Ethics of artificial intelligence and machine learning

Kevin Larsen, MD, FACP

SVP Clinical Innovation and Translation, OptumLabs

OPTUM Labs®

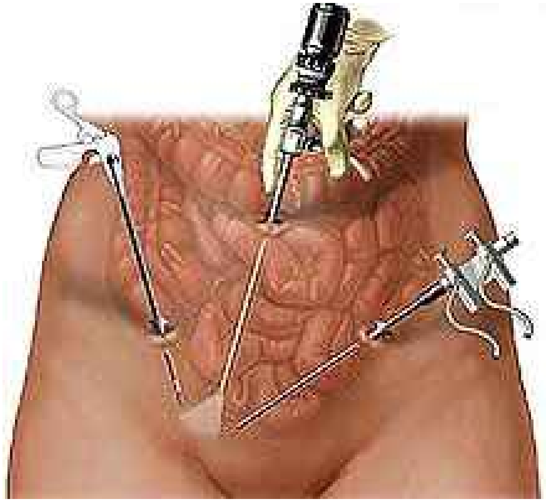
Disclosures

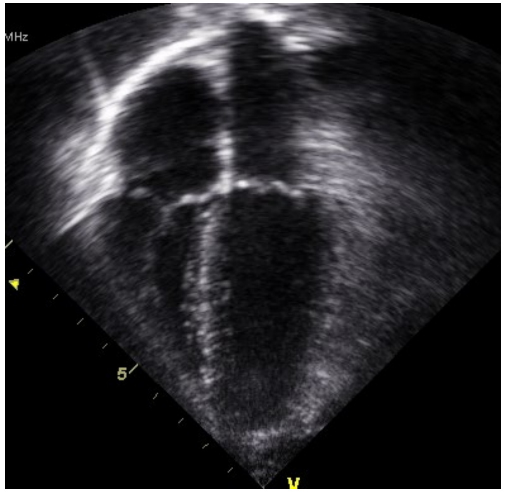
Financial Disclosures:

- None

Machine learning/AI is a powerful clinical tool

Powerful clinical tools require extensive training and critical thinking to be safe and maximally effective





Machine learning/AI is a powerful clinical tool

If our jobs were simple, we could use simple tools

Our jobs are complex

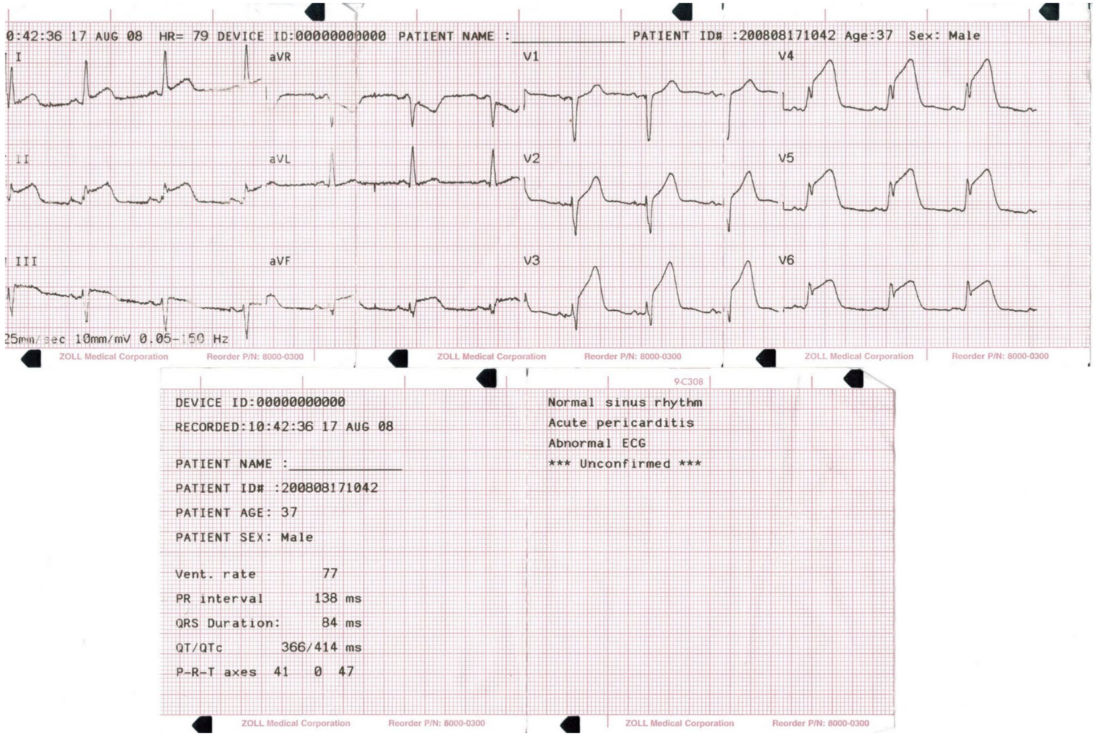
They require sophisticated, complex tools

