Introduction au logiciel R

Capsule: Pivotage

Pivoter jeu de données: en format long et en format large/court

Auteur: Benoît Mâsse

Format des données : long et large

- Forme typique d'un jeu de données: forme matricielle ligne par colonne
- **Format large/court**: c'est le format qu'on retrouve le plus souvent
  - Une ligne par patient/participant
  - Chaque colonne représente une variable
  - La cellule donne la valeur d'une variable pour un patient/participant
- Exemple: Gynuity_Argentina_oxytocin_IVIM_PPH_Prevention_DB

Gynuity_Argentina_oxytocin_IVIM_...

Filter | Cols: « ‹ 1 – 50 › »

	patientID patient unique study ID	inicialesm Participant's initials	inicialesp Staff initials	fecha Date of enrollment (prior to delivery) dd-mmm-yyyy	dar_a_luz Arrived at hospital to give birth?	parto Cesarean delivery planned?	enfermedad Any illness that would prevent participation in study?	enfermedad_especifique Specify illness	elegible woman is eligible to partic
1	3001	AYS	FO	2016-12-01	1	0	0		1
2	3002	SNM	MAG	2016-12-03	1	0	0		1
3	3003	VNR	FO	2016-12-03	1	0	0		1
4	3004	SPD	MLJ	2016-12-05	1	0	0		1
5	3005	CEN	MLR	2016-12-07	1	0	0		1
6	3006	LGM	MLJ	2016-12-07	1	0	0		1
7	3007	MES	JEB	2016-12-08	1	0	0		1
8	3008	MAB	JEB	2016-12-09	1	0	0		1
9	3009	KMO	LB	2016-12-09	1	0	0		1

Showing 1 to 9 of 543 entries, 148 total columns

Format des données : long et large

- **Format long**: ... l'inverse du format large
  - Plusieurs lignes par patient/participant
  - La cellule peut donner une information sur une variable ou une valeur
- Exemple:

Format large

id	y1	y2	y3	y4
1	3.5	4.5	7.5	7.5
2	6.5	5.5	8.5	8.5

2 lignes x 5 colonnes

Format long

id	time	y
1	1	3.5
1	2	4.5
1	3	7.5
1	4	7.5
2	1	6.5
2	2	5.5
2	3	8.5
2	4	8.5

8 lignes x 3 colonnes

Création des jeux de données

- Dans un contexte réglementaire, il est plutôt rare que les jeux de données pour une étude clinique sont créés directement. Plutôt on utilise des logiciels pour la capture et la saisie des données (ex: REDCap).

- Pour l'analyse des données, on doit exporter les données et le format usuel d'exportation est le format large.
  - Par contre, certaines données auxiliaires (ex: résultats de tests de laboratoire) peuvent être en format long.

- **Pivotage**: heureusement, il y a un package en R qui permet de passer d'un format à l'autre
  - Pivoter du format large au format long
  - Pivoter du format long au format large

Quel format choisir?

- Bien souvent le format **long** permet une lecture des données plus compréhensibles.
- Le format **long** est requis pour certaines analyses statistiques:
  - Analyse avec mesures répétées, analyse avec le modèle de Cox avec expositions qui varient dans le temps, analyse avec des unités regroupées en grappe, etc.
- Le format **long** est plus adapté à la création de certains types de tableaux, graphiques et de figures.
- Il est n'est pas nécessaire (ni souhaitable) de pivoter un jeu de données complet dans un format ou l'autre. Typiquement, pour une analyse donnée, on pivotera seulement les variables/données qui sont nécessaires pour l'analyse ou la création d'une figure.

Standardisation des formats de données

- Dans le contexte de la recherche clinique, il existe plusieurs initiatives pour standardiser les formats des données. Par exemple voir les standards du *'Clinical Data Interchange Standards Consortium*' , https://www.cdisc.org/standards

- Un des avantages de cette standardisation est de faciliter le partage et l'exploitation des données entre la communauté scientifique.

