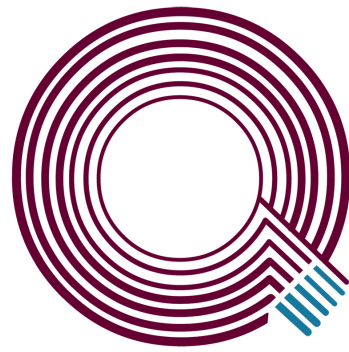




QUALIANCE

Trusted Partner since 2010
5 Years

Data Transparency – Data De-Identification and The Role of the Programmer

PhUSE 2014, London

Jean-Marc Ferran

Consultant & Owner

Data De-Identification

Risk

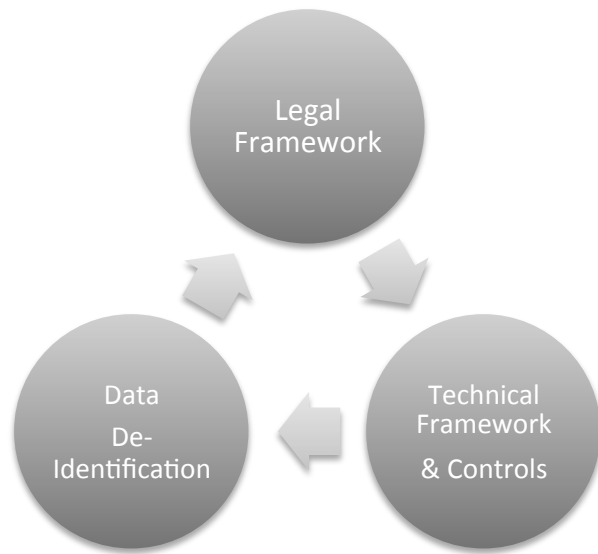
Data
Utility

Our
Role

PhUSE
DI WG



Data Sharing

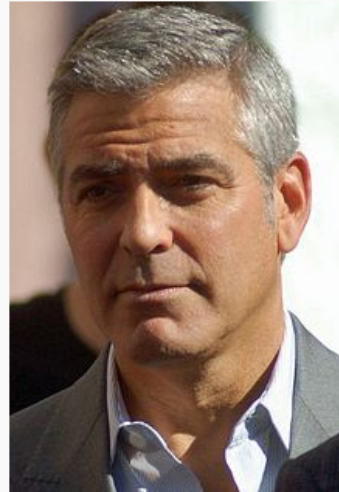


Who is Georges?



Patient ID	DoB	Age	G			Country	Partner Age
1	12APR1963	51	M	 Clooney at a ceremony for John Wells to receive a star on the Hollywood Walk of Fame in January 2012 Born George Timothy Clooney May 6, 1961 (age 53) Lexington, Kentucky, U.S. Occupation Actor, filmmaker Years active 1978–present Spouse(s) Talia Balsam (m. 1989–93) Amal Alamuddin (m. 2014) Parents Nick Clooney Nina Warren Relatives Rosemary Clooney (aunt) Jose Ferrer (uncle) Miguel Ferrer (cousin) Rafael Ferrer (cousin)		Canada	48
2	28MAY1974	40	M			France	41
3	06MAY1961	53	M			United States	36
4	28MAY1954	60	F			Spain	65
5	14JUL1969	45	M			Brazil	41
6	13AUG1964	50	F			Argentina	45
7	18MAR1961	53	M			United States	48
8	22JAN1961	53	M			United States	37
9	27SEP1924	90	M			Canada	73
10	07FEB1956	58	Male		White	Canada	62

George Clooney



Clooney at a ceremony for John Wells to receive a star on the [Hollywood Walk of Fame](#) in January 2012

Born George Timothy Clooney
May 6, 1961 (age 53)
[Lexington, Kentucky, U.S.](#)

Occupation Actor, filmmaker

Years active 1978–present

Spouse(s) [Talia Balsam](#) (m. 1989–93)
[Amal Alamuddin](#) (m. 2014)

Parents [Nick Clooney](#)
[Nina Warren](#)

Relatives [Rosemary Clooney](#) (aunt)
[Jose Ferrer](#) (uncle)
[Miguel Ferrer](#) (cousin)
[Rafael Ferrer](#) (cousin)



Who is Georges?



