

Le pratiche della ricerca scientifica aperta e il loro impatto: dati aperti, codice software aperto, pubblicazioni aperte

Davide Chicco

davide.chicco@unimib.it

www.DavideChicco.it



LINUX DAY MILANO 2024



Mi presento

Sono ricercatore presso il Dipartimento
di Informatica Sistemistica e Comunicazione
dell'Università di Milano-Bicocca



E collaboratore con l'University of Toronto (Canada)

Utente Linux Ubuntu dal 2004 (ora attualmente Xubuntu)
Programmatore R e Python

Ricerca scientifica principalmente in informatica biomedica

Su cosa si basa una scoperta scientifica?

- Una scoperta scientifica inizia da un'ipotesi teorica (eureka!) che viene poi confermata dagli esperimenti

Su cosa si basa una scoperta scientifica?

- Una scoperta scientifica inizia da un'ipotesi teorica (eureka!) che viene poi confermata dagli esperimenti
- Tutto qua? No: la scoperta scientifica dev'essere anche **documentata e replicabile**



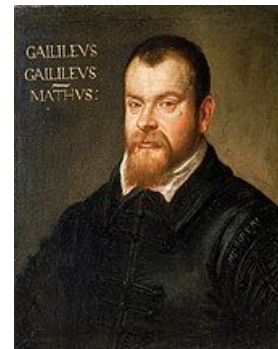
Su cosa si basa una scoperta scientifica?

- Una scoperta scientifica inizia da un'ipotesi teorica (eureka!) che viene poi confermata dagli esperimenti
- Tutto qua? No: la scoperta scientifica dev'essere anche **documentata e replicabile**
- La documentazione dev'essere scritta dai ricercatori e dalle ricercatrici che hanno condotto gli esperimenti e deve fornire tutti i dettagli necessari a spiegare le ipotesi e gli esperimenti fatti



Su cosa si basa una scoperta scientifica?

- Galileo Galilei (1564-1642) fu il primo scienziato nella storia a fornire una documentazione delle sue osservazioni sperimentali mentre osservava Giove e le sue orbite
- *Sidereus nuncius*: il suo diario scientifico con dati, metadati, disegni, date, informazioni e descrizioni delle osservazioni da lui fatte



Immagini Licenza aperta Wikimedia Commons



Immagine Licenza aperta PLOS Computational Biology
<https://doi.org/10.1371/journal.pcbi.1003542.q001>

Su cosa si basa una scoperta scientifica?

- Il *Sidereus nuncius* non permette solamente di sapere come Galileo aveva fatto i suoi esperimenti, ma bensì permetteva a chiunque in possesso d'una tecnologia simile (telescopio) e di competenze scientifiche sufficienti di **riprodurre** le sue osservazioni sperimentali, e quindi di verificarle
- Il metodo usato nel *Sidereus nuncius* e il lavoro di Galileo Galilei in generale sono considerate **le basi del metodo scientifico moderno**



Immagine Licenza aperta PLOS
Computational Biology
<https://doi.org/10.1371/journal.pcbi.1003542.g001>

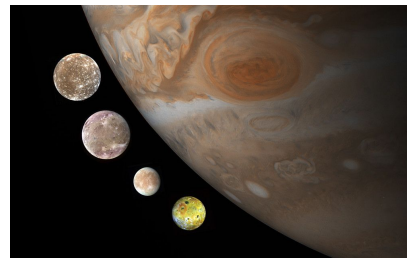


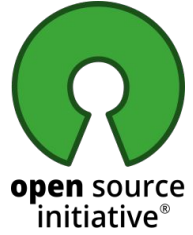
Immagine Licenza aperta [Flickr](#)

Cos'è scoperta scientifica?

