Data Collection and Analysis: Available Resources

Apeksha Gupta, MPH

Qualities of Good Data Collection

- Consistency of methodology
- Reliable and error free
- Utility (understandable and interpretable variables)
- Completeness
- If an exempt study, no PHI is collected

Nicklaus Children's Hospital
Research Institute

Nicklaus Children's Hospital
Research Institute

Advantages of using REDCap

- **Features -** Electronic questionnaires, public survey, and data repository (much like SurveyMonkey)
- **Fast setup** - A simplified methodology for building databases quickly and easily
- **Secure and web-based** - Secure data input within MCHS network
- **Multi-site access** - REDCap projects can be used by researchers from multiple sites and institutions
- **Flexible** - Fully customizable. You are in total control of shaping your database
- **Mid-study modifications** - You may modify the database at any time during the course of your study
- **Easy in** - Data may be imported from external data sources to begin a study or to provide mid-study data uploads
- **Easy out** - Export data to common data analysis packages
- **Audit trails** - To track data manipulation and export procedures

Nicklaus Children's Hospital
Research Institute

REDCap Best Practices

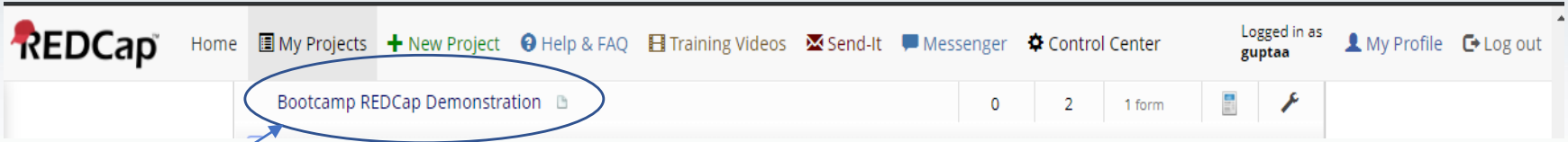
- Plan your data collection
- Document your data collection needs in a study protocol
- Group your data fields/questions based on how they are being collected.
  - Data source (chart review, patient assessment, lab report, etc.)
  - Time point (baseline clinic visit, follow-up visit, etc).
- Use categorical field types (yes/no, multiple choice, etc.) whenever possible
- Keep forms fairly short to minimize risk of data loss
- Involve a statistician early during the development of the database
- Test and Retest the Project
- Do not change variable names or values of categorical field types once you begin collecting real data

REDCap should be used when your study has…

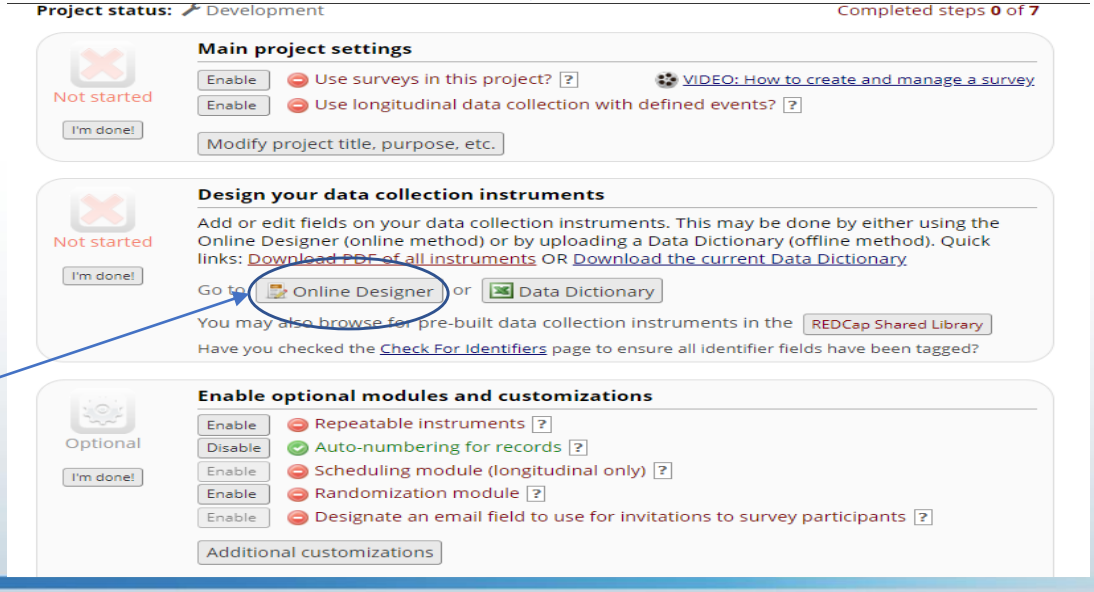
- More than 1 individual collecting data
- Many variables
- Large number of subjects/participants
- Providing/distributing a survey

Nicklaus Children's Hospital
Research Institute

REDCap Demonstration Data Collection Sheet



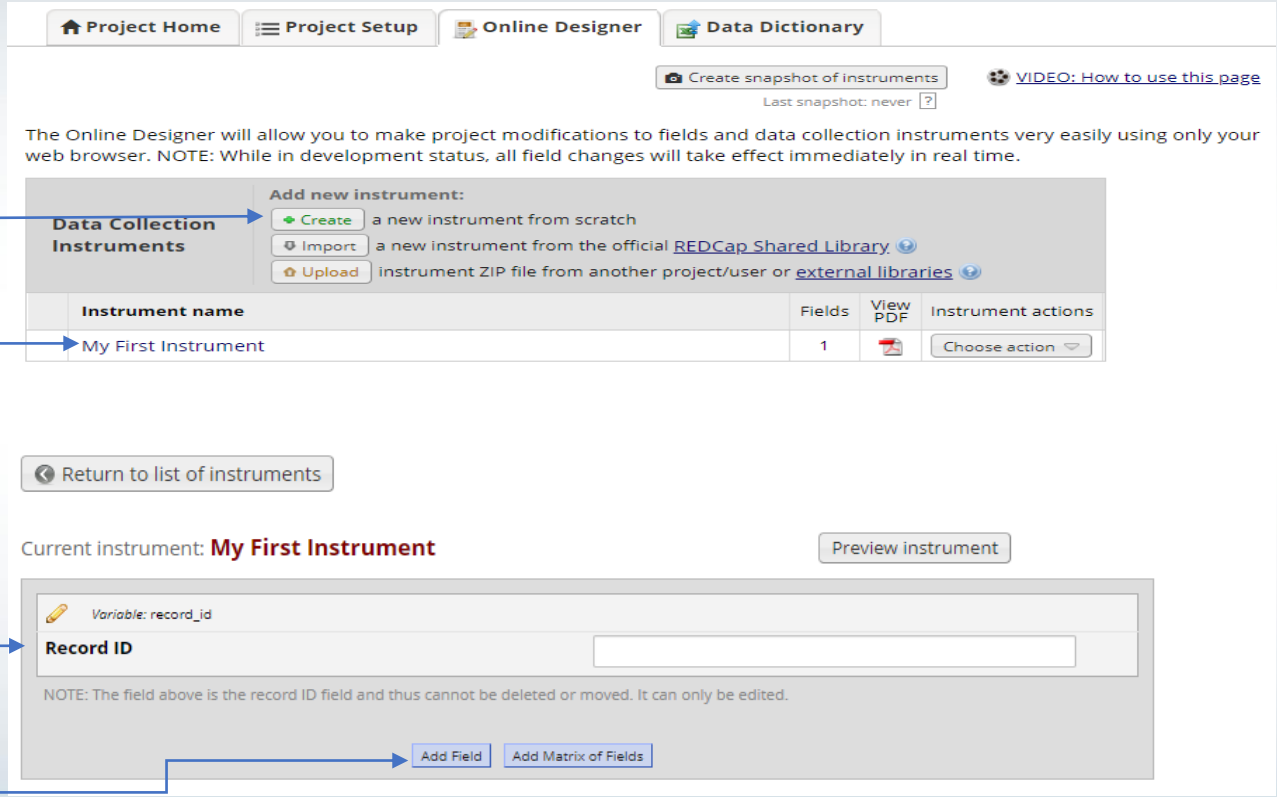
First page in REDCap when you login with your project already created by RAD



Click on Online designer

Nicklaus Children's Hospital
Research Institute

REDCap Demonstration Data Collection Sheet



Create new data collection instruments if you have more than 1

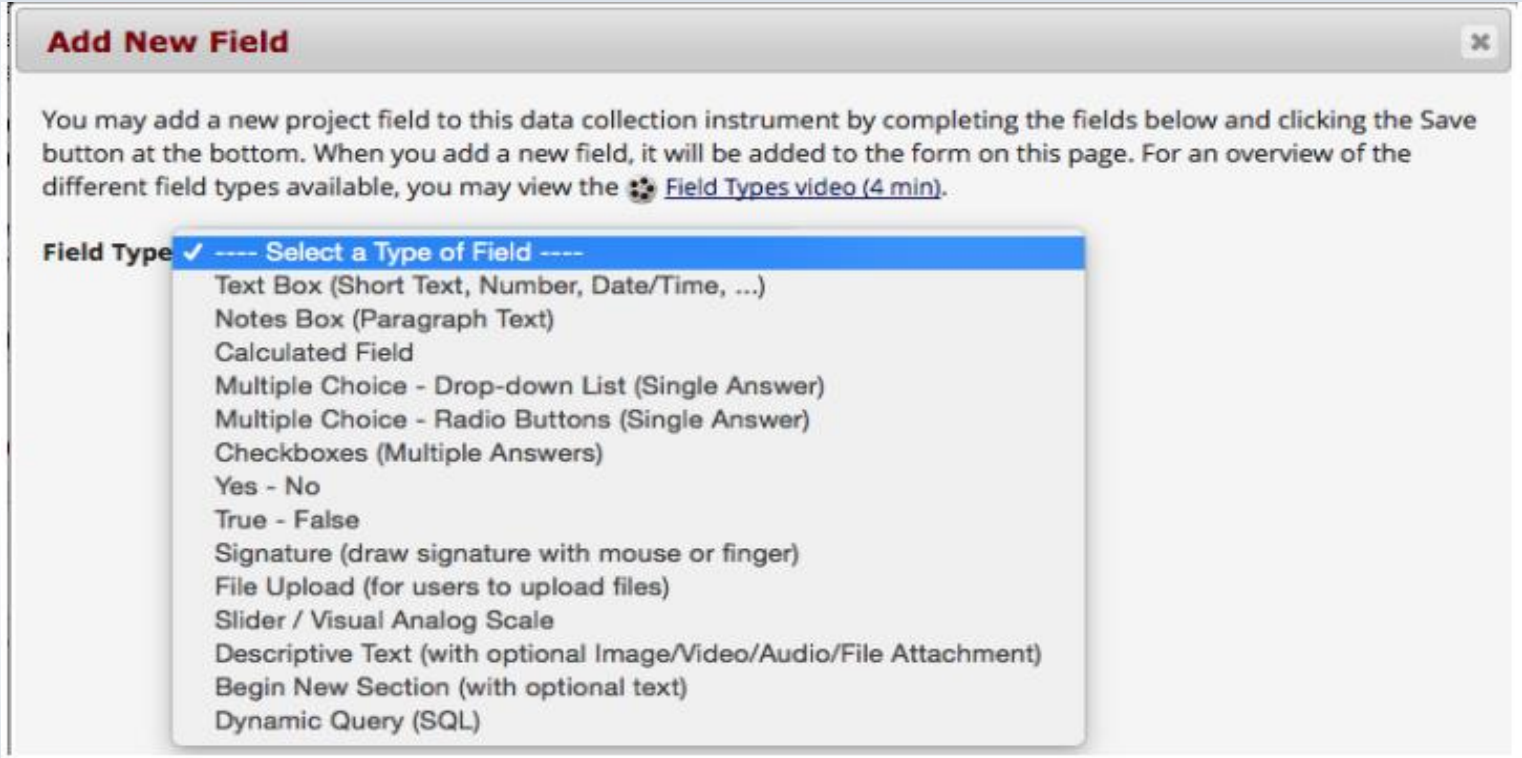
Click on My First Instrument. You can also rename the instrument

