

19



OFICINA ESPAÑOLA DE
PATENTES Y MARCAS

ESPAÑA



11 Número de publicación: **2 996 267**

51 Int. Cl.:

G16H 20/17

(2008.01)

12

TRADUCCIÓN DE PATENTE EUROPEA

T3

86 Fecha de presentación y número de la solicitud internacional: **14.07.2017** **PCT/IB2017/054283**

87 Fecha y número de publicación internacional: **18.01.2018** **WO18011766**

96 Fecha de presentación y número de la solicitud europea: **14.07.2017** **E 17754790 (8)**

97 Fecha y número de publicación de la concesión europea: **06.11.2024** **EP 3485409**

54 Título: **Estimación de la insulina basada en aprendizaje por refuerzo**

30 Prioridad:

15.07.2016 WO PCT/IB2016/054242

45 Fecha de publicación y mención en BOPI de la traducción de la patente:

12.02.2025

73 Titular/es:

**UNIVERSITÄT BERN (100.00%)
Verwaltungsdirektion Hochschulstrasse 6
3012 Berne, CH**

72 Inventor/es:

**MOUGIAKAKOU, STAVROULA;
DASKALAKI, ELENI y
DIEM, PETER**

74 Agente/Representante:

ELZABURU, S.L.P

ES 2 996 267 T3

Aviso: En el plazo de nueve meses a contar desde la fecha de publicación en el Boletín Europeo de Patentes, de la mención de concesión de la patente europea, cualquier persona podrá oponerse ante la Oficina Europea de Patentes a la patente concedida. La oposición deberá formularse por escrito y estar motivada; sólo se considerará como formulada una vez que se haya realizado el pago de la tasa de oposición (art. 99.1 del Convenio sobre Concesión de Patentes Europeas).

DESCRIPCIÓN

Estimación de la insulina basada en aprendizaje por refuerzo

Campo técnico

La invención se refiere al campo de la administración de insulina y a un páncreas artificial correspondiente.

5 Antecedentes de la invención

La diabetes de tipo 1 (T1D) es una enfermedad metabólica caracterizada por niveles de glucosa en sangre no controlados, debido a la ausencia o a un mal funcionamiento de la insulina. El sistema de Páncreas Artificial (AP, por sus siglas en inglés) tiene como objetivo simular la función del páncreas fisiológico y servir como un sistema de regulación automático externo de la glucosa. El AP combina un monitor continuo de glucosa (CGM, por sus siglas en inglés), una bomba de infusión subcutánea continua de insulina (CSII, por sus siglas en inglés) y un algoritmo de control que cierra el bucle entre los dos dispositivos y optimiza la tasa de infusión de insulina.

Un reto importante en el diseño de algoritmos de control eficientes para el AP es el uso de la vía subcutánea tanto para la medición subcutánea de glucosa como para la infusión subcutánea de insulina; esto introduce retrasos de hasta 30 minutos para la medición subcutánea de glucosa y hasta 20 minutos para la absorción de insulina. Por lo tanto, un retardo total de casi una hora restringe tanto la monitorización como la intervención en tiempo real. Además, la glucosa se ve afectada por múltiples factores, que pueden ser genéticos, estilo de vida y ambientales. Con la mejora en la tecnología de sensores, se puede proporcionar más información al algoritmo de control (por ejemplo, lecturas de glucosa y niveles de actividad física más precisos); sin embargo, el nivel de incertidumbre sigue siendo muy alto. Por último, pero no menos importante, uno de los desafíos más importantes surge de la alta variabilidad inter- e intra-paciente, que impone un tratamiento personalizado con insulina.

Junto con las mejoras de los programas informáticos, los retos del AP se están abordando gradualmente con el desarrollo de estrategias algorítmicas avanzadas; las estrategias más investigadas clínicamente son los algoritmos de Derivada Integral Proporcional (PID, por sus siglas en inglés), Controlador Predictivo de Modelo (MPC, por sus siglas en inglés) y lógica difusa (por ejemplo, la lógica MD). Un desarrollo reciente ha sido el AP bihormonal, que emplea tanto insulina como glucagón. El número creciente de ensayos clínicos ha llevado a una extensa evaluación en el ámbito hospitalario y, más recientemente en el hogar, de la viabilidad del AP fuera de un entorno hospitalario controlado. La mayoría de los estudios se restringen a la evaluación algorítmica de una cohorte de pacientes en condiciones inciertas, tales como ingesta errónea de comida y cambios de sensibilidad frente a la insulina (SI, por sus siglas en inglés) (por ejemplo, actividad física).

A pesar de estos resultados prometedores, ninguna de las estrategias de control propuestas actualmente está intrínsecamente diseñada para manejar incertidumbres y una personalización. La PID está diseñada para sistemas lineales, el MPC resuelve un problema de optimización en bucle abierto que ha demostrado ser subóptimo en presencia de incertidumbre y la lógica MD es un enfoque basado en reglas, sometido directamente a la experiencia del diseñador. De cara a la variabilidad entre pacientes, los algoritmos se han mejorado con componentes adaptativos, que se basan principalmente en la identificación personalizada de modelos implicados o la correlación de parámetros algorítmicos con una o múltiples características específicas del paciente, tales como el peso corporal, el factor de corrección o la SI. Sin embargo, el rendimiento exitoso de los algoritmos de AP del estado de la técnica prueba que el desarrollo de AP es factible y viable y allana el camino para una nueva era de investigación algorítmica más avanzada hacia un tratamiento potente y personalizado con insulina.

El aprendizaje por refuerzo (RL) es una rama del aprendizaje automático (ML) y es un campo de investigación intensamente activo que abarca algoritmos que son capaces de aprender a partir de datos y realizar una optimización dentro de entornos inciertos. El campo del RL se sitúa entre el aprendizaje supervisado y no supervisado e incluye problemas en los que un agente intenta mejorar su rendimiento en una tarea dada a lo largo del tiempo mediante una interacción continua con su entorno. Durante ese esfuerzo, no se proporcionan explícitamente datos de entrada-salida, y la mejora del rendimiento se traduce en minimizar alguna noción de coste esperado a largo plazo sobre la trayectoria del agente. El RL comenzó a desarrollarse como una rama independiente a principios de los años 1980 y se inspiraba en la psicología animal y la idea de aprender a través de prueba y error. Fue rápidamente adoptado por el campo del control óptimo como una manera muy eficiente de resolver problemas de programación dinámica para los cuales la "evolución de la dimensionalidad" de Bellman restringía una solución analítica. En el RL, el agente sigue una cierta política de control en base a la cual elige acciones u y se mueve desde el estado actual x al siguiente. Cada transición de estado va seguida de un coste local c que indica la calidad de esa transición basándose en algunos criterios predefinidos. El objetivo del agente es encontrar una política de control óptima para minimizar el coste esperado total a lo largo de su trayectoria. El RL es un campo con antecedentes teóricos ampliamente investigados, que ahora está encontrando su camino hacia la aplicación práctica, debido a los avances modernos en la capacidad computacional. En este contexto, la aplicación a problemas de la vida real sobresale como una de las tendencias actuales del RL. En medicina, el RL se investiga principalmente para pronóstico, clasificación y diagnóstico por medio de una recogida, fusión y análisis de datos grandes/heterogéneos, con menos informes sobre estudios de tratamiento. Las ventajas de ML y RL ilustran una trayectoria prometedora hacia la resolución de los desafíos del AP, como se ha reconocido e

informado recientemente. Para evaluar el algoritmo, se usó un modelo glucorregulador determinista, pero aumentado con incertidumbre para simular la variabilidad entre pacientes. Este procedimiento puede no ser representativo de la variabilidad real entre pacientes y limita la solidez de la capacidad adaptativa presentada.

5 Un factor que complica el uso del RL en medicina es el alto número de parámetros constantes y adaptativos que necesitan ser ajustados o inicializados. Elegir los valores óptimos para esos parámetros es una tarea desafiante y normalmente se realiza manualmente, basándose en características específicas del problema. Sin embargo, ante la variabilidad interindividual, ese proceso manual puede ser poco fiable o incluso inviable. Además, otra crítica del RL es la dificultad para una generalización o explicación cualitativa tanto del proceso de aprendizaje como de la solución final (caja negra).

10 En el presente enfoque, se propone un algoritmo basado en el RL para la infusión personalizada de insulina y la regulación de la glucosa en T1D. Se desarrolla y evalúa *in silico* un algoritmo Actor-Crítico (AC) sin modelo para estudiar su capacidad para mantener la normoglucemia dentro de una gran cohorte de pacientes y bajo incertidumbres ambientales variables. El alcance del enfoque es doble:

i) investigar la aplicabilidad del RL en el contexto de un AP personalizado y

15 ii) conseguir un diseño de AC que pueda generalizarse y traducirse directamente en experiencia médica.

Con el fin de superar la restricción del ajuste analizada anteriormente, el algoritmo AC se mejora con un método novedoso para el ajuste automático y personalizado, basado en la estimación de la transferencia de información (IT, por sus siglas en inglés) de las señales de insulina a glucosa.

