research into Hispanic undocumented populations. Given the sensitive nature of documentation status, it may not be feasible or advisable for researchers to gather that information regularly or directly from participants. While there is a burgeoning literature on this population and the development of predictive models of documentation status within other disciplines, it remains almost entirely undeveloped within health sciences. With the challenges of collecting documentation status, the ability to utilize a proxy created via predictive modeling is useful.

Because of their vulnerable status, better understanding this population is important, especially if it is possible to do so via means that do not put individuals at direct risk or require identifying information. As these predictive models are built utilizing basic demographic, health behavior, and health outcome variables readily available in population and community research projects, they would be widely adaptable and could be utilized to help researchers gain a better understanding of the health status or health outcomes of a range of diseases experienced by potentially undocumented Hispanic individuals.

Strengths and Limitations

The current prediction model complements and builds upon established prediction models for undocumented status used in economics and demonstrates the feasibility of incorporating health behavior and health outcome variables in models related to documentation status. While the model used in economics (MICE) fared relatively poorly, the usage of machine learning, particularly support vector machine algorithms showed significant promise. One benefit of the current study is the assessment of the model's performance via precision, recall, and Matthews correlation coefficient, which is not something typically done with CSMI within economics literature as documentation status in one of the samples is always unknown. Finally, the current analysis builds off of predictive modeling related to documentations status from a different discipline and found that while MICE may be useful and accurate in economics, the integration of health behavior and health status variables in addition to ones typically used previously appears to have had deleterious effects on its applicability to health data.

This study is subject to several limitations. First, each model uses split-sample validation, in which the data set was divided into two, and half of it was used to create the prediction model, and the other half to validate it. This has efficiency losses because it does not use all available data for the model development.[98]. Secondly, these models were created using samples from southern Arizona. These specific models, and the predictors used to create it, might have limited generalizability to different populations, and future researchers would likely need to validate it for their own populations. Finally, these models were not able to be externally validated, and would benefit from further validation with an unrelated data set in which documentation status and all predictor variables are known.

The support vector machine algorithm in particular shows promise and seems a fruitful area of further research. While the rate of false positives remained relatively high, the effect of this in a real-world analysis may have minimal impact on the overall results. Because documented individuals are, relatively, healthier than undocumented individuals, incorrectly assigning an undocumented status to an individual who is not undocumented will likely only attenuate the effect of the object of study. This could lead to higher p-values and make it more difficult to determine statistical significance, but attenuating an effect towards the null is preferable to moving it away from the null in the opposite direction, which could happen with false negatives. The extremely low rate of false negatives, therefore, is encouraging, and overall support vector machine algorithms appear to warrant further investigation.

Conclusions

The expansion of predictive modeling of documentation status into health sciences presents a unique method for understanding the profile and potential needs of undocumented individuals. Utilizing predictive models could protect the safety of vulnerable individuals as well as make possible investigation into potentially undocumented individuals in settings where it would be impractical or impossible to gather information on documentation status first-hand. The results of the current analysis demonstrate some promise and further investigation, and application, of these models appear warranted.

Acknowledgements

- Author contributions: PR conducted analysis and designed and drafted the manuscript; XS, DOG, KPB, SCC, and DGM contributed to the initial conception and revised the manuscript.
- Funding information: No funding was used to support this research.
- Role of sponsors: Not applicable.

Conflicts of interest

None declared.