- Certains organismes réglementaires, comme la FDA, exigent que les données supportant une demande d'homologation d'un nouveau traitement soient soumises avec les standards CDISC.
  - Le format **long** est privilégié par les standards CDISC

Exemple: standard CDISC

Example 2

Data are collected about the occurrence of specific asthma-related conditions. If the event occurred, a start date is collected. The data was collected in the following CDASH-compliant form:

Asthma-Related History

Has the subject had sinusitis?	Yes	No	If yes, start date: _____
Has the subject had atopic dermatitis?	Yes	No	If yes, start date: _____
Has the subject had eczema?	Yes	No	If yes, start date: _____
Has the subject had seasonal allergic rhinitis?	Yes	No	If yes, start date: _____
Has the subject had allergic conjunctivitis?	Yes	No	If yes, start date: _____

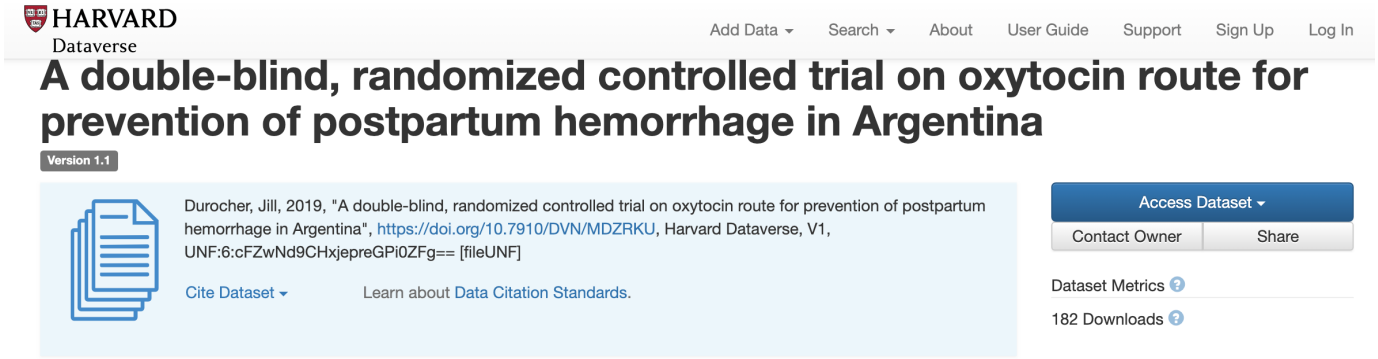
Row 1: The subject had a history of sinusitis.
Rows 2-3: The subject did not have a history of atopic dermatitis or eczema.
Row 4: The subject had a history of seasonal allergic rhinitis.
Row 5: The subject did not have a history of allergic conjunctivitis.

mh.xpt

Row	STUDYID	DOMAIN	USUBJID	MHSEQ	MHTERM	MHDECOD	MHCAT	MHPRESP	MHOCCUR	MHDTC	MHSTDTC	MHEVINTX
1	ABC123	MH	456	1	Sinusitis	Sinusitis	Asthma Related	Y	Y	2013-05-24	2009	Time before query
2	ABC123	MH	456	1	Atopic dermatitis	Dermatitis, atopic	Asthma Related	Y	N	2013-05-24		Time before query
3	ABC123	MH	456	1	Eczema	Eczema	Asthma Related	Y	N	2013-05-24		Time before query
4	ABC123	MH	456	1	Seasonal allergic rhinitis	Rhinitis, seasonal	Asthma Related	Y	Y	2013-05-24	1999	Time before query
5	ABC123	MH	456	1	Allergic conjunctivitis	Conjunctivitis, allergic	Asthma Related	Y	N	2013-05-24		Time before query

Pivotage avec R

- On utilise le jeu de données pour l'étude l'ocytocine en Argentin3
  - https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/MDZRKU