Record ID is populated automatically when you start collecting the data

Click on add field to add data collection variables

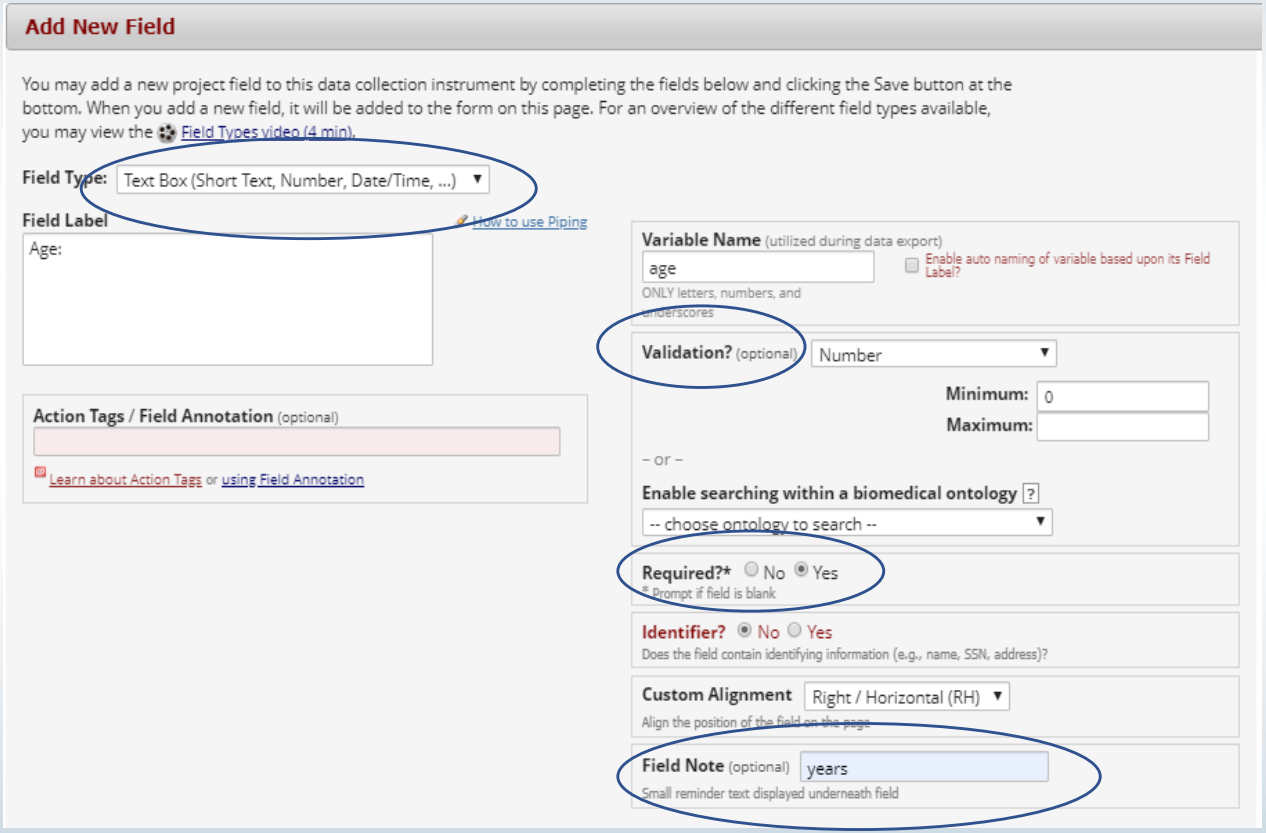
Nicklaus Children's Hospital
Research Institute

REDCap Demonstration Data Collection Sheet



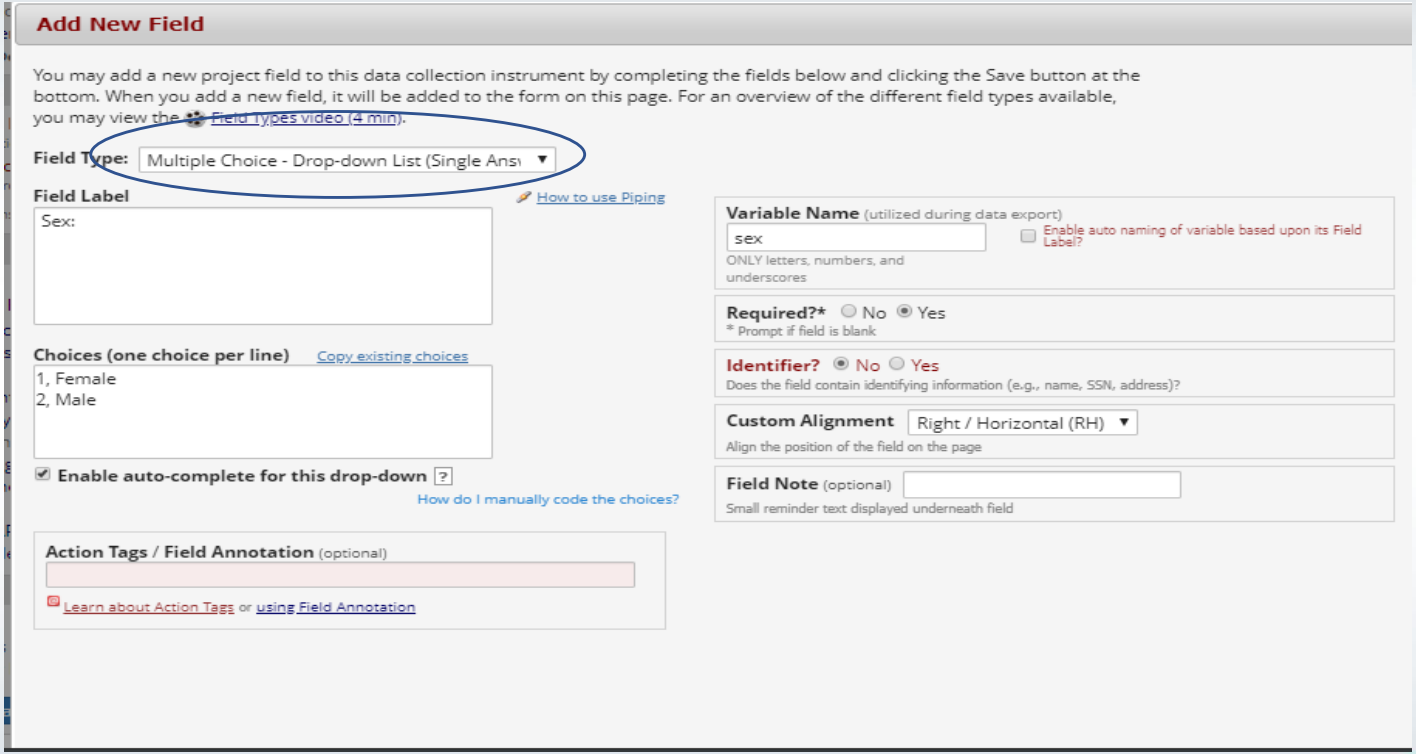
Nicklaus Children's Hospital
Research Institute

REDCap Demonstration Text Box



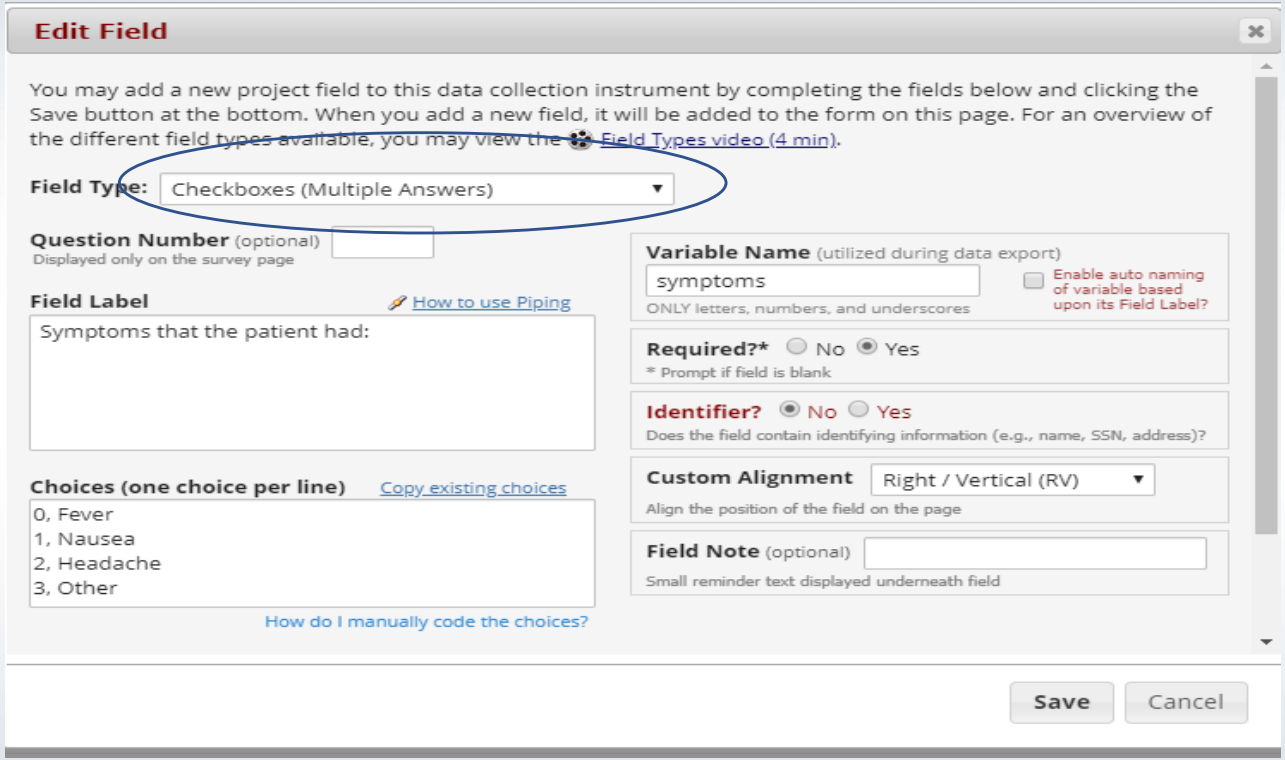
Nicklaus Children's Hospital
Research Institute

