ESR Reference Range Study

Code

Background and Objectives

One of our clinicians let us know that a lot of his patients, especially healthy elderly ones, receive ESR results flagged as abnormally high. He wonders whether our reference ranges are too stringent for this population.

The ESR is the measurement of the rate at which red blood cells (RBCs) settle in anticoagulated blood. RBCs settle faster when pro-sedimentation factors, such as fibrinogen, cause RBCs to stick together to form rouleaux. Pro-sedimentation factors tend to be upregulated in inflammatory states. Thus, the ESR is a nonspecific marker for inflammation.

It is known that a normal ESR value varies with age (higher in younger individuals) and gender (higher in females). However, there are multiple reference ranges in use by different clinical laboratories.

It is questionable whether inspection of a mixed population of ESR values (i.e., from healthy and sick individuals) allows classification of normal vs abnormal values. This may be the case if there is clear separation of normal and abnormal values. In any case, it may be useful to visualize a sample of ESR values and highlight different options for cutoffs to better understand whether changing our reference range might be beneficial.

MGH

Age	Male	Female
all	<=13 mm/h	<=20 mm/h

ARUP Laboratories

Age	Male	Female
all	<=10 mm/h	<=20 mm/h

Quest Diagnostics

Age	Male	Female
<=50	<=15 mm/h	<=20 mm/h
>50	<=20 mm/h	<=30 mm/h

LabCorp

Age	Male	Female
<=50	<=15 mm/h	<=32 mm/h
>50	<=30 mm/h	<=40 mm/h

Bakerman's ABCs of Interpretive Laboratory Data

Age	Male	Female
Newborn	<=2 mm/h	<= 2 mm/h
Neonates and children	3 - 13 mm/h	3 - 13 mm/h
<40	1 - 15 mm/h	1 - 20 mm/h
>=40	Age / 2	(Age / 2) + 5

Data Acquisition

Connect to the MGH Pathology Datamart (`phssql2057.partners.org`), and select the table `MGHLABUTIL_LabResults` .

Hide

```
username <- function() {
    .rs.api.showPrompt(title = "Username",
                       message = "Please enter your Partners User Name")
}
password <- function() {
    .rs.api.askForPassword(prompt = "Please enter your Partners Password")
}
con <- dbConnect(odbc(),
                 driver = "{FreeTDS}",
                 dsn = "PHSSQL2057",
                 uid = str_c("PARTNERS\\", username()),
                 pwd = password())

labresults_tbl <- tbl(con, "MGHLABUTIL_LabResults")
```

Retrieve all ESR values from 2017, along with the following columns:

1. **CollAge**. Patient age at time of collection.
2. **CollectDateTime**. Date and time of collection.
3. **PtName**. Patient name.
4. **PtNumber**. Patient MRN (MGH).
5. **PtSex**. Patient sex.
6. **Result**. ESR result value.
7. **TstOrderName**. Test order name.
8. **UserQAFlags**. This column marks “high” and “low” results.

Hide

```
esr <- labresults_tbl %>%
  select(CollAge,
         CollectDateTime,
         PtName,
         PtNumber,
         PtSex,
         Result,
         TstOrderName,
         UserQAFlags) %>%
  filter(TstOrderName == "ESR",
         year(CollectDateTime) == "2017") %>%
  arrange(CollectDateTime) %>%
  collect()
```

The ESR data was stored in the data frame `esr`. Save the unaltered data frame on disk in case the database server goes down.

Hide

```
esr %>%
    write_rds(str_c("esr_data_", today(), ".rds"))
```

Disconnect from the database and clean up.

Hide

```
dbDisconnect(con)
```

Data Exploration and Cleaning

Missing values

Count the number of `NA`s (missing values) in each column of `esr`.

Hide

```
esr %>%
    map_df(function(x) sum(is.na(x))) %>%
    gather(column, NAs)
```

column	NAs
<chr>	<int>

column <chr>	NAs <int>
CollAge	0
CollectDateTime	0
PtName	0
PtNumber	0
PtSex	0
Result	0
TstOrderName	0
UserQAFlags	0

