University of BRISTOL

Switching from Excel to R: Experiences of a first-time R user developing a Markov model to compare coeliac screening strategies

Edna Keeney, Howard Thom

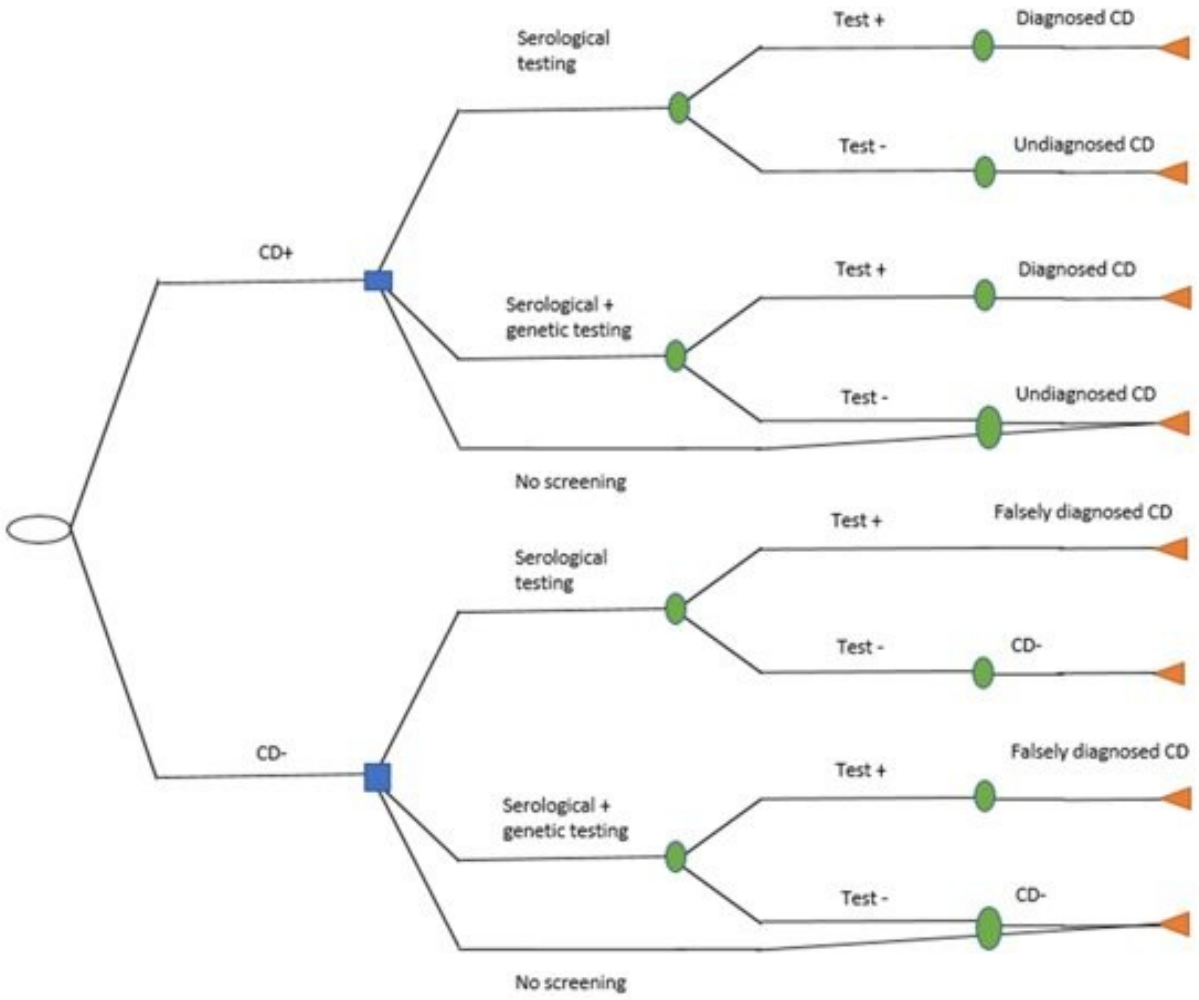
bristol.ac.uk

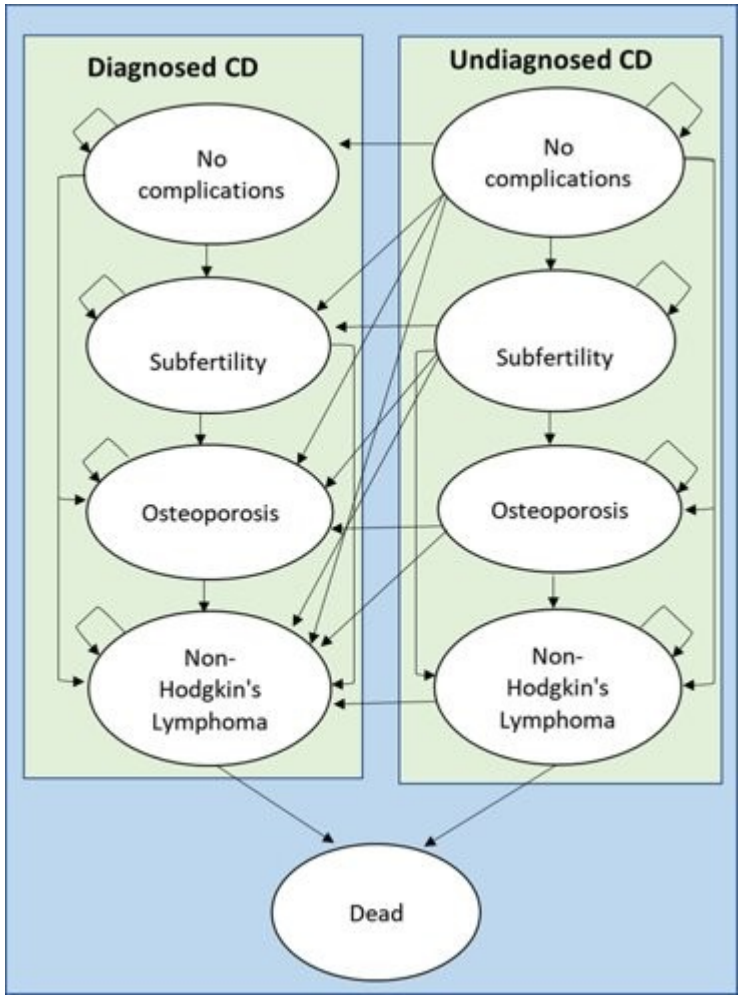
Decision question

- What is the most cost-effective combination of sensitivity and specificity for a risk factor test above which all patients should be screened for coeliac disease?
- What is the most cost-effective testing strategy?
- Stratified by age group (adults and children <16 years old)
- Over a lifetime horizon and from the perspective of the England & Wales NHS.

Model

- Decision tree/cohort Markov model





Inputting parameter estimates - Excel

Name	Live model value	Probabilistic	Mean	Standard error	Alpha	Beta	Distribution
Utilities							
Utility GFD adults	0.85	0.85	0.85	0.01	4162.48	734.56	Beta
Disutility subfertility	0.16	0.15	0.158	0.01	358.73	1911.70	Beta
Costs							
Cost hip fracture	$19,073	19136	19073.00	124.76	23372	1	Gamma
Cost IDA	$18	17	17.89	$2	100	0	Gamma
Cost GFP	$100	95	100	$10	100	1	Gamma

Pros	Cons
Can see estimates produced, good to give a sense check	Formulae are hidden
Visually pleasing	Errors easily produced

Inputting parameter estimates - R

```r
##############################################################################
## Utilities #################################################################
##############################################################################

utility_GFD_adults <- 0.85
