Paper AR02

Python to automate an AR project identifying discrepancies across fields from a gov. open-source dataset of covid-19 vaccine registrations.

Aldir MEDEIROS FILHO, Independent, London, UK

ABSTRACT

This paper presents a Clinical Research Operations' case study & proof of concept[i] to assess if open-source software programming language and open-source datasets could be used as clinical operations managerial tools to deal with longitudinal 'big data' during clinical trials life cycles. *Shares learning process and challenges of using open-source triad Google Collab's IDE & Pandas & Python, to conceive a Clinical Risk Based Monitoring project to identify, track AND resolve intra & inter discrepancies across government opensource datasets documenting distribution & application of vaccines against covid-19 covering a population of 7 million across 295 counties.*

Introduction

Clinical Operations (ClinOps) is part of the constellation of pharma & biotech R&D departments working on clinical drug programme development. It supports the clinical trial process from feasibility to close out. ClinOps is tasked with the planning, implementation, management, and execution of the clinical trial process.

For the moment, ClinOps professionals are not usually part of the PHUSE crowd. Possibly because, even for the big pharma companies, Phase III studies are very rare and very occasionally involves more than a few thousands of patients, so not really 'big data' for many.

During my 25 years career as ClinOps manager the largest clinical trial I dealt with enrolled 3200 patients very unevenly distributed across 180 sites around de world (26 countries, again unevenly spread across 4 continents). The primary endpoint was survival status at 12 months. The overall clinical trial study lasted for about 5 years.

Microsoft Excel always fulfilled all my clinical operations managerial needs. Crossing key data points from different sources helped to set up meta-spreadsheets. Also, to develop my own risk-based remote monitoring excel tools that allowed me to be, most of the time, ahead of the local teams in the field around the world.

However, as digital healthcare becomes the norm worldwide, that brings potential changes to the amount of data possibly available for clinical trials and how often we can have access to it. 'RWD', 'RWE', 'Big Data', 'Data Sciences', 'synthetic study arms', PPRL, etc, are concepts that ClinOps will need to learn, integrate, and deal with. Therefore, Excel, the tool of excellence during the paper paradigm era of clinical trials will need some help in the digital era of big data. That is the purpose of this personal case study.

Here after I share how I combined my ClinOps Project Manager mindset – referred simply as ClinOps from now on - with the concomitant learning from Python for data sciences to create a monthly iteration of my clinical data sciences project named 'Combinations of Concern – CoC' using the CRBM framework.

The CoC project was created using exclusively official government opensource datasets related to registrations of vaccination against covid-19 from the Brazilian state of Santa Catarina (SC) located in the south of Brazil. SC has an estimated population of around 7 million very unevenly distributed across 295 counties.

First, I will summarise how the CoC project concept took shape and evolved by applying key ClinOps premises that guided the project design and tools selection.

		ClinOps premises	Tools
1.1	RISK ID	IN & OUT, CoC, Compliance	CRBM
1.2	MONITORING TOOLS	Sustainability and auditability of the data	Google Collab IDE & Pandas & Python
1.3	DELIVER	Know your customer needs & constraints.	Excel

Then, I briefly share main findings of interest for the PHUSE community from actual 4 parts of the CoC project.

Followed by a warning about the 'Swiss cheesberg' of digital healthcare data during and post-pandemic era.

Finally, I will present key learnings from this project that early November 2022 was on its 18th monthly iteration.

Part I – CoC Project Conception

1.1 Clinical RBM set up

Clinical Risk Based Monitoring (CRBM) is a technique for remote monitoring of clinical trials I was already using for my ClinOps duties long before Transcelerate formalized the RBM concept back in 2012[ii].

From my point of view, simply speaking, CRBM is a process of minimization. It is the identification and selection of a minimal set of variables that could suggest & reveal an issue or point of interest within a specified context. A similar process to the way doctors will look for symptoms and syndromes to pave the way for differential diagnostics algorithms and propose treatments according to resources available.

RISK(S) IDENTIFICATION - Think globally act locally

Back in the second half of 2020 I realised that the mass vaccinations against covid-19 would be a real GLOBAL BIG REAL-WORLD DATA clinical trial. I grabbed the opportunity to put in practice the skills gathered during my journey towards data sciences started a few years before.

In a think globally, act locally mindset, I realized the region I was at the time, the south Brazilian state of Santa Catarina[iii], offered a very interesting 'case study' opportunity. With an estimated population of around 7 million very unevenly distributed across 295 counties offered a nice scope for a simulated study case of 'RWD big data' clinical trial, that fit my limited material hardware resources, a simple laptop and internet with reasonable speed.

I geared up my ClinOps' feasibility mindset, for the first and crucially important phase of a clinical trial lifecycle.

Back end of 2020, the most important challenge ahead was the immense 'logistical diversity' of all sorts that we could expect from around half a dozen potential vaccines still to be approved. It is important to remember that due to context, all those vaccines would be first authorised under a so-called 'conditional approval' scheme.

For ClinOps that means, 'safety study', therefore the need for efficient operational 'pharmacovigilance' (PV) monitoring tools was already very high on the scope of my embryonic RBM strategy.

Still back end 2020, due to global, national, and local constraints of all sorts, it became clear the conditional approvals and their arrival at the vaccination points of deliver was to be VERY UNPREDICATABLE. There was also very high incertitude about regular supply chains.

To remediate all sort of incertitude, it was already envisaged at the time that mixed up of vaccines from different suppliers and different AND ADAPTATIVE regimen timelines between doses was a very possible strategy to face the covid-19 pandemic crisis.

TASK AT HAND:

As a ClinOps I was already very familiar with the digital limitations of public healthcare worldwide.

To help remediate such limitations, for me the task at hand should be the set-up of reliable track tools enabling longitudinal documentation of vaccines supplied & applied to support compliance of specific regimen timelines from very different vaccines for each of the estimated 7 million individuals living in the 295 SC counties in January 2021.

So, I started the design of my Risk Based Monitoring plan around three main aspects:

- Accountability – 'IN & OUT' and
- Correct documentation of vaccines applied
- Compliance – adherence to mixed vaccine regimens.

ACCOUNTABILITY: IN (vaccines & doses distribution) & OUT (vaccines & doses applied)

Due to ICH & GCP and Regulatory Authorities (RA) requirements, ClinOps people must be able to account at any time of the clinical trial lifecycle for EVERY SINGLE tablet, vial, etc of the study drug. Usually, the first document asked aby RA auditors are about study drug accountability during the study lifecycle. Study drug accountability usually is amongst the most time consuming of the ClinOps tasks.

My very ambitious CRBM design intended to enable accountability for each vaccine dose supplied and applied to each individual in the State of SC. That was a significant challenge because the untied Brazilian 3-layer system for the supply chain of vaccines and registrations recording of vaccination: Federal, State and County. Briefly:

LAYER		Datasets of interest:
Federal	Brazil is composed of 27 states.[iv] The federal government was responsible to approve, buy and distribute the vaccines to the capital of each state. The first federal vaccination strategy versions planned to prioritize the distribution and application by age ranges from the oldest to youngest. Initially, also for the health care professionals and essential services. Due to the scarcity of vaccine supply chain for the first months, and to ensure fair distribution across all 27 states, the federal gov. dispatched vaccines to each state based on stepwise proportion of each population range scheduled to be vaccinated. For example, the first shipment was done to all 27 states to deliver Dose 1 vaccine for 10% of the oldest above 90-year-old. The second shipment, days later, would be for the next 10 % of this age range, and so on. These % changed all the time depending on vaccine supply by manufacturers.	For my project I used 4 federal government opensource datasets: First IN layer: Brazilian Ministry of Health (BMoH) 'Technical Notes'': Pdf doc containing a table with the total nr of doses supplied to each Brazilian state at each federal dispatch. Usually twice a week during the first year. Second OUT layer: OpenDataSus (OPDS) csv docs, per Brazilian state, centralizing all the individual registers of applied vaccines for all Brazilian states. Update once a day. The OPDS became the core dataset I used to build my CoC project. Third OUT layer: LocalizaSUS 'vacinometro'. Central list of doses applied per county. Third 'IN' layer: LocalizaSUS 'distribution' list with only the total of confirmed supplied vaccine doses per main provider per county. Both 'LocalizaSUS' websites, controlled by the BMoH, came months after vaccination started, are updated very irregularly, had several different versions, were out during the harkening of MoH, and relatively updated only months after this event. **UNTIL END Oct 22, LOCALIZASUS SITES DOES NOT SHOW ANY SPECIFIC DATA ABOUT PEDIATRIC POPULATION.**
State	Each Brazilian government state had the responsibility to ensure the fair distribution of the vaccines to each local county. They followed the same stepwise proportion of each population range scheduled to be vaccinated A particular challenge at early 2021 was that due covid-19 as well political and administrative issues, each state used different administrative bodies, public or private, to ensure vaccine distribution. For SC, DIVE (Diretoria de Vigilancia Epidemiologica – Epidemiological Vigilance Center) was selected to ensure the central reception and distribution of vaccines. This department was brilliantly led by an exceptionally efficient team of nurses. Despite the stormiest political conditions, they steadily issued, weekly excel spreadsheets detailing the vaccines delivered to each of the 295 counties per population dose, age range & category.	Second IN layer DIVE used its own regional mapping of counties to set up and ensure the vaccine distribution. It is their 17 UDVE regions that I used to structure my RBM project. First OUT layer: 'SC Vacinometro': Only after several months the vaccination plan was going on, the official local state government health department set up a PowerBi 'SC Vacinometro'. Important facts about this document. • none of the 15 pages showed an IN and OUT accountability of the vaccines. • Some regions have different borders and names than DIVE's map. That make very challenge to track IN and OUT. This document was update in a 'lively' basis. • It did not use the same terminology than the federal or DIVE sources to label the doses. Fourth OUT Layer: DIVE's monthly epidemiological reports. Most reliable. Often with very different nrs from the above.
County	There was about 8 different digital platforms & mask available across the state of SC. Some public some private. Some appeared & disappeared out of the blue. The nr of layers between the point of vaccine deliver and governmental central database varied from one to the other.	Most of 295 counties did not have access to internet at improvised (due to covid_19 safety constraints) point of vaccine deliver. So, each registration needed to be entered on two steps, what caused delays for the update of the central databases. 166 out of the 295 counties have a population of less than 10.000.

Each government administration layer uses different systems, masks, mapping of administrative regions and terminology to refer to the distribution supply chain and vaccines dose identification. So, trying to get a holistic IN and OUT accounting system of vaccines received and applied ON REAL TIME was not possible.

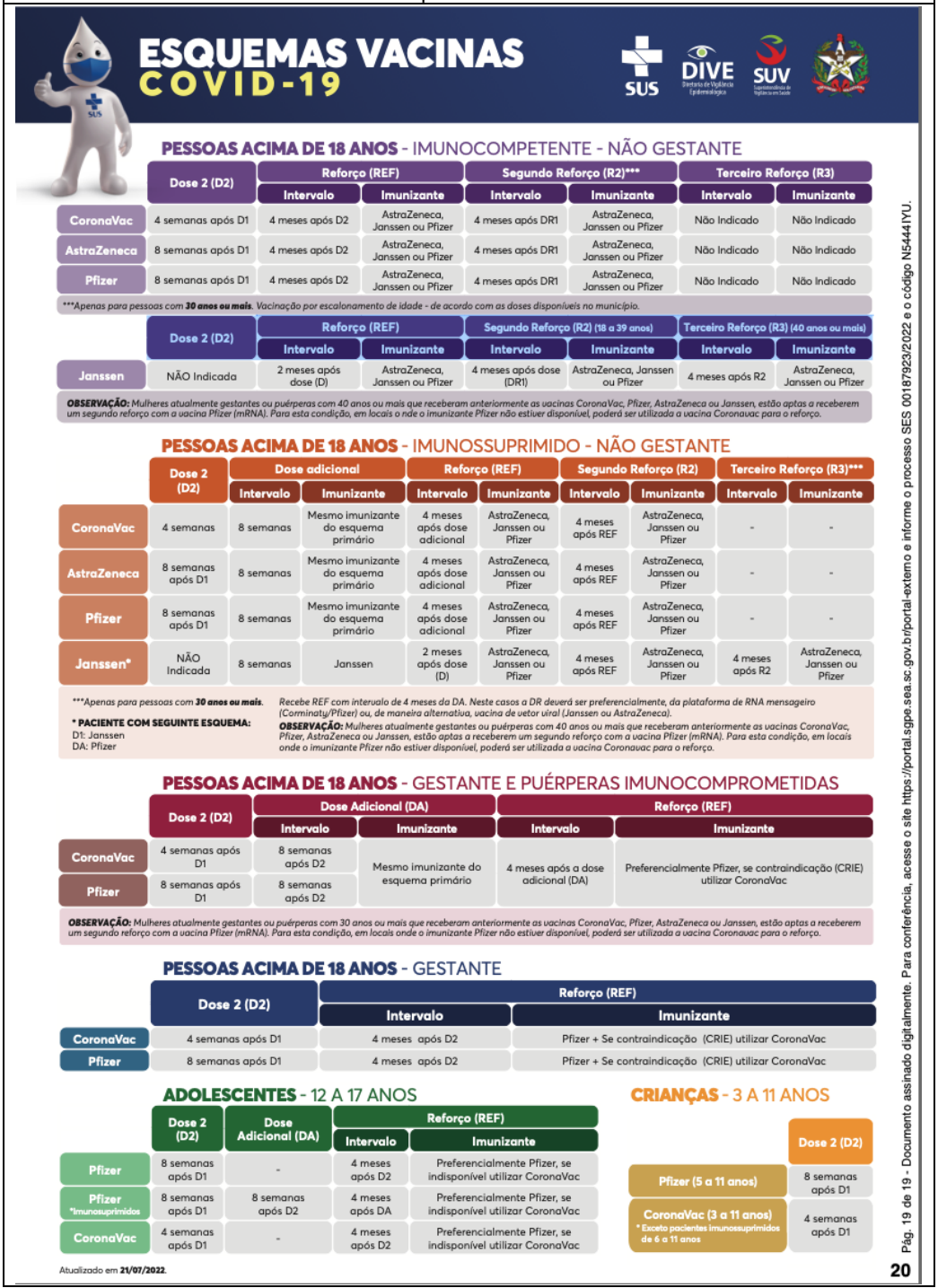
Therefore, at the start of every month I downloaded and archived original datasets used on my project. To enable comparison across the sources mentioned above, I also took screen-picture of official website relevant pages. So that could understand different terminology and numbers across sources and track progress.

After several trials and errors, and steep learn of local processes and documents, the framework for IN and OUT accountability was consolidated in the first notebook.

This IN and OUT accountability became more urgent from the start of the vaccination plan for many reasons.

- The 'wild' 'extension' on the nr of doses obtained per vial that challenged the IN and OUT account. More details in 1.3 here after.
- the frequent change in vaccine intervals and possible permutations between vaccine regimen and population category. Table here after shows the most recent version provided by DIVE. That changed many times since January 2021 to help adapt to vaccines supplies chain availability.