8 rows

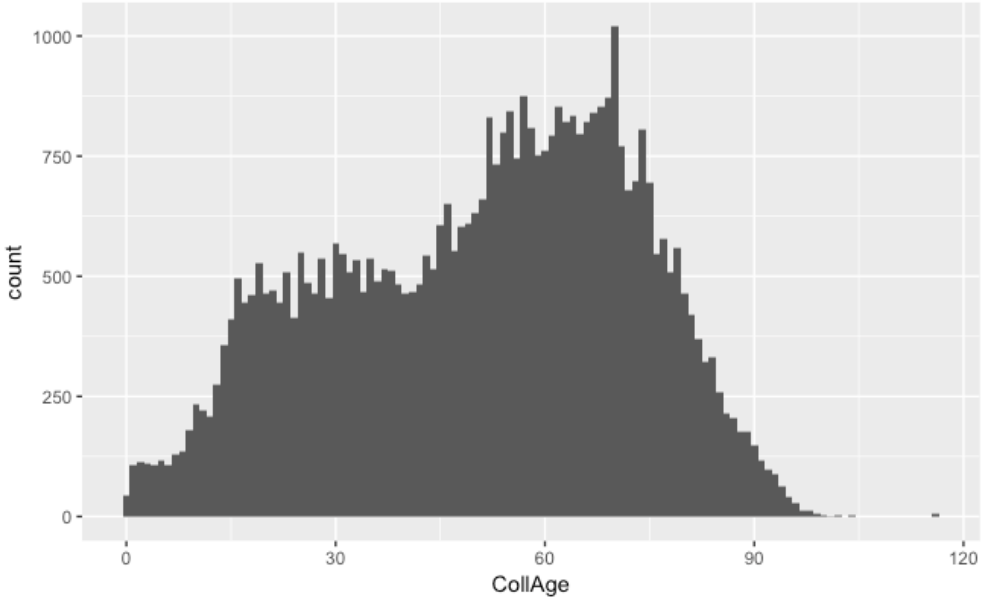
None of the columns of `esr` contain any missing values.

CollAge

`CollAge` , the age at collection, is an **integer** column, which is appropriate.

Hide

```
ggplot(data = esr) +
    geom_histogram(mapping = aes(x = CollAge),
                   binwidth = 1)
```



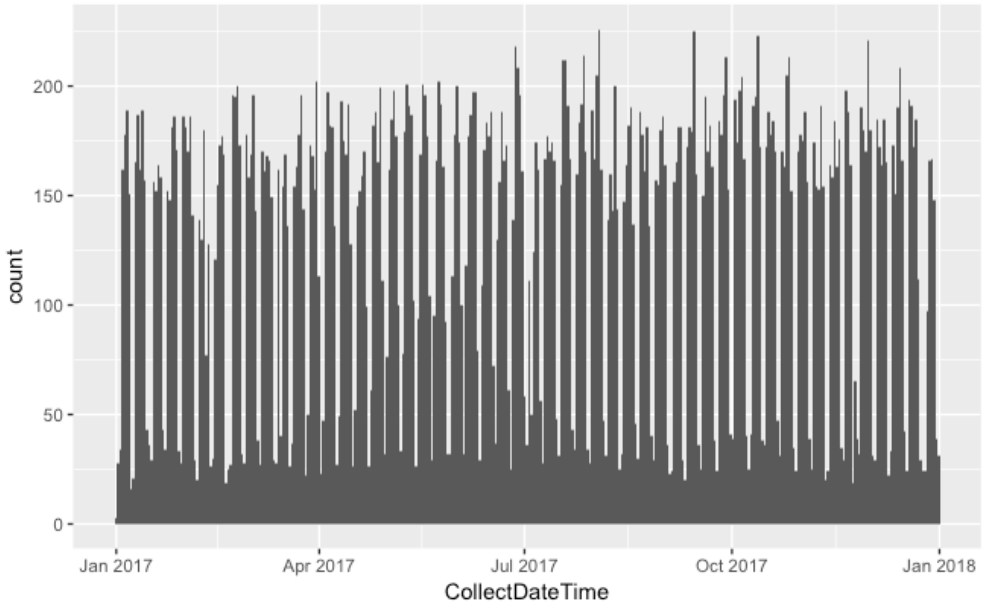
The distribution of `CollAge` is bimodal, with peaks at around 30 and at around 70 years.

CollectDateTime

`CollectDateTime` is a **POSIXct** (datetime) column, which is appropriate.

