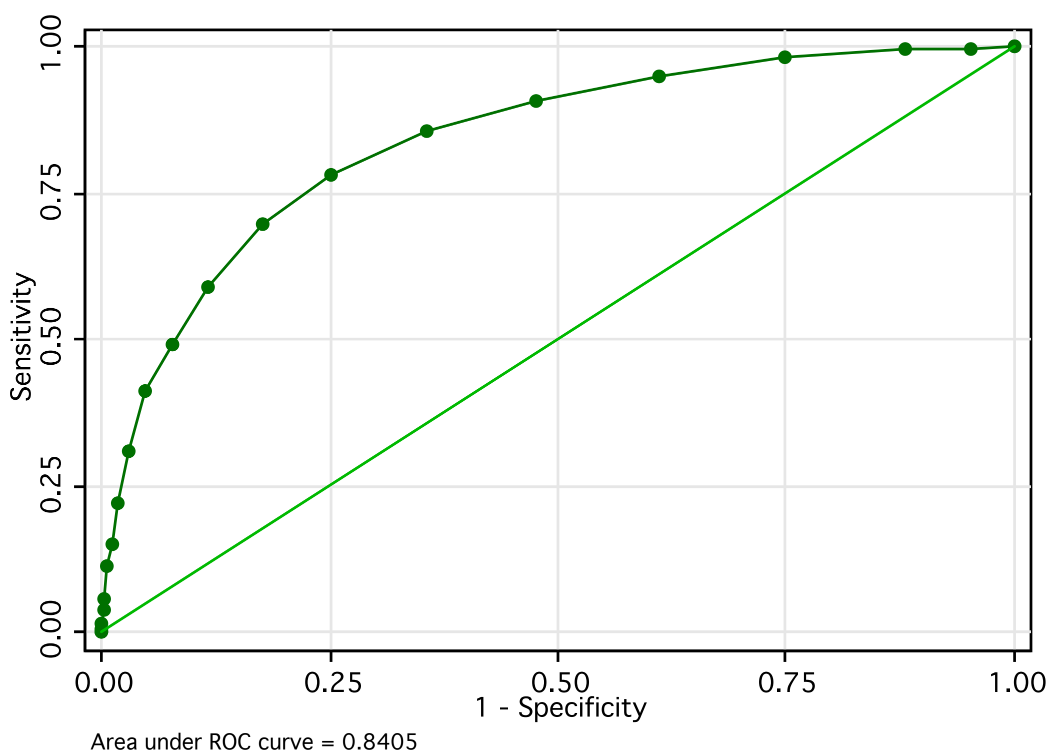
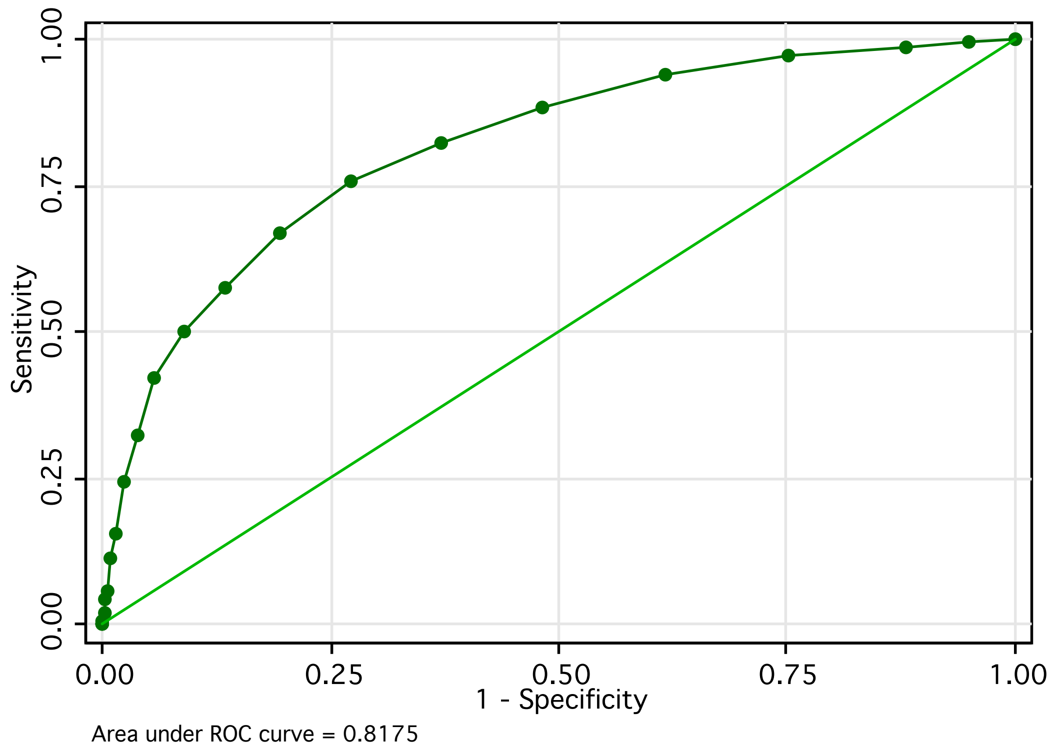
Model Validation