- On télécharge les fichiers de données et la description des variables

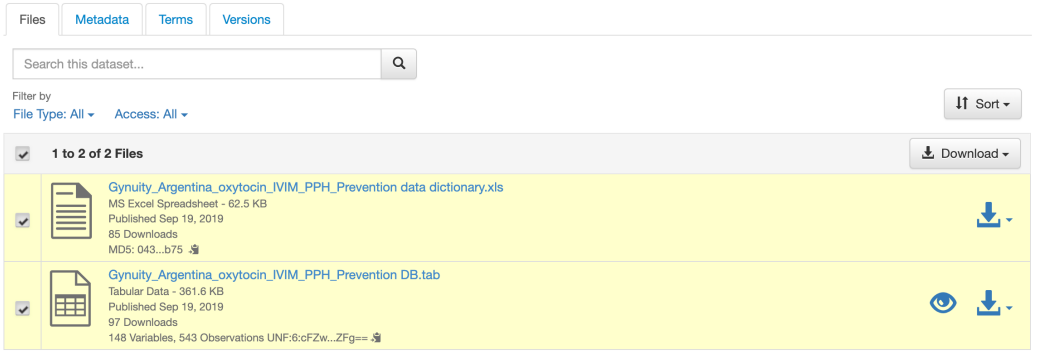
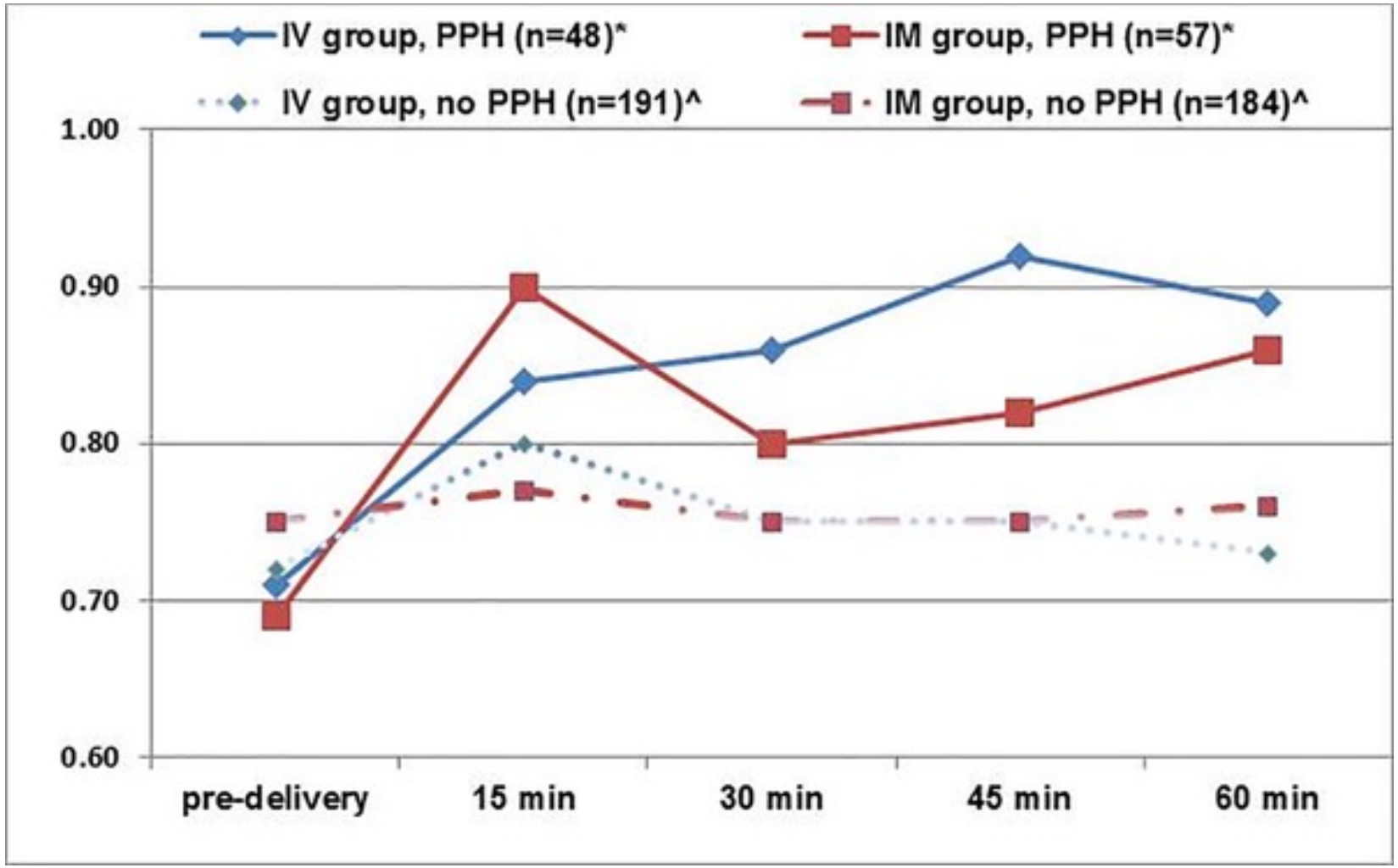