Sophisticated tools require research, testing and critical analysis

Like other medical technologies, these must be designed, run and trained by clinicians

Vision of the future: computer-assisted flying

Machine learning algorithms



Computer assisted radiology reads

Computer assisted EKG reads

Early sepsis identification algorithms

Natural Language Processing

Complex clinical risk scores

What do these have in common?

They were empirically derived using statistical analysis – like regression, correlation and statistical significance

However, the volume of data and the number of variables analyzed is orders of magnitude greater than in a simple regression

They often use a gold standard of “human read” or “human interpretation”

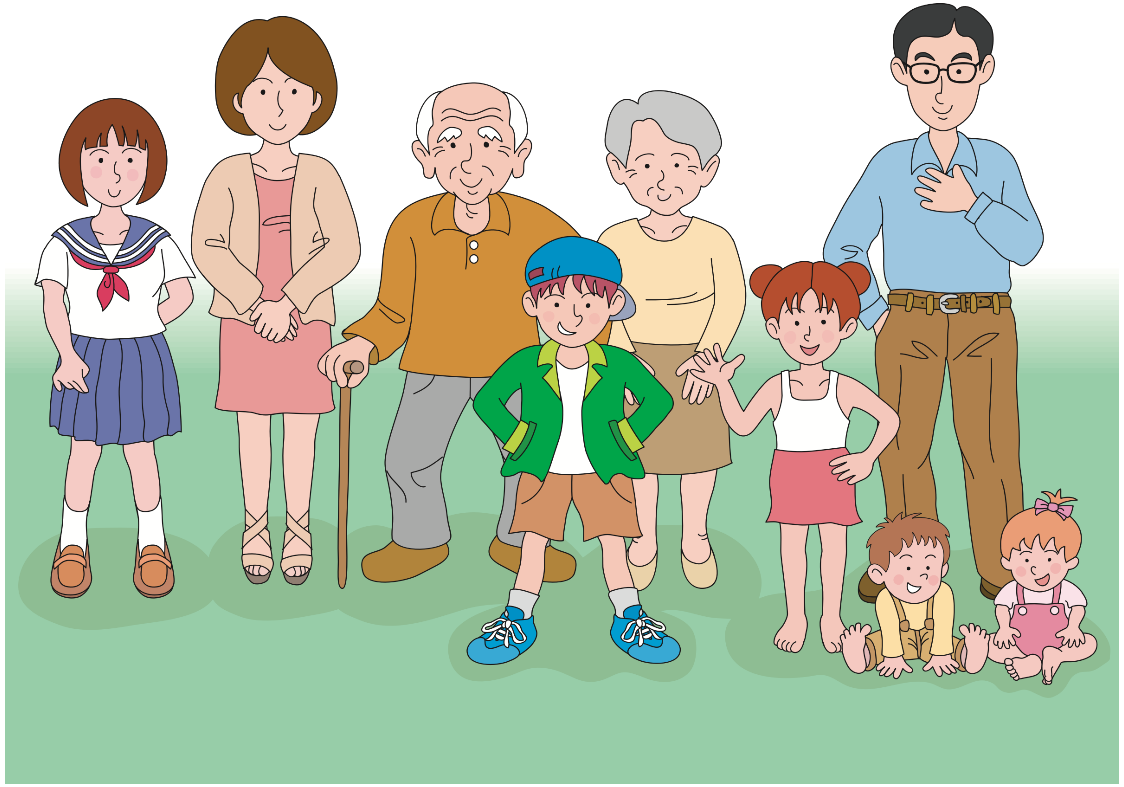
Human decision making is flawed

There are many ways that the human gold standard has systemic bias

These systematic biases are then “learned” by the algorithm and become “invisible”

Generalizability

As with any study, we look to generalize from the AI model



For example- if we use Medicare data (over 65-year-olds) to build our model- can we generalize this to:

50-year-olds?

40-year-olds?

20-year-olds?

Non- Americans?

Infants?