REFERENCES

1. Lopez G, Bialik K, Radford J. *Key Findings about U.S. Immigrants.* Pew Research Center; September 14 2018.
2. Radford J, Budiman A. *Facts on U.S. Immigrants, 2016: Statistical Portrait of the Foreign-Born Population in the United States.* Pew Research Center; September 14 2018.
3. Ruggles S, Flood S, Foster S, Goeken R. Jose Pacas, Megan Schouweiler and Matthew Sobek. IPUMS USA: Version 11.0 [dataset]. Minneapolis, MN: IPUMS, 2021. In.
4. Grit K, den Otter JJ, Spreij A. Access to health care for undocumented migrants: a comparative policy analysis of England and the Netherlands. *J Health Polit Policy Law.* 2012;37(1):37-67.
5. Hacker K, Chu J, Leung C, et al. The impact of Immigration and Customs Enforcement on immigrant health: perceptions of immigrants in Everett, Massachusetts, USA. *Soc Sci Med.* 2011;73(4):586-594.
6. Rodriguez C, Chishti M, Capps R, John LS. *Program in Flux: New Priorities and Implementation Challenges for 287 (g).* Migration Policy Institute Washington, DC; 2010.
7. Edward J. Undocumented Immigrants and Access to Health Care: Making a Case for Policy Reform. *Policy Polit Nurs Pract.* 2014;15(1-2):5-14.
8. Derose KP, Escarce JJ, Lurie N. Immigrants and health care: sources of vulnerability. *Health affairs (Project Hope).* 2007;26(5):1258-1268.
9. Karp RJ, Rhee D, Feldman D, Bouchkouj N. Outreach to immigrant communities: teaching pediatric residents about access to health care. *J Health Care Poor Underserved.* 2007;18(3):510-515.
10. Stimpson JP, Wilson FA, Su D. Unauthorized immigrants spend less than other immigrants and US natives on health care. *Health Aff (Millwood).* 2013;32(7):1313-1318.
11. Stimpson JP, Wilson FA, Zallman L. ED visits and spending by unauthorized immigrants compared with legal immigrants and US natives. *Am J Emerg Med.* 2014;32(6):679-680.
12. Rodríguez MA, Bustamante AV, Ang A. Perceived quality of care, receipt of preventive care, and usual source of health care among undocumented and other Latinos. *J Gen Intern Med.* 2009;24 Suppl 3(Suppl 3):508-513.
13. Hacker K, Chu J, Arsenault L, Marlin RP. Provider's perspectives on the impact of Immigration and Customs Enforcement (ICE) activity on immigrant health. *J Health Care Poor Underserved.* 2012;23(2):651-665.
14. Asch S, Leake B, Anderson R, Gelberg L. Why do symptomatic patients delay obtaining care for tuberculosis? *Am J Respir Crit Care Med.* 1998;157(4 Pt 1):1244-1248.
15. Argeseanu Cunningham S, Ruben JD, Narayan KM. Health of foreign-born people in the United States: a review. *Health Place.* 2008;14(4):623-635.
16. Markides KS, Eschbach K. Aging, migration, and mortality: current status of research on the Hispanic paradox. *J Gerontol B Psychol Sci Soc Sci.* 2005;60 Spec No 2:68-75.
17. Hunt LM, Schneider S, Comer B. Should "acculturation" be a variable in health research? A critical review of research on US Hispanics. *Soc Sci Med.* 2004;59(5):973-986.
18. Lopez-Class M, Castro FG, Ramirez AG. Conceptions of acculturation: a review and statement of critical issues. *Soc Sci Med.* 2011;72(9):1555-1562.
19. Sohn H, Pebley A. New Approaches to Estimating Documentation Status in Survey Data. Population Association of America; 2018; Denver, Colorado.
20. Carter-Pokras O, Zambrana RE. Collection of legal status information: caution! *American journal of public health.* 2006;96(3):399; author reply 399-400.
21. Prenctice J, Pebley A, Sastry N. Prentice Et Al. Respond. *American journal of public health.* 2006;96(3):399-400.