utility_GFdse_adults <- ((0.86-0.84)/3.92)
utility_GFDalpha_adults <- (utility_GFD_adults ^ 2 * (1 - utility_GFD_adults)/utility_GFdse_adults ^ 2) - utility_GFD_adults
utility_GFDbeta_adults <- (utility_GFDalpha_adults / utility_GFD_adults) - utility_GFDalpha_adults
utility_GFD_adults <- rbeta(n = n_samples, shape1 = utility_GFDalpha_adults, shape2 = utility_GFDbeta_adults)

disutility_subfertility <-  0.158
disutility_subfertility_se <- (0.173 - 0.143)/3.92
disutility_subfertility_alpha <- (disutility_subfertility ^ 2 * (1 - disutility_subfertility)/disutility_subfertility_se ^ 2) - disutility_subfertility
disutility_subfertility_beta <- (disutility_subfertility_alpha/disutility_subfertility) - disutility_subfertility_alpha
disutility_subfertility <- rbeta(n = n_samples, shape1 = disutility_subfertility_alpha, shape2 = disutility_subfertility_beta)

##############################################################################
## Costs #####################################################################
##############################################################################

cost_hipfracture <- 19073
cost_hipfractureSE <- ((16515 * 1.17) - (16097 * 1.17)) / 3.92
cost_hipfracture_alpha <- (cost_hipfracture / cost_hipfractureSE) ^ 2
cost_hipfracture_beta <- (cost_hipfractureSE ^ 2) / cost_hipfracture
cost_hipfracture <- rgamma(n = n_samples, shape = cost_hipfracture_alpha, scale = cost_hipfracture_beta)

cost_IDA <- if(perspective == "NHS") 0 else 17.89
cost_gfp <- if(perspective == "NHS") 0 else 100
```