El trabajo en fase temprana en algoritmos de AC ya se ha presentado, junto con resultados de evaluación preliminares:

20 • DASKALAKI ELENA ET AL: "An Actor-Critic based controller for glucose regulation in type 1 diabetes", COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE, ELSEVIER, AMSTERDAM, NL, vol. 109, no. 2, 12 de abril de 2012 (2012-04-12), páginas 116-125, XP028971483, ISSN: 0169-2607, DOI: 10.1016/J.CMPB.2012.03.002.

25 • Daskalaki E, Diem P, Mougiakakou S. Personalized tuning of a reinforcement learning control algorithm for glucose regulation, Proceedings of the 35th IEEE Eng Med Biol Soc Conference; 2013 3-7 de julio; Osaka, Japón.

30 • "Mariano De Paula, Luis Omar Ávila, Ernesto C. Martínez, Controlling blood glucose variability under uncertainly using reinforcement learning and Gaussian processes, Applied Soft Computing, Volumen 35, 2015, páginas 310-332, ISSN 1568-4946" describen un enfoque novedoso para controlar la variabilidad en la concentración de glucosa en sangre usando un aprendizaje basado en simulación de una política de control fuerte. Para tener en cuenta la variabilidad entre pacientes, se propone un algoritmo de iteración de políticas que combina el aprendizaje por refuerzo con la aproximación de proceso gaussiano. Se aprende fuera de línea una política de control genérica usando un modelo de proceso estocástico de Ito de la dinámica de glucosa-insulina que simula la variabilidad glucémica exhaustivamente. Para la seguridad y el rendimiento, solo se muestrean datos relevantes a través del aprendizaje activo bayesiano.

En la presente invención, se introduce un diseño novedoso para el controlador de AC. Específicamente:

- La política de control ha aumentado mediante una política exploratoria, con el fin de incrementar el espacio de búsqueda del algoritmo.
- Se ha incorporado una política de control supervisor para mejorar la seguridad del algoritmo.
- 40 • El controlador de AC está directamente relacionado con parámetros fisiológicos y/o acciones extraídas de la experiencia médica.
- El método de ajuste automático se ha extendido y asociado a características específicas del paciente.

En resumen, el esquema de control novedoso propuesto es capaz de afrontar los siguientes desafíos:

- 45 • Variabilidad inter-/intra-paciente y personalización del tratamiento con insulina a través del uso de un algoritmo de aprendizaje adaptativo en tiempo real
- Solidez del uso de un algoritmo de control que es adecuado para una optimización bajo incertidumbre
- Fácil transferencia a la puesta en práctica en hospitales y en casa ya que
 - se basa en suposiciones a priori limitadas que contrarrestan la alta variabilidad entre pacientes

- se inicializa basándose en parámetros fisiológicos

Compendio de la invención

La presente invención describe un método tal y como se define en las reivindicaciones 1 a 17.

La invención proporciona un método para estimar una terapia óptima, por ejemplo, la insulina óptima que se va a administrar mediante un dispositivo de administración de insulina, tal como una bomba de infusión. La terapia puede determinarse usando un algoritmo de aprendizaje por refuerzo que tiene como objetivo la regulación personalizada de la glucosa. El algoritmo puede optimizar la tasa basal diaria de insulina y la relación de insulina a carbohidrato para cada paciente, basándose (por ejemplo) en su perfil de glucosa medida. Ese algoritmo puede ser capaz de aprender en tiempo real características específicas del paciente, capturadas en el perfil diario de glucosa y proporcionar un tratamiento individualizado con insulina. Un método de ajuste automático y personalizado contribuye a la optimización del rendimiento del algoritmo. Vale la pena señalar que el diseño del algoritmo puede permitir un uso directo a través del paciente sin ninguna intervención o etapas intermedias que impliquen médicos o técnicos.

Por ejemplo, durante los cuatro primeros días operativos, el sistema proporciona el tratamiento estándar del paciente tal y como lo ha definido su médico y, en paralelo, puede recopilar su monitor de glucosa continuo y los datos de la bomba de insulina. Al final de ese período, el algoritmo inicializa automáticamente diversos parámetros. A continuación, el algoritmo puede continuar la personalización del tratamiento con insulina con una adaptación diaria de la tasa basal diaria de insulina y la relación insulina-carbohidrato. Preferentemente, para todos los cálculos implicados, el algoritmo necesita un tiempo computacional mínimo y puede ejecutarse sin contratiempos en un dispositivo móvil/portátil.

Uno de los aspectos de la invención es un método para determinar una terapia con insulina para un paciente. Ese método puede comprender las etapas siguientes:

- utilizar un algoritmo de aprendizaje por refuerzo para manejar un controlador adaptativo, comprendiendo el algoritmo de aprendizaje por refuerzo:
 - a. un crítico que evalúa una política de control de insulina (S) que incluye al menos uno entre una tasa de infusión de insulina y una relación de insulina a carbohidrato, y
 - b. un actor que mejora la política de control de insulina (S);
- aplicar, durante un periodo de tiempo predeterminado ($k-1$), una política de control de insulina predeterminada (S_{k-1});
- recopilar, durante el período de tiempo predeterminado ($k-1$), datos o el perfil de glucemia del paciente;
- determinar un estado del sistema regulador de glucosa basado, por ejemplo, en los datos de glucosa en sangre recogidos;
- evaluar a través del crítico la política de control de insulina predeterminada (S_{k-1}) basándose, por ejemplo, en el estado del sistema regulador de glucosa;
- actualizar por medio del actor la política de control de insulina (S), basándose, por ejemplo, en una acción de control determinada, en donde la política de control de insulina actualizada (S_k) está previsto que se aplique durante un periodo de tiempo posterior (k);

La acción de control determinada tiene en cuenta al menos una de las siguientes acciones de control:

- una acción de control determinista lineal (P_a);
- una acción de control supervisora (P_s) y
- una acción de control exploratoria (P_e).

La acción de control determinista lineal (P_a) puede definirse como una combinación lineal del estado del sistema regulador de glucosa y un vector de parámetros de política determinado. La acción de control supervisora (P_s) puede corresponder a una sugerencia conservadora de una política previa de control de insulina. Puede actuar como módulo de seguridad frente a cambios extremos de la política de control de insulina (S) por el algoritmo y/o como guía de la dirección de la actualización. La acción de control exploratoria (P_e) puede permitir ampliar el espacio de búsqueda del estado del sistema regulador de glucosa y/o respaldar una convergencia más rápida.

Preferentemente, la etapa de evaluación se realiza al final de cada período de tiempo y el vector de parámetros de política puede actualizarse al final del período de tiempo determinado. Además, la duración del período de tiempo predeterminado y/o del periodo de tiempo posterior puede estar comprendida entre 1 hora y 36 horas, preferentemente sustancialmente igual a 24 horas.

La acción de control determinada puede definirse como una tasa del cambio de la política de control de insulina (S) desde el período de tiempo predeterminado (k-1) hasta el período de tiempo posterior (k).

La acción de control determinada puede tener en cuenta al menos dos entre la acción de control determinista lineal (P_a), la acción de control supervisora (P_s) y la acción de control exploratoria (P_e).

- 5 La acción de control determinada puede tener en cuenta una acción de control determinista total (P_d) que se determina por medio de un modelo matemático que tiene en cuenta la acción de control determinista lineal (P_a) y la acción de control supervisora (P_s). Preferentemente, la acción de control exploratoria (P_e) se determina por medio de un modelo matemático que tiene en cuenta al menos la acción de control determinista total (P_d). El vector de parámetros de política puede depender de la diferencia entre la acción de control determinista total (P_d) y la acción de control exploratoria (P_e).

10 El crítico puede evaluar el estado del sistema regulador de glucosa tal y como se ha descrito mediante dos características relacionadas con el estado hiper- e hipoglucémico ocurrido durante el periodo de tiempo predeterminado (k-1). Además, el crítico puede evaluar el estado del sistema regulador de glucosa como se ha descrito por la hipoglucemia diaria promedio y el error de hiperglucemia ocurrido durante el período de tiempo predeterminado (k-1).

15 Preferentemente, el crítico calcula un coste local del estado del sistema regulador de glucosa durante el periodo de tiempo predeterminado. Y el actor se puede configurar para minimizar el coste esperado promedio para cada estado.

20 Una política inicial de control de insulina puede determinarse teniendo en cuenta los valores individuales del paciente optimizados por su médico y/o un simulador *in silico*. El valor o el valor inicial del vector de parámetros de política puede determinarse usando un cálculo de entropía de transferencia entre la insulina y la glucosa. Además, la entropía de transferencia entre la insulina y la glucosa puede estimarse basándose en los datos de glucosa en sangre y los datos de administración de insulina durante un periodo de tiempo comprendido entre 1 y 7 días, preferentemente 4 días.

25 El método puede comprender una fase de inicialización durante la cual una política de control de insulina inicial y/o un valor inicial del vector de parámetros de política se pueden determinar y adaptar al paciente. Y la duración de la fase de inicialización puede estar comprendida entre 1 y 7 días, preferentemente 4 días.

