Bayesian Uncertainty Modelling in A Mass Cytometry Experiment

Eric Barnhill and Shravan Vasishth

February 19, 2018

Some additional code and analyses by Shravan Vasishth (vasishth at uni-potsdam.de).

Introduction

The goal of this experiment was to estimate robustness and uncertainty of effects in a mass cytometry experiment. In this experiment, the response of contrast agents in six cell types was studied. As the cell types behave biologically differently, a separate statistical model was built for each.

In each model, three experiments were run, in which cell samples were subjected to three differing contrast agents:

- Magnevist
- Dotarem
- Gadevist

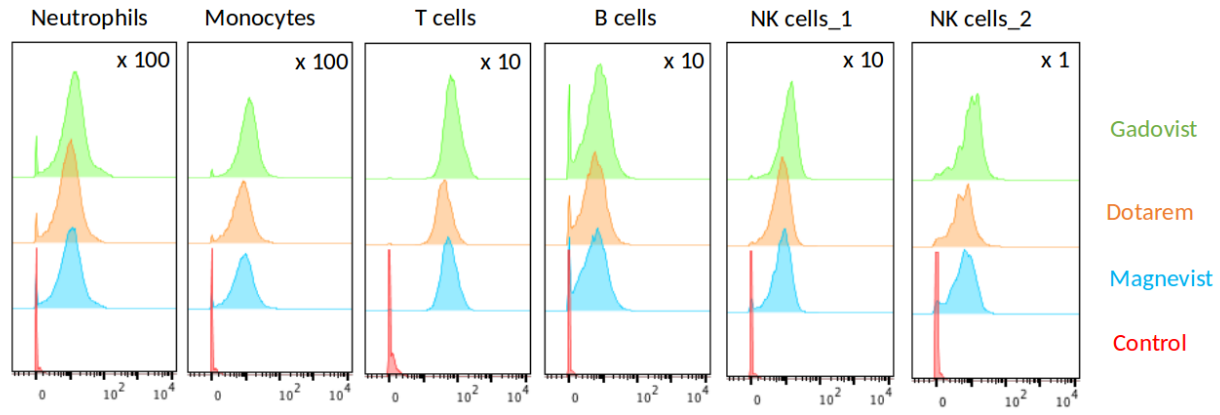
and a control condition. Cells were subjected to these agents at three differing concentrations: 0.1, 0.3, and 1, and the control condition was considered a concentration of 0.0. Thus a fully specified model for the experiment would be:

$$E(Y_{ijk}) = \beta_0 + u_{0j} + w_{0k} + (\beta_1 + u_{1j} + w_{1k})conc_i + \epsilon_i$$

where β_0 is the global intercept, u_{0j} is a random intercept for experiment, w_{0k} is a random intercept for contrast agent, and similarly, β_1, u_{1j} and w_{1k} are adjustments to the slope of the continuous concentration variable *conc*.

The data were acquired with a mass cytometer which provides several thousand measurements for each category. A summary plot from the mass cytometer for one experiment is below:

Figure 2: ^{158}Gd signal on major leukocyte subsets



For this experiment we asked the scientific question: Does the cell signal change with contrast agent, for a given concentration?

Interpretation of this data with a rigorous statistical model posed two major challenges:

- Mass cytometry data is rarely handled in its raw form, which contains millions of samples. Rather, the cytometer outputs four summary statistics: 10% quantile, mean, median, and 90% quantile. To model uncertainty, the underlying distribution of each experimental condition had to be estimated.
- Once the distributions for each condition were estimated, iterative model-building was required to estimate impact of contrast agent.

Data exploration

Data were cleaned and converted to tall format:

```
source('masscyto_clean_gather_dataNEW.R')
mass_cyto <- read.xls(xls="experimental_data.xlsx",
                      header=FALSE, skip=3, nrows=30)[,-(5:10)]
mass_cyto_tall <- clean_gather_data(mass_cyto)
#str(mass_cyto_tall)
#head(mass_cyto_tall)
#summary(mass_cyto_tall)
```

The research question: Does the cell signal change with contrast agent, for a given concentration?