Generalizability

The field is full of examples of training on sample data not representative of the whole population

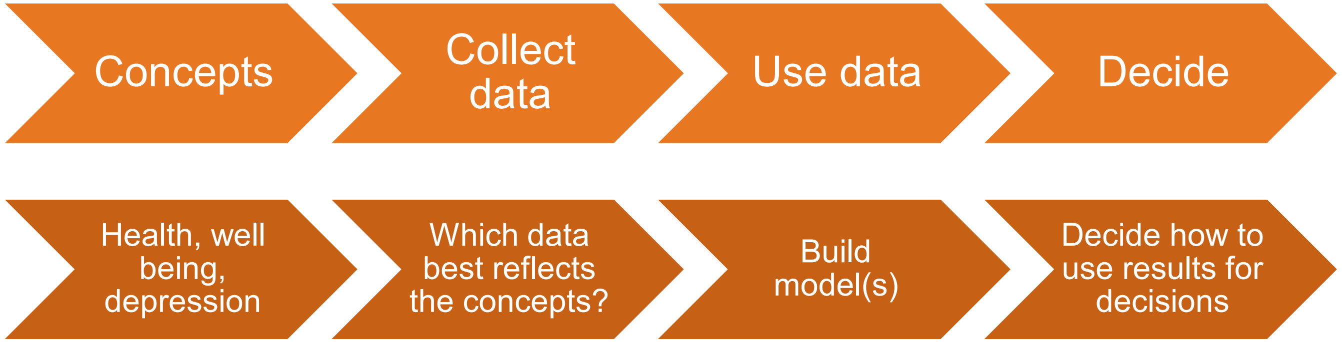
- Data on only men
- Data on only people with insurance
- Data on only people at an academic referral center
- Data on only white people

The Decision Process Framework

Goal → Improve a decision making process using data and models

These decisions are about an action regarding a human

- Approve for college admission, approve for loan, outreach for a health intervention, lower prison sentence, …
- The decision to approve conveys a benefit
- Denial is harmful or removes an opportunity to benefit



Steve Few - ethical principle of data analysis

The ethical practices that can serve as a code of conduct for data sensemaking professionals are, in my opinion, built upon a single fundamental principle. It is the same principle that medical doctors swear as an oath before becoming licensed: Do no harm.

http://www.stephen-few.com/blog/2019/07/22/ethical-data-sensemaking/

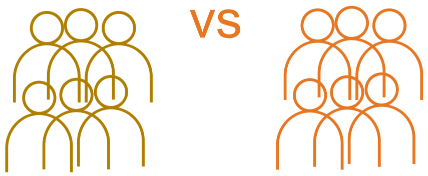
Equity: one example framework

Equality is having the same result, Equity is the process that treats everyone justly according to their circumstances*

Individual Parity

Do two people with same data get same decision?

Demographic parity

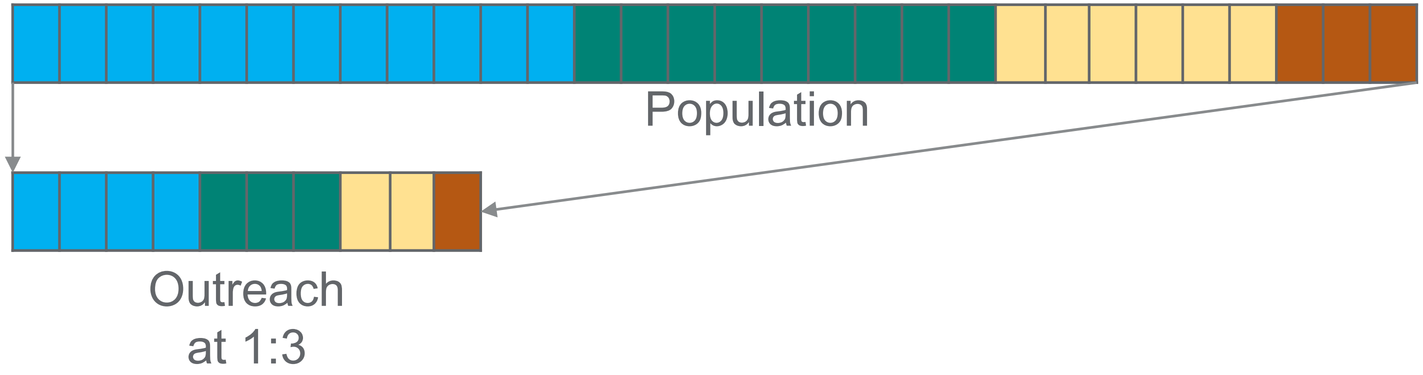


Do two groups have same approval rates?

https://publichealthonline.gwu.edu/blog/equity-vs-equality/

We have 1000 offers to make

Demographic parity says make offers to each protected class in proportion to their population proportion.