Dave Glidden
3 December 2013

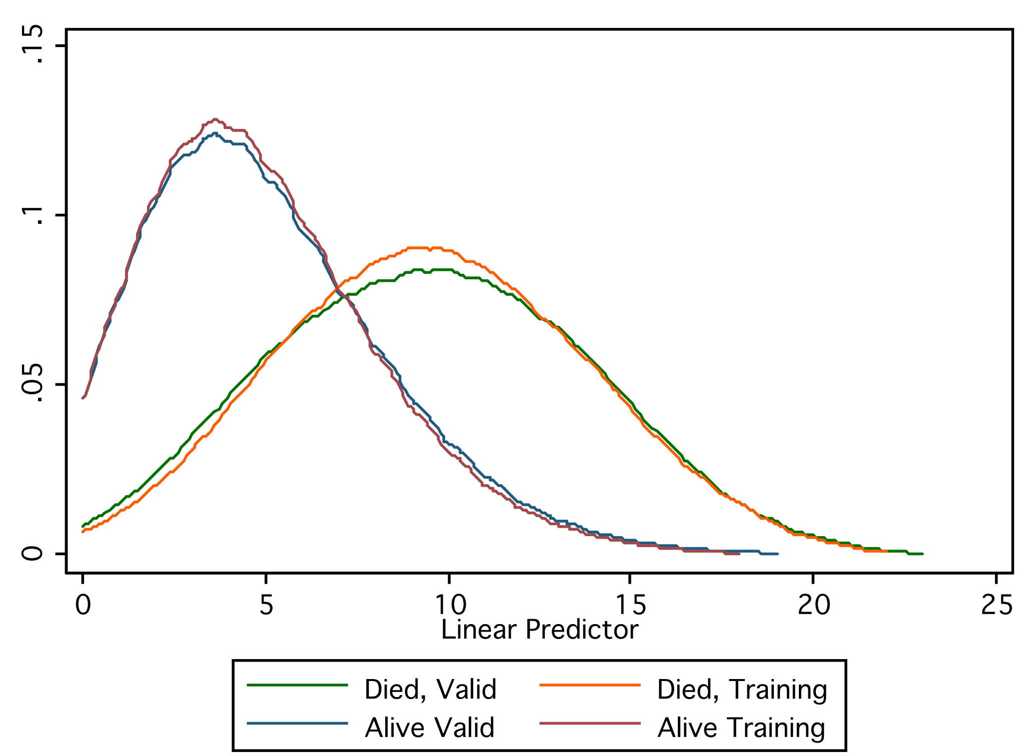
ROC Lee Model: Learn



ROC: Test



Density



Calibration

- Model: predicted probability
- Does predicted probability match truth
- Important for individual decision making
- Classic model-checking problem
- Incorporate statistical tests
  and descriptions of magnitude
- Find model misspecified in some way