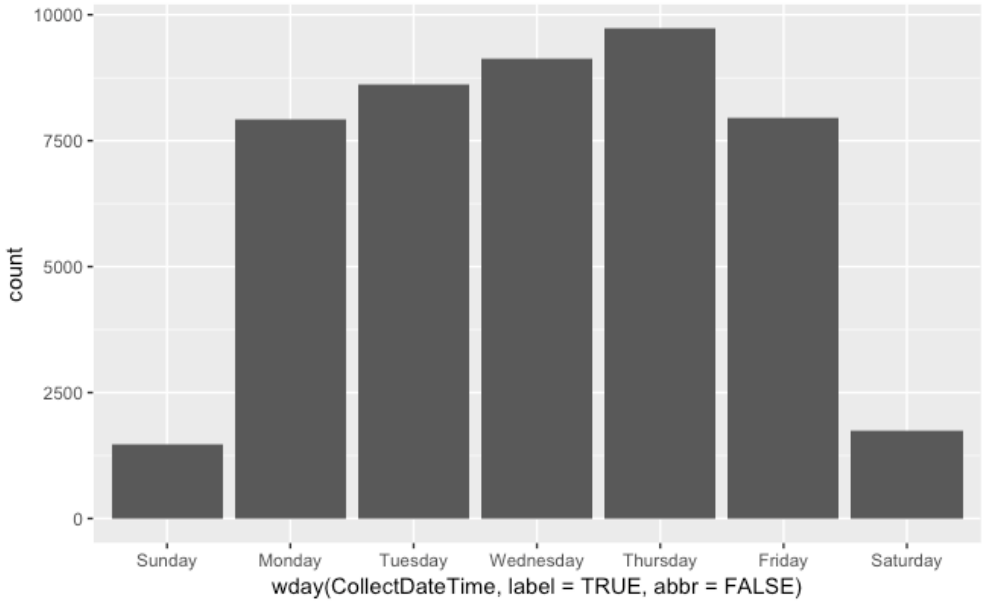
Hide

```
ggplot(data = esr) +
    geom_histogram(mapping = aes(x = CollectDateTime),
                   bins = 365)
```



Hide

```
ggplot(data = esr) +
    geom_bar(mapping = aes(x = wday(CollectDateTime, label = TRUE, abbr = FALSE)))
```



Weekly cycling of ESR test volume is apparent. The weekly ESR test volume was approximately constant throughout 2017, without clear peaks or troughs.

PtName

`PtName` is a **character** column, which is appropriate.

Hide

```
set.seed(1)
esr %>%
    sample_n(10) %>%
    pull(PtName)
```

```
 [1] "[redacted]"                "[redacted]"
 [3] "[redacted]"                "[redacted]"
 [5] "[redacted]"                "[redacted]"
 [7] "[redacted]"                "[redacted]"
 [9] "[redacted]"                "[redacted]"
```

A random sample of 10 names is as expected.

PtNumber

`PtNumber` , the MGH MRNs of the patients, is a **character** column, which is appropriate.

Note: if `PtNumber` were converted to integer, MRNs with leading zeroes would be altered.

Hide

```
set.seed(1)
esr %>%
    sample_n(10) %>%
    pull(PtNumber)
```

```
 [1] "[redacted]"
 [9] "[redacted]"
```

A random sample of 10 MRNs shows that all are seven digits long, as expected for MGH patients.

PtSex

`PtSex` is a **character** column, but since this is a categorial variable, we will convert it to a **factor**.

Hide

```
esr <- esr %>%
    mutate(PtSex = as_factor(PtSex))

esr %>%
  pull(PtSex) %>%
  summary()
```

```
    F     M     U
26701 19857    17
```

The majority of ESR tests were performed on female (`F`) patients. Unexpectedly, in addition to `M` and `F` , there are some rows with `PtSex` denoted as `U` .

Hide

```
esr %>%
    filter(PtSex == "U")
```

CollAge <int>	CollectDateTime <S3: POSIXct>	PtName <chr>	PtNumber <chr>	PtSex <fctr>	Result <chr>	TstOrderName <chr>	UserQAFlags <chr>
35	[redacted]		[redacted]	U	5	ESR	
35	[redacted]		[redacted]	U	7	ESR	
35	[redacted]		[redacted]	U	2	ESR	
39	2017-06-01 09:20:00	MGW CAP,ESR 01	QAPR-7475	U	69	ESR	H
33	2017-06-01 09:21:00	MGW CAP,ESR 02	QAPR-7476	U	9	ESR	
35	2017-06-01 09:21:00	MGW CAP,ESR 03	QAPR-7477	U	8	ESR	
5	2017-06-01 11:30:00	CHC,CAP ESR1	CHCQA-5786	U	60	ESR	H
5	2017-06-01 11:34:00	CHC,CAP ESR2	CHCQA-5787	U	8	ESR	
5	2017-06-01 11:35:00	CHC,CAP ESR3	CHCQA-5788	U	7	ESR	
35	[redacted]		[redacted]	U	2	ESR	

1-10 of 17 rows Previous **1** 2 Next

It appears that the rows with `PtSex` equal to `U` belong to a small number of patients as well as CAP survey samples. Since these rows represent less than 0.1% of the whole data set, it will be safe to remove them.

