MEDIA 2017;17:134-138

Cosa Cambieranno i Big Data
nella prevenzione e gestione del diabete

Summary

Artificial Intelligence (AI) is one of the most powerful phenomena fuelled by Big Data: initial phases of AI in healthcare have already shown enormous potential. A radical change is clearly occurring in medical science, where AI and Big Data are making personalized care and prevention a concrete reality. In this disruption, diabetes is an area of strong focus, with promising results.

Rita Zilich[1], Nicoletta Musacchio[2], Gerardo Medea[3], Gaudenzio Stagno[4]

[1] Partner della Società Mix-x; [2] Presidente Nazionale AMD; [3] Area di interesse metabolica, endocrinologica e diabetologica della SIMG; [4] Dirigente Medico Diabetologia e Malattie Metaboliche, Ospedale "Giovanni XXIII", Gioia Tauro, ASP di Reggio Calabria

Gli strumenti di analisi predittiva sono i nuovi protagonisti della scienza medica che utilizza i Big Data

Strumenti di analisi dei dati di tipo descrittivo, predittivo e prescrittivo

Di Big Data se ne parla ovunque. Il significato, però, non è univoco e per questo le loro caratteristiche vengono sintetizzate nelle 4 "V": Volume, è intuitivo, sono tanti; Varietà: sono eterogenei; Velocità: se ne producono in continuazione; Veridicità: le fonti che li generano sono molteplici, è necessario valutarne la credibilità. Gli strumenti informatici di gestione dei Big Data si chiamano *Business Analytics* e, in base agli obiettivi dell'analisi, vengono classificati in: *Descriptive, Predictive* e *Prescriptive Analytics*. Pensiamo, per esempio, a dei dati epidemiologici analizzati con l'intenzione di produrre una fotografia e, tramite grafici e indicatori statistici, evidenziare l'incidenza della patologia diabetica in vari sottogruppi di popolazioni. Questa è un'analisi di tipo descrittivo. Oppure, si potrebbe avere l'intenzione di approfondire la conoscenza sui fattori di rischio delle complicanze nel diabete. In questo caso si parla di analisi predittiva, perché si basa su strumenti in grado di individuare i collegamenti fra diversi fattori, indicando anche la probabilità con cui quei fattori possono presentarsi congiuntamente. Un esempio può essere la ricerca dei fattori di rischio della malattia renale cronica nel paziente diabetico: in una monografia AMD si riporta che un minor filtrato renale e l'ipertensione sono presenti in percentuali significative fra i pazienti che successivamente sviluppano una malattia renale cronica. Lo stesso ragionamento sulla ricerca di correlazioni si può fare anche in positivo, chiedendosi, per esempio, quali caratteristiche abbiano in comune i pazienti che effettuano un buon autocontrollo strutturato, e capire se sia possibile favorire la presenza dei fattori positivi. La ricerca medica è uno dei campi in cui gli strumenti di analisi predittiva potranno esprimere il loro forte potenziale, perché è un ambito caratterizzato dalla presenza di grandi volumi di dati che contengono ancora molta conoscenza "nascosta": fino a pochi anni

Parole chiave

Big Data
Diabete
Intelligenza Artificiale
Scienze 'Omiche

Indirizzo per la corrispondenza

RITA ZILICH
rita.zilich@mix-x.com

GAUDENZIO STAGNO
gaudenzio.stagno@tin.it

fa era impensabile avere a disposizione strumenti dalle capacità computazionali e matematiche come quelle attuali.
Un'ulteriore sofisticazione nell'esame dei dati è l'analisi prescrittiva: oltre a ricercare regole predittive, si fanno delle simulazioni di tipo *what-if* per capire se e come, attraverso la modifica di alcuni fattori, si possano migliorare gli outcome.
In sostanza: con i *descriptive analytics* si crea reportistica sul passato; con i *predictive* si creano modelli basati sul passato per prevedere il futuro; con i *prescriptive* si utilizzano i modelli creati per selezionare i comportamenti ottimali.