Lee Model

validation set

```
. xi: logistic dead i.agecat male hrsdm hrsca hrslung hrschf bmi smoke bath money walkjog push
i.agecat          _Iagecat_0-6        (naturally coded; _Iagecat_0 omitted)

Logistic regression                               Number of obs   =       7973
                                                  LR chi2(17)     =    1413.63
                                                  Prob > chi2     =     0.0000
Log likelihood = -2423.8528                       Pseudo R2       =     0.2258

------------------------------------------------------------------------------
        dead | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Iagecat_1 |   1.537822   .2259016     2.93   0.003       1.1531    2.050903
  _Iagecat_2 |   2.248315   .3291624     5.53   0.000     1.687476    2.995549
  _Iagecat_3 |   2.586506    .381196     6.45   0.000     1.937602     3.45273
  _Iagecat_4 |   4.109389   .5813912     9.99   0.000     3.114228    5.422558
  _Iagecat_5 |   7.608406    1.14937    13.43   0.000     5.658564    10.23013
  _Iagecat_6 |   14.62968   2.233807    17.57   0.000     10.84588    19.73353
        male |   2.166964   .1712289     9.79   0.000     1.856058    2.529949
       hrsdm |   1.761092   .1639891     6.08   0.000     1.467303    2.113704
       hrsca |   1.664218   .1689592     5.02   0.000      1.36393    2.030617
     hrslung |   1.587757   .2179521     3.37   0.001     1.213219     2.07792
      hrschf |   2.136519   .3437415     4.72   0.000     1.558685    2.928569
      bmicat |   1.736036   .1361793     7.03   0.000     1.488635    2.024553
       smoke |   1.594206   .1606953     4.63   0.000     1.308409    1.942429
        bath |   1.474593   .1711166     3.35   0.001     1.174616    1.851179
       money |   1.420027   .1534369     3.25   0.001     1.149006    1.754974
     walkjog |   1.995648   .1817115     7.59   0.000      1.66947    2.385553
        push |    1.49066   .1350732     4.41   0.000     1.248098    1.780363
------------------------------------------------------------------------------
```

Inspecting Calibration

- Form series of groups based on risk
- Compare expected mortality with observed
- Expected: average(probs)
- Observed: average(death) = *proportion who died*
- Within each group

our point total furnishes such a group

Test Set

```
. table pts2 if valid==0, c(n dead mean dead mean prob)
```

Points	N	mean(dead)	mean(prob)
0	486	.0041152	.0073689
1	739	.0067659	.0126751
2	1,366	.0095168	.017859
3	1,474	.0325645	.0277967
4	1,445	.0394464	.0407945
5	1,330	.0503759	.0589664
6	1,162	.0886403	.0849856
7	886	.1230248	.1186558
8	758	.1912929	.1653885
9	551	.2413793	.2242661
10	407	.2727273	.2922267
11	320	.43125	.384265
12	244	.4836065	.4604374
13	174	.5344828	.5641574
14	107	.5046729	.6368158
15+	203	.7438424	.7780437