22. Brown A. *The Unique Challenges of Surveying U.S. Latinos.* Pew Research Center;2015.
23. Doran KM, Castelblanco DG, Mijanovich T. Undocumented Latino Immigrants and Research: New Challenges in Changing Times. *J Health Care Poor Underserved.* 2018;29(2):645-650.
24. Mather M. *Children in immigrant families chart new path.* Washington, DC: Population Reference Bureau;2009.
25. Mertens DM. Ethics and social justice in ethnocultural qualitative research. 2012.
26. Bledsoe KL, Hopson RK. Conducting ethical research and evaluation in underserved communities. In: *The handbook of social research ethics.* SAGE Publishing; 2009:391-406.
27. Hernández MG, Nguyen J, Casanova S, Suárez-Orozco C, Saetermoe CL. Doing no harm and getting it right: guidelines for ethical research with immigrant communities. *New Dir Child Adolesc Dev.* 2013;2013(141):43-60.
28. Carter A, Lawrence M, Morse A. *2011 immigration-related laws, bills and resolutions in the states: January 1–March 31, 2011.* Washington, DC: National Conference of State Legislators;2011.
29. Bachmeier JD, Van Hook J, Bean FD. Can We Measure Immigrants' Legal Status? Lessons from Two U.S. Surveys. *Int Migr Rev.* 2014;48(2):538-566.
30. Link BG, Phelan J. Social conditions as fundamental causes of disease. *J Health Soc Behav.* 1995;Spec No:80-94.
31. Phelan JC, Link BG, Tehranifar P. Social conditions as fundamental causes of health inequalities: theory, evidence, and policy implications. *J Health Soc Behav.* 2010;51 Suppl:S28-40.
32. Castañeda H, Holmes SM, Madrigal DS, Young ME, Beyeler N, Quesada J. Immigration as a social determinant of health. *Annu Rev Public Health.* 2015;36:375-392.
33. Asad AL, Clair M. Racialized legal status as a social determinant of health. *Soc Sci Med.* 2018;199:19-28.
34. Martinez O, Wu E, Sandfort T, et al. Evaluating the impact of immigration policies on health status among undocumented immigrants: a systematic review. *J Immigr Minor Health.* 2015;17(3):947-970.
35. Torres JM, Young MD. A life-course perspective on legal status stratification and health. *SSM Popul Health.* 2016;2:141-148.
36. Hall M, Greenman E. The Occupational Cost of Being Illegal in the United States: Legal Status, Job Hazards, and Compensating Differentials. *Int Migr Rev.* 2015;49(2):406-442.
37. Flippen CA. Laboring Underground: The Employment Patterns of Hispanic Immigrant Men in Durham, NC. *Soc Probl.* 2012;59(1):21-42.
38. Ortega AN, Fang H, Perez VH, et al. Health care access, use of services, and experiences among undocumented Mexicans and other Latinos. *Archives of internal medicine.* 2007;167(21):2354-2360.
39. Valdivia C. Expanding geographies of deportability: how immigration enforcement at the local level affects undocumented and mixed-status families. *Law & Policy.* 2019;41(1):103-119.
40. Hilfinger Messias DK, McEwen MM, Clark L. The impact and implications of undocumented immigration on individual and collective health in the United States. *Nursing outlook.* 2015;63(1):86-94.
41. Marshall KJ, Urrutia-Rojas X, Mas FS, Coggin C. Health status and access to health care of documented and undocumented immigrant Latino women. *Health care for women international.* 2005;26(10):916-936.
42. DeVoe JE, Wallace LS, Pandhi N, Solotaroff R, Fryer GE, Jr. Comprehending care in a medical home: a usual source of care and patient perceptions about healthcare communication. *Journal of the American Board of Family Medicine : JABFM.* 2008;21(5):441-450.

