**An online tool for correcting verification bias when validating electronic phenotyping
algorithms.**
Ajay Bhasin, MD^{1,2}; Suzette J. Bielinski, PhD, MEd³ Abel N. Kho, MD^{4,5}; Nicholas B. Lar.
PhD, MS^{*6}; Laura Rasmuss

- Ajay Bhasin, MD^{1,2}; Suzette J. Bielinski, PhD, MEd³ Abel N. Kho, MD^{4,5}
- PhD, MS *6 ; Laura Rasmussen-Torvik, MPH, PhD *7
- ¹Department of Medicine, Division of Hospital Medicine, Northwestern University Feinberg
- ²Department of Pediatrics, Division of Hospital-Based Medicine, Northwestern University
-
- 2 **algorithms.**
3 Ajay Bhasin
4 PhD, MS^{*6};
¹Departmen
6 School of M
7 ²Departmen
Feinberg Sc
³Division of Medicine an Ajay Bhasin, MD^{1,2}; Suzette J. Bielinski, PhD, MEd³ Abel N. Kho, MD^{4,3}; Nicholas B. Larson

PhD, MS^{*6}; Laura Rasmussen-Torvik, MPH, PhD^{*7}

¹Department of Medicine, Division of Hospital Medicine, Northwestern U 56789012 ¹Department of Medicine, Division of Hospital Medicine, Northwestern University Feinberg

School of Medicine, Chicago, IL

²Department of Pediatrics, Division of Hospital-Based Medicine, Northwestern University

Feinbe 6 School of Medicine, Chicago, IL

² Department of Pediatrics, Divisio

³ Teinberg School of Medicine, Ch

³ Division of Epidemiology, Depa

³ Medicine and Science, Rochester

⁴ Center for Health Information Pa
 ² Department of Pediatrics, Division of Hospital-Based Medicine, Northwestern University

³ Division of Epidemiology, Department of Quantitative Health Sciences, Mayo Clinic Coll

Medicine and Science, Rochester, Minne Feinberg School of Medicine, Chicago, IL

³Division of Epidemiology, Department of

Medicine and Science, Rochester, Minnesc

⁴Center for Health Information Partnership

School of Medicine, Northwestern Univers

⁵Div ³Division of Epidemiology, Department of Quantitative Health Sciences, Mayo Clinic College of ⁹ Division of Epidemiology, Department of Quantitative Health Sciences, Mayo Clinic College of

1 ⁴ Center for Health Information Partnerships, Institute for Public Health & Medicine, Feinberg

5 School of Medicine, No
- ⁴ Center for Health Information Partnerships, Institute for Public Health & Medicine, Feinberg
-
- ⁵Division of General Internal Medicine, Department of Medicine, Feinberg School of Medicine,
-
- Medicine and Science, Rochester, Minnesota, USA

⁴ Center for Health Information Partnerships, Institut

School of Medicine, Northwestern University, Chica

⁵ Division of General Internal Medicine, Department

⁵ Nort ⁴ Center for Health Information Partnerships, Institute for Public Health & Medicine, Feinberg

5 School of Medicine, Northwestern University, Chicago, IL, USA.

⁵ Division of General Internal Medicine, Department of M School of Medicine, Northwestern University, Chicago, IL, USA.

⁵ Division of General Internal Medicine, Department of Medicine, Northwestern University, Chicago, IL, USA.

⁶ Division of Clinical Trials and Biostatisti ² Division of General Internal Medicine, Department of Medicine, Feinberg School of Medicine,

14 ⁵ Division of Clinical Trials and Biostatistics, Department of Quantitative Health Sciences, Mayo

15 ⁶ Division of Cl 14 Northwestern University, Chicago, IL, USA.

⁶Division of Clinical Trials and Biostatistics,

16 Clinic College of Medicine and Science, Rock

⁷Department of Preventive Medicine, Divisio

Feinberg School of Medicine
 ⁶Division of Clinical Trials and Biostatistics, Department of Quantitative Health Sciences, Mayo
-
- ²Division of Clinical Trials and Biostatistics, Department of Quantitative Health Sciences, Mayo

²Clinic College of Medicine and Science, Rochester, MN, USA

²Department of Preventive Medicine, Division of Epidemiol Clinic College of Medicine and Science, Rochester, MN, USA

⁷ Department of Preventive Medicine, Division of Epidemiology

Feinberg School of Medicine

^{*} These authors contributed equally
 Abstract word Count: 153
 Department of Preventive Medicine, Division of Epidemiology, Northwestern University
-
- Feinberg School of Medicine

^{*}These authors contributed ec

20 Abstract word Count: 153
 Word count: 1293
 **Tables: 2

Conflict of Interest Disclosu**

24 Conflict of Interest Disclosu

25 Funding/Support: None.
-
-
-
- The ² Tepartment of Preventive Medicine, Division of Epidemiology, Northwestern University