Validation Set

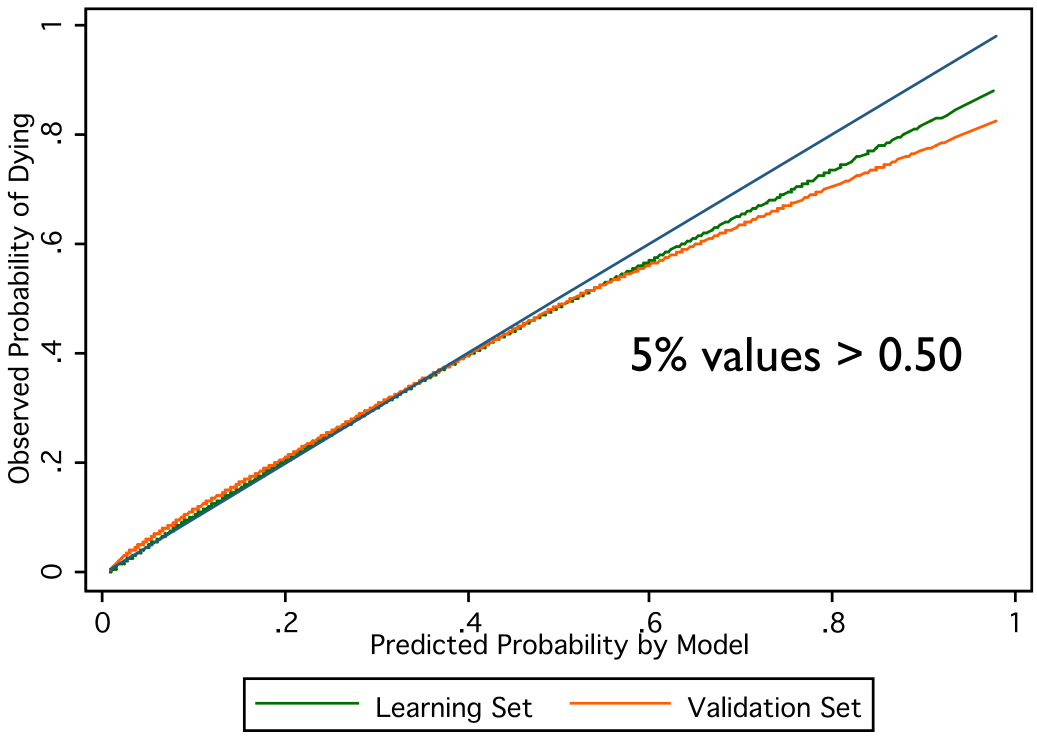
```
. table pts2 if valid==1, c(n dead mean dead mean prob)
```

Points	N	mean(dead)	mean(prob)
0	354	.0056497	.0073689
1	489	.0204499	.0127029
2	889	.0168729	.0174934
3	971	.0360453	.0275617
4	986	.0598377	.0408728
5	842	.0771971	.0585236
6	758	.0936675	.0851236
7	637	.1507064	.1188091
8	498	.1967871	.1658374
9	401	.19202	.2229141
10	308	.275974	.2988213
11	232	.4568965	.3782594
12	192	.4427083	.4634854
13	159	.591195	.5602076
14	77	.5584416	.6408411
15+	180	.6777778	.789714

Smoothing

- Alternative graphical approach
- Draw smooth curve of mortality by predicted probability
- Should look like straight line
- Assessed in test and validation set

Smoothed Calibration



Summary

- See classic overfitting pattern
  - mortality lower than pred at high end
  - mortality higher than pred at low end
- Large calibrations at high end
  - represent 5% of the prediction
  - fixing would complicate model

Validation

- Calibration: examination of model spec
  - largely internal
- Discrimination: measure separation
  - largely external
- Calibration can be fixed
- Discrimination dictates utility of model

Examining Improvement in Risk Prediction

Risk Model

- Existing set of risk factors for disease
- Novel marker or test
  genetic, proteomic, imaging, biomarker
- New marker add predictive information
- How much better are predictions?
  do they aid in clinical decision making

Examples

- Breast density for breast cancer risk
  Tice et al. 2008
- C-reactive protein for CV risk
  Cook et al. 2006
- PSA for prostate cancer risk

The Easy Question

and least interesting

- Statistically significant contribution?
- Fit a logistic/Cox regression model
- Examine p-value for new measure
  adjusting for other factors
- Gives p-value
- Implies better prediction
  most compelling in validation data

Why not interesting?