REDCap Demonstration Single Answer



Nicklaus Children's Hospital
Research Institute

REDCap Demonstration Multiple Answer



Nicklaus Children's Hospital
Research Institute

REDCap Demonstration Calculated Field

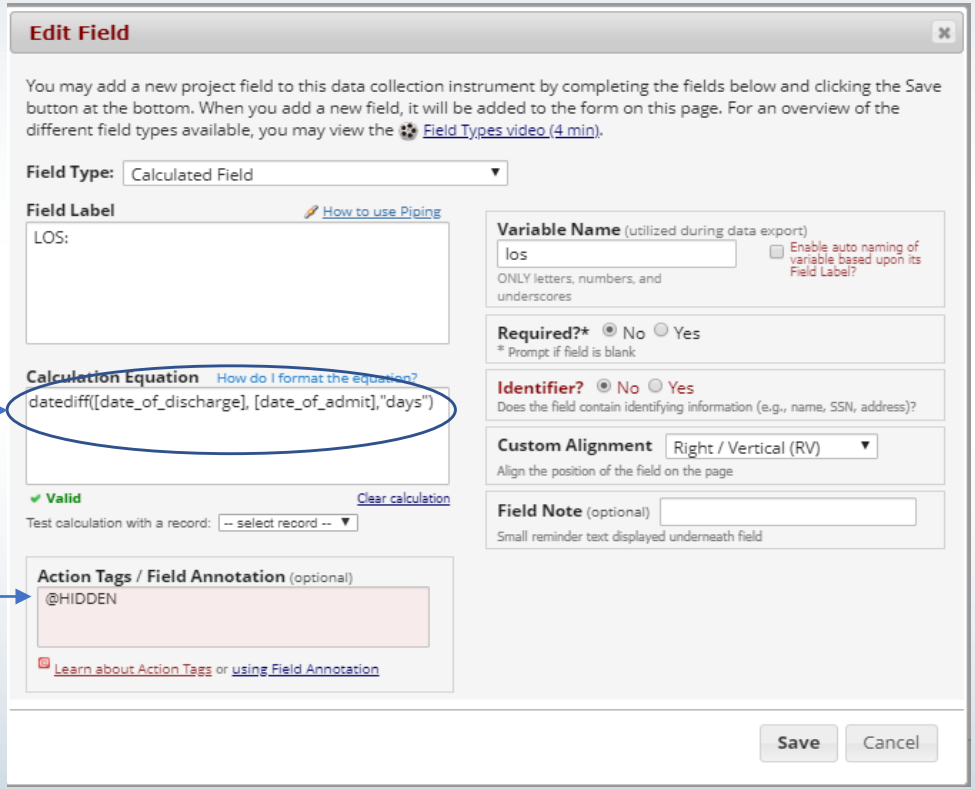
Create date of admit and date of discharge to calculate LOS.

Nicklaus Children's Hospital
Research Institute

REDCap Demonstration Calculated Field

Date difference formula to create LOS

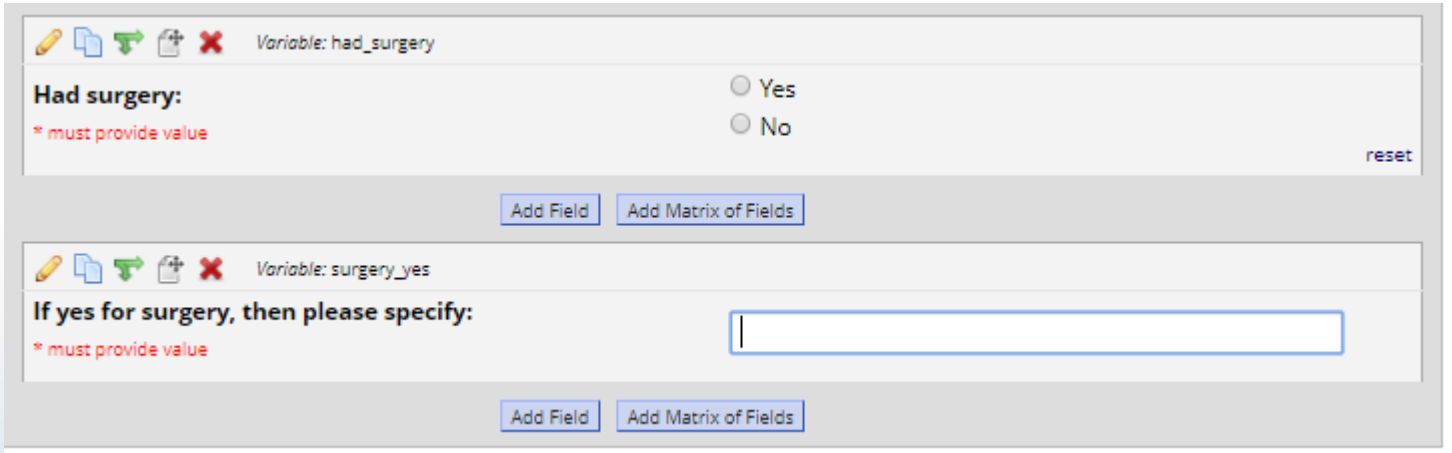
There are many action tags. @HIDDEN hides LOS from the survey link. So the respondent will not be able to see LOS



Nicklaus Children's Hospital
Research Institute

REDCap Demonstration Branching Logic

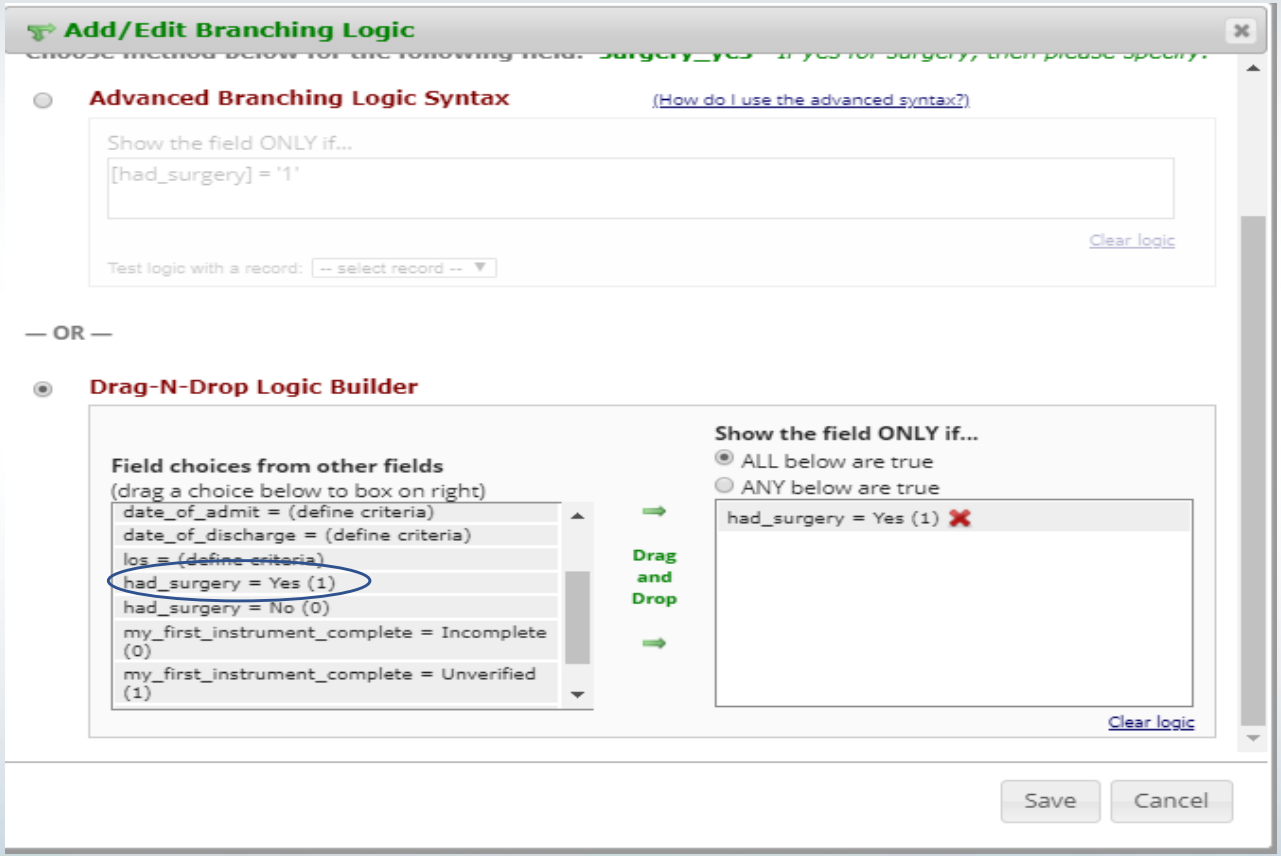
Branching Logic: may be employed when fields/questions need to be hidden for data entry under certain conditions. For instance, you may want to hide the question “If yes for surgery, then please specify?” until a “Yes” answer is checked for a previous question, “Had surgery?”