Pros	Cons
Can clearly see all formulae being used	Can't automatically see resulting estimates

Generating transition probabilities - Excel

- Excel worksheet showing probabilities of developing subfertility, osteoporosis or Non-Hodgkin's Lymphoma (NHL) for coeliac disease patients on a Gluten Free Diet (GFD) at different ages.
- Can explicitly see estimated probabilities alongside parameters.
- Stored in a worksheet in excel file so easily accessible.

Name	Live model value	Probabilistic	Mean	Standard error	Alpha	Beta	Distribution
Transition probabilities							
Subfertility probability GFD 10-20	0.0009	0.00091	0.00089	0.00008	118.75	133364.46	Beta
Subfertility probability GFD 20-30	0.0015	0.00161	0.00149	0.00010	236.22	158421.91	Beta
Subfertility probability GFD 30-40	0.0010	0.00103	0.00104	0.00008	184.43	177246.93	Beta
Subfertility probability GFD 40-50	0.0001	0.00006	0.00008	0.00002	12.10	151203.54	Beta
Osteoporosis probability GFD 0-10	0.0000	0.00003	0.00003	1.02038E-06	752.95	26890793.35	Beta
Osteoporosis probability GFD 10-20	0.0000	0.00004	0.00004	1.53055E-06	752.95	17927033.73	Beta
Osteoporosis probability GFD 20-30	0.0001	0.00012	0.00013	4.59126E-06	752.95	5975437.12	Beta
Osteoporosis probability GFD 30-40	0.0004	0.00038	0.00036	1.32605E-05	752.95	2068171.05	Beta
Osteoporosis probability GFD 40-50	0.0010	0.00093	0.00097	3.51701E-05	752.95	779077.66	Beta
Osteoporosis probability GFD 50-60	0.0032	0.00317	0.00321	0.00011697	752.95	233458.19	Beta
Osteoporosis probability GFD 60-70	0.0074	0.00724	0.00739	0.000268409	752.95	101099.81	Beta
Osteoporosis probability GFD 70-80	0.0138	0.01354	0.01376	0.00049815	752.95	53949.84	Beta
Osteoporosis probability GFD 80-90	0.0208	0.02076	0.02078	0.000749403	752.96	35480.03	Beta
Osteoporosis probability GFD 90-10	0.0180	0.01826	0.01804	0.000651303	752.96	40995.90	Beta
NHL probability GFD	0.0001	0.00016	0.0001	0.00009	1.14	11546.99	Beta

Generating transition probabilities - R

- In R, transition matrices are actual matrices, rather than rows of individual probabilities
- transition_matrices <- array(dim = c(n_samples, n_cycles, n_states, n_states),
  dimnames = list(NULL, NULL, state_names, state_names))
- One transition matrix for each sample and each cycle.