Hide

```
esr <- esr %>%
    filter(PtSex != "U")
```

Result

`Result` is a **character** column but should be an **integer**, so we will convert it.

Before doing so, examine any rows that do *not* contain an integer value:

Hide

```
esr %>%
    filter(is.na(as.integer(Result))) %>%
    group_by(Result) %>%
    summarize(n = n()) %>%
    arrange(desc(n))
```

```
NAs introduced by coercion
```

Result <chr>	**n** <int>
REFUS	1292
CREDIT	118
>140	14
QNS	12
CANCEL	6
	3
LABERR	2
>144	1
COMPR	1
CRDUP	1

1-10 of 19 rows Previous **1** 2 Next

About 1400 rows (out of about 47,000) have a non-integer value in `Result` - about **3%**. The large majority of them have a value of `REFUS` or `CREDIT`, indicating that the sample was not run or reported. These rows should be discarded.

Only a very small number of samples had excessive values (`>140`, `>144`). It is acceptable to discard these rows as well.

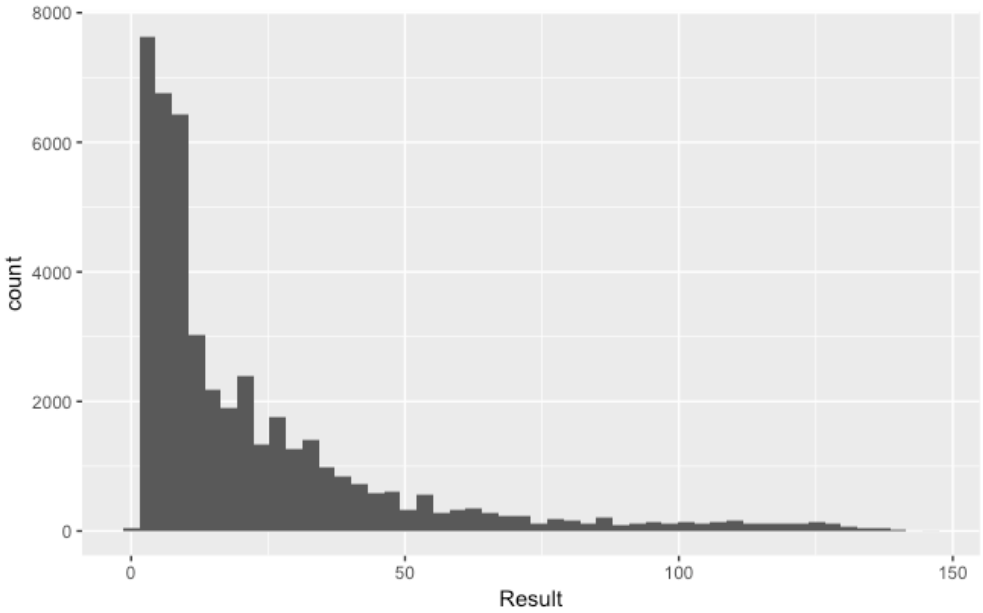
Hide

```
esr <- esr %>%
    mutate(Result = as.integer(Result)) %>%
    drop_na()
```

```
NAs introduced by coercion
```

Hide

```
ggplot(data = esr) +
    geom_histogram(mapping = aes(Result),
                   bins = 50)
```



The distribution of ESR values appears to have a mode at around 10, with a long right tail, which represents the abnormal values.

TstOrderName

`TstOrderName` should be `ESR` for each entry. It is a **character** column, which is acceptable.

Hide

```
esr %>%
    pull(TstOrderName) %>%
    as_factor() %>%
    summary()
```

```
  ESR
45099
```

Each entry is equal to `ESR`, as expected.

UserQAFlags

`UserQAFlags` should be `H` for a high value and `L` for a low value, and empty otherwise. It is a **character** column, but since it is a categorial variable, we will convert it to a **factor**.

Hide

```
esr <- esr %>%
    mutate(UserQAFlags = as_factor(UserQAFlags))

esr %>%
  pull(UserQAFlags) %>%
  summary()
```

```
         H
27297 17802
```

The majority of ESR values was reported as normal. A large fraction was classified as high (`H`). No ESR values were classified as low (`L`). This is expected, as the current MGH ESR reference ranges include zero, which is the lowest possible ESR result value.