Reference: https://www.unmc.edu/vcr/_documents/unmc_redcap_usage.pdf

Nicklaus Children's Hospital
Research Institute

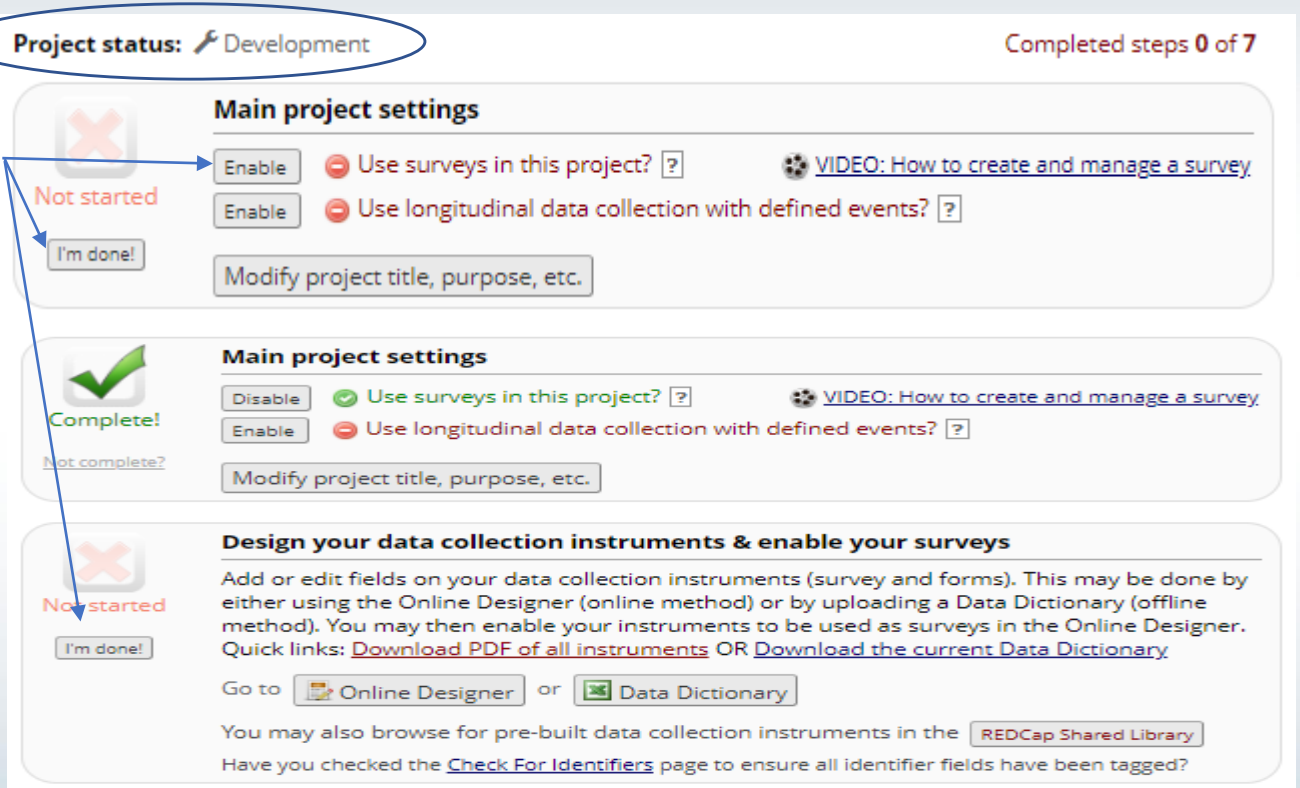
REDCap Demonstration Branching Logic



Nicklaus Children's Hospital
Research Institute

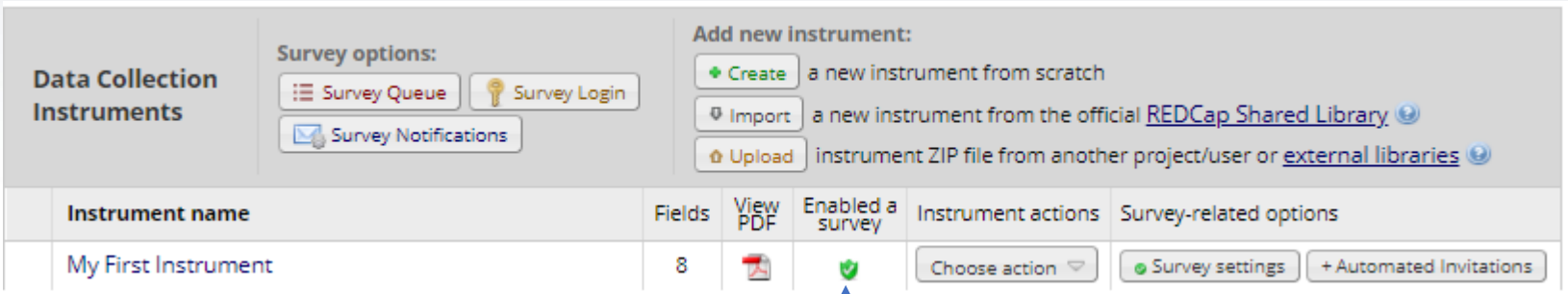
REDCap Demonstration Data Collection Sheet

Enable the survey and click on I am done once the data collection sheet is final



Nicklaus Children's Hospital
Research Institute

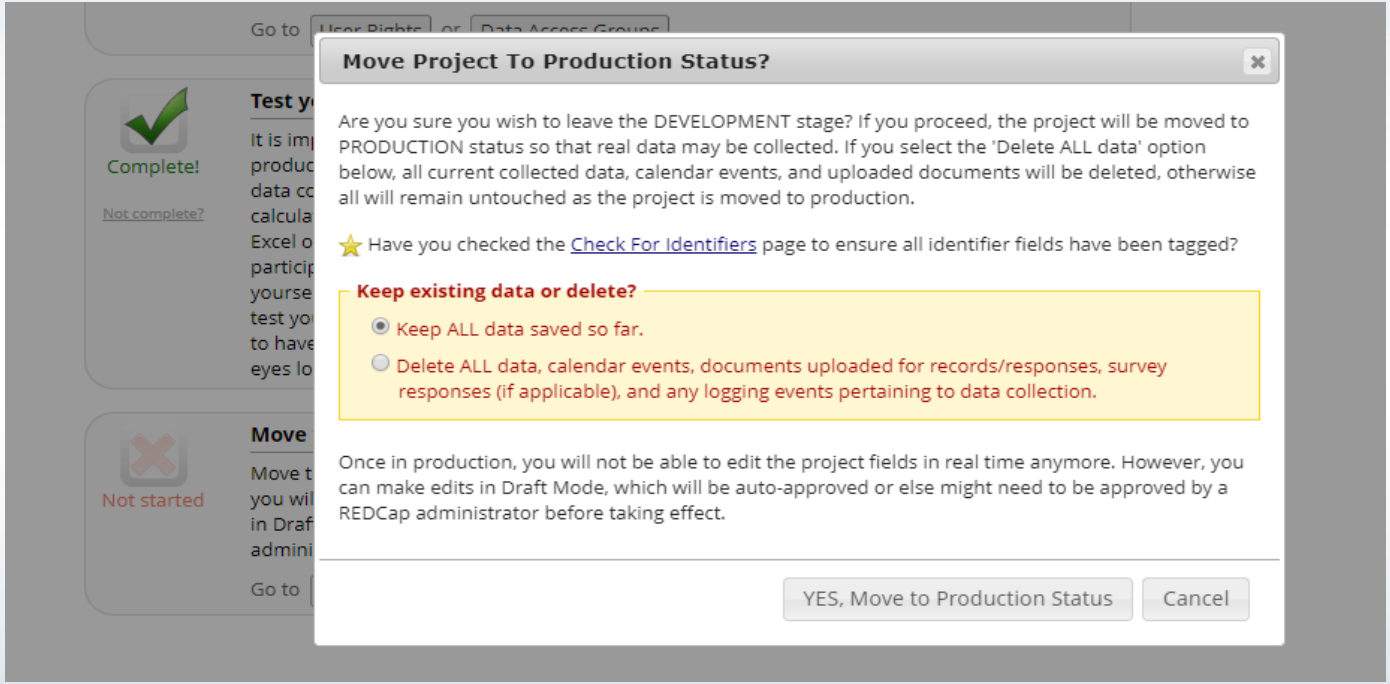
REDCap Demonstration Data Collection Sheet



Enable the survey under Online Designer

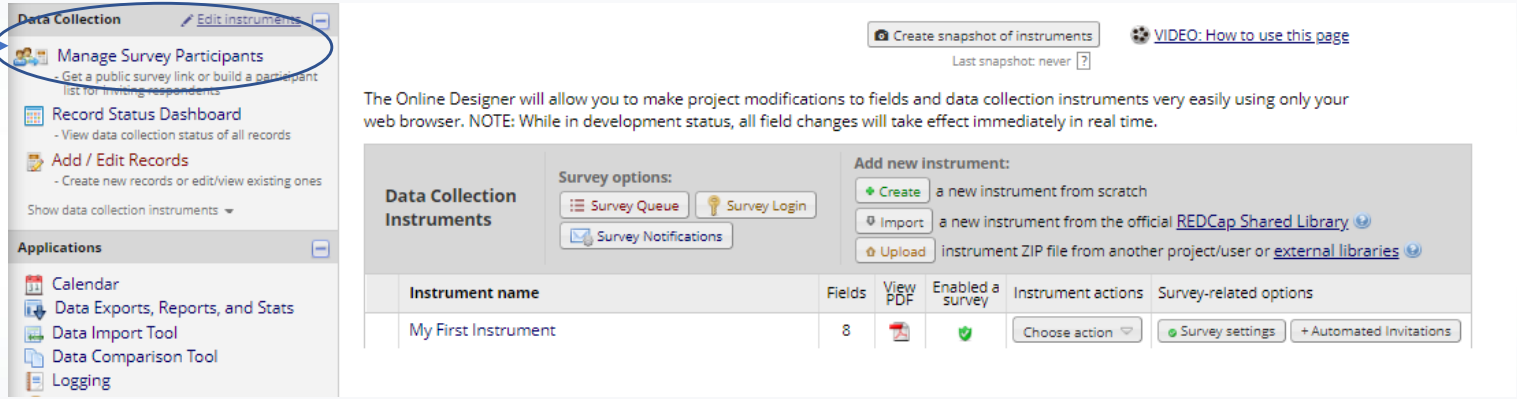
Nicklaus Children's Hospital Research Institute