Breve descripción de los dibujos

La invención se comprenderá mejor de cara a la descripción de las realizaciones preferidas y haciendo referencia a las figuras, en donde

- 30 La Fig. 1 ilustra el sistema, páncreas artificial, en el que el sistema de control para la bomba o la bomba de parche está integrado en un dispositivo portátil conectado inalámbricamente tanto con el o los monitores continuos de glucosa como con la bomba de insulina para la transmisión de datos en tiempo real;
- la Fig. 2 es un diagrama de bloques que ilustra la implementación general del sistema de control actor-crítico de la Fig. 1 para la implementación del sistema de bucle cerrado según el estado de la técnica;
- 35 la Fig. 3 es un diagrama de bloques que ilustra las entradas/salidas del sistema de control actor-crítico de la Fig. 1 para la implementación del sistema de bucle cerrado según la invención;
- la Fig. 4 es un diagrama de bloques que ilustra una implementación del sistema de control actor-crítico de Fig. 1 para la implementación del sistema de bucle cerrado según la invención; y
- la Fig. 5 ilustra un ejemplo del proceso a seguir para ajustar el sistema actor-crítico según la invención.

40 Descripción detallada de las realizaciones preferidas

La invención se expone y se caracteriza en las reivindicaciones independientes, mientras que las reivindicaciones dependientes describen otras características de la invención.

La presente solicitud reivindica el beneficio de la prioridad del documento PCT/IB2016/054242 presentado el 15 de julio de 2016 a nombre de University Bern.

45 Todos los términos científicos y técnicos empleados en el presente documento tienen los significados usados comúnmente en la técnica a menos que se especifique lo contrario. Las definiciones proporcionadas en el presente documento son para facilitar la comprensión de ciertos términos usados frecuentemente en el presente documento y no pretenden limitar el alcance de la presente divulgación.

50 Tal y como se usa en esta memoria descriptiva y en las reivindicaciones adjuntas, las formas singulares "un", "una" y "el, ella" incluyen realizaciones que tienen referentes plurales, a menos que el contenido indique claramente lo contrario.

Tal y como se usa en esta memoria descriptiva y en las reivindicaciones adjuntas, cualquier dirección a la que se haga referencia en el presente documento, tal como "arriba", "abajo", "izquierda", "derecha", "superior", "inferior" y otras direcciones u orientaciones, se describe en el presente documento para mayor claridad haciendo referencia a las figuras y no pretende ser limitante de un dispositivo o un sistema real. Los dispositivos y sistemas descritos en el presente documento pueden usarse en varias direcciones y orientaciones.

Tal y como se usa en el presente documento, "tener", "que tiene", "incluir", "que incluye", "comprender", "que comprende" o similares, se usan en su sentido amplio, y generalmente significa "que incluye, pero no se limita a".

Tal y como se usa en esta memoria descriptiva y en las reivindicaciones adjuntas, el término "o" se emplea generalmente en su sentido que incluye "y/o" a menos que el contenido indique claramente lo contrario.

El sistema de control Actor-Crítico para cerrar el bucle entre monitores continuos de glucosa y bombas de infusión de insulina, tal y como se ilustra en la Fig. 1, puede ejecutarse en uno cualquiera de un dispositivo portátil/móvil y teléfono inteligente, y se presenta en los siguientes párrafos.

El algoritmo Actor-Crítico

El algoritmo AC pertenece a la clase de RL y se caracteriza por la separación del agente en dos partes complementarias mostradas en la Fig. 2: el crítico, que es responsable de evaluar la política de control y el actor, que es responsable de mejorar la política de control. Dentro de la familia de RL, los algoritmos AC difieren de los métodos de solo actor o solamente crítico en que poseen mejores propiedades de convergencia. Además, su coste computacional es menor, ya que estiman intrínsecamente gradientes de baja varianza y política de parametrización para permitir una optimización en tiempo continuo.

En el aprendizaje del AC, el agente sigue una política de control específica y realiza transiciones entre estados dentro de un entorno incierto. La Fig. 2 muestra una vista esquemática de un sistema controlado por un algoritmo AC.

En el caso de sistemas estocásticos, la política de control es una función de probabilidad condicional $\mu(u | x, \theta)$ a partir de la cual se extraen acciones de control u , dados los estados actuales de x . El objetivo del agente es encontrar una política de control óptima, con el fin de minimizar el coste esperado a lo largo de su trayectoria. Una transición entre los estados x e y depende de la acción de control elegida u y sigue una distribución de probabilidad de transición $p(y | x, u)$. Un coste local $c(x, u)$ está asociado con cada estado y acción. Preferentemente, en un ajuste de recompensa promedio, el objetivo del algoritmo AC es encontrar una política de control óptima para minimizar el coste esperado promedio para cada estado en todos los estados. Esto puede definirse como:

$$\bar{a}(\theta) = \sum_{x \in \mathcal{X}, u \in \mathcal{U}} c(x, u) \eta_{\theta}(x, u) \quad (1)$$

en donde $\eta_{\theta}(x, u)$ es la probabilidad estacionaria de la cadena de Markov $\{X_k, U_k\}$.

El crítico

El agente crítico es responsable de la evaluación de la política de control actual basándose en la aproximación de un coste esperado asociado. Uno de los métodos más potentes usados para este fin es el aprendizaje por diferencia temporal (TD), en donde el coste esperado total de un proceso que comienza en el estado x , toma como primera acción u y sigue la política $\mu(u | x, \theta)$ que puede definirse a través de las funciones de valor y de acción-valor $V_{\theta}(x)$ y $Q_{\theta}(x, u)$ respectivamente:

$$V_{\theta}(x) = E\left[\sum_{k=0}^{\infty} \gamma^k c(x_k, u_k) \mid x_0 = x\right] \quad (2)$$

$$Q_{\theta}(x, u) = E\left[\sum_{k=0}^{\infty} \gamma^k c(x_k, u_k) \mid x_0 = x, u_0 = u\right] \quad (3)$$

Las funciones de valor y de acción-valor pueden satisfacer las ecuaciones siguientes:

$$Q_{\theta}(x, u) = c(x, u) + \gamma \sum_y p(y | x, u) V_{\theta}(y) \quad (4)$$

$$V_{\theta}(x, u) = \sum_u \mu(u | x, \theta) \left[c(x, u) + \gamma \sum_y p(y | x, u) V_{\theta}(y) \right] \quad (\text{Ecuación de Bellman}) \quad (5)$$

Para los estados observados dados, $x = x_{k-1}$, $y = x_k$ y la acción $u = u_{k-1}$, y la ecuación de Bellman (5) puede reducirse a:

$$V_{\theta}(x) = c(x, u) + \gamma V_{\theta}(y) \quad (6)$$

La maldición de la dimensionalidad de Bellman restringe la solución analítica de (6) en espacios dimensionales altos y requiere el uso de métodos de aproximación. En el marco de TD, la función de valor $V(x)$ puede aproximarse mediante una función parametrizada $V_w(x)$ con $w \in R^K$. La arquitectura más comúnmente usada para la función parametrizada puede ser la aproximación lineal definida como:

$$\tilde{V}_\theta^w(x) = \sum_{i=1}^K w_i g_i(x) = w^T g_\theta(x) \quad (7)$$

en donde $g_\theta(x)$ es un vector de funciones base de dimensión K . La notación w^T indica transponer. La aproximación de la función de valor puede realizarse a través de la estimación del error de TD d , definida como la desviación de la función de valor aproximado $\tilde{V}^w(x)$ a partir de su estimación posterior $\tilde{V}^w(y)$:

$$d = c(x, u) + \gamma \tilde{V}_\theta^w(y) - \tilde{V}_\theta^w(x) \quad (8)$$

Basándose en el error de TD, el vector de parámetros w puede actualizarse según la fórmula:

$$w_{k+1} = w_k + \alpha_k d_k \sum_{n=0}^k \lambda^{k-n} \nabla_w \tilde{V}_\theta^w(x) = w_k + \alpha_k d_k z_k \quad (9)$$

en donde α_k es una secuencia de tasa de aprendizaje no creciente positiva, $0 < \lambda < 1$ es constante y z_k puede ser el vector de elegibilidad definido como:

$$z_k = \sum_{n=0}^k \lambda^{k-n} g_\theta(x_n) \quad (10)$$

y se puede actualizar según la fórmula siguiente:

$$z_{k+1} = \lambda z_k + g_\theta(x_{k+1}) \quad (11)$$

Para la aproximación de la función acción-valor puede seguirse un proceso similar $Q_\theta(x, u)$:

$$\tilde{Q}_\theta^r(x, u) = \sum_{i=1}^L r^i \phi_\theta^i(x, u) = r^T \phi_\theta(x, u) \quad (12)$$

en donde $\phi_\theta(x, u)$ es el vector de funciones base y $r \in R^L$ es el vector del parámetro respectivo. Una elección comúnmente usada de las funciones base es $\phi_\theta(x, u) = \psi_\theta(x, u)$, en donde $\psi_\theta(x, u) = \nabla_\theta \ln \mu(u | x, \theta)$ es la derivada de la relación de probabilidad de la política de control.