18 Feinberg School of Medicine

²⁸ These authors contributed equally

20 Abstract word Count: 153

21 **Word count**: 1293

22 ^{*}These authors contributed equally

20 Abstract word Count: 153
 21 Word count: 1293
 **22 Tables: 2

Conflict of Interest Disclosures: T

24 or other, exists with respect to the ir

725 Funding/Support:** None.
 Addre
-
- 21 **Word count**: 1293

22 **Tables**: 2
 23 Conflict of Interest

24 or other, exists with

25 **Funding/Support**: 1

26 **Address for Corres**

27 Ajay Bhasin, MD; A

28 Northwestern Unive 22 Tables: 2

23 Conflict of

24 or other, e

25 **Funding/92**

26 **Address f

27** Ajay Bhas

28 Northwest

29 Address: 2

20 Abstract word Count: 153

21 **Word count**: 1293

22 **Tables**: 2
 Conflict of Interest Discle

24 or other, exists with respec

25 **Funding/Support**: None.

26 **Address for Corresponde**

27 Ajay Bhasin, MD; Assistan **Funding/Support:** None.
 Address for Corresponde

27 Ajay Bhasin, MD; Assista

28 Northwestern University F

29 Address: 257 E Huron, Su

30 Email: <u>ajay bhasin@nm.or</u>

29 ORCiD: 0000-0001-5577-

- **Conflict of Interest Disclosures**: To the best of our knowledge, no conflict of interest, financial

24 or other, exists with respect to the information provided in this report.

25 **Funding/Support**: None.

26 **Address f** or other, exists with respect to the information provided in this report.
 Funding/Support: None.
 Address for Correspondence:

Ajay Bhasin, MD; Assistant Professor; Department of Medicine and D

Northwestern Universit **Address for Correspondence:**

27 Ajay Bhasin, MD; Assistant Prc

28 Northwestern University Feinbe

29 Address: 257 E Huron, Suite 16

30 Email: <u>ajay.bhasin@nm.org</u>

21 ORCiD: 0000-0001-5577-2065
-
- Northwestern University Feinberg School of Medicine

29 Address: 257 E Huron, Suite 16-738, Chicago, IL 6061

20 Email: <u>ajay.bhasin@nm.org</u>

21 ORCiD: 0000-0001-5577-2065 | Twitter: @Ajaybhasin

21 ORCiD: 0000-0001-5577-2
-
- 27 Ajay Bhasin, MD; Assistant Professor; Department of Medicine and Department of Pediatrics;

28 Northwestern University Feinberg School of Medicine

29 Address: 257 E Huron, Suite 16-738, Chicago, IL 60611

20 Email: <u>aj</u> Address: 257 E Huron, Suite 16-738, Chicago, IL 60611

20 Email: <u>ajay.bhasin@nm.org</u>

21 ORCiD: 0000-0001-5577-2065 | Twitter: @Ajaybhasin1

21 ORCiD: 0000-0001-5577-2065 | Twitter: @Ajaybhasin1

21 NOTE: This preprint re 30 Email: <u>ajay.bhasin@nm.org</u>
31 ORCiD: 0000-0001-5577-20
NOTE: This preprint reports new rese 31 ORCiD: 0000-0001-5577-2065 | Twitter: @Ajaybhasin19 | Phone: (312) 926-5893
NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide

32 **Abstract**
33 Computat
quality im
35 classificat
36 negative p
37 performed 33 Computable or electronic phenotypes of patient conditions are becoming more commonplace in
34 quality improvement and clinical research. During phenotyping algorithm validation, standard
35 classification performance me quality improvement and clinical research. During phenotyping algorithm validation, standard

classification performance measures (i.e., sensitivity, specificity, positive predictive value,

negative predictive value, and classification performance measures (i.e., sensitivity, specificity, positive predictive value,
negative predictive value, and accuracy) are commonly employed. When validation is
performed on a randomly sampled patient pop negative predictive value, and accuracy) are commonly employed. When validation is
performed on a randomly sampled patient population, direct estimates of these measure
valid. However, it is common that studies will sample 37 performed on a randomly sampled patient population, direct estimates of these measures are
38 valid. However, it is common that studies will sample patients conditional on the algorithm
39 result, leading to a form of b valid. However, it is common that studies will sample patients conditional on the algorithm

139 result, leading to a form of bias known as verification bias. The presence of verification bias

140 requires adjustment of p 139 result, leading to a form of bias known as verification bias. The presence of verification bias
140 requires adjustment of performance measure estimates to account for this sampling bias. Here
141 we describe the appro requires adjustment of performance measure estimates to account for this sampling bias. Herein,

41 we describe the appropriate formulae for valid estimates of sensitivity, specificity, and accuracy

42 to account for veri 41 we describe the appropriate formulae for valid estimates of sensitivity, specificity, and accuracy

42 to account for verification bias. We additionally present an online tool to adjust algorithm

44 performance measure 42 to account for verification bias. We additionally present an online tool to adjust algorithm
performance measures for verification bias by directly taking the sampling strategy into
consideration and recommend use of th 43 performance measures for verification bias by directly taking the sampling strategy into

44 consideration and recommend use of this tool to properly estimate algorithm performanc

45 phenotyping validation studies. 44 consideration and recommend use of this tool to properly estimate algorithm performance for
phenotyping validation studies. 45 phenotyping validation studies.