  	Patient ID	Age Category	Age	Gender	Race	Country	Partner Age
	1	<89	51	Male	White	Canada	
	2	<89	40	Male	Asian	France	
	3	<89	53	Male	White	United States	
	4	<89	60	Female	Black	Spain	
	5	<89	45	Male	Black	Brazil	
	6	<89	50	Female	White	Argentina	
	7	<89	53	Male	White	United States	
	8	<89	53	Male	White	United States	
	9	≥89	.	Male	White	Canada	
	10	<89	58	Male	White	Canada	

Who is Georges?



	Patient ID	Age Category 2	Age	Gender	Race	Continent	Partner Age
	1	50-59		Male	White	North America	
	2	40-49		Male	Asian	Europe	
	3	50-59		Male	White	North America	
	4	60-69		Female	Black	Europe	
	5	40-49		Male	Black	South America	
	6	50-59		Female	White	South America	
	7	50-59		Male	White	North America	
	8	50-59		Male	White	North America	
	9	≥89		Male	White	North America	
	10	50-59		Male	White	North America	



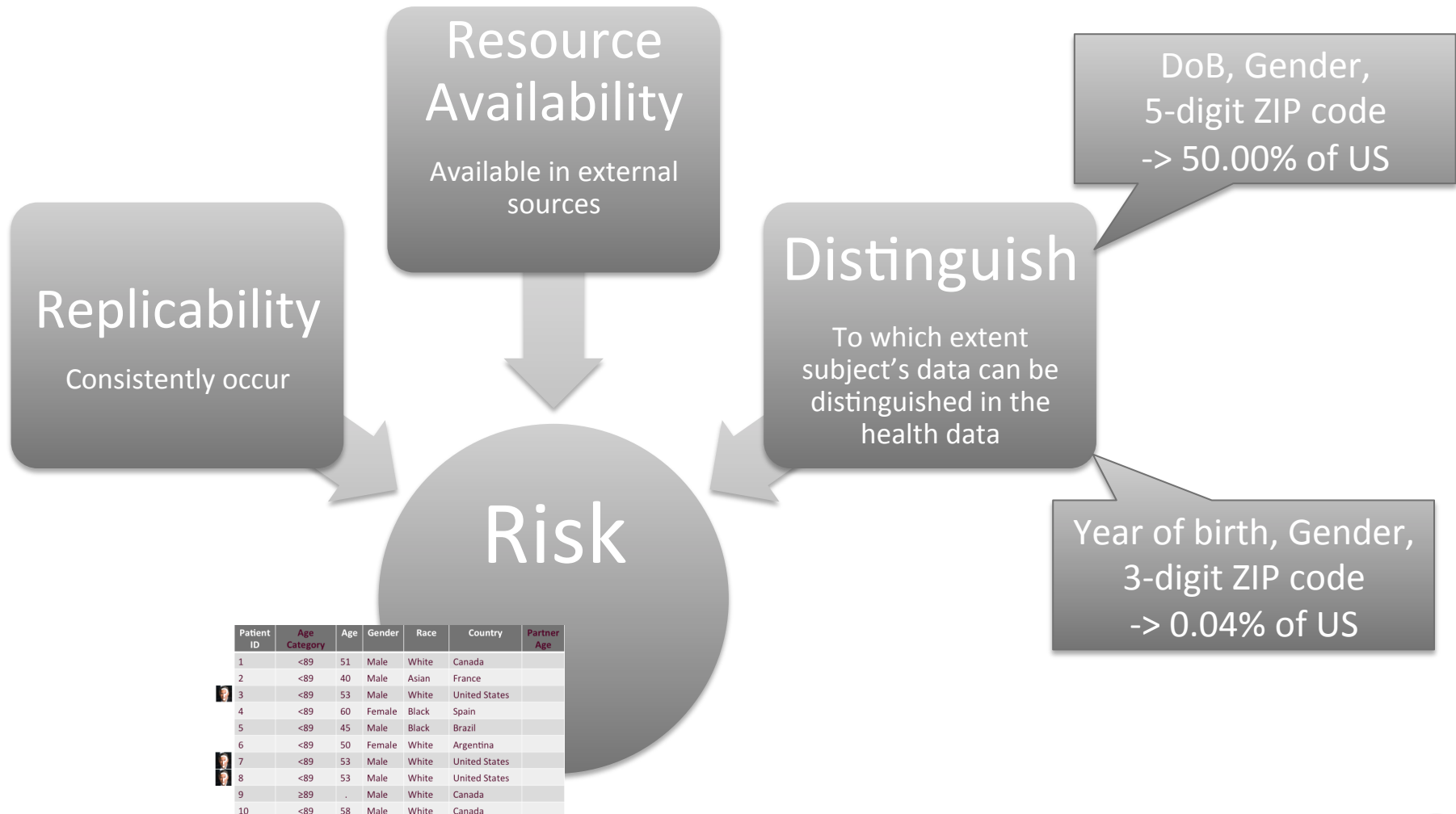
Who is Georges?



