BIOS 648 Analysis of High Dimensional Data

Final Project

Instructor: Chengcheng Hu, Ph.D.

March 5, 2020

Final Project

- Use any method(s) to analyze **one** dataset:

  1. A dataset of your choice (with instructor's prior approval), or
  2. ACTG 384 HIV resistance data (regression problem), or
  3. Bladder cancer karyometry data (classification problem)

- Write a report about your findings consisting of

  1. Background
  2. Description of data
  3. Statistical Methods
  4. Results
  5. Conclusions

- Report due by 5pm on Tuesday, 5/12

Option 1: Data of Your Choice

For this option you will need to

- Identify a high-dimensional dataset
- Have a clear objective
- Upload to the course website on or before Thursday 3/26 a brief description of the project of your choice
- Carry out the analysis and submit a report

Option 2: HIV Resistance Mutations

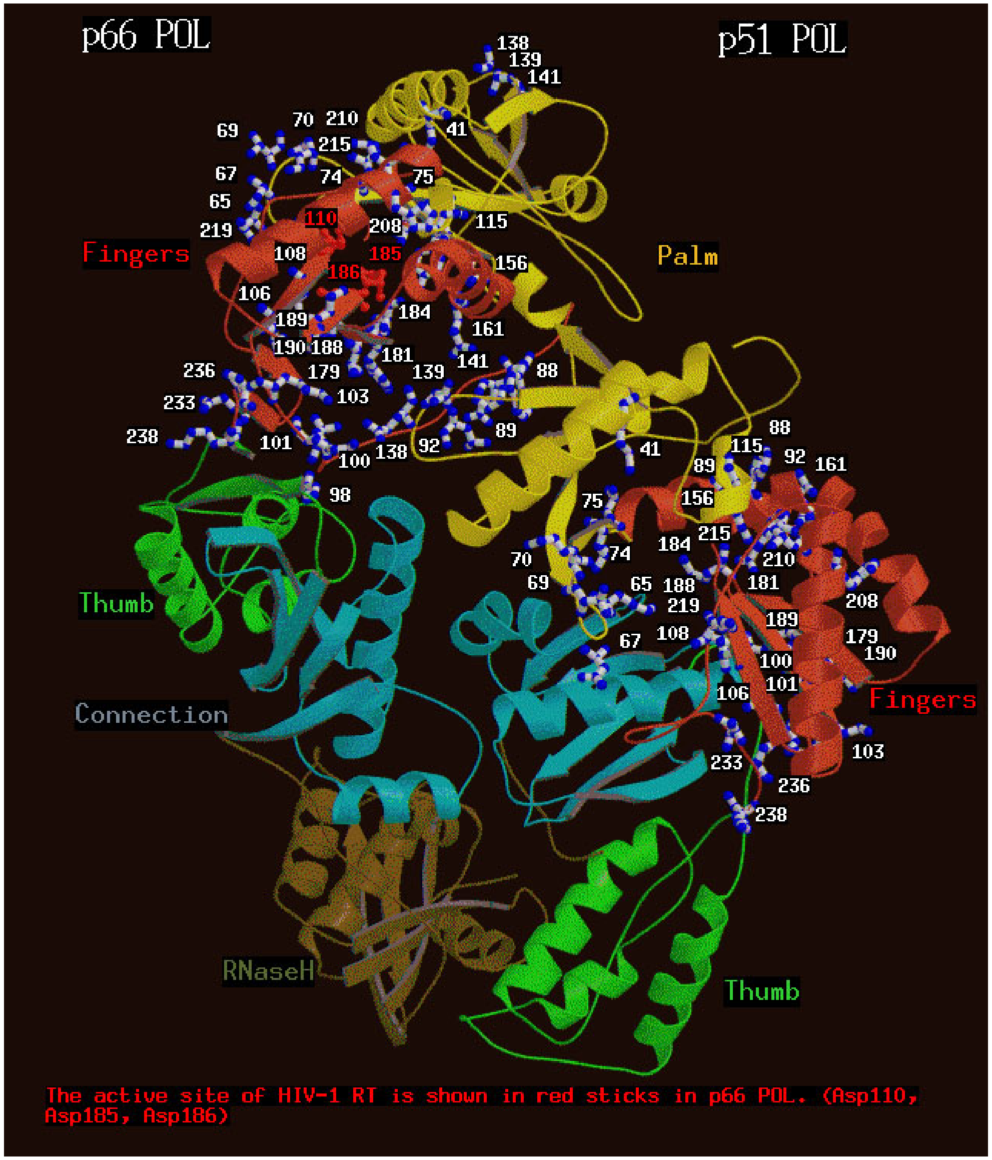
- Many antiretroviral drugs available
  - Nucleoside/Nonnucleoside reverse transcriptase inhibitors (NRTIs and NNRTIs)
  - Protease inhibitors (PIs)
  - Fusion inhibitors, entry inhibitors, etc.