- An easy threshold
- Many factors can reach this
- Statistical significant ≠ clinical significance
- Relevant summary needed
  on which to gauge clinical implications

Measure of Prediction

- TPF = true positive fraction
  proportion of case "positive"
  aka: sensitivity
- FPF = false positive fraction
  proportion of control "positive"
  1-specificity
- Trade-off
- Two kinds of errors
  context dependent trade off

Prediction/OR Disconnect

- Potentially misleading measure of predict.
  -- odds ratio (OR)
- Moderate to large OR not enough for prediction
- Need a huge odds ratio for a variable to imply good prediction

Odds Ratio of 3.0

Test

	+	-	
Disease +	90	10	100
-			900
			1000

OR = 3.0

Sens = 0.90

Odds Ratio of 3.0

Test

	+	-	
Disease +	90	10	100
-	675	225	900
	765	235	1000

OR = 3.0

Sens = 0.90

Spec = 0.25

PVP = 0.12

Odds Ratio of 36.0

Test

	+	-	
+	90	10	100
-			900
			1000

Disease

OR = 36.0

Sens = 0.90

Odds Ratio of 36.0

Test

	+	-	
+	90	10	100
-	720	180	900
	810	235	1000

Disease

OR = 36.0

Sens = 0.90

Spec = 0.80

PVP = 0.89

Dist of Marker and OR

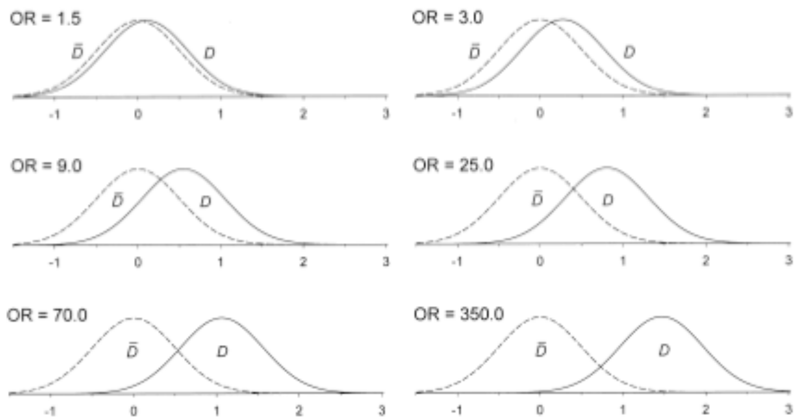


FIGURE 2. Probability distributions of a marker, X, in cases (solid curves) and controls (dashed curves) consistent with the logistic model logit$P(D = 1|X) = \alpha + \beta X$. It has been assumed that X has a mean of 0 and a standard deviation of 0.5 in controls so that a unit increase represents the difference between the 84th and 16th percentiles of X in controls. The marker is normally distributed, with the same variance in cases. The odds ratio (OR) per unit increase in X is shown.

Am J Epidemiol 2004;159:882–890

ROC Curves

- Two tests graphed =>
  FPF v. TPF (varying threshold)
  aligned by FPF
- Don't display absolute risk values
- C-statistic doesn't relate to clinical decisions
  measure ability to classify

Tice at al 2008

- Breast cancer risk
- SEER data
- age, race/ethnicity, family hx, biopsy hx
- added value of breast density
  HR 4.2 (less than 65), 2.2 (65 or older)
  statistically significant
- 5 year risk model (Cox model)
- How much does density add to risk assessment?

Examine Risk Strata

- Based on results of a model, calculate risks
- Group risks in context dependent way
  Tice: <1%, 1-1.66%, 1.67-2.5%, > 2.5%
- Compare between models
  what happens with good predictive model

Without Breast Density

Strata	<1%	1.0-1.66%	1.67-2.5%	> 2.5%
No. Women	215,402	201,972	148,795	63,105
% event	0.7%	1.3%	1.8%	2.9%

With Density

Strata	<1%	1.0-1.66%	1.67-2.5%	> 2.5%
No. Women	249,959	186,106	124,420	68,744
% event	0.7%	1.3%	2%	3.1%

Calibration

- % events by category
- Both models well calibrated
- % events observed falls within expected
- Calibration has subjective aspect
  if model predicts 2% but is really 3%
  close enough?