43. Berk ML, Schur CL, Chavez LR, Frankel M. Health care use among undocumented Latino immigrants. *Health affairs (Project Hope).* 2000;19(4):51-64.
44. Heyman JM, Nunez GG, Talavera V. Healthcare access and barriers for unauthorized immigrants in El Paso County, Texas. *Family & community health.* 2009;32(1):4-21.
45. Holman H, Lorig K. Patients as partners in managing chronic disease. Partnership is a prerequisite for effective and efficient health care. *Bmj.* 2000;320(7234):526-527.
46. Do EK, Matsuyama RK. Healthcare utilization among Hispanic immigrants with diabetes: investigating the effect of US documentation status. *Journal of immigrant and minority health.* 2014;16(2):189-194.
47. Killilea T. Long-term consequences of type 2 diabetes mellitus: economic impact on society and managed care. *Am J Manag Care.* 2002;8(16 Suppl):S441-449.
48. Peters MD, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. *Int J Evid Based Healthc.* 2015;13(3):141-146.
49. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. *Implement Sci.* 2010;5:69.
50. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Annals of internal medicine.* 2018;169(7):467-473.
51. Peters MD. In no uncertain terms: the importance of a defined objective in scoping reviews. *JBI Database System Rev Implement Rep.* 2016;14(2):1-4.
52. Coritsidis GN, Khamash H, Ahmed SI, et al. The initiation of dialysis in undocumented aliens: the impact on a public hospital system. *Am J Kidney Dis.* 2004;43(3):424-432.
53. Hamilton ER, Hale JM, Savinar R. Immigrant Legal Status and Health: Legal Status Disparities in Chronic Conditions and Musculoskeletal Pain Among Mexican-Born Farm Workers in the United States. *Demography.* 2019;56(1):1-24.
54. Iten AE, Jacobs EA, Lahiff M, Fernandez A. Undocumented immigration status and diabetes care among Mexican immigrants in two immigration "sanctuary" areas. *Journal of immigrant and minority health.* 2014;16(2):229-238.
55. Linden EA, Cano J, Coritsidis GN. Kidney transplantation in undocumented immigrants with ESRD: a policy whose time has come? *American journal of kidney diseases : the official journal of the National Kidney Foundation.* 2012;60(3):354-359.
56. Madden EF, Qeadan F. Dialysis Hospitalization Inequities by Hispanic Ethnicity and Immigration Status. *Journal of health care for the poor and underserved.* 2017;28(4):1509-1521.
57. Raghavan R, Sheikh-Hamad D. Descriptive analysis of undocumented residents with ESRD in a public hospital system. *Dialysis and Transplantation.* 2011;40(2):78-81.
58. Ruhnke SA, Reynolds MM, Wilson FA, Stimpson JP. A healthy migrant effect? Estimating health outcomes of the undocumented immigrant population in the United States using machine learning. *Soc Sci Med.* 2022;307:115177.
59. Shen JI, Hercz D, Barba LM, et al. Association of Citizenship Status With Kidney Transplantation in Medicaid Patients. *American journal of kidney diseases : the official journal of the National Kidney Foundation.* 2018;71(2):182-190.
60. Riosmena F, Kuhn R, Jochem WC. Explaining the Immigrant Health Advantage: Self-selection and Protection in Health-Related Factors Among Five Major National-Origin Immigrant Groups in the United States. *Demography.* 2017;54(1):175-200.
61. Lopez G, Radford J. *2015, Foreign-born population in the United States statistical portrait.* Pew Research Hispanic Trends;2017.
62. Riosmena F, Everett BG, Rogers RG, Dennis JA. Negative Acculturation and Nothing More? Cumulative Disadvantage and Mortality during the Immigrant Adaptation Process among Latinos in the United States. *Int Migr Rev.*49(2):443-478.