	Patient ID	DoB	Age	Gender	Race	Country	Partner Age
	1						
	2						
	3						
	4						
	5						
	6						
	7						
	8						
	9						
	10						



Approach to Risk



Approach to Risk



	Direct	Quasi Level 1	Quasi Level 2	Quasi Level 3
Type:	A Combination Uniquely Identify	Demographics	Longitudinal Events & Findings	Longitudinal Sensitive Information
Examples:	• Subject ID • DoB • Death Date • (Address) • (Name)	• Age • Country • Race • Sex • Ethnicity	• Lab • Outcome • Adverse Event • Medications • Medical History	• Abortions • Drug abuse • Mental/Venereal Diseases
Replicability:	High	High	Low	Low
Resource Availability:	High	High	Low	Low
Distinguish:	High	Medium	High	Medium/High

Equivalence Classes



Patients having same characteristics for important quasi identifiers

Size 3:
33.3%



Patient ID	Age Category	Age	Gender	Race	Country	Partner Age
1	<89	51	Male	White	Canada	
2	<89	40	Male	Asian	France	
3	<89	53	Male	White	United States	
4	<89	60	Female	Black	Spain	
5	<89	45	Male	Black	Brazil	
6	<89	50	Female	White	Argentina	
7	<89	53	Male	White	United States	
8	<89	53	Male	White	United States	
9	≥89	.	Male	White	Canada	
10	<89	58	Male	White	Canada	

Equivalence Classes



Patients having same characteristics for important quasi identifiers

Size 5:
20.0%

	Patient ID	Age Category 2	Age	Gender	Race	Continent	Partner Age
	1	50-59		Male	White	North America	
	2	40-49		Male	Asian	Europe	
	3	50-59		Male	White	North America	
	4	60-69		Female	Black	Europe	
	5	40-49		Male	Black	South America	
	6	50-59		Female	White	South America	
	7	50-59		Male	White	North America	
	8	50-59		Male	White	North America	
	9	≥89		Male	White	North America	
	10	50-59		Male	White	North America	