- Mutations resistant to drugs very common under drug pressure

- Combination therapy: potent in suppressing viral load

HIV RT Protein

- P51 (codons 1 - 428)
- P66 (codons 1 - 558)
- Resistance mutation sites labeled
- The beginning 200 ∼ 300 codons are often sequenced in HIV clinical trials



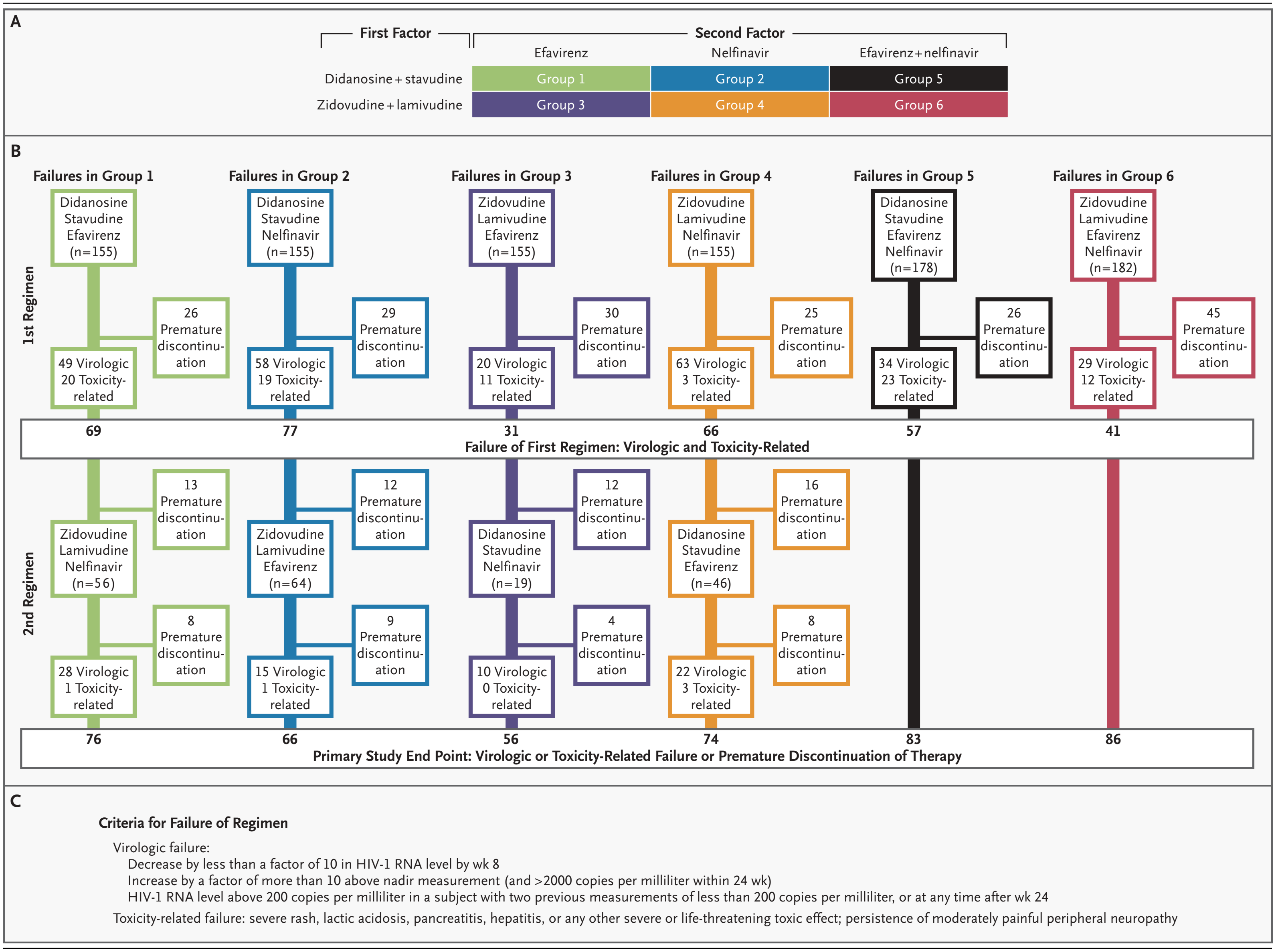
The active site of HIV-1 RT is shown in red sticks in p66 POL. (Asp110, Asp185, Asp186)