REDCap Demonstration Move Project to Production



Nicklaus Children's Hospital
Research Institute

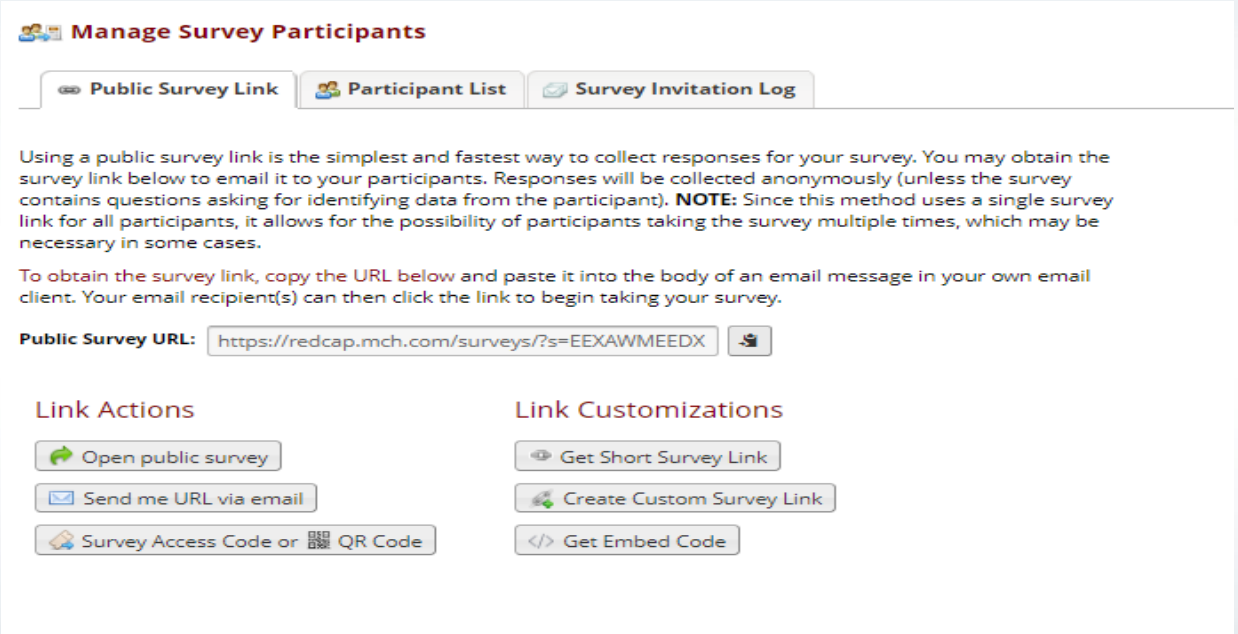
REDCap Demonstration Manage Survey Participant



Once the survey is enabled there will be a survey participant link generated by REDCap

Nicklaus Children's Hospital
Research Institute

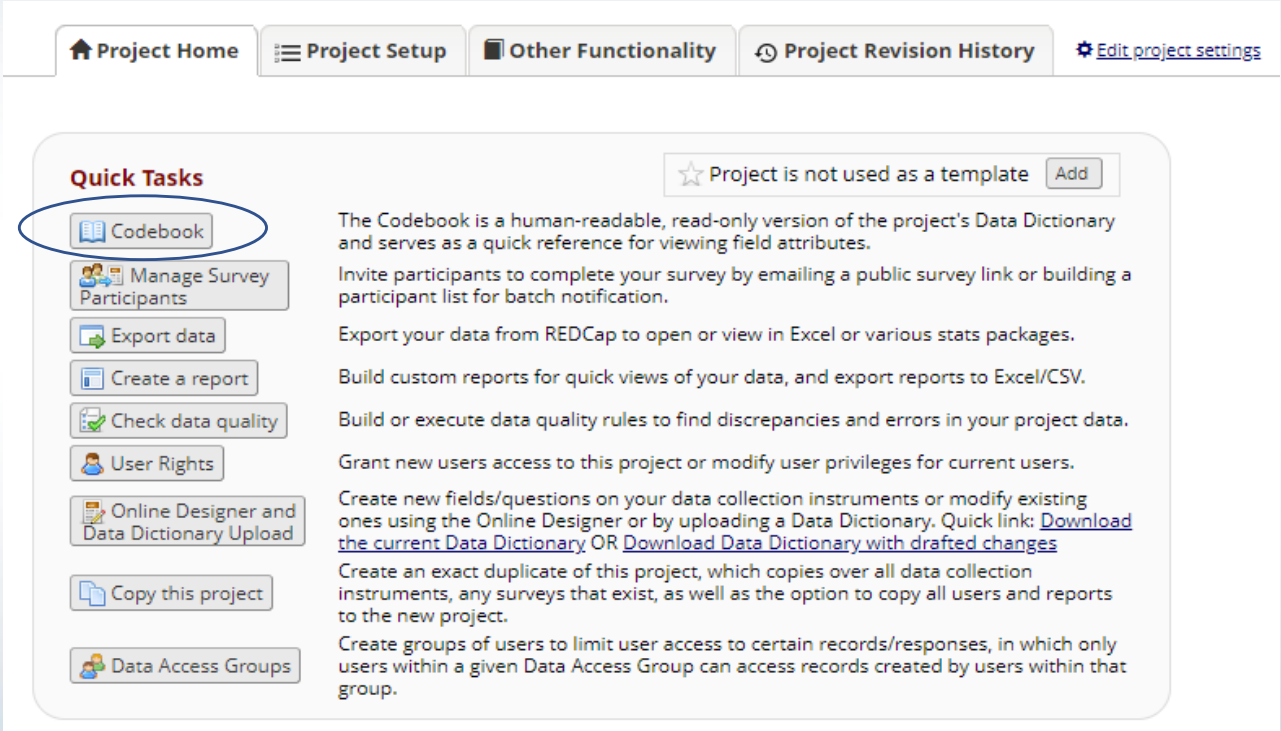
REDCap Demonstration Manage Survey Participant



https://redcap.mch.com/surveys/?s=EEXAWMEEDX

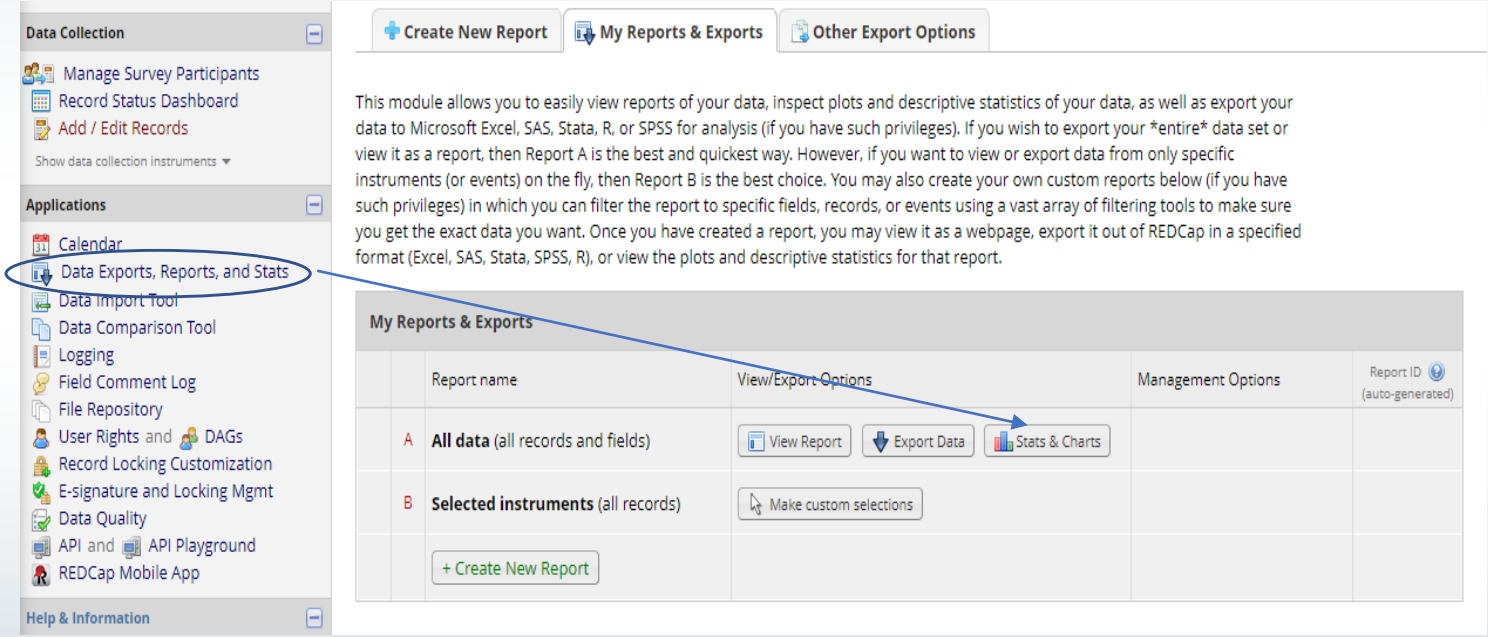
Nicklaus Children's Hospital
Research Institute