L'accelerazione impressa dall'IOT, o internet delle cose, e dalle tecnologie "indossabili"

La grande quantità e varietà di dati a disposizione sulla salute, unita alla possibilità di analizzarli con strumenti informatici sempre più potenti e meno costosi, stanno rivoluzionando la medicina. Da questo punto di vista, assistiamo al dilagare di un altro fenomeno: l'IOT, o internet delle cose. Riguarda tutti i dispositivi in grado di interagire con la rete. Gli oggetti di uso comune che diventano "smart", perché contengono dei sensori che raccolgono le informazioni e le inviano in rete. E, infatti, le tecnologie indossabili, o *wearable technologies*, saranno uno dei protagonisti indiscussi della medicina digitale, definita anche Salute 4.0. Si tratta di dispositivi connettibili alla rete, in grado di rilevare un'ampia varietà d'informazioni dal corpo umano, come attività fisica, temperatura, segni vitali – ECG, EEG –, funzionalità respiratorie, livelli di glucosio e di ossigeno nel sangue. L'elenco non è esaustivo e continua ad allungarsi. Con strumenti che vanno oltre ogni fantasia: esiste una pillola-sensore dotata di tecnologia wireless in grado di inviare al medico informazioni sui farmaci ingeriti da un paziente e quindi sull'aderenza alla terapia.
Il mondo dei dispositivi di monitoraggio vede la diabetologia in una posizione di vantaggio rispetto ad altre aree terapeutiche: glucometri e microinfusori rappresentano da molto tempo una realtà ben conosciuta per il team diabetologico, come anche le implicazioni organizzative e informatiche legate alla necessità di condividere i dati dei pazienti in un'ottica di continuità assistenziale.
L'ampliamento della sensoristica personale porterà, da un lato, a un aumento esponenziale dei dati generati, ma consentirà anche di ricavare informazioni sempre più precise e dettagliate, migliorando la capacità di prevenzione e personalizzazione delle cure.

L'intelligenza artificiale è l'espressione massima delle capacità predittive: impara dal passato, allertandoci su ciò che potrebbe succedere in futuro

L'intelligenza artificiale (IA) studia i fondamenti teorici, le metodologie e le tecniche informatiche che consentono ai computer di svolgere funzioni e ragionamenti tipici della mente umana. Riguarda la capacità di estrapolare, da informazioni riferite al passato, delle regole da utilizzare per risolvere problemi nuovi, anche quelli che il sistema non ha mai affrontato, sebbene possa averne incontrati di simili in passato.
Una delle aree di maggior interesse dell'IA in medicina è il Machine Learning [1]. I software che lo implementano sono in grado, con specifici algoritmi, di individuare dei "pattern" che caratterizzano i dati analizzati. La macchina identifica delle correlazioni fra i dati e, in base a esse, riesce a esprimere delle "predizioni", con una logica di tipo induttivo: parte dai singoli casi per stabilire una regola generale, indicando anche la probabilità con cui quella regola possa verificarsi. È un campo in forte evoluzione, che genera anche un po' di diffidenza per il funzionamento "a scatola chiusa" di molti motori predittivi che intendono sostituirsi al giudizio umano [2]. Esistono però soluzioni di nuova generazione, come gli algoritmi di *Logic Learning Machine* che non cercano di rimpiazzare il giudizio umano, bensì di potenziarlo, creando regole "intelligibili" che possono essere valutate da esperti del settore prima di essere attuate [3].
Un esempio di applicazione del Machine Learning in medicina può essere quello di un'analisi su dati epidemiologici contenenti, oltre a un numero imprecisato d'informazioni sugli individui di una certa popolazione, l'indicazione sulla presenza di obesità e diabete. Il software potrebbe riscontrare questo pattern: l'80% dei pazienti obesi è anche affetto da diabete tipo 2 (T2D), desumendo così una regola per cui, per un paziente obeso vi è l'80% di probabilità di ammalarsi. In uno studio sono stati analizzati i dati contenuti in migliaia di cartelle elettroniche di pazienti per identificare nuovi fattori di rischio (Fig. 1) [4].
La capacità di ricavare "nuova conoscenza" dalle correlazioni presenti nei dati assume una valenza determinante quando si devono analizzare database con milioni di record e centinaia di variabili.