El actor

El objetivo del actor es optimizar la política de control a lo largo del tiempo hacia una minimización del coste esperado promedio para cada estado $\bar{a}(\theta)$. Para la minimización pueden utilizarse métodos de gradiente de política, que implican la estimación del gradiente $\nabla_\theta \bar{a}(\theta)$ con respecto al vector de parámetros de política θ . La función de actualización de política general puede tener la siguiente forma:

$$\theta_{k+1} = \theta_k - \beta_k \nabla_\theta \bar{a}(\theta) \quad (13)$$

en donde β_k es una secuencia positiva de las tasas de aprendizaje. Se pueden proponer diversas versiones de actor, distinguidas principalmente por la estrategia de aproximación para el gradiente de $\nabla_\theta \bar{a}(\theta)$. En el presente ejemplo, se puede utilizar el actor en donde:

$$\nabla_\theta \bar{a}(\theta) = \sum_{x, u} \eta_\theta(x, u) d | p_\theta(x, u) \quad (14)$$

Regulación de la glucosa en T1D basándose en un algoritmo AC

El algoritmo AC está diseñado para optimizar el régimen de insulina para cada paciente con T1D. El régimen de insulina puede definirse como la combinación de la tasa basal de insulina (BR) y la relación entre insulina-carbohidrato (IC) (Fig. 3). Esa elección se tomó para que fuera coherente con la práctica médica; sin embargo, se pueden usar otros perfiles de régimen de insulina. La relación IC se usa para el cálculo de la dosis en bolo (I_{bolus}) según el tamaño conocido de los carbohidratos (CHO) de la comida próxima como:

$$IC = I_{bolus} / CHO \quad (15)$$

Antes del diseño de los agentes críticos y actores, es necesario definir dos parámetros importantes del algoritmo:

- i) la ventana de aprendizaje, que se corresponde con la tasa de actualización del algoritmo, y
- ii) el estado del sistema.

Ambos se analizan en los siguientes párrafos.

Ventana de aprendizaje

- 5 La ventana de aprendizaje se define en este documento como el período de tiempo proporcionado para la recopilación de datos antes de una actualización del perfil de insulina. Hay varias consideraciones que influyen en esa decisión. La ventana de aprendizaje no puede ser comparable al retraso del bucle introducido por el CGM y la absorción subcutánea de insulina. Además, debe considerarse el compromiso entre un aprendizaje rápido y lento. Las actualizaciones frecuentes pueden realizar un seguimiento eficaz de la dinámica rápida de la glucosa, pero pueden perder la "visión general" que es portadora de la información más básica o genérica sobre las características del paciente. Teniendo esto en cuenta, este documento describe un ejemplo en el que la ventana de optimización es de un día (24 horas). Se entiende que "un día" se refiere a un período de tiempo de 24 horas que puede comenzar en cualquier momento del día o la noche, preferentemente el período de tiempo puede comenzar la medianoche de un día dado y puede terminar al final de ese día dado. Ese período de tiempo también puede ser más largo o más corto. La elección de ese "día 1" también considera el ciclo de 24 horas del cuerpo humano, que es portador de información adecuada sobre el estado glicémico general del paciente. Como resultado, la política de insulina se evalúa y se actualiza una vez al día basándose en el perfil de glucosa diario respectivo (Figs. 3 y 4).

Estado del sistema

- 20 La dinámica del sistema glucorregulador se representa como un proceso de decisión de Markov, en donde el estado x_k es el estado del sistema en términos de hipo- e hiperglucemia para el día k . El error de glucosa EG en cada momento t puede definirse como:

$$EG(t) = \begin{cases} G(t) - G_h & \text{si } G > G_h \\ G(t) - G_l & \text{si } G < G_l \\ 0 & \text{si no} \end{cases} \quad (16)$$

- 25 en donde $G(t)$ es el valor de glucosa en el momento t y $G_h = 180 \text{ mg / dl}$, $G_l = 70 \text{ mg / dl}$ son los límites de hiper- e hipoglucemia, respectivamente. El perfil glucémico del día k puede describirse por dos características relacionadas con el estado hiper- e hipoglucémico de ese día y más específicamente con la hipoglucemia diaria promedio y el error de hiperglucemia (Fig. 4):

$$x_k^1 = \frac{1}{N_1} \sum_{t \in \text{dia } k} H(EG(t)) \quad (17a)$$

$$x_k^2 = \frac{1}{N_2} \sum_{t \in \text{dia } k} H(-EG(t)) \quad (17b)$$

- 30 en donde $H(\cdot)$ es la función del Heaviside y N_i es el número de muestras de tiempo por encima del umbral de hiperglucemia ($i = 1$) o por debajo del umbral de hipoglucemia ($i = 2$). En primer lugar, las características se normalizan en $[0, 1]$. Las características normalizadas formulan el estado del día k .

Diseño del crítico

$$x_k = \begin{bmatrix} x_k^1 & x_k^2 \end{bmatrix}^T$$

- 35 La formulación matemática del crítico se ha proporcionado anteriormente. El CGM mide la glucosa en sangre del paciente durante el periodo de tiempo (por ejemplo 24 h) y esas mediciones definen el perfil de glucosa del periodo de tiempo. Todos o una parte de esos datos pueden registrarse en un dispositivo de memoria. Al final del día k (se entiende que el día es un ejemplo del período de tiempo), se recoge el perfil de glucosa del día (y puede recuperarse a partir del dispositivo de memoria) y el estado x_k puede calcularse, por ejemplo, teniendo en cuenta las mediciones de la glucemia o el perfil de glucemia. Basándose en el estado, se puede asignar un coste local $c(x_k)$ (Fig. 4), definido como:

$$40 \quad c(x_k) = a_h x_k^1 + a_l x_k^2 \quad (18)$$

- 45 Los pesos a_h y a_l se usan para ajustar los componentes de hipoglucemia e hiperglucemia. a_h y a_l pueden estar comprendidos entre 0 y 50, por ejemplo $a_h = 1$ y $a_l = 10$. La función de acción-valor se aproxima linealmente como se ha descrito en (12). Las funciones básicas $\phi(\cdot)$ pueden establecerse igual a la derivada de la relación de probabilidad (LRD) de la política de control que se derivará en una fase posterior. Para la actualización del crítico, las constantes γ y λ pueden estar comprendidas cada una entre 0 y 1, preferentemente $\gamma = 0,9$ y $\lambda = 0,5$ para todos los pacientes. La tasa de aprendizaje del crítico puede estar comprendida entre 0 y 1, preferentemente $\alpha_k^c = 0,5$ para todos los pacientes.

Estos valores se encontraron experimentalmente. Los parámetros iniciales r_0 pueden fijarse para valores aleatorios en $[-1 \ 1]$ y los parámetros iniciales z_0 para valores cero para todos los pacientes.

Diseño del actor

5 El actor implementa una política de control estocástica doble $\mu(u_k | x_k, \theta_k)$ para la optimización diaria de BR y la relación IC a partir de un valor inicial de BR (relación IC) (Fig. 4). Para disociar la acción del nivel absoluto del régimen de insulina actual, la acción de control u_k puede definirse como la tasa de cambio de BR (relación IC) desde el día $k-1$ al día k . El beneficio de esta elección se revelará más adelante. Por lo tanto, la BR (relación IC) puede actualizarse de la siguiente manera:

$$S_k = S_{k-1} + P_k^S S_{k-1} \quad (19)$$

$$10 \quad P_k^S = u_k^S \sim \mu(u_k^S | x_k, \theta_k^S) \quad (20)$$

en donde $S = \{BR, IC\}$, S_k es la política de control para el día k y P_k^S es la acción de control, es decir, la tasa de cambio S_k desde el día $k-1$ hasta el día k . La acción de control final aplicada P_k^S se extrae de la distribución de probabilidad $\mu(u_k^S | x_k, \theta_k^S)$ de la política de control basada en el estado actual x_k y del vector de parámetros de política θ_k^S . Preferentemente, para el diseño de la distribución de probabilidad, se sigue un procedimiento de tres etapas, basado en la generación de tres tipos diferentes de acciones de control:

- i) determinista lineal,
- ii) supervisora y
- iii) acción exploratoria.

20 Sin embargo, para el diseño de la distribución de probabilidad solo se pueden utilizar uno o dos de esos tipos de acciones de control. A continuación, se indican las notaciones k y S se omiten por motivos de claridad. El procedimiento puede ser el mismo para BR y la relación IC.

Acción de control determinista lineal:

La acción de control determinista lineal P_a se define como la combinación lineal del estado actual y el vector de parámetros de política:

$$25 \quad P_a = x^T \theta \quad (21)$$

En otras palabras, esa acción de control asocia el estado hipoglucémico e hiperglucémico diario con el porcentaje necesario de cambio de BR (relación IC) para el día siguiente.