Equivalence Classes



Patients having same characteristics for important quasi identifiers

Size 1:
100.0%

Patient ID	DoB	Age	Gender	Race	Country	Partner Age
1	12APR1963	51	Male	White	Canada	48
2	28MAY1974	40	Male	Asian	France	41
3	06MAY1961	53	Male	White	United States	36
4	28MAY1954	60	Female	Black	Spain	65
5	14JUL1969	45	Male	Black	Brazil	41
6	13AUG1964	50	Female	White	Argentina	45
7	18MAR1961	53	Male	White	United States	48
8	22JAN1961	53	Male	White	United States	37
9	27SEP1924	90	Male	White	Canada	73
10	07FEB1956	58	Male	White	Canada	62



Simple Risk Measures

$$\textit{Average}_i \left(\frac{1}{\textit{Size}(\textit{EquivalenceClass}[i])} \right)$$

$$\textit{Max}_i \left(\frac{1}{\textit{Size}(\textit{EquivalenceClass}[i])} \right)$$

Hrynaskiewicz et al., BMJ 2010: Less than 3 quasi identifiers

Data Utility

Proactive Outside a Request

Use Company/Industry
Guidelines

Compare to SAP

Good common sense...

Reactive Based on a Request

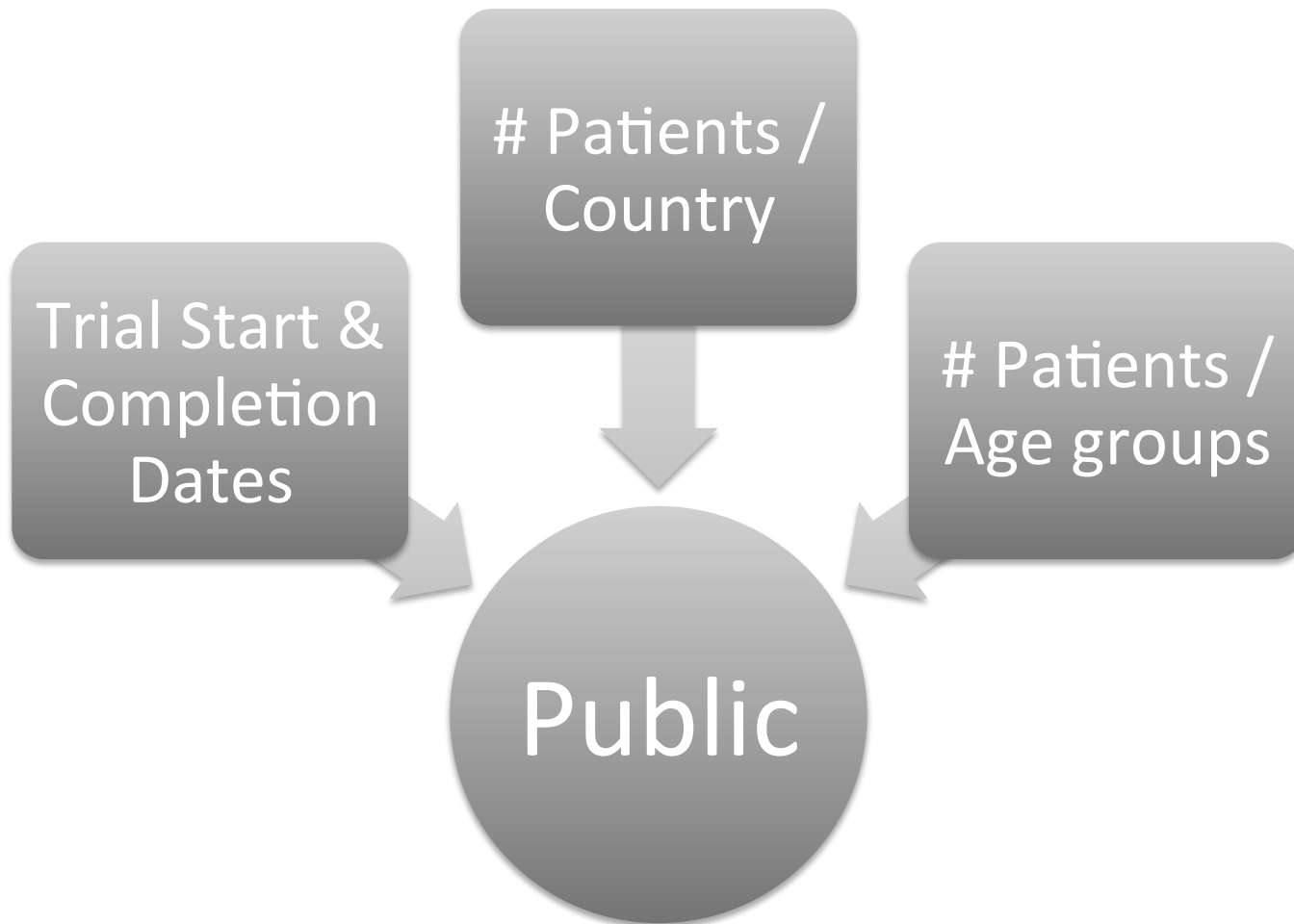
Use Company/Industry
Guidelines

Focus on what is needed

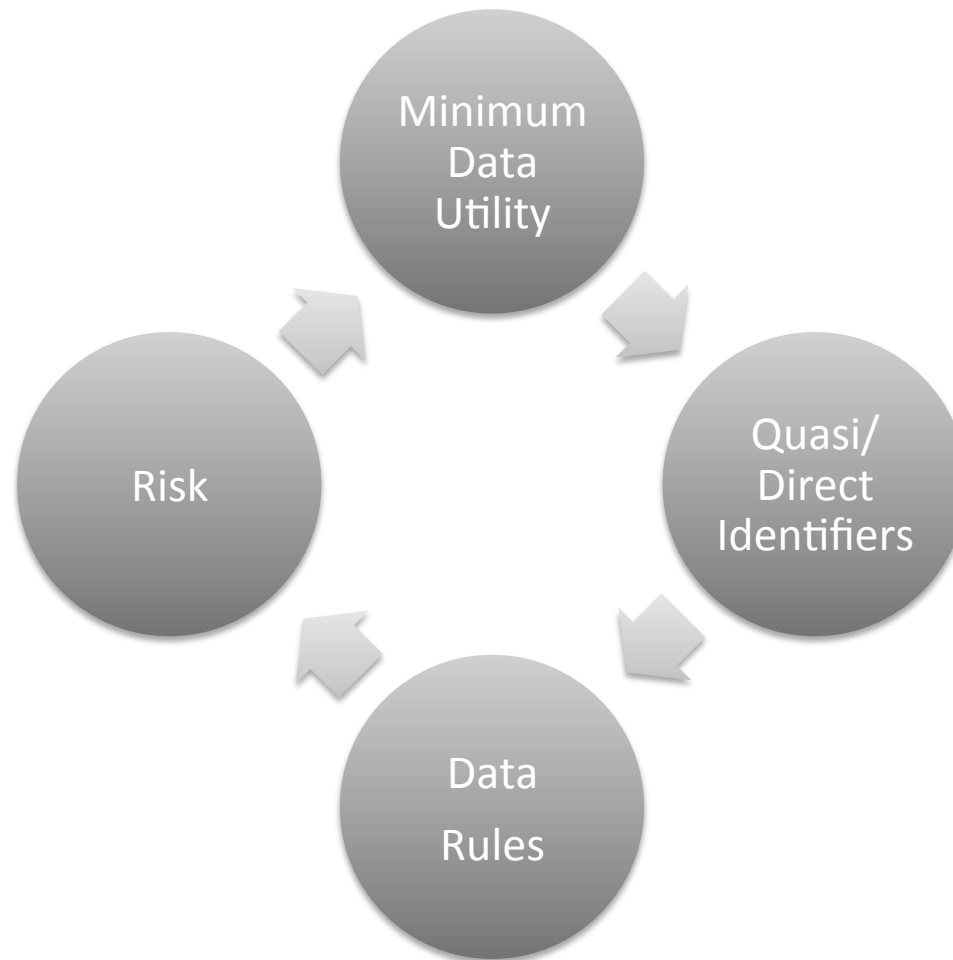
Negotiate with researcher

Public Information to Consider

from clinicaltrials.gov and (soon) [eudraCT](https://eudraCT.europa.eu/)