Biologia dei sistemi e scienze "omiche": un nuovo paradigma nello studio degli organismi viventi

La biologia dei sistemi, che solo fino a pochi anni fa era un concetto abbastanza vago, sembra rappresentare il nuovo paradigma nello studio del funzionamento degli organismi viventi. È una disciplina che analizza le funzioni biologiche e i meccanismi che regolano le dinamiche delle reti intra- e intercellulari, facendo ricorso a competenze e strumenti ingegneristici e informatici.
La biologia dei sistemi comprende tutte le scienze definite "omiche", le quali, analizzando gruppi di molecole biologiche – come ioni, acidi nucleici, proteine ed enzimi – provenienti da campioni biologici, studiano con una visione d'insieme: i geni e le loro funzioni (con la genomica); i trascritti del DNA, ovvero l'RNA (con la trascrittomica); le proteine (con la proteomica) e i metaboliti

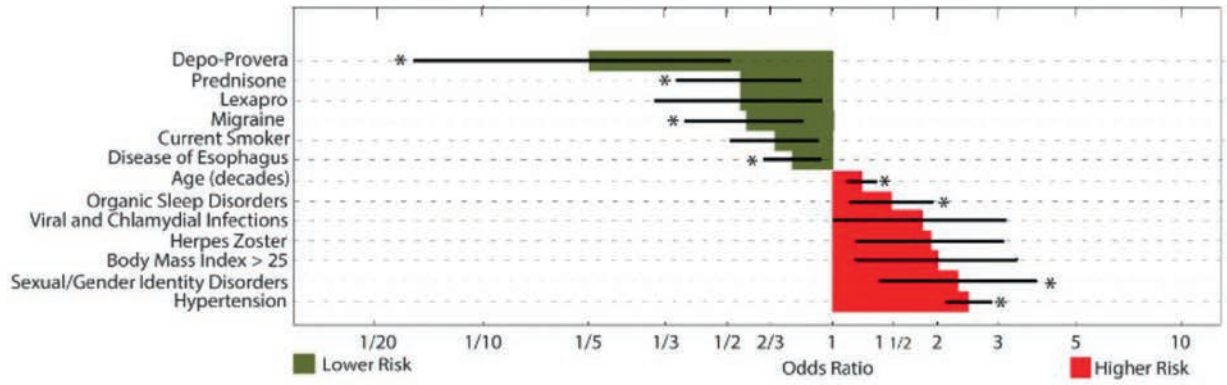

Figura 1.
Il *Machine Learning* applicato all'analisi delle cartelle elettroniche può aiutare a identificare fattori di rischio o elementi predittivi per il diabete tipo 2 (da http://www.scpr.org/news/2016/02/16/57470/new-research-finds-suprising-type-2-diabetes-links/).

presenti nell'organismo (con la metabolomica). È una visione olistica in base alla quale si cerca di comprendere, con un approccio integrato, principi operativi di livello più elevato, che nel loro insieme definiscono la biologia dei sistemi. L'obiettivo è di trovare risposta a domande biologiche con gerarchie complesse (come per esempio: patogenesi, storia naturale e prognosi di una malattia). Vengono utilizzate tecniche definite "ad alto rendimento" (*high throughput*: generano grandi moli di dati) come l'analisi genetica comparativa (basata su Microarray) o di sequenziamento del DNA (*Next Generation Sequencing techniques*, NGS) o strumenti computazionali che possono analizzare dati di migliaia di molecole/campioni. Per avere un'idea della quantità di dati elaborati, si pensi che un singolo genoma umano contiene circa 3,2 miliardi di coppie di basi azotate.
La confluenza fra l'evoluzione delle biotecnologie e lo sviluppo informatico, oltre a creare un fulcro che ha generato innumerevoli scoperte in ambito medico, ha anche determinato un nuovo modo di fare ricerca, in cui c'è una fortissima integrazione fra le competenze di biologi molecolari, bioinformatici, ingegneri e medici specialisti del campo in esame [5].
Nell'area della diabetologia, con queste nuove tecniche, si è evidenziato come i complessi processi fisiopatologici alla base dei T1D (diabete tipo 1), T2D e GDM siano causati da perturbazioni nell'espressione genica, che portano ad alterazioni dei processi fisiologici dei tessuti coinvolti nell'omeostasi del glucosio. Anche l'interazione fra fattori ambientali e varianti genetiche ed epigenetiche sta rivelando nuovi ambiti d'indagine. La complessità del sistema è acuita dal fatto che il contributo relativo di ciascun componente, rispetto all'espressione genica collegata alla patogenesi del diabete, è fortemente individuale: la comprensione dei meccanismi molecolari sottostanti a queste interazioni è cruciale per poter sviluppare nuove strategie di prevenzione e cura personalizzate. Con questi presupposti si stanno moltiplicando gli studi che, con l'ausilio di strumenti di analisi predittiva e *Machine Learning*, cercano di tracciare gli eventi molecolari attraverso "strati" d'informazioni biologiche, per risolvere il complesso puzzle dell'eziologia del diabete (Fig. 2) [6].

