Multinomial Logistic Regression for CRC Stage based on Size

CRC-SPIN v1.0 uses a multinomial logistic regression model to predict CRC stage, based on the American Joint Committee on Cancer (AJCC) guidelines, that are conditional on CRC size.[1] The model was developed using SEER data for colorectal cancer cases diagnosed between 1975-1979. However, the parameterization of CRC-SPIN v1.0 is not described in the literature.

Although the authors of the CRC-SPIN CISNET profile admit that the dependency between CRC stage and size at clinical detection is weak,[1] linking both variables enables a direct method to confer a greater survival benefit for early (ie, preclinical) CRC detection based on size. Upon initiation of a preclinical cancer, CRC-SPIN v1.0 assigns an initial size (0.5 mm) and a size upon transitioning to clinically detectable cancer (ie, the cancer size when ST expires). CRC-SPIN uses an exponential growth function to define CRC size at any time during ST. If CRC is detected during ST (ie, while it is preclinical), it will be smaller and could be assigned a less advanced AJCC stage compared to the same CRC upon clinical detection based on the logistic regression formula.

One consequence of this method is that it generates a different distribution of AJCC staging in a natural history modeling scenario compared to the other CISNET models (MISCAN and SimCRC) and the empirically observed distribution from SEER 1975-1979. In the absence of screening, CRC-SPIN generates fewer stage IV cancers and more stage II and stage III cancers.[2] In contrast, CRC-SPIN v2.x first simulates CRC stage at clinical detection, then the size at clinical detection, which is stratified by stage.[3] The authors explain that this inverted approach allows for greater flexibility in

specifying stage distribution, although their exact methodology is not clearly characterized.[3]

Here, we describe in detail our approach to develop a multinomial logistic regression model for CRC-AIM to predict AJCC stage conditional on CRC size.

CRC-AIM's multinomial logistic regression formula is based on 1975-1979 SEER data, which uses the code developed by Deborah Schrag to convert pre-1988 SEER registry data to AJCC staging categories[4] (see "AJCC staging for 1975-1979 SEER data" for more information.) The SEER query resulted in approximately 33,485 CRC cases with reported sizes from 1 mm to 97 mm (inclusive) and about 1,453 results (~4% of the data) coded as "≥98 mm," which we recoded to 98 mm (**Figure 1A**). We noted a monotonically decreasing proportion of stage I CRCs and a monotonically increasing proportion of stage IV CRCs conditioned on size (up to 85 mm). There are low counts for smaller CRC sizes—only 113 (0.34%) CRCs in the 1-5 mm inclusive size category, and only 42 (0.13%) CRCs in the 1-3 mm inclusive size category—which indicates uncertainty in the stage distribution of smaller CRCs. The implication of the empirical size distribution for these smaller lesions is that even if a 1 mm CRC were detected (roughly the size of CRC initiation in the model), the cancer would only have a 60% chance of being stage I. (Notably, the discrepancy in small polyp sizes could reflect changes in endoscopic technology[5] or variability in polyp size estimation,[6] although these hypotheses are beyond the scope of this analysis.)

To better understand the staging of small lesions, we referred to two other date ranges within the SEER dataset: 1988-1992 and 2011-2015 (**Figure 1B-C**). Both ranges indicate that very small CRC lesions (1 mm to 5 mm) are associated with ~90%

chance of being designated as stage I. We specified the following *a priori* conditions for the multinomial logistic regression model:

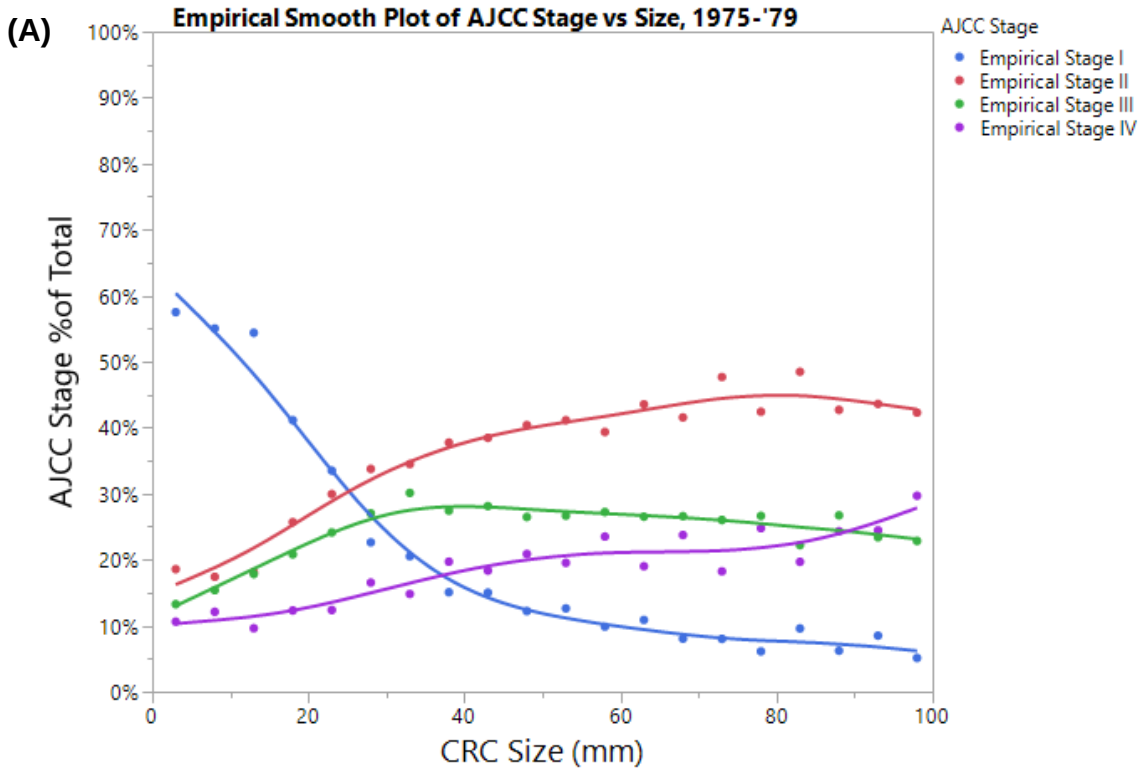
- allowing for 90% stage I CRC for very small lesions (1 mm)
- conforming to a monotonic function for stage I CRC, decreasing probability as size increases
- conforming to a monotonic function (≤85 mm) for stage IV CRC, increasing probability as size increases
- accurately extrapolating to CRCs ≥98 mm, since cancers are modeled up to 140 mm in size
- basing the model on the 1975-1979 SEER dataset

We parameterized the regression by fitting numerous multinomial logistic regression models to the SEER 1975-1979 data and considered various size transformations, including using a fractional polynomial approach. In all models, we treated size as a continuous factor and recoded a size of "≥98 mm" as 98 mm.

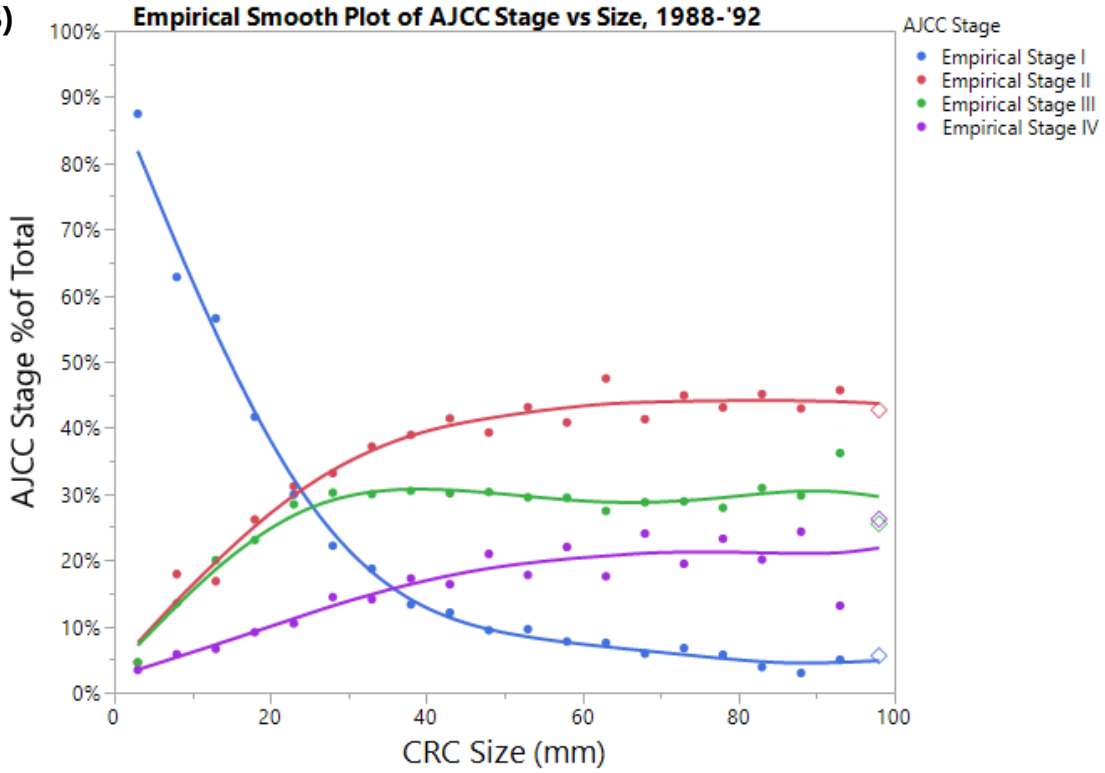
The final model that satisfies our pre-specified conditions uses fractional polynomials (-0.5, 1) for size, and the model is summarized in **Table 1**. Wald tests indicate both size terms in the model are statistically significant. Model RSquare is poor and the lack of fit is statistically significant. Area under the curve (AUC) is very poor for stage II (0.56) and stage III CRC (0.52), and only modestly informative for stage I (0.69) and stage IV (0.57) CRCs (**Figure 2**). Very similar outputs were obtained across other models, however, and as mentioned previously, the weak association between CRC size and stage was noted by the CRC-SPIN authors.[1]