ACTG 384 Study

- Randomized clinical trial of six arms (different drugs and/or different orders of drugs)

- Nearly 900 subjects with viral sequence data at study entry

- Outcome measures for effectiveness of treatment: viral load and CD4 cell count



A

	First Factor	Second Factor		
		Efavirenz	Nelfinavir	Efavirenz + nelfinavir
	Didanosine + stavudine	Group 1	Group 2	Group 5
	Zidovudine + lamivudine	Group 3	Group 4	Group 6

B

1st Regimen
2nd Regimen
Failures in Group 1
Didanosine Stavudine Efavirenz (n=155)
26 Premature discontinu-ation
49 Virologic 20 Toxicity-related
69
13 Premature discontinu-ation
Zidovudine Lamivudine Nelfinavir (n=56)
8 Premature discontinu-ation
28 Virologic 1 Toxicity-related
76
Failures in Group 2
Didanosine Stavudine Nelfinavir (n=155)
29 Premature discontinu-ation
58 Virologic 19 Toxicity-related
77
12 Premature discontinu-ation
Zidovudine Lamivudine Efavirenz (n=64)
9 Premature discontinu-ation
15 Virologic 1 Toxicity-related
66
Failures in Group 3
Zidovudine Lamivudine Efavirenz (n=155)
30 Premature discontinu-ation
20 Virologic 11 Toxicity-related
31
12 Premature discontinu-ation
Didanosine Stavudine Nelfinavir (n=19)
4 Premature discontinu-ation
10 Virologic 0 Toxicity-related
56
Failures in Group 4
Zidovudine Lamivudine Nelfinavir (n=155)
25 Premature discontinu-ation
63 Virologic 3 Toxicity-related
66
16 Premature discontinu-ation
Didanosine Stavudine Efavirenz (n=46)
8 Premature discontinu-ation
22 Virologic 3 Toxicity-related
74
Failures in Group 5
Didanosine Stavudine Efavirenz Nelfinavir (n=178)
26 Premature discontinu-ation
34 Virologic 23 Toxicity-related
57
83
Failures in Group 6
Zidovudine Lamivudine Efavirenz Nelfinavir (n=182)
45 Premature discontinu-ation
29 Virologic 12 Toxicity-related
41
86
Failure of First Regimen: Virologic and Toxicity-Related
Primary Study End Point: Virologic or Toxicity-Related Failure or Premature Discontinuation of Therapy

C

Criteria for Failure of Regimen

Virologic failure:
Decrease by less than a factor of 10 in HIV-1 RNA level by wk 8
Increase by a factor of more than 10 above nadir measurement (and >2000 copies per milliliter within 24 wk)
HIV-1 RNA level above 200 copies per milliliter in a subject with two previous measurements of less than 200 copies per milliliter, or at any time after wk 24

Toxicity-related failure: severe rash, lactic acidosis, pancreatitis, hepatitis, or any other severe or life-threatening toxic effect; persistence of moderately painful peripheral neuropathy

From: Shafer *et al.* *NEJM* **349**, 2304 - 2315 (2003)

ACTG 384 Study: Predictors

- Treatment arm (categorical): A, B, C, D, E, F
- Baseline viral load (on the log10 scale) and CD4 count
- Mutation status (binary: 1=mutated and 0=wild-type) at the first 240 codons of the reverse transcriptase (RT) region and the first 99 codons of the protease (PR) region