- Un'affermazione può dirsi una scoperta scientifica solo se ci sono le informazioni necessarie per **riprodurre gli esperimenti fatti**, e se questa riproduzione porta agli stessi risultati affermati
- **Se non c'è la documentazione** su com'è stata fatta una presunta scoperta scientifica, **non si tratta d'una scoperta scientifica**
- I principi e le pratiche della **scienza aperta** possono portare a scoperte più solide e affidabili

Principi della scienza aperta

1. Scelta di linguaggi di programmazione e software aperti (open source)
2. Utilizzo dati aperti (open data)
3. Condivisione degli scripts online apertamente (code sharing)
4. Condivisione dei dati online apertamente (data sharing)
5. Pubblicazione degli articoli scientifici su riviste ad accesso aperto (open access)



GitHub



Principi della scienza aperta

1. Scelta di linguaggi di programmazione e software aperti (open source)
 - In un progetto scientifico computazionale, la scelta d'usare linguaggi di programmazione gratis e aperti come **R** o **Python** permette a chiunque
 - Lo stesso vale per sistemi operativi e software, come Linux e LibreOffice



Principi della scienza aperta

2- Utilizzo dati aperti (open data)

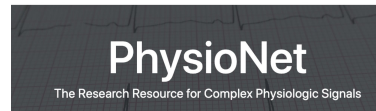
- Datasets gratuiti con licenza aperta possono essere trovati online su varie piattaforme come Figshare, Zenodo, University of California Irvine Machine Learning Repository, PhysioNet, ecc
- Esistono anche motori di ricerca per dati aperti: Google Dataset Search e re3data.org



Ricerca di dataset

Quando si cerca un dataset online però, occorre prestare attenzione: non tutti sono di buona qualità

Controlla se c'è una pubblicazione scientifica associata al dataset, e se tutte le caratteristiche del dataset sono documentate



Qualità dei dati

Kaggle per esempio dà un voto da 0 a 10 (Kaggle Usability Score) ai datasets che pubblica, per indicarne la qualità



MYSAR AHMAD BHAT · UPDATED 2 YEARS AGO



382

New Notebook



Download (2 kB)



Lung Cancer

Does Smoking cause Lung Cancer.

kaggle



Data Card

Code (107)

Discussion (5)

Suggestions (0)

About Dataset

The effectiveness of cancer prediction system helps the people to know their cancer risk with low cost and it also helps the people to take the appropriate decision based on their cancer risk status. The data is collected from the website online lung cancer prediction system .

Total no. of attributes:16

No .of instances:284

Usability ⓘ

10.00

License

CC0: Public Domain

Expected update frequency

Never

Qualità dei dati

Kaggle per esempio dà un voto da 0 a 10 (Kaggle Usability Score) ai datasets che pubblica, per indicarne la qualità

About Dataset

The effectiveness of cancer prediction system helps the people to know their cancer risk with low cost the people to take the appropriate decision based on their cancer risk status. The data is collected from online lung cancer prediction system .

Total no. of attributes:16

No .of instances:284

Attribute information:

1. Gender: M(male), F(female)
2. Age: Age of the patient
3. Smoking: YES=2 , NO=1.
4. Yellow fingers: YES=2 , NO=1.
5. Anxiety: YES=2 , NO=1.
6. Peer_pressure: YES=2 , NO=1.
7. Chronic Disease: YES=2 , NO=1.
8. Fatigue: YES=2 , NO=1.
9. Allergy: YES=2 , NO=1.

This score is calculated by Kaggle.

Completeness · 100%

- ✓ Subtitle
- ✓ Tag
- ✓ Description
- ✓ Cover Image

Credibility · 100%

- ✓ Source/Provenance
- ✓ Public Notebook
- ✓ Update Frequency

Compatibility · 100%

- ✓ License
- ✓ File Format
- ✓ File Description
- ✓ Column Descriptio

kaggle

Ricerca di dataset

I datasets si possono anche trovare nel materiale supplementare d'articoli scientifici, tipo PLOS One. Ad esempio:

Circulating osteocalcin as a bone-derived hormone is inversely correlated with body fat in patients with type 1 diabetes

Yuichi Takashi , Masashi Ishizu , Hiroyasu Mori, Kazuyuki Miyashita, Fumie Sakamoto, Naoto Katakami, Taka-aki Matsuoka, Tetsuyuki Yasuda, Seiichi Hashida, Munehide Matsuhisa , Akio Kuroda