Acción de control supervisor:

30 La acción de control supervisora P_s es una información conservadora basada en reglas para el algoritmo y sirve principalmente como guía de la dirección del cambio a seguir y como módulo de seguridad frente a cambios extremos de la insulina mediante el algoritmo. La acción supervisora se define como:

$$P_s = \begin{cases} 0 & \text{si } x_1 = x_2 = 0 \\ \pm 0.1x_1 & \text{si } x_1 > 0 \text{ y } x_2 = 0 \\ \mp 0.1x_2 & \text{si } x_2 > 0 \end{cases} \quad (22)$$

en donde el signo superior se refiere a BR y el signo inferior a la relación IC.

35 Preferentemente, la suma ponderada de las dos acciones previas puede definir la acción de control determinista total P_d :

$$P_d = hP_a + (1-h)P_s \quad (23)$$

en donde h es un factor que nos permite ajustar la contribución de cada parte con la salida final. h puede estar comprendido entre 0 y 1. Por ejemplo, en este documento, el factor de ponderación que se ha elegido es $h = 0,5$ y asigna por tanto contribuciones iguales a las dos acciones.

40 Acción de control exploratoria:

La acción de control exploratoria P_e puede producirse añadiendo ruido blanco a la política determinista final como se indica a continuación:

$$P_e = P_d + N(0, \sigma) \quad (24)$$

en donde $N(0, \sigma)$ puede ser ruido blanco gaussiano con media de cero y desviación estándar σ . El objetivo del procedimiento de exploración es ampliar el espacio de búsqueda del algoritmo para optimizar el rendimiento y la tasa de convergencia. El resultado del procedimiento de exploración es la acción de control final que se va a aplicar.

- 5 Basándose en el análisis previo, ahora estamos listos para derivar la política de control $\mu(u | x, \theta)$ como la distribución de probabilidad a partir de la cual se extrae la acción de control final $u = P_e$:

$$\mu(u | x, \theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{u - P_d(x)}{\sigma}\right)^2\right) \quad (25)$$

La política de control es una distribución de probabilidad gaussiana con una media igual a la acción determinista total $P_d(x)$ y una desviación estándar σ . Finalmente, la LRD $\psi_\theta(x, u)$ tiene que derivarse. Tomando el gradiente de la política de control con respecto a θ , tenemos

$$\nabla_\theta \mu(u | x, \theta) = \mu(u | x, \theta) \frac{u - P_d(x)}{\sigma^2} \nabla_\theta P_d(x) \quad (26)$$

A partir de (25, 26), la LRD se convierte en:

$$\psi_\theta(x, u) = \nabla_\theta \ln \mu(u | x, \theta) = \frac{u - P_d(x)}{\sigma^2} \nabla_\theta P_d(x) \quad (27)$$

y la actualización del parámetro de política del actor se define como sigue:

$$\theta_{t+1} = \theta_t - \beta_k d_t \psi_\theta(x_t, \mu_\theta(u_t | x_t)) = \theta_t - \beta_k d_t \frac{u - P_d(x)}{\sigma^2} \nabla_\theta P_d(x) \quad (28)$$

En (28) se puede ver que la actualización del vector de parámetros de política depende de la diferencia entre la política determinista total y la exploratoria, es decir, de la varianza de ruido σ^2 . Cuando se ha encontrado una política óptima, que da como resultado un estado $x_k \sim 0$, querríamos reducir la exploración, ya que esto puede conducir al sistema alejándolo de la solución encontrada. Para ello, la varianza σ^2 se define en función del estado x_k :

$$\sigma^2 = KS \|x_k\|^2 \quad (29)$$

Cuanto mayor es el estado x_k , mayor es el tiempo empleado en hipoglucemia/hiperglucemia el día k , es decir, cuanto mayor es el espacio de exploración para una mejor política de control. La constante KS puede estar comprendida entre 0 y 1, preferentemente se establece en 0,05 (después de un procedimiento de prueba y error). La tasa de aprendizaje del actor β_k puede establecerse igual a la varianza σ^2 utilizando el mismo razonamiento. De esta manera, el algoritmo AC puede ser un aprendizaje en todo momento, para compensar los cambios temporales o permanentes en el sistema glucorregulador de cada paciente.

Ajuste personalizado del algoritmo AC

El diseño del algoritmo AC, como se ha descrito en la sección anterior, implica varios parámetros que tienen que ajustarse (Fig. 5). Teniendo en cuenta la variabilidad entre pacientes, podría requerirse un ajuste personalizado para algunos de los parámetros. El ajuste manual para cada paciente es inviable o podría comprometer la seguridad de los pacientes, por lo que es necesario investigar métodos automáticos.

Basándose en simulaciones preliminares y bajo diferentes configuraciones de ajuste, los parámetros de AC pueden dividirse primero en dos clases, tales como fuertes (R) o sensibles (S):

- Los parámetros incluidos en la clase R estaban asociados con una baja sensibilidad frente a la variabilidad del paciente y pueden ajustarse manualmente mediante métodos empíricos, con valores comunes para todos los pacientes administrados en la sección anterior.
- La clase S incluía los parámetros que se encontró que eran sensibles a características específicas del paciente. Los parámetros identificados en esa clase pueden ser los valores iniciales de la BR y la relación IC y el vector de parámetros de política inicial del actor θ_0 . Para los dos primeros parámetros, no es posible un ajuste universal, ya que los requisitos de insulina difieren naturalmente entre diferentes pacientes diabéticos. Se mostrará que el vector de parámetros de política θ puede estar estrechamente relacionado con características específicas del paciente y su ajuste inicial afecta tanto al rendimiento como a la tasa de convergencia del algoritmo. Por lo tanto, se pueden seguir procedimientos de ajuste individualizados automáticos para los parámetros de la clase S. La Tabla 1 resume un ejemplo de parámetros del algoritmo AC junto con su descripción, valores y clase de ajuste.

Tabla 1: Ejemplo de parámetros del algoritmo AC

Parámetro	Descripción	Valor	Clase
α_h	Peso por hiperglucemia de coste local	1	R
α_l	Peso por hipoglucemia de coste local	1	R
γ	Coste a largo plazo del factor de descuento	0,9	R
λ	Constante de aprendizaje de TD	0,5	R
α_k^c	Tasa de aprendizaje del crítico	0,5	R
α_k^a	Tasa de aprendizaje del actor	1	R
r_0	Vector de parámetros iniciales del crítico	Aleatorio en [-1 1]	R
z_0	Vector de elegibilidad inicial del crítico	Cero	R
σ	Acción exploratoria de desviación estándar	0,05	R
S_0	Relación inicial BR/IC del actor	Específico del paciente	S
θ_0	Vector de parámetros iniciales del actor	Específico del paciente	S

Inicialización de BR y relación IC

Para garantizar la seguridad, los valores iniciales para la BR y la relación IC deben ser específicos y apropiados para cada paciente. La experiencia clínica en el tratamiento de la diabetes ha desarrollado diversas reglas empíricas para la estimación de los perfiles de BR y las relaciones IC para pacientes bajo terapia de bombeo con CSII, en base a su peso corporal, SI y factores de estilo de vida. Esas reglas proporcionan un régimen de insulina en bucle abierto que puede no ser óptimo pero asegura una regulación primaria de la glucosa. Por lo tanto, cuando se aplica en la práctica clínica, la BR y la relación IC del algoritmo AC se pueden inicializar usando los valores individuales del paciente optimizados por su médico (Fig. 5). Esta práctica tiene la ventaja adicional de que la transición de un paciente desde CSII a AP puede ser más suave tanto para él/ella mismo/a como para el médico.

Inicialización del vector de parámetros de política θ

La inicialización del vector de parámetros de política θ puede basarse en la investigación de su representación natural dentro del algoritmo de control de infusión de insulina diseñado (Fig. 5). Los valores óptimos del vector de parámetros de política θ responden a la pregunta: "¿cuánto debería cambiar la BR y la relación IC basándose en la hiper/hipoglucemia diaria observada?" La respuesta puede estar relacionada con el SI del paciente y depende de su índice de masa corporal (BMI, por sus siglas en inglés), las necesidades diarias totales de insulina (TDI), el estilo de vida y factores genéticos.

La estimación del SI se realiza actualmente en un entorno clínico usando ensayos de tolerancia a la glucosa por pinzamiento o intravenoso, que requieren tiempo y son costosos. En los últimos años, se han realizado esfuerzos para lograr una estimación en línea del SI que se incorporará en los algoritmos AP, usando datos de CGM y de la bomba de insulina y basándose en la solución inversa de un modelo fisiológico de diabetes.