The final multinomial logistic regression function is plotted as an overlay with the 1975-1979 SEER empirical data (**Figure 3**), with a vertical red line indicating the model-predicted stage distribution for CRC-AIM's extrapolated CRC sizes (ie, 98 mm to 140 mm).

Figure 1. Smoothed empirical distribution of American Joint Committee on Cancer (AJCC) stage vs size for different SEER datasets. The plots were generated from SEER data for the following date collection ranges: (A) 1975-1979; (B) 1988-1992; and (C) 2011-2015. For (B) and (C), the AJCC stage percentages for CRC sizes from 98 mm to 998 mm were combined and coded as 98 mm (symbolized as an open diamond) to visually match (A).

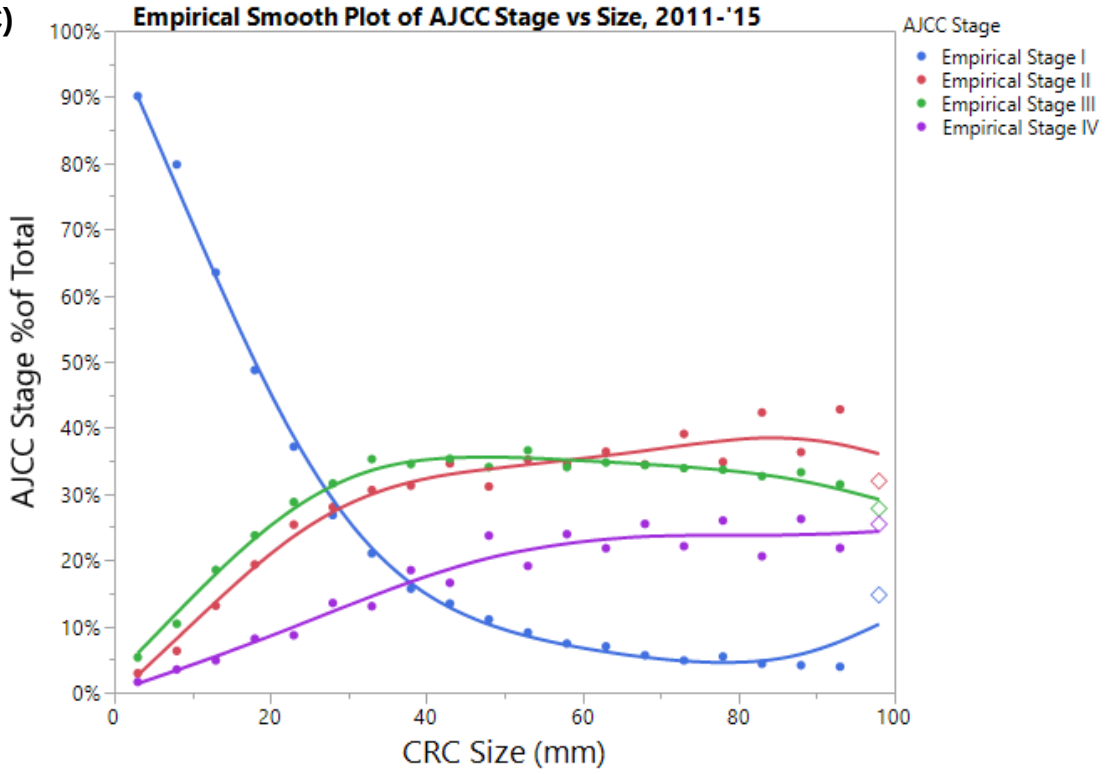


(B)