Nuove frontiere nella prevenzione e nella gestione del diabete

Diabete e genomica

Le nuove tecniche di analisi del genoma e delle interazioni nelle espressioni geniche stanno delineando con sempre maggior precisione i contributi dei geni nella patogenesi del diabete.
Nel T1D, a oggi, sono state identificate più di 57 regioni, o loci di suscettibilità, associate alla malattia e, in un nuovo studio in fase di prepubblicazione, si suggerisce la presenza di 5 nuove regioni, oltre a 17 nuove varianti (SNPs: polimorfismi a singolo nucleotide) a carico delle 57 regioni già censite [7].
Per il T2D, fino a poco tempo fa, si conoscevano 76 loci. Era stato anche rilevato, però, che quasi nessuna di queste variazioni era presente fra gli afroamericani, fra i quali l'incidenza del T2D è quasi doppia rispetto agli euroamericani. Poi, in un recentissimo studio, sono stati mappati ben 111 loci aggiuntivi, di cui 93 sono comuni agli afroamericani e agli europei, e 18 sono specifici degli europei [8].
Altri studi, basati sulla tecnica GWAS (*Genome Wide Association Study*, ovvero studi di associazione sull'intero genoma) hanno individuato dei nuovi fattori di rischio che sembrano essere coinvolti nella patofisiologia delle complicanze micro-

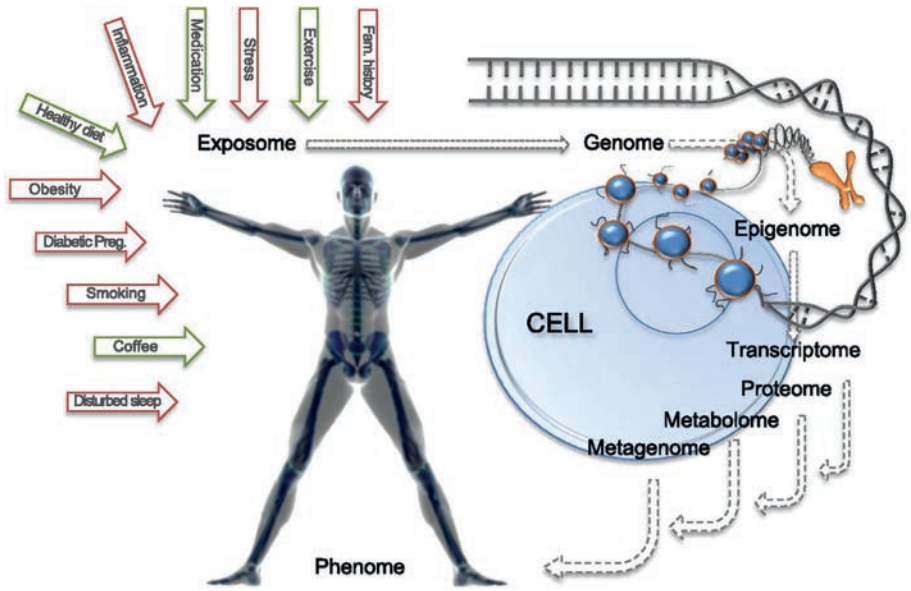

Figura 2.
Il futuro della ricerca sistemica sul diabete, alla scoperta delle interazioni fra l'esposoma (i fattori non genetici a cui l'individuo è esposto) e gli elementi quantificabili del fisioma umano (da http://care.diabetesjournals.org/content/36/5/1413) [6].

vascolari. È stata inoltre individuata una mutazione genetica che presenta un effetto protettivo nei confronti di retinopatia e insufficienza renale.

Diabete e trascrittomica

È stato analizzato il trascrittoma di pazienti diabetici (T1D, T2D, e GDM) per individuare le firme di espressione genica di ciascun tipo di diabete. I risultati hanno evidenziato che i profili di espressione genetica sono specifici per ciascuna delle tre patologie, anche se il profilo del GDM è molto più vicino a quello del T1D rispetto a quello del T2D.
Inoltre, osservando l'influenza dei geni coinvolti in varie funzioni biologiche, si è osservato che i geni indotti possono essere raggruppati in 5 macro-categorie di funzioni (sviluppo degli organismi multicellulari, trasduzione del segnale, risposta allo stress, processi di differenziazione delle cellule, processi del sistema immunitario), mentre i geni repressi fanno riferimento a 3 categorie di processi biologici (regolazione dei processi metabolici, processi di biosintesi, processi di trascrizione) [9].
Uno studio dell'*American Diabetes Association* integra gli aspetti di trascrittomica ed epigenomica per comprendere meglio la patogenesi del T2D col presupposto che nella vita pre- e post-natale i fattori ambientali (e anche l'alimentazione) interagiscono, nel microambiente di ciascun tessuto, con l'architettura di regolazione genetica, modulando così l'espressione genica (Fig. 3)[10].