ACTG 384 Study: Outcome Measures

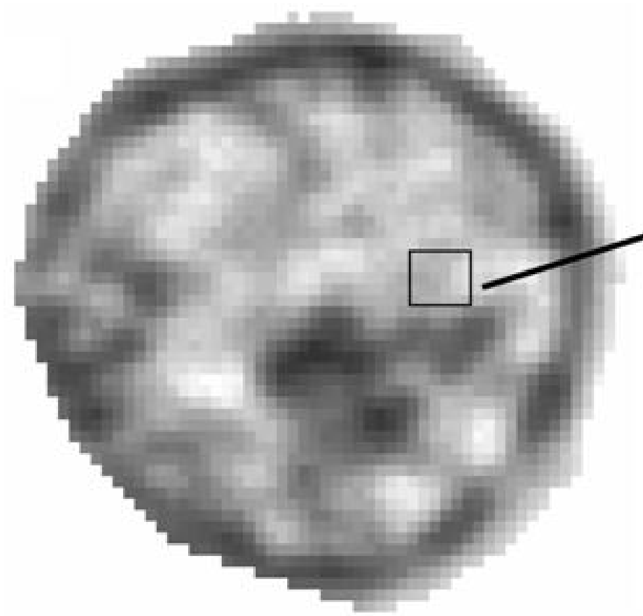
Build models to predict each of the following outcome measures **separately**:

- Last viral load (on the log10 scale) observed in the first year
- Last viral load (on the log10 scale) observed in the first two years
- Last CD4 counts observed in the first year
- Last CD4 counts observed in the first two years

Option 3: Karyometry

- Cancer diagnosis: visual examination of tumor tissues by pathologists
- Early detection is difficult: progression to cancer has started but might not be visible
- Karyometry: digital imaging of nuclear chromatin pattern to detect subtle deviation from normal

```
000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 001 000 000 000 000 000 000
000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000
000 000 000 000 000 000 000 000 006 019 021 012 010 012 008 000 000 000 000 000 000 000 000
000 001 000 000 000 000 000 009 027 026 028 023 016 021 023 025 017 000 000 000 000 000 000
000 000 000 000 000 000 015 017 025 028 033 026 026 024 027 022 026 013 000 000 000 000 000
000 000 000 000 000 000 024 027 035 030 030 027 030 028 036 033 031 023 000 000 000 000 000
000 000 000 000 000 000 025 036 042 037 029 034 035 035 038 037 032 021 000 000 000 000 000
000 000 000 000 000 009 035 045 045 043 037 034 039 039 035 044 031 017 000 000 000 000 001
000 000 000 000 004 016 039 049 052 045 032 036 037 034 047 049 031 005 000 000 000 000 000
000 000 000 000 001 017 037 051 050 048 031 037 037 038 043 043 031 000 000 000 000 000 000
000 001 000 000 000 017 036 048 043 053 053 047 044 042 047 037 028 000 000 000 000 000 000
000 000 000 000 000 014 027 039 044 057 060 056 056 051 044 028 010 000 000 000 000 001 000
000 000 000 000 000 009 023 028 041 046 049 051 051 037 033 011 000 000 000 000 000 001 000
000 000 000 000 000 001 004 017 017 027 031 035 029 022 007 000 000 000 000 000 000 000 000
000 000 000 000 000 000 002 001 001 000 000 003 004 000 000 000 000 000 000 000 000 000 000
000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000
000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000
000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000
000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000
```