DIVE Nota Tecnica No 0057/2022 – 26 September 2022



ESQUEMAS VACINAS COVID-19

SUS · DIVE · SUV

PESSOAS ACIMA DE 18 ANOS - IMUNOCOMPETENTE - NÃO GESTANTE

	Dose 2 (D2)	Reforço (REF)		Segundo Reforço (R2)***		Terceiro Reforço (R3)	
		Intervalo	Imunizante	Intervalo	Imunizante	Intervalo	Imunizante
CoronaVac	4 semanas após D1	4 meses após D2	AstraZeneca, Janssen ou Pfizer	4 meses após DR1	AstraZeneca, Janssen ou Pfizer	Não Indicado	Não Indicado
AstraZeneca	8 semanas após D1	4 meses após D2	AstraZeneca, Janssen ou Pfizer	4 meses após DR1	AstraZeneca, Janssen ou Pfizer	Não Indicado	Não Indicado
Pfizer	8 semanas após D1	4 meses após D2	AstraZeneca, Janssen ou Pfizer	4 meses após DR1	AstraZeneca, Janssen ou Pfizer	Não Indicado	Não Indicado

***Apenas para pessoas com **30 anos ou mais**. Vacinação por escalonamento de idade - de acordo com as doses disponíveis no município.

	Dose 2 (D2)	Reforço (REF)		Segundo Reforço (R2) (18 a 39 anos)		Terceiro Reforço (R3) (40 anos ou mais)	
		Intervalo	Imunizante	Intervalo	Imunizante	Intervalo	Imunizante
Janssen	NÃO Indicada	2 meses após dose (D)	AstraZeneca, Janssen ou Pfizer	4 meses após dose (DR1)	AstraZeneca, Janssen ou Pfizer	4 meses após R2	AstraZeneca, Janssen ou Pfizer

OBSERVAÇÃO: Mulheres atualmente gestantes ou puérperas com 40 anos ou mais que receberam anteriormente as vacinas CoronaVac, Pfizer, AstraZeneca ou Janssen, estão aptas a receberem um segundo reforço com a vacina Pfizer (mRNA). Para esta condição, em locais o nde o imunizante Pfizer não estiver disponível, poderá ser utilizada a vacina Coronavac para o reforço.

PESSOAS ACIMA DE 18 ANOS - IMUNOSSUPRIMIDO - NÃO GESTANTE

	Dose 2 (D2)	Dose adicional		Reforço (REF)		Segundo Reforço (R2)		Terceiro Reforço (R3)***	
		Intervalo	Imunizante	Intervalo	Imunizante	Intervalo	Imunizante	Intervalo	Imunizante
CoronaVac	4 semanas	8 semanas	Mesmo imunizante do esquema primário	4 meses após dose adicional	AstraZeneca, Janssen ou Pfizer	4 meses após REF	AstraZeneca, Janssen ou Pfizer	-	-
AstraZeneca	8 semanas após D1	8 semanas	Mesmo imunizante do esquema primário	4 meses após dose adicional	AstraZeneca, Janssen ou Pfizer	4 meses após REF	AstraZeneca, Janssen ou Pfizer	-	-
Pfizer	8 semanas após D1	8 semanas	Mesmo imunizante do esquema primário	4 meses após dose adicional	AstraZeneca, Janssen ou Pfizer	4 meses após REF	AstraZeneca, Janssen ou Pfizer	-	-
Janssen*	NÃO Indicada	8 semanas	Janssen	2 meses após dose (D)	AstraZeneca, Janssen ou Pfizer	4 meses após REF	AstraZeneca, Janssen ou Pfizer	4 meses após R2	AstraZeneca, Janssen ou Pfizer

***Apenas para pessoas com **30 anos ou mais**.

* **PACIENTE COM SEGUINTE ESQUEMA:**
D1: Janssen
DA: Pfizer

Recebe REF com intervalo de 4 meses da DA. Neste casos a DR deverá ser preferencialmente, da plataforma de RNA mensageiro (Corminaty/Pfizer) ou, de maneira alternativa, vacina de vetor viral (Janssen ou AstraZeneca).

OBSERVAÇÃO: Mulheres atualmente gestantes ou puérperas com 40 anos ou mais que receberam anteriormente as vacinas CoronaVac, Pfizer, AstraZeneca ou Janssen, estão aptas a receberem um segundo reforço com a vacina Pfizer (mRNA). Para esta condição, em locais onde o imunizante Pfizer não estiver disponível, poderá ser utilizada a vacina Coronavac para o reforço.

PESSOAS ACIMA DE 18 ANOS - GESTANTE E PUÉRPERAS IMUNOCOMPROMETIDAS

	Dose 2 (D2)	Dose Adicional (DA)		Reforço (REF)	
		Intervalo	Imunizante	Intervalo	Imunizante
CoronaVac	4 semanas após D1	8 semanas após D2	Mesmo imunizante do esquema primário	4 meses após a dose adicional (DA)	Preferencialmente Pfizer, se contraindicação (CRIE) utilizar CoronaVac
Pfizer	8 semanas após D1	8 semanas após D2			

OBSERVAÇÃO: Mulheres atualmente gestantes ou puérperas com 30 anos ou mais que receberam anteriormente as vacinas CoronaVac, Pfizer, AstraZeneca ou Janssen, estão aptas a receberem um segundo reforço com a vacina Pfizer (mRNA). Para esta condição, em locais onde o imunizante Pfizer não estiver disponível, poderá ser utilizada a vacina Coronavac para o reforço.

PESSOAS ACIMA DE 18 ANOS - GESTANTE

	Dose 2 (D2)	Reforço (REF)	
		Intervalo	Imunizante
CoronaVac	4 semanas após D1	4 meses após D2	Pfizer + Se contraindicação (CRIE) utilizar CoronaVac
Pfizer	8 semanas após D1	4 meses após D2	Pfizer + Se contraindicação (CRIE) utilizar CoronaVac

ADOLESCENTES - 12 A 17 ANOS

	Dose 2 (D2)	Dose Adicional (DA)	Reforço (REF)	
			Intervalo	Imunizante
Pfizer	8 semanas após D1	-	4 meses após D2	Preferencialmente Pfizer, se indisponível utilizar CoronaVac
Pfizer *Imunossuprimidos	8 semanas após D1	8 semanas após D2	4 meses após DA	Preferencialmente Pfizer, se indisponível utilizar CoronaVac
CoronaVac	4 semanas após D1	-	4 meses após D2	Preferencialmente Pfizer, se indisponível utilizar CoronaVac

CRIANÇAS - 3 A 11 ANOS

	Dose 2 (D2)
Pfizer (5 a 11 anos)	8 semanas após D1
CoronaVac (3 a 11 anos) * Exceto pacientes imunossuprimidos de 6 a 11 anos	4 semanas após D1

Atualizado em **21/07/2022.**

20

Pág. 19 de 19 - Documento assinado digitalmente. Para conferência, acesse o site https://portal.sgpo.sea.sc.gov.br/portal-externo e informe o processo SES 00187923/2022 e o código N5444IYU.

Once the framework to help account for the IN and OUT vaccinations per dose was set up on the notebook 01 (and later extended on ntbk_04), I was then able to address the CoC issue on the project notebook nr 02.

CoC – Combination of Concern

The OpendDataSus's data dictionary revealed that out of 34 fields, five were reserved to document 'the vaccine' applied to an individual. This ClinOps knew by experience that this kind of dataset granularity was a challenge.

Column nr - c	Description in English
25	Vaccine batch nr
26	Vaccine manufacturer name
27	Vaccine provider's official registration nr with BR gov.
30	Specific vaccine provider code
31	Vaccine's name

Such concern became real after reviewing the first 3 months of the vaccination registrations from SC, I noticed a month-to-month trend of about 12 % of all the registrations of what I named 'Combinations of Concern – CoC'. In other words, when conflicting data within a specific register across the 5 mentioned fields, impedes to confirm which vaccine an individual really received at a specific moment.

The clearest and most courant CoC example from the first 3 months was:

c26_manufacturer's name	c30_code and/or c31_name
Serum AstraZeneca covid-19 vaccine (From India)	Sinovac/Coronavac (From China).