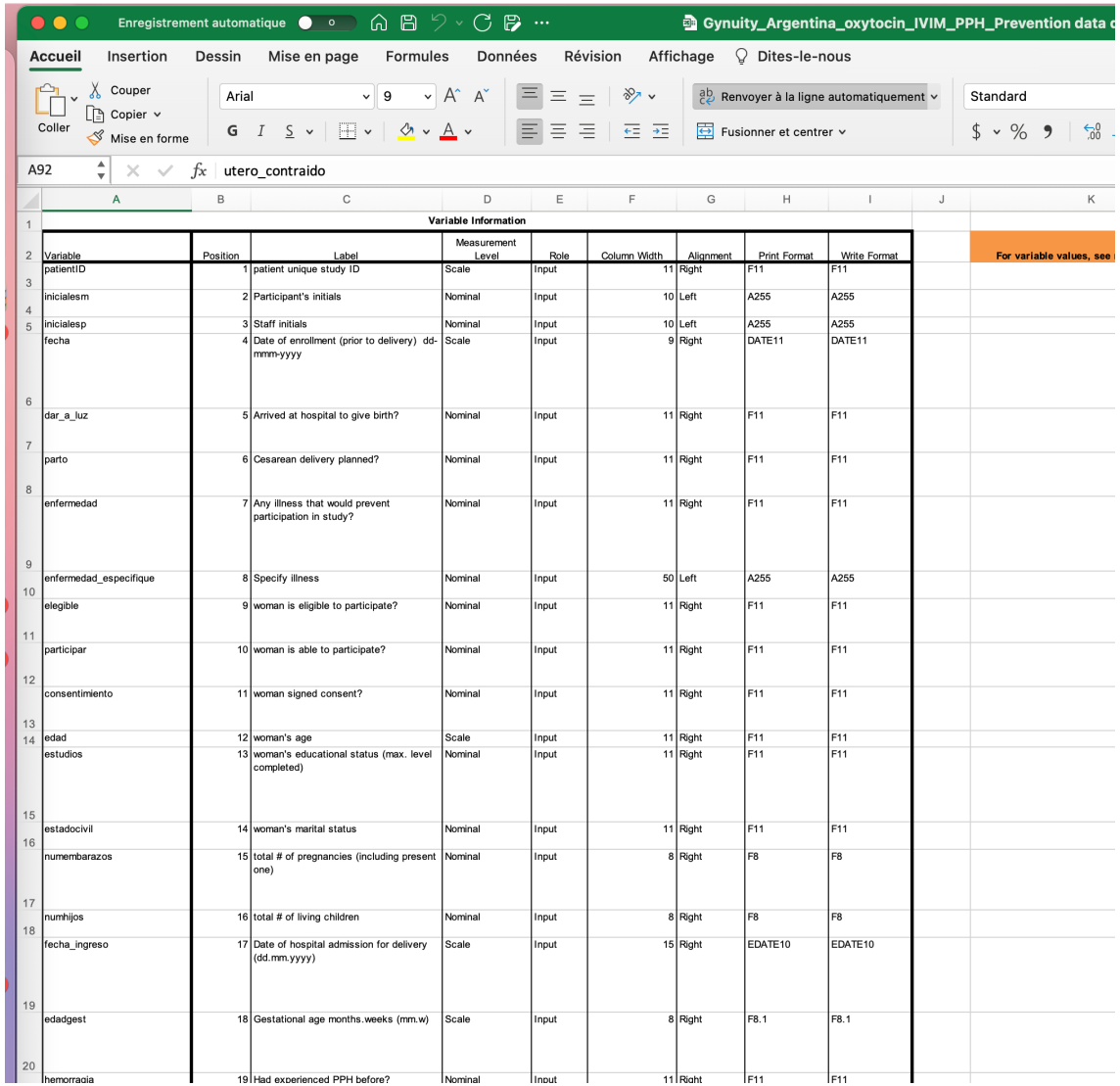
Fig 2. Median shock index values pre-delivery and during the first hour postpartum for PPH cases and non-PPH cases by study group.