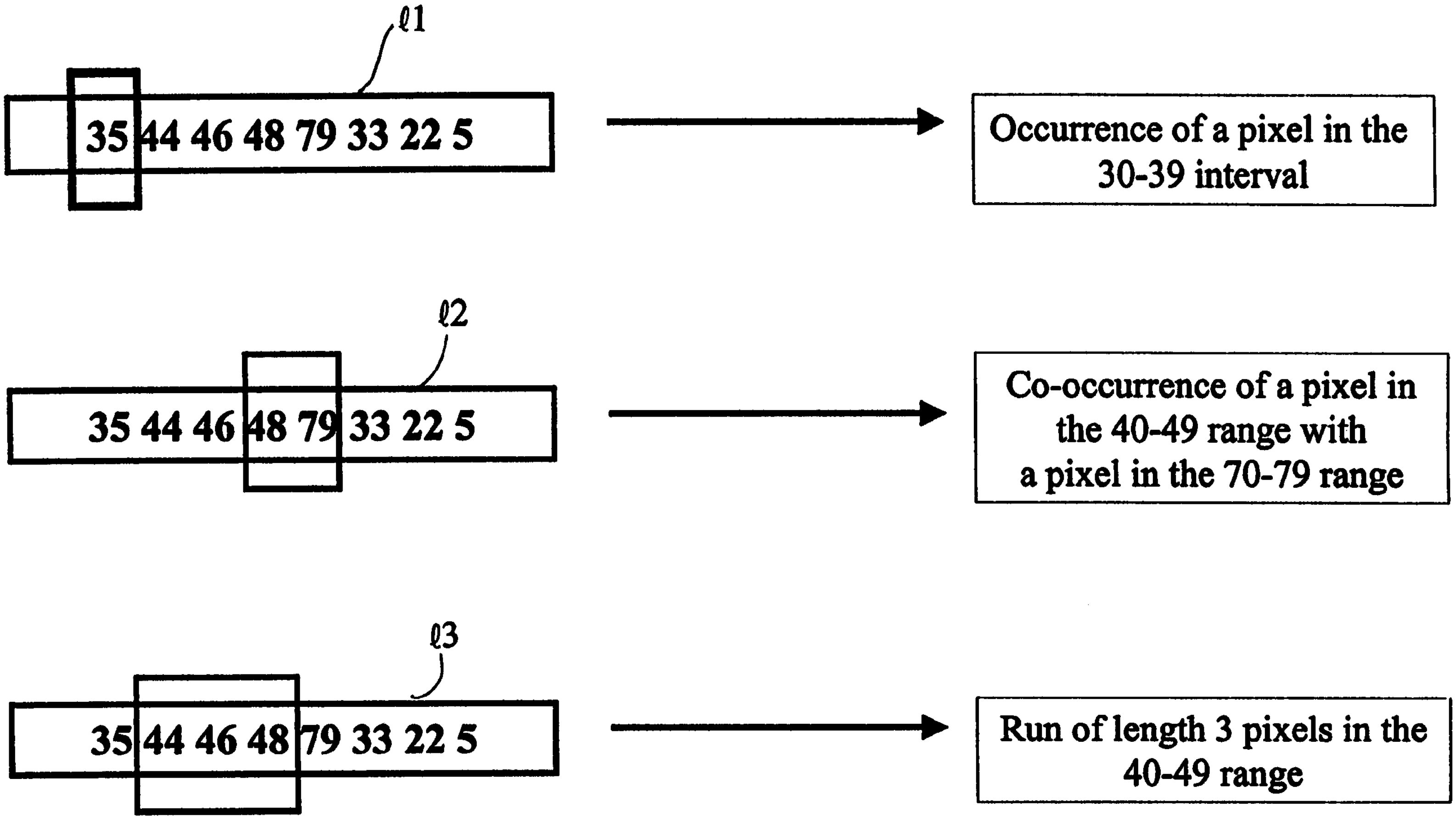
1 micron^2

Summary of Dataset

- A total of 40 bladder cancer patients had the tumor removed
    - 20 subjects were free of recurrence during a follow-up period of at least eight years
    - The other 20 subjects had one or more recurrences

- Around 100 nuclei from the original tumor were imaged for each subject

- A total of 92 features were extracted for each nucleus, such as total optical density (OD), nuclear area, nuclear roundness, OD std. dev., min axi/max axis, OD histogram, co-occurrence matrix, run length matrix, etc.