46 **Introduction**
47 Computable p
48 improvement a
50 Medicaid Serv
51 performance n Computable phenotypes of patient conditions are becoming more commonplace in quality

improvement and clinical research.¹ These phenotypes are algorithmically derived from da

sources such as electronic health record (EH improvement and clinical research.¹ These phenotypes are algorithmically derived from data improvement and clinical research.¹ These phenotypes are algorithmically derived from data
sources such as electronic health record (EHR), insurance claims, or centers for Medicare and
Medicaid Services data, and can emp sources such as electronic health record (EHR), insurance claims, or centers for Medicare and
Medicaid Services data, and can empower research and improve patient care.^{2.3} Algorithm
performance measures, such as sensitiv Medicaid Services data, and can empower research and improve patient care.^{2,3} Algorithm Medicaid Services data, and can empower research and improve patient care.^{2,3} Algorithm
performance measures, such as sensitivity, specificity, and positive and negative predictive
values (PPV and NPV) are common measur 51 performance measures, such as sensitivity, specificity, and positive and negative predictive
52 values (PPV and NPV) are common measures of validity obtained by comparing the algorit
53 result to a "gold standard" (e.g. values (PPV and NPV) are common measures of validity obtained by comparing the algorithm
result to a "gold standard" (e.g. manual chart review). A common validation study design
strategy when the condition of interest has result to a "gold standard" (e.g. manual chart review). A common validation study design
strategy when the condition of interest has low prevalence is to sample based on the algorit
result (e.g. 50 predicted cases and 50 p 54 strategy when the condition of interest has low prevalence is to sample based on the algorithm
55 result (e.g. 50 predicted cases and 50 predicted non-cases).^{4,5} This strategy is both cost-effectiv
56 and statistical result (e.g. 50 predicted cases and 50 predicted non-cases). 4.5 This strategy is both cost-effective result (e.g. 50 predicted cases and 50 predicted non-cases).^{4,3} This strategy is both cost-effective
and statistically efficient by enriching for likely true positives and improving the expected
precision of positive-cla and statistically efficient by enriching for likely true positives and improving the expected
precision of positive-class performance measures (e.g., sensitivity, PPV). However, this
sampling strategy also results in a for precision of positive-class performance measures (e.g., sensitivity, PPV). However, this
sampling strategy also results in a form of selection bias known as verification bias, whic
commonly encountered in diagnostic test e sampling strategy also results in a form of selection bias known as verification bias, which is

commonly encountered in diagnostic test evaluation.⁶⁻⁸ Under these conditions, estimates of

sensitivity, specificity, and commonly encountered in diagnostic test evaluation.⁶⁻⁸ Under these conditions, estimates of commonly encountered in diagnostic test evaluation.⁶⁻⁸ Under these conditions, estimates of
sensitivity, specificity, and accuracy can be biased if the sampling design is not taken into
consideration. Herein, we illustra 60 sensitivity, specificity, and accuracy can be biased if the sampling design is not taken into
61 consideration. Herein, we illustrate the effects of verification bias on performance estimati
62 through an example valida 61 consideration. Herein, we illustrate the effects of verification bias on performance estimation
62 through an example validation study and develop a user-friendly online tool to facilitate
63 adjustment of performance m

through an example validation study and develop a user-friendly online tool to facilitate
63 adjustment of performance measures under these validation study scenarios.
Methods
65 Given that EHR-based phenotyping algorith 63 adjustment of performance measures under these validation study scenarios.
64 **Methods**
65 Given that EHR-based phenotyping algorithms can be prone to error, it is oft
66 characterize classification performance relative 64 **Methods**
65 Given tha
66 characteri
67 abstraction
68 and specif 65 Given that EHR-based phenotyping algorithms can be prone to error, it is often of interest to characterize classification performance relative to ground truth based on manual chart abstraction. Formulae for defining the 66 characterize classification performance relative to ground truth based on manual chart
distanction. Formulae for defining these performance measures adjusting estimates of
and specificity for verification bias are avail 67 abstraction. Formulae for defining these performance measures adjusting estimates of sensitivity
68 and specificity for verification bias are available in Figure 1. Detailed explanations of these
3 68 and specificity for verification bias are available in Figure 1. Detailed explanations of these

Greenes.⁹

derivations, along with formulae for calculating corresponding asymptotic CI's, are provided by Begg and

There is a Greenes.⁹

The Greenes.⁹

The Validation Study Sampling Design

The Validation Study Sampling Design
 71
72
73
74
--72
73
74
75
76 Validation Study Sampling Design

73 For phenotyping algorithms, the tot

74 tends to be very large due to ease of

75 medical institution). Given the pot

76 studies are often performed on a relatively

77 expected preval For phenotyping algorithms, the total number of patients with available classification results
tends to be very large due to ease of implementation (e.g., the entire patient population at a
medical institution). Given the tends to be very large due to ease of implementation (e.g., the entire patient population at a

75 medical institution). Given the potential laborious nature of chart review, algorithm validat