	CD GFD no complications	CD GFD subfertility	CD GFD osteoporosis	CD GFD NHL	Undiagnosed CD no complications	Undiagnosed CD subfertility	Undiagnosed CD osteoporosis	Undiagnosed CD NHL	Death
CD GFD no complications	0.9988688	0.0009226411	4.077212e-05	9.230989e-05	0.0000000	0.0000000000	0.000000e+00	0.000000e+00	0.00007550
CD GFD subfertility	0.0000000	0.9997914180	4.077212e-05	9.230989e-05	0.0000000	0.0000000000	0.000000e+00	0.000000e+00	0.00007550
CD GFD osteoporosis	0.0000000	0.0000000000	9.996797e-01	9.230989e-05	0.0000000	0.0000000000	0.000000e+00	0.000000e+00	0.00022795
CD GFD NHL	0.0000000	0.0000000000	0.000000e+00	8.886990e-01	0.0000000	0.0000000000	0.000000e+00	0.000000e+00	0.11130104
Undiagnosed CD no complications	0.2452858	0.0002608423	2.013561e-05	2.477695e-06	0.7534844	0.0008013298	6.185831e-05	7.611691e-06	0.00007550
Undiagnosed CD subfertility	0.0000000	0.2455517996	2.013561e-05	2.477695e-06	0.0000000	0.7542806171	6.185831e-05	7.611691e-06	0.00007550
Undiagnosed CD osteoporosis	0.0000000	0.0000000000	2.455719e-01	2.477695e-06	0.0000000	0.0000000000	7.541900e-01	7.611691e-06	0.00022795
Undiagnosed CD NHL	0.0000000	0.0000000000	0.000000e+00	2.455744e-01	0.0000000	0.0000000000	0.000000e+00	6.431245e-01	0.11130104
Death	0.0000000	0.0000000000	0.000000e+00	0.000000e+00	0.0000000	0.0000000000	0.000000e+00	0.000000e+00	1.00000000

Generating transition probabilities - R

```
subfertility_probability <- read.csv("data/subfertility.csv")
subfertility_probability_GFD_0  <- 0
subfertility_probability_GFD_10 <- rbeta(n=n_samples, shape1 = subfertility_probability$subfertility_GFD_alpha[2], shape2 = subfertility_probability$subfertility_GFD_beta[2])
subfertility_probability_GFD_20 <- rbeta(n=n_samples, shape1 = subfertility_probability$subfertility_GFD_alpha[3], shape2 = subfertility_probability$subfertility_GFD_beta[3])
subfertility_probability_GFD_30 <-  rbeta(n=n_samples, shape1 = subfertility_probability$subfertility_GFD_alpha[4], shape2 = subfertility_probability$subfertility_GFD_beta[4])
subfertility_probability_GFD_40 <- rbeta(n=n_samples, shape1 = subfertility_probability$subfertility_GFD_alpha[5], shape2 = subfertility_probability$subfertility_GFD_beta[5])
subfertility_probability_GFD_50 <- 0
subfertility_probability_GFD_60 <- 0
subfertility_probability_GFD_70 <- 0
subfertility_probability_GFD_80 <- 0
subfertility_probability_GFD_90 <- 0
subfertility_probability_GFD_all <- data.frame(subfertility_probability_GFD_0, subfertility_probability_GFD_10, subfertility_probability_GFD_20, subfertility_probability_GFD_30,
                                               subfertility_probability_GFD_40, subfertility_probability_GFD_50, subfertility_probability_GFD_60,
                                               subfertility_probability_GFD_70, subfertility_probability_GFD_80, subfertility_probability_GFD_90)

for(i_age_category in c(0:n_agecategories)) {
  transition_matrices[, (c(1:10) + i_age_category * 10), "CD GFD no complications", "CD GFD subfertility"] <- subfertility_probability_GFD_all[, starting_age_column + i_age_category]
  transition_matrices[, (c(1:10) + i_age_category * 10), "CD GFD no complications", "CD GFD osteoporosis"] <- osteoporosis_probability_GFD_all[, starting_age_column + i_age_category]
  transition_matrices[, (c(1:10) + i_age_category * 10), "CD GFD subfertility", "CD GFD osteoporosis"] <- osteoporosis_probability_GFD_all[, starting_age_column + i_age_category]
}
```

- Subfertility data stored in separate CSV file, not as transparent as having all data in one excel file.
- Data often provided in excel files so this is a limitation of using R

Cohort simulation - Excel

Age	Cycle	CD GFD no complications	CD GFD osteoporosis	CD GFD subfertility	CD GFD Non-Hodgkins Lymphoma
30	1	1000	0	0	0
31	2	998.498207	0.363934	1.039459	0.0984
32	3	996.9986694	0.363387446	1.037897948	0.098252224
33	4	995.5013838	0.362841714	1.03633924	0.098104669
34	5	994.0063467	0.362296801	1.034782873	0.097957336
35	6	992.513555	0.361752706	1.033228843	0.097810225
36	7	991.0230051	0.361209428	1.031677147	0.097663334
37	8	989.5346937	0.360666966	1.030127782	0.097516664
38	9	988.0486174	0.360125319	1.028580743	0.097370214
39	10	986.5647729	0.359584486	1.027036028	0.097223984
40	11	985.0831568	0.359044464	1.025493632	0.097077974
41	12	983.9562867	0.951131281	0.078806653	0.096932183
42	13	982.8307056	0.950043249	0.078716503	0.096821299
43	14	981.7064122	0.948956463	0.078626456	0.096710541
44	15	980.5834048	0.947870919	0.078536513	0.096599911
45	16	979.4616821	0.946786617	0.078446672	0.096489407

In Excel, need to be careful with transition probabilities that differ by age.