Durocher J, Dzuba IG, Carroli G, Morales EM, Aguirre JD, et al. (2019) Does route matter? Impact of route of oxytocin administration on postpartum bleeding: A double-blind, randomized controlled trial. PLOS ONE 14(10): e0222981. https://doi.org/10.1371/journal.pone.0222981
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0222981

Liste des variables

- Jeu de données en format large
- Mesures répétées sur certaines variables (tension artérielle, *shock index*) : 0, 15, 30, 45, et 60 minutes.



Enregistrement automatique — Gynuity_Argentina_oxytocin_IVIM_PPH_Prevention data c

Accueil | Insertion | Dessin | Mise en page | Formules | Données | Révision | Affichage | Dites-le-nous

Couper | Copier | Mise en forme | Coller | Arial | 9 | Renvoyer à la ligne automatiquement | Standard | Fusionner et centrer

A92 | utero_contraido

Variable Information

Variable	Position	Label	Measurement Level	Role	Column Width	Alignment	Print Format	Write Format
patientID	1	patient unique study ID	Scale	Input	11	Right	F11	F11
inicialesm	2	Participant's initials	Nominal	Input	10	Left	A255	A255
inicialesp	3	Staff initials	Nominal	Input	10	Left	A255	A255
fecha	4	Date of enrollment (prior to delivery) dd-mmm-yyyy	Scale	Input	9	Right	DATE11	DATE11
dar_a_luz	5	Arrived at hospital to give birth?	Nominal	Input	11	Right	F11	F11
parto	6	Cesarean delivery planned?	Nominal	Input	11	Right	F11	F11
enfermedad	7	Any illness that would prevent participation in study?	Nominal	Input	11	Right	F11	F11
enfermedad_especifique	8	Specify illness	Nominal	Input	50	Left	A255	A255
elegible	9	woman is eligible to participate?	Nominal	Input	11	Right	F11	F11
participar	10	woman is able to participate?	Nominal	Input	11	Right	F11	F11
consentimiento	11	woman signed consent?	Nominal	Input	11	Right	F11	F11
edad	12	woman's age	Scale	Input	11	Right	F11	F11
estudios	13	woman's educational status (max. level completed)	Nominal	Input	11	Right	F11	F11
estadocivil	14	woman's marital status	Nominal	Input	11	Right	F11	F11
numembarazos	15	total # of pregnancies (including present one)	Nominal	Input	8	Right	F8	F8
numhijos	16	total # of living children	Nominal	Input	8	Right	F8	F8
fecha_ingreso	17	Date of hospital admission for delivery (dd.mm.yyyy)	Scale	Input	15	Right	EDATE10	EDATE10
edadgest	18	Gestational age months.weeks (mm.w)	Scale	Input	8	Right	F8.1	F8.1
hemorragia	19	Had experienced PPH before?	Nominal	Input	11	Right	F11	F11