REDCap Demonstration Codebook



Nicklaus Children's Hospital
Research Institute

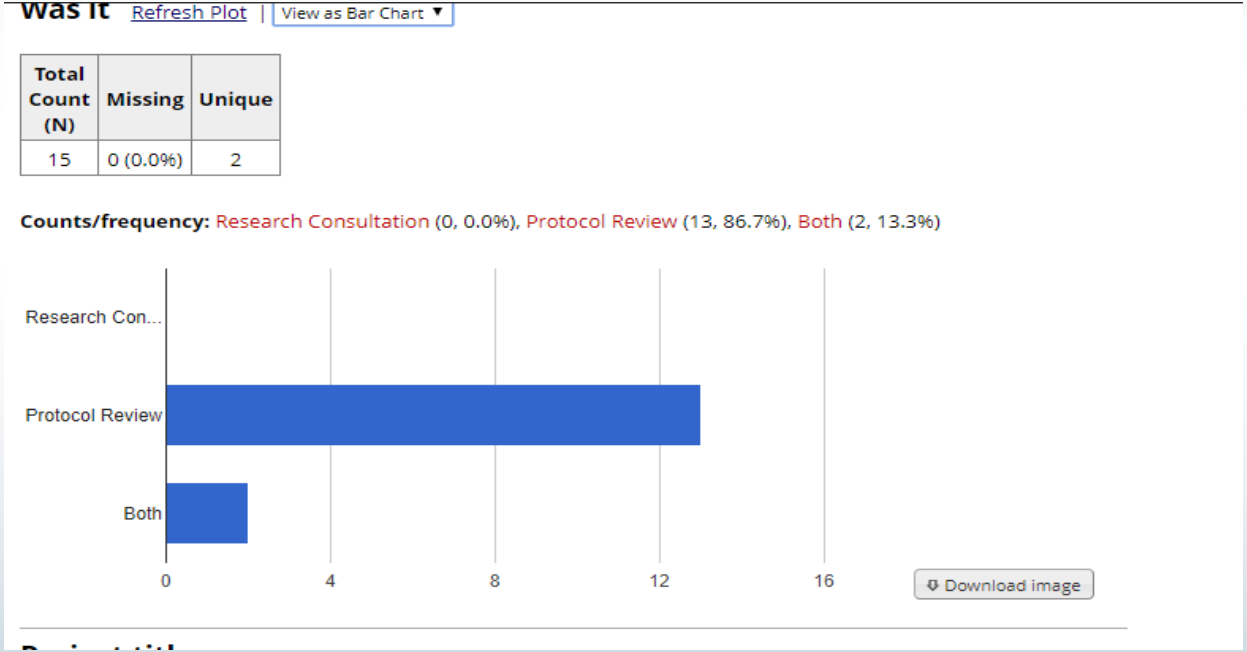
REDCap Demonstration Data Stats Feature



Nicklaus Children's Hospital
Research Institute

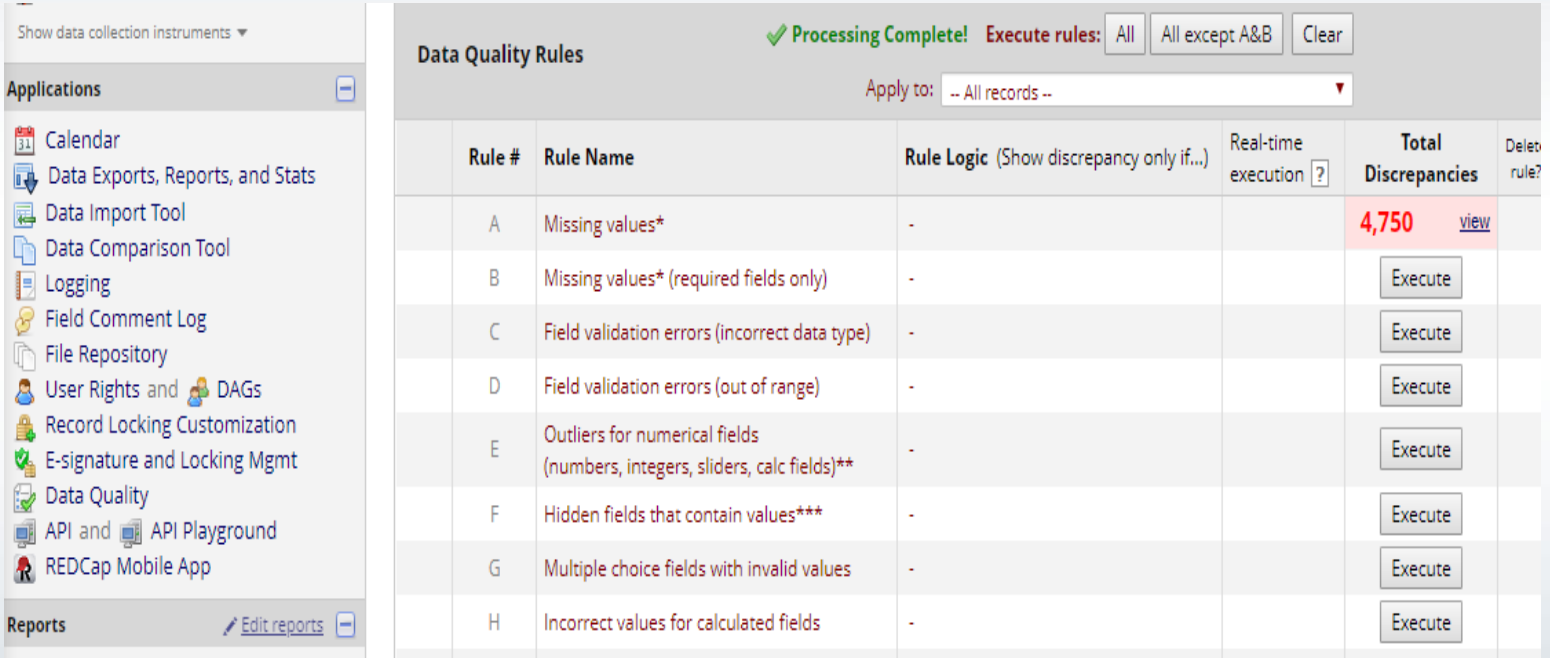
REDCap Demonstration Data Stats Feature

Data Stats will provide you the below statistics for each variable.



Nicklaus Children's Hospital
Research Institute

REDCap Demonstration Data Quality Feature



Nicklaus Children's Hospital
Research Institute

REDCap References

1. https://projectredcap.org/resources/videos/
2. https://www.unmc.edu/vcr/_documents/unmc_redcap_usage.pdf
3. https://www.ctsi.ufl.edu/wordpress/files/2019/02/Project-Creation-User-Guide.pdf
4. http://cri.uchicago.edu/wp-content/uploads/2015/12/REDCap-Beginners-Guide.pdf

Nicklaus Children's Hospital
Research Institute

REDCap Access Request Form

Once the study is approved by regulatory team, please inform RAD at rad@nicklaushealth.org to get access to REDCap to design the survey

Simultaneously please complete the REDCap access request form survey http://bit.ly/REDCapAccessRequestForm which needs to be signed by PI

After RAD team receives signed REDCap Access form along with approval letter from regulatory team, you will get access to design your project in REDCap

Once the survey is ready in REDCap, notify RAD at rad@nicklaushealth.org

RAD team will make sure that the survey is compliant with Data Collection Sheet

RAD team will move the survey to Production and change your access so you can start collecting the data. No further modifications will be possible

Once the data collection is over you can e-mail RAD team at rad@nicklaushealth.org to export the data

Nicklaus Children's Hospital
Research Institute

DATA ANALYSIS IN EXCEL

Nicklaus Children's Hospital
Research Institute

Descriptive Statistics

Mean	• The mean (or average) can be used with both discrete and continuous data. • Mean is particularly susceptible to the influence of outliers so should not be used if data has outliers.
Median	• The median is the middle score for a set of data that has been arranged in order of magnitude. • The median is less affected by outliers and skewed data.
Standard Deviation	• The standard deviation, like the mean, is appropriate when the continuous data is not significantly skewed or has outliers.
Quartile & Interquartile Range	• Quartiles are often reported along with the median when dealing with skewed and/or data with outliers.

Nicklaus Children's Hospital
Research Institute

Poll Question

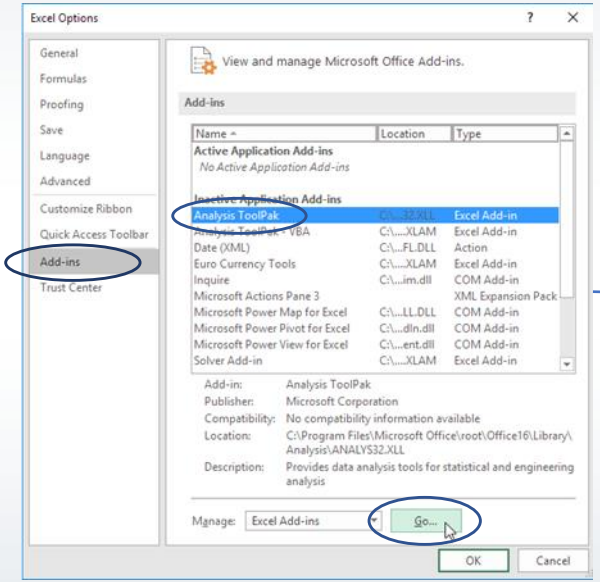
1. Which one of these statistics is not affected by outliers?

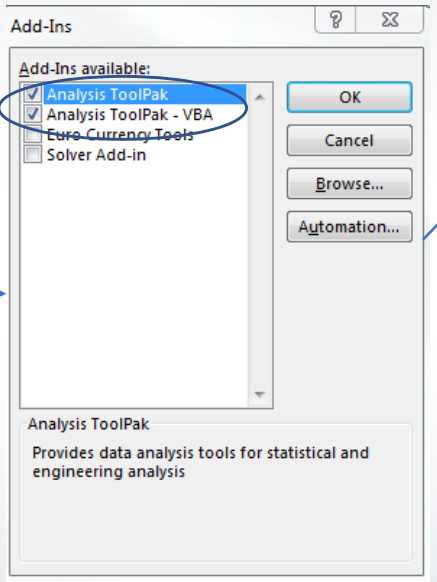
a. Mean
b. Interquartile range
c. Standard deviation
d. Range

Nicklaus Children's Hospital
Research Institute

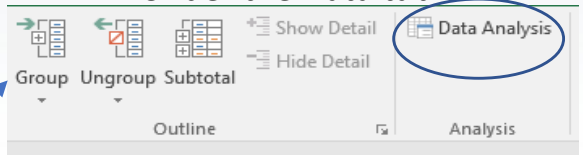
Reference: https://study.com/academy/answer/which-one-of-these-statistics-is-unaffected-by-outliers-a-mean-b-interquartile-range-c-standard-deviation-d-range.html

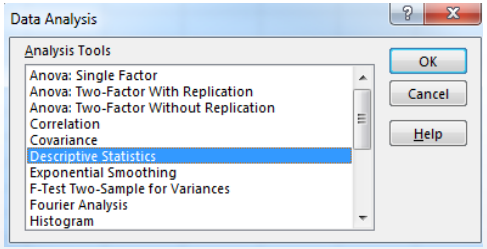
Downloading Data Analysis ToolPak in Excel





Under the Data tab

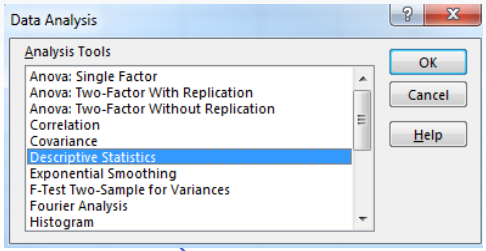




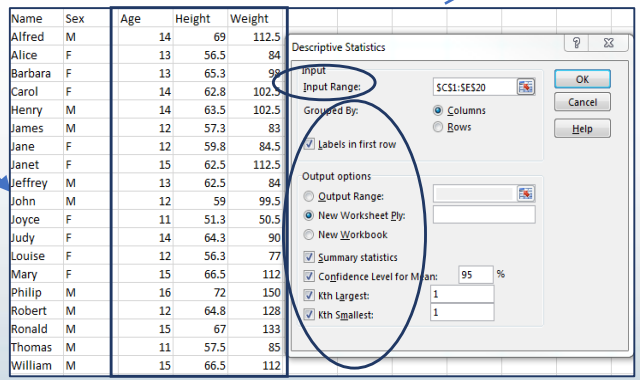
Nicklaus Children's Hospital
Research Institute