Net Reclassification

- Compare two prediction models
  one includes extra predictor
  like breast density
- Cross-tabulate risk predictions based on model
- Examine who gets reclassified
- What % and in which direction?

Table 1. Five-Year Risks for Breast Cancer as Predicted by Models That Do and Do Not Include Breast Density*

5-Year Risk from Model without Breast Density	5-Year Risk from Model with Breast Density				Total
	0% to <1%	1% to 1.66%	1.67% to 2.5%	>2.5%	
0% to <1%					
Women, *n*	176 831	38 500	71	0	215 402
Events, *n*	1161	415	0	0	1576
Nonevents, *n*	175 670	38 085	71	0	213 826
Proportion of women with events	0.7	1.1	0.0	–	0.7
1% to 1.66%					
Women, *n*	64 297	99 456	37 149	1025	201 927
Events, *n*	526	1328	754	32	2640
Nonevents, *n*	63 771	98 128	36 395	993	199 287
Proportion of women with events	0.8	1.3	2.0	3.1	1.3
1.67% to 2.5%					
Women, *n*	8741	45 478	71 309	23 267	148 795
Events, *n*	74	609	1419	621	2723
Nonevents, *n*	8667	44 869	69 890	22 646	146 072
Proportion of women with events	0.9	1.3	2.0	2.7	1.8
>2.5%					
Women, *n*	90	2672	15 891	44 452	63 105
Events, *n*	0	38	340	1467	1845
Nonevents, *n*	90	2634	15 551	42 985	61 260
Proportion of women with events	0.0	1.4	2.1	3.3	2.9
Total					
Women, *n*	249 959	186 106	124 420	68 744	629 229
Events, *n*	1761	2390	2513	2120	8784
Nonevents, *n*	248 198	183 716	121 907	66 624	620 445
Proportion of women with events	0.7	1.3	2.0	3.1	1.4

Reclassification Index

breast cancer cases

Include Density

		<1%	1.0-1.66%	1.67-2.5%	> 2.5%
No	<1%	1161	415	0	0
Dens	1.0-1.66%	526	1328	754	32
	1.67-2.5%	74	609	1419	621
	> 2.5%	0	38	340	1467

Reclassification

cases

- Risk strata with density
  - same: 5375 (61%)
  - lower: 1587 (18%)
    less predictive direction
  - higher: 1822 (21%)
    more predictive direction
- Net reclassification = 21%-18% = 3%
  in the more predictive direction
  3% = 235/8784

Reclassification Index

non-cases

Include Density

		<1%	1.0-1.66%	1.67-2.5%	> 2.5%
No	<1%	175,670	38,085	71	0
Dens	1.0-1.66%	63,771	98,128	36,395	993
	1.67-2.5%	8,667	44,869	69,890	22,646
	> 2.5%	90	2,634	15,551	42,985

Reclassification

controls

- Risk strata with density
- same: 386,673 (62%)
- lower: 135,882 (22%)
  more predictive direction
- higher: 98,190 (16%)
  less predictive direction
- Net reclassification = 22%-16% = 6%
  in the more predictive direction
  6% = 37,392/606,455

Cutpoints and Net Reclassification

- Depends intimately on the cutpoints
- Risk 2.9% (no density) v. 10% (density)
  no reclassification -- both in high strata
- Is this a bug or is this a feature?
- Some changes may not
  50% risk v. 70% risk

Hypothetical Clinical Trial

- Suppose wanted to know public health impact of adding marker
- RCT: standard v. augmented risk assessment
- Outcome: development of disease
- Presupposes that is some intervention

Public Health Impact

- Draws attention to issues which skulk in background in these debates
- If there is no plausible intervention, what is the health value of improved prediction?
- Is there a reasonable threshold for intervention?
- How much could intervention reduce risk in the population?