Empirical Smooth Plot of AJCC Stage vs Size, 1988-'92
AJCC Stage
Empirical Stage I
Empirical Stage II
Empirical Stage III
Empirical Stage IV
AJCC Stage %of Total
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0
20
40
60
80
100
CRC Size (mm)

81

(C)


Empirical Smooth Plot of AJCC Stage vs Size, 2011-'15
AJCC Stage
Empirical Stage I
Empirical Stage II
Empirical Stage III
Empirical Stage IV
AJCC Stage %of Total
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0
20
40
60
80
100
CRC Size (mm)

82

83

Figure 2. Receiver-operating characteristic (ROC) analysis and associated area under the curve for the final CRC-AIM multinomial logistic regression model.

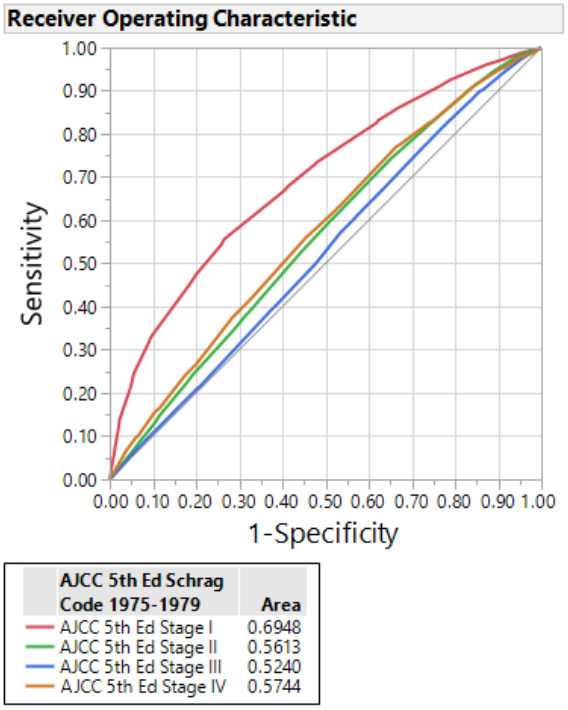


Figure 3. Multinomial logistic regression function overlaying the 1975-1979 SEER data. The vertical red line indicates the threshold for model-predicted stage distribution for CRC-AIM's extrapolated CRC sizes.