Eric wrote:

"The main experimental question is whether certain of the contrast agents, which are safer, produce signal comparable to other contrast agents which are more dangerous. However, we do also want to investigate whether this statement holds at varying concentrations."

To me (SV), this sounds like we need to know if there is an interaction between ContrastAgent and concentration, but the crucial issue is: is there a specific expectation for how ContrastAgent will affect the dependent variable value? For now, I will assume that there is a natural ordering in level of danger:

- Magnevist is less dangerous than Dotarem
- Dotarem is less dangerous than Gadevist

These comparisons can be changed to reflect reality (if the above ordering is incorrect).

We will use sliding contrasts to reflect the above ordering:

```
library(MASS)
## reversing the signs by multiplying by -1:
comparisons<- -1*round(ginv(contr.sdif(3)))
rownames(comparisons)<-c("MvsD","DvsG")
comparisons
```

```
##      [,1] [,2] [,3]
## MvsD    1   -1    0
## DvsG    0    1   -1
```

Also, we would have to remove control from that particular analysis, because controls have only one level of the concentrations. But the effect of control can be estimated and used as a baseline (to-do).

Experiment needs to be modeled as a random effect because each experiment is actually a patient.

Computing standard deviation of the dependent variable "value" in preparation for measurement error modeling

We can just take logs on the mean and the quantile measurements and then get the measurement error on the log scale:

```
x<-rlnorm(1000,meanlog=1.2,sdlog=1)
qnorm(0.05,mean=1.2,sd=1) # 5% quantile
```

```
## [1] -0.4448536
```

```
qnorm(0.95,mean=1.2,sd=1) # 95% quantile
```

```
## [1] 2.844854
```

```
mean(log(x)) ## recovers MLE of lognormal
```

```
## [1] 1.222272
```

```
sd(log(x))   ## recovers sd of lognormal
```

```
## [1] 0.9870951
```

```
log(quantile(x,prob=c(0.05,0.95))) ## recovers 5th and 95th quantile
```

```
##         5%        95%
## -0.3221943  2.7734429
```

So, given the upper 95th percentile, we take the distance between this percentile value and the sample mean, and divide that by 1.64 or so to get an approximate standard deviation. We will use this in the measurement error model.

Modeling effect of concentration

In preparation for the measurement error model, we first
prepare the data so that there is a column for the uncertainty of each mean value in each row of the data frame:

```
## extract means:
means<-subset(mass_cyto_tall,MeasurementType=="Mean")
## extract lower quantiles:
qlow<-subset(mass_cyto_tall,MeasurementType=="pct_05")
## extract upper quantiles:
qhigh<-subset(mass_cyto_tall,MeasurementType=="pct_95")

## log scale difference between upper percentile and mean:
d<-log(qhigh$value)-log(means$value)

means$SD<-d/1.64

## center concentration:
means$cconc<-scale(as.numeric(as.character(means$Concentration)),scale=FALSE)

#head(means)

## rename cell type as numerical values:
means$typ<-unique(as.numeric(means$CellType))
```