Proteomica

Il T1D e il T2D sono patologie multifattoriali complesse collegate ad alterazioni di molti geni e dei relativi prodotti. Ma non tutte le alterazioni trascrizionali inducono variazioni proteiche: è quindi di fondamentale importanza riuscire a decodificare, oltre che l'espressione del mRNA, le variazioni delle proteine cellulari.
Nel 2013 è stato lanciato dalla HUPO (*Human Proteome Organization*) il programma *The Human Diabetes Proteome Project* (HDPP), gestito da un consorzio che intende valorizzare il mix di risorse e competenze utili a identificare le proteine e le isoforme proteiche (forme diverse della stessa proteina che possono essere causate da SNPs, polimorfismi a singolo nucleotide) associate ai meccanismi di azione nei vari tipi di diabete. Uno dei principali obiettivi è quello di mettere a fattor comune tutti i dati in una rete condivisa dalla comunità scientifica. HDPP si è focalizzato inizialmente sulle isole di Langerhans, ponendosi anche l'obiettivo di indagare, in una fase successiva, gli altri tessuti target in cui il glucosio e i lipidi possono causare disfunzioni delle proteine. Nell'ambito di queste iniziative ci si avvarrà di nuovi strumenti di bioinformatica per scoprire gli effetti dei lipidi e del glucosio nell'insorgenza e nella progressione del diabete.
Uno dei "prodotti" di questo progetto è un database, HDPP-1000, contenente l'elenco di più di 1.000 proteine coinvolte nell'eziologia del diabete.

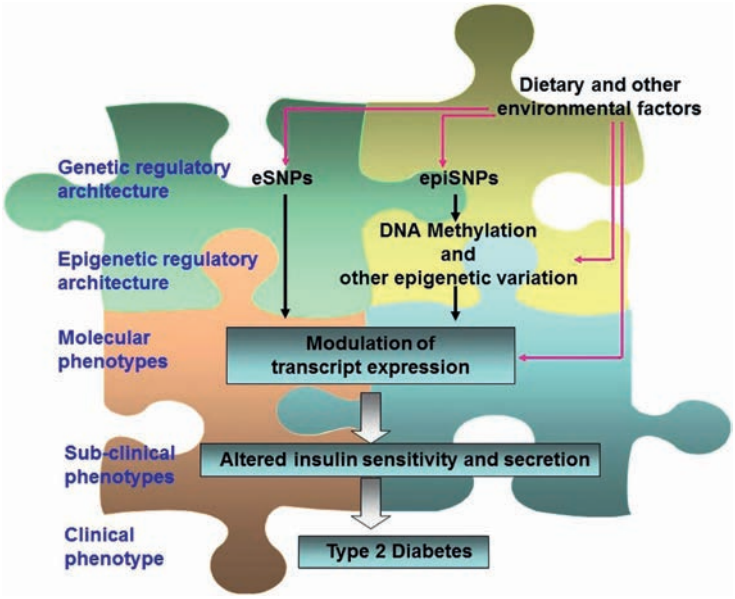

Figura 3.
Un modello causale della patogenesi del diabete 2 (da http://diabetes.diabetesjournals.org/content/63/9/2901) [10].

Metabolomica

In una recente meta-analisi [11] si sono analizzate 46 pubblicazioni (27 studi trasversali e 19 studi prospettici) che evidenziano associazioni fra metaboliti e le condizioni di prediabete e T2D.
Dagli studi esaminati è risultato che negli individui con T2D vi sono livelli alterati nei metaboliti di: carboidrati (glucosio e fruttosio), lipidi (fosfolipidi, sfingomielina e trigliceridi) e amminoacidi (amminoacidi ramificati, amminoacidi aromatici, glicina e glutammina).
Gli studi prospettici hanno suggerito che la concentrazione sanguigna di diversi metaboliti (esosi, amminoacidi ramificati e aromatici, fosfolipidi e trigliceridi) è associata all'incidenza di prediabete e di T2D. Maggiori concentrazioni di isoleucina, leucina, valina, tirosina e fenilalanina sono associate a un maggior rischio di T2D. Glicina e glutammina presentano invece un'associazione inversa: alla patologia vengono collegate minori concentrazioni di queste sostanze.
La meta-analisi conclude che l'utilizzo di tecnologie *high-throughput* nell'analisi metabolomica ha consentito di affermare che particolari concentrazioni di diversi tipi di amminoacidi presenti nel sangue sono associabili al rischio di sviluppare il T2D.