Overview of the Cleaned Data

Hide

```
library(DT)
datatable(esr)
```

Show 10 **entries** **Search:**

	CollAge	CollectDateTime	PtName	PtNumber	PtSex	Result	TstOrderName	UserQAFlags
1	47	[redacted]	[redacted]	[redacted]	F	5	ESR	
2	56	[redacted]	[redacted]	[redacted]	M	39	ESR	H
3	52	[redacted]	[redacted]	[redacted]	M	34	ESR	H
4	68	[redacted]	[redacted]	[redacted]	F	71	ESR	H
5	38	[redacted]	[redacted]	[redacted]	F	28	ESR	H
6	34	[redacted]	[redacted]	[redacted]	F	5	ESR	
7	44	[redacted]	[redacted]	[redacted]	F	34	ESR	H
8	6	[redacted]	[redacted]	[redacted]	F	62	ESR	H
9	29	[redacted]	[redacted]	[redacted]	M	65	ESR	H
10	27	[redacted]	[redacted]	[redacted]	M	78	ESR	H

Showing 1 to 10 of 45,099 entries

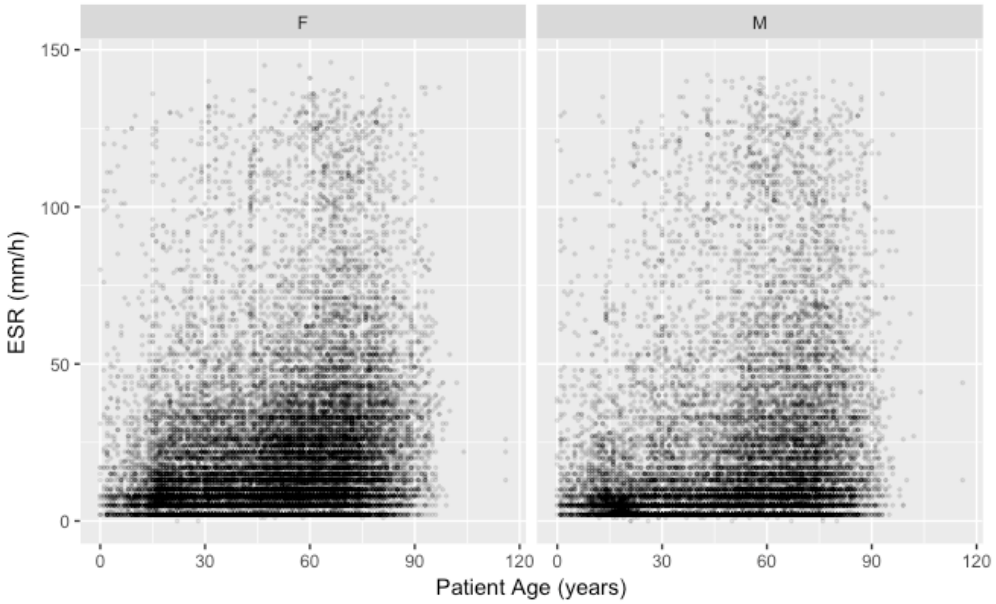
Previous 1 2 3 4 5 … 4510 Next

Visualization and Modeling

The primary goal of this project is to test is whether normal and abnormal ESR values can be separated into distinct populations, so that reference intervals can be set at the boundary. From the literature, I expect that ESR values increase with age and sex. To reduce these confounding effects, we will look at the relationship of ESR and age, broken down by sex.

Hide

```
ggplot(data = esr) +
    geom_point(mapping = aes(x = CollAge, y = Result),
               size = 0.5,
               alpha = 0.1) +
    facet_wrap(~PtSex) +
    labs(x = "Patient Age (years)",
         y = "ESR (mm/h)")
```