Vaccines from different manufacturers / providers also had different regimens for interval between Doses 1 and 2. For the first months of the Brazilian's vaccination plan that was 4 weeks for SINOVAC and 8 weeks for AZ.

So, a high proportion of CoC registers could have negative impact on pharmacovigilance process and planning on the nr of shots need per county for the next dose.

I decided then to use CoC criteria as the first pointer for risk assessment for my CRBM plan. What it reveals & suggests and how can we address it?

Once the framework to secure IN & Out accountability was done on notebook 01 and the design to monitor the CoC consolidated on notebook 02, I extended the now named 'CoC project', to explore compliance matters.

COMPLIANCE: participant adherence to treatment

Participant adherence to treatment (compliance) is an essential part of ClinOps work for every clinical trial.

Unfortunately support from many official government quarters to vaccination against covid-19 was not unanimous in Brazil, particularly in the south region of Brazil. Fake news and awkward statements from government officials about vaccines was demotivating large sector of the population to completely adhere to the vaccination plan.

When we add to this, different vaccine regimens intervals that changed according to supply, unofficial extension of nr of doses per vial, and the unpredictability of vaccines deliver already suggesting using different vaccines to complete any individual vaccination plan, the issue of closely assess compliance became urgent. Particularly after the need for boost third, fourth,… vaccinations doses became clear.

So, after I settled out the CoC track strategy around Feb 2022, I explored a few other risk pointers that could provide clarity about regimen compliance adherence. That resulted on the parts II to IV with 8 new notebooks

1.2 Sustainability and auditability of the data

But before we move on let's return to end May 2021, to explain why and how Python got in the picture.

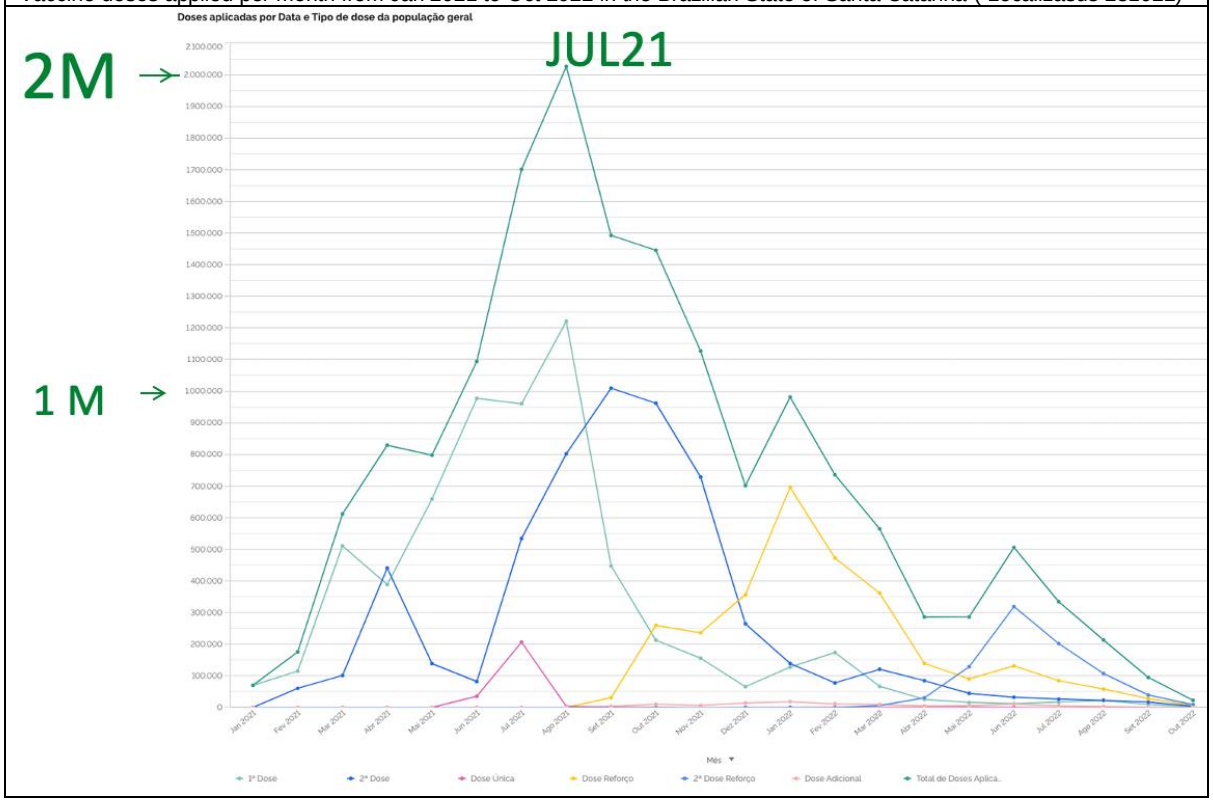
By May 2021 we already had many vaccine 'configurations' from each of the two vaccine providers, 'Aztra&Zeneca" and 'Sinovac'. We knew by then that at least two new providers (Pfizer and Janssen), with totally different regimens and transport & storage constraints were about to be approved by the Brazilian's ANVISA.

The expected increase in complexity and scale of the CoC issue and high-risk impact on compliance motivated me to find appropriate tools to track and address this problem by moving to the next phase of this project.

Register means all the 32 fields containing data related to the application of one individual vaccination shot. Up to April 2021, the number of vaccination registers from SC was under one million lines.

Microsoft Excel was sufficient up to this point to help me explore and become acquainted with the local characteristics (i.e., UDVES, population distribution, geography, etc). However, the situation changed drastically from early May 2021 when one million register cap was quickly passed.

Vaccine doses applied per month from Jan 2021 to Oct 2022 in the Brazilian State of Santa Catarina (Localizasus 231022)



That initiated my search for tools that would help systematize the data review and face the significant speed and scale of the dataset's growth. Initially, I tried with the two computational languages and data sciences techniques I spent learning over the previous 3 years. None was helpful:

- SAS language due to the limitations imposed by their property business model.
- The open-source R language and its R Studio IDE were also not flexible enough to help me deal with the constant increase in registers numbers and changing scope of the data.

In my urgency to progress the CoC project, after reading Nicolas Dupuis [v] paper for the PHUSE EU Connect 2019 – 'SM01 R or Python? Dilemma, Dilemma…' I decided to start with Python from zero.

My learning Python journey started with a 3 months bootcamp, 'Introduction to Python for Applied Data Sciences in Healthcare'. I also invested on Python and Python for Data Sciences online courses from 3 different institutions for the next 10 months. Completing the bootcamp, full program from the 3 institutions, attending a PyData online conference, a few books plus a lot of time spent with blogs like 'Towards Data Sciences', 'Medium Daily Digest', 'Stackoverflow', etc, to help shape my CoC project quickly amounted for over 2000 hours full-time basis self-learning experience. An entertaining way to deal with Covid lockdown restrictions.

For this beginner in programming and DS, one of the most challenging aspects of learning Python online is the wild permutations of IDEs, Python libraries and versions used by TOO MANY instructors. None of the courses I attended had consistent didactic program education policy. It is a kaleidoscope of fragmented possibilities that easily lead to confusion. Such fragmentation stalled the progress of my CoC project many times.

Another challenge is that, as for SAS and R, Python courses have a very careless attitude towards consistent terminology. These terminological cacophonies are rather disorientating for an international ClinOps that know the risks, and therefore, how important a clear project terminology policy is for the success of clinical trials, international or not.

Fortunately, my ClinOps' mindset helped me get out of the didactic chaos from the Python courses and have a breakthrough decision that significantly accelerated my 'programming' and 'DS' skills and finally allowed a significant progress of the CoC project.