Bibliografia

1 http://www.corriere.it/video-articoli/2017/03/29/che-cos-machine-learningcosi-computer-diventanto-intelligenti/586d16f8-148d-11e7-a7c3-077037ca4143.shtml

2 https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/.

3 http://www.rulex.ai/technology/white-papers/.

4 http://www.scpr.org/news/2016/02/16/57470/new-research-finds-suprising-type-2-diabetes-links/.

5 http://www.biochronicles.net/news/tecnologie-omiche-e-integrazione-dei-dati/.

6 Franks PW, Pearson E, Florez JC. *Gene-Environment and Gene-Treatment Interactions in Type 2 Diabetes. Progress, pitfalls, and prospects*. Diabetes Care 2013;36:1413-21.

7 Cooper NJ, Wallace C, Burren OS, Cutler S, et al. *Type 1 diabetes genome-wide association analysis with imputation identifies five new risk regions*. bioRxiv preprint first posted online Mar. 24, 2017. doi: http://dx.doi.org/10.1101/120022.

8 Lau W, Andrew T, Maniatis N. *High-Resolution Genetic Maps Identify Multiple Type 2 Diabetes Loci at Regulatory Hotspots in African Americans and Europeans*. Am J Hum Genet 2017;100:803-16.

9 Passos GA, editor. *Transcriptomics in health and disease*. Springer 20145, p. 141.

10 Das SK. *Integrating transcriptome and epigenome: putting together the pieces of the type 2 diabetes pathogenesis puzzle*. Diabetes 2014;63:2901-3.

11 Guash-Ferrè M, Hruby A, Toledo E, et al. *Metabolomics in prediabetes and diabetes: a systematic review and meta-analysis*. Diabetes Care 2016;39:833-46.

SEZIONE DI AUTOVALUTAZIONE

NGS, microarray e GWA sono:

- ☐ tecnologie di *predictive analytics* che applicano l'intelligenza artificiale alla genomica
- ☐ bioecnologie definite *high throughput* perché generano grossi volumi di dati, successivamente analizzati con software di bioinformatica
- ☐ strumenti di *machine learning* in grado di mappare il genoma
- ☐ tecniche di analisi delle variazioni genetiche basate su sistemi bioinformatici definiti *high throughput*

Lo studio della proteomica nel diabete indaga:

- ☐ l'espressione del mRNA nelle varie forme di diabete
- ☐ quali, fra le possibili alterazioni trascrizionali collegate alla patologia, inducano anche variazioni proteiche
- ☐ le firme di espressione genica nei vari tipi di diabete
- ☐ in che modo il glucosio e i lipidi possano causare disfunzioni delle proteine

La metabolomica applicata allo studio del diabete:

- ☐ suggerisce che nel prediabete e nel T2D i metaboliti di carboidrati e amminoacidi sono sempre più elevati rispetto ai gruppi di controllo
- ☐ cerca di individuare i metaboliti dei carboidrati collegati alla patologia diabetica
- ☐ cerca di capire se le alterazioni del metabolismo dei carboidrati rappresentino un fattore di rischio nel T2D
- ☐ cerca di capire il collegamento fra il diabete e determinate alterazioni nei livelli di specifici metaboliti

L'analisi di tipo predittivo:

- ☐ si basa su strumenti biotecnologici che comprendono le scienze "omiche"
- ☐ si basa su strumenti informatici in grado di individuare i fattori di rischio di una patologia
- ☐ si basa su strumenti in grado di individuare i collegamenti fra diversi fattori, indicando anche la probabilità con cui quei fattori possono presentarsi congiuntamente
- ☐ si basa su strumenti in grado di simulare come, modificando opportunamente alcuni fattori, si può migliorare l'output di un sistema

Verifica subito le risposte on line **www.diabete-rivistamedia.it**

PACINI EDITORE MEDICINA