For variable values, see

Téléchargement du jeu de données

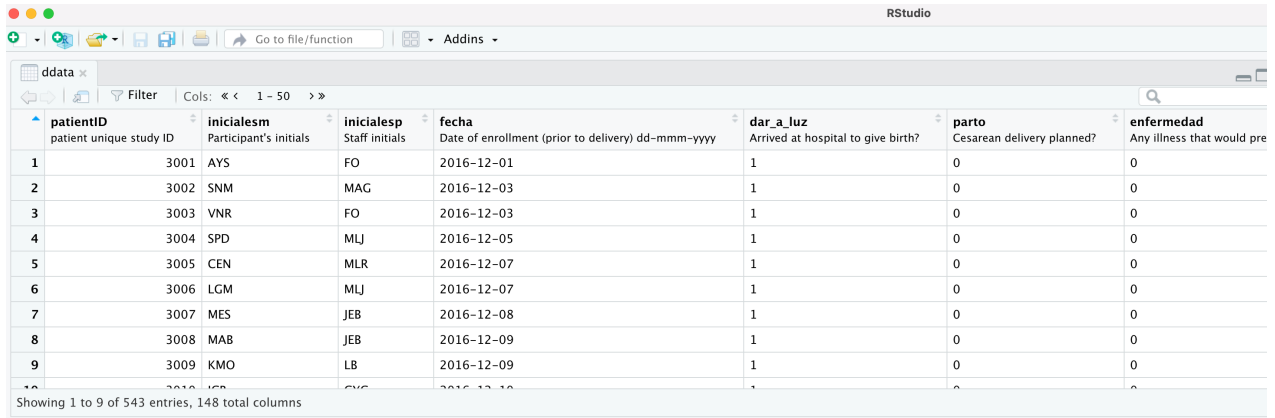
- Monte le package haven qui permet l'importation de fichier SPSS

```
> library(haven)
```

- Téléchargement du jeu de données

```
> ddata <- read_sav("Dropbox/Capsule R/pivotage/dataverse_files/Gynuity_Argentina_oxytocin_IVIM_PPH_Prevention DB.sav")
```

```
> View(ddata)
> attach(ddata)
```



Pivotage en format long de mesures répétées

- Monte les packages tidyr, dplyr, et readr

```
> library(tidyr)
> library(dplyr)
> library(readr)
```

- Sélection des données et variables à pivoter

```
> ddata2 <- ddata[,c(1,125:129)]
> ddata2
```

```
# A tibble: 543 × 6
   patientID SI_baseline SI_15min SI_30min SI_45min SI_60min
       <dbl>       <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
 1      3001       0.621    0.901    0.857    0.76     0.827
 2      3002       0.752    0.673    0.688    0.661    0.708
 3      3003       0.452    0.654    0.657    0.605    0.606
 4      3004       0.640   NA       NA       NA       NA
 5      3005       0.65     0.690    0.678    0.644    0.738
 6      3006       1.01     1.10     0.892    0.842    0.955
 7      3007       0.657    0.724    0.667    0.618    0.603
 8      3008       0.591    0.611    0.576    0.558    0.609
 9      3009       0.730    0.916    1.02     1.09     1.16
10      3010       0.857    0.743    0.716    0.685    0.808
# … with 533 more rows
```

Pivotage en format long de mesures répétées

- Pivotage en format long

```
> ddata3 <- pivot_longer(ddata2, !patientID, names_to="SI time", values_to="Value")
> ddata3

# A tibble: 2,715 × 3
   patientID `SI time`   Value
       <dbl> <chr>       <dbl>
 1     3001 SI_baseline 0.621
 2     3001 SI_15min    0.901
 3     3001 SI_30min    0.857
 4     3001 SI_45min    0.76
 5     3001 SI_60min    0.827
 6     3002 SI_baseline 0.752
 7     3002 SI_15min    0.673
 8     3002 SI_30min    0.688
 9     3002 SI_45min    0.661
10     3002 SI_60min    0.708
# … with 2,705 more rows
```

Pivotage en format long de mesures répétées

- Pivotage en format long : utilisation de l'information contenue dans les noms de variable

```
> ddata3 <- pivot_longer(ddata2, !patientID, names_to=c(".value","Time"),names_sep=("_"))
> ddata3
```

```
# A tibble: 2,715 × 3
   patientID Time        SI
       <dbl> <chr>    <dbl>
 1      3001 baseline 0.621
 2      3001 15min    0.901
 3      3001 30min    0.857
 4      3001 45min    0.76
 5      3001 60min    0.827
 6      3002 baseline 0.752
 7      3002 15min    0.673
 8      3002 30min    0.688
 9      3002 45min    0.661
10      3002 60min    0.708
# … with 2,705 more rows
```