37 studies are often perform medical institution). Given the potential laborious nature of chart review, algorithm validation
37 studies are often performed on a relatively small subset of the total population. When the
37 expected prevalence of the d tudies are often performed on a relatively small subset of the total population. When the
expected prevalence of the disease condition is low (i.e., less than 10%), validation studie
have correspondingly low precision for zo expected prevalence of the disease condition is low (i.e., less than 10%), validation studies may
have correspondingly low precision for estimating sensitivity and PPV if patients are randomly
sampled from the populatio Transformal provides are randomly sampled from the population. For example, for a disease with prevalence of 2%, in a random
sample of 500 patients we expect 10 positive disease patients, on average. Even at a true
algorit 79 sampled from the population. For example, for a disease with prevalence of 2%, in a random

80 sample of 500 patients we expect 10 positive disease patients, on average. Even at a true

81 algorithm sensitivity of 90% (80 sample of 500 patients we expect 10 positive disease patients, on average. Even at a true
81 algorithm sensitivity of 90% (i.e., 9/10 cases correctly identified), the Wilson score 95%
82 confidence interval (CI) would b 81 algorithm sensitivity of 90% (i.e., 9/10 cases correctly identified), the Wilson score 95%

82 confidence interval (CI) would be [0.596,0.995]. In contrast, 90% specificity would corre

83 to a 95% confidence interval o 82 confidence interval (CI) would be [0.596,0.995]. In contrast, 90% specificity would correspond
83 to a 95% confidence interval of [0.870,0.925]. This disparity in precision can be mitigated by
84 oversampling subjects p to a 95% confidence interval of [0.870,0.925]. This disparity in precision can be mitigated by

84 oversampling subjects predicted by the algorithm as a positive case (e.g., 1:1 sampling based c

95 predicted disease statu %84 oversampling subjects predicted by the algorithm as a positive case (e.g., 1:1 sampling based on
85 predicted disease status), leading to a more balanced representation of true disease cases and
86 unaffected non-cases

185 predicted disease status), leading to a more balanced representation of true disease cases and

186 unaffected non-cases within the validation sample.

187 Naïve and Adjusted Validation Performance

188 While the sampl unaffected non-cases within the validation sample.

Naive and Adjusted Validation Performance

While the sampling strategy defined above leads to

algorithm performance, sampling patients for the va

disease status can lea 87 *Naïve and Adjusted Validation Performance*
88 While the sampling strategy defined above le
89 algorithm performance, sampling patients for
90 disease status can lead to biased estimation o 88 While the sampling strategy defined above leads to more statistically efficient estimation of algorithm performance, sampling patients for the validation study based on algorithm-classif
disease status can lead to biase 89 algorithm performance, sampling patients for the validation study based on algorithm-classified
90 disease status can lead to biased estimation of performance measures. Referred to as 90 disease status can lead to biased estimation of performance measures. Referred to as

101 true negatives.

102 Simulation Ana

103 To further illus

104 a broad range c

105 with estimated

106 PPV of 0.70, 0. 102 *Simulation Analysis*
103 To further illustrate 1
104 a broad range of real
105 with estimated true p
106 PPV of 0.70, 0.80, a
107 equal numbers of pre 103 To further illustrate the impact of verification bias on sensitivity and specificity estimates across

104 a broad range of realistic study conditions, we conducted a simple simulation study for a disease

105 with est 204 a broad range of realistic study conditions, we conducted a simple simulation study for a disease
205 with estimated true prevalence between 1% and 50%; true NPV of 0.90, 0.95, and 0.99; and true
206 PPV of 0.70, 0.80, with estimated true prevalence between 1% and 50%; true NPV of 0.90, 0.95, and 0.99; and true

PPV of 0.70, 0.80, and 0.90. For validation, we considered a balanced study design, such that

equal numbers of predicted cases 106 PPV of 0.70, 0.80, and 0.90. For validation, we considered a balanced study design, such that equal numbers of predicted cases and non-cases are selected for chart abstraction. We then calculated the expected bias of n equal numbers of predicted cases and non-cases are selected for chart abstraction. We then

calculated the expected bias of naive estimates of sensitivity and specificity relative to

appropriately adjusted estimates based calculated the expected bias of naive estimates of sensitivity and specificity relative to

109 appropriately adjusted estimates based on expected values of true positive rate (TPR), 1

110 positive rate (FPR), true negati 109 appropriately adjusted estimates based on expected values of true positive rate (TPR), false

110 positive rate (FPR), true negative rate (TNR), and false negative rate (FNR) in the validation

111 study.

112 *Online* 110 positive rate (FPR), true negative rate (TNR), and false negative rate (FNR) in the validation
111 study.
2011 *Online Tool*

111 study.
112 *Online* 112 *Online Tool*

We used Microsoft Visual Studio Code (version 1.78.0) and Python (version 3.10) with the
114 Streamlit package (version 1.13.0) to create a simple tool to calculate sensitivity, specificity
115 PPV, NPV, and accuracy of a

114 *Streamlit* package (version 1.13.0) to create a simple tool to calculate sensitivity, specificity,

115 PPV, NPV, and accuracy of a phenotyping algorithm based on chart validation. The tool is

116 freely available at 115 PPV, NPV, and accuracy of a phenotyping algorithm based on chart validation. The tool is

116 freely available at: https://bit.ly/3tMTJiE.