We need to evaluate who benefits and is harmed

- Are more benefits/harms going to or or … ?

Naïve machine learning reinforces past practices

ProPublica argues the COSMOS recidivism model is biased

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Fig2: The bias in COMPAS. (from Larson et al. ProPublica, 2016)

The mistakes the model makes are different by race, even though race is not in the model.

- The questionnaire collects data that is correlated to race
- Recidivism = a jail booking, not a conviction.
- Jail bookings are most correlated with where police are, not crime

These findings are controversial but are worthy of discussion and understanding

Benefits and harms are necessary to decide!

Benefits:

- People who should get approved, do get approved.

  People who can pay back the loan, get offered the loan

- This is true positive rate (aka recall, sensitivity)

  True positive = do get approved, All positives = should be approved

- Undeserving do not get offers (preserve capacity for deserving)

Harms are more complex

- Is there a harm from a false positive?

  Not paying back the loan limits ability to make new loans

- Add harm from not getting the benefit (false negatives)

  A person is harmed by not getting the loan, not getting an outreach call.

Individual Parity is helped by modeling

Individual Parity means two people with same/similar data get same/similar decision.

Lending example: Bob and Fred are loan officers. Due to their diverse backgrounds, they may reach different conclusions on loan approvals.

By processes and training, the bank limits the variation in their approvals.

Does a predictive model guarantee individual parity?

Model-based Individual Parity is not guaranteed

All models will create the same score for same data*, but the data may differ by unobservables

Two opportunities to violate parity:

- Is the data collection guaranteed to be correct ?
- Does the post-model process to a decision preserve the score parity?

Cynthia Dwork coined *Individual Fairness* to mean that two people with "similar" inputs get "similar" scores

- It's hard to determine the right way to decide what similar means

- Assuming model has no randomized scoring
- Dwork, https://arxiv.org/pdf/1104.3913.pdf

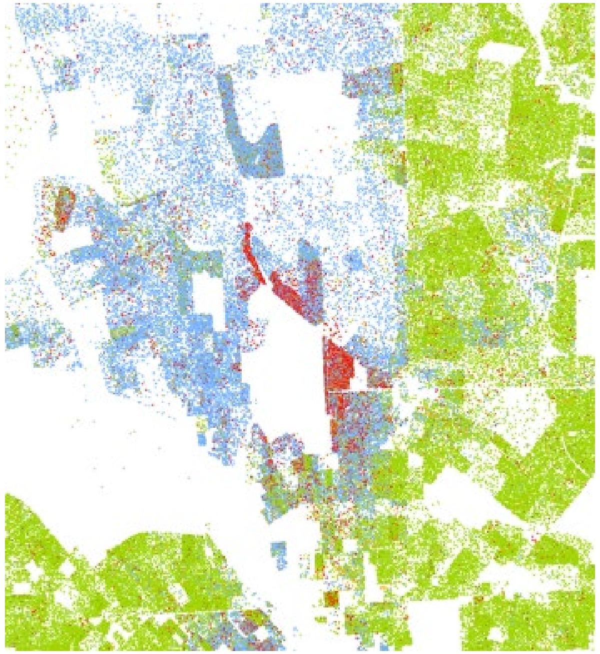
Systemic differences in data collection

Academic Testing

- "Runner is to marathon as oarsman is to regatta"[1]
- Runner is to marathon as drummer is to band?

Loan Redlining as historical bias

- From Racial Dot Map, and Baltimore Sun



Recidivism

- Re-arrests but not convictions
- How many of your friends/acquaintances are taking drugs illegally?

1 From Methods for Identifying Biased Test Items
By Gregory Camilli, Lorrie A. Shepard, Lorrie Shepard

2 https://www.baltimoresun.com/opinion/readers-respond/bs-ed-rr-housing-discrimination-letter-20190213-story.html

3 J. Angwin, J. Larson, S. Mattu, L. Kirchner, "Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks," *ProPublica*, 23 May 2016;

FICO scores are race blind

Omitting race from models does not imply equity!

“Fairness through unawareness” is a fallacy.

The data collection process is almost never unaware of race.

Questions:

- Can the decision process after the scoring overcome issues in the scores?
- What is our obligation as creators of the scores to mitigate and recognize potential issues and misuse?