Pivotage en format large

- Pivotage en format large

```
> ddata4 <- pivot_wider(ddata3,names_from=Time,values_from=SI)
> ddata4
```

```
# A tibble: 543 × 6
   patientID baseline `15min` `30min` `45min` `60min`
       <dbl>    <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
 1      3001    0.621   0.901   0.857   0.76    0.827
 2      3002    0.752   0.673   0.688   0.661   0.708
 3      3003    0.452   0.654   0.657   0.605   0.606
 4      3004    0.640   NA      NA      NA      NA
 5      3005    0.65    0.690   0.678   0.644   0.738
 6      3006    1.01    1.10    0.892   0.842   0.955
 7      3007    0.657   0.724   0.667   0.618   0.603
 8      3008    0.591   0.611   0.576   0.558   0.609
 9      3009    0.730   0.916   1.02    1.09    1.16
10      3010    0.857   0.743   0.716   0.685   0.808
# … with 533 more rows
```

Pivotage un peu plus complexe : tension artérielle

- Sélection et extraction des données

```
> ddata5 <- ddata[,c(1,68:69,72:73,77:78,81:82)]
> ddata5
```

```
# A tibble: 543 × 9
   patientID presion_sist_15 presion_diast_15 presion_sist_30 presion_diast_30 presion_sist_45
       <dbl>           <dbl>            <dbl>           <dbl>           <dbl>           <dbl>
 1      3001             101               60              98              60             100
 2      3002             107               76             109              76             109
 3      3003             133               83             134              83             124
 4      3004              NA               NA              NA              NA              NA
 5      3005             126               77             118              72             118
 6      3006              96               52             111              59             114
 7      3007             123               69             108              70             110
 8      3008             113               63             118              68             120
 9      3009             119               94             101              66             106
10      3010             109               66             109              72             108
# … with 533 more rows, and 3 more variables: presion_diast_45 <dbl>, presion_sist_60 <dbl>,
#   presion_diast_60 <dbl>
```

Pivotage un peu plus complexe : tension artérielle

- Pivotage en format long

```
> ddata6 <- pivot_longer(ddata5, !patientID, names_to=c(".value","Tension","Time"),
names_pattern="(.*)_(.*)_(.*)")

> ddata6
```

```
# A tibble: 4,344 × 4
   patientID Tension Time  presion
       <dbl> <chr>   <chr>   <dbl>
 1      3001 sist    15        101
 2      3001 diast   15         60
 3      3001 sist    30         98
 4      3001 diast   30         60
 5      3001 sist    45        100
 6      3001 diast   45         58
 7      3001 sist    60        104
 8      3001 diast   60         59
 9      3002 sist    15        107
10      3002 diast   15         76
# … with 4,334 more rows
```

Pivotage un peu plus complexe : tension artérielle

- Format long mais avec des colonnes pour systolique et diastolique

```
> ddata7 <- pivot_longer(ddata5, !patientID, names_to=c(".value","Time"), names_pattern="presion_?(.*)_(.*)")
> ddata7
```

```
# A tibble: 2,172 × 4
   patientID Time   sist diast
       <dbl> <chr> <dbl> <dbl>
 1      3001 15      101    60
 2      3001 30       98    60
 3      3001 45      100    58
 4      3001 60      104    59
 5      3002 15      107    76
 6      3002 30      109    76
 7      3002 45      109    76
 8      3002 60      106    73
 9      3003 15      133    83
10      3003 30      134    83
# … with 2,162 more rows
```

Ex: statistiques descriptives avec format long

- Utilisation de la fonction tapply

```
> tapply(ddata7$diast,ddata7$Time,summary)

$`15`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  38.00   60.00   68.00   68.96   76.00  114.00      63

$`30`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  20.00   62.00   69.00   69.28   76.00  111.00      65

$`45`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  30.00   62.00   69.00   69.22   76.00  118.00      66

$`60`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  38.00   62.00   70.00   69.14   76.00  102.00      66
```

Ex: statistiques descriptives avec format long

- Diagramme en boîte

```
> boxplot(diast ~ Time, data = ddata7)
```