117 **Results**

118 The 2x2 contingency table of the example validation study a 116 freely available at: https://bit.ly/3tMTJiE.

117 **Results**

118 The 2x2 contingency table of the example

119 total source cohort are presented in Table

120 corresponding to unadjusted and verificati

121 Unadjusted 117 **Results**
118 The 2x2
119 total sou
120 correspo
121 Unadjus
122 0.942 se The 2x2 contingency table of the example validation study along with projected counts from the
total source cohort are presented in Table 1, while respective performance measure analyses
corresponding to unadjusted and ver total source cohort are presented in Table 1, while respective performance measure analyses

120 corresponding to unadjusted and verification-bias adjusted estimates are presented in Table 2

121 Unadjusted performance est corresponding to unadjusted and verification-bias adjusted estimates are presented in Table 2.

121 Unadjusted performance estimates for the hypothesized phenotyping algorithm corresponded t

122 0.942 sensitivity, 0.979 s Unadjusted performance estimates for the hypothesized phenotyping algorithm corresponded to

122 0.942 sensitivity, 0.979 specificity, and 0.960 accuracy. The disease prevalence in the validation

123 study sample was 0.52 0.942 sensitivity, 0.979 specificity, and 0.960 accuracy. The disease prevalence in the validation

study sample was 0.520, whereas the true prevalence in the source population was 0.091. After

adjusting for verification

123 study sample was 0.520, whereas the true prevalence in the source population was 0.091. After
124 adjusting for verification bias, the updated performance measures for the algorithm corresponde
125 to 0.620 sensitivity adjusting for verification bias, the updated performance measures for the algorithm corresponded
to 0.620 sensitivity, 0.999 specificity, and 0.944 accuracy.
Results from our simulation study are presented in Figure 2. The to 0.620 sensitivity, 0.999 specificity, and 0.944 accuracy.

126 Results from our simulation study are presented in 1

127 substantial positive bias for sensitivity estimation that may

128 decreases toward zero when anal Results from our simulation study are presented in Figure 2. These results illustrate the

127 substantial positive bias for sensitivity estimation that may be observed as disease prevalence

128 decreases toward zero when substantial positive bias for sensitivity estimation that may be observed as disease prevalence
decreases toward zero when analyzing the unadjusted validation study confusion matrix result
This bias relationship is attenua decreases toward zero when analyzing the unadjusted validation study confusion matrix results.

129 This bias relationship is attenuated as the NPV approaches 1.00, but still yields extreme bias at

130 Iower prevalence va This bias relationship is attenuated as the NPV approaches 1.00, but still yields extreme bias at
130 lower prevalence values. For specificity (Figure 2B), we observe similar trends of increased
131 absolute bias with decr 130 lower prevalence values. For specificity (Figure 2B), we observe similar trends of increased
131 absolute bias with decreased prevalence. However, the magnitude of this bias remains largel
132 consistent across realist 131 absolute bias with decreased prevalence. However, the magnitude of this bias remains largely

132 consistent across realistically high values of NPV considered for the simulation study, with lov

133 PPV leading to mod 2132 consistent across realistically high values of NPV considered for the simulation study, with lower

2133 PPV leading to moderate increases in bias. Of note, these results represented expected biases,

2134 and actual 133 PPV leading to moderate increases in bias. Of note, these results represented expected biases,
134 and actual results may vary based on sizes of the total population and sampling cohort due to
135 sampling variability. 134 and actual results may vary based on sizes of the total population and sampling cohort due to
135 sampling variability. 135 sampling variability.

135 sampling variability.

136 **Discussion**
137 The
138 much they
139 algorithm v
140 performanc
141 mitigated b 137 The provided example demonstrates the performance metrics of an algorithm and how

138 much they can change when one does not randomly sample from the source population for

139 algorithm validation. Oversampling of al much they can change when one does not randomly sample from the source population for
139 algorithm validation. Oversampling of algorithm-positive cases for validation can bias mod
140 performance measures, leading to infl 139 algorithm validation. Oversampling of algorithm-positive cases for validation can bias model
140 performance measures, leading to inflated sensitivity and accuracy estimates. The bias can be
141 mitigated by considerin

140 performance measures, leading to inflated sensitivity and accuracy estimates. The bias can be

141 mitigated by considering the prevalence of disease in the source population and adjusting the

142 calculations to acco mitigated by considering the prevalence of disease in the source population and adjusting the

calculations to account for the difference.

While sampling conditional on predicted disease status will lead to valid direct e calculations to account for the difference.