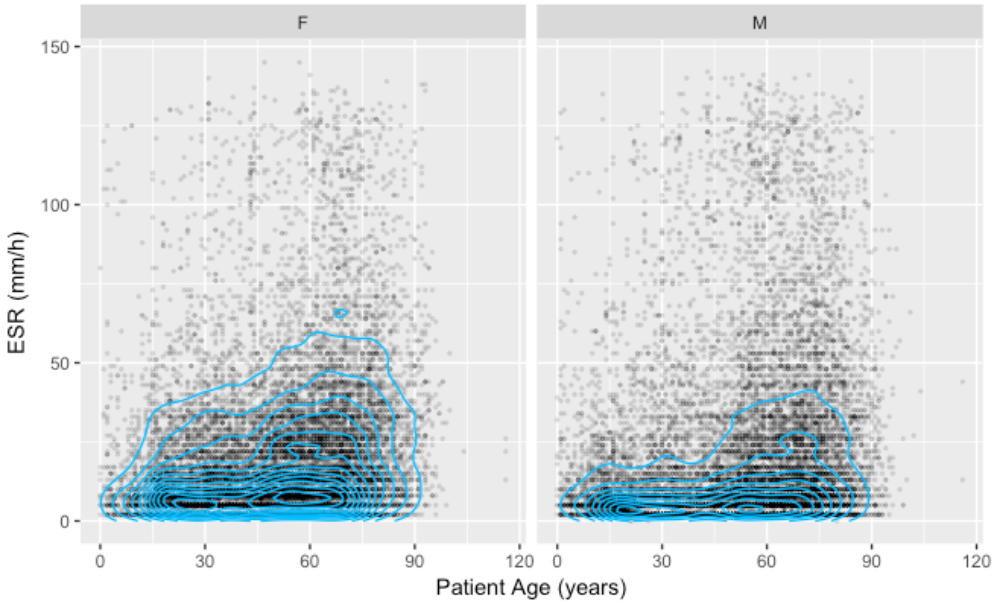


There is no clear separation of normal and abnormal values. In addition, there are more data points on the `female` panel, which makes this graph difficult to interpret.

Will a separation become apparent when equal numbers of points are shown on each graph, and a density contour plot is overlaid?

Hide

```
ggplot(data = esr_sample,
       mapping = aes(x = CollAge,
                     y = Result)) +
    facet_wrap(~PtSex) +
    geom_point(size = 0.5,
               alpha = 0.1) +
    geom_density2d(color = "deepskyblue") +
    labs(x = "Patient Age (years)",
         y = "ESR (mm/h)")
```

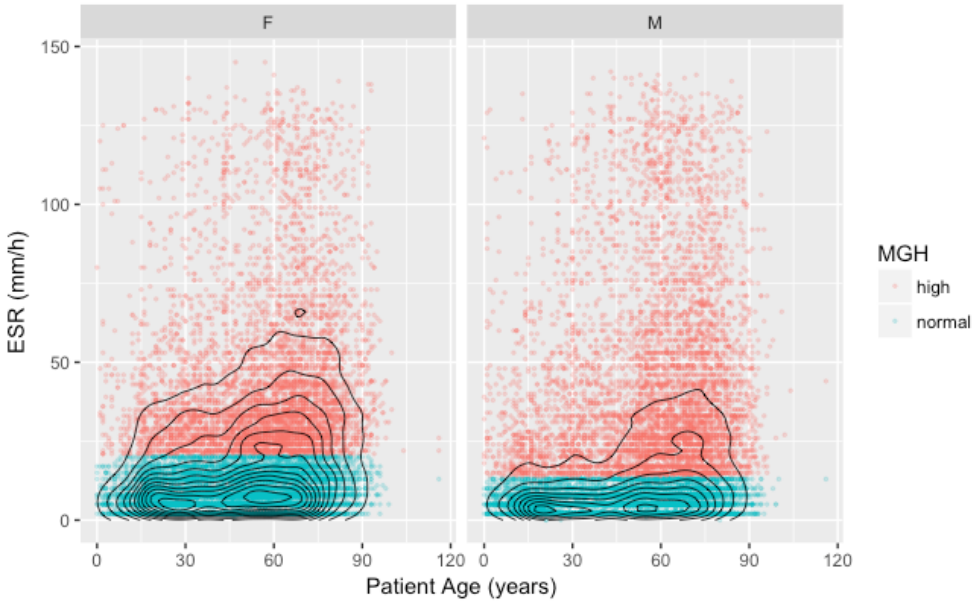


Subsampling and density contour plotting did not reveal separable populations. This graph clearly shows that female patients tended to have higher ESR levels than male patients, and that older patients tended to have higher ESR levels than younger patients. However, this effect was not universal, and many old patients of both genders had low ESR values.

Do the MGH reference ranges appear to appropriately capture patients with normal ESRs? Abnormal ESRs?

Hide

```
ggplot(data = esr_sample,
       mapping = aes(x = CollAge,
                     y = Result,
                     color = MGH)) +
    geom_point(size = 0.5,
               alpha = 0.2) +
    geom_density2d(color = "black",
                   size = 0.3) +
    facet_wrap(~PtSex) +
    labs(x = "Patient Age (years)",
         y = "ESR (mm/h)")
```