Approach to Data De-Identification



Our role within Data De-Identification

Programmer

Different Data Models

Data DI Plan

Program & re-use macros

Work with Metadata

Validate

Document Data De-Identification

Data Scientist

Find the data!

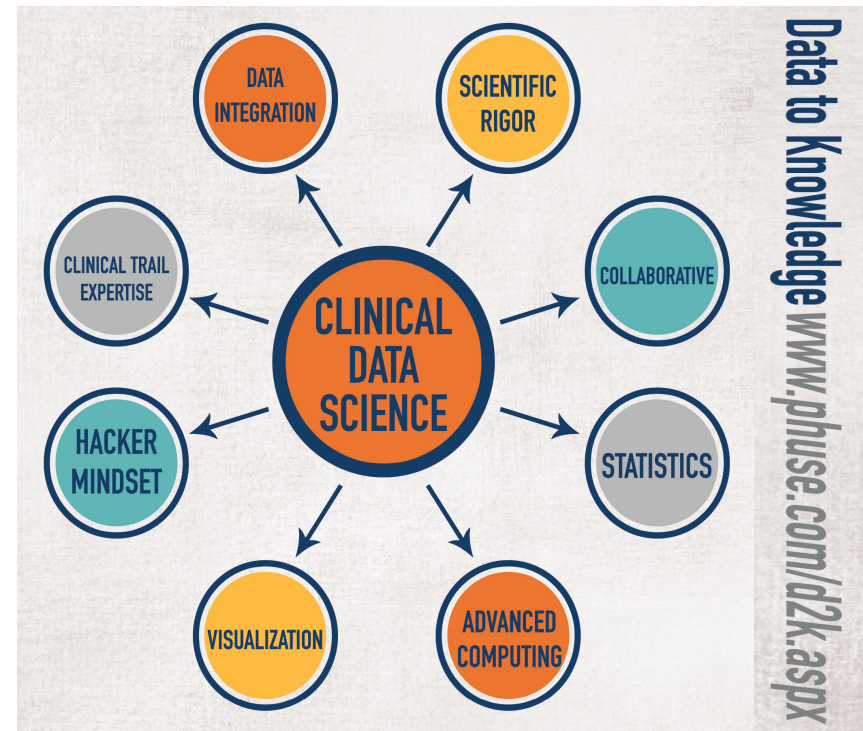
Hack the data!

Understand data privacy and utility!

Pick people brain!

Consider changing guidelines!

You make the rules!





PhUSE De-Identification Working Group Vision

“Develop data de-identification standards for CDISC data models”

20+
Participants
from Pharma,
CROs,
Software and
Academia

Focus first on
SDTM

Data Privacy
Rules &
Rational
Data Utility

Thanks!



QUALIANCE
Trusted Partner since 2010
5 Years

Jean-Marc Ferran

Consultant & Owner, Qualiance ApS



dk.linkedin.com/in/jeanmarcferran/



@QualianceTwitta

References

- [1] **Preparing raw clinical data for publication guidance for journal editors, authors, and peer reviewers**, Hrynaszkiwicz I, Norton M L, et al. - British Medical Journal 2010; 340:304–307
- [2] **Evaluating the Risk of Re-identification of Patients from Hospital Prescription Records**, Khaled El Emam et al. - CJHP – Vol. 62, No. 4 – July–August 2009
- [3] **A De-identification Strategy Used for Sharing One Data Provider's Oncology Trials Data through the Project Data Sphere Repository**, Malin, 2013