Data Analysis – Descriptive statistics using Data Analysis ToolPak

Descriptive statistics (Overall Mean, Median, Standard Deviation, Range, Sum, and Count)



Data selected in Input Range



Age		*Height*		*Weight*	
Mean	13.31578947	Mean	62.33684211	Mean	100.026316
Standard Error	0.342442479	Standard Error	1.17623173	Standard Error	5.22469867
Median	13	Median	62.8	Median	99.5
Mode	12	Mode	62.5	Mode	112.5
Standard Deviation	1.492672159	Standard Deviation	5.127075247	Standard Deviation	22.7739335
Sample Variance	2.228070175	Sample Variance	26.28690058	Sample Variance	518.652047
Kurtosis	-1.11092552	Kurtosis	-0.138969241	Kurtosis	0.68336484
Skewness	0.063611668	Skewness	-0.259669589	Skewness	0.18335097
Range	5	Range	20.7	Range	99.5
Minimum	11	Minimum	51.3	Minimum	50.5
Maximum	16	Maximum	72	Maximum	150
Sum	253	Sum	1184.4	Sum	1900.5
Count	19	Count	19	Count	19
Largest(1)	16	Largest(1)	72	Largest(1)	150
Smallest(1)	11	Smallest(1)	51.3	Smallest(1)	50.5
Confidence Level(95.0%)	0.719444951	Confidence Level(95.0%)	2.471171167	Confidence Level(95.0%)	10.9766846

Nicklaus Children's Hospital
Research Institute

Poll Question

2. A list of 5 pulse rates is: 72, 80, 84, 75, 92. What is the median for this list?

a. 74
b. 76
c. 77
d. 80

Nicklaus Children's Hospital
Research Institute

Reference: https://study.com/academy/answer/a-list-of-5-pulse-rates-is-70-64-80-74-92-what-is-the-median-for-this-list-a-74-b-76-c-77-d-80.html

Data Analysis – Descriptive statistics using Excel

Descriptive statistics (Quartiles and Interquartile)

	A	B	C	D	E	F	G
1	Name	Race	Sex	Age	Height	Weight	To Calculate Quartile for Weight
2	Alice	White	F	13	56.5	84	50.5
3	Barbara	American African	F	13	65.3	98	84.25
4	Carol	Asian	F	14	62.8	102.5	=QUARTILE(F2:F20,2)
5	Jane	White	F	12	59.8	84.5	QUARTILE(array, quart)
6	Janet	American African	F	15	62.5	112.5	150
7	Joyce	Asian	F	11	51.3	50.5	
8	Judy	White	F	14	64.3	90	
9	Louise	American African	F	12	56.3	77	
10	Mary	Asian	F	15	66.5	112	
11	Alfred	American African	M	14	69	112.5	
12	Henry	American African	M	14	63.5	102.5	
13	James	White	M	12	57.3	83	
14	Jeffrey	American African	M	13	62.5	84	
15	John	White	M	12	59	99.5	
16	Philip	American African	M	16	72	150	
17	Robert	Asian	M	12	64.8	128	
18	Ronald	White	M	15	67	133	
19	Thomas	American African	M	11	57.5	85	
20	William	American African	M	15	66.5	112	

IF QUART EQUALS	QUARTILE RETURNS
0	Minimum value
1	First quartile (25th percentile)
2	Median value (50th percentile)
3	Third quartile (75th percentile)
4	Maximum value

Interquartile range

Third quartile – First Quartile

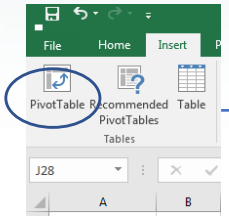
= 112.25-84.25

=28

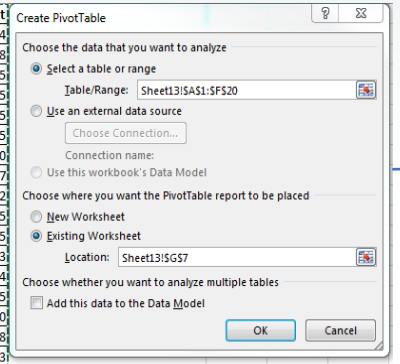
Nicklaus Children's Hospital Research Institute

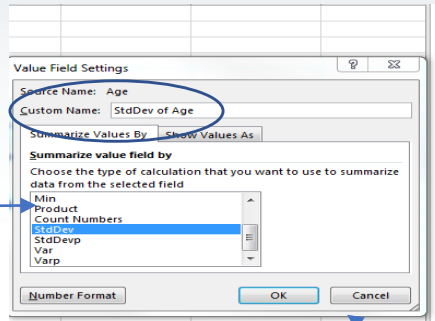
Data Analysis – Basic statistics using PivotTable

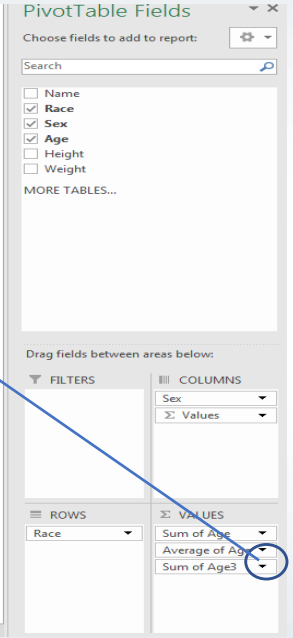
Alternative way: Descriptive statistics (Mean and Standard Deviation)



Name	Race	Sex	Age	Height	Weight
Alice	White	F	13	56.5	84
Barbara	American African	F	13	65.3	98
Carol	Asian	F	14	62.8	102.5
Jane	White	F	12	59.8	84.5
Janet	American African	F	15	62.5	112.5
Joyce	Asian	F	11	51.3	50.5
Judy	White	F	14	64.3	90
Louise	American African	F	12	56.3	77
Mary	Asian	F	15	66.5	112
Alfred	American African	M	14	69	112.5
Henry	American African	M	14	63.5	102.5
James	White	M	12	57.3	83
Jeffrey	American African	M	13	62.5	84
John	White	M	12	59	99.5
Philip	American African	M	16	72	150
Robert	Asian	M	12	64.8	128
Ronald	White	M	15	67	133
Thomas	American African	M	11	57.5	85
William	American African	M	15	66.5	112







Click on the arrow next to "Sum of Age 3" you will get an option for "Value Field Setting". Select "Value Field Setting" to get the above box and select StdDev to get StdDev for Age. Similar process has been done for "Average of Age".

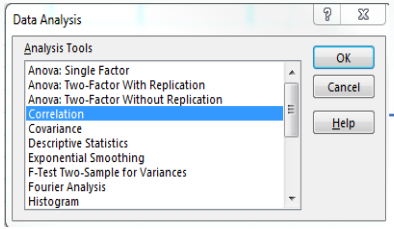
Column Labels	F			M			Total Sum of Age	Total Average of Age	Total StdDev of Age
Row Labels	Sum of Age	Average of Age	StdDev of Age	Sum of Age	Average of Age	StdDev of Age			
American African	40	13.33333333	1.527525232	83	13.83333333	1.722401424	123	13.66666667	1.58113883
Asian	40	13.33333333	2.081665999	12	12	#DIV/0!	52	13	1.825741858
White	39	13	1	39	13	1.732050808	78	13	1.264911064
Grand Total	119	13.22222222	1.394433378	134	13.4	1.646545205	253	13.31578947	1.492672159

Sample size of 1

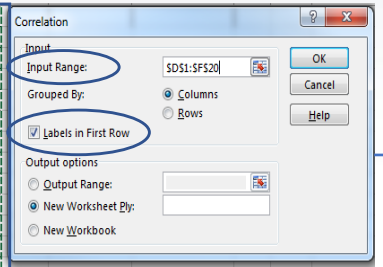
Nicklaus Children's Hospital
Research Institute

Data Analysis – Descriptive statistics using Data Analysis ToolPak

Descriptive statistics (Correlation)



Name	Race	Sex	Age	Height	Weight
Alice	White	F	13	56.5	84
Barbara	American African	F	13	65.3	98
Carol	Asian	F	14	62.8	102.5
Jane	White	F	12	59.8	84.5
Janet	American African	F	15	62.5	112.5
Joyce	Asian	F	11	51.3	50.5
Judy	White	F	14	64.3	90
Louise	American African	F	12	56.3	77
Mary	Asian	F	15	66.5	112
Alfred	American African	M	14	69	112.5
Henry	American African	M	14	63.5	102.5
James	White	M	12	57.3	83
Jeffrey	American African	M	13	62.5	84
John	White	M	12	59	99.5
Philip	American African	M	16	72	150
Robert	Asian	M	12	64.8	128
Ronald	White	M	15	67	133
Thomas	American African	M	11	57.5	85
William	American African	M	15	66.5	112



	Age	Height	Weight
Age	1		
Height	0.811434	1	
Weight	0.740885	0.877785	1

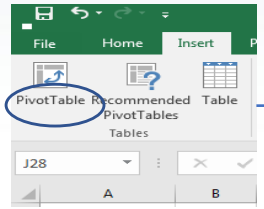
Size of Correlation	Interpretation
.90 to 1.00 (−.90 to −1.00)	Very high positive (negative) correlation
.70 to .90 (−.70 to −.90)	High positive (negative) correlation
.50 to .70 (−.50 to −.70)	Moderate positive (negative) correlation
.30 to .50 (−.30 to −.50)	Low positive (negative) correlation
.00 to .30 (.00 to −.30)	negligible correlation

Reference: Diener E, Wirtz D, Tov W, Kim-Prieto C, Choi DW, Oishi S, Biswas-Diener R. New well-being measures: Short scales to assess flourishing and positive and negative feelings. Social Indicators Research. 2010 Jun 1;97(2):143-56.