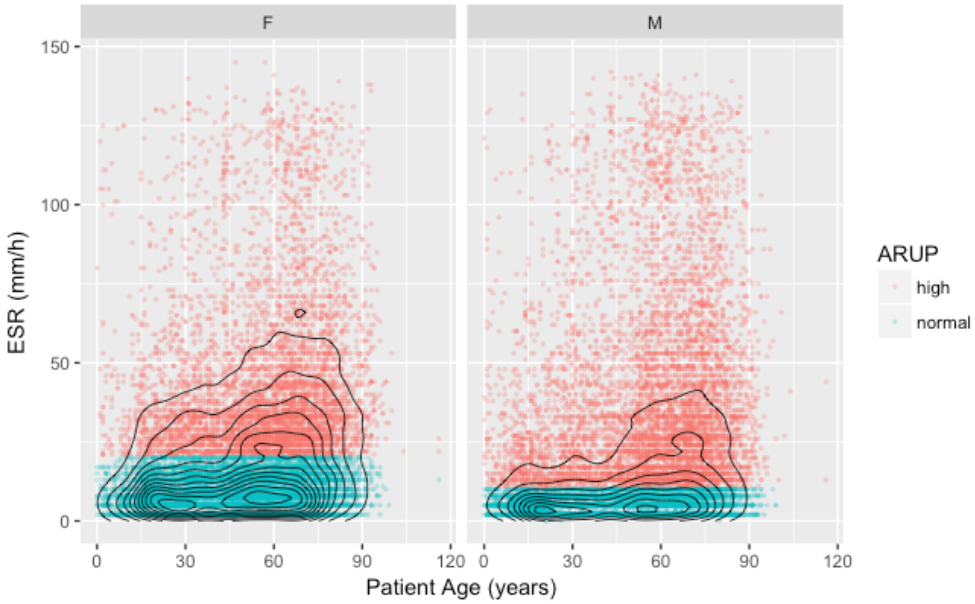


MGH's normal reference range appears to capture the majority of normal ESR results. It does not capture the tendency of older individuals to have higher ESR values. However, this tendency is not universal, and it is uncertain whether high ESR values in older individuals are physiologically normal.

ARUP Laboratories

Hide

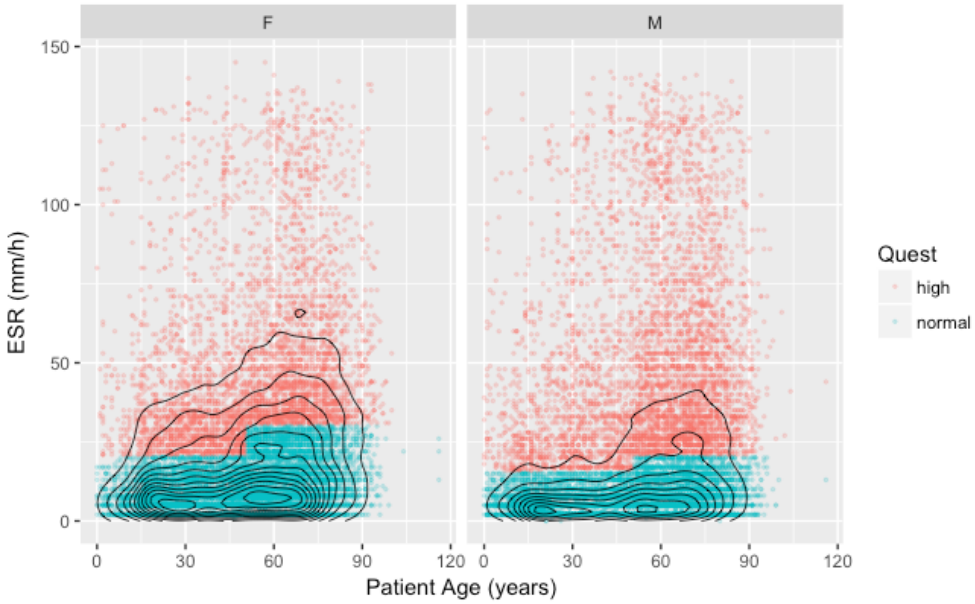
```
esr_sample <- esr_sample %>%
    mutate(
        ARUP = case_when(
            PtSex == "M" & Result >10 ~ "high",
            PtSex == "F" & Result >20 ~ "high",
            TRUE ~ "normal") %>%
                factor(levels = c("high", "normal"))
    )

ggplot(data = esr_sample,
       mapping = aes(x = CollAge,
                     y = Result,
                     color = ARUP)) +
    geom_point(size = 0.5,
               alpha = 0.2) +
    geom_density2d(color = "black",
                   size = 0.3) +
    facet_wrap(~PtSex) +
    labs(x = "Patient Age (years)",
         y = "ESR (mm/h)")
```