Cohort Simulation - R

cohort_vectors <- array(dim=c(n_tests, n_samples, n_cycles, n_states),
dimnames=list(t_names, NULL, NULL, state_names))

```
# Main model code
# Loop over the test options

for (i_test in 1:n_tests)
{
  # Loop over the PSA samples
  for(i_sample in 1:n_samples)
  {

    transition_matrices_sample <- transition_matrices[i_sample, , , ]
    # Loop over the cycles
    # Cycle 1 is already defined so only need to update cycles 2:n_cycles
    for(i_cycle in 2:n_cycles)
    {
      # Markov update
      # Multiply previous cycle's cohort vector by transition matrix
      # i.e. pi_j = pi_(j-1)*P
      cohort_vectors[i_test, i_sample, i_cycle, ] <-
        cohort_vectors[i_test, i_sample, i_cycle-1, ] %*%
        transition_matrices_sample[i_cycle, , ]
    }
```

Few short lines of code compared to whole sheet of probabilities

Cohort Simulation - R

- Although, not automatically visible, R's View() function can be used directly to explore cohort simulation.
- Not as easy to format tables to look nice, as in Excel.

	CD GFD no complications	CD GFD subfertility	CD GFD osteoporosis	CD GFD NHL	Undiagnosed CD no complications	Undiagnosed CD subfertility	Undiagnosed CD osteoporosis	Undiagnosed CD NHL	Death
1	0.01954219	0.0000348785	3.007273e-05	7.053763e-07	0.97710933	0.0017439249	0.0015036367	3.526882e-05	0.000000e+00
2	0.25919115	0.0007359970	4.198221e-04	1.352694e-05	0.73623666	0.0020983956	0.0011945780	3.014442e-05	7.973506e-05
3	0.43948454	0.0016822852	7.384954e-04	4.528901e-05	0.55473770	0.0021727328	0.0009465778	2.501569e-05	1.673652e-04
4	0.57504437	0.0027655912	9.998462e-04	8.856594e-05	0.41797068	0.0020833173	0.0007482636	2.033444e-05	2.790275e-04
5	0.67689574	0.0039160625	1.215190e-03	1.381746e-04	0.31491956	0.0019062642	0.0005902382	1.628059e-05	4.024932e-04
6	0.75332963	0.0050897442	1.393628e-03	1.905380e-04	0.23726572	0.0016900839	0.0004646256	1.288652e-05	5.631388e-04
7	0.81059137	0.0062600601	1.542533e-03	2.432270e-04	0.17874957	0.0014647372	0.0003650367	1.011003e-05	7.733553e-04
8	0.85338783	0.0074119194	1.667819e-03	2.946316e-04	0.13465606	0.0012478230	0.0002862665	7.876525e-06	1.039774e-03
9	0.88524857	0.0085375093	1.774176e-03	3.437248e-04	0.10142920	0.0010488156	0.0002240930	6.102224e-06	1.387810e-03
10	0.90890632	0.0096343693	1.865731e-03	3.898900e-04	0.07640011	0.0008721189	0.0001751695	4.706227e-06	1.751582e-03

Generating transition probabilities and cohort simulation - Comparison

- Transition matrices in R better to manage more complex tasks (such as age-dependent transition probabilities) due to use of higher dimensional arrays.
- Learning curve in understanding matrices and loops.
- However, theory of Markov models (taught as matrices multiplied by vectors) translates more directly to R than to excel.

Conclusions

- Obvious challenges with switching to R:
  - Learning code
  - Understanding loops, matrices, functions etc.
- Negatives of R for cost-effectiveness modelling:
  - Not as easy to visualize estimated transition probabilities and patient cohort to recognize potential errors
  - Data often stored separately
  - Not as widely used as excel so difficult when working with different teams without R expertise
- Positives of R for cost-effectiveness modelling:
  - Code always clearly visible, easier to spot errors in that way
  - Core pieces of code easy to adapt for different models
  - More flexible when requirements are more complex
  - Handy packages for plots etc.