

We can't embrace machine learning in healthcare without 'serious vetting' of the underlying data

Given that machine learning in the health domain can have a direct impact on people's lives, broad claims emerging from this kind of research should not be embraced without serious vetting.

...

One key question to ask is: Whose information is in the data and what do these data reflect?

Common forms of electronic health data, such as billing claims and clinical records, contain information only on individuals who have encounters with the [health care system](#). But many individuals who are sick don't — or can't — see a doctor or other health care provider and so are invisible in these databases. This may be true for individuals with lower incomes.

...

As a [ProPublica report](#) demonstrated, black and Native American patients are drastically underrepresented in cancer clinical trials. This is important to underscore given that randomized trials are frequently highlighted as superior in discussions about machine learning work that leverages nonrandomized electronic health data.

In interpreting results from machine learning research, it's important to be aware that the patients in a study often do not depict the population we wish to make conclusions about and that the information collected is far from complete.

Read full, original post: [Machine learning for clinical decision-making: pay attention to what you don't see](#)