Quest Diagnostics

Hide

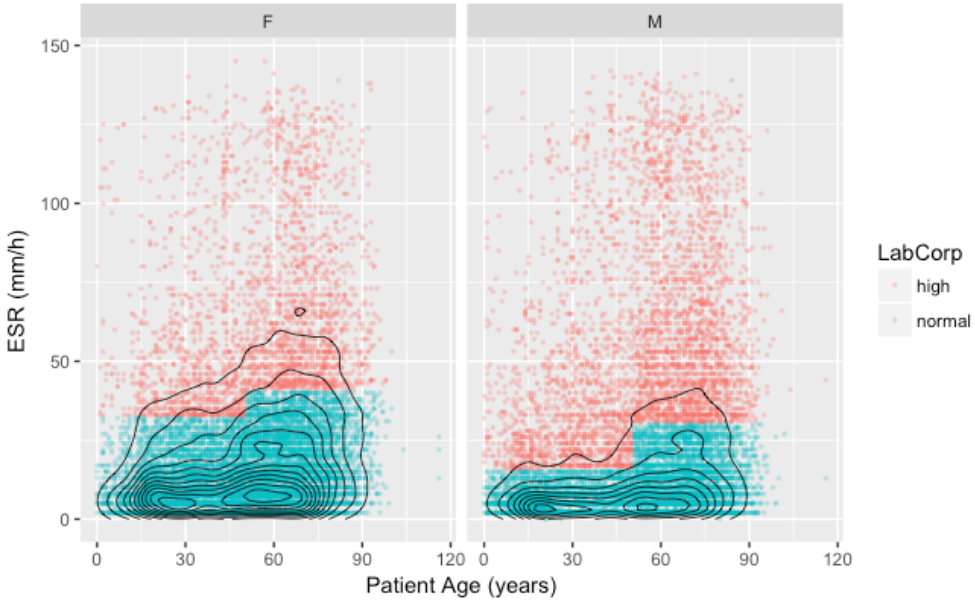
```
esr_sample <- esr_sample %>%
    mutate(
        Quest = case_when(
            (PtSex == "M" & CollAge <=50 & Result >15) |
            (PtSex == "M" & CollAge >50 & Result >20) |
            (PtSex == "F" & CollAge <=50 & Result >20) |
            (PtSex == "F" & CollAge >50 & Result >30) ~ "high",
            TRUE ~ "normal") %>%
                factor(levels = c("high", "normal"))
    )

ggplot(data = esr_sample,
       mapping = aes(x = CollAge,
                     y = Result,
                     color = Quest)) +
    geom_point(size = 0.5,
               alpha = 0.2) +
    geom_density2d(color = "black",
                   size = 0.3) +
    facet_wrap(~PtSex) +
    labs(x = "Patient Age (years)",
         y = "ESR (mm/h)")
```



LabCorp

Hide

```
esr_sample <- esr_sample %>%
    mutate(
        LabCorp = case_when(
            (PtSex == "M" & CollAge <=50 & Result >15) |
            (PtSex == "M" & CollAge >50 & Result >30) |
            (PtSex == "F" & CollAge <=50 & Result >32) |
            (PtSex == "F" & CollAge >50 & Result >40) ~ "high",
            TRUE ~ "normal") %>%
                factor(levels = c("high", "normal"))
    )
ggplot(data = esr_sample,
       mapping = aes(x = CollAge,
                     y = Result,
                     color = LabCorp)) +
    geom_point(size = 0.5,
               alpha = 0.2) +
    geom_density2d(color = "black",
                   size = 0.3) +
    facet_wrap(~PtSex) +
    labs(x = "Patient Age (years)",
         y = "ESR (mm/h)")
```



Bakerman's ABCs

Hide

```
esr_sample <- esr_sample %>%
    mutate(
        Bakerman = case_when(
            (CollAge <18 & Result >13) ~ "high",
            (PtSex == "M" & CollAge >=18 & CollAge <40 & Result >15) |
            (PtSex == "F" & CollAge >=18 & CollAge <40 & Result >20) |
            (PtSex == "M" & CollAge >=40 & Result > (CollAge/2)) |
            (PtSex == "F" & CollAge >=40 & Result > (CollAge/2 + 5)) ~ "high",
            TRUE ~ "normal") %>%
                factor(levels = c("high", "normal"))
    )
ggplot(data = esr_sample,
       mapping = aes(x = CollAge,
                     y = Result,
                     color = Bakerman)) +
    geom_point(size = 0.5,
               alpha = 0.2) +
    geom_density2d(color = "black",
                   size = 0.3) +
    facet_wrap(~PtSex) +
    labs(x = "Patient Age (years)",
         y = "ESR (mm/h)")
```