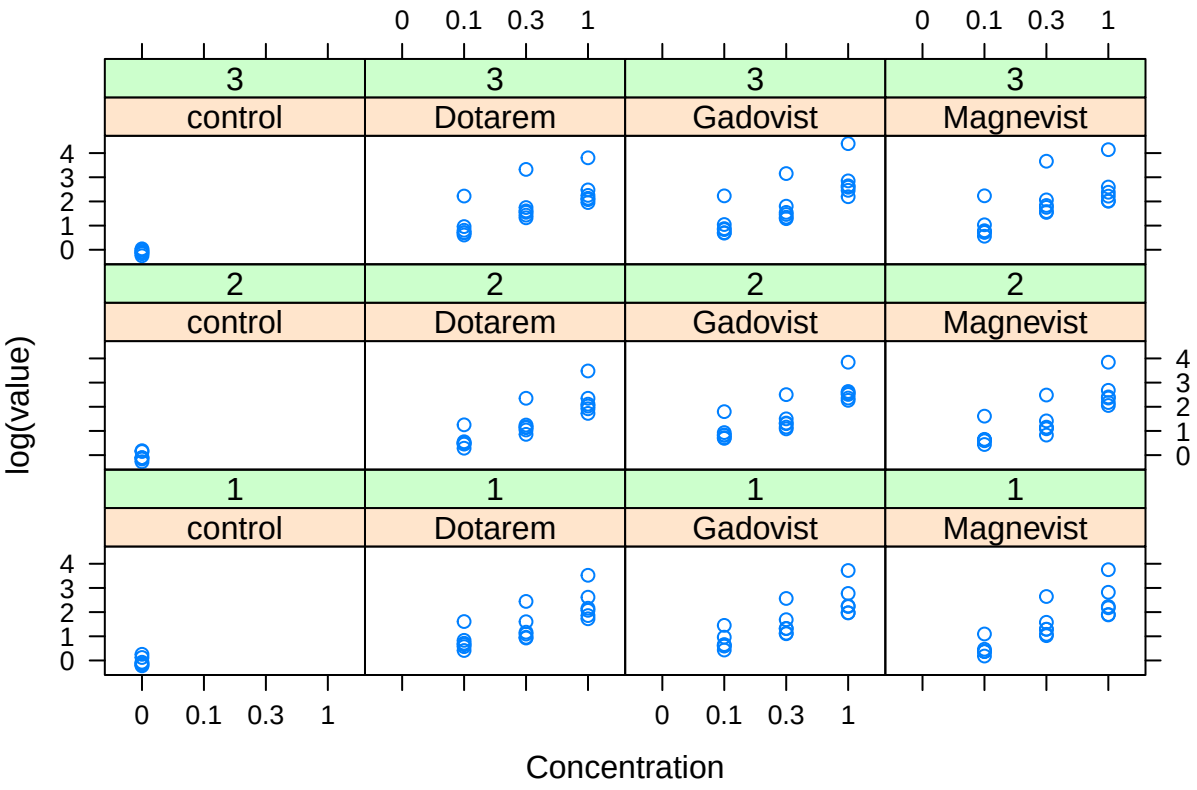
```
dat<-list(cconc=as.vector(means$cconc),
     value=log(means$value),
     SD=means$SD,
     typ=as.integer(means$typ),
     N = dim(means)[1],
     J = length(unique(means$typ))
     )

str(dat)
```

```
## List of 6
##  $ cconc: num [1:180] -0.42 -0.32 -0.12 0.58 -0.32 -0.12 0.58 -0.32 -0.12 0.58 ...
##  $ value: num [1:180] -0.223 0.464 1.579 2.821 0.833 ...
##  $ SD   : num [1:180] 0.527 0.625 0.611 0.621 0.627 ...
##  $ typ  : int [1:180] 3 6 2 1 4 5 3 6 2 1 ...
##  $ N    : int 180
##  $ J    : int 6
```

The first step is to visualize the effect of concentration by contrast agent and experiment, pooling data from all cell types:

```
library(lattice)
xyplot(log(value)~Concentration|ContrastAgent+Experiment,means)
```



We start by just modeling the effect of concentration on log(value). Concentration is centered and scaled to have standard deviation 1; this has the (small) advantage that the intercept now has the interpretation that it reflects the predicted log(value) when concentration is the average value.