A menudo en la práctica, el SI está directamente relacionado con la TDI de un paciente, ya que esa información es fácilmente accesible. Sin embargo, incluso para dos pacientes con el mismo TDI y BMI, el impacto de 1 U de insulina puede ser diferente. En este estudio, captamos esa diferencia a través del IT de las señales de insulina a glucosa. El IT de insulina a glucosa puede medirse usando la noción de entropía de transferencia (TE), un método muy potente para la estimación de IT en procesos aleatorios no lineales. La TE estima el IT a partir de una señal de causa Y (insulina) para una señal de efecto X (glucosa). Este valor es independiente de la magnitud de las dos señales, es decir, la cantidad de insulina y la concentración de glucosa. Para dos pacientes con la misma TE, una TDI mayor corresponde a un SI menor. De manera similar, si dos pacientes tienen la misma TDI, una TE más alta puede traducirse a un SI más bajo. Siguiendo este razonamiento, la información sobre el SI de un paciente puede estimarse como:

$$\hat{SI} = c_1 \frac{TE}{TDI} \quad (30)$$

en donde c_1 es una constante positiva. Dada la definición de SI, si un paciente desea reducir sus niveles de glucosa mediante ΔG , la cantidad necesaria de insulina debe ser:

$$I_{\Delta G} = \frac{\Delta G}{SI} \quad (31)$$

Sustituyendo SI por su estimación $\hat{S}$ / proporcionada en (30), tenemos:

$$I_{sc} = c \frac{\Delta G}{TE} TDI \quad (32)$$

en donde $c = 1 / c_1$.

- 5 En el caso del algoritmo AC, el objetivo es encontrar el cambio óptimo en la BR y la relación IC para eliminar la hipo- e hiperglucemia diaria. Esto puede verse como un paralelo a (32):

$$\Delta S^i = c' \frac{x^i}{TE} TDI \quad (33)$$

- 10 en donde x_i es la característica de hiperglucemia ($i=1$) o hipoglucemia ($i=2$), es decir, el error hipo/hiperglucémico diario promedio como se ha definido en (29a, b), ΔS^i es el cambio en la BR y la relación IC basado en la característica respectiva y c' es una constante positiva. Considerando que la TDI se refleja directamente en la BR diaria y la relación IC, (33) puede reescribirse como:

$$\Delta S^i = c' \frac{x^i}{TE} S \quad (34)$$

Si se establece, $\theta^i = \frac{c'}{TE}$, (34) se convierte en:

$$\Delta S^i = \theta^i x^i S \quad (35)$$

- 15 y el cambio total en la BR o la relación IC basado en las características tanto de hipo como de hiperglucemia, es la combinación lineal de sus contribuciones respectivas como:

$$\Delta S = \Delta S^1 + \Delta S^2 = \theta^1 x^1 S + \theta^2 x^2 S = (\theta^1 x^1 + \theta^2 x^2) S = \theta^T x S \quad (36)$$

en donde $\theta = [\theta_1 \ \theta_2]^T$ y $x = [x_1 \ x_2]^T$ es el vector de las características. Finalmente, si se establece que $P_S = \theta^T x$, entonces (35) se convierte en:

$$\Delta S = P^S S \quad (37)$$

- 20 en donde P^S es el porcentaje de cambio de S y representa la acción de control determinista de AC como se ha definido anteriormente en (33).

- 25 El análisis mencionado anteriormente ilustra que definir la acción de control como la tasa de cambio de insulina, permitía el ajuste de AC, usando la IT de insulina a glucosa y sin la necesidad de estimar el SI, que sería un proceso más engorroso. El análisis puede ser aproximado y puede (solo) usarse como una estimación de borrador de la actualización de BR o IC necesaria. Sin embargo, el alcance es proporcionar un punto de partida mejor para AC con el fin de mejorar el proceso de optimización. Los valores iniciales del vector de parámetros de política para un paciente p puede establecerse como:

$$\theta_i^S(p) = \left[\frac{W_h}{TE(p)} \ \frac{W_l}{TE(p)} \right] \quad (38)$$

- 30 en donde W_h y W_l son pesos relacionados con las características de hiper- e hipoglucemia y pueden estar comprendidos entre -1 y 1, preferentemente $W_h = 0,1$ y $W_l = -0,2$ para todos los pacientes. De nuevo, se asigna un valor más alto al peso hipoglucémico, ya que evitar la hipoglucemia tiene mayor prioridad.

Estimación de la TE de insulina a glucosa

- 35 La TE de insulina a glucosa puede estimarse basándose en los datos de CGM y de la bomba de insulina durante cuatro días (por ejemplo) recogidos de cada paciente. Para elegir el tamaño de datos apropiado, se usaron conjuntos de datos de diferentes duraciones y se calculó la correlación entre los valores de TE respectivos para longitudes de datos sucesivas. Se observó que los datos de cuatro días o más proporcionaban valores de TE altamente correlacionados (>99%).

La estimación de TE puede basarse en la siguiente fórmula:

$$TE_{t,t+d} = \sum_i p(G_t, G_{t+d}, IA_{t+d}) \log \frac{p(G_t | G_{t+d}, IA_{t+d})}{p(G_t | G_{t+d})} \quad (39)$$

en donde G_t , IA_t son la glucosa y la insulina activa en el momento t y d es el retardo temporal de insulina establecido en este caso como $d = 20$ minutos, según el retardo promedio de la absorción de insulina fisiológica para análogos de insulina de acción rápida. La insulina activa puede estimarse como la suma de insulina a bordo (*IOB*, por sus siglas en inglés) relacionada con las dosis de bolo y la infusión de insulina basal:

$$IA(t) = IOB(t) + I_{basal}(t) \quad (40)$$

Para la estimación de las distribuciones de probabilidad, se puede usar el método de partición fija de datos, en donde las series temporales se pueden dividir en intervalos de tamaño igual y las distribuciones de probabilidad se aproximan como histogramas. El tamaño de los intervalos de partición para la glucosa y la insulina puede ser $G_{bin} = 10$ mg/dl y $IA_{bin} = 1$ U, respectivamente.

Conceptos generales de la invención

Según un aspecto tal y como se ha descrito anteriormente, la invención proporciona un método para estimar una terapia con insulina que comprende al menos uno entre la tasa de infusión de insulina (basal) y la relación de insulina a carbohidrato (dosis en bolo = carbohidratos contenidos en la próxima comida x relación de insulina a carbohidrato) adaptada a un paciente en base a uno de los siguientes puntos:

- controlador adaptativo que utiliza un aprendizaje de tipo actor-crítico;
- política de control de insulina evaluada (crítico) y actualizada (actor) una vez al día usando los perfiles de glucosa/insulina de las últimas 24 horas;
- el estado del sistema regulador de glucosa, tal y como se describe por dos características relacionadas con el estado hiper- e hipoglucémico de ese día y más específicamente por el error promedio diario de hipoglucemia y/o hiperglucemia;
- la evaluación de la política de control (crítico) al final de cada día a partir del cálculo del estado del sistema regulador de glucosa y del coste local asignado correspondiente, preferentemente utilizando el aprendizaje por diferencia temporal; y
- la actualización de la política de control (actor), la tasa de cambio de insulina de un primer período de tiempo (en el ejemplo un día) para un período de tiempo posterior, basándose en los efectos de uno, dos o tres tipos diferentes de acciones de control, i) la acción determinista, que asocia el estado hipoglucémico e hiperglucémico diario al porcentaje necesario de tasa basal y el cambio de la relación de insulina a carbohidrato para el día siguiente, ii) la acción supervisora, que se corresponde con la sugerencia conservadora de la política de control y actúa como seguridad y guía para el sistema, y iii) la acción exploratoria, que amplía el espacio de búsqueda del estado del sistema y apoya la convergencia más rápida y mejor; o
- el alcance del actor es minimizar el coste esperado promedio para cada estado del sistema (optimización de la política de control)

El método puede incluir dos etapas sucesivas:

- etapa de inicialización, en la que se inicializa el método;
- etapa de aprendizaje, en la que el método adaptativo aprende del paciente (política de control específica del paciente).

Preferentemente, la etapa de aprendizaje es continua y se ejecuta en cada período de tiempo después de la etapa de inicialización.

La tasa inicial basal de insulina del actor, la relación inicial de insulina a carbohidrato y/o el vector de parámetros de política pueden ser específicos del paciente.

La tasa basal de insulina y la relación de insulina a carbohidrato de la AC se pueden inicializar usando los valores individuales del paciente, optimizados por su médico y/o un simulador *in silico*.

El vector del parámetros de política del controlador de AC puede inicializarse usando la entropía de transferencia de insulina a glucosa estimada, basándose en datos de la glucosa y la bomba de insulina durante un corto período de tiempo, por ejemplo, cuatro días, recogidos de cada paciente.

La insulina puede ser un análogo de insulina.

El método puede llevarse a cabo antes/durante/después de administrar la insulina al paciente.

El método puede recibir datos en tiempo real de interfaces de comunicación inalámbricas/cableadas desde un monitor de glucosa y una bomba de insulina.

La comida puede no estar anunciada o ser introducida manualmente por el paciente.

- 5 La comida puede anunciarse al controlador a través de la interfaz con otras aplicaciones o sensores móviles/inteligentes.

La tasa basal estimada y la relación de insulina a carbohidrato se pueden usar para controlar una bomba de insulina o una bomba de parche de insulina directamente o a través de la unidad de control de la bomba.

- 10 En un aspecto adicional, la invención proporciona un dispositivo para informar al paciente sobre la tasa basal estimada y la relación de insulina a carbohidrato. El dispositivo puede ser cualquier dispositivo portátil/móvil o teléfono inteligente con un procesador configurado para ejecutar el método descrito anteriormente en el presente documento.