A path forward

Be aware of possible sources of bias

- benefit design
- supply of healthcare
- is lack of data an indication of health, or inability to get care?

Measure disparities - in inputs and predictions

- Compare to literature review

Consider the harm of the model

- Not just accuracy
- Who gets left out and at what consequence?

Can you engineer better features, or less biased ones?

Example in large data sets

A model designed for marketing to estimate someone's race/ethnicity is used in clinical research

imputed race uses last name and zip code as part of the model

Imputation downside

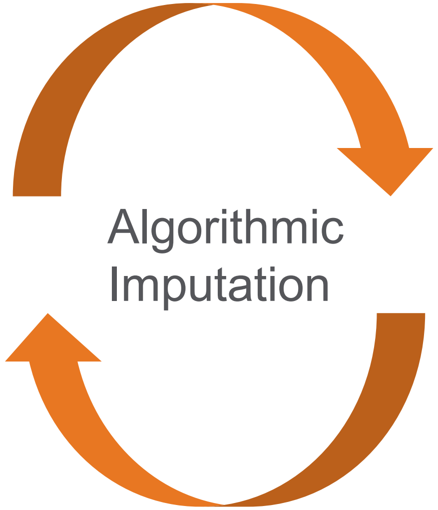
Name: Rosa Williams nee Sanchez

Race: White

Ethnicity: Hispanic

Preferred Language: Spanish

Zip Code: 90305 - Inglewood, CA



Imputed analysis

Race: Black
(Williams in this zip is correlated with Black race)

Ethnicity: non-Hispanic

Preferred language: English

This will affect how we record Rosa within our data and potentially affect how we interact with Rosa.

Ethics questions

- How does self reported race correlate to the imputed race?
- We compared a subset with self reported race from CMS to the company that supplied the marketing imputed race

- For independent analysis of CMS race data see https://journals.lww.com/lww-medicalcare/Fulltext/2020/01000/Validity_of_Race_and_Ethnicity_Codes_in_Medicare.16.aspx

In Western US, only 25% of self reported blacks identified as black by imputed race

Sociodemographic Characteristics from CMS Enrollment Files	Sample Size		Agreement Stratified by CMS Race and Ethnicity				Lack of Agreement: Non-White by CMS, White marketing race		
	N	%	White	Black	Hispanic	Asian	Black	Hispanic	Asian
Census region									
Northeast	611,488	18.5	88.7	63.2	90.1	87.5	31.6	5.9	6.2
Midwest	835,898	25.3	91.9	71.2	91.6	77.9	26.4	5.5	13.2
South	1,527,792	46.3	84.8	72.3	93.1	77.0	24.6	4.1	10.8
West	324,114	9.8	92.0	27.8	92.8	83.2	67.6	4.3	9.2

First shaded column is % agreement for people that self reported as black. It’s never above 75% and very low, <30%, in the West.

Second shaded column is % errors for non-white people.

Similar findings when analyzed imputed vs EHR data

Why does this matter?

Example 1

- Defensible empiric logic
- Regression analyses of large data sets
- Race/ethnicity correlates with an outcome of interest

But if race does appear to correlate with clinical outcomes, does that justify its inclusion in diagnostic or predictive tools?

Medical Algorithms Have a Race Problem

Certain lab tests provide one result if a patient is Black, another if they're white. But debate over 'race adjustments' is heating up.

By Kaveh Waddell
Last updated: September 18, 2020

"No one is saying to throw away science. We just want to make sure that we are not causing harm to our patients."

NWAMAKA ENEANYA
Nephrologist and assistant professor at the University of Pennsylvania

Example 2: eGFR- race “correction”

Impacts:

- Referral to a nephrologist
- Placement on transplant waiting list
- Dosing of medications

Racial disparities:

- End-stage kidney disease
- Death due to kidney failure
- Longer wait times for kidney transplant

VIEWPOINT

Black Kidney Function Matters
Use or Misuse of Race?

Neil R. Powe, MD, MPH, MBA
Priscilla Chan and Mark Zuckerberg San Francisco General Hospital, University of California, San Francisco.

Author Audio Interview

Supplemental content

Racial discrimination has been a lightning rod for passionate discourse and social action in the US for decades, if not centuries. The recent killings of African Americans by law enforcement has amplified the discourse. Health care has not been immune to such tragedies, with past experimentation without informed consent and segregation in health care facilities. These were systemically ingrained, institutional practices without ethical or evidentiary footing. Race was an identifying characteristic used to implement practices that resulted in consequences for health and well-being. The use of race in algorithms for clinical care, including for kidney disease, has generated and now even more so is generating discourse and action about current-day, systemic discrimination in health care.