143 While sampling conditional on pre

144 of PPV and NPV, these measures are then

145 are not necessarily intrinsic properties of a

146 with caution as disease prevalence may While sampling conditional on predicted disease status will lead to valid direct estimates
144 of PPV and NPV, these measures are themselves a function of disease prevalence. Thus, they
145 are not necessarily intrinsic pr of PPV and NPV, these measures are themselves a function of disease prevalence. Thus, they
are not necessarily intrinsic properties of a phenotyping algorithm, and should be interpreted
with caution as disease prevalence m are not necessarily intrinsic properties of a phenotyping algorithm, and should be interpreted

146 with caution as disease prevalence may vary across validation populations.¹⁰ Likewise,

147 alternative performance meas with caution as disease prevalence may vary across validation populations.¹⁰ Likewise, with caution as disease prevalence may vary across validation populations.¹⁰ Likewise,
alternative performance measures that are in part functions of sensitivity and/or specific:
as F1-score and positive/negative likeli alternative performance measures that are in part functions of sensitivity and/or specificity, such
as F1-score and positive/negative likelihood ratios, will also likely be biased and require similar
corrections. Stratifie as F1-score and positive/negative likelihood ratios, will also likely be biased and require similar
149 corrections. Stratified study designs can also be adopted when there are covariates that may
150 correlate with differ references for how to address adjustment under these conditions.^{6,9}

149 corrections. Stratified study designs can also be adopted when there are covariates that may

150 correlate with differential algorithm performance, and we refer the reader to appropriate

151 references for how to ad correlate with differential algorithm performance, and we refer the reader to appropriate
151 references for how to address adjustment under these conditions.^{6,9}
152 For accurate adjustment and algorithm calibration, the ---
152
153
154
155
156 152 For accurate adjustment and algorithm calibration, the source population should be
153 defined prior to application of an algorithm. Ideally, a very high percentage of the source
154 population will be characterized by defined prior to application of an algorithm. Ideally, a very high percentage of the source
154 population will be characterized by the algorithm: if a high percentage of patients are not
155 classified as either disease p 154 population will be characterized by the algorithm: if a high percentage of patients are not

155 classified as either disease positive or negative by the algorithm, then the performance me

156 the algorithm will be d 2155 classified as either disease positive or negative by the algorithm, then the performance metrics of
2156 the algorithm will be difficult to interpret and this will significantly increase the difficulty of
2157 cross-i 156 the algorithm will be difficult to interpret and this will significantly increase the difficulty of cross-institutional validation.¹¹⁻¹³ cross-institutional validation. $11-13$

158 This tool will enable clinicians, informaticists, and data scientists to appropriately

159 characterize performance of computable phenotype algorithms.

160 **References:**

161 1. Richesson RL, Smerek MM, Blake Cameron characterize performance of computable phenotype algorithms.

160 **References:**

161 1. Richesson RL, Smerek MM, Blake Cameron C. A Framework to

162 Computable Phenotype Definitions Across Health Care Delivery and (

163 160 **References:**

161 1. Riches

162 Computable P

163 EGEMS (Wash

164 2. Bielins

165 Failure with D

166 and Genomics

167 3. Carrol

168 arthritis in ele

169 4. Jackso

170 phenotype alg

171 and controls fa

Computable Phenotype Definitions Across Health Care Delivery and Clinical Research Applications.

163 EGEMS (Wash DC) 2016;4:1232.

2. Bielinski SJ, Pathak J, Carrell DS, et al. A Robust e-Epidemiology Tool in Phenotyping EGEMS (Wash DC) 2016;4:1232.

164 2. Bielinski SJ, Pathak J, Carrell DS, et al. A Robust e-Epidemiology Tool in Phenotyping Heart

164 Failure with Differentiation for Preserved and Reduced Ejection Fraction: the Electroni 2. Bielinski SJ, Pathak J, Car
164 2. Bielinski SJ, Pathak J, Car
165 Failure with Differentiation for P
166 and Genomics (eMERGE) Networ
167 3. Carroll RJ, Thompson WK
168 arthritis in electronic health reco
169 4. Jackso Failure with Differentiation for Preserved and Reduced Ejection Fraction: the Electronic Medical Report and Genomics (eMERGE) Network. J Cardiovasc Transl Res 2015;8:475-83.

167 3. Carroll RJ, Thompson WK, Eyler AE, et al 166 and Genomics (eMERGE) Network. J Cardiovasc Transl Res 2015;8:475-83.

167 3. Carroll RJ, Thompson WK, Eyler AE, et al. Portability of an algorithm to identify rheumatoid

168 arthritis in electronic health records. J 3. Carroll RJ, Thompson WK, Eyler AE, et al. Portability of an algorithm

168 arthritis in electronic health records. J Am Med Inform Assoc 2012;19:e162

169 4. Jackson KL, Mbagwu M, Pacheco JA, et al. Performance of an el arthritis in electronic health records. J Am Med Inform Assoc 2012;19:e162-9.

169 4. Jackson KL, Mbagwu M, Pacheco JA, et al. Performance of an electronic health record-base

170 phenotype algorithm to identify community 4. Jackson KL, Mbagwu M, Pacheco JA, et al. Performance of an electronic phenotype algorithm to identify community associated methicillin-resistant St
170 phenotype algorithm to identify community associated methicillin-re

provided methods of Diagnostic Tests Part 1: Patient Selection
170 phenotype algorithm to identify community associated methicillin-resistant Staphylococcus aureus can
171 and controls for genetic association studies. BMC 171 and controls for genetic association studies. BMC Infect Dis 2016;16:684.