ClinOps's MINDSET: SUSTAINABILITY = <u>Auditable, Traceable = KEEP IT SIMPLE</u>

The conception, design and continuous adaptative development of the Python CoC project was always done with ClinOps's ICH & GCP & ALCOA+[vi] SUSTAINABILITY PRINCIPLES in mind.

ClinOps people are trained to ensure that any clinical trial related document & data point must be auditable, traceable at any time, from conception to up to 25 years, as required by regulatory authorities.

So, throughout the interactive roller coaster processes of learning Python for DS AND developing my CoC project to cope with fast evolving source datasets I was always focused on trying to ensure that my notebooks' sequency and code strategy and its derived datasets were following ALCOA+ principles.

The most pragmatical and breakthrough decision derived from this ClinOps & ALCOA+ mindset was to simplify LONG TERM maintenance by restricting to the minimalist triad IDE Google's Collab & Pandas & Python.

In other words, keep the 'complications' (from horology) to a minimum by intentionally avoiding, IDEs hard to navigate & maintain, abstract codes, 'fancy' visualization packages, hard to maintain webpages interfaces, etc.

Hopping that easy to maintain & adapt make the overall project time resilient & reproducible.

1.3 KNOW YOUR CUSTOMER NEEDS AND CONSTRAINTS.

My CRBM project's aims to be useful and efficient to enable remote monitoring and improve scientific quality by reducing entropy or uncertainty of the data, while it's being produced, collected, and used. Identify and facilitate corrective actions on an ongoing basis where and when necessary.

The blood clot risk raised during the first 2021 months of vaccination in Europe [vii] reminded the importance of ALWAYS have a clear picture who received what and when in case a large-scale safety response was required.

In Brazil, that was a challenge because we had the issue "of one brand, delivering 'different' vaccines". For example, the AZ brand had 5 different vaccine configurations circulating. Manufactured in different countries and/or delivered in different containers (i.e., 1 ready-to-use syringe, or vial with 5 or 10 doses). See slide presentation for examples.

Additionally, from the first months of vaccination in SC I started to notice, particularly at certain counties, a worrying imbalance in the IN and OUT rate. After investigation it was identified that due to the scarcity of vaccine, some nurses, first un-officially and then somewhat officially, started 'to extend' the nr of doses per vial. For a vial with officially 5 doses, I could see some vaccination counties, delivering 6 to 8 doses. For a vial with officially 10 doses, deliver up to 15 doses. That without much proper consideration how that surplus of dose vaccination would affect the supply for the next dose.

As all those vaccines were delivered in small batch quantities to over a thousand points of vaccine deliver across SC that used a plethora of digital platforms with masks unfit to proper register the complexity of the vaccine manufacturing sources. Therefore, it was important to have meta tools that could be feedbacked to local professionals at the field to identify, track who received what and when.

Despite a scaring top-down (Federal government to county-point-of-deliver) chronic cross-institutional crisis of roles and responsibilities in the chain of command for the PV surveillance of the covid vaccines in SC and Brazil, as ClinOps, I had a clear idea who were my customer, their needs and, and most important, their constraints. That helped to think pragmatically in the best way I could design my CoC project for efficient deliver.

In SC, as in most of the health public services worldwide, Microsoft or Google Excel is the format that can be mostly easily understood & used at the point of care. One should be aware that most often those point of healthcare deliver have access only to, unstable internet, old Microsoft Excel version, very occasionally printers.

So, my notebooks were conceived in ways that, when a risk is identified, I can deliver tailored reports in SIMPLE and LIGHT excel format files (no fancy graphs, colours, fonts, etc) to each nurse in charge at the vaccination point of deliver in SC (~ 1200 during the first months of vaccination).

And most importantly, be able to easily monitor the progress of issues' resolution.

Part II – Moving on to more technical aspects and key findings

Historic of Google Collab IDE & Pandas & Python CoC monthly iterations								
Original iteration nr	simplified CoC monthly iteration nr	Dataset Freeze	OPDS original raw file size (GB) on hard drive	TOTAL rows (registrations)	added registrations since last CoC	Catch up' from previous months before last CoC.	% 'catch up' per added registrations	Nr of vaccinations effectively applied since previous CoC iteration
1 & 2	1 & 2	01-Jun-21	1.1	2.059.426				
3	3	01-Jul-21	1.7	3.133.344	1.116.119			
4	4	01-Aug-21	2.6	4.854.329	1.720.985	282.647	16%	1.438.338
5	5	01-Sep-21	3.6	6.822.321	1.967.992	241.618	12%	1.726.374
6	6	01-Oct-21	4.7	8.839.765	2.017.444	341.322	19%	1.676.122
7	7	01-Nov-21	5.4	10.148.626	1.308.861	146.408	10%	1.162.453
8	8	17-Jan-22	5.9	11.921.415	1.772.789	123.801	8%	1.648.988
9	9	01-Feb-22	6.1	12.481.710	560.295	112.719	14%	447.576
10	10	01-Mar-22	6.5	13.317.696	835.985	310.159	36%	525.826
14	11	01-Apr-22	6.9	14.092.523	774.827	375.503	48%	399.324
15	12	01-May-22	7.0	14.387.231	294.708	56.036	19%	238.672
16	13	01-Jun-22	7.2	14.858.441	471.21	231.373	49%	239.837
17	14	01-Jul-22	7.5	15.339.609	481.168	25.72	5%	438.664
18	15	01-Aug-22	7.7	15.756.690	417.081	115.105	28%	301.976
19	16	01-Sep-22	8.3	17.321.648	however, ~ 1.275.000 registers not covid related			
19_BIS	16_BIS	02-Sep-22		16.045.255	288.565	80.239	28%	208.326
20	17	01-Oct-22	7.9	16.163.979	118.724	26.343	22%	92.381
21	18	31-Oct-22	7.9	16.220.784	56.805	9.567	17%	47.238

To keep regular interval between CoC iterations and follow ALCOA+ I froze and archived datasets at fix dates.

Registration: each registration corresponds to one vaccine dose shot an individual received. An individual who received 4 doses, we expect to find 4 registrations in the core dataset.

Original iteration nr 11 to 13 were testing versions to adapt to changes on how the original datasets were provided from Mars 2022. More details here after.

'Catch up': while working on the issue of compliance to vaccination regimen, I found out that there was always significant ongoing 'catch up' of the datasets with registrations being added from vaccinations that happened at every month since the beginning of the vaccination on 17 January 2021. I found important to document it to help have on a more accurate review of results for any potential time-series analysis.

On early Sep 22 there was 'a leakage' that added to the covid-19 vaccinations registrations dataset over one million registers of vaccine that were not covid related. I was obliged to clean my working datasets to continue to be able to complete the CoC notebooks. Fortunately, by early October 22 this 'leakage' was addressed.

From '01 December 2021' up to 16 of January 2022 the databases from the Brazilian Ministry of Health (BMoH) that documented the covid-19 pandemic were 'hacked' and unavailable for health professionals and public.

When it became available again on 16 Jan 22, the data from fields & columns 26 and 27, that I used for the CoC tracking until November 2021, were 'simplified' in a way that it was no longer reliable.

That obliged me to re-think my CoC codes to focus on data from column 25 (vaccine manufacturer batch number) to be able to have traceability to the source manufacturer of each vaccine. As result I created a CoC version 02, with a new set of flags code, that consolidated from CoC_11 on early Apr 2022.