Published: May 3, 2019 • <https://doi.org/10.1371/journal.pone.0216416>

Article	Authors	Metrics	Comments	Media Coverage
				

Abstract

Introduction

Subjects and methods

Results

Discussion

Supporting information

References

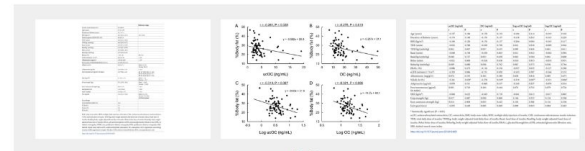
Reader Comments

Figures

Abstract

The objective of the present study was to investigate the correlations between serum undercarboxylated osteocalcin (ucOC) or osteocalcin (OC) concentrations and %body fat, serum adiponectin and free-testosterone concentration, muscle strength and dose of exogenous insulin in patients with type 1 diabetes. We recruited 73 Japanese young adult patients with childhood-onset type 1 diabetes. All participants were receiving insulin replacement therapy. The correlations between logarithmic serum ucOC or OC concentrations and each parameter were examined. Serum ucOC and OC concentrations were inversely correlated with %body fat ($r = -0.319$, $P = 0.007$; $r = -0.321$, $P = 0.006$, respectively). Furthermore, multiple linear regression analyses were performed to determine whether or not serum ucOC or OC concentrations were factors associated with %body fat. Serum ucOC and OC concentrations remained significant factors even after adjusting for gender, HbA1c, body weight-adjusted total daily dose of insulin and duration of diabetes ($\beta = -0.260$, $P = 0.027$; $\beta = -0.254$, $P = 0.031$, respectively). However, serum ucOC and OC concentrations were not correlated with serum adiponectin or free-testosterone concentrations, muscle strength or dose of exogenous insulin. In conclusion, our study demonstrates the inverse correlation between serum ucOC or OC concentrations and body fat in patients with type 1 diabetes.

Figures



1,740 View	4 Share
---------------	------------

Download PDF	
Print	Share

 Check for updates

ADVERTISEMENT



Ricerca di dataset

I datasets si possono anche trovare nel materiale supplementare d'articoli scientifici, tipo PLOS One. Ad esempio:

Circulating osteocalcin as a bone-derived hormone is inversely correlated with body fat in patients with type 1 diabetes
Yuichi Takashi, Masashi Ishizu, Hiroyasu Mori, Kazuyuki Miyashita, Fumie Sakamoto, Naoto Katakami, Taka-aki Matsuoka, ...

Abstract
Introduction
Subjects and methods
Results
Discussion
Supporting information
References

Reader Comments
Figures

Citation: Takashi Y, Ishizu M, Mori H, Miyashita K, Sakamoto F, Katakami N, et al. (2019) Circulating osteocalcin as a bone-derived hormone is inversely correlated with body fat in patients with type 1 diabetes. PLoS ONE 14(5): e0216416. <https://doi.org/10.1371/journal.pone.0216416>

Editor: Masaki Mogi, Ehime University Graduate School of Medicine, JAPAN

Received: January 26, 2019; **Accepted:** April 21, 2019; **Published:** May 3, 2019

Copyright: © 2019 Takashi et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability: All relevant data are within the manuscript and its Supporting Information files.

Funding: M.M. received the grant from Japan Agency for Medical Research and Development (No. 17k1010002h0003) (<https://www.amed.go.jp/en/index.html>). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Subject Areas

- Insulin
- Adipose tissue
- Diabetes mellitus
- HbA1c
- Osteocalcin
- Adiponectin
- Diabetic retinopathy
- Type 2 diabetes

Introduction

Bones perform several functions, such as supporting the body, protecting the internal organs and central nervous system and contributing to hematopoiesis. Recently, it has been reported that the skeleton also functions as an endocrine organ and systemically regulates the functions of other organs [1].

Several bone-derived hormones have been defined, including fibroblast growth factor 23 and lipocalin 2, among others [2–5]. Osteocalcin (OC), another such hormone, is reported to affect glucose and energy metabolism [6]. It is produced by osteoblasts and acts as a bone matrix protein. The serum level of OC was reported to be sustained after adulthood, and the reference value of serum OC concentrations is 8.4–33.1 ng/ml in males, 7.8–30.8 ng/ml in premenopausal females and 14.2–54.8 ng/ml in postmenopausal females [7]. Undercarboxylated osteocalcin (ucOC) is considered to be the active form of circulating OC, exerting endocrine functions according to experimental studies [8, 9]. However, whether or not ucOC is the active form in humans as well is unclear. The reference value of the serum ucOC concentrations is <4.5 ng/ml in general, although Shiraki et al. reported that the mean serum ucOC concentrations was 3.0 ng/ml in their Japanese cohort study [10].