63. Gagnier JJ, Kienle G, Altman DG, Moher D, Sox H, Riley D. The CARE Guidelines: Consensus-based Clinical Case Reporting Guideline Development. *Glob Adv Health Med.* 2013;2(5):38-43.
64. Balakrishnan A, Jordan N. The Undocumented Elderly: Coverage Gaps and Low Health Care Use. *J Health Care Poor Underserved.* 2019;30(3):891-898.
65. Bohan JN. Illegal immigrants with ESRD: The impact on U.S. and Canadian health care facilities. *Nephrol News Issues.* 2005;19(4):27-28, 30, 71.
66. Cantu PA, Hayward MD, Hummer RA, Chiu CT. New estimates of racial/ethnic differences in life expectancy with chronic morbidity and functional loss: evidence from the National Health Interview Survey. *J Cross Cult Gerontol.* 2013;28(3):283-297.
67. Carrillo JE, Shekhani NS, Deland EL, et al. A regional health collaborative formed By NewYork-Presbyterian aims to improve the health of a largely Hispanic community. *Health Aff (Millwood).* 2011;30(10):1955-1964.
68. Cervantes L, Mundo W, Powe NR. The Status of Provision of Standard Outpatient Dialysis for US Undocumented Immigrants with ESKD. *Clin J Am Soc Nephrol.* 2019;14(8):1258-1260.
69. Cervantes L, O'Hare A, Chonchol M, et al. Circumstances of Death among Undocumented Immigrants Who Rely on Emergency-Only Hemodialysis. *Clin J Am Soc Nephrol.* 2018;13(9):1405-1406.
70. Cervantes L, Tuot D, Raghavan R, et al. Association of Emergency-Only vs Standard Hemodialysis With Mortality and Health Care Use Among Undocumented Immigrants With End-stage Renal Disease. *JAMA Intern Med.* 2018;178(2):188-195.
71. Diaz-Perez Mde J, Farley T, Cabanis CM. A program to improve access to health care among Mexican immigrants in rural Colorado. *J Rural Health.* 2004;20(3):258-264.
72. Garcini LM, Renzaho AMN, Molina M, Ayala GX. Health-related quality of life among Mexican-origin Latinos: the role of immigration legal status. *Ethn Health.* 2018;23(5):566-581.
73. Hogan AN, Fox WR, Roppolo LP, Suter RE. Emergent Dialysis and its Impact on Quality of Life in Undocumented Patients with End-Stage Renal Disease. *Ethn Dis.* 2017;27(1):39-44.
74. Hsueh L, Vrany EA, Patel JS, et al. Associations between immigrant status and pharmacological treatments for diabetes in U.S. adults. *Health Psychol.* 2018;37(1):61-69.
75. Kang YJ, McCormick D, Zallman L. Affordability of and Access to Information About Health Insurance Among Immigrant and Non-immigrant Residents After Massachusetts Health Reform. *J Immigr Minor Health.* 2017;19(4):929-938.
76. LEE H, ZAKHARY BL, FIREK MA, LANGE H, GOMEZ CA, FIREK A. The Prevalence and Impact of Diabetes Mellitus among Undocumented Immigrants in an Indigent Care Program in Riverside, California. *Diabetes.* 2018;67(Supplement_1).
77. Ortega AN. When Politics Trumps Health: Undocumented Latino Immigrants and US Health Care. *MEDICC Rev.* 2015;17(4):59.
78. Peota C. CARE IN THE SHADOWS: THE CHALLENGES FOR UNDOCUMENTED IMMIGRANTS. *Minn Med.* 2017;100(1):12-15.
79. Ross J, Hua S, Perreira KM, et al. Association between immigration status and anxiety, depression, and use of anxiolytic and antidepressant medications in the Hispanic Community Health Study/Study of Latinos. *Ann Epidemiol.* 2019;37:17-23.e13.
80. Sher SJ, Aftab W, Moorthi RN, et al. Healthcare outcomes in undocumented immigrants undergoing two emergency dialysis approaches *Clin Nephrol.* 2017;88(10):181-192.
81. Thomsen M. Caring for the Uninsured, Undocumented Population: Challenges of Chronic Disease Management and Access to Care. *Prof Case Manag.* 2015;20(6):330-334.
82. Migration Policy Institute. Profile of the Unauthorized Population: United States. https://www.migrationpolicy.org/data/unauthorized-immigrant-population/state/US. Accessed March 29, 2020.