A number of institutions have taken steps to remove the use of race in equations involving estimated glomerular filtration rates (eGFRs). In 2017, the Beth Israel Deaconess Medical Center discarded race from reporting of eGFR in laboratory reports after concerns from medical students and inclusive vetting by clinical leaders and administrators. In 2019, Zuckerberg San Francisco General Hospital moved to substitute muscle mass for race in reporting eGFR after a small group of faculty and trainees lobbied the clinical laboratory. In 2020, the University of Washington, Brigham and Women's Hospital, Massachusetts General Hospital, and Vanderbilt removed race from eGFR reporting. Social media are now flooded with requests by aspiring and established medi-

There will be continued tension about whether the use of race in medicine constitutes misuse.

equation, developed in 2009 (estimates function as glomerular filtration rate in milliliters/minute/1.73 m²).[1] The latter 2 do not include weight but incorporate a coefficient that reflects that measured glomerular filtration rate was 21% or 16% greater in Black participants in the MDRD and CKD-EPI research studies, respectively, and afford greater precision in estimating kidney function. The application of these coefficients based on race is causing great consternation and appeals to expunge them from eGFR and clinical reporting.[2] Black kidney function matters because Black adults in the US are nearly 3 times more likely to develop end-stage kidney failure, and on average 5 years earlier than White adults.[3]

Appreciating contrasting views on the imprecise concept of race is fundamental to understanding the controversy on race in eGFR reporting. Race, a concept invented by humans, was first used to group people with certain observable physical characteristics, such as skin color or facial features, who evolved from different geographies in the world. It changed to be associated with people's self-identities that include customs and ways of life, factors that are cultural and social. It is also an unclear concept because classification is self-identified and can be wrongly assigned by others. Genomics shows that ancestry is more informative than race when biology is examined. In 26 studies that pooled data that included a gold standard of directly measured glomerular filtration rate among 8254 participants for derivation and 3896 participants for validation, a signal was discovered that distinguished people who self-reported their race as Black compared with other races.[4] The equations derived from evidence are recommended in international guidelines and used worldwide.

“We need to slow down as a community of physicians to figure out how best to do this.”

NEIL POWE

Chief of medicine, Zuckerberg San Francisco General Hospital

Hospital, 1001 Potrero Ave, Room 5F40, San Francisco, CA 94110 (neil.powe@ucsf.edu).

Cockcroft-Gault equation, developed in 1973 (estimates function as creatinine clearance in milliliters/minute); the MDRD equation, developed in 1999; and the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI)

dosing of medications (eg, analgesics, opioids, metformin, angiotensin-converting enzyme inhibitors, chemotherapy) to considering kidney donation and research study participation for new therapies, any of which might exacerbate

Example 3: VBAC algorithm with race

Impacts:

- Likelihood of TOLAC
  - Surgical complications
  - Recovery time
  - Subsequent pregnancy complications
- Marital status and insurance type

Racial disparities:

- C-section rates
- Maternal mortality rates

VAGINAL BIRTH AFTER CESAREAN Height & weight optional; enter them to automatically calculate BMI	
Maternal age	18 years
Height (range 54-80 in.)	in
Weight (range 80-310 lb.)	lb
Body mass index (BMI, range 15-75)	25 kg/m^2
African-American?	no
Hispanic?	no
Any previous vaginal delivery?	no
Any vaginal delivery since last cesarean?	no
Indication for prior cesarean of arrest of dilation or descent?	no
Calculate	

A new calculator without race and ethnicity is under development.

This calculator is based on the equation published in the article "Development of a nomogram for prediction of vaginal birth after cesarean" cited below. It is designed for educational use and is based on a population of women who received care at the hospitals within the MFMU Network. Responsibility for its correct application is accepted by the end user.

Grobman WA, Lai Y, Landon MB, Spong CY, Leveno KJ, Rouse DJ, Varner MW, Moawad AH, Caritis SN, Harper M, Wapner RJ, Sorokin Y, Miodovnik M, Carpenter M, O'Sullivan MJ, Sibai BM, Langer O, Thorp JM, Ramin SM, Mercer BM; National Institute of Child Health and Human Development (NICHD) Maternal-Fetal Medicine Units Network (MFMU), "Development of a nomogram for prediction of vaginal birth after cesarean delivery," Obstetrics and Gynecology, volume 109, pages 806-12, 2007.