<table>
<tr><th colspan="3">CoC_v2_flag_codes</th></tr>
<tr><td colspan="2">CoC_v2 flag code</td><td>AMF Rationale</td></tr>
<tr><td>G</td><td>Green</td><td>‘c25_v_lote’ identification is correctly spelt AND ‘consistent’ with available data in columns ‘c30’ and ‘c31’</td></tr>
<tr><td>LG</td><td>Light Green</td><td>‘c25_v_lote’ identification is ‘lightly misspelled’ (a zero instead O or vice-versa’). However cross-checking other sources suggests it matches a documented vaccine shipment AND/OR is ‘consistent’ with ‘c30’ or ‘c31’</td></tr>
<tr><td>Y</td><td>Yellow</td><td>Per original new dataset from 17 Jan 2022 (after the ‘hack’ of BMoH websites), the AZ brand is now the only provider (out of the 4) with two identifications for the population over 18 yo.
<table><tr><th>c30_v_codigo</th><th>c31_v_nome</th></tr><tr><td>85</td><td>ASTRAZENECA/FIOCRUZ - COVISHIELD</td></tr><tr><td>89</td><td>ASTRAZENECA - ChAdOx1-S</td></tr></table>
Code yellow is when review of fields c25, c30 and c31 suggest that there was a mix up between ‘ChAdOx1” and ‘COVISHIELD’ data entered in the register.</td></tr>
<tr><td>R</td><td>Red</td><td>When we have a correct spelled or misspelt batch number that is a mismatch with the data in columns c30 and/or c31. In other words, batch number from one manufacturer with the vaccine code nr from another manufacturer & provider.</td></tr>
<tr><td>U</td><td>Unknown</td><td>When the batch number is misspelt to a point that we cannot match it with any c30 and/or c31.</td></tr>
<tr><td>UR</td><td>U RED</td><td>When we have names, phone or ID numbers that are clearly not batch numbers</td></tr>
<tr><td>P_ALERTE</td><td>Pediatric Alert</td><td>Since Jan 22 (up to Oct 22) we have only 2 providers approved for pediatric covid- 19 vaccines. Back on Jan 2022 Pfizer started to also provide specific pediatric vaccines batches. For the Pediatric SINOVAC we still have a confusing picture.
<table><tr><th colspan="2">c30_v_codigo</th><th colspan="2">c31_v_nome</th></tr><tr><td>Over 18 yo</td><td>5 to 11 yo</td><td>Over 18 yo</td><td>5 to 11 yo</td></tr><tr><td>86</td><td>98</td><td>SINOVAC/BUTANTAN - CORONAVAC</td><td>SINOVAC - CORONAVAC</td></tr><tr><td>87</td><td>99</td><td>PFIZER - COMIRNATY</td><td>PEDIÁTRICA - PFIZER COMIRNATY</td></tr></table>
P_ALERTE is when my review shows that the batch number in c25 does not matches the c30 or c31 or c3 patient age. Due to the population in question the most critical flag.</td></tr>
</table>

Data Scientists may consider average of 1.2 % of High Flag Alerts (Y, R, U, UR and P_ALERTE) as ‘acceptable’. For several reasons a ‘ClinOps Data Scientist’ cannot share this point of view. Particularly because udve 13 contains Florianopolis, an ‘island’ and 2nd most populated SC city with a worrying high nr of High Flag Alerts. One must not forget how covid-19 is transmitted. On part IV here after I share how after investigations it was found that udve 13 high number is a key ‘revealing symptom’ of an important local disfunction. Source ntbk_04.

CoC_v2 c35_SC_UDVE_Nr	G	LG	Y	R	U	UR	P_ALERTE	Total	G_%	TOT_HIGH_FLAG_ALERTS	HIGH_FLAG_ALERTS_%
udve01	464,520	1,602	97	17	46	0	343	466,625	99.55%	503	0.11%
udve02	350,978	3,601	2,287	260	493	61	1,940	359,620	97.6%	5,041	1.4%
udve03	1,110,175	11,417	2,868	82	1,369	236	1,966	1,128,113	98.41%	6,521	0.58%
udve04	383,835	2,132	16	7	25	17	468	386,500	99.31%	533	0.14%
udve05	476,473	1,332	82	1	118	5	685	478,696	99.54%	891	0.19%
udve06	591,190	30,104	6,345	133	2,314	136	1,494	631,716	93.58%	10,422	1.65%
udve07	753,690	39,296	27,370	636	2,904	272	9,207	833,375	90.44%	40,389	4.85%
udve08	497,440	19,548	5,805	16	333	72	22	523,236	95.07%	6,248	1.19%
udve09	1,616,989	21,864	7,721	24	629	289	5,236	1,652,752	97.84%	13,899	0.84%
udve10	1,544,868	15,578	5,679	5	1,086	93	3,758	1,571,067	98.33%	10,621	0.68%
udve11	612,258	2,645	1,031	0	354	0	325	616,613	99.29%	1,710	0.28%
udve12	1,586,532	18,000	8,617	62	853	200	2,999	1,617,263	98.1%	12,731	0.79%
udve13	2,829,100	81,834	25,634	354	1,820	292	52,504	2,991,538	94.57%	80,604	2.69%
udve14	667,838	1,405	313	7	75	138	1,279	671,055	99.52%	1,812	0.27%
udve15	870,755	2,970	611	2	479	8	1,572	876,397	99.36%	2,672	0.3%
udve16	961,935	2,577	514	16	471	326	1,251	967,090	99.47%	2,578	0.27%
udve17	444,124	1,619	0	5	353	0	247	446,348	99.5%	605	0.14%
Total	15,762,700	257,524	94,990	1,627	13,722	2,145	85,296	16,218,004	97.19%	197,780	1.22%

From March 22, due to the large increase on nr of registers, the BMoH, decided to split the data from each of the 27 Brazilians states, across 3 different datasets per state. That obliged me to rethink again my code structure. After trying some learnings from a PyData event (i.e.; Parquet) and decided to use only the minimalist triad Collab & Pandas & Python I started to be more confident to explore new potential compliance issue matters, and therefore, increased the number of notebooks from 2 to 11.

For review of key findings per notebook please ask the author for the Executive & Summary notebook.

CoC_project structure as per CoC_18 from early Nov 2022.			
PART	Focus	Google Collab Ntbk nr and title	
I Data Wrangling &EDA	IN_OUT	1	Downloading, data wrangling, adding UDVE column and Parqueting.
			Added a new column based on UDE regionalization for IN & OUT accountability
	CoC	2	Defining & Reviewing 'Combinations of Concern (CoC)' versions 1 and 2, first EDA by crosschecking of UDVEs & CoCs.
			Added a CoC_v2 variable populated with the ALERT flag for quality of documentation for vaccine manufacturer.
II Checking Compliance related variables (columns)	Population categorization	3	Review columns c21_to_c24 for missing values
	Vaccine Providers distribution	4	Doses_Types_and_V_Manufacturers
	Subject ID issues	5	Subject_ID_AGE_crosschecks
	Regimen compliance	6	MULTIDOSE_UNIDOSE_DUPLICATES
III. Focusing on data from 'FLORIANOPOLIS' (SC's capital)	BIG CITY COMMUTE "BIAS"	7	FLORIPA_c_18_Multidose_Unidose_Duplicated
		8	FLORIPA_c_13_Multidose_Unidose_Duplicated
		9	FLORIPA_c_13_within_FLORIPA_c18_Multidose_Unidose_Duplicated
IV Reviewing 'Pediatric' data	'5 to 11 yo'	10	PEDIATRICS - SC and FLORIANOPOLIS
Executive Summary	Per iteration - key finding per notebook.		

PART IV: THE 'SWISS CHEESBERG' HEALTHCARE DATA AND POST COVID_19 PANDEMIC ERA.

One of the most worrying findings from my CoC project is that for the paediatric population (5 to 11 yo), 46% of registers from all SC and 99% of registers from Florianopolis (second most SC populated city) shows up as having received an adult version of the vaccine instead of the specific paediatric one. For Pfizer vaccines the dose for adult and paediatric are different (I,e Pfizer adult = 0.3 ml and Pfizer paediatric = 0.2ml). For Coronavac/Sinovac it is the same dose (0.5 ml) for adults and paediatrics.