Here is the standard hierarchical linear model of effect of concentration on log value. Cell Types and Experiment are treated as random effects. Here I am making the simplifying assumption that Cell Types and Experiment are independent—is this reasonable?

```
m1<-lmer(log(value)~cconc+(1+cconc|Experiment)+
          (1+cconc|CellType),
        subset(means,ContrastAgent!="control"))

print(summary(m1))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: log(value) ~ cconc + (1 + cconc | Experiment) + (1 + cconc |
##     CellType)
##    Data: subset(means, ContrastAgent != "control")
##
## REML criterion at convergence: 54.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.9462 -0.8274  0.0046  0.6786  4.0490
##
## Random effects:
##  Groups     Name        Variance Std.Dev. Corr
##  CellType   (Intercept) 0.32794  0.5727
##             cconc       0.04512  0.2124   1.00
##  Experiment (Intercept) 0.03271  0.1809
##             cconc       0.01808  0.1345   -1.00
##  Residual               0.06430  0.2536
## Number of obs: 162, groups:  CellType, 6; Experiment, 3
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   1.5599     0.2568   6.074
## cconc         1.7435     0.1273  13.693
##
## Correlation of Fixed Effects:
##       (Intr)
## cconc 0.368
```

The above frequentist model has some problems. In particular, it cannot estimate the correlations (notice that these are ± 1). I would therefore fit a simpler frequentist model, assuming no correlations between the varying intercepts and slopes:

```
m2<-lmer(log(value)~cconc+(1+cconc||Experiment)+
          (1+cconc||CellType),
        subset(means,ContrastAgent!="control"))

print(summary(m2))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## log(value) ~ cconc + ((1 | Experiment) + (0 + cconc | Experiment)) +
##     ((1 | CellType) + (0 + cconc | CellType))
##    Data: subset(means, ContrastAgent != "control")
##
## REML criterion at convergence: 67.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -3.1918 -0.7972  0.0205  0.6100  3.9187
##
## Random effects:
##  Groups       Name        Variance Std.Dev.
##  CellType     cconc       0.03443  0.1855
##  CellType.1   (Intercept) 0.33146  0.5757
##  Experiment   cconc       0.01184  0.1088
##  Experiment.1 (Intercept) 0.03170  0.1780
##  Residual                 0.06645  0.2578
## Number of obs: 162, groups:  CellType, 6; Experiment, 3
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   1.5599     0.2573   6.062
## cconc         1.7435     0.1115  15.632
##
## Correlation of Fixed Effects:
##       (Intr)
## cconc -0.004
```

The full model above (m1) can be fit in Stan quite easily. As priors for the fixed effects we choose Cauchy(0,5), to allow them to have extreme values, and for the others, we choose Normal(0,1), which seems reasonable as the dependent variable is on the log scale and the standard deviations are all going to be less than 1. The correlation parameters have as priors the LKJ(2) prior, which downweights the extreme values ± 1.

```
priors<-c(set_prior("cauchy(0,5)",
                    class = "b"),
                      set_prior("cauchy(0,5)", class = "b",coef="cconc"),set_prior("normal(0,1)", class
          set_prior("lkj(2)", class = "cor"))

m1brm<-brm(formula = log(value) ~ cconc+(1+cconc|Experiment)+
           (1+cconc|CellType),
             data = subset(means,ContrastAgent!="control"),
             family = gaussian(),
             prior = priors,
             warmup = 1000,
             iter = 2000,
             chains = 4,
             control = list(adapt_delta = 0.99,max_treedepth=15))
```

```
## Compiling the C++ model
```

```
## Start sampling
```

```
print(summary(m1brm))
```