83. Carvajal S, Kibor C, McClelland DJ, et al. Stress and Sociocultural Factors Related to Health Status Among US-Mexico Border Farmworkers. *Journal of Immigrant Minority Health.* 2014;16(6):1176-1182.
84. Carvajal S, Rosales C, Rubio-Goldsmith R, et al. The Border Community and Immigration Stress Scale: A Preliminary Examination of a Community Responsive Measure in Two Southwest Samples. *Journal of Immigrant Minority Health.* 2013;15(2):427-436.
85. Capps R, Fix M, Hook JV, Bachmeier JD. *A Demographic, Socioeconomic, and Health Coverage Profile of Unauthorized Immigrants in the United States.* Washington, DC: Migration Policy Institute;2013.
86. Capps R, Bachmeier JD, Van Hook J. Estimating the Characteristics of Unauthorized Immigrants Using U.S. Census Data: Combined Sample Multiple Imputation. *The ANNALS of the American Academy of Political and Social Science.* 2018;677(1):165-179.
87. Rendall MS, Ghosh-Dastidar B, Weden MM, Baker EH, Nazarov Z. Multiple Imputation For Combined-Survey Estimation With Incomplete Regressors In One But Not Both Surveys. *Sociol Methods Res.* 2013;42(4).
88. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med.* 2011;30(4):377-399.
89. Breiman L. Random Forests. *Machine Learning.* 2001;45:5-32.
90. Liu Y, Wang Y, Zhang J. New Machine Learning Algorithm: Random Forest. In: Liu B, Ma M, Chang J, eds. *Information Computing and Applications.* Vol 7473. Berlin: Springer, Berlin, Heidelberg; 2012.
91. Kuhn M, Johnson K. *Applied Predictive Modeling.* Vol 1: Springer; 2013.
92. C. C, V. V. Suport-Vector Networks. *Machine Learning.* 1995;20:273-297.
93. Claesen C., De Smet F., Suykens JAK., De Moor B. A robust ensemble approach to learn from positive and unlabeled data using SVM basse models. *Nurocomputing.* 2014.
94. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med.* 2015;162(1):W1-73.
95. J. G. *Psychometric Methods.* 2nd ed. New York: McGraw-Hill; 1954.
96. Chicco D., Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics.* 2020;21.
97. *R: A language and environment for statistical computing* [computer program]. Vienna, Austria: R Foundation for Statistical Computing; 2021.
98. Steyerberg EW, Harrell FE, Jr. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol.* 2016;69:245-247.
99. Centers for Disease Control and Prevention. COVID Data Tracker. New Admissions of Patients with Confirmed COVID-19, United States, August 01, 2020 - March 07, 2022 Web site. https://covid.cdc.gov/covid-data-tracker/#new-hospital-admissions. Published 2022. Accessed March 9th, 2022.
100. Iuliano AD, Brunkard JM, Boehmer TK, et al. Trends in Disease Severity and Health Care Utilization During the Early Omicron Variant Period Compared with Previous SARS-CoV-2 High Transmission Periods - United States, December 2020-January 2022. *MMWR Morb Mortal Wkly Rep.* 2022;71(4):146-152.
101. Ward IL, Bermingham C, Ayoubkhani D, et al. Risk of covid-19 related deaths for SARS-CoV-2 omicron (B.1.1.529) compared with delta (B.1.617.2): retrospective cohort study. *Bmj.* 2022;378:e070695.
102. Ejaz H, Alsrhani A, Zafar A, et al. COVID-19 and comorbidities: Deleterious impact on infected patients. *J Infect Public Health.* 2020;13(12):1833-1839.