172 5. Kho AN, Hayes MG, Rasmussen-Torvik L, et al. Use of diverse electronic medical record systems

173 to identify genetic risk for type 2 d 172 5. Kho AN, Hayes MG, Rasmussen-Torvik L, et al. Use of diverse election dentify genetic risk for type 2 diabetes within a genome-wide association assoc 2012;19:212-8.
174 Assoc 2012;19:212-8.
175 6. Gaffikin L, McGrath to identify genetic risk for type 2 diabetes within a genome-wide association study. J Am Med Inform

174 Assoc 2012;19:212-8.

175 6. Gaffikin L, McGrath J, Arbyn M, Blumenthal PD. Avoiding verification bias in screening

174 Assoc 2012;19:212-8.

174 Assoc 2012;19:212-8.

175 6. Gaffikin L, McGrath J, Arbyn M, Blumenthal PD. Avoiding verification bias in screening test

176 evaluation in resource poor settings: a case study from Zimbabwe. 175 6. Gaffikin L, Mc
176 evaluation in resource
177 7. O'Sullivan JW,
178 2018;23:54-5.
179 8. Hall MK, Kea E
180 Emerg Med J 2019;36:
181 9. Begg CB, Gree
182 selection bias. Biomet
183 10. Grunau G, Lini
184 Pliabilitie

evaluation in resource poor settings: a case study from Zimbabwe. Clin Trials 2008;5:496-503.

177 7. O'Sullivan JW, Banerjee A, Heneghan C, Pluddemann A. Verification bias. BMJ Evid Based N

2018;23:54-5.

179 8. Hall MK, 77 7. O'Sullivan JW, Banerjee A, Heneghan C, Pluddemann A. Verification bias. BMJ Evid Bas

178 2018;23:54-5.

179 8. Hall MK, Kea B, Wang R. Recognising Bias in Studies of Diagnostic Tests Part 1: Patient

180 Emerg Med J 77 7. O'Sullivan JW, Banerjee A, Heneghan C, Pluddemann A. Verification bias. BMJ Evid Based Med
178 2018;23:54-5.
179 8. Hall MK, Kea B, Wang R. Recognising Bias in Studies of Diagnostic Tests Part 1: Patient Selection
18 179 8. Hall M
180 Emerg Med J 2
181 9. Begg C
182 selection bias.
183 10. Grunal
184 Pliabilities, anc
185 11. Rasmu
185 11. Rasmu
185 12. Newto
187 12. Newto
188 phenotyping a
189 2013;20:e147-Example 188. Hall Magnostic tests when disease verification is subject to
180 Emerg Med J 2019;36:431-4.
181 9. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to
182 selection bias 181 9. Begg CB, Greenes RA
182 selection bias. Biometrics 198
183 10. Grunau G, Linn S. Cor
184 Pliabilities, and Pitfalls in Rese
185 11. Rasmussen-Torvik LJ,
186 on study sample selection in
187 12. Newton KM, Peissig F
 selection bias. Biometrics 1983;39:207-15.

183 10. Grunau G, Linn S. Commentary: Sensitivity, Specificity, and Predictive Values: Foundations,

184 Pilabilities, and Pitfalls in Research and Practice. Front Public Health 183 10. Grunau G, Linn S. Commentary: Ser
184 Pliabilities, and Pitfalls in Research and Praces
185 11. Rasmussen-Torvik LJ, Furmanchuk
186 on study sample selection in electronic hea
187 12. Newton KM, Peissig PL, Kho AN, Pilabilities, and Pitfalls in Research and Practice. Front Public Health 2018;6:256.

183 11. Rasmussen-Torvik LJ, Furmanchuk A, Stoddard AJ, et al. The effect of number of healthcare

185 11. Rasmussen-Torvik LJ, Furmanch 185 11. Rasmussen-Torvik LJ, Furmanchuk A, Stoddard AJ, et al. The effect of nun
186 an study sample selection in electronic health record data. Int J Popul Data Sci 2C
187 12. Newton KM, Peissig PL, Kho AN, et al. Validat

on study sample selection in electronic health record data. Int J Popul Data Sci 2020;5.

187 12. Newton KM, Peissig PL, Kho AN, et al. Validation of electronic medical record-based

188 phenotyping algorithms: results and 187 12. Newton KM, Peissig PL, Kho AN, et al. Validation of electronic medical record-b
188 phenotyping algorithms: results and lessons learned from the eMERGE network. J Am I
189 2013;20:e147-54.
190 13. Desai JR, Wu P, N 187 12. Newton KM, Peissig PL, Kho AN, et al. Validation of electronic medical record-based 189 2013;20:e147-54.

189 2013;20:e147-54.