Breast Density

- Action point: risk of 1.67% over 5 years
  risk < 1.67: no intervention
  risk ≥ 1.67: intervention
- Intervention: chemoprophylaxis, enhanced screening, ?
- Intervention effect: 50% reduction in risk
- Starting point for a scenario

Table 1. **Five-Year Risks for Breast Cancer as Predicted by Models That Do and Do Not Include Breast Density***

5-Year Risk from Model without Breast Density	5-Year Risk from Model with Breast Density				Total
	0% to <1%	1% to 1.66%	1.67% to 2.5%	>2.5%	
0% to <1%					
Women, *n*	176 831	38 500	71	0	215 402
Events, *n*	1161	415	0	0	1576
Nonevents, *n*	175 670	38 085	71	0	213 826
Proportion of women with events	0.7	1.1	0.0	–	0.7
1% to 1.66%					
Women, *n*	64 297	99 456	37 149	1025	201 927
Events, *n*	526	1328	754	32	2640
Nonevents, *n*	63 771	98 128	36 395	993	199 287
Proportion of women with events	0.8	1.3	2.0	3.1	1.3
1.67% to 2.5%					
Women, *n*	8741	45 478	71 309	23 267	148 795
Events, *n*	74	609	1419	621	2723
Nonevents, *n*	8667	44 869	69 890	22 646	146 072
Proportion of women with events	0.9	1.3	2.0	2.7	1.8
>2.5%					
Women, *n*	90	2672	15 891	44 452	63 105
Events, *n*	0	38	340	1467	1845
Nonevents, *n*	90	2634	15 551	42 985	61 260
Proportion of women with events	0.0	1.4	2.1	3.3	2.9
Total					
Women, *n*	249 959	186 106	124 420	68 744	629 229
Events, *n*	1761	2390	2513	2120	8784
Nonevents, *n*	248 198	183 716	121 907	66 624	620 445
Proportion of women with events	0.7	1.3	2.0	3.1	1.4

Population

- Regular v. Augmented Screening
- 629, 229 women in each group
- 8,784 breast cancer cases anticipated

Regular Screening

- N=629,229
- 211, 900 have risk > 1.67%
  have some kind of intervention
- 4216 cancers in < 1.67%
- 4568 cancers in > 1.67%

Augmented Screening

- N=629,229
- 193,164 have risk > 1.67%
  have some kind of intervention
- 4151 cancers in < 1.67%
- 4633 cancers in > 1.67%

Effect of Screening

- 65 additional cases would get intervention
- With 50% benefit of intervention
- 32.5 cases of cancer averted
- Net efficacy: 33/ 8784 = 0.4% of cases averted
- Population RR = 0.996

Case Reclassification

- 2 x 2 table using 1.67%
- Net case reclassification is 0.007
- TPF (std) = 4568/8784 = 0.52
- TPF (aug) = 4633/8784 = 0.527
- Difference in TPF = 0.007

Net Reclassification

- Consider 2 x 2 classification
- NRI_{case} is difference of TPF
- $NRI_{control}$ is difference of FPF
- $RR_{pop} = 1 - (NRI_{case})*RR$
  $0.996 = 1 - (0.007)*0.5$
- 100% prevention only RRpop = 0.993

Public Health Impact

- 65 cases prevented
- NNT = 2,971
- Treated standard: 211,900
- Treated augmented: 193,164
- Reduces treated by 3% in non-cases

Odds Ratio

- Risk factor may have elevated OR
- Individual risk is modified by risk factor
- Population risk may not be greatly modified
- Especially true if risk factors cluster
- Both embody some level of truth

C-statistic

- Very difficult to budge
- Measure pure ability to classify
- However, strict classification is rarely of interest
- More interested in risk stratification

Net Reclassification

- Net reclassification tied to cutpoints
- Map to clinical decision rule
- Has correspondence to hypothetical clinical trial
- Devil is in the details:
  where and how many cutpoints

Summary

- Must balance false positives and false negatives
- Utilities may be very different for different people
- Attempt to relate to clinically relevant measures: individual or population level
- NRI: population level effects