103. Cheng YJ, Kanaya AM, Araneta MRG, et al. Prevalence of Diabetes by Race and Ethnicity in the United States, 2011-2016. *JAMA.* 2019;322(24):2389-2398.
104. Kaiser Family Foundation. Health Coverage and Care of Immigrants. https://www.kff.org/racial-equity-and-health-policy/fact-sheet/health-coverage-and-care-of-immigrants/. Published 2022. Accessed.
105. Page KR, Venkataramani M, Beyrer C, Polk S. Undocumented U.S. Immigrants and Covid-19. *N Engl J Med.* 2020;382(21):e62.
106. Cervantes L, Martin M, Frank MG, et al. Experiences of Latinx Individuals Hospitalized for COVID-19: A Qualitative Study. *JAMA Netw Open.* 2021;4(3):e210684.
107. Huang BZ, Creekmur B, Yoo MS, Broder B, Subject C, Sharp AL. Healthcare Utilization Among Patients Diagnosed with COVID-19 in a Large Integrated Health System. *J Gen Intern Med.* 2022;37(4):830-837.
108. Czeisler M, Marynak K, Clarke KEN, et al. Delay or Avoidance of Medical Care Because of COVID-19-Related Concerns - United States, June 2020. *MMWR Morb Mortal Wkly Rep.* 2020;69(36):1250-1257.
109. Soto Mas F, Nervi L, Rosero DV, Sebastian R, Guldan S, Casanova V. COVID-19 and essential workers: Healthcare delays among organic farmers. *J Healthc Qual Res.* 2022;37(5):283-290.
110. Hayward SE, Deal A, Cheng C, et al. Clinical outcomes and risk factors for COVID-19 among migrant populations in high-income countries: A systematic review. *J Migr Health.* 2021;3:100041.
111. Strully K, Yang TC, Liu H. Regional variation in COVID-19 disparities: connections with immigrant and Latinx communities in U.S. counties. *Ann Epidemiol.* 2021;53:56-62.e52.
112. Kim HN, Lan KF, Nkyekyer E, et al. Assessment of Disparities in COVID-19 Testing and Infection Across Language Groups in Seattle, Washington. *JAMA Netw Open.* 2020;3(9):e2021213.
113. Perez MC, Fortuna L. Psychosocial stressors, psychiatric diagnoses and utilization of mental health services among undocumented immigrant Latinos. In: *Mental Health Care for New Hispanic Immigrants.* Routledge; 2013:107-124.
114. Ornelas IJ, Yamanis TJ, Ruiz RA. The Health of Undocumented Latinx Immigrants: What We Know and Future Directions. *Annu Rev Public Health.* 2020;41:289-308.
115. Ortega AN, McKenna RM, Kemmick Pintor J, et al. Health Care Access and Physical and Behavioral Health Among Undocumented Latinos in California. *Med Care.* 2018;56(11):919-926.
116. Vargas Bustamante A, Fang H, Garza J, et al. Variations in healthcare access and utilization among Mexican immigrants: the role of documentation status. *J Immigr Minor Health.* 2012;14(1):146-155.
117. Galvan T, Lill S, Garcini LM. Another Brick in the Wall: Healthcare Access Difficulties and Their Implications for Undocumented Latino/a Immigrants. *J Immigr Minor Health.* 2021;23(5):885-894.
118. Lutrick K, Ellingson KD, Baccam Z, et al. COVID-19 Infection, Reinfection, and Vaccine Effectiveness in a Prospective Cohort of Arizona Frontline/Essential Workers: The AZ HEROES Research Protocol. *JMIR Res Protoc.* 2021;10(6).
119. Thompson MG, Yoon SK, Naleway AL, et al. Association of mRNA Vaccination With Clinical and Virologic Features of COVID-19 Among US Essential and Frontline Workers. *Jama.* 2022;328(15):1523-1533.
120. Centers for Disease Control and Prevention. COVID Data Tracker: Variant Proportions. https://covid.cdc.gov/covid-data-tracker/#variant-proportions. Published 2023. Accessed February 16, 2023.
121. Cortes C, Vapnik V. Support-vector networks. *Machine Learning.* 1995;20(3):273-297.
122. Cameron AC, Trivedi PK. *Microeconomics using Stata.* College Station, TX: Stata Press; 2009.

123. Pitlik SD. COVID-19 Compared to Other Pandemic Diseases. *Rambam Maimonides Med J.* 2020;11(3).
124. Zheng Q, Ma P, Wang M, et al. Efficacy and safety of Paxlovid for COVID-19:a meta-analysis. *J Infect.* 2023;86(1):66-117.