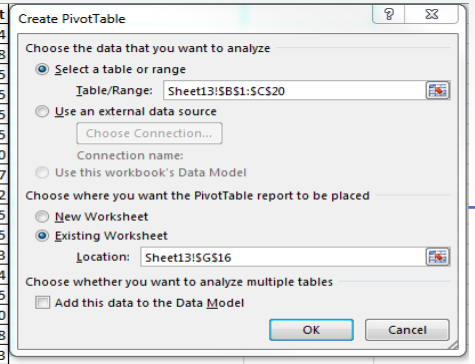
Nicklaus Children's Hospital
Research Institute

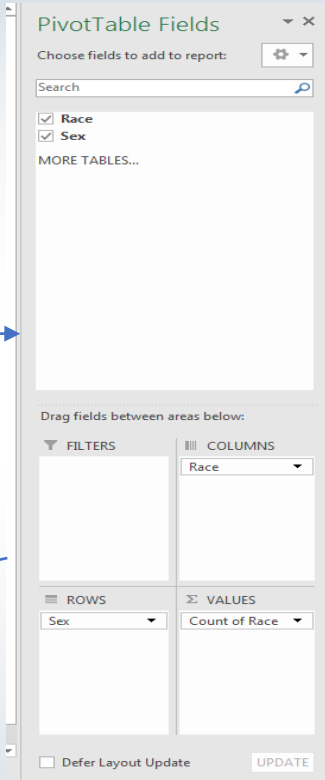
Data Analysis – Descriptive statistics using Pivot Table

Descriptive statistics (Frequencies)



Name	Race	Sex	Age	Height	Weight
Alice	White	F	13	56.5	84
Barbara	American African	F	13	65.3	98
Carol	Asian	F	14	62.8	102.5
Jane	White	F	12	59.8	84.5
Janet	American African	F	15	62.5	112.5
Joyce	Asian	F	11	51.3	50.5
Judy	White	F	14	64.3	90
Louise	American African	F	12	56.3	77
Mary	Asian	F	15	66.5	112
Alfred	American African	M	14	69	112.5
Henry	American African	M	14	63.5	102.5
James	White	M	12	57.3	83
Jeffrey	American African	M	13	62.5	84
John	White	M	12	59	99.5
Philip	American African	M	16	72	150
Robert	Asian	M	12	64.8	128
Ronald	White	M	15	67	133
Thomas	American African	M	11	57.5	85
William	American African	M	15	66.5	112





Count of Race	Column Labels			
Row Labels	American African	Asian	White	Grand Total
F	3	3	3	9
M	6	1	3	10
Grand Total	9	4	6	19

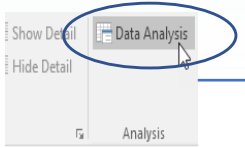
Nicklaus Children's Hospital
Research Institute

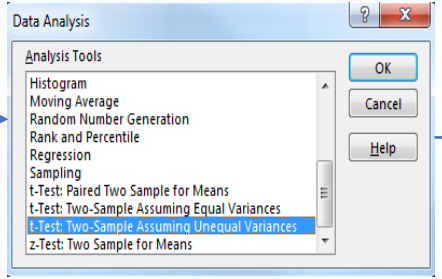
Data Analysis – Basic statistics using Data Analysis ToolPak

Testing a difference in means between groups

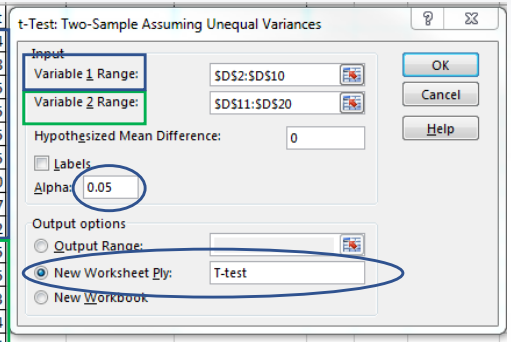
(Null Hypothesis [Ho]: There no difference in the Weight between Males and Females)

Under the Data tab





Name	Race	Sex	Age	Height	Weight
Alice	White	F	13	56.5	84
Barbara	American African	F	13	65.3	98
Carol	Asian	F	14	62.8	102.5
Jane	White	F	12	59.8	84.5
Janet	American African	F	15	62.5	112.5
Joyce	Asian	F	11	51.3	50.5
Judy	White	F	14	64.3	90
Louise	American African	F	12	56.3	77
Mary	Asian	F	15	66.5	112
Alfred	American African	M	14	69	112.5
Henry	American African	M	14	63.5	102.5
James	White	M	12	57.3	83
Jeffrey	American African	M	13	62.5	84
John	White	M	12	59	99.5
Philip	American African	M	16	72	150
Robert	Asian	M	12	64.8	128
Ronald	White	M	15	67	133
Thomas	American African	M	11	57.5	85
William	American African	M	15	66.5	112



t-Test: Two-Sample Assuming Unequal Variances		
	Female	*Males*
Mean	90.111111	108.95
Variance	375.73611	516.525
Observations	9	10
Hypothesized Mean Difference	0	
df	17	
t Stat	-1.949304	
P(T<=t) one-tail	0.0339772	
t Critical one-tail	1.7396067	
P(T<=t) two-tail	0.0679545	
t Critical two-tail	2.1098156	

Note: Reject Ho when P is <0.05

We conclude that there is no significant difference in Weight between males and females. (p=0.0679545> Alpha 0.05)

Nicklaus Children's Hospital Research Institute

Poll Question

3. A result is called “statistically significant” whenever:

a. The null hypothesis is true.
b. The alternative hypothesis is true.
c. The p-value is less or equal to the significance level.
d. The p-value is larger than the significance level.

Reference: https://www.ics.uci.edu/~jutts/8/SampleFinalMCKey.pdf

Data Analysis – Basic statistics using Excel

Calculating Odds Ratio (OR) with confidence interval

Exposure	Lung Cancer	Without Lung Cancer	Grand Total
Smokers	a 17	b 5	22
Non-Smokers	c 9	d 69	78
Grand Total	26	74	100

OR = (a/c)/(b/d) or (a*d)/(b*c) = 17*69/5*9= 1,173/45 = 26.07

Upper = $e^{\ln(OR) + 1.96*\sqrt{1/a + 1/b + 1/c + 1/d}}$ =EXP(LN(26.07)+(1.96*(SQRT((1/17)+(1/5)+(1/9)+(1/69))))) = 87.87

Lower = $e^{\ln(OR) - 1.96*\sqrt{1/a + 1/b + 1/c + 1/d}}$ = EXP(LN(26.07)-(1.96*(SQRT((1/17)+(1/5)+(1/9)+(1/69))))) = 7.73

You can directly copy paste the formula in excel and change numbers to get Upper and Lower Confidence Interval.

Interpretation: Patients who smoked were 26.07 times more likely to have lung cancer than patients who did not smoke.

Note: If confidence interval includes 1 then the result is not statistically significant

Nicklaus Children's Hospital Research Institute

Poll Question

4. A case-control study of 1700 participants looked at the association between Tamoxifen and uterine cancer. The study included 689 cases. There were 139 cases and 58 controls taking Tamoxifen.
Calculate the odds ratio of the above study

	Uterine Cancer	
Tamoxifen	**Yes**	**No**
Yes	139	58
No	550	953

a. 3.20
b. 4.15
c. 4.20
d. 4.00

Nicklaus Children's Hospital
Research Institute

Reference Unit 4 Practice Questions eta.health.usf.edu

Poll Question

5. Odds Ratio of Tamoxifen and risk of developing uterine cancer is 4.15. Choose the correct interpretation:

a. Tamoxifen has higher risk of developing uterine cancer
b. Tamoxifen has lower risk of developing uterine cancer

Effect Size

- Measures the strength of the relationship between two variables on a numeric scale.
- Statistic effect size helps us in determining if the difference is real.
- In hypothesis testing, effect size, power, sample size, and critical significance level are related to each other.

Nicklaus Children's Hospital
Research Institute

Reference https://www.statisticssolutions.com/statistical-analyses-effect-size/

Effect size calculation

Cohen'd:

$$d = M_1 - M_2 / s_{pooled}$$

M1 = mean of group 1
M2 = mean of group 2
Spooled = pooled standard deviations for the two groups

Effect size of 0.2 is a small effect, 0.5 is a medium effect, and 0.8 is a large effect.

Pooled standard deviation equation:

$$\sqrt{[(s_1^2 + s_2^2) / 2}$$

Where S_1 and S_2 are the standard deviation of group 1 and 2, respectively.

Nicklaus Children's Hospital
Research Institute

Reference https://www.leeds.ac.uk/educol/documents/00002182.htm

Effect size calculation

Phi (ϕ): 2 x 2 Contingency table

Phi = √(X2 / n)

X2 is the Chi-Square test statistic
n = total number of observations

How to Interpret

A value of ϕ = 0.1 is considered to be a small effect, 0.3 a medium effect, and 0.5 a large effect.

Nicklaus Children's Hospital
Research Institute

Reference https://www.leeds.ac.uk/educol/documents/00002182.htm

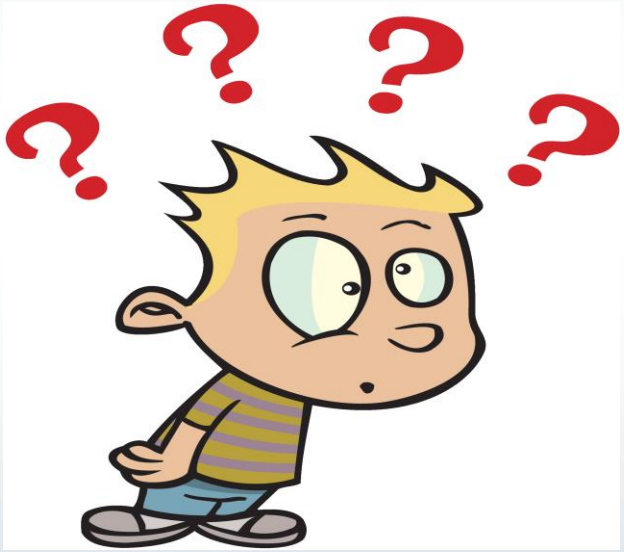
Other Statistical Assistance

- If your study requires more complicated statistical approach please contact the RAD team for additional support.
- If the above is the case, please adhere to the research project deadlines so that we can best provide assistance.

2 years research project deadlines

October 12th, 2020	Present research concept: https://is.gd/research_consultation
December 11th, 2020	Submit draft protocol in protocol builder for Regulatory and RAD review and let us know about it by emailing to: rad@nicklaushealth.org and RegulatoryAffairs@nicklaushealth.org
January 11th, 2021	Submit application to conduct research with final Research Protocol and all supporting documents https://rebrand.ly/research_eApp
July 12th, 2021	Complete all data collection **in REDCap** and contact biostatistician to start data analysis (analysis can take up to 6 weeks)
October 1st, 2021	Complete manuscript draft and send to rad@nicklaushealth.org and all co-authors for review.
October 2021 - May 2022	Finalize and submit/re-submit manuscript for publication

Nicklaus Children's Hospital
Research Institute



Nicklaus Children's Hospital
Research Institute