After some 'informal investigation' it was suggested the cause of this 'possible mixed up' might be an interoperability issue. Particularly in Florianopolis where there are known interoperability and other issues with 2 out the 4 digital platforms used at points of vaccination. Check CoC_Executive_Summary_ntbk_10 key findings.

Unfortunately, this kind of worrying finding, still not resolved at end of Oct 22, is not a SC exclusivity. Missing data and interoperability systems issues in all directions happens worldwide without exception. That led me to share my ClinOps Clinical Data Scientist perspective on the matter.

'Data Science' 'education' not usually fit for Healthcare Data:

It is not scientific the very cavalier way how 'Data Sciences' and 'Machine Learning' courses usually 'teach' how to deal with missing and outlier data. As a hardwired ClinOps that was the most important cultural shock I had when I start learning SAS programming and SAS for Data Sciences. Unfortunately, I saw the same unscientific attitude related to missing & outlier's data on the R and Python courses.

The way 'Imputation' techniques (simply deleting, transforming' missing and/or outliers' data) are careless presented on such courses might be acceptable in banking, retail, physics, or other business domains. However, in healthcare, at least for Clinical Research & Development that must not be the case. Missing data and outliers in healthcare require a far more sophisticated, cautious, scientifically well thought, and documented approach.

Diseases are outliers. 'The Laws of Medicine' – Field Notes from an Uncertain Science' – Siddhartha Mukherjee

Clinical Trials Data are outliers out of outliers. (me and many others)

Safety Clinical Trials Serious Adverse Events (SAEs) are outlier out of outlier out of outlier. Please read 'The progress of Experiment' – Harry M. Marks to understand why clinical trials are regulated as they are today.

Pharma Biometrics perception

The PHUSE community is essentially composed by members from the Pharma & Biotech R&D 'Biometric departments' (schematically, R- 2 to R - 4 on my table here after). As a ClinOps I would like to remind the 'Biometrics' crowd that their 'bread and butter' material, on top of being outliers, are, from the data integrity point of view, very far from the real picture of raw source clinical data in real life healthcare.

Far from Reality	Pharma & Biotech Clinical Development Departments	
R -4	Data Scientists, programmers,	
R -3	Data Management (DM), Statisticians	Reduce source data entropy by ~ 10% but fragment the holistic clinical picture. Often fragmented DM process between sponsors and vendors contributes to slow recruitment speed in clinical trials.
R -2	Pharmacovigilance department (PV)	When a SAE happen, PV gets the best picture of the event. Unfortunately, PV not always closely works with ClinOps and DM, so it may increase entropy, upset, and increase nurses & doctors' burden.
R -1	Clinical Operations Managers	
Reality	Clinical Research Assistants (CRA) Site Clinical Trial Nurses (SCTN)	Reduce study patient source data entropy & 'missing data' by at least 80% before it gets to CRF & DM.

Normally, at least for NDA[viii] phase II and III clinicals trials, the ClinOps CRA and SCTN (very rare bird in healthcare)) spend hours to ensure that all the sources' documents are well placed, ordered and chronologically tied up and clear in the patient source file BEFORE the required data in entered in the CRF. They usually reduce clinical source data (patient file) entropy by at least 50%, most often up to 80%, before it gets to a CRF and then Data Management.

Per example, clinical trials involving acute conditions at enrolment (i.e., stroke, MI, schizophrenia, or depression exacerbation). In the UK, ask your ClinOps colleagues how long it usually takes for a SCTN to obtain a proper documented medical history record & concomitant medications from the patient's GP, that would be acceptable for regulatory audit purposes. Or ask your PV department how long it takes to close an SAE.

Pre-pandemic reality of healthcare data

99,99% of healthcare patient's file data worldwide will never have the luxury to 'be cared' by a CRA & SCTN. Therefore, for this ClinOps with extensive international experience, the premise is always that, raw healthcare data is at first sight, by Clinical Data Sciences terms, 'garbage data', 'Swiss cheesberg' (definition here after).

Before the covid-19 pandemic, during the reign of the paper paradigm of source documents, patient files anywhere in the world and for any therapeutic area were already, at first sight, unreliable for scientific needs (some worse than others). The amount of work done by CRAs and SCTNs during clinical trials are the proof of it.

A patient health-care file is not one page written by one nurse and one doctor. In public healthcare, particularly for chronic patients and acute events, patients' files are fuzzy kaleidoscopes[ix] with very fragmented moving docs, uncoordinatedly filled, most often by dozens of overloaded and tired healthcare workers (each with their own jargon and style) into a plethora of not interoperable digital platforms of all sorts and sizes.

After digital paradigm start to expand in healthcare, people start to talk about the wonders and promises of 'Big Data', RWD, REW, PPRL. etc. Pharma started to hire Data Scientists & Engineers to deal with 'huge datasets of 'anonymised data' (most often 'audited' data), to create synthetic arms for clinical trials and so on.

Very little effort was made, even within the PHUSE community, to remind and highlight the big 'GARBAGE' (from the scientific point of view) that, in general, raw data (digital or paper) is in real life healthcare. For any therapeutic area and anywhere in the world.

Post-pandemic reality of healthcare data

There's absolutely no doubt about how, during the covid-19 pandemic, digital technology positively contributed to what is the most triumphant medical science response in the history of humanity.

However, as Clinical Data Scientists, we must be aware about the actual 'health' of our 'bread and butter' material: digital healthcare data. Since the start of covid-19 pandemic, the entropy of healthcare digital data is getting rapidly worse than before, and it will not get better anytime soon. A few reasons for that.

- Healthcare staff reduction: before the covid-19 pandemic, without exception in the world, overloaded healthcare staff was already a key contributor factor for the 'entropy', 'garbage', 'swiss chessberg' of patient files. As already warned by Dr Topol back in 2019, overloaded healthcare professionals, working on crazy shifts with 'unfriendly' and often not interoperable tools does not contribute to deliver 'clear patients files free of 'missing data'. He personally paid a high price for that.[x] Covid-19 pandemic drastically reduced healthcare staff during the pandemic. We all know that these numbers in the next years will be far below than before the pandemic despite increase of demand (i.e., 'long covid'). Consequently the 'quality of health of patient healthcare' data has been suffering a significant deterioration and is not close to get better anytime soon.

- Global disruptions of Medical Supply and 'Gold Standards': Clinical Data Scientists need also to be aware about the massive logistical disruption of all sorts of medical supplies worldwide. Public and private hopsitals were obliged to change their 'gold standards treatments'. Such issues will take long time to be resolved. Clinical Data Scientists, when working with datasets generated after covid started ('anonymised datasets', 'synthetic arms for clinical trials', PPRL, ML for prediction in healthcare, etc), must first check if they are still 'representative' of the pre-pandemic 'gold standards of treatment'. [xi]

- Cyber-attacks: Another important issue adding to the above is the vulnerability of digital healthcare. As I saw happen during my CoC project and in the UK recently[xii], hack of healthcare services has an immediate deleterious impact on the quality of patient healthcare datasets. The work overload created by the digital hack on already very stretched healthcare resources certainly has very negative impact on the accuracy of the patient medical records. Such events must be carefully considered when performing time-series research or selecting datasets for Machine Learning prediction in healthcare.