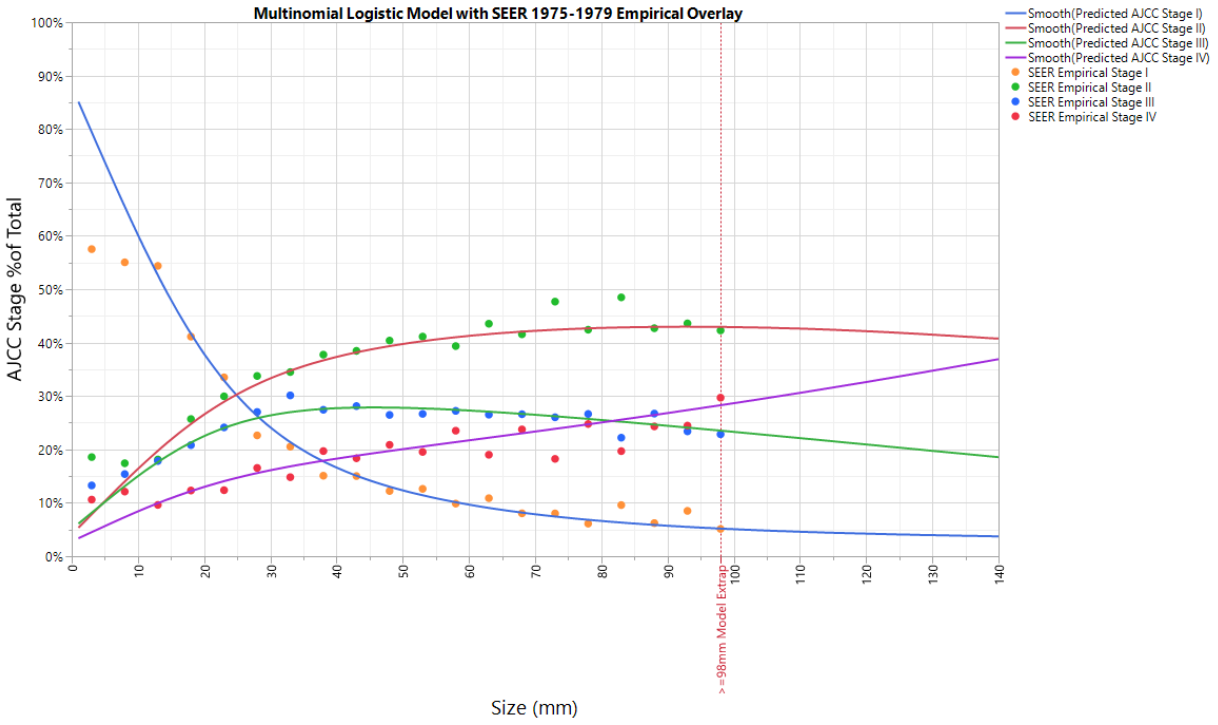


Table 1. Diagnostics, coefficients and effect tests for final multinomial logistic regression model of colorectal cancer (CRC) stage conditioned on CRC size.

Probabilities that achieved statistical significance (α ≤ 0.05) are in bold.

Whole Model Test				
Model	**-LogLikelihood**	**DF**	**ChiSquare**	**Prob>ChiSq**
Difference	1347.124	6	2694.248	**<.0001**
Full	43493.676			
Reduced	44840.800			
RSquare (U)	0.0300			
AICc	87005.4			
BIC	87081.1			
Observations (or Sum Wgts)	33485			
Lack of Fit				
Source	**DF**	**-LogLikelihood**	**ChiSquare**	**Prob>ChiSq**
Lack of Fit	282	245.133	490.2656	**<.0001**
Saturated	288	43248.543		
Fitted	6	43493.676		
Parameter Estimates				
Term	**Estimate**	**Std Error**	**ChiSquare**	**Prob>ChiSq**
Intercept $[\alpha_1]$	4.45118078	0.2818035	249.49	**<.0001**
Sqrt(Size (mm)) $[\beta_1]$	-0.8845893	0.0854211	107.24	**<.0001**
Size (mm) $[\gamma_1]$	0.02665437	0.0063795	17.46	**<.0001**
Intercept $[\alpha_2]$	0.29839421	0.2763527	1.17	0.2802
Sqrt(Size (mm)) β_2	0.16069904	0.0788434	4.15	**0.0415**
Size (mm) $[\gamma_2]$	-0.015023	0.0055023	7.45	**0.0063**
Intercept $[\alpha_3]$	0.41562304	0.2894097	2.06	0.1510
Sqrt(Size (mm)) β_3	0.10670585	0.0833045	1.64	0.2002
Size (mm) $[\gamma_3]$	-0.0169	0.0058694	8.29	**0.0040**
Effect Wald Tests				
Source	**Nparm**	**DF**	**Wald ChiSquare**	**Prob>ChiSq**
Sqrt(Size (mm))	3	3	242.02246	**<.0001**
Size (mm)	3	3	69.3294212	**<.0001**

References

1. CISNET Colorectal Cancer Collaborators. RAND Corporation (CRC-SPIN), 2015. HI.001.03112015.70373:https://cisnet.cancer.gov/colorectal/profiles.html.
2. Zauber A, Knudsen AB, Rutter CM, Lansdorp-Vogelaar I, Kuntz KM. Technical Report: Evaluating the benefits and harms of colorectal cancer screening strategies: a collaborative modeling approach. 2015; AHRQ Publication No. 14-05203-EF-2:https://www.uspreventiveservicestaskforce.org/Home/GetFile/1/16540/cisnet-draft-modeling-report/pdf.
3. CISNET Colorectal Cancer Collaborators. RAND Corporation (CRC-SPIN), 2018. HI.001.11302018.9737:https://cisnet.cancer.gov/colorectal/profiles.html.
4. Schrag D. AJCC Staging: Staging Colon Cancer Patients using the TNM system from 1975 to the present. 2007; https://www.mskcc.org/departments/epidemiology-biostatistics/epidemiology/ajcc-staging.
5. Sivak MV. Gastrointestinal endoscopy: past and future. *Gut.* 2006;55(8):1061-1064.
6. Summers RM. Polyp size measurement at CT colonography: what do we know and what do we need to know? *Radiology.* 2010;255(3):707-720.