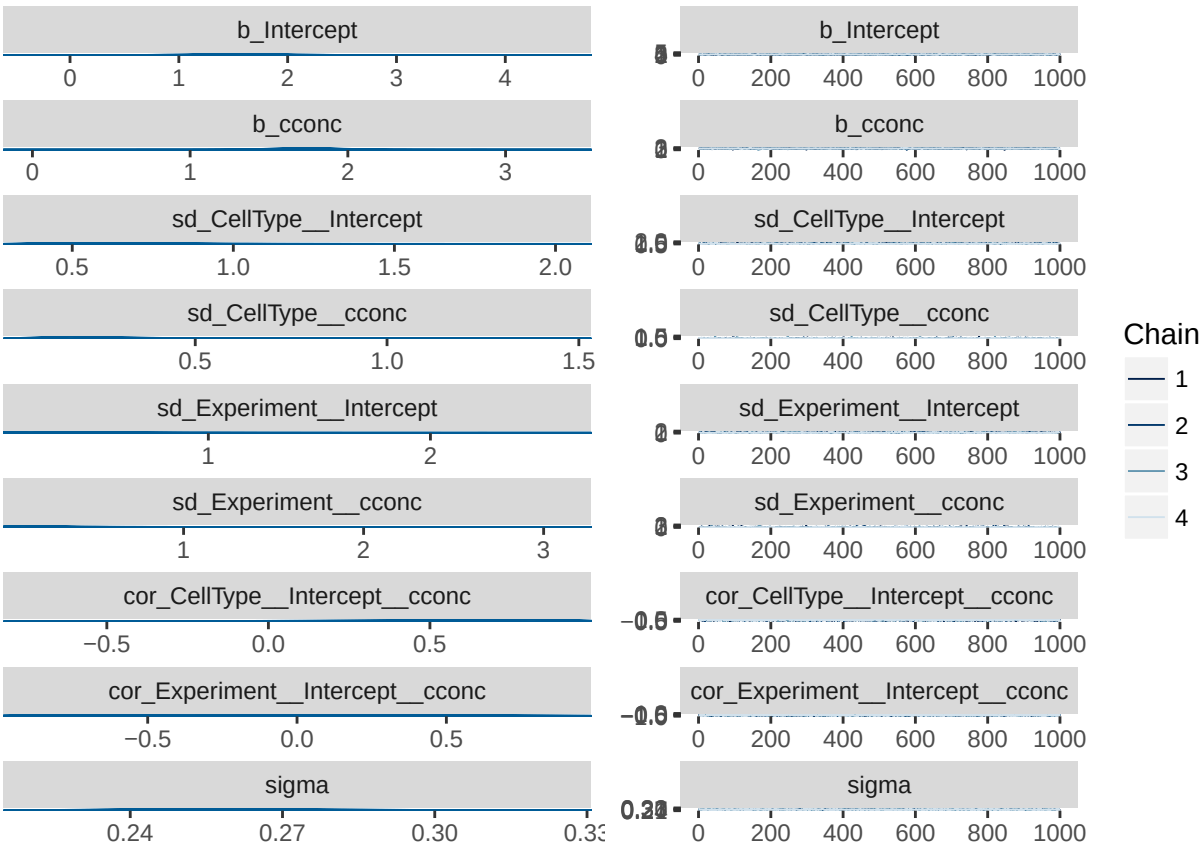
```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: log(value) ~ cconc + (1 + cconc | Experiment) + (1 + cconc | CellType)
##    Data: subset(means, ContrastAgent != "control") (Number of observations: 162)
## Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup samples = 4000
##     ICs: LOO = NA; WAIC = NA; R2 = NA
##
## Group-Level Effects:
## ~CellType (Number of levels: 6)
```

```
##                       Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
## sd(Intercept)             0.66      0.23     0.36     1.26       1634 1.00
## sd(cconc)                 0.23      0.13     0.06     0.54       1661 1.00
## cor(Intercept,cconc)      0.54      0.33    -0.28     0.96       2890 1.00
##
## ~Experiment (Number of levels: 3)
##                       Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
## sd(Intercept)             0.42      0.31     0.11     1.28       1292 1.00
## sd(cconc)                 0.29      0.29     0.02     1.07       1282 1.00
## cor(Intercept,cconc)     -0.15      0.45    -0.89     0.74       3285 1.00
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
## Intercept     1.57      0.41     0.74     2.39       1192 1.01
## cconc         1.73      0.27     1.20     2.23       1306 1.00
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
## sigma     0.26      0.02     0.23     0.29       4000 1.00
##
## Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
## is a crude measure of effective sample size, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

Notice some important differences between the frequentist model (FM) and the Bayesian model (BM): The FM fails to estimate the intercept-slope correlations; the BM is able to estimate the correlations, but also shows very high uncertainty in the posterior distributions of these correlations. The BM also provides 95% credible intervals for each parameter; the FM only does this for the fixed effects parameters (the intercept and the effect of concentration).

Plotting the posteriors:

```
plot(m1brm,N=9)
```



Modeling effect of contrast agent and its interaction with concentration

Because of the complexity of the model, I will not attempt to fit correlations in the frequentist model:

```
## hand-coded sliding contrasts:
means$MvsD<-ifelse(means$ContrastAgent=="Magnevist",1,
           ifelse(means$ContrastAgent=="Dotarem",-1,0))
means$DvsG<-ifelse(means$ContrastAgent=="Dotarem",1,
           ifelse(means$ContrastAgent=="Gadovist",-1,0))

m3<-lmer(log(value)~ cconc +MvsD+DvsG+cconc:MvsD + cconc:DvsG+(1+cconc +MvsD+DvsG+cconc:MvsD + cconc:Dvs
                 subset(means,ContrastAgent!="control"))
print(summary(m3))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: log(value) ~ cconc + MvsD + DvsG + cconc:MvsD + cconc:DvsG +
##     ((1 | Experiment) + (0 + cconc | Experiment) + (0 + MvsD |
##         Experiment) + (0 + DvsG | Experiment) + (0 + cconc:MvsD |
##         Experiment) + (0 + cconc:DvsG | Experiment)) + ((1 |
##     CellType) + (0 + cconc | CellType) + (0 + MvsD | CellType) +
##     (0 + DvsG | CellType) + (0 + cconc:MvsD | CellType) + (0 +
##     cconc:DvsG | CellType))
##    Data: subset(means, ContrastAgent != "control")
##
## REML criterion at convergence: 64.4
##
```

```
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.2662 -0.5969 -0.0337  0.5395  4.2093
##
## Random effects:
##  Groups       Name        Variance  Std.Dev.
##  CellType     cconc:DvsG  3.494e-18 1.869e-09
##  CellType.1   cconc:MvsD  0.000e+00 0.000e+00
##  CellType.2   DvsG        0.000e+00 0.000e+00
##  CellType.3   MvsD        0.000e+00 0.000e+00
##  CellType.4   cconc       3.636e-02 1.907e-01
##  CellType.5   (Intercept) 3.313e-01 5.756e-01
##  Experiment   cconc:DvsG  0.000e+00 0.000e+00
##  Experiment.1 cconc:MvsD  5.535e-16 2.353e-08
##  Experiment.2 DvsG        2.032e-03 4.508e-02
##  Experiment.3 MvsD        2.067e-04 1.438e-02
##  Experiment.4 cconc       1.291e-02 1.136e-01
##  Experiment.5 (Intercept) 3.197e-02 1.788e-01
##  Residual                 5.841e-02 2.417e-01
## Number of obs: 162, groups:  CellType, 6; Experiment, 3
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)  1.559926   0.257373   6.061
## cconc        1.743527   0.113075  15.419
## MvsD        -0.003335   0.028293  -0.118
## DvsG        -0.089993   0.037536  -2.398
## cconc:MvsD   0.059804   0.069592   0.859
## cconc:DvsG  -0.093508   0.069592  -1.344
##
## Correlation of Fixed Effects:
##            (Intr) cconc  MvsD   DvsG   ccn:MD
## cconc      -0.004
## MvsD        0.000  0.000
## DvsG        0.000  0.000  0.344
## cconc:MvsD  0.000  0.000 -0.115 -0.043
## cconc:DvsG  0.000  0.000 -0.057 -0.087  0.500
```

Stan version:

```
m3brm<-brm(formula = log(value)~ cconc +MvsD+DvsG+cconc:MvsD + cconc:DvsG+
             (1+cconc +MvsD+DvsG+cconc:MvsD + cconc:DvsG|Experiment)+
             (1+cconc +MvsD+DvsG+cconc:MvsD + cconc:DvsG|CellType),
           data = subset(means,ContrastAgent!="control"),
           family = gaussian(),
           prior = priors,
           warmup = 1000,
           iter = 2000,
           chains = 4,
           control = list(adapt_delta = 0.99,max_treedepth=15))
```

```
## Compiling the C++ model
```

```
## Start sampling
```

```r
print(summary(m3brm))
```

```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: log(value) ~ cconc + MvsD + DvsG + cconc:MvsD + cconc:DvsG + (1 + cconc + MvsD + DvsG + cco
##    Data: subset(means, ContrastAgent != "control") (Number of observations: 162)
## Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup samples = 4000
##     ICs: LOO = NA; WAIC = NA; R2 = NA
##
## Group-Level Effects:
## ~CellType (Number of levels: 6)
##                             Estimate Est.Error l-95% CI u-95% CI Eff.Sample
## sd(Intercept)                   0.69      0.25     0.37     1.29       1652
## sd(cconc)                       0.26      0.14     0.08     0.62       1266
## sd(MvsD)                        0.04      0.03     0.00     0.13       4000
## sd(DvsG)                        0.03      0.03     0.00     0.12       3338
## sd(cconc:MvsD)                  0.08      0.08     0.00     0.29       4000
## sd(cconc:DvsG)                  0.08      0.08     0.00     0.29       4000
## cor(Intercept,cconc)            0.32      0.29    -0.31     0.81       4000
## cor(Intercept,MvsD)             0.05      0.34    -0.59     0.68       4000
## cor(cconc,MvsD)                 0.05      0.32    -0.58     0.65       4000
## cor(Intercept,DvsG)             0.02      0.33    -0.61     0.65       4000
## cor(cconc,DvsG)                 0.01      0.34    -0.64     0.66       4000
## cor(MvsD,DvsG)                  0.03      0.33    -0.61     0.66       4000
## cor(Intercept,cconc:MvsD)       0.04      0.32    -0.59     0.65       4000
## cor(cconc,cconc:MvsD)           0.03      0.32    -0.59     0.64       4000
## cor(MvsD,cconc:MvsD)           -0.00      0.34    -0.65     0.63       4000
## cor(DvsG,cconc:MvsD)           -0.00      0.34    -0.63     0.65       4000
## cor(Intercept,cconc:DvsG)      -0.01      0.33    -0.61     0.61       4000
## cor(cconc,cconc:DvsG)           0.01      0.34    -0.64     0.64       4000
## cor(MvsD,cconc:DvsG)           -0.01      0.34    -0.65     0.63       4000
## cor(DvsG,cconc:DvsG)           -0.01      0.33    -0.63     0.62       4000
## cor(cconc:MvsD,cconc:DvsG)      0.02      0.34    -0.63     0.66       2700
##                             Rhat
## sd(Intercept)               1.00
## sd(cconc)                   1.00
## sd(MvsD)                    1.00
## sd(DvsG)                    1.00
## sd(cconc:MvsD)              1.00
## sd(cconc:DvsG)              1.00
## cor(Intercept,cconc)        1.00
## cor(Intercept,MvsD)         1.00
## cor(cconc,MvsD)             1.00
## cor(Intercept,DvsG)         1.00
## cor(cconc,DvsG)             1.00
## cor(MvsD,DvsG)              1.00
## cor(Intercept,cconc:MvsD)   1.00
## cor(cconc,cconc:MvsD)       1.00
## cor(MvsD,cconc:MvsD)        1.00
## cor(DvsG,cconc:MvsD)        1.00
## cor(Intercept,cconc:DvsG)   1.00
## cor(cconc,cconc:DvsG)       1.00
## cor(MvsD,cconc:DvsG)        1.00
```