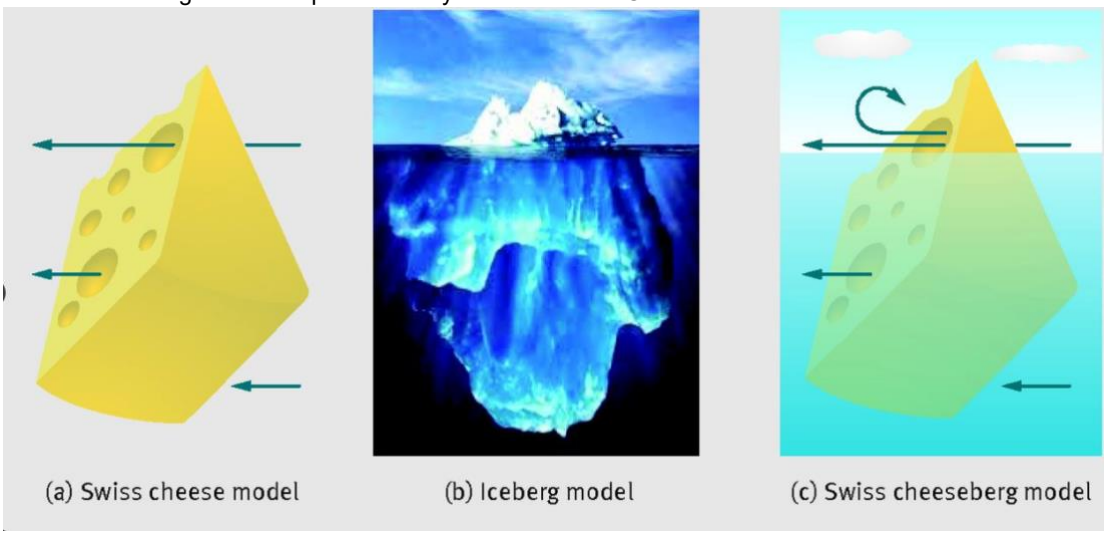
Data Scientists knows the mantra 'garbage in, garbage out'. It is important that Clinical Data Scientists are aware that healthcare data generated since the pandemic started and for the years to come, need to be audited and worked with the most extreme caution. Clinical Data Scientists must learn to be thoroughly inquisitive about the source and source dataset's quality if you want to use it for scientific purposes.

As a food for thought to conclude this section, I will borrow and expand the concept of "Swiss cheesberg model' for patient safety from Professor Coiera[xiii]. I am approaching it from the statistical meaning of 'population representativity'.

We need to think about digital healthcare data generated since the covid-19 started as a 'Swiss cheesberg' having much bigger 'holes' (ie. missing data) than before. The visible over the water point, healthcare dataset, are no longer, 'representative', 'adequately supported by what is 'underneath'.

The overall 'Swiss chessberg' is now far more holed, cracked, unstable and prone to sudden flipping over and fragmentation than before. So, when using RWD, PPRL, etc, to be safe about the quality of our analysis, we must first be aware about the real status of our datasets 'holes' and document them carefully.

"Swiss cheesberg model' for patient safety from Professor Coiera[xiv]



Conclusions

Good news now to conclude this paper. A skilled, down to earth, combination of CRBM and open sources tools could be a good start point to help Clinical Data Scientists to deal with the 'Swiss cheesberg'.

Proof of Concept

The simple triad, Google Collab IDE & Pandas and Python passed my 'proof of concept' as potential tools to be used for ClinOps managerial purposes for clinical trials. With discipline and collaborative work with regulatory and Quality Assurance departments we can make them better suited for ICH & GCP & ALCOA+ requirements and therefore acceptable to regulatory authorities.

The possible next step would be to think strategically AND tactically how to use those sorts of tools for the big scope of clinical drug program development to speed up process for the later purpose of regulatory submission. Lab results across the clinical program development of a study drug would be a good case study.

Due to the multidisciplinary background of Data Scientists using very disparate jargons, a kind of 'terminology chart & policy' for project & program clinical drug development documentation may help minimise cross-departmental and international communication confusion. Also develop pragmatic good practices how to upfront deal with 'missing data' and 'outliers', in a different way that's done by DM, throughout program development is highly recommended.

The OpenStudyBuilder[xv] project recently initiated by Novo Nordisk seems to go in this direction.

Did you said automate?

This paper's title says, '*Python's ecosystem <u>to automate</u> a Clinical Risk Based Monitoring'.* Did that happen? Depend on what we mean by 'automate'. As mentioned, healthcare data is a very untied and changeable without warning. Therefore, a lot of think should be made to ensure that the overall project code strategy can spot and warn about any dataset changes since last run.

That said, the triad Google Collab IDE & Pandas & Python make the code maintenance and replication from notebook or project iteration to the other much easier and friendly. That allowed a 'semi-automation' of the updates. If no big surprises, the actual monthly update, careful analysis, and proper documentation of the actual 12 notebooks and support document filings requires about 4 working hours days. I expect that with better hardware and internet, programming experience, and time I could reduce this time by half.

I leave the reader with two citations as food for thought:

'*The science of medicine* *tells us that* *uncertainty, imperfections,* *priors, outliers, and biases,* *are endemic.* *They are not going to go away today,* *tomorrow or 50 or 100 years from now.* *The sooner we instill these ideas* *in ourselves, and our patients,* *the more likely it is that* *we will have a much more collaborative effort* *to solve the major problems of human disease.* *We need to do that*.'	*'…, ML models* *are usually not used in a static and isolated way,* *but are embedded in some process or product,* *and interact with people.* *A more dynamic and holistic view of the entire process,* *<u>from data collection</u>* *to the final consumption of the explained prediction is needed.* *This includes thinking how* *to explain predictions* *to individuals* *with diverse knowledge and backgrounds* *and* *about the need of interpretability* *on the level of an institution or society in general*. '
Dr Siddhartha Mukherjee **'Eric J. Topol. 'Emperor's Mukherjee on 'Three Laws of Medicine'. Medscape, Oct 12, 2015'**	'Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges' Christoph Molnar, Giuseppe and Bernd Bischl

RECOMMENDED READING

Books:

'The progress of Experiment' – Harry M. Marks - Science and Therapeutic Reform in the United States, 1900-1990 - Cambridge University Press - This book examines the science and the politics of drug evaluation. That will help Data Scientists to understand the need for ICH & GCP and ALCOA+ guidelines.

"The Laws of Medicine' – Field Notes from an Uncertain Science' – Dr Siddhartha Mukherjee

'DEEP MEDICINE' – Dr Eric Topol

Articles:

'Eric J. Topol. 'Emperor's Mukherjee on 'Three Laws of Medicine'. Medscape, Oct 12, 2015'

Coiera E, Collins S, Kuziemsky C. A unified model of patient safety (or "Who froze my cheese?") *BMJ* 2013; 347 :f7273 doi:10.1136/bmj.f7273

CONTACT INFORMATION

Aldir Medeiros Filho
Clinical Operations Project & Program Manager and Clinical Data Scientist
amfilho60@gmail.com

REFERENCES

[i] https://en.wikipedia.org/wiki/Proof_of_concept

[ii] https://www.transceleratebiopharmainc.com/?s=Risk+based+monitoring&submit=submit

[iii] https://en.wikipedia.org/wiki/Santa_Catarina_(state)

[iv] https://en.wikipedia.org/wiki/Federative_units_of_Brazil

[v] https://phuse.global/Communications/PHUSE_Archive/1

[vi] https://youtu.be/L45Ux7a7FyU

[vii] https://www.bbc.co.uk/news/health-61010090

[viii] https://www.fda.gov/drugs/types-applications/new-drug-application-nda

[ix] https://en.wikipedia.org/wiki/Kaleidoscope

[x] https://youtu.be/ZWAnjdQy5kE

[xi] https://www.bbc.co.uk/news/health-63573718

[xii] https://www.bbc.co.uk/news/technology-62725363

[xiii] https://www.bmj.com/content/bmj/347/bmj.f7273.full.pdf

[xiv] https://www.bmj.com/content/bmj/347/bmj.f7273.full.pdf

[xv] https://novo-nordisk.gitlab.io/nn-public/openstudybuilder/project-description/

https://www.gov.br/saude/pt-br/centrais-de-conteudo/publicacoes/publicacoes-svs/coronavirus/plano-nacional-de-operacionalizacao-da-vacinacao-contra-a-covid-19-pno-2a-edicao-com-isbn