190 13. Desai JR, Wu P, Nichols GA, Lieu TA, O'Connor PJ. Diabetes and asthma case identification,

validation, and representativeness when using electronic health data to constr

190 13. Desai JR, V
191 validation, and rep
192 comparative effec
193

190 13. Desai JR, Wu P, Nichols GA, Lieu TA, O'Connor PJ. Diabetes and asthma case identification, validation, and representativeness when using electronic health data to construct registries for comparative effectiveness

192 comparative effectiveness and epidemiologic research. Med Care 2012;50 Suppl:S30-5.

193

194 112 comparative effectiveness and epidemiologic research. Med Care 2012;
193
194

194
|
|

- 198
199
200
201 ---
199
200
201
-
-

203
|
|

-
-

204 **Figures** 205 **Figure 1.** 206 207 **A. Mathematical definition of performance measures:** 208 **Five primary performance measures of interest:** -  Pr 1| 1 -  Pr 0| 0  Pr 1| 1  Pr 0| 0 
  Pr

210

215
216 ---
216 216

b) Adjusted estimates of sensitivity and specificity correcting for verification bias

\n
$$
\text{Sensitivity}^{adj} = \frac{TPR}{TPR + FNR} = \frac{PPV \times \tau^{+}}{PPV \times \tau^{+} + (1 - NPV) \times \tau^{-}}
$$
\n
$$
\text{Specificity}^{adj} = \frac{TNR}{TNR + FPR} = \frac{NPV \times \tau^{-}}{NPV \times \tau^{-} + (1 - PPV) \times \tau^{+}}
$$
\n218

\n219

219
220
221
2223
2223
2225
2225
2227 219 220 **Legend**

221 Consider

222 patient's

223 and $X \in$

224 unaffect

225 classific:

225 classific:

227 true nega

228 equation

728 τ^+ and τ

231 rates of 1 221 Consider a phenotyping algorithm for predicting the presence of a given disease condition based on a

222 patient's EHR data. We designate $Y \in \{0,1\}$ to be the true underlying disease status for a given patien

223 222 patient's EHR data. We designate $Y \in$

223 and $X \in \{0,1\}$ to be the predicted disease

224 unaffected and affected disease statuse

classifications of disease relative to true

227 true negatives (TN), false positi patient's EHR data. We designate $Y \in \{0,1\}$ to be the true underlying disease status for a given patient 223 and $X \in$

224 unaffect

225 classific

225 classific

227 true neg

228 equation

229 rate TPI

230 τ^+ and τ

231 rates of

232 these pa

233 above th

234 $Pr(Y =$ and $X \in \{0,1\}$ to be the predicted disease status by the algorithm, such that 0 and 1 respectively denote unaffected and affected disease statuses. For disease phenotyping on a patient cohort of size N, the unaffected and affected disease statuses. For disease phenotyping on a patient cohort of size *N*, the classification results can be summarized using a standard 2x2 contingency table, which tabulates pa classifications of classification results can be summarized using a standard 2x2 contingency table, which tabulates patient

classifications of disease relative to true disease status into four distinct categories: true positives (TP),

tru classifications of disease relative to true disease status into four distinct categories: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), as indicated in Table 1. Counts in the eq 227 true negatives (TN), false positives (FP), and false negatives (FN), as indicated in Table 1. Counts in the equations above can be replaced by corresponding rates by simply factoring out *N* (e.g., the true positive r equations above can be replaced by corresponding rates by simply factoring out *N* (e.g., the true positive rate *TPR* = $\frac{TP}{N}$ = Pr (*Y* = 1, *X* = 1)). Given that unbiased estimates of test positive and negative rate rate $TPR = \frac{1}{N}$ τ^+ and τ^- , are 229 rate $TPR = \frac{17}{N} = Pr (Y = 1, X = 1)$). Given that unbiased estimates of test positive and negative rates,

230 τ^+ and τ^- , are available from the algorithm classifications for the original source cohort, the expected 230 τ^+ and τ^- , are available from the algorithm classifications for the original source cohort, the expected

231 rates of TPR, TNR, FPR, and FNR in the source cohort can actually be calculated as simple functions 231 rates of TPR, TNR, FPR, and FNR in the source cohort can actually be calculated as simple functions of these parameters and the PPV and NPV estimates from the validation study. For example, recall from a bove that TPR 232 these parameters and the PPV and NPV estimates from the validation study. For example, recall from

233 above that TPR can be framed as the joint probability Pr ($Y = 1, X = 1$). Since Pr($Y = 1, X = 1$) = $Pr(Y = 1 | X = 1) \times Pr(X =$ 233 above that TPR can be framed as the joint probability Pr $(Y = 1, X = 1)$. Since Pr $(Y = 1, X = 1) =$ Pr $(Y = 1 | X = 1) \times$ Pr $(X = 1)$ by basic rules of conditional probability, and Pr $(Y = 1 | X = 1) =$ Pl and Pr $(X = 1) = \tau^+$ per our def 234 Pr(Y = 1|X = 1) × Pr (X = 1) by basic rules of conditional probability, and Pr(Y = 1|X = 1) = PPV and Pr(X = 1) = τ^+ per our definitions above, it follows that $TPR = PPV \times \tau^+$
236 and $Pr(X = 1) = \tau^+$ per our definitions above, it follows that $TPR = PPV \times \tau^+$

235

239