```
## cor(DvsG,cconc:DvsG)       1.00
## cor(cconc:MvsD,cconc:DvsG) 1.00
##
## ~Experiment (Number of levels: 3)
##                            Estimate Est.Error l-95% CI u-95% CI Eff.Sample
## sd(Intercept)                  0.43      0.32     0.12     1.29       1374
## sd(cconc)                      0.31      0.29     0.02     1.10       1235
## sd(MvsD)                       0.15      0.20     0.00     0.75       1345
## sd(DvsG)                       0.18      0.22     0.01     0.85       1245
## sd(cconc:MvsD)                 0.30      0.32     0.01     1.23       1168
## sd(cconc:DvsG)                 0.29      0.31     0.01     1.15       1637
## cor(Intercept,cconc)          -0.09      0.32    -0.68     0.54       4000
## cor(Intercept,MvsD)            0.04      0.34    -0.61     0.67       4000
## cor(cconc,MvsD)               -0.01      0.33    -0.65     0.63       4000
## cor(Intercept,DvsG)            0.03      0.34    -0.61     0.68       4000
## cor(cconc,DvsG)               -0.04      0.33    -0.67     0.60       4000
## cor(MvsD,DvsG)                 0.01      0.34    -0.64     0.64       4000
## cor(Intercept,cconc:MvsD)     -0.06      0.33    -0.68     0.57       4000
## cor(cconc,cconc:MvsD)          0.04      0.33    -0.60     0.66       4000
## cor(MvsD,cconc:MvsD)          -0.03      0.35    -0.68     0.62       4000
## cor(DvsG,cconc:MvsD)          -0.02      0.33    -0.65     0.62       3354
## cor(Intercept,cconc:DvsG)     -0.07      0.34    -0.70     0.59       4000
## cor(cconc,cconc:DvsG)          0.04      0.34    -0.60     0.68       4000
## cor(MvsD,cconc:DvsG)          -0.02      0.33    -0.65     0.63       4000
## cor(DvsG,cconc:DvsG)          -0.03      0.33    -0.65     0.60       3348
## cor(cconc:MvsD,cconc:DvsG)     0.06      0.34    -0.59     0.70       3038
##                            Rhat
## sd(Intercept)              1.00
## sd(cconc)                  1.00
## sd(MvsD)                   1.00
## sd(DvsG)                   1.00
## sd(cconc:MvsD)             1.01
## sd(cconc:DvsG)             1.00
## cor(Intercept,cconc)       1.00
## cor(Intercept,MvsD)        1.00
## cor(cconc,MvsD)            1.00
## cor(Intercept,DvsG)        1.00
## cor(cconc,DvsG)            1.00
## cor(MvsD,DvsG)             1.00
## cor(Intercept,cconc:MvsD)  1.00
## cor(cconc,cconc:MvsD)      1.00
## cor(MvsD,cconc:MvsD)       1.00
## cor(DvsG,cconc:MvsD)       1.00
## cor(Intercept,cconc:DvsG)  1.00
## cor(cconc,cconc:DvsG)      1.00
## cor(MvsD,cconc:DvsG)       1.00
## cor(DvsG,cconc:DvsG)       1.00
## cor(cconc:MvsD,cconc:DvsG) 1.00
##
## Population-Level Effects:
##            Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
## Intercept      1.57      0.41     0.76     2.39        970 1.01
## cconc          1.75      0.29     1.13     2.30        871 1.00
## MvsD          -0.01      0.14    -0.31     0.27       1140 1.00
```

```
## DvsG          -0.09      0.15    -0.38     0.24        842 1.00
## cconc:MvsD     0.05      0.26    -0.50     0.57        747 1.01
## cconc:DvsG    -0.09      0.26    -0.62     0.41       1061 1.00
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
## sigma     0.25      0.02     0.22     0.28       4000 1.00
##
## Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
## is a crude measure of effective sample size, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

Measurement error on log(value)

to-do