Occurrence of a pixel in the 30-39 interval

Co-occurrence of a pixel in the 40-49 range with a pixel in the 70-79 range

Run of length 3 pixels in the 40-49 range

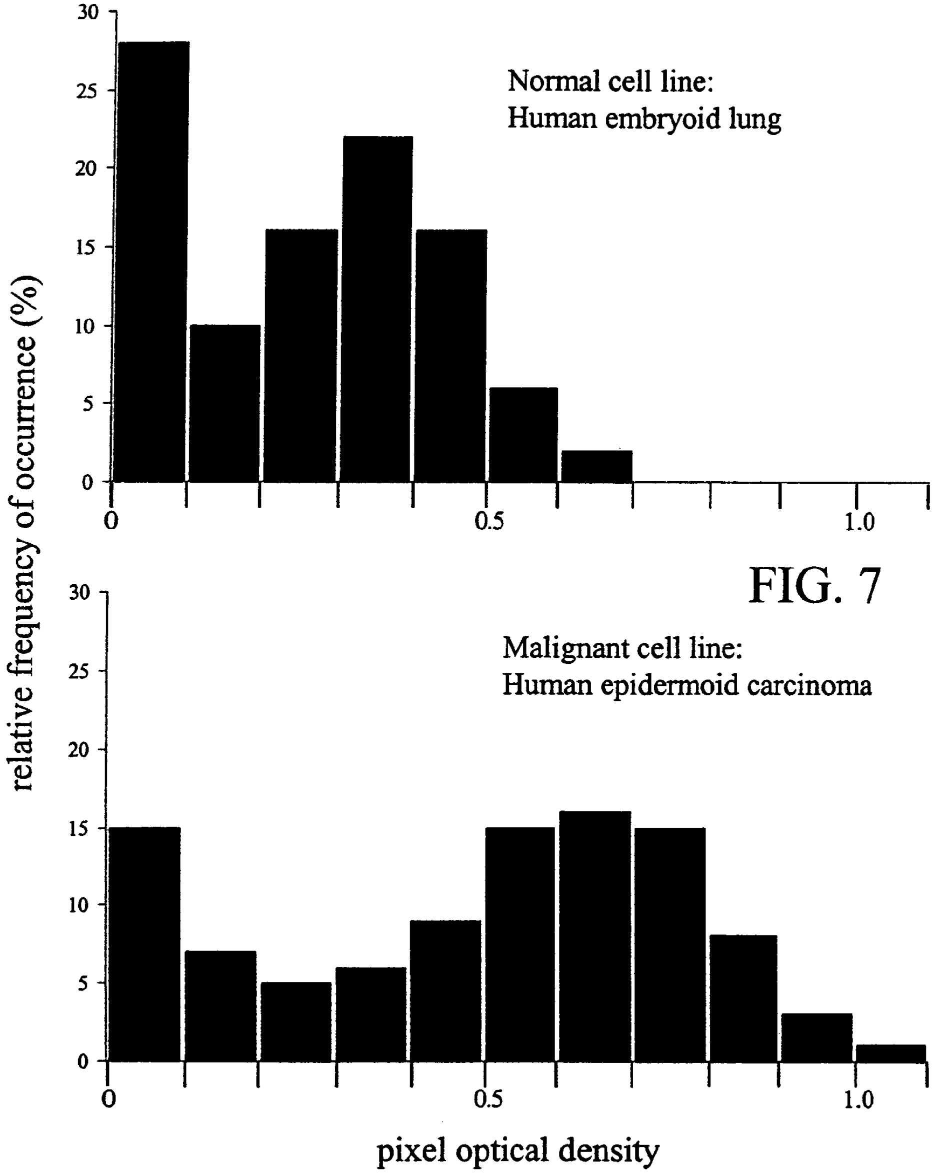


FIG. 7

Normal cell line:
Human embryoid lung

Malignant cell line:
Human epidermoid carcinoma

Objective

To predict recurrence based on karyometric features

Notes

1. The response (recurrent or not) is measured on the patient level, not on the nucleus level

2. Predictors include the 92 features, but they are measured about 100 times for each subject, once for each nucleus; thus a summary measure for each feature can be calculated for each subject, like the mean or median value across all $\sim$ 100 nuclei

3. Mean or median may or may not be the best predictor since cancer recurrence is likely to be caused by a small proportion of cells; thus you might also want to try variables like the 10th or 90th percentile of a certain feature within a subject, etc.

4. You can also include variability of certain features as potential predictors; be creative!

5. An alternative way is to work on the nucleus level and then summarize across all nuclei of the same patient to arrive at a predicted class for the patient

6. Performance of your model should be measured by a valid estimate of the misclassification error