Ricerca di dataset

I datasets si possono anche trovare nel materiale supplementare d'articoli scientifici, tipo PLOS One. Ad esempio:

Circulating osteocalcin as a bone-derived hormone is inversely correlated with body fat in patients with type 1 diabetes

Yuichi Takashi, Masashi Ishizu, Hiroyasu Mori, Kazuyuki Miyashita, Fumie Sakamoto, Naoto Katakami, Taka-aki Matsuoka, ...

Supporting information

[Abstract](#)

[Introduction](#)

[Subjects and methods](#)

[Results](#)

[Discussion](#)

[Supporting information](#)

[References](#)

[Reader Comments](#)

[Figures](#)

Circulating osteocalcin as a bone-derived hormone is inversely correlated with body fat in patients with type 1 diabetes

Showing 3/3: pone.0216416.s003.xlsx

	A	B	C	D	E	
1	no	gender (male=1, female=0)	age	duration of diabetes	BMI	insulin regi
2	1	0	31	17	22.4376704418629	0
3	2	1	31	25	25.5569364310525	1
4	3	0	32	26	20.9765625	0
5	4	0	25	17	21.2977601232862	1
6	5	1	39	35	32.6089403451714	0
7	6	0	37	25	22.6318993047242	0
8	7	0	38	30	20.6619637650979	1
9	8	1	35	26	23.3920939998693	1

data set



3 / 3



Download

Minimal data set of this study.

(XLSX)

Principi della scienza aperta

3. Condivisione degli scripts online apertamente (code sharing)

- Importantissimo: una volta finito il progetto, ricercatori e ricercatrici **dovrebbero pubblicare su GitHub, GitLab o altre piattaforme tutto il codice software necessario** per riprodurre gli esperimenti e riottenere gli stessi risultati che hanno descritto



GitHub

Principi della scienza aperta

4. Condivisione dei dati online apertamente (data sharing)

- Insieme al codice software, ricercatori e ricercatrici dovrebbero i datasets utilizzati per il progetto scientifico (se in possesso della licenza per la pubblicazione libera)
- Sia nel caso di dati nuovi (per esempio, ottenuti da un ospedale, sia nel caso di dati preprocessati)



Principi della scienza aperta

Research Article | [Open access](#) | Published: 03 February 2020

Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone

[Davide Chicco](#) ✉ & [Giuseppe Jurman](#)

BMC Medical Informatics and Decision Making 20, Article number: 16 (2020) | [Cite this article](#)

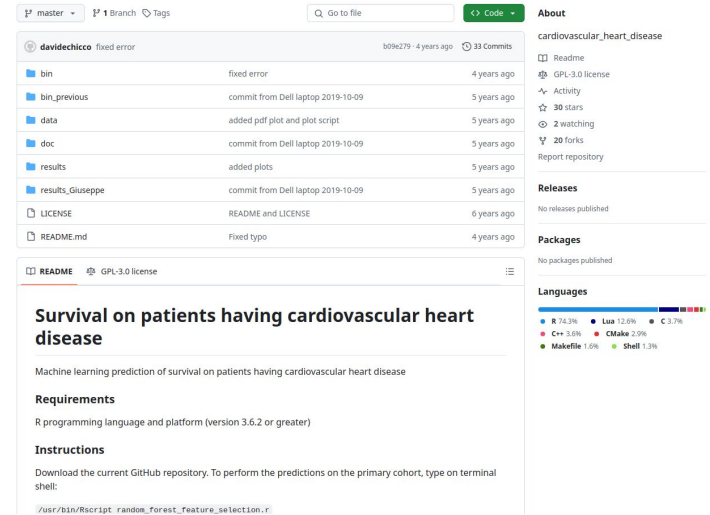
Availability of data and materials

The dataset used in this project [66] is publicly available under the Creative Commons Attribution 4.0 International (CC BY 4.0) license at:

https://plos.figshare.com/articles/Survival_analysis_of_heart_failure_patients_A_case_study/5227684/1

Our software code is publicly available under the GNU General Public License v3.0 at:

https://github.com/davidechicco/cardiovascular_heart_disease



The screenshot shows the GitHub repository page for 'cardiovascular_heart_disease' by user 'davidechicco'. The repository is in the 'master' branch and has 33 commits. The file list includes 'bin', 'bin_previous', 'data', 'doc', 'results', 'results_Giuseppe', 'LICENSE', and 'README.md'. The 'README' file is selected, showing the title 'Survival on patients having cardiovascular heart disease' and a description: 'Machine learning prediction of survival on patients having cardiovascular heart disease'. It also lists requirements for R programming language and platform (version 3.6.2 or greater) and provides instructions for downloading the repository and performing predictions. The 'About' section on the right shows the repository is licensed under GPL-3.0, has 30 stars, and 20 forks. The 'Languages' section shows the repository is primarily in R (74.3%), with other languages like Lua, C++, C, and Shell also present.