El dispositivo puede incluir todas las pantallas e interfaces necesarias para informar al paciente sobre la terapia óptima con insulina y transferir los datos de forma cableada/inalámbrica a una nube/servidor u otros dispositivos relacionados con la tecnología de la diabetes.

- 15 El dispositivo puede incluir interfaces con otras aplicaciones móviles (por ejemplo, libro electrónico de diabetes, aplicaciones para entrenamiento físico/alimentos, etc.).

El dispositivo puede incluir todas las pantallas e interfaces necesarias para informar al paciente sobre la terapia óptima con insulina y transferir los datos de forma cableada/inalámbrica a una nube/servidor u otros dispositivos relacionados con la tecnología de la diabetes.

REIVINDICACIONES

1. Un método para determinar una terapia con insulina para un paciente mediante un dispositivo que tiene un procesador, comprendiendo el método:

- 5 • usar, mediante el procesador, un algoritmo de aprendizaje por refuerzo para hacer funcionar un controlador adaptativo del dispositivo, comprendiendo el algoritmo de aprendizaje por refuerzo:
 - a. un crítico que evalúa una política de control de insulina (S) que incluye al menos uno entre una tasa de infusión de insulina y una relación de insulina a carbohidrato, calculando un coste relacionado con un estado del sistema regulador de glucosa del paciente para un periodo de tiempo predeterminado, y
 - 10 b. un actor que mejora la política de control de insulina (S) para minimizar el coste esperado relacionado con un estado posterior del sistema regulador de glucosa;
- aplicar, durante un periodo de tiempo predeterminado (k-1), una política de control de insulina predeterminada (S_{k-1});
- recopilar, durante el período de tiempo predeterminado (k-1), datos de la glucosa en sangre del paciente;
- 15 • determinar un estado del sistema regulador de glucosa basado en los datos de glucosa en sangre recogidos;
- evaluar a través del crítico la política de control de insulina predeterminada (S_{k-1}) basándose en el estado del sistema regulador de glucosa;
- actualizar por medio del actor, la política de control de insulina (S), basándose en una acción de control determinada, en donde la política de control de insulina actualizada (S_k) está previsto que se aplique durante un periodo de tiempo posterior (k);
- 20

en donde la acción de control determinada tiene en cuenta al menos dos de las siguientes acciones de control:

- 25 - una acción de control determinista lineal (P_a) que actualiza la política de control de insulina (S) por medio de un modelo matemático que tiene en cuenta una combinación lineal del estado del sistema regulador de glucosa y un vector de parámetros de política determinado para asociar el estado del sistema regulador de glucosa al cambio necesario para el periodo de tiempo posterior (k);
- una acción de control supervisora (P_s) que actualiza la política de control de insulina (S) mediante un modelo matemático que tiene en cuenta una política de control de insulina previa para actuar como módulo de seguridad frente a cambios extremos de la política de control de insulina (S) a través del algoritmo y/o como guía de la dirección de la actualización, y
- 30 - una acción de control exploratoria (P_e) que actualiza la política de control de insulina (s) mediante un modelo matemático que toma un ruido blanco para ampliar el espacio de búsqueda del algoritmo del estado del sistema regulador de glucosa con el fin de optimizar un rendimiento y una tasa de convergencia.

2. Método según la reivindicación 1, en donde la acción de control determinada se define como una tasa de cambio de la política de control de insulina (S) desde el período de tiempo predeterminado (k-1) hasta el período de tiempo posterior (k).

3. Método según la reivindicación 1, en donde la acción de control determinada tiene en cuenta una acción de control determinista total (P_d) que se determina por medio de un modelo matemático que tiene en cuenta la acción de control determinista lineal (P_a) y la acción de control supervisora (P_s).

4. Método según la reivindicación 3, en donde la acción de control exploratoria (P_e) se determina por medio de un modelo matemático que tiene en cuenta al menos la acción de control determinista total (P_d).

5. Método según una cualquiera de las reivindicaciones precedentes, en donde el crítico evalúa el estado del sistema regulador de glucosa tal como se describe mediante dos características relacionadas con el estado hiperglucémico e hipoglucémico ocurrido durante el periodo de tiempo predeterminado (k-1).

6. Método según una cualquiera de las reivindicaciones precedentes, en donde el crítico evalúa el estado del sistema regulador de glucosa tal como se describe mediante el error promedio diario de hipoglucemia e hiperglucemia ocurrido durante el periodo de tiempo predeterminado (k-1).

7. Método según una cualquiera de las reivindicaciones precedentes, en donde el crítico calcula además un coste local del estado del sistema regulador de glucosa durante el periodo de tiempo predeterminado.

8. Método según la reivindicación 7, en donde el actor está configurado para minimizar el coste promedio esperado para cada estado.

9. Método según una cualquiera de las reivindicaciones precedentes, en donde la etapa de evaluación se realiza al final de cada periodo de tiempo.
10. Método según una cualquiera de las reivindicaciones precedentes 3 a 4, en donde el vector de parámetros de política depende de la diferencia entre la acción de control determinista total P_d y la acción de control exploratoria P_e .
- 5 11. Método según una cualquiera de las reivindicaciones precedentes, en donde el vector de parámetros de política se actualiza al final del periodo de tiempo determinado.
12. Método según una cualquiera de las reivindicaciones precedentes, en donde la política de control de insulina inicial se determina teniendo en cuenta los valores individuales del paciente, optimizados por su médico y/o un simulador *in silico*.
- 10 13. Método según una cualquiera de las reivindicaciones precedentes, en donde el valor o el valor inicial del vector de parámetros de política se determina usando un cálculo de la entropía de transferencia de insulina a glucosa.
14. Método según la reivindicación 13, en donde la entropía de transferencia de insulina a glucosa se estima basándose en los datos de glucosa en sangre y los datos de administración de insulina durante un periodo de tiempo comprendido entre 1 y 7 días, preferentemente 4 días.
- 15 15. Método según una cualquiera de las reivindicaciones precedentes, que comprende además una fase de inicialización durante la cual se determinan y se adaptan al paciente una política de control de insulina inicial y/o un valor inicial del vector de parámetros de política.
16. Método según la reivindicación 15, en donde la duración de la fase de inicialización está comprendida entre 1 y 7 días, preferentemente 4 días.
- 20 17. Método según una cualquiera de las reivindicaciones precedentes, en donde la duración del período de tiempo predeterminado y/o del período de tiempo posterior está comprendida entre 1 hora y 36 horas, preferentemente es sustancialmente igual a 24 horas.