File	Commit	Time
bin	fixed error	4 years ago
bin_previous	commit from Dell laptop	5 years ago
data	added pdf plot and plot script	5 years ago
doc	commit from Dell laptop	5 years ago
results	added plots	5 years ago
results_Giuseppe	commit from Dell laptop	5 years ago
LICENSE	README and LICENSE	6 years ago
README.md	Fixed typo	4 years ago

Survival on patients having cardiovascular heart disease

Machine learning prediction of survival on patients having cardiovascular heart disease

Requirements

R programming language and platform (version 3.6.2 or greater)

Instructions

Download the current GitHub repository. To perform the predictions on the primary cohort, type on terminal shell:

```
/usr/bin/Rscript random_forest_feature_selection.R
```

About

cardiovascular_heart_disease

Readme
GPL-3.0 license
Activity
30 stars
2 watching
20 forks
Report repository

Releases

No releases published

Packages

No packages published

Languages

R 74.3%
Lua 12.0%
C++ 3.0%
C 3.7%
CMake 2.0%
Makefile 1.0%
Shell 1.3%

Principi della scienza aperta

5. Pubblicazione degli articoli scientifici su riviste ad accesso aperto (open access)

- Una volta ottenuti risultati scientifici interessanti, è prassi scrivere un articolo scientifico da inviare ad una rivista per la pubblicazione
- Esistono migliaia di riviste scientifiche, e gli autori e le autrici dello studio devono scegliere tra due categorie: ad accesso chiuso o ad accesso aperto (open access)



PLOS
COMPUTATIONAL
BIOLOGY



BioData Mining

PeerJ
Computer Science

Principi della scienza aperta

5. Pubblicazione degli articoli scientifici su riviste ad accesso aperto (open access)

- Le riviste ad accesso chiuso permettono agli autori e alle autrici d'un articolo di pubblicarlo gratis, ma poi fanno pagare chi vuole leggerlo (anche per biblioteche, università, ecc)
- Le riviste ad accesso aperto chiedono il pagamento dei costi di pubblicazioni agli autori e alle autrici (in media 3.000€) e poi rende liberamente disponibile l'articolo per chiunque



PLOS
COMPUTATIONAL
BIOLOGY



BioData Mining

PeerJ
Computer Science

Principi della scienza aperta

5. Pubblicazione degli articoli scientifici su riviste ad accesso aperto (open access)

- Da una ventina d'anni s'è anche diffusa l'abitudine di condividere preprints (versioni finali di articoli, pre-invio alle riviste scientifiche) su piattaforme gratuite come arXiv, bioRxiv, medRxiv, e Research Square
- Utili per la disseminazione, ma non offrono revisione



arXiv

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

medRxiv

THE PREPRINT SERVER FOR HEALTH SCIENCES



Research
Square

Principi della scienza aperta

1. Scelta di linguaggi di programmazione e software aperti (open source)
2. Utilizzo dati aperti (open data)
3. Condivisione degli scripts online apertamente (code sharing)
4. Condivisione dei dati online apertamente (data sharing)
5. Pubblicazione degli articoli scientifici su riviste ad accesso aperto (open access)



GitHub



Principi della scienza aperta

Seguire queste buone abitudini può portare a risultati scientifici più solidi, affidabili, che possono portare al progresso della società e a nuove scoperte

Non seguire questi Principi porta a: opacità, mancanza di trasparenza, conseguenze problematiche sulla società, mancanza d'informazioni

Grazie a tutte e a tutti per l'attenzione e l'ospitalità!

Le pratiche della ricerca scientifica aperta e il loro impatto: dati aperti,
codice software aperto, pubblicazioni aperte

Davide Chicco

davide.chicco@unimib.it

www.DavideChicco.it