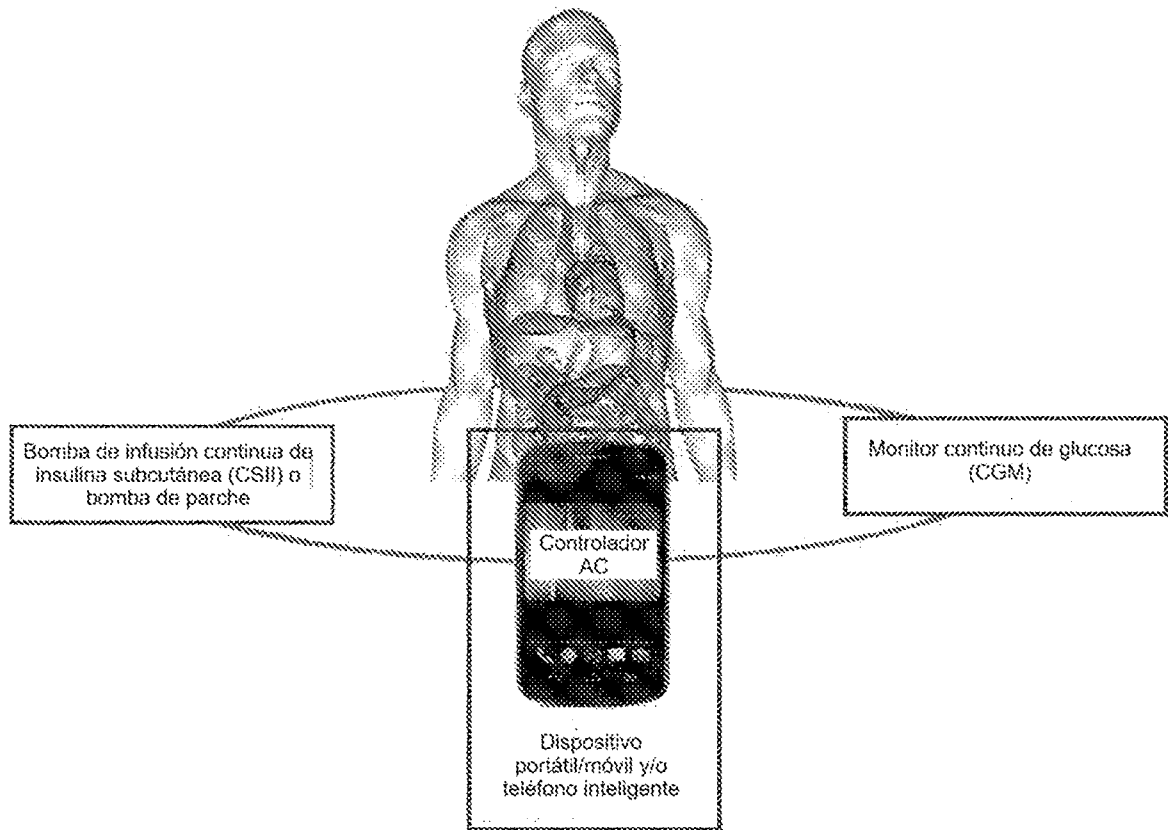


Fig. 1

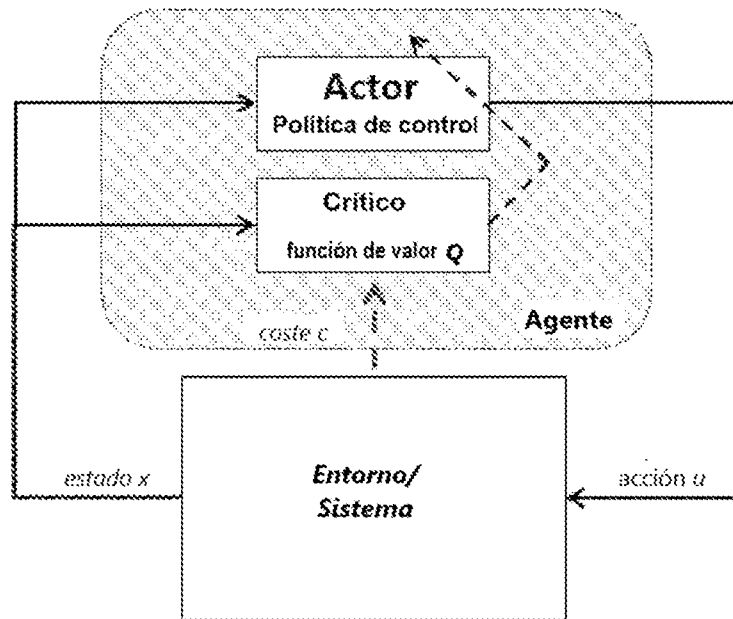


Fig. 2 (Técnica Anterior)

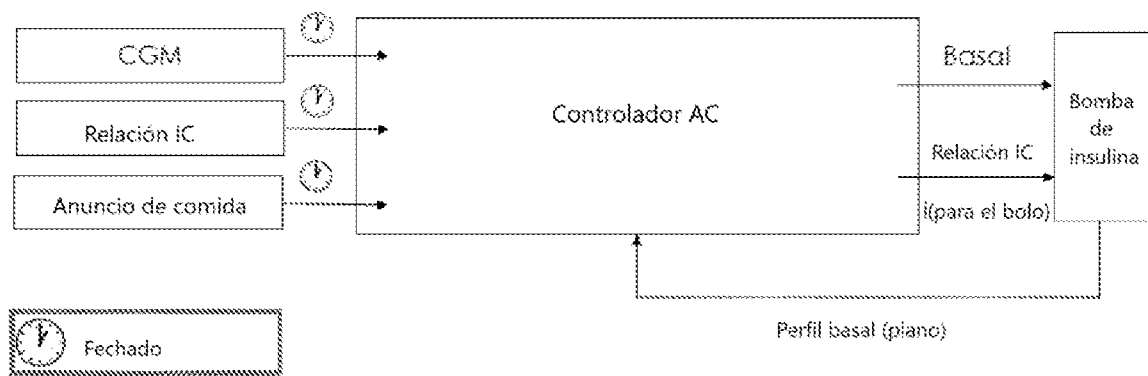


Fig. 3

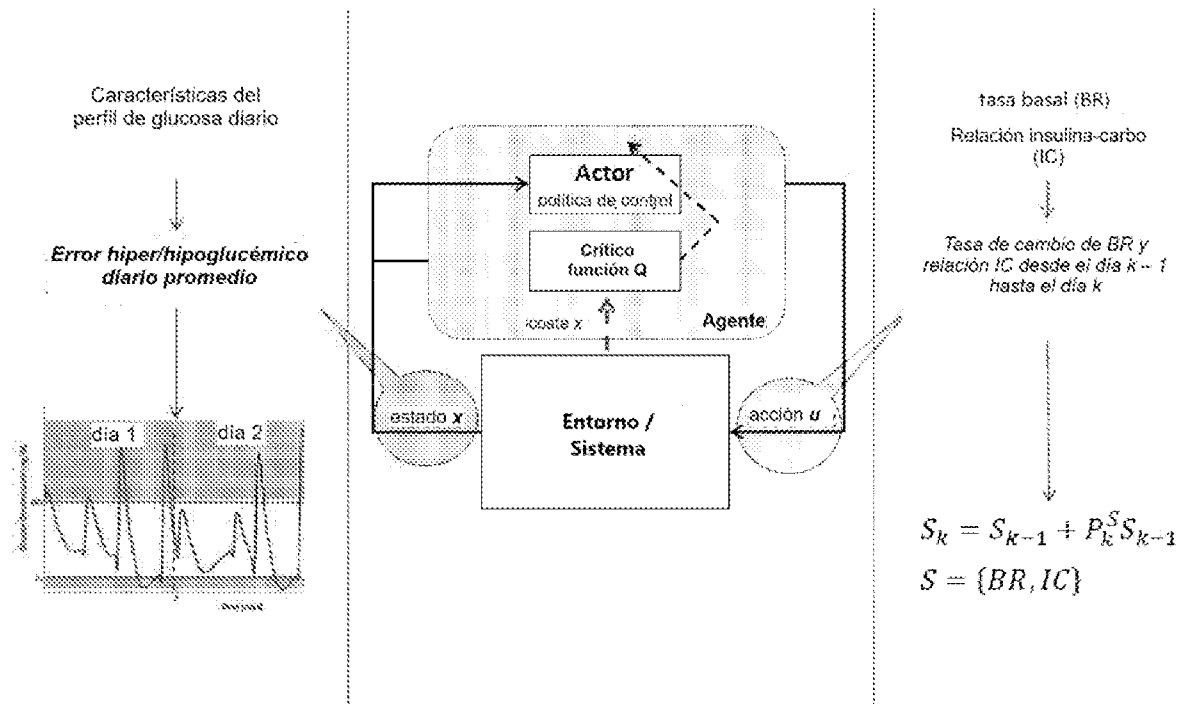


Fig. 4

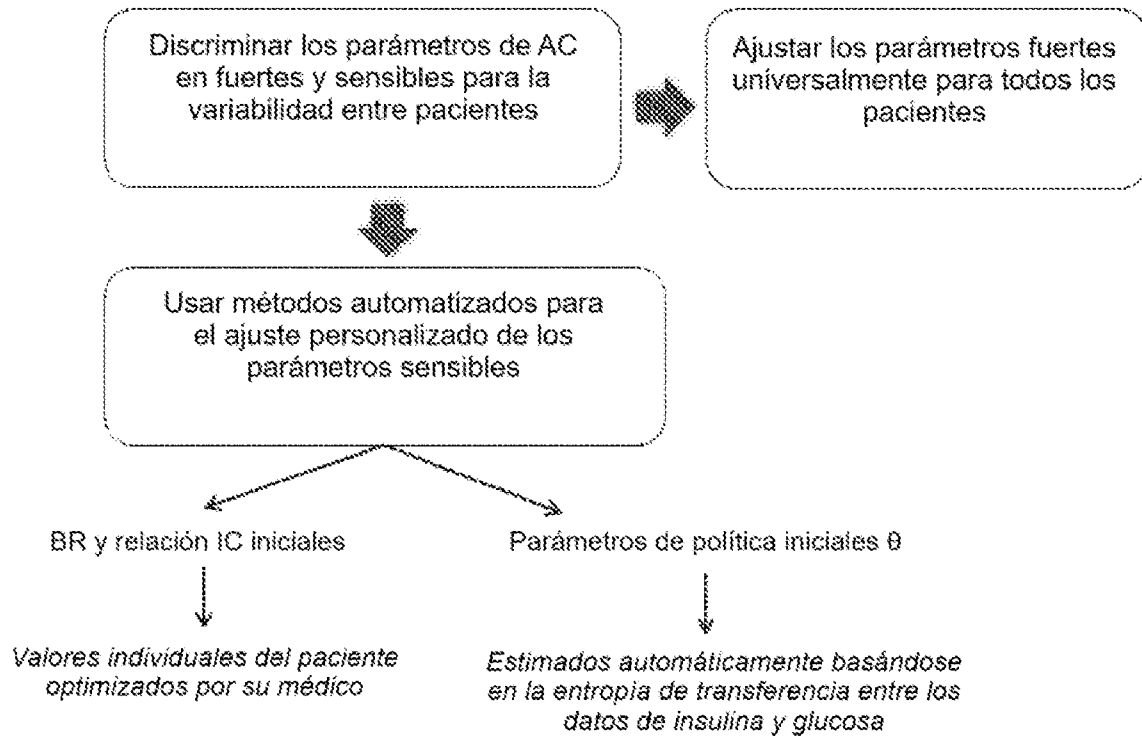


Fig. 5